Small decisions, mostly separate parts: how small language models indent code, choose ordinal suffixes and write in capitals

Richard Jeevash
Independent researcher
jeevash@richardjeevash.com
ORCID 0009-0005-7812-0939

November 2026

Abstract

Small language models make many small formatting and morphology decisions correctly: how far to indent the next line of Python, whether a number takes *st* or *th*, whether the next word continues a passage written in capitals. We asked how they make three of these decisions, and then whether the mechanisms share parts. Each study explored under a frozen charter, then wrote down what the explored account predicts and tested those predictions, with thresholds and layer-matched random controls, once on held-out material, so the tests show whether each account repeats on new material. *Indentation* (Qwen3-0.6B-Base and OLMo-2-0425-1B): dedents reach the model by two routes. After `return` and similar statements the keyword is read directly from the decision token through many heads. Unmarked dedents depend on a small set of late heads, chief among them one that attends to the line-ending token of an open block's header and copies that header's indent; which open header it attends to sets how far to dedent, and steering its attention moves the prediction. *Ordinal suffixes* (Qwen3-0.6B-Base, OLMo-2-0425-1B and five further models): to write "112th", Qwen3-0.6B-Base uses an early previous-token head to bring the tens digit to the last digit; the first MLP reads it, and a late MLP writes a teen override towards *th* that does not consult the last digit, while a different MLP writes the last digit's own suffix. In OLMo-2-0425-1B, where the tokeniser can put the tens digit in another token, the signal arrives too faint, and a thousands separator repairs the errors; early previous-token heads carry part of the check in three further models. In the ordinal study's second round two tests failed, the two number lengths of one prediction. *Capitals* (Qwen3-1.7B-Base and Gemma-4-E2B): the next word is chosen in a lower-case-like form and turned into capitals in the last layers along one direction, which is written by a few late components and fed by heads that carry the passage's case; the switch to capitals needs the MLPs after it. A prospective cross-study test (line-breaking, indentation and ordinals in Qwen3-0.6B-Base; capitals in Qwen3-1.7B-Base, with heads matched by index) finds less sharing than earlier reports of component reuse. The early heads the studies named are general movers of the previous or the current token. Of the later heads, two have a small second role: L17H6 is shared between line-breaking and indentation, and L11H9 serves both the ordinal suffix choice and the dedent after a terminal statement, a role first noticed after the cross-study test and then confirmed by a frozen test on new material. The groups that do most of each task's work harm the other tasks no more than random groups do. A third round re-ran every code-based test of the indentation and cross-study work on permissively licensed and newly generated code, which can be redistributed, and all of them held again.

1 Introduction

Much of what a language model writes is decided by small rules of form. A model writing Python must choose how far to indent each line before it sees the next keyword. A model writing an ordinal must apply the exception that makes "11th" and "112th" rather than "11st" and "112nd". A model continuing a passage in capitals must keep the capitals and still choose the right word, although its vocabulary spells upper-case words differently. Small models get all three mostly right. This paper asks how, in small open-weight base models, and then asks a second question that only a set of such studies can answer: do these mechanisms share parts?

Each behaviour has close relatives in earlier work, and we state our claims against them (Section 7).

- *Indentation.* Indentation has been studied as a representation [3], in models trained for the purpose [16], and in recurrent networks and very small transformers [26, 28]. A late head that attends to the still-open bracket is known from parenthesis closing in a pretrained code model [46], and attention to the top of a stack of open brackets from models of bracket languages [11, 48, 58]. To our knowledge, no earlier work gives a causal account of which components of a pretrained language model choose the next line's indent, or of how such a model chooses how far to dedent among open blocks.
- *Ordinal suffixes.* To our knowledge no earlier work explains how a model applies the teen exception. Its parts are familiar: early layers that assemble multi-token units [13, 29], digit-wise and modular number representations [2, 20, 31], and tokenisation effects on numeracy [49].
- *Capitals.* Earlier work found that an in-context "capitalise this word" function is applied late by an additive feed-forward update along one transferable vector [36], and sparse-autoencoder studies report features that promote all-capitals tokens [8]. To our knowledge, no earlier work explains how a model keeps writing a passage in capitals: where the passage's case is carried, which components write and read it, and how it interacts with the choice of word.
- *Shared parts.* That circuits reuse components across tasks is established [30, 32, 35]. We test reuse prospectively across four behaviours whose parts were named by separate studies; the fourth is line-breaking in prose, which a companion preprint studies with the same design [27]. We find general early heads, as those reports do, but less reuse among the task-specific parts: the groups that do most of each task's work are separate, and two heads have a small role in a second behaviour, both confirmed by frozen tests; the second was first read post hoc and then tested on new material.

All four studies followed one design (Section 2). Under a frozen charter we first required behavioural competence, then explored with ablation, patching, attention analysis, the logit lens and steering until a candidate mechanism appeared. We then wrote down what the candidate predicts, especially effects nobody would predict without it, froze those predictions with their thresholds, controls and code, and tested them once on material sealed before any model read it. The indentation, ordinal and capitals studies each ran two such rounds, the second tracing the first round's account into its parts. Because the predictions were written after exploration, these tests show that each explored account repeats on held-out material, not that it predicted effects nobody had looked for; the counts of tests passed are a record of that replication, not a measure of how surprising the accounts are. We count *tests*, the units that each frozen protocol scores as pass or fail: a test is one prediction read in one model and, where a protocol reads number lengths separately, at one length, so one prediction can be several tests. Throughout, we label results *exploratory* (chose what to test), *prospective* (tested once against frozen predictions) or *post hoc* (computed after a test, and changing no verdict). Intervals, and a few values that we call descriptive, accompany prospective results and decide no test. The code of the first two indentation rounds and of the cross-study test came from public repositories, many without a licence that allows it to be redistributed. A third round therefore re-ran every test that used natural code, with the frozen thresholds and code, on permissively licensed code and newly generated programs, and

Table 1: The four studies. The round columns count tests passed, as each frozen protocol scores them, passed on held-out material and summed over models: a record of held-out replication of each explored account. A test is one frozen prediction, written after exploration, read in one model and, where the protocol reads number lengths separately, at one length. The 2 failed tests are one prediction of the ordinal study's second round at its two number lengths (Section 4.2). Round 3 re-ran the code-based tests of the first two columns on shareable material and, for shared parts, adds the six tests of L11H9's second role (Section 3.3); totals elsewhere in the text are over rounds 1 and 2.

Study	Decision and main finding	Models	Round 1	Round 2	Round 3
Indentation	next line's indent: a terminal-keyword route and a set of block-end heads, the chief of which reads open headers	Qwen3-0.6B-Base, OLMo-2-0425-1B	9 of 9	25 of 25	34 of 34
Ordinals	suffix after a number: an early copy of the tens digit, a teen override at a late MLP (path traced in Qwen3-0.6B-Base)	Qwen3-0.6B-Base, OLMo-2-0425-1B and five others	14 of 14	39 of 41	–
Capitals	case of the next word: lower case first, capitals added late along one direction	Qwen3-1.7B-Base, Gemma-4-E2B	19 of 19	23 of 23	–
Shared parts	the named heads, switched off at each other's decision positions	Qwen3-0.6B-Base (capitals: Qwen3-1.7B-Base)	22 of 22	–	15 of 15

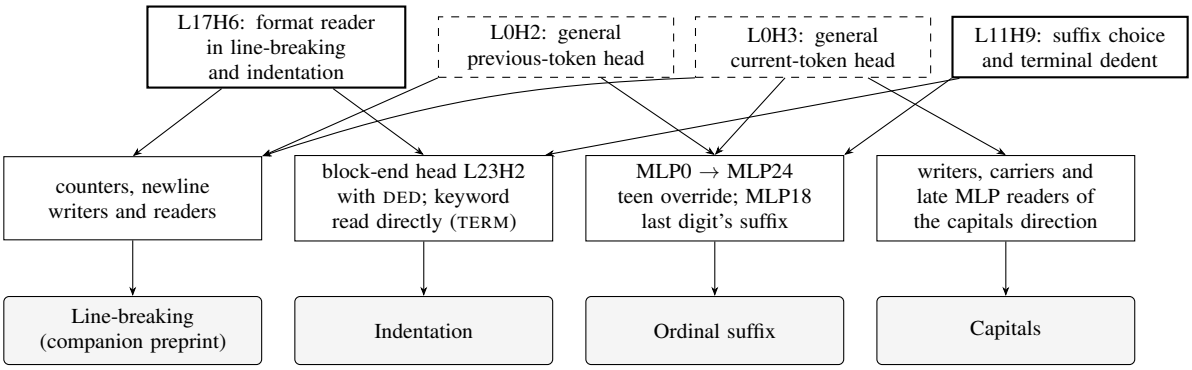


Figure 1: The parts found by the three studies (middle row) and the heads tested across studies (top). Solid frames: heads confirmed by frozen tests as shared by two behaviours; L11H9's role in indentation was first read post hoc and then confirmed on new material in round 3 (Sections 3.3 and 6). Dashed frames: the two general early movers. Arrows mark effects confirmed on held-out material in Qwen3-0.6B-Base. Capitals were tested in Qwen3-1.7B-Base with heads matched by index, and the line-breaking parts come from the companion preprint [27].

added a frozen test of a post hoc reading from the cross-study test (Section 3.3); this is the material the code release ships. Table 1 summarises the four studies; Fig. 1 shows the parts they found and where they meet.

2 Shared methods

Models. We used small base models. Seven have fewer than two billion parameters: Qwen3-0.6B-Base and Qwen3-1.7B-Base [57], OLMo-2-0425-1B [51], SmolLM2-1.7B [6], Pythia-410M and Pythia-1B [7], and GPT-2 [45]. The eighth, Gemma-4-E2B [17], has 2.3 billion "effective" parameters by its developers' count, and 5.1 billion when its per-layer embedding tables are counted. Each study's charter named its candidate

models and a behavioural gate that a model had to pass before any internal analysis, with a fallback order fixed in advance. Gates decided which models were studied: Pythia-410M failed the indentation gate on the first indented line of a chunk of code, GPT-2 failed it entirely, and Qwen3-0.6B failed the capitals gate on choosing the same word (Appendix A).

Decision positions. Each behaviour is measured at one position, and every intervention acts there unless stated. For indentation it is the token that ends the previous line; for ordinals, the number's last token; for capitals, the position that predicts the start of the next word. Measures are each study's own: the log-probability of the right indent under q, the model's next-token distribution renormalised over the candidate indents; the logit margin m between the suffix the last digit calls for and *th*, with the *teen effect* the mean m of non-teens minus that of teens and the last digit's *pull* the mean m of non-teen numbers ending in 1–3; and, for capitals, whether the top next token is upper case and whether it is the same word as in the lower-case passage.

Interventions and controls. *Mean ablation* replaces chosen heads' outputs at the decision position with their mean over decision positions of the same material [24, 59]. Every named set is compared with twenty random sets that replace each head by another head of the same layer; a set counts as special only if it beats every random set. *Activation patching* [34] copies a component's output from a source run into a target run; *path patching* [19, 54] lets the change reach one receiver only. For single heads we also edit attention directly: an *attention knock-out* removes one source position and renormalises, and *attention steering* sets a head's attention all on one source token. The *logit lens* [4, 40] reads each layer's residual stream through the final normalisation and unembedding. Directions found as differences of mean activations are added or removed to steer the model [42, 53]. Probes are regularised linear models [1].

Explore, then confirm. Each study's charter fixed the question, models, tools, budget, what would count as a candidate and when to stop, and was frozen with its digest before exploration. Held-out material was split by a hash (of repository, article, frame or number) and sealed before any model read it: Python from repositories created after every model studied was trained, which we call *post-cut-off* code, and programs from a seeded generator (indentation), hand-written sentence frames and numbers up to ten digits (ordinals), and paragraphs and headings from English Wikipedia articles created in 2025–26, after the models were trained (capitals). New repositories and pages can contain older code and text, so overlap with the training data remains possible. Each protocol froze its predictions with thresholds set from exploration values with margins, the controls and the code, and was run once. Decisions use point estimates; intervals, where given, resample repositories, programs, numbers or paragraphs and are descriptive. All runs used bfloat16, checked against float32 on exploration scans.

3 Indentation: two routes to a dedent

Python marks blocks by indentation, so after each newline a model must choose the next line's indent: deeper after a header, the same inside a block, or shallower when a block ends. A dedent is a bet on the program's structure, made before the next keyword appears. We call a dedent after `return`, `raise`, `break`, `continue` or `pass` *terminal*, and any other dedent *unmarked*. Qwen3-0.6B-Base was the main model; OLMo-2-0425-1B was the second model in the second round, chosen by a gate order fixed in advance.

3.1 Round 1: two routes

Exploration found that no single component matters much for terminal dedents, while unmarked dedents depend on a group of late heads, DED = {L23H2, L17H3, L17H4, L17H5, L18H6}, most of all on L23H2. The signal that the previous line was a terminal statement reaches the decision token from layer 5 and is mostly in place by layer 17, carried by mid-layer heads (CARRY and the related set TERM, Appendix B). Two accounts failed in exploration and were dropped. The terminal signal is not read from the previous line's first token: a name such as `return_value` moved the mass on shallower open levels to 0.109, against 0.085 for a control prefix and 0.737 after a real `return`. Nor is how far to dedent held as a linear count at the decision token: probes reached at most R^2 0.31 for the number of levels dedented.

The round-1 protocol (frozen at 06:31 on 8 November 2026) was run once on 1,606 held-out chunks of natural code and 2,008 generated programs, and all 9 tests passed (Table 7). Without DED, mean log-probability of the true indent at unmarked dedents fell by 1.09 nats (change -1.091 [-1.168, -1.015]; no random set below -0.029), and the probability of keeping the current level rose to 0.473; terminal dedents kept 0.991 of their probability on open levels. Ablating TERM instead changed terminal dedents by -0.044 and left unmarked dedents almost unchanged (-0.027). Removing both routes changed terminal dedents by -0.188, more than the -0.103 expected from the two alone, so the routes back each other up. In return-versus-assignment pairs, CARRY carried 0.650 of the gap by necessity and 0.373 by sufficiency, against at most 0.075 and 0.032 for random sets.

3.2 Round 2: how each route gets its information

The block-end head reads the open header. At unmarked dedents, L23H2 puts much of its attention on the token that ends the header of the innermost open block (in Qwen3's tokeniser a merged colon and newline, `:\n`). Knocking that one token out of its attention lowers the log-probability of the true indent, while the same knock-out in any other head of the layer barely matters; knocking out the most recent *closed* header instead does little. The head's direct output from an open header favours that header's own indent among the candidates. So the head reads the header of the block that is ending and copies its indent. What tells it to look is spread over many earlier components: the largest change in its attention to the enclosing header from ablating any single earlier head or MLP was -0.044 (exploratory).

Its attention is the choice of how far to dedent. When several blocks are open, the open header that L23H2 attends to most is, in most cases, the level the line returns to, including where that is not the innermost level. Setting the head's attention all on one open header moves the prediction to that header's indent. This answers the question round 1 left open: how far to dedent is not a count at the decision token but the indent of the header the block-end head reads.

The terminal route reads the keyword directly. Removing only the attention edge from the decision token to the keyword, in every layer, closes much of the gap between a `return` line and a neutral line, and no single layer carries much of this. The share differs by keyword and is small for `pass`.

Held-out results. The round-2 protocol (frozen at 09:55 on 9 November 2026) was run once on 2,712 new chunks from 225 repositories and 1,510 targeted programs, which open a chain of blocks and return to a randomly chosen open level. All 25 tests passed (Table 2; Tables 8 and 9 list them). In Qwen3-0.6B-Base, knocking the innermost open header out of L23H2's attention cost -0.148 nats at unmarked dedents, while no other layer-23 head changed by more than 0.044 under the same edit. Where the target level was not the

Table 2: Indentation, round 2: selected held-out results in the two models (Qwen3-0.6B-Base: head L23H2; OLMo-2-0425-1B: head L13H1 and the block-end set DED-O). Changes are in mean log-probability of the true indent. Every test passed; Appendix B gives all 25 with their thresholds. The OLMo-2-0425-1B entry in the row on all multi-header dedents is not a tested prediction: there the head does not beat the innermost rule, shown in brackets.

Measure	Qwen3-0.6B-Base	OLMo-2-0425-1B
Attention to the innermost open header, unmarked dedents (same-level lines)	0.458 (0.118)	0.290 (0.181)
Knocking that token out: change (largest absolute change, other heads of the layer)	−0.148 (0.044)	−0.132 (0.002)
Knocking out the latest closed header instead	−0.001	−0.004
Steering to another open header: rise in its indent (control head)	+2.72 (+0.08)	+1.58 (+0.15)
Most-attended header is the target, all multi-header dedents (innermost rule)	0.914 (0.822)	0.752 (0.822)
Keyword edge removed, `return`: share of the gap closed	0.716	0.895
Block-end set ablated, unmarked dedents (twenty random sets)	–	−0.775 (−0.116 to +0.100)

innermost, the most-attended open header was the target in 0.893 of cases, and over all dedents with two or more open headers the head beat the rule "return to the innermost level" (0.914 against 0.822). Steering raised the log-probability of the steered header's indent by +2.72 nats in natural code (another layer-23 head: +0.08) and by +1.41 in targeted programs (control −0.01). The keyword edge closed 0.716 of the gap for `return`, with no single layer above 0.098; for `pass` it closed −0.078.

A second model. OLMo-2-0425-1B (16 layers of 16 heads) passed the round-1 gate and has both routes. Four heads, DED-O = {L13H1, L10H12, L11H0, L15H4}, are needed for unmarked dedents and not for terminal ones (−0.775 against −0.045). L13H1, at the same relative depth as L23H2, reads open headers, and steering it moves the dedent. One difference held: its most-attended header is the target more often than the innermost header where the two differ, but over all multi-header dedents it does not beat the innermost rule (no prediction was made about this). The keyword edge carries even more of the terminal route than in Qwen3-0.6B-Base.

The OLMo-2 checkpoints come from another run (exploratory). OLMo-2-0425-1B's developers publish intermediate checkpoints of a training run of the same recipe. In those checkpoints terminal dedents are learnt first (Table 10): top-1 accuracy at terminal dedents was 0.942 by 21 billion tokens, while unmarked dedents rose slowly (0.582 at 21 billion, 0.789 by 1,007 billion). A block-end reader forms during stage 1, but at L15H9 (its most-attended header is the target at 0.82 of multi-level dedents at the end of stage 1), and all three published stage-2 runs keep it there. In the released model, by contrast, L13H1 is the reader (attention 0.31) and L15H9 is silent (0.00). A weight comparison explains this: the cosine between the released model's token embeddings and those at the end of stage 1 is 0.006, against 0.9996 between the end of stage 1 and a stage-2 run that continues from it. The released weights do not descend from the published checkpoints, so the checkpoints do not show the released model's reader forming. They show instead that two training runs of one recipe put the block-end reader in different heads with nearly the same behaviour.

3.3 Round 3: replication on shareable material

The natural code of the first two rounds, and of the cross-study test (Section 6), was cut from public repositories, many of which carry no licence that allows their code to be redistributed. So that the material

behind every code-based result can be shared, a third round re-ran those tests on new material. The new natural code is 1,000 Python files from 274 repositories created after the models were trained, each under an MIT, Apache, BSD, CC0 or Unlicense licence and pinned to a commit, which gave 2,764 chunks from 225 repositories; it was joined by 2,000 new programs from the round-1 generator and 1,500 from the round-2 generator, all with unused seeds. A charter and protocol, frozen with digests on 10 November 2026 before any model read this material, fixed the tests: every code-based test of rounds 1 and 2 and the indentation side of the cross-study test, run by byte-identical copies of the frozen scripts with their frozen thresholds and verdict rules, plus six new tests of L11H9 (Section 6). Thresholds for the new tests that no earlier protocol fixed were set from runs on material that earlier rounds had already read.

All 9 round-1 tests passed again, so by the frozen rule the two-route account is confirmed again: without DED, unmarked dedents changed by -0.994 (round 1: -1.091; no random set below -0.028), and the terminal route's set TERM changed terminal dedents by -0.032, just past its frozen bar of -0.03 (round 1: -0.044). All 25 round-2 tests passed, with values close to the originals, and all 5 questions were confirmed again in both models: knocking out the open header cost -0.117 at unmarked dedents in Qwen3-0.6B-Base (round 2: -0.148), steering raised the steered header's indent by +2.74 nats (round 2: +2.72), and OLMo-2-0425-1B's block-end set changed unmarked dedents by -0.776 (random -0.109 to +0.098). All 9 indentation-side tests of the cross-study test passed again (Section 6). On the new code the gate reads as before: 3 of the 5 models gated pass every item, Pythia-410M again misses only the header item (0.874), and GPT-2 fails every item. Table 11 lists each test with its original and round-3 value.

4 Ordinal suffixes: an early copy and a late override

English ordinals follow a short rule with one exception: the last digit chooses the suffix (1 *st*, 2 *nd*, 3 *rd*, otherwise *th*) unless the tens digit is 1, in which case the suffix is *th*. A model must combine two digits that its tokeniser may have put in different tokens. Qwen3, SmolLM2 and Gemma 4 write one token per digit; OLMo 2 cuts a run of digits into chunks of three from the left, so "1112" becomes `111` and `2`; GPT-2 and Pythia use byte-pair chunks of varying length. Sentences come from hand-written frames in five families (dates, centuries, rankings, anniversaries and positions) that end just before the suffix. The *teen AUROC* is the probability that a non-teen number ending in 1–3 has a higher margin m than a teen number; it is unaffected by a model's overall pull towards *th*. 6 of 7 models passed the first round's gate, and Qwen3-1.7B-Base, added for the second round, passed the same gate before any internal analysis (Appendix A). Besides Qwen3-0.6B-Base and OLMo-2-0425-1B, the study used five further models: GPT-2, Pythia-1B, SmolLM2-1.7B and Gemma-4-E2B in the first round, and Qwen3-1.7B-Base, SmolLM2-1.7B and Gemma-4-E2B for the early heads in the second.

4.1 Round 1: token boundaries and an early copy

The first round's protocol was frozen on 8 November 2026 and run once. All 14 tests passed: the 13 predictions of Table 12, of which the last on token boundaries is read separately in two models.

Behaviour follows token boundaries. OLMo-2-0425-1B fails exactly where its chunks of three leave the last digit alone in its token. On held-out items its teen AUROC was 0.775 at four digits and 0.729 at seven, while every length at which one token holds both digits stayed at 0.918 or above (Fig. 2 and Table 13). Plain four-digit teens were given *th* in only 0.220 of cases. With a thousands separator, which puts the last three digits in one token at every length, the AUROC rose to 0.989 and 0.953. GPT-2 and Pythia-1B show the same failure on the few numbers their tokenisers split before the last digit (teen accuracy 0.113 and 0.175, against

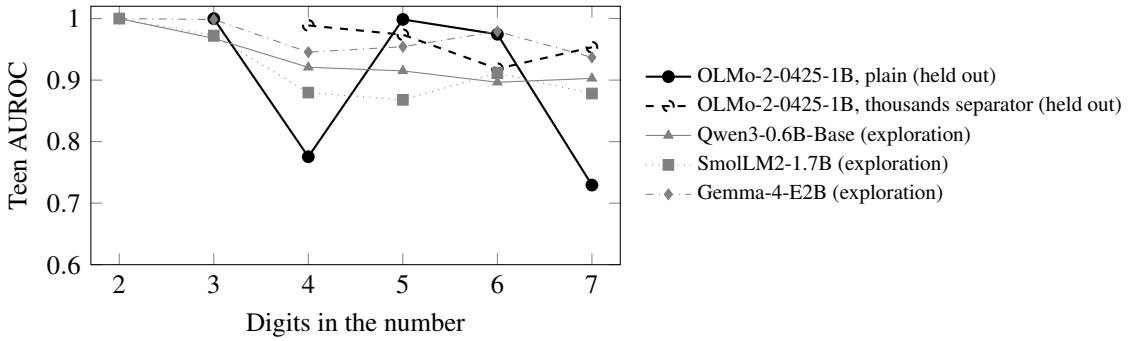


Figure 2: Teen AUROC by number length. OLMo-2-0425-1B falls only at four and seven digits, where its chunks of three leave the last digit in a token of its own; the separator removes the dip. Models with one token per digit show no such pattern.

0.950 and 0.965 on controls), though these sets hold only 10 distinct teen numbers each: the boundary test passes its frozen criterion in both models, but on too few numbers to show that the failure holds beyond them.

Qwen3-0.6B-Base: an early copy of the tens digit. In Qwen3-0.6B-Base, the layer-0 head L0H2 attends from the last digit to the tens digit and was the top head when patching teen pairs (0.599 of the logit difference at three digits). Ablating it with L1H6, a weaker head with the same pattern, left 0.430 and 0.459 of the teen effect at three and four digits, while every random pair of heads from layers 0–2 kept at least 0.994 (Table 14). Two-digit teens largely survived (0.709), so they have a further route.

OLMo-2-0425-1B: heads needed only across the boundary. Three layer-0 heads that attend to the previous token (L0H5, L0H0, L0H9) carry almost all of what remains of the teen effect where the tens digit is in the previous token (kept 0.110 and 0.076 at four and seven digits) and nothing where one token holds both digits (0.995 and 0.993; Table 15). One margin was narrow: at seven digits the lowest random triple kept 0.856, just above the 0.85 bar.

4.2 Round 2: from the copied digit to the suffix

The second round asked what happens after the copy. New frames were written and split, and numbers of six to ten digits were set aside, before any model read them; three- to five-digit numbers could not be fresh, so the held-out tests use round-1 confirmation and reserve numbers in new frames. The protocol was frozen on 9 November 2026 and run once. Of its 41 tests, 39 passed (Table 16).

MLP0 reads the copy and MLP24 decides. Of the teen effect that L0H2's ablation removes, 0.913 (four digits) and 0.913 (six digits) is removed when the ablation reaches MLP0's input alone. The signal then spreads thinly over many middle-layer components (exploratory) and comes together at MLP24: among MLP12–MLP27 it is the largest mediator of L0H2's teen effect (0.588 and 0.560 of it), the logit lens shows the teen effect rising most sharply from layer 23 to 24 (by 2.889 and 2.384 logits), and MLP24's own output accounts directly for 0.781 and 0.916 of the teen effect, with no other MLP above 0.201.

Two writers that do not consult each other. For teens, MLP24 raises *th* above the mean of the other three suffixes by 1.873–2.893 logits whatever the last digit, and favours the last digit's own suffix by at most 0.278: it says "teen, so *th*" without knowing which suffix it overrides. The last digit's suffix is written by MLP18,

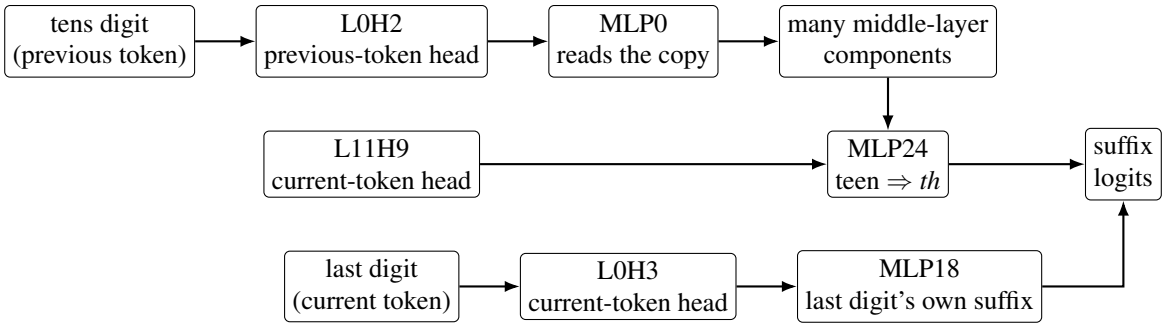


Figure 3: The path from digits to suffix in Qwen3-0.6B-Base. The teen route (top) is copied early, read by MLP0, spread over many components and gathered at MLP24, which writes a teen override. The last digit's suffix is written separately (bottom); L0H3's role rests on exploration and on the cross-study test (Section 6). The two routes meet only in the logits.

which favours it by 0.793–1.619 logits to much the same degree for teens and non-teens. The model does not compute the rule as a conditional: one component writes the last digit's suffix, another writes a teen override, and the override wins in the logits when the tens digit is 1. This fits the picture of MLPs as stores of input patterns that promote output tokens [18]. Figure 3 shows the path.

Backup heads. When L0H2 and L1H6 are removed, two later previous-token heads, L2H4 and L5H2, carry part of the tens digit, mostly for two-digit numbers: removing them as well lowers the two-digit teen effect kept from 0.701 to 0.482, below every control pair (Table 17). Even then 0.482 of the two-digit effect remains, and we did not find the route that carries it.

OLMo-2-0425-1B: a weak route, not a missing one. On held-out numbers of three to ten digits the token account held at every new length (Table 18). Where the tens digit sits in the previous token, the tens digit does arrive at the last token, but faintly: the teen direction there is at most 0.503, 0.484 and 0.440 of the three-digit one in size at four, seven and ten digits. The late layers can still use a signal of normal size: adding the three-digit teen direction at layer 8 to non-teen numbers lowers the suffix margin by 2.962, 2.359 and 1.986 logits, while random directions of the same size move it by at most 0.290, 0.401 and 0.242. The reader works; what reaches it is too faint.

Other models, and the two failed tests. In Qwen3-1.7B-Base, SmolLM2-1.7B and Gemma-4-E2B, heads in layers 0–1 that put at least half of the last token's attention on the previous token carry part of the teen effect, while random heads from the same layers carry almost none (Table 19). Of the 8 tests for these models, 6 passed. The two that failed are one prediction at its two number lengths: that in Qwen3-1.7B-Base these previous-token heads would leave the last digit's pull intact. They also removed much of it (0.548 and 0.233 kept at three and four digits, against the 0.55 required). In that model the two kinds of early head divide the work less cleanly than in Qwen3-0.6B-Base.

Checkpoints of another run (exploratory). Six stage-1 checkpoints of the published OLMo-2 training run were measured on exploration items (Table 20). As the indentation study found, the released OLMo-2-0425-1B weights do not descend from this run: the released model's token embeddings have cosine 0.006 with those at the end of stage 1, while the four published stage-2 checkpoints keep at least 0.9996 (a post hoc comparison). In this run the in-token teen check forms early and the cross-token check stays near chance until about a trillion tokens and is still weak at the end of the stage. This shows how the recipe can learn

the check, not when the released model learnt it, and the checkpoints support no inference about how the released model's cross-token route formed.

5 Capitals: lower case first, capitals last

Text in capitals is a small fraction of what a model reads, and its tokens differ: a vocabulary may have one token for "the" but spell "THE" and longer upper-case words out of pieces. The study gated its four candidate models on English Wikipedia paragraphs, and two passed: Gemma-4-E2B, whose vocabulary has many whole-word upper-case tokens, and Qwen3-1.7B-Base, which splits upper-case words into pieces. All four kept the case; the cost of capitals lay in choosing the word, and SmolLM2-1.7B (same word 0.595) and Pythia-1B (0.480) failed the 0.60 bar on it. Two more models were later put through the same gate and failed the same condition: OLMo-2-0425-1B, for the question of training dynamics below, and Qwen3-0.6B-Base, for the cross-study test (Section 6). In all, two of six gated models passed. The material was split by article into exploration, confirmation and reserve parts before any model read it; round 1 was confirmed on the confirmation part and round 2 on the reserve part and on headings from the same articles. U is the coming word's first token in capitals, L_c the same word's first token in lower case, and L_r the lower-case first token of another word of the paragraph.

5.1 Round 1: one direction carries the case

All 19 tests of the first round passed, 9 in Gemma-4-E2B and 10 in Qwen3-1.7B-Base (Table 21).

Lower case first. On held-out upper-case passages, the logit lens ranked L_c above U at the direction layer ℓ (24 of 35 in Gemma-4-E2B, 22 of 28 in Qwen3-1.7B-Base) at 0.707 and 0.753 of positions, while another word's lower-case token beat U at only 0.162 and 0.127. U took over for good only from layer 32 and 25, the last three layers of each model (Fig. 4). Read post hoc across depth, the lens shows a general lean towards lower-case tokens at the first layers and the coming word only from mid depth; "lower case first" means that once the word can be read through the lens, it reads in lower case.

One direction switches the case. The *capitals direction* d at a layer is the mean residual at word-start prediction positions in upper-case passages minus that in lower-case passages, estimated on exploration paragraphs. Adding it at ℓ to a lower-case passage made the next word upper case at 0.981 of positions in Gemma-4-E2B and 0.956 in Qwen3-1.7B-Base, and subtracting it from an upper-case passage made it lower case at 0.961 and 0.967; random directions of the same norm flipped at most 0.003. The direction is needed, though not alone: setting the residual's component along d to the midpoint of the two passages' means left the next word in capitals at 0.540 and 0.527 of upper-case positions. Adding d in ordinary mixed-case prose also capitalised the next word (0.957 and 0.941). The case follows the last few words: a lower-case passage whose last two words were in capitals gave an upper-case next word at 0.708 and 0.683 of positions.

5.2 Round 2: the parts that use the direction

The second round asked which components write the direction, which carry the passage's case to the prediction position, which read it, why the edit sometimes changes the word, and whether other capitals use it. A *context-only* passage (A) is the upper-case passage with its last word in lower case; it ends in the same token as the lower-case passage (L), so what separates A from L at the prediction position came through

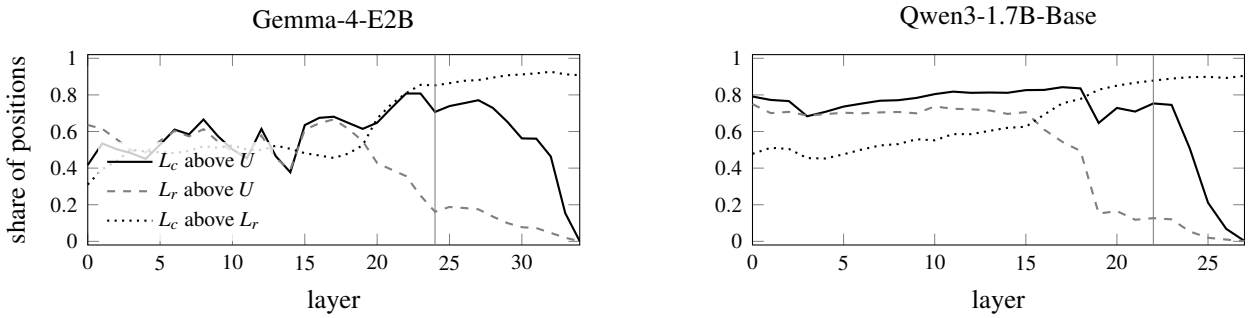


Figure 4: Held-out upper-case passages, logit lens by layer; the frozen tests read the lens at ℓ, and the reading across all layers is post hoc. Solid: the coming word's lower-case first token above its upper-case one. Dashed: another word's lower-case token above it. Dotted: the coming word's lower-case token above the other word's. The vertical line marks the direction layer ℓ.

Table 3: Capitals, round 2: selected held-out results. "Random" gives the extreme of twenty matched random sets. All tests passed; Appendix D gives every measure with its threshold and interval.

Measure	Qwen3-1.7B-Base	Gemma-4-E2B
Writers' writes along d moved to lower-case values: upper-case top-1 (random, lowest)	0.029 (0.985)	0.009 (0.949)
Add d after ℓ: flips to capitals; with the reader MLPs frozen	0.968; 0.002	0.989; 0.000
Carriers patched in A: share of the excess along d removed (random, highest)	0.872 (0.166)	0.859 (0.322)
Edit changes the word: where real capitals change it; where they do not	0.694; 0.120	0.549; 0.080
Remove d: acronyms kept in capitals; headings in capitals kept	1.000; 0.669	1.000; 0.768

attention. *Directional ablation* moves only a component's write along d to its value in L. All 23 tests passed, 12 in Qwen3-1.7B-Base and 11 in Gemma-4-E2B (Table 3; all are listed in Table 23).

Writers. The writes along d come almost entirely from a few components shortly before ℓ: in Qwen3-1.7B-Base MLPs 19–22 and head L19H0, in Gemma-4-E2B MLPs 22–24 and head L24H6. Moving only their writes along d to their lower-case values left an upper-case next word at 0.029 and 0.009 of positions, against at least 0.985 and 0.949 for random sets. The token embedding writes almost nothing along d, so the direction is not the current token's case itself.

Carriers. Four or five heads bring the earlier words' case to the prediction position (in Qwen3-1.7B-Base L19H0, L17H6, L18H1, L18H0; in Gemma-4-E2B L14H6, L14H7, L14H1, L15H2, L24H6). One in each model looks mostly at the last few words (L17H6 puts 0.929 of its attention on the last eight words, L14H7 0.918); the others look further back (L19H0: 0.284). Patching them from L into A removed 0.872 and 0.859 of A's excess along d. In Gemma-4-E2B random sets reach higher (0.322) because they share layer 14 with several named heads: attention at that layer carries case more widely than the named heads alone. L17H6 alone removed 0.363, more than any of twenty random layer-17 heads.

Readers. With d added after ℓ, freezing the MLPs after ℓ at their unedited outputs stopped the flip (0.002 and 0.000, from 0.968 and 0.989), while freezing matched heads did not. Freezing also stops whatever else these MLPs write, so this shows that the flip needs them, not that they do nothing else.

The current token. The current token's own case enters separately. In Qwen3-1.7B-Base, the layer-0 head L0H3 attends almost only to the current token, and its output from another paragraph's lower-case passage moved 0.352 of the gap between upper- and lower-case passages, while its output from another upper-case passage moved 0.029 (Table 22). The cross-study test (Section 6) showed that this head attends to the current token in every task, so case is one of the token's properties that it passes on.

Changed words and other capitals. Where adding d changes the word, it does so mostly where real capitals change it too (0.694 and 0.549 of flips, against 0.120 and 0.080 where the upper-case passage keeps the word), and more often for words whose capitals form needs an extra token. The direction thus reproduces how the model writes capitals, word changes included. Acronyms and initial capitals do not use it: removing d left acronyms in capitals at 1.000 and 1.000 of positions. Headings written in capitals do: removing d kept all capitals inside them at only 0.669 and 0.768, and at the first word after a heading the projection on d had fallen back to 0.034 and 0.116 of the gap, so the line break ends the capitals.

Training dynamics. The only checkpointed candidate, OLMo-2-0425-1B, failed the round-1 gate on the same-word condition (0.510), so the question of when the pattern forms stopped there, as the charter required.

6 Shared parts: a prospective test across four behaviours

Question and design. The three studies, and a companion study of line-breaking in prose with the same design [27], named groups of heads with a role in one behaviour each. The line-breaking groups, all in Qwen3-0.6B-Base, are the counters COUNT = {L4H0, L4H1, L5H13, L6H2, L7H15}, the newline writers WRITE = {L5H12, L5H13, L6H2} and the readers READ = {L7H14, L15H3, L17H6, L22H3}; the indentation sets DED, CARRY and TERM are listed in Section 3 and Appendix B. Three leads suggested that some might be the same parts: L17H6 was a reader in line-breaking and a member of the indentation sets CARRY and TERM; the early heads L0H2 and L0H3 appeared in the ordinal and capitals accounts; and L11H9, also a member of CARRY and TERM, had appeared in the ordinal study's exploration. The cross-study test asked, under its own frozen charter and protocol, whether the groups overlap inside one model. Each named group was switched off by mean ablation at the decision positions of every task, with its own study's measures, and compared with twenty random groups that replace each head by an unnamed head of the same layer. Line-breaking, indentation and ordinals were tested in Qwen3-0.6B-Base. Capitals used Qwen3-1.7B-Base with heads matched by index, because Qwen3-0.6B-Base failed the capitals gate (same word 0.547); the charter fixed this fallback in advance. All material was new and sealed before any model read it: a new batch of Wikipedia articles (150 line-width items, 60 paragraphs; 11,489 capitals word starts in 100 paragraphs), Python from repositories created after 1 October 2026 and generated programs with unused seeds (504 unmarked and 532 terminal dedents in natural code, 1,587 and 1,371 in generated programs), and the ordinal study's unread reserve (96 teen and 96 non-teen numbers at four digits, and the same at five). The protocol predicted an effect by a threshold and, in most cases, by requiring the named group to fall below all twenty of its random groups; it predicted the absence of an effect by a fixed bound on the change, without reference to the random groups. It was frozen and run once, and all 22 tests passed (Table 4; full matrices in Appendix E).

Line-breaking and indentation meet at L17H6. Without L17H6 the model kept 0.706 of the effect of line width on where it breaks a line (twenty random layer-17 heads: 0.988 to 1.036). The head also adds a little to the dedent after `return`: its ablation changed terminal dedents in generated programs by -0.022, small but below every random head of its layer (-0.011 to 0.002). This was the protocol's criterion for a shared part: a change of at most -0.010 that is also below every random head. On the new material of round 3 it was

Table 4: Shared parts: each named group ablated at the decision positions of every task, held out. Line-breaking: width effect and break/non-break separation kept. Indentation: change in mean log-probability of the right indent at unmarked dedents in natural code and terminal dedents in generated programs. Ordinals: teen effect and last-digit pull kept at four digits. Capitals (Qwen3-1.7B-Base, heads by index): D_{up} and D_{low} kept, the upper- and lower-case log-probability gaps. COUNT, WRITE and READ are the line-breaking counters, newline writers and readers (members in the text); "–": not run. Post hoc marks compare each value with the range of its twenty random groups, given in Appendix E: † below the range (more harm than any random group), ‡ above it.

	Line-breaking (kept)		Indentation (change)		Ordinals, 4 digits (kept)		Capitals (kept)	
Group	width	sep.	code unm.	gen. term.	teen	pull	D_{up}	D_{low}
COUNT	$0.150^{\dagger}$	$0.040^{\dagger}$	0.007	-0.003	1.013	1.008	1.009	1.009
WRITE	$0.466^{\dagger}$	$0.254^{\dagger}$	0.001	-0.001	0.999	0.997	–	–
READ−L17H6	$0.568^{\dagger}$	$0.512^{\dagger}$	-0.003	-0.003	$1.065^{\ddagger}$	$1.111^{\ddagger}$	–	–
L17H6	$0.706^{\dagger}$	$0.836^{\dagger}$	-0.010	$-0.022^{\dagger}$	0.990	1.002	$0.927^{\dagger}$	$0.957^{\dagger}$
CARRY	$0.715^{\dagger}$	$0.864^{\dagger}$	–	–	$0.714^{\dagger}$	$0.742^{\dagger}$	$0.860^{\dagger}$	$0.880^{\dagger}$
TERM	$0.675^{\dagger}$	$0.831^{\dagger}$	–	–	$0.704^{\dagger}$	$0.708^{\dagger}$	–	–
DED	0.990	0.999	$-1.112^{\dagger}$	$-0.266^{\dagger}$	0.976	0.932	0.996	0.999
L0H2	0.973	$0.871^{\dagger}$	-0.008	0.000	$0.507^{\dagger}$	$0.823^{\dagger}$	1.004	$0.970^{\dagger}$
L0H3	$0.391^{\dagger}$	$0.000^{\dagger}$	-0.009	-0.007	$1.077^{\ddagger}$	$0.155^{\dagger}$	$0.923^{\dagger}$	$0.331^{\dagger}$
L11H9	$1.023^{\ddagger}$	$1.027^{\ddagger}$	-0.009	$-0.030^{\dagger}$	$0.726^{\dagger}$	$0.733^{\dagger}$	$0.947^{\dagger}$	$0.936^{\dagger}$

met again (-0.025, random -0.010 to $+0.004$). Almost all of the indentation sets' effect on line-breaking passes through this head: CARRY keeps 0.715 of the width effect with it and 1.015 without it; TERM keeps 0.675 and 0.966. Both remainders pass the frozen bound (at least 0.92 kept), but, read post hoc against their random groups, they are not nil: TERM without L17H6 lies just below its random range for the width effect (0.975 to 1.028), and CARRY without L17H6 raises the separation to 1.033, above its range (0.960 to 1.014). In Qwen3-1.7B-Base, by index, L17H6 trims the capitals gap (0.927 kept) without changing which word is chosen (word log-probability change -0.002), and in the capitals study it is the local carrier of passage case. Each role moves a fact about the recent text's format to where the format is decided.

L11H9: read post hoc, then confirmed on new material. L11H9 sits in the indentation sets CARRY and TERM, attends to the current token (0.90 to 0.96 of its attention) and holds part of the ordinal suffix choice in general rather than the teen check: teen effect kept 0.726 and pull kept 0.733 at four digits, each below every random head. The protocol predicted that it would not matter for line-breaking or indentation, and tested this with fixed bounds: at least 0.93 of both line-breaking measures kept, and an absolute change of at most 0.05 on each of the four indentation measures. Both tests passed, and the frozen verdict is a suffix head not shared with line-breaking or indentation. That bound was looser than the criterion applied to L17H6, and the frozen tests did not compare L11H9's indentation values with its random heads. Read post hoc, its change at terminal dedents in generated programs, -0.030 (random -0.007 to 0.002), meets the criterion that made L17H6 a shared part, and is the larger of the two. Because that reading was made after the run, round 3 tested it under a protocol frozen before any model read the new material (Section 3.3). There L11H9 alone changed terminal dedents in generated programs by -0.027, below all twenty random layer-11 heads (-0.004 to $+0.003$): the criterion that made L17H6 a shared part. CARRY and TERM, run on held-out indentation material for the first time, met the same criterion (-0.303 and -0.224), and leaving L11H9 out made their changes less negative by +0.039 and +0.052. The head spared unmarked dedents (-0.010 in natural code, -0.012 in generated programs). In natural code its effect on terminal dedents was much smaller, -0.003, though still below its random range (-0.001 to $+0.001$), because terminal dedents there

are nearly certain. In return-versus-assignment pairs it carried 0.122 of the gap by necessity, against at most 0.033 for random layer-11 heads. By the frozen rule (confirmed if the single-head and set tests both pass) its second role is confirmed: L11H9 is shared, with a small role, between the ordinal suffix choice and the dedent after a terminal statement (`return`, `raise`, `break`, `continue` or `pass`), which agrees with the indentation study's having named it in CARRY and TERM. The cross-study test's own verdict, which bounded the head's indentation effect at 0.05, also stands; the two tests asked different questions. For comparison, L17H6 carried 0.069 in the same pairs, above nineteen of its twenty random layer-17 heads; the twentieth was L17H4, itself a CARRY head (0.261), because the frozen draw rule excludes only the head under test. This comparison was not scored. Three further post hoc readings concern L11H9. The indentation sets lower the ordinal measures by about as much as L11H9 alone, with or without L17H6 (teen effect kept 0.723 for CARRY and 0.714 for TERM without L17H6, below their random ranges). Its ablation raised both line-breaking measures a little above their random ranges (width effect 1.023, random 0.985 to 1.006; separation 1.027). And in Qwen3-1.7B-Base, by index, it lowered both capitals gaps (D_{up} kept 0.947, D_{low} kept 0.936), each below its random range.

The groups that do most of each task's work are separate. The line-breaking groups remove much of a line-breaking measure (COUNT keeps 0.040 of the separation, WRITE 0.466 of the width effect, READ without L17H6 0.568) and leave indentation inside the protocol's bounds (largest absolute change for COUNT: 0.007 in natural code). The block-end set DED changes unmarked dedents by -1.112 (random -0.027 to 0.014) and leaves line-breaking as it was (0.990 and 0.999 kept). Read post hoc against the random ranges, none of these four groups harms another task's measure more than its random groups do, on any measure in Appendix E. The newline writers of line-breaking do not serve indentation, although both decisions are made at the same newline token.

The early heads are general movers, not task parts. L0H2 attends mainly to the previous token in prose, numbers and capitals text (0.60, 0.90 and 0.70 of its attention at decision positions). It carries the tens digit for the teen check (0.507 and 0.546 of the teen effect kept at four and five digits) and also matters for line-breaking (separation kept 0.871). L0H3 attends to the current token in every task (0.97 to 0.99). It carries the last digit (pull kept 0.155 and 0.314), is needed for line-breaking decisions (separation kept 0.000), and in Qwen3-1.7B-Base matters far more for choosing the next word than for keeping capitals (D_{low} kept 0.331, D_{up} kept 0.923; word log-probability change -1.816). The capitals study's finding that it reads a token's case is one use of reading the current token. Both heads passed the frozen bounds on indentation.

What "separate" means here. In the frozen tests, a part was predicted not to serve a task if its ablation stayed within a bound fixed in advance: for indentation an absolute change of at most 0.03, or 0.10 at unmarked dedents in generated programs, and 0.05 for L11H9. Every such test passed, on these measures, at the decision positions. The range of twenty layer-matched random groups is a stricter standard, which the frozen tests applied only where they predicted an effect; Table 4 and Appendix E mark every value outside its range, and these marks are post hoc. Beyond the cases above, the marked cross-task values belong to the general early movers or to sets that contain L17H6 or L11H9 (for example CARRY in capitals, D_{up} kept 0.860), or are small: READ without L17H6 raises the ordinal measures (teen effect kept 1.065), and L17H6 in Qwen3-1.7B-Base lowers D_{low} to 0.957. Effects that pass through earlier positions are outside the test: in exploration, ablating L0H2 at every position did reduce terminal dedents, but the indentation study's second round showed that this effect does not travel through the terminal line (L0H2 ablated on that line and the decision token changed the probability q of staying at the current indent by +0.029). Sharing through MLPs and the residual stream is not excluded.

7 Related work

Brackets, code and indentation. Theory and experiments on bracket languages show that attention can track the top of a stack of open brackets [11, 58]; simplified models of such tasks can mislead about the full model [15]; and in trained bracket models masking attention to the true top of the stack hurts long-distance accuracy, while residual-stream subspaces that decode depth matter much less [48]. In pretrained models, parenthesis closing has been traced to components of varying reliability, among them, in a code model, a late head that attends to the open parenthesis that has yet to be closed [46], and to heads that keep track of the number of parentheses already closed [37], whose authors suggest that similar counting may serve Python indentation and propose to test such reuse. L23H2 is structurally of this kind, but it copies an indent from the header token rather than a bracket type, and its attention sets how far to dedent. Indentation itself has been studied in recurrent networks, where some cells grow with the depth of indentation [28], in a four-layer attention-only model with a head that attends to earlier tokens of similar indentation [26], in weight-sparse models trained on Python code, whose hand-made tasks include predicting the indent after a `for` line [16], and as an ordered low-dimensional geometry of indent level in larger models, tested by patching probe accuracy [3]. Whether the indent that L23H2 copies lives in that geometry is untested.

Numbers and morphology. Mechanistic studies of numbers have found circuits that compare years [23], attention that moves operands to the last token before MLPs compute the result [50], neurons that each respond to a numerical pattern [39], digit-wise base-10 representations [31], digit-position MLP neurons [2], mod-10 features in successor heads [20] and shared circuits for number sequences [30]. Early layers assemble multi-token units [13, 29], and tokenisation changes arithmetic accuracy, with commas helping [49]. The ordinal account has this overall shape: L0H2 and MLP0 are a specific case of early assembly, and the separator result gives one concrete reason why commas help. Agreement and morphology circuits have been traced in other settings [12, 14]. To our knowledge, what is new is the teen override at MLP24, kept separate from MLP18's last-digit write, and the weak cross-token route in OLMo-2-0425-1B.

Case, latent forms and directions. Merullo et al. [36] found, for in-context uppercasing of single words among other tasks, that the word appears first and a mid-to-late feed-forward update along one transferable vector then makes it upper case. The capitals account shares this shape and extends it to passages in natural text, with named writers, carriers and readers, the split between headings and word-bound capitals, and the same roles in two tokenisers. Function and task vectors include capitalisation tasks [25, 52]; sparse-autoencoder features in GPT-2 small split the vocabulary by capitalisation and promote all-capitals tokens [8]. Models represent the answer in one form and write it late in another in multilingual settings [47, 55]. Direction-based steering and erasure [5, 33, 42, 43, 53] supply the tools. On the input side, representing case apart from the word helps a masked model on upper-cased text [44], and upper-case or alternating-case spans in a prompt concentrate attention on themselves, with accuracy falling mainly for alternating case [9]; neither concerns how a model writes in capitals.

Reuse of components. Merullo et al. [35] found that about 78% of the heads in the indirect-object circuit of GPT-2 Medium are reused for a colour-object task. Lan et al. [30] report shared sub-circuits for related sequence continuations, and Mondorf et al. [38] composable circuits for related compositional tasks. Li and Subramani [32], across six tasks and five models including Qwen3-4B, find that component-level circuits share most of their parts across tasks, related or not, mostly MLP blocks and a few attention heads with general-purpose roles. Xu [56], with matched random controls in three 1B-class models, propose a previous-token substrate under task circuits in a mixture-of-experts model. Dunefsky et al. [10] find a countdown

subcircuit in a 70B model that serves tasks from sentence planning to aligning cells in ASCII tables, and Gurnee et al. [22] describe the line-breaking mechanism that the companion study [27] follows. General-purpose units recur across models [21], and previous-token heads are the substrate of induction [41]. Our test differs in being prospective and in comparing parts named by separate studies of unrelated behaviours. Its result agrees with the general-purpose roles of shared heads found by Li and Subramani [32] and Xu [56], but finds the task-specific parts mostly unshared: the groups that do most of each task's work are separate, and beyond the general early movers two heads have a small role in a second behaviour, L17H6 by the cross-study test and L11H9 by a later frozen test on new material.

8 Discussion

One pattern, three behaviours. In each behaviour the decision combines a general signal, moved by heads that do the same job in every task, with a task-specific step late in the model. Indentation reads the open block's header with a late head, the chief of a small set, and copies its indent; ordinals copy the tens digit with a general previous-token head and decide "teen" at one late MLP; capitals carry the passage's case with a few heads and turn the chosen word into capitals with the last MLPs. Two of the three decisions are written as an override rather than a conditional: MLP24 writes *th* without consulting the last digit, and the capitals direction turns a lower-case choice into capitals rather than choosing among upper-case words. The cross-study test shows that the task-specific parts are mostly not shared, even between line-breaking and indentation, which are decided at the same newline token. Two heads have a small second role: L17H6 moves format facts in line-breaking and indentation, and L11H9, confirmed on new material, reads the current token for the suffix choice and adds to the dedent after a terminal statement. Which head carries a role can differ between training runs, as the two OLMo-2 runs show for indentation, while the role recurs.

Limitations.

- *Models and settings.* Small base models (the largest, Gemma-4-E2B, has 2.3 billion effective parameters), English text, Python only for indentation, numbers up to ten digits, and Wikipedia paragraphs and headings for capitals. Indentation was tested in two models, the ordinal path traced only in Qwen3-0.6B-Base, and capitals in two models.
- *Partial accounts.* In indentation, what tells the block-end head to look is spread over many components, and we did not trace what writes the indent it reads. In ordinals, the signal between MLP0 and MLP24 is spread thinly and untraced; 0.482 of the two-digit teen effect survives all four copying heads; the late components that fail to read OLMo-2-0425-1B's faint cross-token signal were not found; and the frame-dependent errors of Qwen3-0.6B-Base remain unexplained. In capitals, removing *d* in round 1 left about half of the words in capitals, so case is also carried in ways one mean difference does not capture; the writers, carriers and readers were tested as sets, not path by path.
- *Tools.* Mean ablation replaces outputs with averages that need not occur on real inputs; direct attribution linearises the final normalisation; the logit lens is brittle [4]; and patching choices can change which components look important [24, 59]. Random-set controls bound these concerns but do not remove them.
- *The cross-study test* used one model family, with capitals heads matched by index, and ablation at decision positions only. Its frozen tests judged the absence of an effect by fixed bounds; the comparison of every value with its random range is post hoc, except for L11H9's second role and the indentation sets CARRY and TERM, which round 3 tested prospectively on new material. Round 3 re-ran only the code-based tests; the line-breaking, ordinal and capitals sides rest on the cross-study test's material alone.

- *Held-out material* is held out from these analyses, not necessarily from the models' training data. The natural code and the Wikipedia articles come from repositories and pages created after the models were trained, but such repositories and pages can contain older code and text, so overlap with the training data remains possible.
- *Training dynamics* are exploratory and come from a published training run that the released OLMo-2-0425-1B weights do not descend from.

Departures from plan. Every threshold was set before its test from values measured on exploration material, and no test was rerun. The indentation gate's first measure left out decision points after blank lines for two tokenisers; a dated amendment corrected the measure, and no gate verdict changed (Table 6 gives both measures). The indentation study's second round set its thresholds from runs of the confirmation scripts on exploration material, and a dated amendment, made before the protocol was frozen, permitted the download of OLMo-2 checkpoints, which the charter had excluded. In the ordinal study's second round, two draft thresholds were reset after a full run of the tests on exploration items and a third from the last exploration run, all before the freeze; a dated amendment moved the freeze later than planned, let one exploration run that was already queued finish, and allowed the checkpoint analysis to be completed as an exploratory result. No held-out material was read before the freeze. The indentation and ordinal studies planned their checkpoint analyses on the assumption that the published OLMo-2 checkpoints are ancestors of the released model. A post hoc weight comparison shows that they are not, so the ordinal study's report no longer infers from them when the released model's cross-token route formed, and no claim here rests on that assumption. In the capitals study the checkpoint question stopped at the gate. In round 3 of the indentation study, the frozen queue's cap on resumed passes ran out for OLMo-2-0425-1B's set ablation before all its runs were saved; the same command was resumed from the saved record, which changed no computation. Appendix F lists the remaining notes.

Code and data availability. The code, the frozen charters and protocols with their file digests, the saved outputs, the exact model identifiers and the held-out material that may be redistributed are at `https://github.com/Jeevash23/interp-small-decisions`, archived at `https://doi.org/10.5281/zenodo.24150049` (release v1.0; later releases are listed under the same record). The natural code of round 3, all under permissive licences, is included with each repository's licence text. The natural code of the first two indentation rounds and of the cross-study test is not distributed: for each file the repository lists its repository, path, licence and SHA-256, and the fetch scripts retrieve it again.

Use of language-model assistants. Language-model assistants were used throughout this work, under the author's direction, to help design the studies, to write and run the code for the exploration, the experiments and the analyses, to draft and revise the text and to check the references against their online records. The study designs, frozen documents and saved outputs are recorded with digests. All reported results come from the saved study outputs, and every number in the text was generated from those outputs and checked by an independent script.

References

[1] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016. URL `https://arxiv.org/abs/1610.01644`.

[2] Tanja Baeumel, Daniil Gurgurov, Yusser al Ghussin, Josef van Genabith, and Simon Ostermann. Modular

arithmetic: Language models solve math digit by digit. *arXiv preprint arXiv:2508.02513*, 2025. URL https://arxiv.org/abs/2508.02513.

[3] Saksham Bassi and Sharvi Tomar. Geometry of ordinal representations in language models. In *Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning (ICML)*, Seoul, South Korea, 2026. URL https://arxiv.org/abs/2607.04167.

[4] Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023. URL https://arxiv.org/abs/2303.08112.

[5] 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 (NeurIPS)*, 2023. URL https://arxiv.org/abs/2306.03819.

[6] Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, et al. SmolLM2: When smol goes big – data-centric training of a small language model. *arXiv preprint arXiv:2502.02737*, 2025. URL https://arxiv.org/abs/2502.02737.

[7] Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pages 2397–2430. PMLR, 2023. URL https://proceedings.mlr.press/v202/biderman23a.html. arXiv:2304.01373.

[8] Joseph Bloom and Johnny Lin. Understanding SAE features with the logit lens. LessWrong and AI Alignment Forum, 2024. URL https://www.lesswrong.com/posts/qykrYY6rXXM7EEs8Q/understanding-sae-features-with-the-logit-lens. Posted 11 March 2024; accessed 9 November 2026.

[9] Maximilian Dillitzer, Tin Stribor Sohn, Jason J. Corso, and Michael Auerbach. Attention is case-sensitive. *arXiv preprint arXiv:2608.03711*, 2026. URL https://arxiv.org/abs/2608.03711. Accepted at ECCV 2026.

[10] Jacob Dunefsky, Wes Gurnee, and Emmanuel Ameisen. A shared subcircuit lets LLMs count down across tasks. *arXiv preprint arXiv:2607.12279*, 2026. URL https://arxiv.org/abs/2607.12279.

[11] Javid Ebrahimi, Dhruv Gelda, and Wei Zhang. How can self-attention networks recognize Dyck-n languages? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020. URL https://arxiv.org/abs/2010.04303.

[12] Javier Ferrando and Marta R. Costa-jussà. On the similarity of circuits across languages: a case study on the subject-verb agreement task. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024. URL https://arxiv.org/abs/2410.06496.

[13] Sheridan Feucht, David Atkinson, Byron Wallace, and David Bau. Token erasure as a footprint of implicit vocabulary items in LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9727–9739, 2024. doi: 10.18653/v1/2024.emnlp-main.543. URL https://aclanthology.org/2024.emnlp-main.543/.

[14] Matthew Finlayson, Aaron Mueller, Sebastian Gehrmann, Stuart Shieber, Tal Linzen, and Yonatan Belinkov. Causal analysis of syntactic agreement mechanisms in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL-IJCNLP)*, 2021. URL https://arxiv.org/abs/2106.06087.

[15] Dan Friedman, Andrew Lampinen, Lucas Dixon, Danqi Chen, and Asma Ghandeharioun. Interpretability illusions in the generalization of simplified models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. URL https://arxiv.org/abs/2312.03656.

[16] Leo Gao, Achyuta Rajaram, Jacob Coxon, Soham V. Govande, Bowen Baker, and Dan Mossing. Weight-sparse transformers have interpretable circuits. *arXiv preprint arXiv:2511.13653*, 2025. URL https://arxiv.org/abs/2511.13653.

[17] Gemma Team. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026. URL https://arxiv.org/abs/2607.02770.

[18] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*

(EMNLP), 2021. URL https://arxiv.org/abs/2012.14913.

[19] Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. Localizing model behavior with path patching. *arXiv preprint arXiv:2304.05969*, 2023. URL https://arxiv.org/abs/2304.05969.

[20] Rhys Gould, Euan Ong, George Ogden, and Arthur Conmy. Successor heads: Recurring, interpretable attention heads in the wild. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2312.09230.

[21] Wes Gurnee, Theo Horsley, Zifan Carl Guo, Tara Rezaei Kheirkhah, Qinyi Sun, Will Hathaway, Neel Nanda, and Dimitris Bertsimas. Universal neurons in GPT2 language models. *arXiv preprint arXiv:2401.12181*, 2024. URL https://arxiv.org/abs/2401.12181.

[22] Wes Gurnee, Emmanuel Ameisen, Isaac Kauvar, Julius Tarng, Adam Pearce, Chris Olah, and Joshua Batson. When models manipulate manifolds: The geometry of a counting task. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/linebreaks/index.html. Also arXiv:2601.04480 (2026).

[23] Michael Hanna, Ollie Liu, and Alexandre Variengien. How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023. URL https://arxiv.org/abs/2305.00586. arXiv:2305.00586.

[24] Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*, 2024. URL https://arxiv.org/abs/2404.15255.

[25] Roee Hendel, Mor Geva, and Amir Globerson. In-context learning creates task vectors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023. URL https://arxiv.org/abs/2310.15916.

[26] Jett Janiak, cmathw, and StefanHex. Polysemantic attention head in a 4-layer transformer. LessWrong, 2023. URL https://www.lesswrong.com/posts/nuJFTS5iiJKT5G5yh/polysemantic-attention-head-in-a-4-layer-transformer. Posted 9 November 2023; accessed 9 November 2026.

[27] Richard Jeevash. Counting by dilution and a width stored at the newline: a partial account of line-breaking in Pythia-410M, 2026. Companion preprint, archived at https://doi.org/10.5281/zenodo.24112443.

[28] Andrej Karpathy, Justin Johnson, and Li Fei-Fei. Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078*, 2015. URL https://arxiv.org/abs/1506.02078.

[29] Vedang Lad, Jin Hwa Lee, Wes Gurnee, and Max Tegmark. The remarkable robustness of LLMs: Stages of inference? *arXiv preprint arXiv:2406.19384*, 2024. URL https://arxiv.org/abs/2406.19384.

[30] Michael Lan, Philip Torr, and Fazl Barez. Towards interpretable sequence continuation: Analyzing shared circuits in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 12576–12601, 2024. doi: 10.18653/v1/2024.emnlp-main.699. URL https://aclanthology.org/2024.emnlp-main.699/.

[31] Amit Arnold Levy and Mor Geva. Language models encode numbers using digit representations in base 10. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL), Short Papers*, pages 385–395, 2025. doi: 10.18653/v1/2025.naacl-short.33. URL https://aclanthology.org/2025.naacl-short.33/.

[32] Michael Li and Nishant Subramani. How much do circuits tell us? Measuring the consistency and specificity of language model circuits. *arXiv preprint arXiv:2605.08348*, 2026. URL https://arxiv.org/abs/2605.08348.

[33] 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 (COLM)*, 2024. URL https://arxiv.org/abs/2310.06824.

[34] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022. URL https://arxiv.org/abs/2202.05262. arXiv:2202.05262.

[35] Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. Circuit component reuse across tasks in transformer language models. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2310.08744.

[36] Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. Language models implement simple Word2Vec-style vector arithmetic. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for*

Computational Linguistics (NAACL), 2024. URL https://arxiv.org/abs/2305.16130.

[37] Samuel Miller, Daking Rai, and Ziyu Yao. Mechanistic understanding of language models in syntactic code completion. In *AAAI 2025 Workshop on Towards Knowledgeable Foundation Models*, 2025. URL https://arxiv.org/abs/2502.18499.

[38] Philipp Mondorf, Sondre Wold, and Barbara Plank. Circuit compositions: Exploring modular structures in transformer-based language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025. URL https://arxiv.org/abs/2410.01434.

[39] Yaniv Nikankin, Anja Reusch, Aaron Mueller, and Yonatan Belinkov. Arithmetic without algorithms: Language models solve math with a bag of heuristics. *arXiv preprint arXiv:2410.21272*, 2024.

[40] nostalgebraist. interpreting GPT: the logit lens. LessWrong, 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens. Posted 31 August 2020; accessed 8 November 2026.

[41] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022. URL https://arxiv.org/abs/2209.11895.

[42] Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*, 2023. URL https://arxiv.org/abs/2312.06681.

[43] Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. URL https://arxiv.org/abs/2311.03658.

[44] Rafal Powalski and Tomasz Stanislawek. UniCase – rethinking casing in language models. *arXiv preprint arXiv:2010.11936*, 2020. URL https://arxiv.org/abs/2010.11936.

[45] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf. Accessed 6 November 2026.

[46] Daking Rai, Samuel Miller, Kevin Moran, and Ziyu Yao. Failure by interference: Language models make balanced parentheses errors when faulty mechanisms overshadow sound ones. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://arxiv.org/abs/2507.00322.

[47] Alan Saji, Jaavid Aktar Husain, Thanmay Jayakumar, Raj Dabre, Anoop Kunchukuttan, and Ratish Puduppully. RomanLens: The role of latent romanization in multilinguality in LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26410–26429, 2025. doi: 10.18653/v1/2025.findings-acl.1354. URL https://arxiv.org/abs/2502.07424.

[48] Aryan Sharma, Cutter Dawes, and Shivam Raval. Dissociating decodability and causal use in bracket-sequence transformers. *arXiv preprint arXiv:2604.22128*, 2026. URL https://arxiv.org/abs/2604.22128.

[49] Aaditya K. Singh and DJ Strouse. Tokenization counts: the impact of tokenization on arithmetic in frontier LLMs. *arXiv preprint arXiv:2402.14903*, 2024. URL https://arxiv.org/abs/2402.14903.

[50] Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023. URL https://arxiv.org/abs/2305.15054.

[51] Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Allyson Ettinger, Michal Guerquin, David Heineman, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Jake Poznanski, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 OLMo 2 furious. *arXiv preprint arXiv:2501.00656*, 2024. URL https://arxiv.org/abs/2501.00656. A shorter version was accepted to COLM 2025.

[52] Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function vectors in large language models. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2310.15213.

[53] Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023. URL https://arxiv.org/abs/2308.10248.

[54] Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2211.00593.

[55] Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. Do Llamas work in English? On the latent language of multilingual transformers. *arXiv preprint arXiv:2402.10588*, 2024. URL https://arxiv.org/abs/2402.10588.

[56] Yongzhong Xu. Pattern selectivity is not task-causal structure: A cross-architecture mechanistic study of composed-task circuits in 1B-class language models. *arXiv preprint arXiv:2606.05378*, 2026. URL https://arxiv.org/abs/2606.05378.

[57] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

[58] Shunyu Yao, Binghui Peng, Christos Papadimitriou, and Karthik Narasimhan. Self-attention networks can process bounded hierarchical languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021. URL https://arxiv.org/abs/2105.11115.

[59] Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=Hf17y6u9BC. arXiv:2309.16042.

A Gates

Each study gated its candidate models on exploration material before any internal analysis, with bars and a fallback order fixed in its charter. Tables 5 and 6 give the indentation gate. In the ordinal study 6 of the 7 models gated in the first round qualified; Pythia-410M missed the second condition by 0.006 and Pythia-1B took its place. Qwen3-1.7B-Base was not gated in the first round; the second round's charter required the same gate of it, on exploration items, before any internal analysis, and it passed. So 7 of the 8 models gated for ordinals passed. In the capitals study, all four candidates kept the case (0.994 to 1.000 of cased top-1 tokens upper case) and two chose the same word often enough: Gemma-4-E2B (0.697) and Qwen3-1.7B-Base (0.620); SmolLM2-1.7B (0.595) and Pythia-1B (0.480) did not, and the bar was not loosened although SmolLM2 missed it narrowly. Two further models were put through the capitals gate for other purposes and failed the same-word condition: OLMo-2-0425-1B (0.510), the only candidate with published checkpoints, and Qwen3-0.6B-Base (0.547), for the cross-study test. So two of the six models put through the capitals gate passed.

Table 5: The gate, on the exploration part of the post-cut-off code (items 1–3) and of the generated programs (item 4). Values use the corrected measure; Table 6 gives both measures. "Top-1 minus copy" is the top-1 accuracy of the indent minus the accuracy of copying the previous line's indent; "never-opened mass" is the share of q at dedents that falls on shallower indents that no open block has, for the indent width (two spaces, three spaces or mixed) where it is largest. OLMo-2-0425-1B was gated in round 2 on the same material.

Model	points	header top-1	terminal top-1	terminal mass on open levels	top-1 minus copy	never-opened mass (worst width)
Pythia-410M	22,854	0.870	0.973	0.994	0.358	0.073
Qwen3-0.6B-Base	22,854	1.000	0.989	0.997	0.406	0.065
OLMo-2-0425-1B	22,854	0.997	0.966	0.984	0.392	0.071
Gemma-4-E2B	22,854	0.995	0.986	0.996	0.406	0.061
GPT-2	22,854	0.078	0.337	0.696	−0.243	0.208
bar		≥ 0.90	≥ 0.5	≥ 0.7	≥ 0.10	≤ 0.10

Table 6: Gate on post-cut-off code under the first measure, which left out decision points after blank lines for two tokenisers, and under the corrected measure. Gemma-4-E2B has no value under the first measure.

Model	measure	points	header top-1	terminal top-1	terminal mass on open levels	top-1 minus copy
Pythia-410M	first	20,249	0.870	0.965	0.991	0.340
Pythia-410M	corrected	22,854	0.870	0.973	0.994	0.358
Qwen3-0.6B-Base	first	22,854	1.000	0.989	0.997	0.406
Qwen3-0.6B-Base	corrected	22,854	1.000	0.989	0.997	0.406
GPT-2	first	21,129	0.078	0.382	0.639	−0.247
GPT-2	corrected	22,854	0.078	0.337	0.696	−0.243

B Indentation: all frozen predictions

Table 7 lists the round-1 tests, Tables 8 and 9 the round-2 tests in the two models, Table 10 the exploratory checkpoint measurements, and Table 11 the round-3 replication. Round 1 head sets: DED as in Section 3; CARRY = {L17H4, L11H9, L5H6, L17H6, L16H2, L6H5}; TERM = {L11H9, L5H6, L17H6, L16H2, L23H6, L22H8, L23H3, L25H9, L19H6}. Exploratory: patching the whole residual stream at the decision token between a terminal statement and `value = value` showed sufficiency near zero up to layer 4 (0.04), 0.15 at layer 5 and 0.87 at layer 17.

Table 7: The frozen predictions and their results on held-out material. Changes are in mean log-probability of the true indent unless stated. Intervals resample repositories or programs (2,000 resamples); they are descriptive and were not used for verdicts.

	Prediction	Result	Random sets	Verdict
H1	DED, natural, other dedents: change ≤ -0.40, below every random set	-1.091 $[-1.168, -1.015]$	lowest -0.029	pass
H2	same points: probability of keeping the current level rises ≥ 0.10, above every random set	+0.306 $[0.286, 0.323]$	highest +0.016	pass
H3	DED, terminal dedents: mass on open levels ≥ 0.95; change at most a quarter of H1's	0.991; -0.059	–	pass
H4	DED, same-level lines: change ≥ -0.01	+0.025 $[0.022, 0.028]$	–	pass
H5	TERM, terminal dedents: change ≤ -0.03, below every random set	-0.044 $[-0.054, -0.035]$	lowest -0.009	pass
H6	TERM, other dedents: absolute change ≤ 0.10	-0.027 $[-0.044, -0.012]$	–	pass
H7	TERM+DED, terminal dedents: change below the sum of the two alone	-0.188 (sum -0.103)	–	pass
H8	DED, generated, other dedents: change ≤ -0.25, below every random set	-0.631 $[-0.646, -0.615]$	lowest -0.096	pass
H9	CARRY, generated pairs: necessity ≥ 0.40, sufficiency ≥ 0.20, each above every random set	0.650, 0.373	0.075, 0.032	pass

Table 8: Round 2, Qwen3-0.6B-Base, head L23H2: frozen predictions and results on held-out material. Natural code: 3,934 unmarked dedents, 18,570 same-level lines and 3,504 terminal dedents; generated: 782 programs ending a block with a terminal statement. Changes are in mean log-probability of the true indent unless stated. Intervals resample repositories (2,000 resamples) and are descriptive.

	Prediction	Result	Verdict
Q1	Attention to the innermost open header's `:\n` at unmarked dedents ≥ 0.30 and $\geq 2.5\times$ that at same-level lines	0.458 [0.439, 0.476]; 0.118	pass
Q2	Knocking out that token, unmarked dedents: ≤ -0.05, below every other layer-23 head, and at least twice the largest absolute change among them	-0.148 [-0.162, -0.132]; others -0.004 to +0.044	pass
Q3	Same knock-out: same-level lines ≥ -0.02; terminal dedents at most a quarter of Q2 in size	+0.013; -0.005	pass
Q4	Knocking out the latest closed header instead: at most half of the open-header knock-out on the same points	-0.001 against -0.122	pass
Q5	Most-attended open header is the target where the target is not the innermost: ≥ 0.60; over all multi-header dedents, beats the innermost rule by ≥ 0.02	0.893 [0.873, 0.913]; 0.914 against 0.822	pass
Q6	Steering, natural unmarked dedents: rise ≥ 1.0 and $\geq 5\times$ the control; becomes top in a share $\geq$ control $+0.05$	+2.72 (control +0.08); 0.125 (0.004)	pass
Q7	Knocking out the true-level header: ≤ -0.10	-0.210	pass
Q8	Steering, generated terminal targets: rise ≥ 1.0 and $\geq 5\times$ control; top in ≥ 0.20 and $\geq$ control $+0.10$	+1.41 (control -0.01); 0.347 (0.020)	pass
Q9	`return`: removing the decision-token-to-keyword edge in all layers closes ≥ 0.40 of the gap to the neutral line	0.716 [0.699, 0.733]	pass
Q10	Largest single-layer share at most $0.35\times$ the Q9 share	0.098 (layer 17)	pass
Q11	L0H2 ablated on the terminal line and decision token: the probability q of staying at the current indent changes by at most 0.05	+0.029 [0.023, 0.034]	pass

Table 9: Round 2, OLMo-2-0425-1B: head L13H1 (O1–O10) and the block-end set DED-O = {L13H1, L10H12, L11H0, L15H4} (G1–G4), on the same held-out material. O-predictions repeat the Q-predictions with the thresholds shown.

	Prediction	Result	Verdict
O1	As Q1 with ≥ 0.20 and $\geq 1.4\times$	0.290; 0.181	pass
O2	As Q2 (other layer-13 heads)	-0.132; others +0.000 to +0.002	pass
O3	As Q3	+0.016; -0.009	pass
O4	As Q4	-0.004 against -0.119	pass
O5	Target is most attended in ≥ 0.45 of non-innermost cases, and at least 0.15 more often than the innermost	0.694 against 0.139	pass
O6	As Q6 with rise ≥ 0.8 and $\geq 4\times$ control	+1.58 (+0.15); 0.178 (0.015)	pass
O7	True-level knock-out ≤ -0.08	-0.166	pass
O8	As Q8 with rise ≥ 0.8, $\geq 4\times$ control, top ≥ 0.15	+1.45 (+0.11); 0.393 (0.024)	pass
O9	As Q9	0.895	pass
O10	As Q10	0.158	pass
G1	DED-O ablated, unmarked dedents: ≤ -0.30 and below all twenty random sets	-0.775 [-0.831, -0.713]; random -0.116 to +0.100	pass
G2	Terminal dedents: at most a quarter of G1 in size; open-level mass ≥ 0.90	-0.045 (ratio 0.057); 0.971	pass
G3	Same-level lines: ≥ -0.02	+0.033	pass
G4	Unmarked dedents: mass on the current level rises ≥ 0.08	+0.194	pass

Table 10: Exploratory. Checkpoints of a published training run of the OLMo-2 recipe, from which the released OLMo-2-0425-1B weights do not descend, and the released model (last row), on the same post-cut-off code. Left: top-1 accuracy of the indent at unmarked dedents, terminal dedents and same-level lines. Right, for three heads: attention to the innermost open header's `:\n` at unmarked dedents, and how often the head's most-attended open header is the target level at multi-level dedents. Stage-1 rows are labelled by tokens seen; stage-2 rows by tokens seen in that run.

	top-1 accuracy			attention / pick		
Checkpoint	unmarked	terminal	same	L11H9	L13H1	L15H9
Stage 1 (tokens seen)						
1B	0.331	0.386	0.279	0.01 / 0.24	0.01 / 0.18	0.01 / 0.19
21B	0.582	0.942	0.963	0.35 / 0.37	0.02 / 0.22	0.07 / 0.19
84B	0.727	0.963	0.957	0.34 / 0.40	0.01 / 0.25	0.17 / 0.52
420B	0.764	0.966	0.959	0.28 / 0.61	0.00 / 0.30	0.27 / 0.81
1,007B	0.789	0.966	0.954	0.19 / 0.55	0.01 / 0.24	0.35 / 0.81
4,001B (end)	0.782	0.963	0.969	0.23 / 0.64	0.00 / 0.31	0.29 / 0.82
Stage 2 (three runs from the end of stage 1)						
run 1, +3B	0.687	0.966	0.977	0.21 / 0.55	0.00 / 0.34	0.31 / 0.73
run 1, +51B	0.756	0.956	0.970	0.18 / 0.54	0.00 / 0.34	0.29 / 0.70
run 2, +51B	0.738	0.956	0.969	0.18 / 0.55	0.00 / 0.33	0.29 / 0.69
run 3, +51B	0.727	0.966	0.969	0.18 / 0.54	0.00 / 0.28	0.31 / 0.69
Released	0.767	0.969	0.975	0.03 / 0.39	0.31 / 0.72	0.00 / 0.18

Table 11: Round 3: every re-run and new test on the shareable material, with the original value (round 1, round 2 or the cross-study test) and the round-3 value. Changes are in mean log-probability of the true indent unless stated; the thresholds are those of the original protocols (Tables 4 and 7 to 9). R2 rows give the main tested quantity of each test. Every test passed.

Test	Quantity	Orig.	Round 3	Test	Quantity	Orig.	Round 3
H1	DED, unmarked	−1.091	−0.994	Q1	attention to open header	0.458	0.457
H2	DED, mass kept at level	+0.306	+0.283	Q2	knock-out	−0.148	−0.117
H3	DED, terminal open mass	0.991	0.994	Q3	knock-out, same level	+0.013	+0.012
H4	DED, same level	+0.025	+0.024	Q4	closed header	−0.001	0.000
H5	TERM, terminal	−0.044	−0.032	Q5	picking, not innermost	0.893	0.896
H6	TERM, unmarked	−0.027	−0.017	Q6	steering rise	+2.72	+2.74
H7	both, terminal	−0.188	−0.134	Q7	true-header knock-out	−0.210	−0.195
H8	DED, generated	−0.631	−0.652	Q8	steering, generated	+1.41	+1.39
H9	CARRY, pair necessity	0.650	0.652	Q9	keyword edge share	0.716	0.710
XA3	L17H6, gen. terminal	−0.022	−0.025	Q10	largest layer share	0.098	0.102
XB4	DED, code unmarked	−1.112	−0.801	Q11	L0H2 on the line	+0.029	+0.030
XD3	L11H9, largest |change|	–	0.027	O1	attention (OLMo-2-0425-1B)	0.290	0.283
E1	L11H9, gen. terminal	−0.030	−0.027	O2	knock-out	−0.132	−0.115
E2	CARRY; TERM	–	−0.303; −0.224	O3	knock-out, same level	+0.016	+0.014
E3	without L11H9: weaker by	–	+0.039; +0.052	O4	closed header	−0.004	−0.003
E4	L11H9, unmarked	–	−0.010; −0.012	O5	picking, not innermost	0.694	0.715
E5	L11H9, code terminal	−0.003	−0.003	O6	steering rise	+1.58	+1.61
E6	L11H9, pair necessity	–	0.122	O7	true-header knock-out	−0.166	−0.154
				O8	steering, generated	+1.45	+1.41
				O9	keyword edge share	0.895	0.859
				O10	largest layer share	0.158	0.136
				G1	DED-O, unmarked	−0.775	−0.776
				G2	DED-O, terminal	−0.045	−0.044
				G3	DED-O, same level	+0.033	+0.030
				G4	DED-O, mass kept at level	+0.194	+0.199

The indentation-quiet tests of the cross-study test (XA5, XB1–XB3, XC3, XC6) passed again on the new material; their values are in the round-3 report released with the code.

C Ordinal suffixes: all frozen predictions

Table 12 lists the round-1 predictions, with their values in Tables 13 to 15. Table 16 summarises the round-2 tests by account, with their values in Tables 17 to 19; Table 20 gives the exploratory checkpoint measurements.

Table 12: The frozen predictions and their outcome on held-out material. Effect kept is the share of the teen effect left after mean ablation at the last position; random sets are of the same size and from layers 0–2.

Test	Prediction	Result	
B1	OLMo-2-0425-1B plain 4- and 7-digit teen AUROC ≤ 0.80	0.775, 0.729	pass
B2	the seven cells with both digits in the last token ≥ 0.88	lowest 0.918	pass
B3	separator minus plain ≥ 0.10 at 4 and 7 digits	+0.213, +0.224	pass
B4	AUROC(5)−AUROC(4) and AUROC(6)−AUROC(7) ≥ 0.10; teen accuracy ≤ 0.65 at 4 digits, ≥ 0.85 at 3	+0.223, +0.245; 0.220, 1.000	pass
B5	GPT-2 and Pythia-1B: boundary teen accuracy ≤ 0.50, control ≥ 0.85, AUROC gap ≥ 0.10	0.113 vs 0.950, gap 0.574; 0.175 vs 0.965, gap 0.140	pass
Q1	Qwen3-0.6B-Base: L0H2+L1H6 keep ≤ 0.55 at 3 and 4 digits; every random pair ≥ 0.85	0.430, 0.459; random ≥ 0.994	pass
Q2	teen accuracy falls ≥ 0.20, non-teen ≤ 0.10	0.458, 0.385; 0.000, 0.021	pass
Q3	2 digits: kept ≥ 0.55, teen accuracy after ablation ≥ 0.95	0.709; 0.979	pass
Q4	L0H2 alone keeps ≤ 0.70; L1H6 alone ≥ 0.80	0.583, 0.586; 0.942, 0.946	pass
Q5	L0H2 the top head in teen patching, ≥ 0.40, at 3 and 4 digits	0.599, 0.612	pass
Q6	last-digit patching: $\lvert$L0H2$\rvert \leq 0.05$, top head ≥ 0.30	−0.000, −0.000; 0.598, 0.558	pass
O1	OLMo-2-0425-1B: L0H5+L0H0+L0H9 keep ≤ 0.65 at 4 and 7 digits; every random triple ≥ 0.85	0.110, 0.076; random ≥ 0.921, ≥ 0.856	pass
O2	the same heads keep ≥ 0.90 at 3 and 5 digits	0.995, 0.993	pass

Table 13: OLMo-2-0425-1B on held-out items: teen AUROC with 95% intervals (resampling distinct numbers) and plain-form teen accuracy.

Digits	Plain tokens	Plain AUROC	Separator AUROC	Plain teen acc.	Teen numbers
3	one	1.000 [0.999, 1.000]	–	1.000	7
4	3+1	**0.775** [0.722, 0.825]	0.989 [0.979, 0.996]	**0.220**	71
5	3+2	0.999 [0.997, 1.000]	0.974 [0.959, 0.987]	1.000	169
6	3+3	0.974 [0.960, 0.987]	0.918 [0.890, 0.943]	0.980	193
7	3+3+1	**0.729** [0.679, 0.776]	0.953 [0.934, 0.970]	0.535	200

Table 14: Qwen3-0.6B-Base: mean ablation at the last position on held-out items (96 teen and 96 non-teen numbers per length). Effect kept is relative to the unablated teen effect, given in logits.

Digits	Teen effect	Both kept	L0H2 alone	L1H6 alone	Random pairs	Teen acc. drop	Non-teen drop
2	11.69	0.709	0.873	0.984	0.996–1.058	0.021	0.083
3	5.01	**0.430**	0.583	0.942	0.996–1.217	**0.458**	0.000
4	4.25	**0.459**	0.586	0.946	0.994–1.116	**0.385**	0.021

Table 15: OLMo-2-0425-1B: mean ablation of three layer-0 heads at the last position on held-out items (96 teen and 96 non-teen numbers per length, plain form). Effect kept is relative to the unablated teen effect, given in logits.

Digits	Tens digit	Teen effect	Three heads kept	L0H5 / L0H0 / L0H9 alone	Random triples
3	last token	6.24	0.995	0.997 / 0.999 / 0.997	0.989–1.007
4	previous token	1.21	**0.110**	0.608 / 0.781 / 0.724	0.921–1.047
5	last token	3.71	0.993	0.994 / 1.004 / 1.001	0.994–1.020
7	previous token	0.62	**0.076**	0.629 / 0.715 / 0.745	0.856–1.031

Table 16: Second-round accounts and their prospective tests on held-out material.

Tests	Account	Passed	Verdict
R1–R6	Qwen3-0.6B-Base: L0H2 → MLP0 → MLP24, which decides teen; MLP18 writes the last digit's suffix (4 and 6 digits)	12 of 12	supported
S1–S3	Qwen3-0.6B-Base: later previous-token heads back up the copy, mostly at two digits	7 of 7	supported
O3–O8	OLMo-2-0425-1B: a weak route across the token boundary, not a missing one (3–10 digits)	14 of 14	supported
X	three more one-token-per-digit models: early previous-token heads carry the teen check	6 of 8	partly supported

Table 17: Qwen3-0.6B-Base, held out: teen effect kept after mean ablation at the last position. A = L0H2 + L1H6; B = L2H4 + L5H2; controls are A plus 20 random pairs from layers 2–6.

Digits	Teen effect	A	A + B	A + random pair	B alone	Drop from adding B
2	9.874	0.701	**0.482**	0.685–0.706	0.986	**0.219**
3	3.293	0.415	**0.321**	0.390–0.425	0.951	0.095
4	2.624	0.341	**0.243**	0.334–0.372	0.931	0.098

Table 18: OLMo-2-0425-1B, held out (96 teen and 96 non-teen numbers per length, plain form). "Three heads kept" is the teen effect left after mean-ablating L0H5, L0H0 and L0H9 at the last position; random triples are from layers 0–2.

Digits	Tokens	Teen effect	AUROC	Three heads kept	Lowest random triple	Probe, layer 2 / 15
3	3	4.670	0.999	1.004	0.987	0.989 / 1.000
4	3+1	1.186	**0.730**	**0.234**	0.937	1.000 / **0.854**
5	3+2	3.156	0.971	0.996	0.985	1.000 / 0.995
7	3+3+1	0.625	**0.685**	**0.209**	0.930	0.979 / **0.844**
8	3+3+2	2.452	0.912	0.998	0.984	1.000 / 0.995
9	3+3+3	1.539	0.887	0.994	0.982	1.000 / 0.984
10	3+3+3+1	0.561	**0.640**	**0.242**	0.924	0.989 / **0.734**

Table 19: Other one-token-per-digit models, held out: mean ablation of early heads at the last position. "Pull" is the mean suffix margin of non-teen numbers; random sets are 20 sets of the same size from layers 0–1.

Model	Heads	Digits	Teen effect kept	Pull kept	Lowest random set
Qwen3-1.7B-Base	PREV L0H2+L1H6	3 / 4	0.484 / 0.377	**0.548 / 0.233**	0.994 / 0.998 (teen)
	CUR L0H3+L0H7	3 / 4	1.136 / 1.024	−0.567 / −2.754	0.985 / 0.991 (pull)
SmolLM2-1.7B	PREV L0H6+L0H8	3 / 4	0.729 / 0.661	0.801 / 0.588	0.995 / 1.000 (teen)
Gemma-4-E2B	PREV L0H4	3 / 4	0.801 / 0.653	0.906 / 0.821	0.991 / 0.994 (teen)

Table 20: Stage-one checkpoints of a published OLMo-2-0425-1B training run of the same recipe, from which the released weights do not descend; exploration items (64 teen and 64 non-teen numbers per length). The three heads are those at the released model's positions L0H5, L0H0 and L0H9; their ablation is shown only where the four-digit teen effect is at least 0.30 logits.

	Teen AUROC by digits				Four digits		
Tokens (billions)	3	4	5	7	Teen effect	Three heads kept	Probe, layer 2 / 14
1	0.451	0.493	0.585	0.577	0.00	–	0.39 / 0.41
21	0.789	0.516	0.992	0.459	0.03	–	0.94 / 0.64
84	0.982	0.540	0.959	0.519	0.20	–	0.97 / 0.62
420	0.996	0.569	0.978	0.467	0.49	0.997	0.98 / 0.62
1,007	0.998	0.710	0.982	0.584	0.98	1.003	1.00 / 0.74
4,001	0.998	0.817	1.000	0.616	1.63	1.002	0.99 / 0.87

D Capitals: all frozen predictions

Table 21 lists the round-1 tests, Table 22 the values for head L0H3, and Table 23 the round-2 tests.

Table 21: The frozen predictions and their held-out outcome. Shares are of word starts unless stated; d is the capitals direction and r a random direction of the same norm; ℓ is layer 24 in Gemma-4-E2B and 22 in Qwen3-1.7B-Base; edits use the exploration directions at the prediction position. Held-out values are point estimates with 95% intervals from 2,000 bootstrap resamples of paragraphs (150 paragraphs for the lens, 100 for the edits and recasing). The last column gives the exploration values (Gemma-4-E2B / Qwen3-1.7B-Base). Every prediction passed in both models.

	Measure	Threshold	Gemma-4-E2B	Qwen3-1.7B-Base	Exploration
G1	Gate repeated on held-out paragraphs: same word	≥ 0.60	0.707	0.649	0.697 / 0.620
L1	Lens at ℓ: L_c above U	≥ 0.60	0.707 [0.696, 0.717]	0.753 [0.742, 0.764]	0.705 / 0.748
	Lens at ℓ: L_r above U	≤ 0.30	0.162 [0.154, 0.170]	0.127 [0.121, 0.133]	0.164 / 0.130
	Last layer: L_c above U	≤ 0.05	0.001 [0.001, 0.002]	0.002 [0.001, 0.003]	0.001 / 0.002
	First layer from which $L_c > U$ stays below one half	last five	32	25	32 / 25
L2	L_c above L_r at ℓ, upper-case passages	≥ 0.75	0.852 [0.844, 0.860]	0.878 [0.872, 0.884]	0.852 / 0.871
	… lower-case passages (and within 0.05 of the above)	≥ 0.75	0.852 [0.845, 0.859]	0.890 [0.884, 0.896]	0.853 / 0.885
D1	Add d at ℓ (lower case): flips to capitals	≥ 0.80	0.981 [0.964, 0.995]	0.956 [0.929, 0.978]	0.961 / 0.957
	… same word in capitals	≥ 0.45	0.799 [0.753, 0.841]	0.658 [0.601, 0.716]	0.775 / 0.600
	Subtract d at ℓ (upper case): flips to lower case	≥ 0.80	0.961 [0.940, 0.981]	0.967 [0.948, 0.984]	0.955 / 0.971
	… same word in lower case	≥ 0.45	0.785 [0.742, 0.824]	0.668 [0.616, 0.720]	0.756 / 0.657
	Add r: flips	≤ 0.05	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.000 / 0.000
	Subtract r: flips	≤ 0.05	0.003 [0.000, 0.008]	0.003 [0.000, 0.008]	0.000 / 0.000
D2	Midpoint along d (upper case): still upper case	≤ 0.70	0.540 [0.483, 0.595]	0.527 [0.477, 0.584]	0.524 / 0.465
	… same word in either case	≥ 0.60	0.822 [0.781, 0.860]	0.805 [0.758, 0.849]	0.846 / 0.725
	Midpoint along r: still upper case	≥ 0.95	0.997 [0.991, 1.000]	1.000 [1.000, 1.000]	1.000 / 1.000
D3	Add d in original prose: capitals	≥ 0.75	0.957 [0.934, 0.977]	0.941 [0.913, 0.968]	0.952 / 0.959
	… same word in capitals	≥ 0.50	0.821 [0.776, 0.861]	0.678 [0.621, 0.732]	0.823 / 0.639
	Add r in original prose: capitals	≤ 0.05	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.000 / 0.000
D4	Layer 16: add d, flips	≤ 0.15	0.011 [0.003, 0.022]	0.008 [0.000, 0.017]	0.022 / 0.000
	Layer 16: subtract d, flips	≥ 0.40	0.736 [0.682, 0.784]	0.481 [0.433, 0.530]	0.756 / 0.500
R1	Lower case, last 2 words in capitals: upper-case top-1	≥ 0.50	0.708 [0.664, 0.753]	0.683 [0.631, 0.732]	0.689 / 0.687
	… last 8 words in capitals	≥ 0.85	0.932 [0.904, 0.957]	0.925 [0.897, 0.950]	0.933 / 0.933
	Upper case, last 2 words in lower case: upper-case top-1	≤ 0.25	0.086 [0.057, 0.116]	0.131 [0.097, 0.171]	0.078 / 0.115
	… last 8 words in lower case	≤ 0.10	0.017 [0.005, 0.030]	0.029 [0.013, 0.049]	0.014 / 0.036
E1	Changed word in the lower-case passage's top 5	≥ 0.50	0.713 [0.637, 0.785]	0.676 [0.615, 0.742]	0.762 / 0.694
	Control word in the top 5	≤ 0.10	0.021 [0.000, 0.046]	0.029 [0.006, 0.063]	0.014 / 0.022
H1	Qwen3-1.7B-Base only, head L0H3: see Table 22			passed	

Table 22: Qwen3-1.7B-Base, head L0H3 on held-out paragraphs (50 paragraphs; 186 upper-case positions whose next word is upper case). D is $\log p(U) - \log p(L_c)$; the share of D moved is towards the lower-case passage's own D.

Measure	Threshold	Held out	Exploration
Attention on the current token, upper-case passages	≥ 0.90	0.978 [0.975, 0.981]	0.977
Attention on the current token, lower-case passages	≥ 0.90	0.968 [0.965, 0.972]	0.972
Output from another paragraph in lower case: share of D moved	≥ 0.20	0.352 [0.323, 0.381]	0.344
flips to lower case	≥ 0.25	0.495 [0.425, 0.565]	0.543
Output from another paragraph in upper case: share of D moved	≤ 0.08	0.029 [0.011, 0.049]	0.030
flips	≤ 0.05	0.011 [0.000, 0.027]	0.000
Head L0H2 from the same word's lower-case passage: share of D moved	≤ 0.08	0.011 [0.005, 0.016]	0.021
Head L0H0, the same	≤ 0.08	0.001 [0.000, 0.002]	-0.001
L0H3 from the same word's lower-case passage (no threshold)		0.325 [0.300, 0.352]	0.302

Table 23: Round 2: the frozen predictions and their held-out outcome. u is the unit vector of the capitals direction d; U, L and A are the upper-case, lower-case and context-only passages, and $D = \log p(U) - \log p(L_c)$ as in Table 22. Held-out values are point estimates with 95% intervals from 2,000 bootstrap resamples of articles. "Random" gives the extreme of the twenty matched random sets. The last column gives the exploration values (Qwen3-1.7B-Base / Gemma-4-E2B). Every prediction passed in both models.

	Measure	Threshold	Qwen3-1.7B-Base	Gemma-4-E2B	Exploration
W1	Writers' writes along u moved: upper-case top-1	≤ 0.30	0.029 [0.016, 0.044]	0.009 [0.002, 0.019]	0.021 / 0.000
	… same word in either case	≥ 0.55	0.688 [0.651, 0.724]	0.809 [0.773, 0.845]	0.691 / 0.836
	… random sets, lowest upper-case top-1	≥ 0.85	0.985	0.949	0.968 / 0.964
R1	Add d at ℓ: flips to capitals	≥ 0.80	0.968 [0.952, 0.981]	0.989 [0.979, 0.998]	0.958 / 0.966
	… readers frozen	≤ 0.20	0.002 [0.000, 0.006]	0.000 [0.000, 0.000]	0.000 / 0.000
	… random sets frozen, lowest	see note	0.966	0.985	0.937 / 0.948
C1	Carriers patched in A: excess along u removed	≥ 0.60	0.872 [0.860, 0.884]	0.859 [0.845, 0.874]	0.880 / 0.839
	… excess D removed	≥ 0.30	0.424 [0.414, 0.434]	0.576 [0.562, 0.588]	0.459 / 0.584
	… random sets, highest	see note	0.166	0.322	0.144 / 0.344
C2	Carriers patched in U: upper-case top-1	≤ 0.60	0.309 [0.273, 0.345]	0.255 [0.213, 0.298]	0.351 / 0.291
	… random sets, lowest	≥ 0.90	0.996	0.982	0.989 / 1.000
C3	L17H6 alone in A: excess removed (above every random head)	≥ 0.20	0.363 [0.346, 0.380]		0.359
C4	Local carrier: attention on the last eight words	≥ 0.60	0.929 [0.925, 0.933]	0.918 [0.913, 0.923]	0.929 / 0.914
	Broad carrier: the same	≤ 0.50	0.284 [0.269, 0.299]	0.206 [0.195, 0.218]	0.282 / 0.193
E1	Edit changes the word where real capitals do	see note	0.694 [0.646, 0.744]	0.549 [0.480, 0.619]	0.627 / 0.514
	… where they do not	see note	0.120 [0.096, 0.145]	0.080 [0.059, 0.101]	0.105 / 0.058
	Changed word in the lower-case top 5	≥ 0.40	0.538 [0.483, 0.593]	0.619 [0.543, 0.695]	0.683 / 0.520
E2	Word change, capitals need ≥ 1 extra token	see note	0.541 [0.482, 0.595]	0.424 [0.360, 0.486]	0.432 / 0.406
	… no extra token	see note	0.205 [0.173, 0.236]	0.122 [0.097, 0.147]	0.250 / 0.110
A1	Acronyms: remove d, still all capitals	≥ 0.85	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	0.957 / 1.000
	… remove r	≥ 0.90	0.991 [0.971, 1.000]	0.974 [0.948, 0.994]	1.000 / 1.000
A2	Initial capitals: remove d, still capitalised	≥ 0.90	0.968 [0.958, 0.977]	0.982 [0.974, 0.990]	0.979 / 1.000
	… remove r	≥ 0.95	0.993 [0.988, 0.997]	0.992 [0.987, 0.997]	0.995 / 1.000
A3	Headings in capitals: remove d, still all capitals	≤ 0.85	0.669 [0.609, 0.722]	0.768 [0.717, 0.816]	0.554 / 0.694
	… remove r	≥ 0.95	0.994 [0.985, 1.000]	1.000 [1.000, 1.000]	1.000 / 1.000
A4	Excess along u inside the heading (share of the gap)	≥ 0.40	0.728 [0.697, 0.757]	0.753 [0.728, 0.776]	0.713 / 0.749
	… at the first word after it	≤ 0.25	0.034 [0.027, 0.041]	0.116 [0.108, 0.124]	0.038 / 0.123
	… top-1 all capitals there	≤ 0.10	0.018 [0.004, 0.039]	0.022 [0.006, 0.044]	0.013 / 0.025

Note. R1: every random set at least the unfrozen flip rate minus 0.15. C1: every random set at most the named value minus 0.25. E1: the two change rates differ by at least 0.30. E2: they differ by at least 0.10.

E Shared parts: the full held-out matrices

Each cell gives the named group's value, with the range of its twenty random groups in brackets. "Kept" is the ablated value divided by the unablated one; "change" is ablated minus unablated. Unablated line-breaking values: width effect 1.383, separation 3.284. Post hoc marks, judged on unrounded values: † the named group's value is below the range of its random groups (more harm than any of them), ‡ above it. The frozen tests used the random range only where they predicted an effect, and fixed bounds elsewhere (Section 6).

Line-breaking, Qwen3-0.6B-Base (effect kept).

Group	width effect	separation
COUNT	0.150† (0.892 to 1.071)	0.040† (0.888 to 1.028)
WRITE	0.466† (0.914 to 1.058)	0.254† (0.973 to 1.037)
READ−L17H6	0.568† (0.966 to 1.008)	0.512† (0.990 to 1.008)
L17H6	0.706† (0.988 to 1.036)	0.836† (0.994 to 1.010)
CARRY	0.715† (0.937 to 1.068)	0.864† (0.966 to 1.037)
CARRY−L17H6	1.015 (0.924 to 1.065)	1.033‡ (0.960 to 1.014)
TERM	0.675† (0.958 to 1.061)	0.831† (0.968 to 1.021)
TERM−L17H6	0.966† (0.975 to 1.028)	0.990 (0.978 to 1.025)
DED	0.990 (0.898 to 1.028)	0.999 (0.909 to 1.013)
L0H2	0.973 (0.909 to 1.066)	0.871† (0.959 to 1.018)
L0H3	0.391† (0.969 to 1.044)	0.000† (0.966 to 1.015)
L11H9	1.023‡ (0.985 to 1.006)	1.027‡ (0.993 to 1.001)

Indentation, Qwen3-0.6B-Base (change in mean log-probability of the right indent).

Group	code, unmarked	code, terminal	generated, unmarked	generated, terminal
COUNT	0.007 (−0.007 to 0.011)	−0.001 (−0.001 to 0.001)	0.000 (−0.021 to 0.063)	−0.003 (−0.009 to 0.012)
WRITE	0.001 (−0.003 to 0.001)	0.001‡ (−0.001 to 0.001)	0.001 (−0.005 to 0.025)	−0.001 (−0.012 to 0.005)
READ−L17H6	−0.003 (−0.029 to 0.012)	−0.001 (−0.003 to 0.002)	−0.007 (−0.117 to 0.073)	−0.003 (−0.046 to 0.015)
L17H6	−0.010 (−0.011 to 0.008)	−0.003† (−0.003 to 0.000)	0.021‡ (−0.028 to 0.015)	−0.022† (−0.011 to 0.002)
DED	−1.112† (−0.027 to 0.014)	−0.058† (−0.004 to 0.001)	−0.584† (−0.117 to 0.053)	−0.266† (−0.021 to 0.007)
L0H2	−0.008 (−0.059 to 0.002)	−0.001 (−0.014 to 0.000)	−0.003 (−0.055 to 0.111)	0.000 (−0.043 to 0.001)
L0H3	−0.009 (−0.059 to 0.001)	−0.004† (−0.003 to 0.000)	0.010‡ (−0.055 to 0.002)	−0.007 (−0.024 to 0.001)
L11H9	−0.009 (−0.010 to 0.003)	−0.003† (−0.002 to 0.000)	−0.001 (−0.014 to 0.019)	−0.030† (−0.007 to 0.002)

Ordinal suffixes, Qwen3-0.6B-Base (effect kept).

Group	teen, 4 digits	teen, 5 digits	pull, 4 digits	pull, 5 digits
COUNT	1.013 (0.929 to 1.045)	1.009 (0.916 to 1.030)	1.008 (0.902 to 1.133)	1.000 (0.932 to 1.028)
WRITE	0.999 (0.915 to 1.034)	0.999 (0.957 to 1.032)	0.997 (0.951 to 1.021)	0.994 (0.992 to 1.039)
READ	1.047‡ (0.966 to 1.038)	1.021 (0.954 to 1.064)	1.113‡ (0.901 to 1.096)	1.055 (0.880 to 1.108)
READ−L17H6	1.065‡ (0.984 to 1.040)	1.015 (0.979 to 1.106)	1.111‡ (0.948 to 1.066)	1.040 (0.968 to 1.064)
L17H6	0.990 (0.973 to 1.014)	1.012 (0.954 to 1.019)	1.002 (0.904 to 1.032)	1.010 (0.899 to 1.025)
CARRY	0.714† (0.912 to 1.069)	0.757† (0.879 to 1.086)	0.742† (0.814 to 1.163)	0.675† (0.832 to 1.126)
CARRY−L17H6	0.723† (0.953 to 1.076)	0.742† (0.962 to 1.077)	0.736† (0.945 to 1.144)	0.653† (0.945 to 1.118)
TERM	0.704† (0.842 to 1.091)	0.785† (0.803 to 1.075)	0.708† (0.754 to 1.127)	0.662† (0.847 to 1.136)
TERM−L17H6	0.714† (0.930 to 1.046)	0.771† (0.887 to 1.146)	0.697† (0.865 to 1.120)	0.631† (0.894 to 1.140)
DED	0.976 (0.957 to 1.066)	1.005 (0.948 to 1.081)	0.932 (0.879 to 1.140)	0.977 (0.884 to 1.111)
L0H2	0.507† (0.988 to 1.033)	0.546† (0.973 to 1.002)	0.823† (0.895 to 0.994)	0.901 (0.784 to 1.000)
L0H3	1.077‡ (0.988 to 1.003)	1.132‡ (0.992 to 1.005)	0.155† (0.980 to 1.000)	0.314† (0.988 to 1.001)
L11H9	0.726† (0.988 to 1.029)	0.738† (0.973 to 1.021)	0.733† (0.981 to 1.046)	0.643† (0.973 to 1.013)

Capitals, Qwen3-1.7B-Base, heads by index (D kept; word log-probability change).

Group	D_{up} kept	D_{low} kept	word log-prob. change
COUNT	1.009 (0.989 to 1.014)	1.009 (0.989 to 1.016)	$0.000^{\ddagger}$ (-0.019 to -0.003)
READ	$0.931^{\dagger}$ (0.993 to 1.017)	$0.923^{\dagger}$ (0.950 to 1.021)	-0.013 (-0.039 to 0.000)
L17H6	$0.927^{\dagger}$ (0.995 to 1.017)	$0.957^{\dagger}$ (0.996 to 1.019)	-0.002 (-0.014 to 0.000)
CARRY	$0.860^{\dagger}$ (0.993 to 1.018)	$0.880^{\dagger}$ (0.984 to 1.015)	-0.062 (-0.073 to -0.010)
DED	0.996 (0.966 to 1.025)	0.999 (0.977 to 1.020)	-0.026 (-0.083 to -0.009)
L0H2	1.004 (0.996 to 1.007)	$0.970^{\dagger}$ (0.992 to 0.999)	$-0.038^{\dagger}$ (-0.027 to 0.000)
L0H3	$0.923^{\dagger}$ (0.997 to 1.003)	$0.331^{\dagger}$ (0.994 to 1.000)	$-1.816^{\dagger}$ (-0.027 to 0.001)
L11H9	$0.947^{\dagger}$ (0.998 to 1.008)	$0.936^{\dagger}$ (0.996 to 1.007)	$-0.033^{\dagger}$ (-0.007 to 0.002)

F Process notes

- *Freezes.* Each charter and protocol was frozen, with file digests, before the material it governs was read, in November 2026. Indentation: round-1 charter on the 7th and protocol on the 8th; round-2 charter on the 8th and protocol on the 9th. Ordinals: round-1 protocol on the 8th; round-2 charter on the 8th and protocol on the 9th. Capitals and the cross-study test: charters and protocols on the 7th and 8th. Each protocol was run once, after its freeze.
- *Indentation, round 1.* A trial of the behaviour script on 30 programs read the generated exploration part before that part's digests were recorded; no held-out file was opened. One attention scan on exploration material ran before the natural-code gate result was known, although the charter placed internal analysis after the gate; the remaining internal analyses waited for the gate.
- *Indentation, round 2.* The confirmation ran once, in several sessions that resumed from saved records; every script covered all of its material. The four stage-2 OLMo-2 checkpoints were downloaded after the protocol was frozen, for the exploratory checkpoint analysis only, and the weight comparison was made after all three stage-2 runs were found to keep L15H9.
- *Ordinals.* A dated amendment, written before any model read a number above 9,999, extended the hash split to seven-digit numbers. The second round's held-out date frames remain unread. Post hoc analyses are the make-up of the GPT-2 and Pythia-1B boundary sets, exploration AUROCs recomputed with ties counted as one half, and the weight comparison; the checkpoint analysis is exploratory.
- *Capitals.* Round 2 used the round-1 confirmation part as exploration material, and the reserve part, unread until then, as held-out material. No run needed repeating in either round.
- *Indentation, round 3.* Charter and protocol were frozen on 10 November and run once, on the 10th and 11th, by byte-identical copies of the frozen scripts. The material was fetched and sealed before any model read it; repositories used by any earlier round were excluded, and files whose own header named another licence were dropped. The new tests' margins were set from runs on material already read in earlier rounds.
- *Cross-study test.* Capitals used Qwen3-1.7B-Base by head index under the charter's fallback rule after Qwen3-0.6B-Base failed the capitals gate. The ordinal reserve held too few three-digit teens, so four- and five-digit numbers were used. One exploratory all-position run was left unfinished; the protocol did not use it. The marks against the random ranges, and the reading of L11H9 by the criterion frozen for L17H6, were made after the run and are post hoc; they change no verdict. The reading of L11H9 was then tested prospectively in round 3.