DECIDED WHILE READING: WHY A LANGUAGE MODEL STOPS USING A FACT PLACED BEFORE ITS RULE

Anonymous authors
Paper under double-blind review

ABSTRACT

Language models are often asked to apply a rule to a few facts and give an immediate verdict, as guardrails, classifiers and judges do. Such prompts place the rule before, among or after the data. They assume that a model with the whole prompt in view applies the rule to the fact it names, wherever the rule appears. We find that this assumption fails. Across 18 open models, moving one unchanged rule sentence just past the fact it names, or after all the facts, sharply reduces that fact's influence on the answer, and the answer follows the facts read after the rule instead. Yet the models still know the fact. Asked separately, they mostly answer correctly what the rule compares and whether the fact matches. Causal experiments explain the gap. As a model reads each fact after the rule, it forms the rule's outcome at that fact's own position, and its direct answer later reads the outcome mainly from those positions. A fact read before the rule has no outcome there, and the answer, which sees both the rule and the fact, does not make up for it. In short, the verdict is decided while reading. For systems that answer directly, the rule must come before the data it governs, and explaining before answering restores the lost fact.

1 INTRODUCTION

Many deployed uses of language models give the model a rule and a few facts and ask for an immediate verdict. A guardrail checks a request against a policy (Inan et al., 2023), a classifier routes a ticket, and a model grades an answer against a rubric (Zheng et al., 2023). The rule may sit in the system prompt, before the data, after it, or among it. The usual picture is that the model has the whole prompt in view when it answers, so the rule's position should matter only through the distance and salience of each part, as in position effects on retrieval (Liu et al., 2024a) and instruction placement (Liu et al., 2024b). We find that position matters in a more basic way. A rule governs the facts that a model reads after it.

Figure 1 shows the effect. A request lists three facts, a desk, a room and a slot, each with an assigned and a requested value. The rule says that the action is PERMIT if the requested desk matches the assigned one and BLOCK if it differs. With the rule written first, the model sees that the desks differ and answers BLOCK. Moving the same rule one sentence later, just past the desk, makes it answer PERMIT. It now applies the rule to the room and the slot, which match. Yet the model has not lost the desk. Asked separately on the same prompt, it says that the rule is about the desk and that the desks differ. Asked to give its letter first and explain afterwards, it gives the wrong letter and then derives the correct action from the desk. We call a rule placed after the fact it names a *late rule*.

The effect is general. We move one unchanged rule sentence among three facts and enumerate every combination of which facts hold, so that each placement of the rule yields the model's complete answer table (Section 2). We measure how much the named fact still drives the answer, on a scale where 100 means that it decides the answer as the rule says. Across 18 open models from eight families, this influence falls from 76 to 34 when the rule moves from just before the fact to just after it, and it falls in every model. We call this gap *the step*. The same loss appears when the rule comes after all the facts, both in these requests and in support tickets written as realistic forms (Section 3). Distance to the answer does not explain it, because a fact read just after a rule at the start of the

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

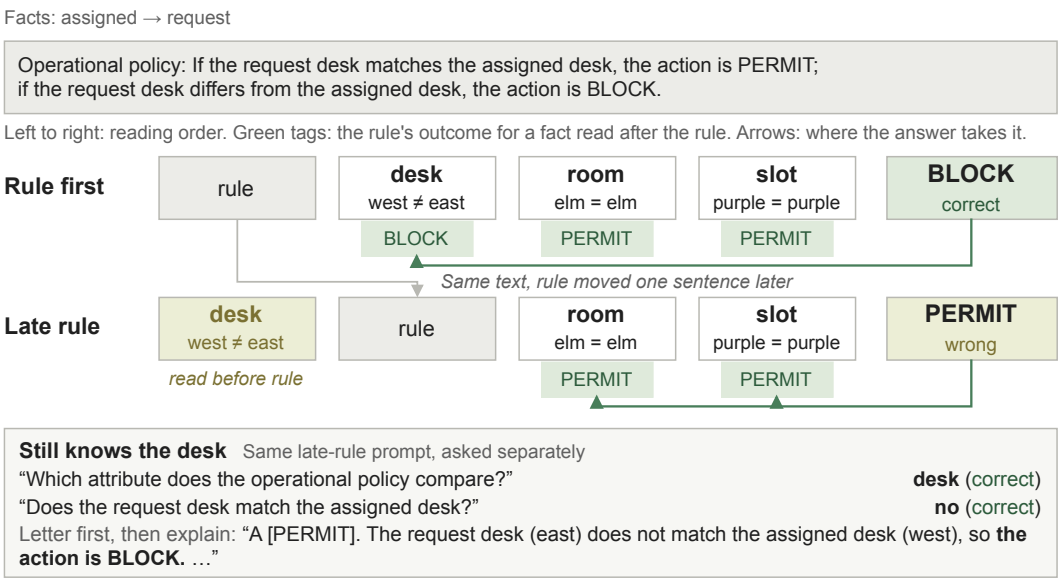


Figure 1: Moving the rule one sentence later makes the model stop using the fact it names, although the model still knows that fact. Each row is the same request in reading order, and only the rule sentence moves. Green tags show our account of where the rule's outcome for a fact is available. A fact read after the rule carries it at its own tokens, and a fact read before the rule carries none. Arrows show where the direct answer takes the outcome from, so with the late rule the answer follows the room and the slot and becomes PERMIT (Table 1 gives the full answer tables, and Section 4.2 tests the tags and arrows). On the same late-rule prompt, the model still answers questions about the desk correctly (bottom). Example from Qwen3-14B, answering with one letter.

prompt, far from the answer, is used more than a fact read just before a rule at the end, right next to it. Nor is the late rule ignored, since the answer applies it to the facts that follow.

Why would a model not use a fact that it knows? Our answer is that the verdict is decided while reading (the tags and arrows in Figure 1). When a model reads a fact with the rule already in view, it works out the rule's outcome for that fact at the fact's own tokens. Reading the desk after the rule, it registers that the desks differ, so the action is BLOCK. The direct answer later takes the outcome mainly from there. We call this *fact-side reading*. A desk read before the rule has no outcome at its tokens, because those tokens could not attend to a rule that had not yet appeared. This alone need not cost the fact, since the answer can attend to both the rule and the desk. But the answer mostly does not combine them afresh. It takes the outcomes of the facts read after the rule instead, the room and the slot, even though the rule does not name them. A weaker route through the rule's own tokens also exists, and the fact keeps part of its influence. Fact-side reading places rules among the in-context computations that happen as text is read, such as the binding of attributes to entities (Feng & Steinhardt, 2024) and the evaluation of list items against a question (Sen Sharma et al., 2026).

Two tests support this account in the models that show the step (Section 4.2). Re-reading only the named fact right after the rule restores its use, while the same sentence placed just before the rule does not. Transplanting hidden states moves the answer much more when the states are written into the fact's tokens than into the rule's tokens. Llama-3.3-70B, which can do the task but has almost no step, shows the opposite. It takes a late rule's outcome mainly from the rule (Section 4.3). The architecture therefore allows another way of reading, and causal attention does not impose fact-side reading.

Our contributions are three.

(1) **A placement law for rules.** In 18 models and in realistic tickets, a rule governs the facts read after it, while the models still know the fact they stop using (Sections 3 and 4.1).

(2) **An account of the law, tested causally.** The direct answer takes a rule's outcome mainly from the facts' own tokens (Section 4.2).

Table 1: The answer tables behind Figure 1. The rule gives PERMIT when the desks match. Each column sets which facts match, and the rows give the model's direct answer with the rule read first and with the late rule. With the rule first, the answer follows the desk. With the late rule, it is PERMIT exactly when the rooms and the slots both match. The last column gives each fact's influence (defined below).

Desks match	no	yes	no	yes	no	yes	no	yes	Influence of
Rooms match	no	no	yes	yes	no	no	yes	yes	desk / room / slot
Slots match	no	no	no	no	yes	yes	yes	yes	
Rule first	BLOCK	PERMIT	BLOCK	PERMIT	BLOCK	PERMIT	BLOCK	PERMIT	100 / 0 / 0
Late rule	BLOCK	BLOCK	BLOCK	BLOCK	BLOCK	BLOCK	PERMIT	PERMIT	0 / 50 / 50

(3) **A model that reads at the rule.** Llama-3.3-70B shows that fact-side reading is not imposed by the architecture (Section 4.3).

The account also tells us what helps. Explaining before answering removes the step, because the explanation restates the fact after the rule. A correction of the rule, by contrast, has little effect once the facts have been read (Section 5). For guardrails, classifiers and judges that answer directly, the rule belongs before the data it governs, not among or after it.

2 MEASURING WHICH FACT A RULE GOVERNS

Three facts, one rule, one moving sentence. Each request pairs one rule with three facts, and only the rule moves. The facts come in a fixed order, and each gives an assigned and a requested value for one attribute (`Assigned desk: east. Request desk: west.`). The rule, an "operational policy", names one of the three attributes and maps "differs" and "matches" to PERMIT and BLOCK, and an output mapping assigns the answer letters. The rule can stand before all facts or after the first, second or third fact, and it can name any of the three, which gives twelve combinations that we call *layouts*. Two layouts that name the same fact differ in nothing but the place of the rule sentence. Within each layout we counterbalance six factors, which yields 192 structures. The rule maps "differs" to PERMIT or to BLOCK, its two clauses come in either order, and the letters are A/B with PERMIT first or X/Y with PERMIT second. The attributes come in six orders, either value can be the assigned one, and the value words are paired with the attributes in both ways. Every structure is run with all eight truth assignments of the facts (below), so that a model answers $12 \times 192 \times 8 = 18{,}432$ prompts per *vocabulary*, a set of value words such as *west* and *east*. We report the first of two vocabularies and give the second in the appendix. Models answer directly with one letter, and we read the next-token logit difference between the two instructed letters.

Complete answer tables and influence. For each structure we enumerate all eight truth assignments of the three facts, where a fact is "true" when it satisfies the rule's PERMIT condition. The model's eight answers form a complete answer table, which says exactly which facts the answer depends on (Table 1). The *influence* of a fact is the share of its four false-to-true flips that change the answer from BLOCK to PERMIT, minus the share that change it from PERMIT to BLOCK, in points. Writing $y(t)$ for the answer to truth assignment t and (t_f^0, t_f^1) for the four pairs of assignments that differ only in fact f, false in t_f^0 and true in t_f^1,

$$\begin{aligned}\mathrm{infl}(f) = 25 \sum_{(t_f^0,\, t_f^1)} \mathbf{1}\big[y(t_f^0) = \textsc{block},\ y(t_f^1) = \textsc{permit}\big] \\ - 25 \sum_{(t_f^0,\, t_f^1)} \mathbf{1}\big[y(t_f^0) = \textsc{permit},\ y(t_f^1) = \textsc{block}\big].\end{aligned} \tag{1}$$

Influence thus runs from 100, for a fact that decides the answer as the rule says, through 0, for a fact that does not matter or an answer drawn at random, to -100, for an answer that follows the fact against the rule. In the running example, the desk's influence falls from 100 to 0 when the rule moves past it, while the influence of the room and of the slot rises from 0 to 50 each. For a fact that the rule does not name, a positive influence therefore means that the model applies the rule's comparison to it.

Table 2: The placement law, where the answer takes the rule's outcome from, and what restores the fact. Entries are the influence of a fact on the answer (100 = decides it as the rule says, 0 = no net effect), medians over models, first vocabulary. A is the fact the rule names, and B and C follow it. An arrow marks the change from a baseline, and the last column counts the models in the stated direction. A transplant writes a token span's states from another prompt, the donor, at one layer (Section 4.2).

Test	Contrast	Result	Models
I. The placement law (18 models, 8 families)			
Step at the rule	A with the rule just before it vs. just after it	76 vs. 34	18/18
Redirection	rule moved past A, change in A / B / C	-46 / $+18$ / $+13$	18/18
Support tickets	rule after the named field / the ticket (change)	-28 / -14	17/18, 14/18
II. Where the fact must be read (models with the step)			
Re-reading	rule after A + copy of A after / A before / B after	22 $\rightarrow$ 80 / 23 / 21	12/12
III. Where the answer takes the outcome from (models with the step; transplants at one layer per model)			
Outcome at A	A's post-rule states into A's tokens	12/3/25 $\rightarrow$ 54/71/63	3/3
	opposite-value states into A's tokens	donor 33–82	10/10
A vs. rule	largest change in A, into A's vs. rule's tokens	-62 vs. -23	10/10
Llama-3.3-70B	same test	-29 vs. -89	reversed
IV. What restores the fact (models with the step)			
Explaining first	rule after A, reasoning before vs. after the letter	97 vs. 28	11/11
Correction	correction read before vs. after the facts	25 vs. 5	12/12

Controls. Over the full grid of rule and fact positions, every cell with A before the rule is below every cell with A after it (Figure 2b). A's states read after a same-noun sentence that is not a rule restore little (20, 8 and 28), and opposite-value states of A read before the rule change A's influence much less than those read after it (median -27 against -62). Models of part III are listed in Figure 4b.

Because an exact tie gives no answer, we leave out the structures in which any of the eight answers ties (5.8%).

Models. We test 18 open instruction-tuned models from eight families: Qwen3 4B to 32B including the 30B mixture-of-experts model (Qwen Team, 2025), Gemma-3 4B to 27B (Gemma Team, 2025), Llama-3.x 3B to 70B (Grattafiori et al., 2024), R1-Distill-Llama-70B (DeepSeek-AI, 2025), Phi-4 (Abdin et al., 2024), Mistral-Small-24B (Mistral AI Team, 2025), OLMo-2 13B and 32B (OLMo et al., 2024), Granite-4-H, a hybrid attention and state-space model (IBM Research, 2025), and GLM-5.3-Flash (GLM-5-Team et al., 2026). Thirteen of them can do the task, with an influence of the named fact of at least 61 when the rule comes first. Twelve of these show a step of at least 15 points in both vocabularies, and we call them the models with the step. The thirteenth, Llama-3.3-70B, has a step of 2.5 and serves as the contrast in the tests of Sections 4.1 to 5. The other five cannot do the task even with the rule first (25 to 47).

Support tickets. To test the law outside this format, we write support tickets for three scenarios, a hotel room change, a parcel return and a travel expense claim. Each ticket names a customer, places unrelated notes between the facts, and writes the facts as form fields (`Ordered size: medium. Delivered size: large.`). A company policy maps to ESCALATE or CLOSE and appears in the system prompt, at the start of the message, right after the named field, or after the whole ticket.

3 A RULE GOVERNS THE FACTS READ AFTER IT

The named fact is used much more when it is read after the rule than when it is read before it (Table 2, first row, and Figure 2a). With the rule just before the named fact, its influence is 76, and with the rule just after it, 34. Both values average each model over the three positions of the named fact and then take the median over models. Because only the rule sentence moves between the two layouts, the drop cannot come from a change in the fact or in what follows it. The gap appears in all 18 models and in both vocabularies (79 and 33 in the second).

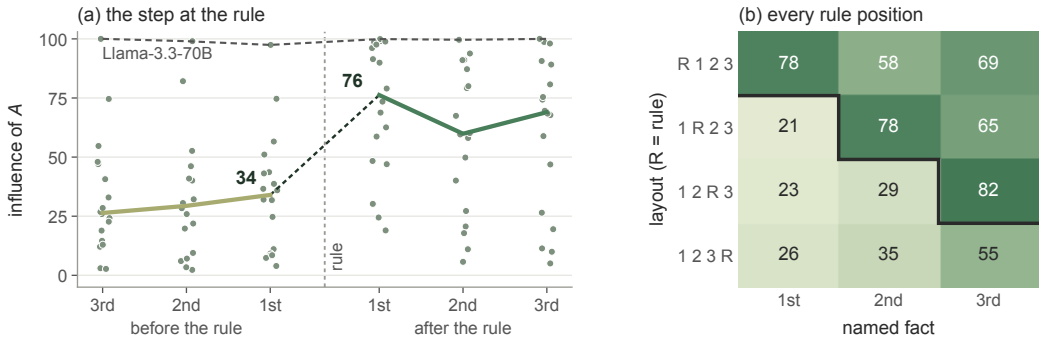


Figure 2: The named fact is used when it is read after the rule and much less when it is read before it. (a) Influence of the named fact A against its position relative to the rule, counted in facts from the rule. Points are the 18 models and lines their median. The dashed line is Llama-3.3-70B. The two positions next to the rule hold the same fact, and only the rule moves. (b) Median influence over the full grid of rule positions (rows, R marks the rule) and named positions (columns). Above the staircase the named fact is read after the rule. First vocabulary, the second is in Appendix A.

Distance to the answer does not explain the step. If it did, the named fact would be used more the closer it sits to the end of the prompt. The full grid of four rule positions and three named positions shows otherwise. Every cell with the named fact before the rule (21 to 55) is below every cell with it after the rule (58 to 82), in both vocabularies (Figure 2b and Appendix A), and the first fact after a rule at the start, farthest from the answer, is used more (78) than the last fact before a rule at the end, right next to it (55).

A late rule is applied to the facts that follow it. Because influence is signed by the rule's mapping, a positive influence of an unnamed fact means that the answer applies the rule's comparison to that fact. Moving the rule from before the named fact A to between A and the next fact B lowers A's influence by 46 and raises the influence of the two later facts by 18 and 13 (medians, the same direction in all 18 models). The gain goes to the facts read after the rule. B gains 19 when it is read just after the rule and only 9 when it is read just before it (14 of 18 models). The complete answer tables show this redirection directly. Tables that equal exactly "A" fall from 63% to 7% of all structures, while tables that depend on B and C alone rise from 0.1% to 21%, as in Table 1 (Appendix A). One could object that a rule placed among the facts reads as a heading for those that follow. But this rule names the desk, as the models themselves say when asked (Section 4.1), so applying it to the room and the slot is an error.

The law also holds outside the synthetic format. In support tickets, a rule inserted right after the named field costs that field 28 points against a rule at the start of the message (17 of 18 models), and a rule after the whole ticket costs 14 (14 of 18, Table 2 and Appendix E). The loss remains when the fields are paraphrased as free prose (15 points, 16 of 18), and the step remains when the model decodes to the end or answers with the action words instead of letters (Appendix B). The answer, then, uses the facts read after the rule. Why does it not use the others?

4 Why the named fact goes unused

4.1 The model still knows the fact

The first possibility is that the model loses the fact, and it does not. On the same prompts, we ask separately which attribute the rule compares, whether the requested and assigned desk match, and what the action is. In the twelve models with the step, moving the rule past A lowers action accuracy from 94% to 61% (medians), while the answer to "which attribute does the rule compare?" goes only from 97% to 95%, and "do the desks match?" from 99% to 87%. Even on the prompts with a wrong action, the model answers both questions correctly in a median 76% of cases (Figure 3 and Appendix C).

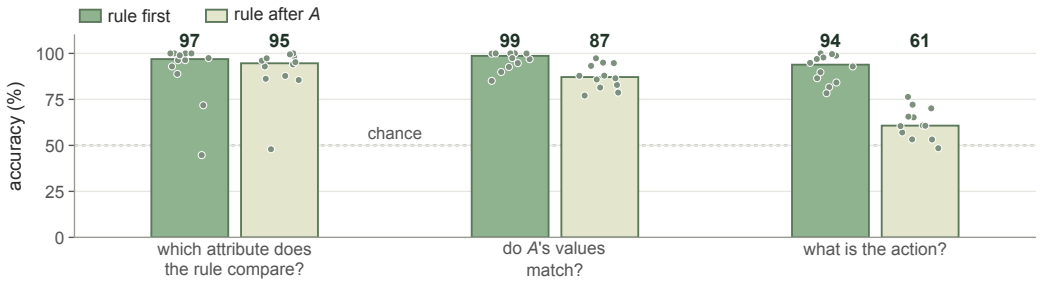


Figure 3: Moving the rule past A breaks the action while the model still knows what the rule compares and whether A matches. Accuracy of three separate questions asked on the same context, with the rule first and with the rule after A. Bars are medians over the 12 models with the step, and points are single models. The low outlier on the first question is Llama-3.1-8B, which is at chance on it in both layouts. First vocabulary.

The model's own words show the same split. When it gives the letter first, its explanation of a wrong letter states the desk comparison correctly in 66% to 100% of cases (eleven models, Figure 1 and Appendix D). The model holds the fact, then, and the failure lies between the fact and the answer. That leaves two places where the answer could find the rule's outcome for the desk, the desk's own tokens or the rule's tokens.

4.2 The answer takes the rule's outcome from the fact

The two places predict different results. If the answer takes the outcome from the fact's tokens, where it exists only when the rule was read first, then reading the fact again after the rule should restore its use, and the fact's post-rule states should carry what the answer needs. If it takes the outcome from the rule's tokens, the rule's states should carry it instead. Our account is the first, with a weaker contribution from the second (Figure 1), and we test it by re-reading the fact in all twelve models with the step and by transplanting hidden states in ten of them (Appendix H).

Re-reading only the fact after the rule restores it. With the rule between A and B, the influence of A is 22. A copy of A's sentence inserted right after the rule raises it to 80 (median over twelve models). Two controls show that the gain comes from reading A after the rule. The same copy inserted right before the rule adds the text without the rule in view and leaves the influence at 23, and a copy of B inserted after the rule adds a post-rule sentence that is not A and leaves it at 21. In all twelve models, the copy of A after the rule raises A's influence, and by more than the same copy before the rule (Figure 4a and Appendix F).

The fact's post-rule state carries what the answer uses. We transplant hidden states over A's tokens (Meng et al., 2022; Zhang & Nanda, 2024), one layer at a time, and report the layer with the largest effect (Appendix H). The recipient has the rule between A and B, padded so that A's tokens sit where they sit in the rule-first prompt, and we call the prompt that supplies the states the *donor*. With the rule-first prompt as donor, where A was read after the rule, the transplant restores A's influence in all three models tested (Table 2). A's states read after a sentence with the same nouns that is not a rule restore little, so the effect needs the rule and not only its words. A stronger test takes the donor from the rule-first request in which A has the opposite value. The answer then follows the donor's A instead of its own. The donor's A reaches an influence of 33 to 82 across the ten models tested, and the recipient's A loses 62 points (median). The same opposite-value A read before the rule lowers it by only 27, so what the answer takes from A's tokens forms after the rule.

The rule's own tokens matter much less. The rule's tokens are read after A and could hold the rule's outcome for A. Transplanting into them the rule's states from the request with the opposite A and the same layout, whose rule has read that A, moves the answer much less than the transplant into A's tokens at the same layers. The largest change in A's influence is -62 at A against -23 at the rule (medians over ten models), larger at A in all ten models and in all nine run with the second

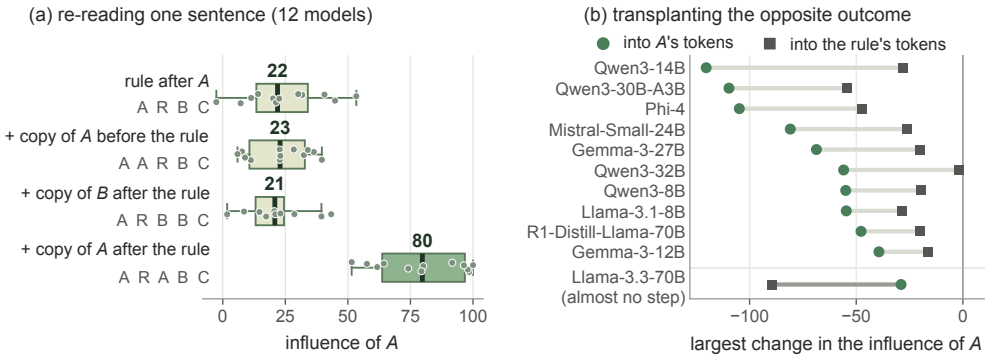


Figure 4: Where the fact must be read, and where the answer takes the outcome from. (a) Influence of the named fact A with the rule after A, and after inserting a copy of A before the rule, a copy of B after it, or a copy of A after it (R marks the rule). Only the copy of A read after the rule restores A. Boxes span the middle half of the models, bars inside mark the medians and points are single models. (b) Largest change in A's influence when states carrying the opposite value are transplanted into A's tokens or into the rule's tokens, one row per model. Influence runs from -100 to 100, so a change can exceed 100. Models with the step move with the transplant at the fact, and Llama-3.3-70B moves with the transplant at the rule.

vocabulary (Figure 4b). After the transplant into the rule's tokens, the answer follows the donor's value with an influence of at most 27. Both effects are partial, since the transplant into A's tokens moves the answer about half the time and depends on the layer (Appendix H), and a linear probe in Qwen3-8B agrees with them (Appendix I).

4.3 LLAMA-3.3-70B TAKES THE OUTCOME FROM THE RULE

Llama-3.3-70B is the one model that can do the task and has almost no step (99.9 with the rule just before the named fact, 97.4 with it just after). Its transplants reverse the pattern of the other models. States transplanted into the rule's tokens change A's influence by -89, and states transplanted into A's tokens by -29, so Llama-3.3-70B takes a late rule's outcome from the rule. The difference does not lie in the architecture or the pretrained weights, since R1-Distill-Llama-70B, the same model fine-tuned on reasoning traces and run here without reasoning, has the step (73 against 44). Nor does this reading reach every format. In the support tickets, whose facts are separated by notes, Llama-3.3-70B loses 22 points when the rule is inserted after the named field (Appendix J). The step is therefore not imposed by causal attention, although the ten models with the step that we tested use the rule-side route only weakly.

5 WHAT RESTORES THE FACT

Explaining before answering removes the step. When the instruction asks for one or two sentences of reasoning before the letter, the named fact's influence with the rule between A and B is 97, and when it asks for the letter first and the explanation after, it is 28, close to the 22 of a direct answer (medians over eleven models with the step, all eleven in this direction, Figure 5a and Appendix D). The explanation restates the fact after the rule. On the prompt of Figure 1, Qwen3-14B answers BLOCK when it explains first, and when asked for the letter first it answers PERMIT and only then writes the same comparison. Other ways of reading the fact after the rule help as well. Step-by-step reasoning (Wei et al., 2022; Kojima et al., 2022) and native thinking remove the step in a two-fact version of the task (Appendix K), repeating the whole prompt restores the fact (Appendix F), and so does stating the fact as a verdict in the rule's own words (Appendix G).

A correction of the rule, by contrast, has little effect once the facts have been read. When the rule names one attribute and a later sentence replaces it with another ("replace *slot* with *room*"), the corrected attribute's influence is 5 if the correction comes after the facts and 25 if the same sentence comes before them (medians over the twelve models with the step, all twelve in this direction,

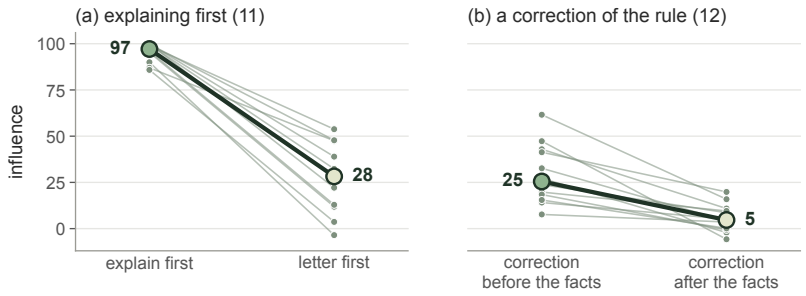


Figure 5: Explaining before answering restores the named fact, and a correction reaches the facts only when it is read before them. Thin lines join the two values of one model, and the thick line joins the medians. (a) Influence of A with the rule between A and B, when the instruction asks for one or two sentences of reasoning before the letter or for the letter first. (b) The rule names one attribute, and a later sentence replaces it with another. Influence of the new attribute when that sentence is read before the facts or after them. First vocabulary.

Figure 5b). Combining a rule with its correction is harder in either position than naming the attribute from the start (68), but only a correction read before the facts reaches them. A correction read after the facts comes too late to be applied to them.

6 Related work

Order effects in prompts. Order is known to matter. Accuracy depends on where relevant information sits in a long context (Liu et al., 2024a), on whether instructions come before or after the input (Liu et al., 2024b), on the order of premises (Chen et al., 2024) and of few-shot examples (Lu et al., 2022; Zhao et al., 2021), and on the position of answer options (Zheng et al., 2024). LLM judges change their verdicts when two candidates swap places (Wang et al., 2024; Shi et al., 2025) and favor score options at particular positions (Xu et al., 2026). These studies measure accuracy or bias as a function of position. We ask which of several facts the answer uses when a rule names one of them, and why. We found no study that places a rubric before or after the response it grades, a choice our results bear on.

Causal attention and repetition. In a decoder, text read early cannot attend to text that comes later. Position bias has been traced to causal attention and position embeddings and removed by letting documents attend to each other (Wang et al., 2025), and an order effect in multiple-choice prompts has been traced to options read before the context (Ok & Lee, 2026). Repetition is the usual remedy. Repeating the question (Xu et al., 2024), the options (Ok & Lee, 2026) or the whole prompt (Leviathan et al., 2025) improves answers, and a second copy lets the representation of early text reflect what follows (Springer et al., 2025). Our re-reading test shows that the copy helps only when it is read after the rule (Section 4.2). Causal attention is also not the whole story, because one model takes the outcome from the rule (Section 4.3).

Computation during reading. Models do much of their in-context work while they read. They bind attributes to entities (Feng & Steinhardt, 2024), track the states of entities (Kim & Schuster, 2023), retrieve bound information through lookbacks (Prakash et al., 2026), pull a subject's attributes from its own tokens (Geva et al., 2023), and evaluate list items against a question either while reading them or at the answer (Sen Sharma et al., 2026). Induction heads let later tokens copy from earlier ones (Olsson et al., 2022), and a few heads carry a representation of the task (Todd et al., 2024) that is recognized at a particular depth (Sia et al., 2024). Timing matters as well. Multi-hop answers fail when an intermediate result is computed too late in depth (Biran et al., 2024). We add rules to this picture, with a failure of timing in position.

Rules and instructions inside the model. For a single in-context conditional rule, the predicate is tested at the queried value before the answer is routed (Thavarasa & Sukumar, 2026), and circuit

analyses of propositional reasoning trace how rules and facts are retrieved (Hong et al., 2025). Instructions appear as directions in the model's states that steer (Stolfo et al., 2025) or predict (Heo et al., 2025) whether it complies. In single-item judgments, blocking the flow from an instruction to the item's tokens barely changes behavior (Waldis et al., 2026). Li (2026) read a decision from notes that a model writes after a field, and the cached states of an operation act mainly as an address for what is read at the answer (Wu & Shomali, 2026). In behavior, models break simple stated rules (Mu et al., 2023), separate instructions from data poorly (Zverev et al., 2025), and can be trained to rank instructions by their source (Wallace et al., 2024). These studies fix the rule's position and mostly test one item. We move the rule among several facts that compete for it, and we compare the facts' tokens with the rule's tokens directly (Section 4.2).

Knowing without using. Large models are largely calibrated about whether their answers are true (Kadavath et al., 2022), truth can be read from their activations (Burns et al., 2023; Azaria & Mitchell, 2023), and shifting a few heads toward truthful directions changes what they say (Li et al., 2023). Models often encode information they do not use in their answers (Orgad et al., 2025; Li et al., 2026), the correct binding can be decoded from the states of items they answer wrongly (Ravulapalli et al., 2026), and stated reasoning need not reflect the computation behind an answer (Turpin et al., 2023; Lanham et al., 2023). We give one reason a known fact goes unused, its position relative to the rule, and a handle that restores it.

Human reading. People read incrementally too. Readers build a representation of a text as it arrives (Kintsch, 1988), and a word heard first shapes how the words after it are processed (Altmann & Kamide, 1999). A passage is understood better when its topic is given before it (Bransford & Johnson, 1972), questions placed before a text change what readers take from it (Rothkopf, 1966; Frase, 1967), and an instruction prepares how later input is handled (Hommel, 2000; Monsell, 2003). The models we study behave like readers who apply a rule only to what they read after it.

7 Conclusion and limitations

We moved one unchanged rule sentence among the facts it governs and recorded each model's complete answer table. Across 18 open models, a rule placed after the fact it names arrives too late. The model stops using that fact, although it still knows the fact when asked separately. Re-reading and transplant experiments trace this to how the rule is applied. As a model reads each fact with the rule in view, it works out the rule's outcome at that fact, and its direct answer later takes the outcome from there. A fact read before the rule has no outcome to give, and in this sense the verdict is decided while reading. This way of reading is common but not forced, since Llama-3.3-70B takes a late rule's outcome from the rule itself.

The account reaches beyond prompt design. If a model settles each fact as it reads it, then an evaluation that states its criterion after the material it grades tests the model under a late rule, as our tickets show when the policy follows the whole ticket. The account also tells interpretability where to look. An instruction is applied at the data read after it rather than at its own tokens, and moving a single sentence gives a controlled contrast between a fact read with and without the rule in view. Finally, Llama-3.3-70B shows that a direct answer can take a late rule's outcome from the rule, which raises the question of what data or objective produces this route and whether other models can acquire it.

Limitations. Our study has several limitations. It concerns direct answers, and with reasoning before the answer the step disappears (Section 5), so the account describes how a model answers without deliberation. The panel covers open models up to 70B parameters, and the mechanistic tests cover the models with the step at a few layers each. The transplants also leave one question open. Because they replace the whole state of a token span, they show where the answer takes what it needs from a fact, but not what form it takes there. The state may hold the action itself, or the comparison the rule asks for, marked as falling under the rule. Donors whose rule exchanges the two actions could separate these, but their effect differs between models (Appendix H). Finally, we wrote the support tickets ourselves. They are realistic in form but not drawn from deployed traffic.

For systems that answer directly, the order of a prompt is part of its meaning, and the rule belongs before the data it governs.

Reproducibility statement

All prompts are generated deterministically from the factors in Section 2, and models answer with their native chat templates, with reasoning disabled except in the explanation conditions. One script regenerates every appendix table from the per-prompt outputs and recomputes each number quoted in the text, and the prompts, outputs and code are retained with checksums.

AI use statement

We used large language model assistants to help search the literature, write code, polish the language and translate. They also served as discussion partners when we refined hypotheses, analyzed data and interpreted results. The authors directed the research and made all final decisions. Code was tested, every reported number was recomputed from the recorded outputs by scripts, and citations were checked against their sources. The authors take full responsibility for the content.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024. URL https://arxiv.org/abs/2412.08905.

Gerry T. M. Altmann and Yuki Kamide. Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, 73(3):247–264, 1999. doi: 10.1016/S0010-0277(99)00059-1.

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.68. URL https://aclanthology.org/2023.findings-emnlp.68/.

Eden Biran, Daniela Gottesman, Sohee Yang, Mor Geva, and Amir Globerson. Hopping too late: Exploring the limitations of large language models on multi-hop queries. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 14113–14130. Association for Computational Linguistics, November 2024. doi: 10.18653/v1/2024.emnlp-main.781. URL https://aclanthology.org/2024.emnlp-main.781/.

John D. Bransford and Marcia K. Johnson. Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior*, 11 (6):717–726, 1972. doi: 10.1016/S0022-5371(72)80006-9.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ETKGuby0hcs.

Xinyun Chen, Ryan Andrew Chi, Xuezhi Wang, and Denny Zhou. Premise order matters in reasoning with large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 6596–6620, 2024. URL https://proceedings.mlr.press/v235/chen24i.html.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Jiahai Feng and Jacob Steinhardt. How do language models bind entities in context? In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/9d1b7fc578c0d2d6431fc26d736ecaf3-Abstract-Conference.html.

Lawrence T. Frase. Learning from prose material: Length of passage, knowledge of results, and position of questions. *Journal of Educational Psychology*, 58(5):266–272, 1967. doi: 10.1037/h0025028.

Gemma Team. Gemma 3. 2025. URL https://goo.gle/Gemma3Report.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12216–12235, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.751. URL https://aclanthology.org/2023.emnlp-main.751/.

GLM-5-Team, Aohan Zeng, Xin Lv, Zhenyu Hou, et al. GLM-5: from vibe coding to agentic engineering, 2026. URL https://arxiv.org/abs/2602.15763.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Juyeon Heo, Christina Heinze-Deml, Oussama Elachqar, Kwan Ho Ryan Chan, Shirley Ren, Andrew Miller, Udhyakumar Nallasamy, and Jaya Narain. Do LLMs "know" internally when they follow instructions? In *The Thirteenth International Conference on Learning Representations*, pp. 81339–81357, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/ca6980a3dba7fb3e4e66925656dba68b-Abstract-Conference.html.

Bernhard Hommel. The prepared reflex: Automaticity and control in stimulus-response translation. In Stephen Monsell and Jon Driver (eds.), *Control of Cognitive Processes: Attention and Performance XVIII*, pp. 247–274. The MIT Press, 2000. doi: 10.7551/mitpress/1481.003.0019.

Guan Zhe Hong, Nishanth Dikkala, Enming Luo, Cyrus Rashtchian, Xin Wang, and Rina Panigrahy. A implies B: Circuit analysis in LLMs for propositional logical reasoning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 30749–30803. Curran Associates, Inc., 2025. doi: 10.52202/085713-1033. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/2c4477572f06102b2ae3bd2b2ffcebf9-Paper-Conference.pdf.

IBM Research. Granite 4.0 language models. https://github.com/ibm-granite/granite-4.0-language-models, 2025. Accessed: 2025-10-01.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, et al. Llama guard: LLM-based input-output safeguard for human-AI conversations. *arXiv preprint arXiv:2312.06674*, 2023. URL https://arxiv.org/abs/2312.06674.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know, 2022. URL https://arxiv.org/abs/2207.05221.

Najoung Kim and Sebastian Schuster. Entity tracking in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3835–3855, 2023. URL https://aclanthology.org/2023.acl-long.213/.

Walter Kintsch. The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, 95(2):163–182, 1988. doi: 10.1037/0033-295X.95.2.163.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems 35*, 2022. URL https://arxiv.org/abs/2205.11916.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023. URL https://arxiv.org/abs/2307.13702.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. Prompt repetition improves non-reasoning LLMs, 2025. URL https://arxiv.org/abs/2512.14982.

Bojie Li. Models take notes at prefill: KV cache can be editable and composable, 2026. URL https://arxiv.org/abs/2606.17107.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 41451–41530. Curran Associates, Inc., 2023. doi: 10.52202/075280-1797. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/81b8390039b7302c909cb769f8b6cd93-Abstract-Conference.html.

Yubo Li, Ramayya Krishnan, and Rema Padman. LLMs know the constraint but do not use it: Activation bottlenecks in pragmatic constraint reasoning, 2026. URL https://arxiv.org/abs/2608.12321.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024a. URL https://aclanthology.org/2024.tacl-1.9/.

Yijin Liu, Xianfeng Zeng, Chenze Shao, Fandong Meng, and Jie Zhou. Instruction position matters in sequence generation with large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 11652–11663. Association for Computational Linguistics, August 2024b. doi: 10.18653/v1/2024.findings-acl.693. URL https://aclanthology.org/2024.findings-acl.693/.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8086–8098, 2022. URL https://aclanthology.org/2022.acl-long.556/.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372. Curran Associates, Inc., 2022. doi: 10.52202/068431-1262. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Mistral AI Team. Mistral Small 3, January 2025. URL https://mistral.ai/news/mistral-small-3.

Stephen Monsell. Task switching. *Trends in Cognitive Sciences*, 7(3):134–140, 2003. doi: 10.1016/S1364-6613(03)00028-7.

Norman Mu, Sarah Chen, Zifan Wang, Sizhe Chen, David Karamardian, Lulwa Aljeraisy, Basel Alomair, Dan Hendrycks, and David Wagner. Can LLMs follow simple rules? *arXiv preprint arXiv:2311.04235*, 2023. URL https://arxiv.org/abs/2311.04235.

Hyunjong Ok and Jaeho Lee. Lost in the prompt order: Revealing the limitations of causal attention in language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 38566–38587. Association for Computational Linguistics, July 2026. doi: 10.18653/v1/2026.findings-acl.1921. URL https://aclanthology.org/2026.findings-acl.1921/.

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

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. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In *The Thirteenth International Conference on Learning Representations*, pp. 66880–66913, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/a712d461e57201efe35d429a6f1731c1-Abstract-Conference.html.

Nikhil Prakash, Natalie Shapira, Arnab Sen Sharma, Christoph Riedl, Yonatan Belinkov, Tamar Shaham, David Bau, and Atticus Geiger. Language models use lookbacks to track beliefs. In *The Fourteenth International Conference on Learning Representations*, pp. 7657–7694, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/0d1941f73833eb107939f62d323d6cc0-Abstract-Conference.html.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Manas Venkata Sai Ravulapalli, Samrath Singh Chadha, and Abhinav M. Hari. Legible failures: Detecting and repairing in-context binding errors, 2026. URL https://arxiv.org/abs/2609.11216.

Ernst Z. Rothkopf. Learning from written instructive materials: An exploration of the control of inspection behavior by test-like events. *American Educational Research Journal*, 3(4):241–249, 1966. doi: 10.3102/00028312003004241.

Arnab Sen Sharma, Giordano Rogers, Natalie Shapira, and David Bau. LLMs process lists with general filter heads. In *International Conference on Learning Representations*, pp. 63227–63263, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/6732332e3c4155da82cb9d4970e88133-Paper-Conference.pdf.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. Judging the judges: A systematic study of position bias in LLM-as-a-judge. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 292–314, Mumbai, India, December 2025. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics. doi: 10.18653/v1/2025.ijcnlp-long.18. URL https://aclanthology.org/2025.ijcnlp-long.18/.

Suzanna Sia, David Mueller, and Kevin Duh. Where does in-context learning happen in large language models? In *Advances in Neural Information Processing Systems*, volume 37, pp. 32761–32786. Curran Associates, Inc., 2024. doi: 10.52202/079017-1030. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/3979818cdc7bc8dbeec87170c11ee340-Abstract-Conference.html.

Jacob Springer, Suhas Kotha, Daniel Fried, Graham Neubig, and Aditi Raghunathan. Repetition improves language model embeddings. In *The Thirteenth International Conference on Learning Representations*, pp. 93543–93579, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/e9b0c9480b03f9a6bca24ed8c08ee058-Abstract-Conference.html.

Alessandro Stolfo, Vidhisha Balachandran, Safoora Yousefi, Eric Horvitz, and Besmira Nushi. Improving instruction-following in language models through activation steering. In *The Thirteenth International Conference on Learning Representations*, pp. 55790–55823, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/8c3262a4c965ba9888f120d4f9e13478-Abstract-Conference.html.

Luxshan Thavarasa and Sivasuthan Sukumar. Test, then route: How language models execute in-context conditional rules across models and languages, 2026. URL https://arxiv.org/abs/2608.04183.

Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron Wallace, and David Bau. Function vectors in large language models. In *The Twelfth International Conference on Learning Representations*, pp. 17282–17333, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/4ae163cb8788970e53b4fd9578141139-Abstract-Conference.html.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems 36*, 2023. URL https://arxiv.org/abs/2305.04388.

Andreas Waldis, Leshem Choshen, Yufang Hou, and Yotam Perlitz. Instructions shape production of language, not processing, 2026. URL https://arxiv.org/abs/2605.11206.

Eric Wallace, Kai Xiao, Reimar Leike, Lilian Weng, Johannes Heidecke, and Alex Beutel. The instruction hierarchy: Training LLMs to prioritize privileged instructions, 2024. URL https://arxiv.org/abs/2404.13208.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9440–9450, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.511. URL https://aclanthology.org/2024.acl-long.511/.

Ziqi Wang, Hanlin Zhang, Xiner Li, Kuan-Hao Huang, Chi Han, Shuiwang Ji, Sham M. Kakade, Hao Peng, and Heng Ji. Eliminating position bias of language models: A mechanistic approach. In *The Thirteenth International Conference on Learning Representations*, pp. 91212–91239, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/e389b15166cf98966ba058965a8c17e3-Abstract-Conference.html.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35*, 2022. URL https://arxiv.org/abs/2201.11903.

Lingfeng Wu and Behzad Shomali. Written as a record, read as an address: What a forward pass leaves in an operation's KV cache, 2026. URL https://arxiv.org/abs/2609.24635.

Xiaohan Xu, Chongyang Tao, Tao Shen, Can Xu, Hongbo Xu, Guodong Long, Jian-Guang Lou, and Shuai Ma. Re-reading improves reasoning in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15549–15575, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.871. URL https://aclanthology.org/2024.emnlp-main.871/.

Yuzheng Xu, Tosho Hirasawa, Tadashi Kozuno, and Yoshitaka Ushiku. Am I more pointwise or pairwise? revealing position bias in rubric-based LLM-as-a-judge, 2026. URL https://arxiv.org/abs/2602.02219.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/06a52a54c8ee03cd86771136bc91eb1f-Abstract-Conference.html.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12697–12706, 2021. URL https://proceedings.mlr.press/v139/zhao21c.html.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*, pp. 19426–19454, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/54dd9e0cff6d9214e20d97eb2a3bae49-Abstract-Conference.html.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623. Curran Associates, Inc., 2023. doi: 10.52202/075280-2020. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

Egor Zverev, Sahar Abdelnabi, Soroush Tabesh, Mario Fritz, and Christoph Lampert. Can LLMs separate instructions from data? and what do we even mean by that? In *The Thirteenth International Conference on Learning Representations*, pp. 67147–67179, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/a77eadda332b6d4a9ae1e0e4024555f2-Abstract-Conference.html.

A THE PLACEMENT GRID AND THE ANSWER TABLES

Figure 6 repeats Figure 2 for the second vocabulary. The step is 79 against 33 at equal distance, and every model shows it. Table 3 gives, per model and vocabulary, the influence of the named fact just before and just after the rule, the first fact's influence with the rule first and last, and the change in the influence of A, B and C when the rule moves from before A to between A and B.

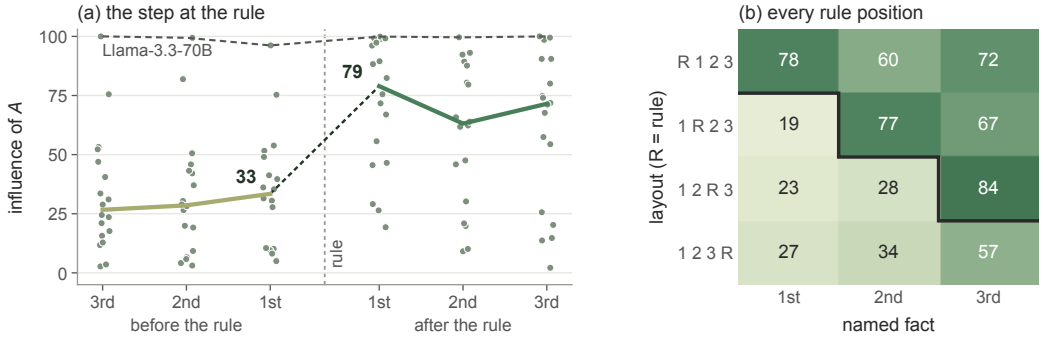


Figure 6: Figure 2 for the second vocabulary. (a) Influence of the named fact against its position relative to the rule, with points for the 18 models and median lines. (b) Median influence over the grid of rule and named positions.

Answer tables. Because every layout is run on all eight truth assignments, each structure gives a complete answer table, and we classify it by the facts it depends on. With the rule first, 63% of the tables equal exactly the named fact A. With the rule between A and B, 7% do, and 21% depend on B and C only, against 0.1% with the rule first. Among the tables that depend on B and C only, conjunctions are about twice as common as disjunctions; separately, tables that always answer BLOCK are about three times as common as tables that always answer PERMIT (Table 6).

Unnamed facts. Part of A's loss goes to the fact nearest the answer, whatever the rule's position. The rule's position still matters for the unnamed facts: B gains 19 when it is read just after the rule and 9 when it is read just before it (14 of 18 models). For C, the fact nearest the end, the corresponding gains are 24 and 21.

Ties. A structure enters the influence of a layout only if none of its eight answers is an exact tie between the two letter logits; 5.8% of structures are excluded this way. Models run in bfloat16.

B OUTPUT AUDITS

The main results read the next-token logit difference between the two instructed letters. Two audits check that this readout reflects the model's answer.

Table 3: The placement law per model. Entries are the influence of a fact on the answer (100 means that the fact decides the answer as the rule says, 0 means no net effect). Under *named fact*, the named fact is read directly after or directly before the rule at the same position among the facts, averaged over the three such pairs. Under *first fact*, the first fact is read with the rule after all three facts or before them. Under *rule past* A, the columns give the change in the influence of A, B and C when the rule moves from before A to between A and B, with A named. The models above the first rule show the step, Llama-3.3-70B is the exception, and the five models below the second rule cannot do the task even with the rule first. Medians are over the 18 models, and the last row counts the models in the expected direction, with after above before, last below first, A down, and B and C up.

	First vocabulary							Second vocabulary						
	Named fact		First fact		Rule past A			Named fact		First fact		Rule past A		
Model	after	before	last	first	ΔA	ΔB	ΔC	after	before	last	first	ΔA	ΔB	ΔC
Qwen3-8B	91	32	24	90	−60	+28	+30	88	31	18	87	−58	+27	+32
Qwen3-14B	99	57	33	98	−57	+40	+31	99	54	31	98	−60	+44	+30
Qwen3-32B	99	32	41	100	−93	+62	+30	99	31	41	100	−96	+65	+30
Qwen3-30B-A3B	99	36	27	99	−78	+35	+45	99	40	34	100	−75	+36	+44
Gemma-3-12B	79	11	19	63	−66	+37	+10	82	10	21	70	−74	+41	+9
Gemma-3-27B	96	37	26	96	−82	+42	+37	96	36	24	95	−82	+44	+35
Mistral-Small-24B	69	25	28	86	−66	+12	+6	72	28	29	89	−66	+13	+8
Phi-4	98	75	75	94	−41	+19	+24	97	75	76	95	−42	+18	+26
Llama-3.1-8B	59	7	3	61	−50	+2	+4	56	10	4	61	−49	+3	+5
R1-Distill-Llama-70B	73	44	47	74	−51	+30	+21	76	41	47	78	−59	+32	+23
GLM-5.3-Flash	63	43	48	72	−40	+17	+6	67	49	52	77	−41	+16	+7
Granite-4-H	90	51	55	82	−39	+10	+10	89	52	53	78	−33	+9	+7
Llama-3.3-70B	100	97	100	100	−7	+7	+6	100	96	100	100	−11	+10	+8
Qwen3-4B	47	9	12	38	−24	+9	+19	47	9	12	36	−24	+9	+18
Gemma-3-4B	19	8	15	26	−12	+11	+7	19	10	16	26	−13	+10	+5
Llama-3.2-3B	24	4	3	47	−42	+20	+6	26	5	3	46	−36	+22	+10
OLMo-2-13B	30	9	13	25	−25	+11	+4	29	8	13	20	−19	+10	+8
OLMo-2-32B	48	39	23	43	−21	+10	+16	46	35	24	38	−18	+11	+9
Median	76	34	26	78	−46	+18	+13	79	33	27	78	−45	+17	+10
Models in direction	18		17		18	18	18	18		17		18	18	18

Table 4: The full placement grid. Entries are the median influence of the named fact over the 18 models, for every position of the rule and of the named fact. Italic cells have the named fact before the rule. In both vocabularies, every italic cell is below every upright cell.

	First vocabulary			Second vocabulary		
	named fact			named fact		
Rule	1st	2nd	3rd	1st	2nd	3rd
Before all facts	78	58	69	78	60	72
After the first fact	*21*	78	65	*19*	77	67
After the second fact	*23*	*29*	82	*23*	*28*	84
After all facts	*26*	*35*	*55*	*27*	*34*	*57*

Table 5: Unnamed facts read just after vs. just before the rule, with A named. Entries are the influence of B with the rule between A and B (*after*) or between B and C (*before*), and of C with the rule between B and C (*after*) or after all facts (*before*). The last row counts the models, out of 18, in which after exceeds before.

	First vocabulary				Second vocabulary			
	B		C		B		C	
Model	after	before	after	before	after	before	after	before
Qwen3-8B	31	13	49	41	31	13	56	44
Qwen3-14B	38	29	48	45	42	30	47	47
Qwen3-32B	62	13	62	37	65	10	66	40
Qwen3-30B-A3B	35	11	59	41	36	12	54	40
Gemma-3-12B	33	−2	39	10	36	−3	42	8
Gemma-3-27B	39	23	53	39	40	23	53	38
Mistral-Small-24B	10	8	12	12	13	9	8	11
Phi-4	19	15	8	25	18	16	7	24
Llama-3.1-8B	3	6	7	1	5	5	15	5
R1-Distill-Llama-70B	31	6	34	31	31	6	34	34
GLM-5.3-Flash	16	21	20	22	15	19	21	22
Granite-4-H	19	21	36	20	21	23	38	23
Llama-3.3-70B	6	2	2	0	10	1	1	0
Qwen3-4B	2	−3	11	−4	3	−4	15	0
Gemma-3-4B	7	2	6	7	7	0	8	7
Llama-3.2-3B	22	4	26	1	24	4	25	2
OLMo-2-13B	12	1	20	11	11	2	22	13
OLMo-2-32B	10	16	21	24	11	13	19	21
Median	19	9	24	21	20	10	23	22
Models in direction	14		14		15		14	

Table 6: Complete answer tables with the named fact A first, pooled over the 18 models, with the rule first or between A and B. Entries are the share of structures (%) whose eight answers equal each function of the truth values of the three facts. A fact is true when it satisfies the rule's PERMIT condition.

	First vocabulary		Second vocabulary	
Eight-row answer table	rule first	after A	rule first	after A
Exactly A	62.9	7.4	62.0	6.8
A and one other fact (AND)	3.2	4.8	4.5	4.9
A and one other fact (OR)	0.7	2.7	1.0	2.6
$A \wedge B \wedge C$	1.2	10.3	1.4	10.0
$A \vee B \vee C$	0.8	6.0	0.8	5.9
Majority of A, B, C	0.0	3.1	0.1	3.0
Not A	0.9	0.0	0.7	0.0
$B \wedge C$	0.0	10.2	0.0	10.9
$B \vee C$	0.0	5.2	0.0	5.5
B or C alone, or its negation	0.1	5.5	0.2	5.2
Always BLOCK	10.5	12.2	10.0	13.3
Always PERMIT	1.7	4.1	1.9	3.7
Other	17.9	28.5	17.3	28.2
Any function of B and C only	0.1	20.8	0.2	21.6
Structures (tie-free)	3,327	3,209	3,337	3,226

Unrestricted greedy decoding. We decoded the prompts that contrast the rule first with the rule between A and B (3,072 prompts per model and vocabulary, 110,592 in all) greedily to the model's end-of-sequence token or 32 new tokens, without restricting the vocabulary. A response counts as valid only if it is exactly one instructed letter followed by end of sequence. For Qwen3-32B, Gemma-3-27B and Llama-3.3-70B, all 18,432 responses are valid, and every answer table equals the one read from the letter logits. Across the panel, the named fact loses influence and the later facts gain in 16 of the 17 models that keep complete tables, in each vocabulary (Phi-4 appends explanations and keeps none). The exception, Gemma-3-4B, does not use the named fact even with the rule first on these prompts (influence 2), so it has nothing to lose; all 12 models that can do the task and keep complete tables show the direction. Invalid responses mostly add an explanation after the letter or state the action in words.

Action words. Answering with the words PERMIT and BLOCK instead of letter codes keeps the direction of the step in all 16 model and vocabulary cells tested.

Table 7: Unrestricted greedy decoding. Each model and vocabulary has 3,072 prompts (192 structures $\times$ 8 truth assignments $\times$ rule first or between A and B), decoded over the full vocabulary to the end-of-sequence token or 32 new tokens. Column *letter* counts responses that are exactly one instructed letter followed by the end of sequence. Column *pairs* counts the structures, out of 192, whose 16 responses are all such letters. The Δ columns give the change in the influence of A, B and C from rule first to rule between A and B on those pairs, and – marks a model without complete pairs (Phi-4 continues with text after the letter in 1,448 and 1,507 prompts). The last row counts the models with complete pairs in which A loses and B and C gain.

	First vocabulary					Second vocabulary				
Model	letter	pairs	ΔA	ΔB	ΔC	letter	pairs	ΔA	ΔB	ΔC
Qwen3-8B	3,072	192	−59	+27	+29	3,072	192	−57	+27	+31
Qwen3-14B	3,072	192	−57	+40	+31	3,072	192	−60	+43	+29
Qwen3-32B	3,072	192	−93	+63	+30	3,072	192	−95	+65	+29
Qwen3-30B-A3B	3,071	191	−77	+36	+44	3,071	191	−75	+36	+43
Gemma-3-12B	3,072	192	−66	+37	+9	3,072	192	−74	+40	+10
Gemma-3-27B	3,072	192	−82	+43	+37	3,072	192	−82	+44	+36
Mistral-Small-24B	2,303	53	−63	+15	+7	2,702	93	−60	+11	+6
Phi-4	1,581	0	–	–	–	1,511	0	–	–	–
Llama-3.1-8B	2,230	59	−31	+1	+1	2,321	85	−39	+3	+10
R1-Distill-Llama-70B	2,175	19	−62	+28	+17	2,249	29	−59	+22	+14
GLM-5.3-Flash	3,069	189	−39	+17	+4	3,071	191	−41	+15	+7
Granite-4-H	3,072	192	−36	+10	+9	3,072	192	−33	+7	+7
Llama-3.3-70B	3,072	192	−7	+7	+6	3,072	192	−11	+10	+8
Qwen3-4B	2,986	163	−37	+15	+24	2,983	164	−34	+14	+21
Gemma-3-4B	2,852	141	+11	−2	−1	2,903	147	+6	0	−1
Llama-3.2-3B	1,794	17	−10	+4	+4	2,034	32	−5	+1	+1
OLMo-2-13B	3,072	192	−22	+14	+7	3,072	192	−18	+11	+9
OLMo-2-32B	3,072	192	−20	+12	+15	3,072	192	−17	+11	+9
A down, B and C up				16/17					16/17	

C THE MODEL KNOWS THE FACT

On each context we ask three separate questions: which attribute the rule compares, whether the request and assigned values of that attribute match, and the action. The first two questions use the same context as the action question and end with their own answer instruction. Table 9 gives the accuracies per model with the rule first and with the rule between A and B, and the share of wrong actions for which both other questions are answered correctly.

Answer-first explanations. When the instruction asks for the letter first and an explanation afterwards, the explanations of wrong letters that state A's comparison state it correctly in 66% to 100% of cases, depending on the model (eleven models). In the example of Figure 1 (Qwen3-14B), the

Table 8: Action words instead of letter codes. The output mapping is replaced by "Answer with one word: PERMIT or BLOCK.", and the answer is the sign of the next-token margin between the first tokens of the two words. Entries are the influence of A with the rule first and between A and B, and the change in the influence of B and C, on structures without ties in either layout. For every model, the top token over the full vocabulary is one of the two words in at least 99.4% of prompts.

	First vocabulary				Second vocabulary			
	A				A			
Model	first	after A	ΔB	ΔC	first	after A	ΔB	ΔC
Qwen3-8B	84	29	+31	+38	77	33	+24	+35
Qwen3-14B	89	19	+54	+34	89	16	+59	+31
Qwen3-32B	99	13	+60	+26	100	9	+62	+27
Gemma-3-12B	72	−4	+23	+8	71	−4	+21	+6
Gemma-3-27B	89	19	+38	+25	88	20	+38	+24
Mistral-Small-24B	100	69	+11	+15	100	74	+11	+11
Phi-4	91	48	+19	+27	92	50	+20	+27
Llama-3.3-70B	100	87	+6	+3	100	91	+7	+3
Median	90	24	+27	+25	91	26	+22	+26

model answers "A" (PERMIT) and continues: "The request desk (east) does not match the assigned desk (west), so the action is BLOCK. However, the request room and slot match the assigned room and slot, but the operational policy only considers the desk match." The explanation reads the desk after the rule and derives the right action; the letter, produced before it, does not follow it. Asked to explain first on the same prompt, the model writes "The request desk (east) does not match the assigned desk (west), so the action is BLOCK" and answers correctly.

D CONSEQUENCES

Explaining before answering. We ask for one or two sentences of reasoning before the letter, or for the letter first and the explanation after, and score the letter. With the rule between A and B, the named fact's influence is 97 when the explanation comes first and 28 when the letter comes first (eleven models with the step, all in this direction); with the rule last, run on the first seven of them, 100 against 35 (7 of 7). Explaining first gives the model a reading of A after the rule, in its own words, before it commits to the letter.

Corrections. The rule names one attribute, and a later sentence replaces it with another ("replace *slot* with *room*"). When the correction is read after the facts, the corrected attribute's influence is 5, and the replaced attribute keeps control (65). When the same correction is read before the facts, the corrected attribute's influence is 25 (medians over the twelve models with the step, and before exceeds after in all twelve). A rule that names the attribute from the start gives 68.

E SUPPORT TICKETS

The tickets use three scenarios: a hotel room change, a parcel return and a travel expense claim. Each ticket names a customer, places unrelated notes between the facts, and writes the facts as form fields (`Ordered size: medium. Delivered size: large.`). The rule maps to ESCALATE or CLOSE, and all eight truth assignments are enumerated as in the main task. Against the rule at the start of the message (median influence 61), a rule inserted right after the named field costs 28 points (17 of 18 models) and a rule after the whole ticket costs 14 (14 of 18). A rule in the system prompt is on par with one at the start of the message (median difference 0, higher in 9 of 18).

Example. One parcel-return ticket with the policy inserted right after the named field, as generated for the experiments. The system prompt reads "You are a customer-service assistant who reviews return requests." The user message with form fields is

Table 9: The fact is known. Entries are the accuracy (%) of three separate questions asked on the same context with the rule first or between A and B. The questions ask which attribute the rule compares (*Attribute*), whether A's requested and assigned values match (*Match*), and what the action is. The last column gives, with the rule between A and B, the share of prompts with a wrong action on which both other questions are answered correctly.

	Attribute		Match		Action		Both right
Model	first	after A	first	after A	first	after A	(wrong actions)
First vocabulary							
Qwen3-8B	96.3	92.9	99.9	78.7	94.9	65.2	61.2
Qwen3-14B	92.9	96.0	99.9	95.0	98.8	70.1	86.4
Qwen3-32B	96.4	97.3	100.0	82.9	100.0	53.3	77.9
Qwen3-30B-A3B	100.0	85.5	94.7	86.5	99.6	60.8	68.4
Gemma-3-12B	100.0	98.6	100.0	81.4	81.6	48.4	74.0
Gemma-3-27B	100.0	100.0	100.0	97.3	97.8	57.0	93.9
Mistral-Small-24B	97.5	94.0	92.6	87.9	92.9	60.5	92.8
Phi-4	71.8	86.2	85.1	85.8	96.9	76.4	50.3
Llama-3.1-8B	44.7	47.9	96.9	77.1	78.4	53.2	32.9
R1-Distill-Llama-70B	88.9	87.8	99.9	93.2	86.5	60.7	74.4
GLM-5.3-Flash	100.0	99.5	97.5	94.7	84.2	65.6	99.2
Granite-4-H	99.0	95.2	89.9	87.8	89.8	72.1	77.4
Llama-3.3-70B	100.0	100.0	100.0	100.0	100.0	96.0	100.0
Qwen3-4B	88.5	97.3	99.9	73.4	68.6	56.6	71.3
Gemma-3-4B	67.3	59.8	100.0	65.5	61.9	56.6	37.5
Llama-3.2-3B	56.8	52.1	64.8	51.6	73.6	52.5	27.2
OLMo-2-13B	52.2	52.9	84.0	60.4	60.6	50.1	28.3
OLMo-2-32B	81.8	83.7	94.5	50.8	71.7	62.0	38.8
Median, 12 with the step	96.9	94.6	98.7	87.1	93.9	60.7	75.9
Median, all 18	94.6	93.5	98.7	84.3	88.2	60.6	72.7
Second vocabulary							
Qwen3-8B	96.7	94.7	100.0	78.2	93.3	64.5	60.7
Qwen3-14B	93.6	98.6	99.9	94.9	98.8	68.6	90.0
Qwen3-32B	96.5	97.9	99.9	82.0	100.0	52.3	76.4
Qwen3-30B-A3B	100.0	87.0	95.1	85.7	100.0	62.2	71.0
Gemma-3-12B	100.0	98.2	100.0	81.3	84.5	47.8	72.6
Gemma-3-27B	99.9	100.0	100.0	97.4	97.3	56.6	94.2
Mistral-Small-24B	98.4	93.0	94.6	90.6	93.6	61.5	94.3
Phi-4	73.5	83.3	84.7	85.5	97.2	76.4	48.4
Llama-3.1-8B	44.2	49.5	97.4	76.7	77.9	53.5	35.4
R1-Distill-Llama-70B	87.6	87.0	100.0	92.4	88.5	59.2	72.9
GLM-5.3-Flash	99.9	99.2	98.1	95.4	87.2	67.5	98.1
Granite-4-H	97.9	92.9	90.4	88.0	88.7	72.3	74.2
Llama-3.3-70B	100.0	100.0	100.0	100.0	100.0	94.5	100.0
Qwen3-4B	85.2	94.3	99.9	75.7	67.6	55.5	72.2
Gemma-3-4B	66.5	59.8	100.0	63.5	62.4	56.4	35.7
Llama-3.2-3B	57.0	53.7	66.5	51.6	72.7	54.8	29.1
OLMo-2-13B	52.5	53.8	82.7	60.0	58.8	50.2	29.2
OLMo-2-32B	78.6	80.0	94.6	51.2	69.3	60.5	38.6
Median, 12 with the step	97.3	93.8	99.0	86.8	93.5	61.9	73.6
Median, all 18	95.0	92.9	99.0	83.8	88.6	59.9	72.4

Table 10: Consequences. The left half gives the influence of the named fact A with the rule between A and B or after all facts, first vocabulary. The instruction asks either for one or two sentences of reasoning before the letter (*expl.*) or for the letter first and the reasoning after it (*ans.*). Column A % gives, among answer-first responses with a wrong letter whose explanation says whether A's values match, the share that say it correctly. The right half concerns a rule that names attribute c and a later sentence that replaces it with b. Columns *bef.* and *aft.* give the influence of b when the correction is read before or after the facts, *repl.* gives the influence of the replaced c with the correction after the facts, and *named* gives the influence of b when the rule names b from the start. The correction before the facts did not complete on Llama-3.3-70B, and the later models were run with the rule between A and B only (–). Medians are over the models with the step that have each entry, and the last row counts the models in which explaining first exceeds answering first and before exceeds after.

	Explain vs. answer first						Correction of the rule							
	after A			last			First vocabulary				Second vocabulary			
Model	expl.	ans.	A %	expl.	ans.	A %	bef.	aft.	repl.	named	bef.	aft.	repl.	named
Qwen3-8B	97	28	88	100	25	84	14	5	63	51	13	7	60	54
Qwen3-14B	100	39	98	100	35	94	18	−1	86	77	15	−1	88	77
Qwen3-32B	96	13	100	100	44	82	47	−6	81	91	49	−5	82	91
Qwen3-30B-A3B	100	32	98	–	–	–	62	16	66	68	61	16	69	71
Gemma-3-12B	90	−4	75	99	26	83	24	9	51	55	26	8	58	59
Gemma-3-27B	95	12	77	100	31	68	43	11	88	87	41	11	89	90
Mistral-Small-24B	99	54	100	100	43	98	27	−2	78	59	29	−2	78	62
Phi-4	100	48	98	100	74	97	33	4	96	89	30	4	96	87
Llama-3.1-8B	86	4	66	–	–	–	8	4	16	26	5	5	19	28
R1-Distill-Llama-70B	100	22	73	–	–	–	15	0	57	68	16	0	56	67
GLM-5.3-Flash	–	–	–	–	–	–	20	10	14	37	22	9	15	44
Granite-4-H	87	48	80	–	–	–	41	20	40	81	42	17	43	80
Llama-3.3-70B	100	91	85	100	99	100	–	2	98	100	–	4	96	100
Median	97	28	88	100	35	84	25	5	65	68	27	6	65	69
Models in direction	11/11			7/7			12/12				11/12			

> Return request #4471 from Sam Okafor, opened through the web form. Ordered size: medium. Delivered size: large. Returns policy: If the delivered size differs from the ordered size, the action is ESCALATE; if the delivered size matches the ordered size, the action is CLOSE. The customer wrote that the courier left the box at the side door. Ordered color: navy. Delivered color: navy. The original packaging was kept. Ordered model: Pro. Delivered model: Lite. The account has two earlier orders. Output mapping: Output A for ESCALATE and B for CLOSE. Output exactly one letter and nothing else.

The free-prose version keeps every other sentence and writes the three fields as "The customer ordered size medium, and the package contained size large", "The order was for the navy color; the item that arrived is navy" and "The order specified the Pro model, and the box holds the Lite model". The correct action is ESCALATE, since the sizes differ.

Free prose. When the same facts are paraphrased as free prose instead of form fields, the rule at the start of the message and in the system prompt is weaker (50 and 47, against 61 and 57 with fields). The inserted rule still costs the named fact 15 points (16 of 18 models), while a rule after the whole ticket costs nothing (median 0, lower in 9 of 18).

Table 11: Support tickets. Entries are the influence of the named field on the answer (ESCALATE vs. CLOSE), pooled over three scenarios, with the rule at the start of the user message, in the system prompt, right after the named field, or after the whole ticket. The left half writes the facts as form fields, and the right half paraphrases the same facts in free prose. The last rows give the median over models, the median change of each model from the start-of-message layout, and the number of models below it.

	Form fields				Free prose			
Model	start of message	system prompt	after field	after ticket	start of message	system prompt	after field	after ticket
Qwen3-8B	69	50	36	54	53	45	26	44
Qwen3-14B	83	76	42	57	43	37	34	53
Qwen3-32B	79	76	48	69	61	54	44	57
Qwen3-30B-A3B	74	46	42	45	52	39	41	52
Gemma-3-12B	38	44	28	60	34	34	25	47
Gemma-3-27B	61	68	45	66	58	63	40	57
Mistral-Small-24B	90	90	62	85	74	69	54	73
Phi-4	56	46	46	74	51	42	40	56
Llama-3.1-8B	50	61	15	38	50	64	20	30
R1-Distill-Llama-70B	83	86	48	82	74	76	54	76
GLM-5.3-Flash	62	83	40	46	45	62	30	34
Granite-4-H	57	52	59	51	47	50	52	55
Llama-3.3-70B	62	61	40	93	60	66	47	86
Qwen3-4B	51	48	18	31	32	30	18	33
Gemma-3-4B	46	53	11	21	44	44	25	24
Llama-3.2-3B	28	37	3	5	18	22	3	16
OLMo-2-13B	28	24	10	7	8	9	16	16
OLMo-2-32B	66	74	38	44	60	61	40	47
Median	61	57	40	52	50	47	37	50
Median change from start		0	−28	−14		+1	−15	0
Lower than start		9/18	17/18	14/18		8/18	16/18	9/18

F READING THE FACT AGAIN

Table 12 gives the re-reading test of Section 4.2 per model and vocabulary, with two further variants. Repeating the whole prompt, as in prompt repetition (Leviathan et al., 2025), raises the named fact's influence from 20 to 93 with the rule between A and B and from 28 to 100 with the rule last (seven models with the step; Table 13). Part of this gain appears even when the first copy lacks the rule (73 over the twelve models), so the repetition also helps in ways other than a second reading of A after

the rule. Repeating only A's sentence isolates the second reading: after the rule it restores A, and before the rule it does not.

Table 12: Re-reading only the fact. Influence of the named fact A with the rule first, with the rule between A and B, and, in the latter layout, after inserting a copy of A's sentence right before the rule, a copy of B's sentence right after it, or a copy of A's sentence right after it. The last two columns read the whole prompt (rule between A and B) twice, and twice with the rule removed from the first copy. Medians are over the twelve models with the step. Llama-3.3-70B, which has almost no step, is shown for reference.

	Rule		Rule after A, plus a copy of			Prompt twice	
Model	first	after A	A before rule	B after rule	A after rule	both copies	1st w/o rule
First vocabulary							
Qwen3-8B	90	30	28	23	92	95	68
Qwen3-14B	98	41	40	22	98	93	73
Qwen3-32B	100	7	6	9	96	90	82
Qwen3-30B-A3B	99	21	23	21	100	100	73
Gemma-3-12B	63	−3	8	2	58	45	35
Gemma-3-27B	96	14	11	15	62	90	74
Mistral-Small-24B	87	20	23	21	64	100	78
Phi-4	94	53	34	29	98	93	88
Llama-3.1-8B	61	11	9	8	52	16	5
R1-Distill-Llama-70B	74	23	23	17	74	95	58
GLM-5.3-Flash	72	32	32	39	80	94	97
Granite-4-H	82	45	36	43	79	86	61
Llama-3.3-70B	100	92	80	88	100	100	100
Median	88	22	23	21	80	93	73
Second vocabulary							
Qwen3-8B	87	29	31	25	87	92	65
Qwen3-14B	98	37	39	20	98	95	67
Qwen3-32B	100	4	4	8	97	90	82
Qwen3-30B-A3B	100	25	29	24	100	100	72
Gemma-3-12B	70	−4	11	3	57	43	29
Gemma-3-27B	95	13	10	13	64	89	72
Mistral-Small-24B	88	22	26	24	70	100	75
Phi-4	95	53	33	22	99	93	88
Llama-3.1-8B	61	12	7	9	46	25	4
R1-Distill-Llama-70B	78	19	21	15	69	93	59
GLM-5.3-Flash	77	36	33	44	85	95	99
Granite-4-H	78	45	35	46	73	86	54
Llama-3.3-70B	100	89	74	85	100	100	100
Median	87	23	28	21	79	93	69

G VERDICT WORDING

If the facts are given as verdicts instead of raw values, the rule has less to apply. Verdicts that repeat the rule's own words ("the request desk differs from the assigned desk"), written in place of the raw facts, are used from before the rule as well as after it (influence 28 with raw facts before the rule, 95 with such verdicts; higher in all seven models with the step). Verdicts in other words ("Desk values: not identical.") show the step like raw facts: 67 with the rule first, 36 with the rule after A, and 32 with the rule last (medians over the same seven models). A fact read before the rule is usable when its outcome can be matched to the rule's words without applying the rule.

Table 13: Reading the whole prompt twice. Influence of the named fact A when the prompt is read once or twice in a row, for three positions of the rule. In the column *twice, 1st w/o rule*, the first copy lacks the rule (from the re-reading run). Medians are over the models with the step, and Llama-3.3-70B is shown for reference.

	Rule first		Rule between A and B			Rule last	
Model	once	twice	once	twice	twice, 1st w/o rule	once	twice
First vocabulary							
Qwen3-8B	90	92	30	95	68	24	96
Qwen3-14B	98	99	41	93	73	33	99
Qwen3-32B	100	100	7	90	82	40	100
Gemma-3-12B	63	90	−3	45	35	19	81
Gemma-3-27B	96	98	14	90	74	26	100
Mistral-Small-24B	87	100	20	100	78	28	100
Phi-4	94	91	53	93	88	75	100
Llama-3.3-70B	100	100	92	100	100	100	100
Median	94	98	20	93	74	28	100
Second vocabulary							
Qwen3-8B	87	91	29	92	65	18	96
Qwen3-14B	98	99	37	95	67	31	98
Qwen3-32B	100	100	4	90	82	40	99
Gemma-3-12B	70	89	−4	43	29	21	76
Gemma-3-27B	95	98	13	89	72	24	100
Mistral-Small-24B	88	100	22	100	75	29	100
Phi-4	95	92	53	93	88	76	100
Llama-3.3-70B	100	100	89	100	100	100	100
Median	95	98	22	92	72	29	99

H TRANSPLANTS

Setup. The recipient prompt has the rule between A and B and a neutral sentence before A of the same token length as the rule, so that A's tokens occupy the same positions as in the rule-first prompt. A donor differs from the recipient only in the rule's place, in A's values, or in both. We replace the hidden states over A's tokens, or over the rule's tokens, at one layer and let the model run to the answer. Table 14 lists every model and layer for the donors used in the main text, and Table 15 the restoring transplants. Every run first writes each span's own states back and requires the output to be unchanged. Granite-4-H fails this check because its forward pass is not bit-for-bit reproducible, so it is not used for transplants.

Layers. The size of both transplants depends on the layer. In seven of the ten models with the step, neither transplant changes the answer at the deepest layer tested; in Qwen3-14B, Qwen3-32B and Qwen3-30B-A3B the transplant into A's tokens still acts there. The transplant into the rule's tokens changes A's influence by up to 55 (Qwen3-30B-A3B), always less than the transplant into A's tokens in the same model. After the transplant into A's tokens the answer follows the donor with influence 33 to 82 at the best layer; after the transplant into the rule's tokens, by at most 27. Qwen3-30B-A3B was run with the first vocabulary only.

Donors. These two donor types were run in Qwen3-8B, Qwen3-32B and Gemma-3-27B. Donors whose rule names another attribute restore A only in part (31 to 43). Donors whose rule exchanges the two actions, which would separate a stored action from a stored comparison, give model-dependent results (5, -8 and 63), so we do not use them to locate the outcome. In the rule-first prompt itself, swapping only A's states for those of the opposite value flips the answer in all 192 structures of Qwen3-32B at layers 32 and 40.

Table 14: Transplants into the fact's tokens vs. the rule's tokens. The recipient has the rule between A and B. Entries are the influence of the recipient's A without a transplant (*none*) and after replacing, at one layer, the hidden states over (i) A's tokens with those of the rule-first prompt whose A has the opposite value, so that this A was read after the rule, (ii) A's tokens with those of the same-layout prompt whose A has the opposite value, so that this A was read before the rule, or (iii) the rule's tokens with those of that same-layout prompt, whose rule has read the opposite-value A. Because the donor's A has the opposite value, the influence of the donor's A, that is, how much the answer follows the transplant, is minus each entry. Qwen3-30B-A3B was run with the first vocabulary only (–).

		First vocabulary				Second vocabulary			
Model	Layer	none	(i)	(ii)	(iii)	none	(i)	(ii)	(iii)
Qwen3-8B	18	12	−43	11	−7	14	−41	10	−11
	27		13	13	12		14	14	14
Qwen3-14B	12	46	−56	−46	46	47	−57	−47	47
	16		−74	−45	45		−75	−48	46
	20		−71	−22	18		−69	−24	22
	24		−22	44	39		−23	45	41
Qwen3-32B	32	3	−35	−11	4	5	−33	−9	5
	48		−53	3	1		−54	2	0
Qwen3-30B-A3B	12	27	−75	−32	28	–	–	–	–
	18		−82	−36	20		–	–	–
	24		−28	19	−27		–	–	–
	30		−60	20	−1		–	–	–
Gemma-3-12B	12	1	−25	−13	0	4	−23	−12	4
	18		−29	−5	−16		−30	−5	−14
	24		−39	−1	−3		−39	3	−1
	30		1	1	0		4	4	4
Gemma-3-27B	23	25	−44	−1	5	23	−36	1	2
	40		25	25	25		22	23	23
Mistral-Small-24B	12	22	−59	−11	10	26	−62	−12	13
	16		−2	18	−5		−2	21	−4
	20		14	22	19		15	25	23
	24		22	22	22		26	26	26
Phi-4	12	33	−71	−28	21	35	−76	−30	23
	16		8	25	−14		7	26	−13
	20		15	33	20		18	36	24
	24		33	34	34		36	36	35
Llama-3.1-8B	8	14	−41	−8	11	15	−38	−10	7
	12		−9	19	−15		−17	16	−17
	16		14	13	15		13	14	15
	20		14	13	17		15	15	18
R1-Distill-Llama-70B	20	15	−33	−14	4	13	−33	−13	4
	30		−13	10	−5		−15	8	−6
	40		16	15	16		13	13	14
Llama-3.3-70B	30	76	47	75	−14	72	41	69	−14
	50		75	75	75		69	70	69

Table 15: Restoring the fact's post-rule state. Influence of A with the rule first, with the rule between A and B, and in the latter prompt after replacing, at one layer, the hidden states over A's tokens with those of the same A read after the rule in the rule-first prompt (←*after rule*) or after a non-rule sentence with the same nouns (←*after nouns*). A neutral sentence of the rule's length before A keeps A's tokens at the same positions. Writing back the recipient's own states reproduces the unpatched value in every row.

		First vocabulary				Second vocabulary			
Model	Layer	rule first	rule after A	←after rule	←after nouns	rule first	rule after A	←after rule	←after nouns
Qwen3-8B	9	90	12	29	18	87	14	29	18
	14			38	19			40	19
	18			54	20			56	21
	22			23	12			22	13
	27			13	13			14	14
Qwen3-32B	16	100	3	12	4	100	5	12	6
	24			23	5			21	6
	32			31	5			30	6
	40			47	13			45	12
	48			71	8			70	8
Gemma-3-27B	16	96	25	25	24	95	23	23	21
	23			63	28			60	25
	31			46	26			45	23
	39			25	25			23	23
	46			25	25			23	23
Llama-3.3-70B	20	100	76	85	81	100	72	83	74
	30			92	78			90	72
	40			75	76			71	71

I LINEAR PROBES

We train linear probes on the hidden states of Qwen3-8B, Qwen3-32B and Gemma-3-27B. At A's own tokens, whether A's values match is read with 100% accuracy in both layouts, while the action implied by A is read with 100% accuracy with the rule first and at chance (50%) with a late rule (Qwen3-8B, layers 18 to 35). With the rule after A, A's outcome is also read at the rule's last token and at the answer position with 98% to 100% accuracy at later layers. At the answer position, however, probes read arbitrary combinations of the available bits (whether A matches, exclusive-or the letter code or the clause order) equally well, in all three models. Decodability there does not show where the answer takes the outcome from, which is why Section 4.2 relies on transplants.

J THE LLAMA-3.3-70B EXCEPTION

Llama-3.3-70B has a step of 2.5 points in the first vocabulary (99.9 with the rule just before the named fact, 97.4 with it just after) and answers correctly with the rule after all facts. At layer 30, the opposite-value transplant into the rule's tokens changes A's influence by -89 and -85 in the two vocabularies, and the transplant into A's tokens by -29 and -30. R1-Distill-Llama-70B is Llama-3.3-70B-Instruct fine-tuned on reasoning traces; answering directly, it shows the full step (73 against 44). The two models share architecture and pretrained weights, so the step is not fixed by either. The exception is not complete: in the support tickets, Llama-3.3-70B loses 22 points when the rule is inserted right after the named field (Table 11).

K TWO FACTS AND REASONING

A two-fact version of the task places a rule that names one of two facts between them. Where the two facts disagree, direct answers follow the unnamed fact in 35% to 100% of cases in five models (Qwen3-8B, Qwen3-32B, Gemma-3-12B, Gemma-3-27B, Llama-3.3-70B). Prompting for

step-by-step reasoning before the answer lowers this to at most 9% in four of them (Gemma-3-12B: 23%), and native thinking to 3% and 0% (Qwen3-8B and Qwen3-32B).