THE GEOMETRY OF CONTEXTUAL RELATIONS: LANGUAGE MODELS ADDRESS FACTS BY ORDER OF MENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Human reasoning depends on how objects are related within propositions. *How do relations organize the language representations of contextual contents?* We give an LLM a list of facts in its context (e.g., *Alice eats an apple. Bob eats a pear.*) and measure how its hidden state changes when the question switches from what Alice eats to what Bob eats. Averaged over many lists, this change is a steering vector, which we call the *ordinal vector*. It points to a fact by its *order of mention*, the order in which the facts were stated in the context. We find that LLMs represent the fact a question asks about by its order of mention, not by the name the question contains. We state this as the *ordinal addressing hypothesis*: each order of mention has a *fact address* in the model's state, shared by all contexts, and a question moves the state to the fact address of the fact it asks about, while the context supplies what that fact says. Across Qwen, Gemma, and Llama, fact addresses are (1) *ordered by mention*: query states are organized by the order of facts, not of names, even when one fact has multiple subjects; (2) *steerable*: added to a question about the first fact of a new list, the ordinal vector makes the model answer with the second fact of that list; (3) *low-rank*: they span a low-rank subspace in which the first-mentioned fact is the easiest to reach, surprisingly similar to human recall; and (4) *emergent*: they are shared in late-middle layers, hold from 1.5B to 32B parameters, and form early in pretraining. Language models reach a stated fact by where it was mentioned, deepening our understanding of LLM reasoning. Code is available here.

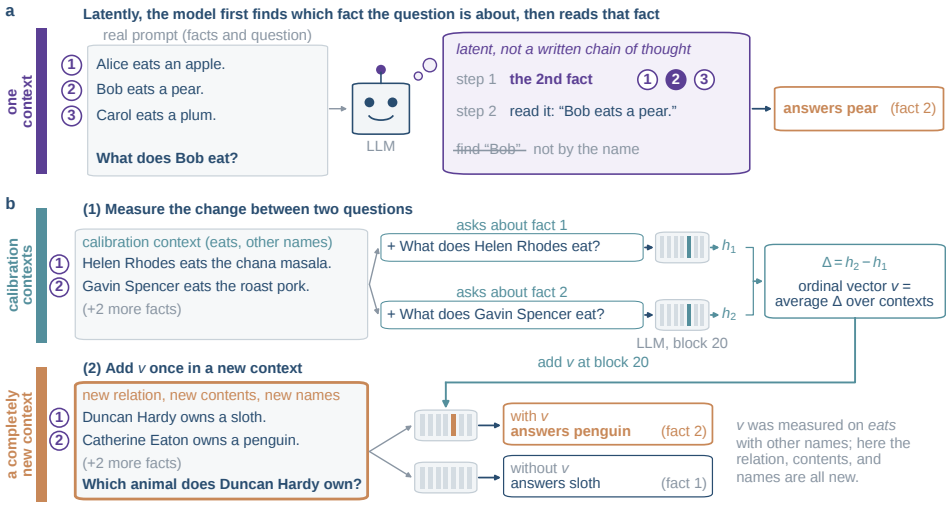

Figure 1: **Language models address stated facts by order of mention.** (a) Asked what Bob eats, the model's state after the question points to the second fact, not to the name, and the answer is then read from that fact. (b) A real Qwen2.5-7B run: (1) the block-20 query change Δ between questions about fact 1 and fact 2, averaged over calibration contexts, is the ordinal vector v; (2) added once in a new context, v moves the answer to fact 2's animal.

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.

1 INTRODUCTION

Human reasoning draws conclusions from relations among objects and events (Halford et al., 2010). The same inference works for any names, yet its conclusion depends on who plays which role: if Alice gives Bob a book, Bob has it; if Bob gives Alice a book, Alice has it. Combining facts therefore requires preserving each element's semantic role, who does what to whom, even as the contents change (Smolensky, 1990).

Large language models (LLMs) face this problem when reasoning from contextual facts (Feng & Steinhardt, 2024). In *Alice loves Bob. Carol loves David.*, the relation *loves* has two *semantic roles*, the person who loves and the person loved (Gildea & Jurafsky, 2002), and a *binding* specifies who fills each role: Bob and David share a role, but whom Alice loves depends on the binding. Semantic structure shapes language-model representations: word-vector analogies and concept geometry reveal directions shared across related meanings (Mikolov et al., 2013; Park et al., 2024; 2025b), and contexts trace recurring trajectories as logically structured information accumulates (Zhou et al., 2026). Relations add a particular challenge: the same relational form recurs across examples, while particular facts depend on which objects fill each role. This leads to our central question:

How do relations organize the language representations of contextual contents?

We study it where a question meets the facts. Given the facts *Alice eats the apple* and *Bob eats the pear*, a listener asked what Bob eats can find the fact by its description, the one about Bob, or by its order of mention, the second one, as *the former* and *the latter* do. People track the order in which facts are mentioned, and the first-mentioned participant is the easiest to retrieve (Gernsbacher & Hargreaves, 1988; Henson, 1998). Which does a language model use? We change the question from what Alice eats to what Bob eats and measure how the hidden state changes before the model answers. Averaged over contexts, this change is a steering vector, which we call the *ordinal vector* (Figure 1). We find that language models address stated facts by order of mention. We state this as a hypothesis (Section 3): each order of mention has a *fact address*, shared by all contexts; a question moves the state to the address of the fact it asks about, and the context supplies what is stated there, the fact's *binding identity*. Its consequences hold in Qwen, Gemma, and Llama:

(1) Questions are organized by order of mention. Query changes keep their direction on new names and reverse when the facts are reversed, in fourteen LLMs with base as well as instruction-tuned weights. Query states group by the queried fact's order of mention, not by the named person, and follow the order of facts, not of names, when one fact has two subjects (Sections 3 and 4).

(2) The fact address transfers across contexts. Added once to a context with new names, contents, and relation, the ordinal vector selects the designated fact among four, and that context supplies the content, also in short paragraphs and from 1.5B to 32B parameters (Section 5; Table 5). The vector that transfers follows order of mention, not names, relation, or token distance (Section 5).

(3) Fact addresses are low-rank. They span a low-rank subspace where ordinal vectors compose; later addresses crowd together and the first is easiest to reach, reminiscent of serial recall (Section 6).

(4) The fact address is shared in a depth window and learned early in pretraining. In late-middle blocks, ordinal vectors and states copied from another context select the receiving context's facts; later states carry the answer's content. In OLMo-2-7B checkpoints, questions are first organized as if by the named person and switch to order of mention by 42B tokens, before the model answers reliably, and the full depth profile forms late (Section 7).

A question thus says where to look, and the context says what is there, a division learned from next-token prediction alone and reminiscent of how people track discourse.

2 RELATED WORK

Mechanistic interpretability. Circuits, features, and causal interventions explain behavior through internal components and the variables they carry (Olah et al., 2020; Elhage et al., 2021; Wang et al., 2023; Bricken et al., 2023; Meng et al., 2022; Geiger et al., 2021), and activation addition and function vectors show that an activation estimated across examples can alter behavior on new inputs (Rimsky et al., 2024; Todd et al., 2024). Closest to our question, studies of contextual

binding find binding identifiers, propositional probes, and positional and lexical contributions to bound-entity retrieval (Feng & Steinhardt, 2024; Feng et al., 2025; Prakash et al., 2026; Gur-Arieh et al., 2026), and editing low-rank subspaces of bound-token states changes bindings (Dai et al., 2024; 2026). Prakash et al. (2026) find ordering IDs that a lookback mechanism uses to retrieve bound information, and Gur-Arieh et al. (2026) find that this positional mechanism grows noisy in the middle of long lists, where lexical and reflexive mechanisms take over. These works locate order in the states of the bound tokens within one context; we study the question's state across contexts, state its geometry as a hypothesis, and show that the address it carries transfers to new contexts and crowds in the middle of a list, a geometric account of that noise.

Representation analysis and geometry. Linear and categorical representation theories tie semantic contrasts and concept hierarchies to directions (Mikolov et al., 2013; Arora et al., 2018; Park et al., 2024; 2025b), language models transform concepts with shared geometry (Hu et al., 2026), and statistical accounts derive such geometry from co-occurrence and its symmetries (Jiang et al., 2024; Karkada et al., 2026; Modell et al., 2025; Huh et al., 2024). Geometry formed while reading structured input has been studied for in-context graphs, sentence trajectories, and logical reasoning (Park et al., 2025a; Hosseini & Fedorenko, 2023; Zhou et al., 2026), and steering can depend on how a context is presented (Opiełka et al., 2026); the geometry we study is set by the order of the stated facts rather than by word meaning.

3 Order of Mention and Fact Addresses

Given the facts *Alice eats the apple* and *Bob eats the pear*, the question *What does Bob eat?* is answered in two steps: find the fact the question is about, then read it. We formalize the first step.

Definition 3.1 (Stated facts, order of mention, and binding identity)**.** A context C states facts $f_1, \ldots, f_n$ in this order, where f_i binds one or more entities e_i and a content x_i under a relation r_i. The index i is the fact's *order of mention*, and the *binding identity* at order of mention i is $\beta_C(i) = (e_i, r_i, x_i)$, what C states there. A question q names part of one fact and asks for another part; $p_C(q)$ is that fact's order of mention, and the answer is read from $\beta_C(p_C(q))$. For a fixed relation, $q_C(p)$ is the question that names the entity of fact f_p and asks for its content.

A question names an entity, but the model could reach the fact by its order of mention instead, as ordinal reference and human serial memory do (Henson, 1998).

Contexts and states. A *word set* is one set of names and contents; a context states one fact per name in fixed sentence frames, and its *arrangement* orders the facts while its *content assignment* pairs names with contents. For a context C and question q, the *query state* $h_\ell(C, q) \in \mathbb{R}^d$ is the residual stream at the last prompt token after Transformer block ℓ.

Definition 3.2 (Query change and ordinal vector)**.** For orders of mention $p \neq p'$ and calibration contexts $\mathcal{C}_{\text{cal}}$ from other word sets, define

$$\Delta_\ell(C; q \to q') = h_\ell(C, q') - h_\ell(C, q), \qquad v_\ell^{p \to p'} = \mathbb{E}_{C \in \mathcal{C}_{\text{cal}}} \Delta_\ell\big(C; q_C(p) \to q_C(p')\big). \quad (1)$$

The mean v is a steering vector, the *ordinal vector*. In a *receiver* C_*, intervene once:

$$h_\ell(C_*, q_{C_*}(p)) \leftarrow h_\ell(C_*, q_{C_*}(p)) + v_\ell^{p \to p'}, \quad (2)$$

keep the question, and ask whether the complete answer is the *designated answer*, the content of the *designated fact* at p' in C_*. Replacing v by the receiver's own $\Delta_\ell(C_*; q_{C_*}(p) \to q_{C_*}(p'))$ gives the *ceiling*.

Hypothesis 3.3 (Ordinal addressing)**.** At an intermediate block ℓ there are context offsets c_C and *fact addresses* $a_1, a_2, \ldots$, one per order of mention and shared by all contexts, such that

$$h_\ell(C, q) = c_C + a_{p_C(q)} + \varepsilon_\ell(C, q) \quad (3)$$

with small ε_ℓ (measured by held-out R^2), and, beyond block ℓ (the *readout clause*), the answer is the content of $\beta_C(p)$ when the last-token state is $c_C + a_p$: a question moves the state to the address of the fact it asks about, and the context supplies the binding identity found there.

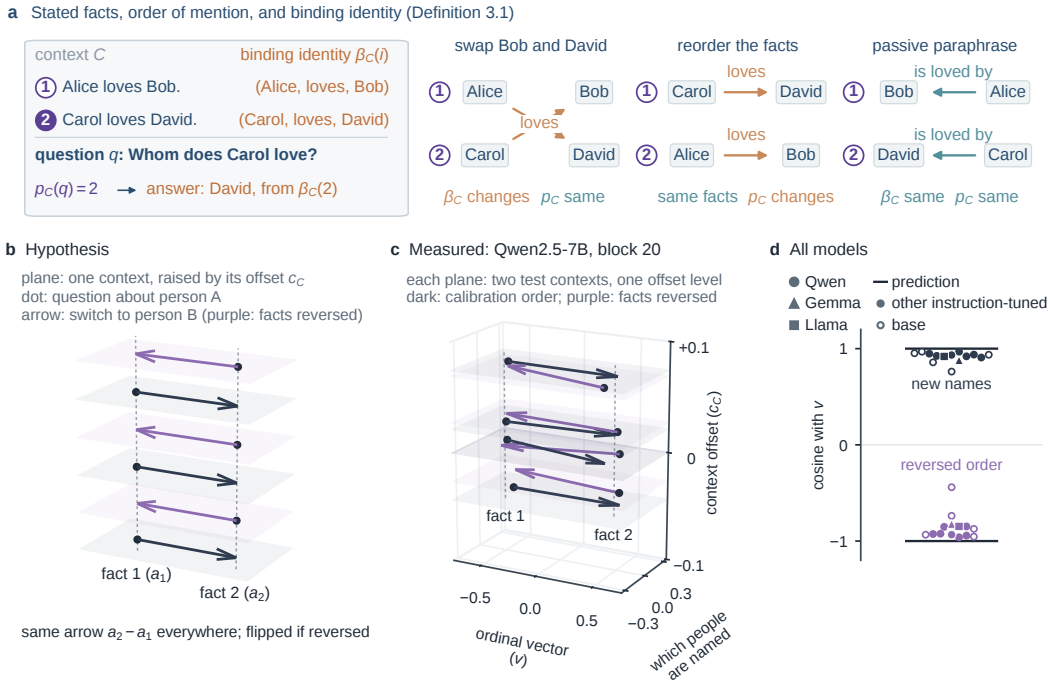

Figure 2: **The problem, the hypothesis, and its first test.** (a) Definition 3.1 on the introduction's example: each stated fact has an order of mention and a binding identity, and a question is answered from the binding identity at its order of mention. Swapping Bob and David changes the binding identities; reordering keeps the facts and changes p_C; the passive paraphrase changes neither. (b) Hypothesis 3.3 as a displacement field: every context shares the fact addresses a_1, a_2, so the question moves each context by the same $a_2 - a_1$, or by $-(a_2 - a_1)$ when the two facts are reversed. (c) The same form measured in Qwen for eight test contexts, two arrangements at each of four offset levels (percentiles of the context-offset coordinate), on three axes fit on calibration word sets: along v, which people are named, and the context offset; units of $\|v\|$, with the offset axis stretched for display (Appendix B). (d) Cosines with v for new names (dark; mean of 64 name pairs) and between the ordinal vectors of reversed arrangements (purple), one dot per model, against the predictions of Proposition 3.4; small dots: eleven further models (Appendix G; open: base weights).

Let $y(C, q) = h_\ell(C, q) - \frac{1}{n}\sum_p h_\ell(C, q_C(p))$ be the *query-dependent state*: the query state centered over the context's n questions, one per fact. With $\varepsilon = 0$, the hypothesis has four consequences (Propositions 3.4 to 3.6, Corollary 3.7), each tested below; Proposition 3.9 relaxes $\varepsilon = 0$.

Proposition 3.4 (Invariance and reversal)**.** *For any contexts C, C_* and orders of mention $p \neq p'$, $\Delta_\ell(C; q_C(p) \to q_C(p')) = \Delta_\ell(C_*; q_{C_*}(p) \to q_{C_*}(p')) = a_{p'} - a_p$, for any names and contents they state. If C'' exchanges the facts at p and p' of C, the same pair of questions gives $-(a_{p'} - a_p)$.*

Proposition 3.5 (Organization by order of mention)**.** *$y(C, q) = a_{p_C(q)} - \bar{a}$ with $\bar{a} = \frac{1}{n}\sum_p a_p$: the query-dependent state depends on q only through $p_C(q)$, not on the entities or relation it names or on token distances. If C' moves the fact at p in C to $\pi(p)$, then $y(C', q) = a_{\pi(p_C(q))} - \bar{a}$.*

Proposition 3.6 (Transfer)**.** *For any receiver C_*, Equation (2) gives $h_\ell(C_*, q_{C_*}(p)) + v_\ell^{p \to p'} = c_{C_*} + a_{p'}$, so by the readout clause the answer is the content of $\beta_{C_*}(p')$.*

Corollary 3.7 (Linear structure)**.** *Every change $\Delta_\ell(C; q \to q')$ lies in the address subspace $A_n = \mathrm{span}\{a_p - a_1 : p \leq n\}$, of dimension at most $n - 1$, shared by all contexts of n facts, and ordinal vectors compose: $v^{p \to p''} = v^{p \to p'} + v^{p' \to p''}$.*

Proof of Propositions 3.4 to 3.6 and Corollary 3.7. Substitute Equation (3): c_C cancels in every within-context difference, so every $\Delta_\ell(C; q_C(p) \to q_C(p'))$ and every mean $v^{p \to p'}$ equals $a_{p'} - a_p$;

exchanging or moving facts changes only which address each question reaches; the readout clause gives the answer; and differences of addresses lie in A_n and telescope. □

Remark 3.8. A relation component, shared by a context's questions, cancels in Δ_ℓ; coding the question by the named entity or the answer's content would contradict Propositions 3.4 to 3.6.

Proposition 3.9 (Spacing bounds transfer)**.** *Let instead $\varepsilon_\ell \sim \mathcal{N}(0, \sigma^2 I)$, let v be exact, $v = a_{p'} - a_p$, and let the readout return the fact whose address is nearest to $h_\ell - c_{C_*}$. The edit of Proposition 3.6 then selects $\beta_{C_*}(p')$ with probability between $1 - (n-1)\,\bar{\Phi}(d_{p'}/2\sigma)$ and $1 - \bar{\Phi}(d_{p'}/2\sigma)$, both rising with $d_{p'} = \min_{p'' \neq p'} \|a_{p''} - a_{p'}\|$, where $\bar{\Phi}$ is the standard normal tail: isolated addresses are reached more reliably. This extends the readout clause, to which it reduces at $\varepsilon = 0$.*

Proof. The edited state is $c_{C_*} + a_{p'} + \varepsilon_\ell$. It is read as fact p'' only if the component of ε_ℓ toward $a_{p''}$ exceeds $\|a_{p''} - a_{p'}\|/2$, with probability $\bar{\Phi}(\|a_{p''} - a_{p'}\|/2\sigma) \leq \bar{\Phi}(d_{p'}/2\sigma)$; a union bound over the $n-1$ other addresses gives the lower bound. A component toward the nearest address beyond $d_{p'}/2$ alone causes an error, which gives the upper bound. □

Models and measurement. We use Qwen2.5-7B-Instruct, Gemma-3-4B-IT, and Llama-3.1-8B-Instruct. The *main blocks* are 20, 22, and 15, fixed on separate word sets before testing. Each edit occurs once, before the first answer token, in native chat templates (Appendix A). Tests use 64 new word sets per model, and calibration uses others. We score complete greedy answers, counting invalid and truncated outputs as failures.

Invariance and reversal hold across models. In two-fact contexts (Appendix B), the query change behaves as Proposition 3.4 predicts (Figure 2c,d). Over 64 new name pairs, its mean cosine with the ordinal vector is 0.93, 0.88, and 0.92 in Qwen, Gemma, and Llama, and the ordinal vectors of the two arrangements have cosines of -0.93, -0.82, and -0.85. Eleven further models show the same signs; the weakest reversal cosine is -0.44, in base Gemma-3-4B (Figure 2d).

4 Questions Are Organized by Order of Mention

We next test Proposition 3.5: the query-dependent state is set by the queried fact's order of mention.

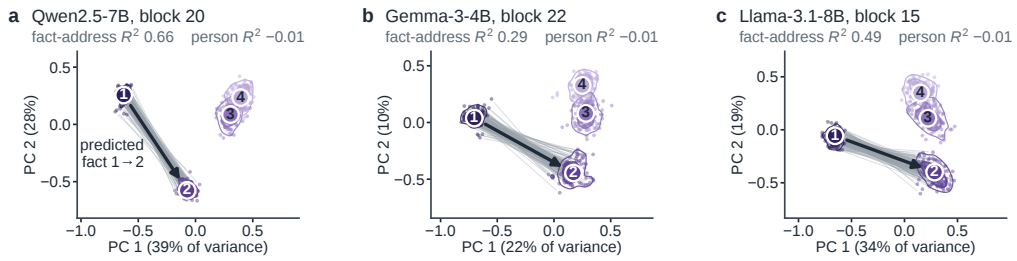

Figure 3: **Query states are organized by order of mention, not by the queried person.** (a–c) Query-dependent states of held-out word sets at each model's main block, projected onto the top two principal directions of the four fact-address templates fit on eight other word sets; axis labels give the share of held-out query-dependent variance on each direction, and the unit is the fact-1 to fact-2 template distance. Each dot is one context under one question, colored by the queried fact's order of mention; gray lines join a context's fact-1 and fact-2 questions, the dark arrow is the predicted change $a_2 - a_1$, and contours enclose 50% and 85% of each order's density. Panel headers: held-out R^2 of fact-address and person templates.

Query states group by the queried fact. A *fact-address template*, the mean query-dependent state over calibration contexts whose queried fact has a given order of mention, explains 66%, 29%, and 49% of the query-dependent variance of held-out word sets in Qwen, Gemma, and Llama, and a *person template*, by the queried person (the k-th name of each word set), explains none (Figure 3). Inside one word set, where the names are fixed, the named person sets the state in early blocks but adds at most 0.10 to the fact-address template's R^2 at the main blocks (Appendix C.1, Figure 7).

The address belongs to the fact, not to the name. When one fact has two subjects, the order of facts and the order of names come apart: Carol is the third name but the subject of the second fact in *Alice and Bob eat the apple. Carol eats the pear.* Such query states lie closer to the fact-address template of their fact than to the template at their name's order in 99.2%, 96.2%, and 96.2% of cases in Qwen, Gemma, and Llama (Appendix C.2). In two-place facts such as *Alice loves Bob*, query states are organized by fact and *argument slot*, which argument is asked for, not by person: a fact-by-slot template has held-out R^2 of 0.82, 0.65, and 0.70 (Appendix C.3, Figure 8).

5 The Fact Address Transfers across Contexts

If the address is shared, it should also move between contexts: by Proposition 3.6, an ordinal vector calibrated on other contexts should select the designated fact in a new context, which supplies its content. With two facts any change reaches the only other content, so we use four and the twelve directed changes between their orders of mention, each averaged over calibration contexts (Appendix A). In a list that begins *Petra eats the buckwheat* and *Freya eats the amaranth*, $v^{1\to2}$ added to the question about Petra should give *amaranth*.

Table 1: **An ordinal vector from other contexts selects the designated fact in a new context.** Designated answers (%), question unchanged; ± is half the 95% word-set bootstrap interval; shading is proportional to the rate. (A) 64 new word sets, 12,288 cases per model; Random: at most 0.01%. (B) 64 further word sets, 1,024 cases per row, all directions scaled to one length; the question and name rows are computed without the facts (Appendix D). Ceiling: the receiver's own state for the designated question; it differs between panels because the word sets differ. (C) The receiver's state replaced by a donor's; 16 donors × 64 receivers; designated: the receiver's fact at the order of mention of the other question.

	Qwen	Gemma	Llama
A. Transfer to new contexts			
Ordinal vector v	**84.1** ±2.0	**53.9** ±1.6	**36.2** ±1.6
Random	0.0	0.0	0.0
Ceiling	96.9 ±1.3	96.4 ±1.2	91.9 ±1.9
B. Where the vector comes from			
With facts (v)	**71.3** ±4.0	**41.7** ±2.4	**10.3** ±2.0
Question alone	0.0	0.0	0.0
Names listed, same order	2.5 ±1.1	1.3 ±0.8	3.9 ±1.4
Names listed, shuffled	0.0	0.1 ±0.1	0.2 ±0.2
Random	0.0	0.0	0.0
Ceiling	94.4 ±2.4	97.9 ±1.1	93.2 ±2.4
C. Whole state from another context			
Donor asked another fact	**80.9** ±3.1	**72.0** ±2.7	**51.3** ±3.1
Donor asked the receiver's question	1.7 ±0.9	13.9 ±1.6	6.1 ±1.1

The answer follows the receiver's facts. Across 64 new word sets, the ordinal vector selects the designated fact in 84.1% of cases in Qwen, 53.9% in Gemma, and 36.2% in Llama (Table 1A), and equally long random directions in at most 0.01% (Appendix D). It still works in short paragraphs that state the facts with varied verbs among unrelated sentences (44.5%, 42.4%, and 14.0%; Appendix G.3).

Facts define the vector. With every direction scaled to one length, the ordinal vector calibrated with the facts selects the designated answer in 71.3%, 41.7%, and 10.3% of cases. Directions computed from inputs without the facts select none from the question alone and at most 3.9% from a list of the names (Table 1B): the vector arises from reading the question against the facts.

A whole state from another context. A stronger test replaces the receiver's query state for question s with a *donor* context's state. If the donor asked question t about another fact, receivers answer with their own fact at the order of mention of question t in 80.9%, 72.0%, and 51.3% of cases, against 1.7%, 13.9%, and 6.1% when the donor asked the receiver's own question (Table 1C). It rarely yields the donor's complete answer (Appendix F.1): the state carries which fact to read, not what the donor read.

Which ordinal vector transfers. It is the one chosen by the receiver's orders of mention, not by its names, relation, or token distances. When one fact has two subjects, a vector chosen by the order of facts selects the designated answer in 65.3% of Qwen cases, and one chosen by the order of names in none (Appendix C.2). When the facts are rearranged, the vector of the original question pair mostly fails, and the one chosen by the new orders of mention raises Qwen's transfer for person

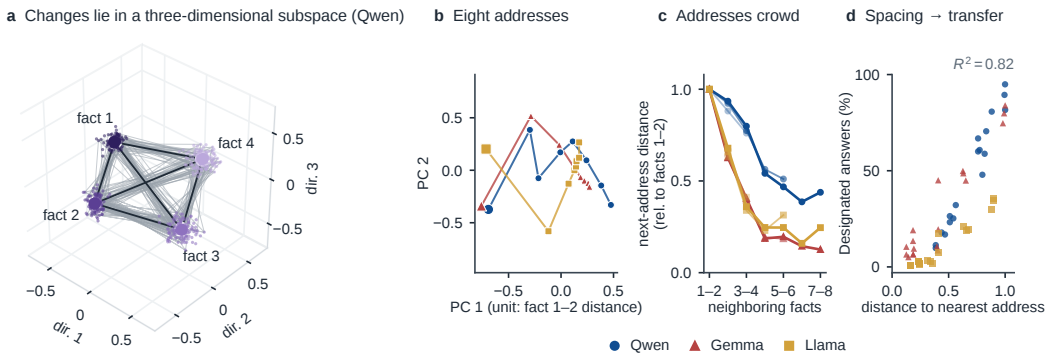

Figure 4: **Fact addresses are low-rank and ordered by mention.** (a) Qwen's held-out contexts in the address subspace fit on eight other word sets: query states (small dots, colored by the queried fact) and changes (gray); large dots: the four fact addresses, joined as a tetrahedron. (b) The eight fact addresses of eight-fact lists in their first two principal directions (unit: fact-1 to fact-2 distance); large marker: fact 1. (c) Distance between neighboring fact addresses, relative to facts 1–2, for four, six, and eight facts (light to dark). (d) Each address's distance to its nearest neighbor (unit: fact-1 to fact-2 distance) against transfer to that fact (Proposition 3.9), for 54 addresses; R^2 after removing each model's mean.

changes from 9.0% to 69.7% (Appendix D.2). Across six relations, a vector calibrated only on *eats* and rescaled to the receiving relation's change length comes within five percentage points of each relation's own (Appendix D.3). When filler separates order from token distance, the vector selects the fact at the calibrated order over the one at the calibrated absolute distance by 72 percentage points in Qwen, and by less in Gemma and Llama (Appendix D.4).

6 Fact Addresses Are Low-Rank

6.1 Fact addresses span a shared low-rank subspace

If one address set serves every context, the addresses should span a small shared subspace. As Corollary 3.7 predicts, the three-dimensional address subspace spanned by fact-address templates from other contexts holds 69%, 33%, and 52% of the squared norm of held-out changes in Qwen, Gemma, and Llama, within 1–8 points of the best three-dimensional subspace, against 0.1% (about $3/d$) for a random one (Figure 4a; Appendix E.1). The subspaces of six relations nearly coincide (mean cosine of the largest principal angle 0.78–0.89), and ordinal vectors estimated on disjoint word sets compose along a path: $v^{p \to p'} + v^{p' \to p''}$ has cosine 0.99, 0.97, and 0.98 with $v^{p \to p''}$. Fact addresses also predict how person and relation changes interact (Appendix E.4).

6.2 Later addresses crowd together, and the first is easiest to reach

In lists of up to eight facts, the fact addresses form an ordered chain whose later links shorten (Figure 4b,c). Each address is fixed once its fact is read: the vector from fact 1 to fact p keeps its direction when later facts are removed (median cosine 0.99 with the eight-fact vector; Appendix E.3), so the direction of $a_p - a_1$ does not depend on the list length. Transfer follows this spacing, as Proposition 3.9 predicts: the farther a fact address lies from its nearest neighbor, the better an ordinal vector reaches it ($R^2 = 0.82$ within models; Figure 4d). Accordingly, transfer is highest to the first fact (82–95% in Qwen), falls toward the middle, and in most lists rises slightly at the last: a serial-position curve with strong primacy, as in human recall (Henson, 1998; Appendix E.2, Figure 11).

7 Where and When the Fact Address Is Shared

Addressing is an intermediate step: inside one word set, the query state carries first the named person, then the fact address, then the requested content (Appendix C.1). Hypothesis 3.3 should



Figure 5: **Through depth, the fact address is first shared across contexts, then held only in whole states, then replaced by the answer's content.** (a) The two interventions and three outcomes on one Gemma example, and each block classified over relative depth: shared address (the ordinal vector selects the receiver's fact in at least 10% of cases), whole-state address (only donor states do, at least 10%), and content (the donor's first token reaches 50%); blocks are classified one by one, so the first two can interleave; lower rows: other Qwen models (Appendix G.1). (b–d) The three outcomes through depth, with the ceiling as reference, on the prompts of Table 1C (one arrangement, two content assignments, four person changes; 16 donors × 64 receivers), where the ordinal vector selects 55.5%, 44.3%, and 25.8% at the main blocks, below Table 1A over all arrangements; bands: 95% intervals resampling donor and receiver word sets. Shading: the depth window, blocks where the ordinal vector selects at least 10%.

therefore fail near the output, where the state carries the retrieved content x_p of $\beta_C(p)$: what it selects then depends on the context it came from. To see where the address is shared, we apply the two interventions of Section 5, the ordinal vector and the donor state, at every block and record three outcomes: the receiver's fact selected by the ordinal vector, the receiver's fact selected by the donor state, and the donor's own first answer token (Figure 5a; Appendix F.1). A state carrying only an address selects the receiver's content; one carrying retrieved content, the donor's.

7.1 The shared address occupies a depth window

In early blocks even the ceiling has no effect, though order of mention is already readable (Figure 13); in a late-middle *depth window*, ordinal vectors and donor states both select receiver contents (Figure 5b–d). The ordinal vector fails first: Gemma's falls from 44.3% at block 22 to zero at block 23 (Figure 5c), while the ceiling stays at 96%. In Llama, from block 17 to 21, the ordinal vector stays below 1% while donor states select 30–36%: each whole state holds the address, which is no longer a shared difference. The donor's own first token then takes over, reaching 67.0% at block 23 in Gemma (*baked apple crumble* joins the donor's first word to the receiver's content). The geometry agrees: fact-address R^2 collapses where the shared address ends, from 0.29 at block 22 to 0.05 at block 23 in Gemma, and each state aligns with its answer's output direction (Figure 13).

7.2 The fact address forms during pretraining

We repeat the tests on ten checkpoints of OLMo-2-7B (Team OLMo et al., 2024), from 5B training tokens to the released model (Figure 6; Appendix F.2). Questions switch from the named person to

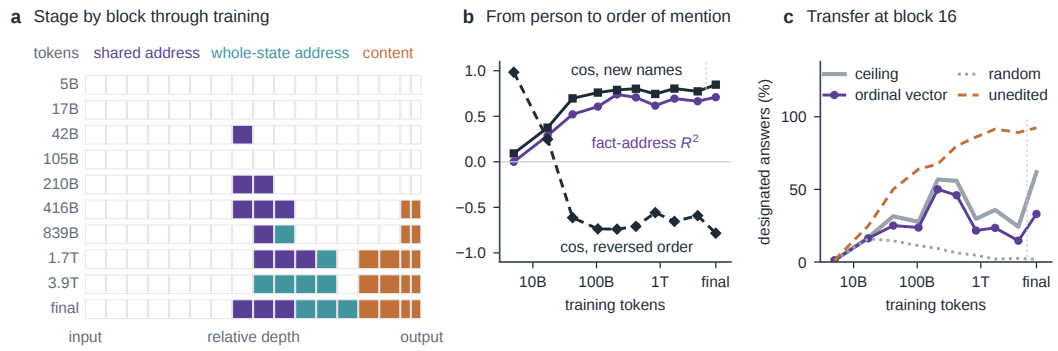

Figure 6: **In pretraining, questions switch from the named person to order of mention before the model answers reliably, and the full depth profile forms late.** Ten checkpoints of OLMo-2-7B, from 5B training tokens to the released model (*final*, after mid-training). (a) Stages as in Figure 5a, with each threshold applied to the gain over the unedited answer; all three stages first co-occur at 839B. (b) Fact-address R^2 (peak over blocks) and the cosines of Proposition 3.4 at block 16. (c) Designated answers at block 16; dashed: unedited accuracy; dotted: random directions.

order of mention: at 5B tokens the query change is the same in both arrangements (cosine $+0.98$), as if questions were represented by the named person, and it reverses by 42B (-0.61), as Proposition 3.4 predicts, while the model answers only half the questions; fact-address R^2 rises from 0.00 to 0.52 over the same span. The depth profile forms later (Figure 6a). Shared-address blocks appear at 42B, are absent at 105B, and are present from 210B through 1.7T; content blocks appear from 416B, and all three stages first co-occur at 839B. At 3.9T no shared-address block reaches threshold, and only the released model shows all three as one contiguous profile. Transfer at block 16, the released model's main block, peaks at 210B (50.1% against a 56.8% ceiling) and then declines as the depth profile moves away from it.

8 Discussion

Order of mention is learned from pretraining. Resolving a question by order of mention separates the structure of a context from its contents: one address, which counts facts rather than tokens, serves any names and contents, and the context supplies them. No objective asks for this separation, yet OLMo-2-7B learns it early in pretraining, after first organizing questions as if by the named person. It also bears signatures reminiscent of human discourse processing: the first-mentioned fact is the easiest to reach, and later ones crowd together, as in serial recall (Gernsbacher & Hargreaves, 1988; Henson, 1998). Unlike linear concept representations (Park et al., 2024), a fact address is a place in the context, not a meaning.

Fact retrieval as addressing. Retrieving a stated fact may then depend on where it stands: crowded addresses in the middle of a list transfer least, which may help explain weaker retrieval in the middle of long contexts (Liu et al., 2024) and reasoning that changes when premises are reordered (Chen et al., 2024). Our edits test the address in the question's state; long lists also use lexical and reflexive retrieval (Gur-Arieh et al., 2026). Our contexts are templated lists and short paragraphs; Gemma's fact-address templates explain only 29% of the variance; Llama transfers weakly in paragraphs and with two-subject facts (12–14%); and each edit touches one token position at one block. Long documents and multi-step reasoning remain open.

9 Conclusion

Language models address stated facts by order of mention: a question moves the state to a fact address shared across contexts, and the context supplies what is stated there. Questions are organized in a shared coordinate system of orders of mention, through which a model reaches particular facts.

AI USE STATEMENT

Generative AI assistants were used for tasks that require disclosure: they helped formulate the hypothesis and mathematical claims and write their proofs, design the experiments and stimuli, implement the methods, and analyze and interpret the results; they also drafted and edited text and wrote figure code. The authors set the research questions, directed and reviewed all AI-assisted work, and take full responsibility for the content.

REFERENCES

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. Linear algebraic structure of word senses, with applications to polysemy. *Transactions of the Association for Computational Linguistics*, 6:483–495, 2018.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Xinyun Chen, Ryan Andrew Chi, Xuezhi Wang, and Denny Zhou. Premise order matters in reasoning with large language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 6596–6620. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/chen24i.html.

Qin Dai, Benjamin Heinzerling, and Kentaro Inui. Representational analysis of binding in language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17468–17493, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.967. URL https://aclanthology.org/2024.emnlp-main.967/.

Qin Dai, Benjamin Heinzerling, and Kentaro Inui. Cell-based representation of relational binding in language models. *arXiv preprint arXiv:2604.19052*, 2026. URL https://arxiv.org/abs/2604.19052.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

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

Jiahai Feng, Stuart Russell, and Jacob Steinhardt. Monitoring latent world states in language models with propositional probes. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/3132d0405fabe24b2a7b6cd7ba9de6b5-Abstract-Conference.html.

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

Morton Ann Gernsbacher and David J. Hargreaves. Accessing sentence participants: The advantage of first mention. *Journal of Memory and Language*, 27(6):699–717, 1988. doi: 10.1016/0749-596X(88)90016-2.

Daniel Gildea and Daniel Jurafsky. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288, 2002. doi: 10.1162/089120102760275983. URL https://aclanthology.org/J02-3001/.

Yoav Gur-Arieh, Mor Geva, and Atticus Geiger. Mixing mechanisms: How language models retrieve bound entities in-context. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/2eeff35664016c7f0f8aa704f0d9a83e-Abstract-Conference.html.

Graeme S. Halford, William H. Wilson, and Steven Phillips. Relational knowledge: the foundation of higher cognition. *Trends in Cognitive Sciences*, 14(11):497–505, November 2010. ISSN 1364-6613. doi: 10.1016/j.tics.2010.08.005. URL http://dx.doi.org/10.1016/j.tics.2010.08.005.

Richard N. A. Henson. Short-term memory for serial order: The Start-End model. *Cognitive Psychology*, 36(2):73–137, 1998. doi: 10.1006/cogp.1998.0685.

Eghbal A. Hosseini and Evelina Fedorenko. Large language models implicitly learn to straighten neural sentence trajectories to construct a predictive representation of natural language. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/88dddaf430b5bc38ab8228902bb61821-Abstract-Conference.html.

Zhimin Hu, Lanhao Niu, and Sashank Varma. Language models represent and transform concepts with shared geometry. *arXiv preprint arXiv:2607.04525*, 2026. URL https://arxiv.org/abs/2607.04525.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20617–20642, 2024. URL https://proceedings.mlr.press/v235/huh24a.html.

Yibo Jiang, Goutham Rajendran, Pradeep Kumar Ravikumar, Bryon Aragam, and Victor Veitch. On the origins of linear representations in large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 21879–21911, 2024. URL https://proceedings.mlr.press/v235/jiang24d.html.

Dhruva Karkada, Daniel J. Korchinski, Andres Nava, Matthieu Wyart, and Yasaman Bahri. Symmetries in language statistics shape the geometry of model representations. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/63405.

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, 2024. doi: 10.1162/tacl_a_00638.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372. Curran Associates, Inc., 2022. doi: 10.52202/068431-1262. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/6f1d43d5a82a37e89b0665b33bf3a182-Paper-Conference.pdf.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 746–751. Association for Computational Linguistics, 2013. URL https://aclanthology.org/N13-1090/.

Alexander Modell, Patrick Rubin-Delanchy, and Nick Whiteley. The origins of representation manifolds in large language models. *arXiv preprint arXiv:2505.18235*, 2025. URL https://arxiv.org/abs/2505.18235.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020. doi: 10.23915/distill.00024.001. https://distill.pub/2020/circuits/zoom-in.

Gustaw Opiełka, Hannes Rosenbusch, and Claire E. Stevenson. Causality $\neq$ invariance: Function and concept vectors in LLMs. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/86b3697c4eb7792c951831636bfdacd5-Abstract-Conference.html.

Core Francisco Park, Andrew Lee, Ekdeep Singh Lubana, Yongyi Yang, Maya Okawa, Kento Nishi, Martin Wattenberg, and Hidenori Tanaka. ICLR: In-context learning of representations. In *International Conference on Learning Representations*, 2025a. URL https://arxiv.org/abs/2501.00070v2.

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*, volume 235 of *Proceedings of Machine Learning Research*, pp. 39643–39666, 2024. URL https://proceedings.mlr.press/v235/park24c.html.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. In *International Conference on Learning Representations*, 2025b. URL https://arxiv.org/abs/2406.01506.

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

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via contrastive activation addition. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Paul Smolensky. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, 46(1–2):159–216, 1990. doi: 10.1016/0004-3702(90)90007-M. URL https://www.sciencedirect.com/science/article/pii/000437029090007M.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, et al. 2 OLMo 2 furious. *arXiv preprint arXiv:2501.00656*, 2024.

Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron Wallace, and David Bau. Function vectors in large language models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 17282–17333, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/4ae163cb8788970e53b4fd9578141139-Paper-Conference.pdf.

Kevin Ro 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 *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=NpsVSN6o4ul.

Yufa Zhou, Yixiao Wang, Xunjian Yin, Shuyan Zhou, and Anru R. Zhang. The geometry of reasoning: Flowing logics in representation space. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ixr5Pcabq7.

APPENDIX CONTENTS

The appendix follows the main text. Appendix A gives the shared protocol, and each later part opens with the main-text claim it supports: invariance and reversal in two-fact contexts (B; Section 3); query states organized by the queried fact's order of mention, including coordinated subjects and two-place relations (C; Section 4); transfer of the fact address, what the facts contribute, whole donor states, and what chooses the vector: rearranged facts, six relations, and token distance (D; Section 5); the address subspace, longer lists, prefixes, and the person–relation interaction (E; Section 6); the depth window and pretraining checkpoints (F; Section 7); and the same tests across model scale, base weights, and prose (G; Sections 3, 5 and 7).

A Setup **14**

B Invariance and Reversal in Two-Fact Contexts **14**

C Questions Are Organized by Order of Mention **15**
C.1 Query-State Geometry . . . 15
C.2 Coordinated Subjects: The Order of Facts, Not of Names . . . 16
C.3 Two-Place Relations . . . 17

D The Fact Address Transfers across Contexts **18**
D.1 What the Facts and Whole Donor States Contribute . . . 18
D.2 Rearranged Facts . . . 18
D.3 Six Relations . . . 19
D.4 Order of Mention versus Token Distance . . . 19

E Fact Addresses Are Low-Rank **21**
E.1 The Address Subspace . . . 21
E.2 Longer Lists . . . 21
E.3 Prefixes of a List . . . 22
E.4 Person and Relation Do Not Add . . . 23

F Where and When the Fact Address Is Shared **24**
F.1 The Depth Window . . . 24
F.2 Pretraining Checkpoints . . . 25

G Generality: Model Scale, Base Weights, and Prose **26**
G.1 Model Scale . . . 26
G.2 Base and Instruction-Tuned Weights . . . 26
G.3 Facts Stated in Prose . . . 27

A SETUP

Every intervention uses one protocol: a change of the query state is calibrated on some word sets, added once to contexts built from new word sets, and judged by the complete answer.

Models and main blocks. We use Qwen2.5-7B-Instruct, Gemma-3-4B-IT, and Llama-3.1-8B-Instruct; block indices are zero-based Transformer block outputs. Qwen's main block 20 and Gemma's block 22 were fixed in earlier runs on separate block-selection word sets. Llama's block 15, and the main block of every model in Appendix G, come from one rule on those word sets. For each tested block, the rule scores the ordinal vector's gain over the unedited answer, minus the mean gain of random directions, minus any rise in invalid answers; it finds the best block and chooses the earliest block within two points of it. The models of Appendix G test every second block, and where no block reaches two points, the block nearest Qwen2.5-7B's relative depth is used. No main block was chosen on test word sets.

Word sets. A *word set* is a set of names and contents: four people and four contents in the four-person experiments. Except where noted, each experiment draws new names and new contents, mostly multiword dishes (*ackee and saltfish*, *pad see ew*). Calibration and test word sets share no complete name or content string, although subwords can recur. Within every word set, the four answers have pairwise distinct first tokens under the model's tokenizer. Every experiment tests on 64 word sets not used for calibration, except the depth curves of Appendix G and Appendix F.2, which use 8 donor and 32 receiver word sets. Unless stated otherwise, calibration uses 16 word sets; Table 1A–B, the two-fact contexts of Appendix B, and the rearranged facts of Appendix D.2 use 8.

A context states one fact per person, *Petra eats the buckwheat.*, followed by a question, *What does Petra eat?*, under the instruction *Read the sentences. Answer the question with only the food, in lowercase, with no explanation.* A *content assignment* assigns the four contents to the four people; an *arrangement* orders the four facts. Both rotate cyclically, so every person and every content occupies every order of mention. For Table 1A, the ordinal vector $v^{p \to p'}$ for a directed pair of orders of mention averages Equation (1) over all calibration contexts, content assignments, and arrangements that realize it. Each test word set crosses four arrangements, four content assignments, and twelve directed changes: 192 cases per word set, 12,288 per model and condition.

Intervention. We edit the query state, the state of the last prompt token, at the output of one block, once, while the prompt is processed. All later blocks and positions are recomputed on the receiving context, and generation continues from there. Prompts use each model's native chat template (Appendix G.2 adds a plain prompt), and computation is in FP32. Random controls are two fixed random directions with the length of the ordinal vector they replace. The *ceiling* replaces only the receiver's last-token state at the edited block with its own state for the designated question q'; the question tokens still read q, so it measures how far replacing this one state moves the answer.

Decoding and scoring. Answers are greedy, with at most 16 new tokens. We score the complete normalized answer string: the *designated* answer is the content that the designated question q' asks for in the receiving context. An answer is *invalid* when its normalized string is not exactly one content of the context. Invalid and truncated answers remain in every denominator, as do cases whose unedited answer is wrong.

Intervals. Intervals are 95% percentile intervals from 5,000 bootstrap draws over test word sets, keeping all content assignments, arrangements, questions, and paired conditions of a word set together. For donor states, donor and receiver word sets are resampled independently. All intervals are pointwise, one per reported rate. In the text, intervals are given in brackets; in tables, $\pm$ gives half the interval. Differences between rates are in percentage points ("points").

B INVARIANCE AND REVERSAL IN TWO-FACT CONTEXTS

Section 3 claims that in two-fact contexts the query change between the two people's questions keeps its direction under new names and reverses with the order of mention (Proposition 3.4); this section gives the materials, the coordinates of Figure 2c, and the cosines of Figure 2d.

Materials. Each context states two facts, *Ruby eats the tuna. Simon eats the guava.*, and asks *What does Ruby eat?* There are 72 word sets, each with two names and four contents in two sets of two, and a context uses the two contents of one set: 8 word sets calibrate, and 64 are test word sets. All 144 names are distinct, calibration and test contents disjoint, and a word set's contents start with different letters. Each word set is realized in the active form and in the passive form (*The tuna is eaten by Ruby.*), with both content assignments and both arrangements, and the question follows the facts. Every model answers every calibration question correctly.

Coordinates. Label the two people A and B within each word set; the labels stay with the people when their sentences are reversed. For arrangement $o \in \{0, 1\}$ ($o = 0$: A's fact first), let v_o be the mean of $h(C, q_B) - h(C, q_A)$ over both content assignments and the eight calibration word sets, so the ordinal vector is $v = v_0$. Figure 2c uses three axes fit on calibration word sets only: the arrangement contrast $u_1 \propto v_0 - v_1$, close to v_0 because $v_1 \approx -v_0$; the component u_2 of $(v_0 + v_1)/2$ orthogonal to u_1 ("which people are named"); and the top principal direction of the calibration context offsets $\frac{1}{2}(h(C, q_A) + h(C, q_B))$ after projecting out u_1, u_2. States are centered by the mean calibration offset and expressed in units of $\|v\|$, the root-mean-square norm of v_0, v_1. Figure 2c draws the offset axis stretched for display; its tick labels give measured values. The eight displayed test contexts are the active contexts with the first content assignment at the 10th, 37th, 63rd, and 90th percentiles of the offset coordinate, four in each arrangement.

Along u_2, test pairs separate by whether a woman or a man is named first (every calibration pair names a woman first), so u_2 carries a contrast between the two named people while u_1 carries the order of mention. The plane of u_1, u_2 holds 91.7%, 88.8%, and 94.8% of the squared norm of the test changes in Qwen, Gemma, and Llama.

Cosines. Figure 2d uses full-dimensional states. For each of the 64 test word sets, the change of that word set in a given arrangement has mean cosine 0.93, 0.88, and 0.92 with the calibration v_o of the same arrangement (minimum 0.85, 0.71, and 0.78). The changes calibrated in the two arrangements are nearly opposite: $\cos(v_0, v_1) = -0.93$, -0.82, and -0.85. The query change keeps its direction when the names change and flips when the order of mention flips, as a difference of two fact addresses should.

C Questions Are Organized by Order of Mention

Section 4 claims that the query-dependent state is set by the queried fact's order of mention, not by the named person, also when one fact has two subjects and when facts have two arguments; this section gives the query-state geometry and the tests behind that claim.

C.1 Query-State Geometry

Query states are organized by the order of mention of the queried fact (Figure 3a–c): fact-address templates predict held-out states, and person templates predict nothing.

Data. This geometry and that of Figure 13 use last-token states saved at every block for the four-person contexts behind Table 1C. Calibration contexts are 16 word sets $\times$ 4 content assignments $\times$ 4 arrangements, 256 contexts, each with all four questions; because the arrangement rotates, every person occupies every order of mention. Test contexts are the 64 test word sets $\times$ 2 content assignments in one arrangement, 128 contexts. For every block, the query-dependent state is $y(C, q) = h(C, q) - \frac{1}{4}\sum_{q'} h(C, q')$ at the last prompt token.

Templates and held-out variance. A fact-address template a_p averages $y(C, q)$ over calibration contexts by the order of mention p of the queried fact; a person template b_e averages by the queried person e. Held-out variance explained is $R^2 = 1 - \sum \|y - a_{p(q)}\|^2 / \sum \|y\|^2$, and likewise for b_e; a negative value means the template predicts worse than zero. We report the leave-one-word-set-out estimate on calibration contexts, here and at every block of Figure 13: templates fit on 15 word sets predict the held-out word set, and sums are pooled over the 16 held-out word sets. Only calibration contexts can separate the two templates: test contexts keep one arrangement, in which

person and order of mention coincide. At the main blocks the estimate is 0.66, 0.29, and 0.49 for the fact-address template and −0.01 for person in all three models.

Projection. The plane of Figure 3a–c is spanned by the top two principal directions of the four fact-address templates fit on eight calibration word sets. The points are the query-dependent states of the other eight calibration word sets, unseen by the fit, in all four arrangements and content assignments. Each model contributes 512 points, and the axis labels give the share of their variance on each displayed direction.

Within one word set. Across word sets the names change, so a person template can only carry what the k-th names of different word sets have in common. Inside one word set the four names and contents stay fixed while its 16 contexts put every person at every order of mention. For each calibration word set and each held-out context, we fit templates on the other 15 contexts of that word set and score the held-out context's four query-dependent states, pooling over the 16 word sets (Figure 7). At the first block the state is set by the person the question names (R^2 0.93, 0.99, and 0.98 in Qwen, Gemma, and Llama); this share falls to zero as the share of the fact address rises. The fact address peaks inside the window of Figure 5 in Qwen and Gemma (blocks 18 and 19) and earlier in Llama (block 9), and the requested content, a template that averages by the content the question asks for, takes over in later blocks. At the main blocks the queried person explains −0.04, −0.03, and 0.02 of the variance and adds 0.07, 0.02, and 0.10 to a fact-address template, which alone explains 0.64, 0.22, and 0.44. The requested content is already present there in Gemma (0.43), whose main block ends its window. At the main blocks the fact address, not the name, sets the state.

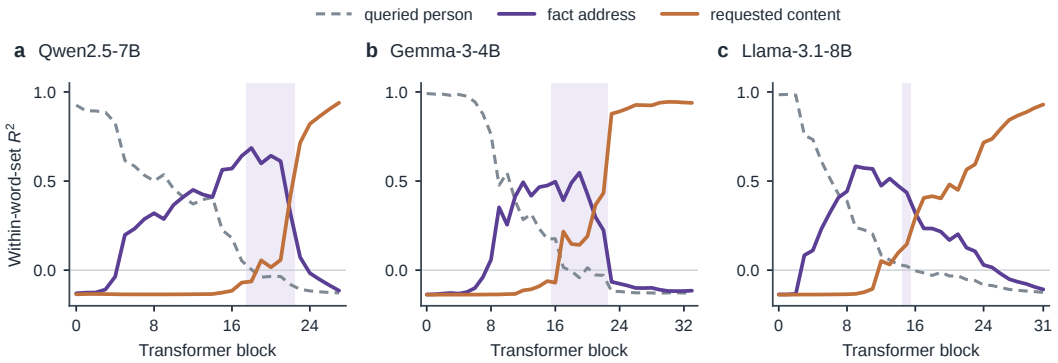

Figure 7: **Inside one word set, the query state passes from the named person to the fact address to the requested content.** Held-out R^2 of query-dependent states under templates for the queried person, the queried fact's order of mention, and the requested content, each fit on the other 15 contexts of the same word set and pooled over 16 calibration word sets. Shading: first to last block where the ordinal vector selects at least 10% of designated answers (Figure 5).

C.2 Coordinated Subjects: The Order of Facts, Not of Names

The address belongs to the fact, not to the name (Sections 4 and 5). In every other experiment each fact has one subject, so the k-th fact and the k-th name coincide. Here one of four facts has two coordinated subjects, *Alice and Bob eat the apple. Carol eats the pear. Dave eats the plum. Emma eats the fig.*, so five names are mentioned in four facts and, after the coordinated fact, a person's name order exceeds the order of their fact by one. Calibration uses ordinary four-fact lists of 16 word sets in four arrangements and four content assignments: fact-address templates a_p of the query-dependent state by the queried fact's order of mention, and the twelve ordinal vectors $v^{p\to p'}$. Test contexts come from 64 new word sets of five people and four multiword contents, with the coordinated fact at each of the four orders of mention and two content assignments, and every single-person question is asked: 2,560 questions, of which 1,280 have a fact order f that differs from their name order m.

Geometry. For each question with $f \neq m$ and $m \leq 4$ (768 questions), the query-dependent state y is compared with the template of its fact order, a_f, and with the template of its name order, a_m, in the address subspace. It is closer to a_f in 99.2% $[98.6, 99.7]$ of questions in Qwen, 96.2% $[94.9, 97.5]$ in Gemma, and 96.2% $[94.8, 97.5]$ in Llama, and the mean advantage $\cos(y, a_f) - \cos(y, a_m)$ is 0.94, 0.61, and 0.68. The full space gives the same fractions. Held-out calibration questions are assigned to their own template in 100%, 100%, and 99.4% of cases.

Behavior. For receivers whose fact and name orders differ and each target order of mention t, an ordinal vector chosen by fact order, $v^{f \to t}$, is compared with one chosen by name order, $v^{m \to t}$. Only targets for which the two orders predict different contents, and for which the name-order prediction is not the original answer, enter the contrast. The fact-order vector selects the content of fact t in 65.3% $[63.0, 67.7]$ of cases in Qwen, 34.0% $[32.4, 35.6]$ in Gemma, and 12.1% $[10.5, 13.7]$ in Llama. The name-order vector selects the content its order predicts in 0.0% of cases in all three models and mostly leaves the original answer. On receivers whose two orders agree, the same vectors select their target in 78.2%, 57.7%, and 21.9% of cases, and the ceiling in 97.9%, 97.1%, and 84.7%; random directions keep the original answer in at least 99.9% of cases, and unedited answers are correct in at least 99.9%. An ordinal vector acts only when it starts at the state's own fact address, so both measures place each question at the address of its fact, not of its name.

C.3 Two-Place Relations

Queries over two-place facts are organized by fact and argument slot, not by person (Section 4). Each context states three facts of one two-place relation among six people, such as *Alice loves Bob. Carol loves David. Emma loves Frank.*, with five relations (*loves*, *trusts*, *helps*, *visits*, *pays*). A question asks for one argument given the other, *Whom does Alice love?* or *Who loves Bob?* Every answerable question is a pair of an order of mention and an argument slot, so a query change can move to another fact (same slot), to the other argument of the same fact, or to both. Calibration uses 16 word sets with all arrangements and all content assignments, whose contents are people. The 64 test word sets, with new names whose first tokens are distinct, are edited once at the main block with each calibration mean, the means of the other four relations, two random directions, and the ceiling. Templates are fit on calibration word sets and scored on the test word sets.

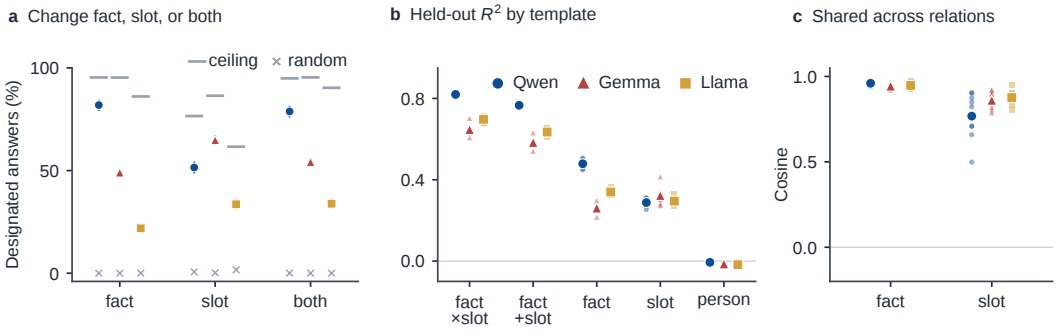

Figure 8: **Queries over two-place facts address a fact and an argument slot.** (a) Designated answers for changes of fact, of argument slot, and of both, against the ceiling (gray ticks) and random directions (×), on 64 new active contexts. (b) Held-out R^2 under fact-by-slot, additive (fact+slot), fact-only, slot-only, and person templates. (c) Cosines of fact and of slot changes across relations. In (b) and (c), small markers: individual relations; large markers: their mean.

All three changes transfer to new names: fact, slot, and joint changes select the designated person in 82%, 52%, and 79% of Qwen cases, 49%, 65%, and 54% of Gemma cases, and 22%, 34%, and 34% of Llama cases, while random directions stay at or below 1.7% (Figure 8a). For fact changes, a change calibrated on the other four relations comes within four points of the receiving relation's own. Query states are organized by the two coordinates: a fact-by-slot template explains 0.82, 0.65, and 0.70 of held-out variance, an additive template 0.77, 0.58, and 0.63, and the queried person nothing (Figure 8b). Fact changes estimated on different relations point the same way (mean cosine

0.94–0.96), and so do slot changes (0.77–0.88). A question about a two-place fact addresses the fact by its order of mention and, within it, the argument slot it asks for.

D THE FACT ADDRESS TRANSFERS ACROSS CONTEXTS

Section 5 claims that an ordinal vector from other contexts selects the designated fact in a new context, that the facts define this vector, that a whole donor state carries only which fact to read, and that the vector that transfers is chosen by the receiver's orders of mention, not by its names, relation, or token distances. Random directions of each ordinal vector's length select the designated fact in at most 0.01% of the 12,288 cases per model (Table 1A). We give the controls of Table 1B–C, then which vector transfers under rearranged facts, six relations, and token distance.

D.1 WHAT THE FACTS AND WHOLE DONOR STATES CONTRIBUTE

No-facts directions. Table 1B compares directions from different inputs at one length. The ordinal vector for a directed pair of orders of mention is the calibration mean of Equation (1) over eight calibration word sets, four arrangements, and four content assignments (128 query changes per pair). Each no-facts direction is the difference between the states for the two questions of the case, computed on an input without the facts: the question alone; the receiving names listed in the order of their facts; or the same names in a shuffled order. Every direction, including the random ones, is scaled to one common length, the root-mean-square length of the twelve factual calibration directions. Each of 64 test word sets contributes two arrangements, two content assignments, and four directed changes, 1,024 cases per row. These word sets and this subset of arrangements and content assignments differ from those of Table 1A, so the two panels' rates are not directly comparable. The no-facts inputs contain the question words and the names but no facts, and their directions select at most 3.9% of designated answers.

Whole donor states. In Table 1C, the donors are the 16 calibration word sets and the receivers the 64 test word sets, all in one fixed arrangement. For each donor, receiver, content assignment, and directed person change $s \to t$, the receiver asks question s, and its state for s is replaced by the donor's state for question t (another fact) or for question s (the receiver's own question); in both rows the designated answer is the receiver's fact at t, which gives 8,192 cases per condition. The paired gain of the first replacement over the second is 79.1 $[75.4, 82.6]$, 58.1 $[54.4, 61.9]$, and 45.2 $[42.0, 48.2]$ percentage points, and it is positive for every donor: the smallest per-donor gain is 68.0, 50.2, and 38.5 points. On the same prompts, which are also those of the depth curves of Figure 5, the ordinal vector selects 55.5%, 44.3%, and 25.8%, below the whole donor state and below its rates in Table 1A (84.1%, 53.9%, and 36.2%), which average over all arrangements. Complete donor answers are rare at the main blocks (Figure 5), and at the final block the same replacement gives them in only 2.1%, 8.2%, and 3.6% of cases, with 91.0%, 87.1%, and 79.5% invalid (Appendix F.1). A donor state carries which fact to read, not the donor's content.

D.2 REARRANGED FACTS

When facts move, the ordinal vector must follow their new orders of mention (Section 5). Each context states four facts about two people and two relations, *Alice eats the apple. Alice carries the pear. Bob eats the plum. Bob carries the peach.*, and a question asks about one of the four *cells*, a pair of a person and a relation; a person change keeps the relation, and a relation change keeps the person. Calibration uses eight word sets in the person-major arrangement above (grouped by person). The test contexts, on 64 word sets, state the facts in two held-out arrangements, $(1, 4, 2, 3)$ and $(3, 2, 4, 1)$, numbering the facts of the person-major arrangement, with four content assignments and four starting questions: 2,048 cases per condition and change type. Three edits target the same designated answer. The *original-pair* edit keeps the ordinal vector for the same pair of questions, which in the new arrangement connects different orders of mention. The *new-order-of-mention* edit takes the ordinal vector for the pair of orders of mention that the two questions occupy in the new arrangement. The *recalibrated* edit is estimated on calibration word sets in the held-out arrangement.

Choosing the ordinal vector by the new orders of mention recovers most of the recalibrated rate in Qwen and nearly all of it for Llama's person changes, and it more than doubles Gemma's rates while

Table 2: **When facts move, the ordinal vector must follow their orders of mention.** Designated answers (%) under two held-out arrangements, 64 test word sets and 2,048 cases per entry; ± is half the 95% word-set bootstrap interval; bold: the new-order-of-mention edit; ceiling: the receiver's own state for the designated question; no facts: the direction computed from the question alone.

Model	Change	Original pair	New order of mention	Recalibrated	Ceiling	Random	No facts
Qwen	Person	9.0 ±1.1	**69.7** ±2.3	80.9 ±2.5	90.2 ±1.7	0.0	0.0
	Relation	22.6 ±1.6	**61.6** ±2.1	68.5 ±1.9	77.3 ±1.9	5.7	7.0
Gemma	Person	17.6 ±0.9	**40.8** ±1.4	61.1 ±2.1	79.2 ±1.7	0.1	0.1
	Relation	5.5 ±1.3	**19.9** ±2.9	41.5 ±2.8	73.6 ±1.6	0.8	0.7
Llama	Person	6.2 ±1.4	**15.7** ±2.8	16.4 ±2.8	29.8 ±3.7	0.0	0.0
	Relation	0.7 ±0.5	**0.6** ±0.3	2.8 ±1.1	11.8 ±2.1	0.3	0.3

leaving a larger gap to recalibration (Table 2). Llama's relation changes do not act in these contexts even when recalibrated, and even the ceiling reaches only 11.8%.

D.3 Six Relations

The same ordinal vector acts under every relation (Section 5). Each context lists four facts under one of six relations, each with its own contents: *eats* a dish, *owns* an animal, *visited* a city, *works as* a job, *plays* an instrument, or *plays* a sport. The question asks for one person's content under the same relation, and the model answers with only that content. Each relation has 16 calibration and 64 test word sets, with new names and contents in the test word sets. Calibration follows Section 5: for each of the twelve directed changes between orders of mention, the mean query change over a relation's calibration contexts at the main block. Each test receiver is edited once with the ordinal vector of every relation, with the *eats* vector rescaled to the receiving relation's own change length, with two random directions of that length, and with the ceiling. Other cross-relation changes keep the source relation's length, so their gap to the receiving relation's own change mixes direction and length. Geometry uses the query-dependent states at this block, with templates fit on calibration and tested on test word sets.

Unedited answers are correct in at least 97.5% of cases for every relation and model, and the ceiling reaches 88–99%. Random directions select the designated fact in at most 0.03% of cases. Averaged over relations, the receiving relation's own vector selects the designated fact in 72%, 47%, and 18% of Qwen, Gemma, and Llama cases, and a vector from another relation in 70%, 46%, and 16% (Figure 9a). Rescaled to the receiving relation's length, the *eats* vector comes within five points of each relation's own vector in every model (paired differences from -4.0 to $+4.6$ points). Fact-address templates give held-out R^2 of 0.61–0.71, 0.26–0.35, and 0.41–0.52, while person templates stay within 0.01 of zero (Figure 9b). The ordinal vectors of two relations point the same way, with cosines of 0.79–0.96 over relation pairs and models, against 0.18–0.38 between mismatched ordinal vectors (Figure 9c); the address subspaces of the six relations nearly coincide as well (Appendix E.1). The relation changes the contents at an order of mention, not the ordinal vector between them.

D.4 Order of Mention versus Token Distance

The ordinal vector follows order of mention, not token distance (Section 5). When all facts have the same length, the k-th fact is also a fixed number of tokens before the question, so an address by order of mention cannot be told apart from one by distance. Each context of 64 new four-person word sets is rendered under eight length conditions with the same arrangement and content assignments: uniform facts; one fact lengthened by a neutral opening clause (four conditions); the second and fourth facts lengthened; filler lines between facts; and a filler paragraph before the question. Ordinal vectors are calibrated on uniform contexts from 16 other word sets (primary) or on mixed-length contexts (secondary) and applied unchanged.

For an edit from order of mention p to p' on a question about fact s, let $d(s)$ be the token distance of s from the question and $d_{\mathrm{ref}}[p]$ the distance of order p in uniform contexts. The *absolute* distance address selects the fact whose distance is nearest $d_{\mathrm{ref}}[p']$; the *relative* one selects the fact whose

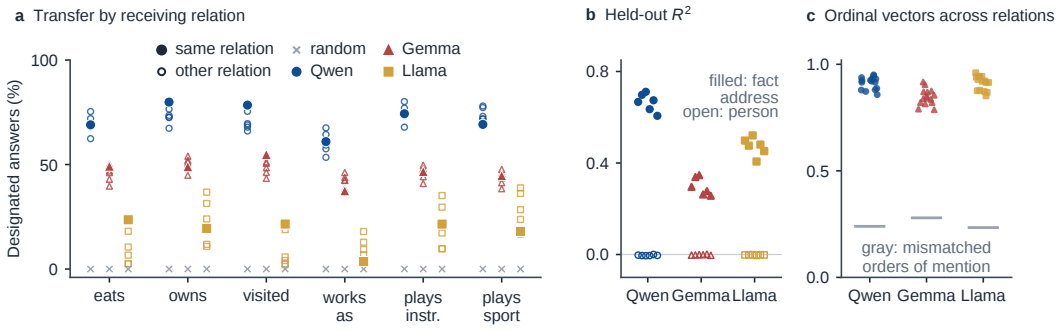

Figure 9: **The same ordinal vector acts under every relation.** (a) Designated answers by receiving relation: an ordinal vector calibrated on the same relation (filled), on each of the five other relations at its own length (open), and random directions (×); 64 test word sets, 3,072 cases per point. Not plotted: the *eats* vector rescaled to the receiving relation's change length comes within five points of the filled marker for every relation (paired differences -4.0 to $+4.6$ points). (b) Held-out R^2 of query states under fact-address (filled) and person (open) templates, one point per relation. (c) Cosine between ordinal vectors fit separately on two relations, one point per relation pair, averaged over pairs of orders of mention; gray bars are the cosine between mismatched ordinal vectors.

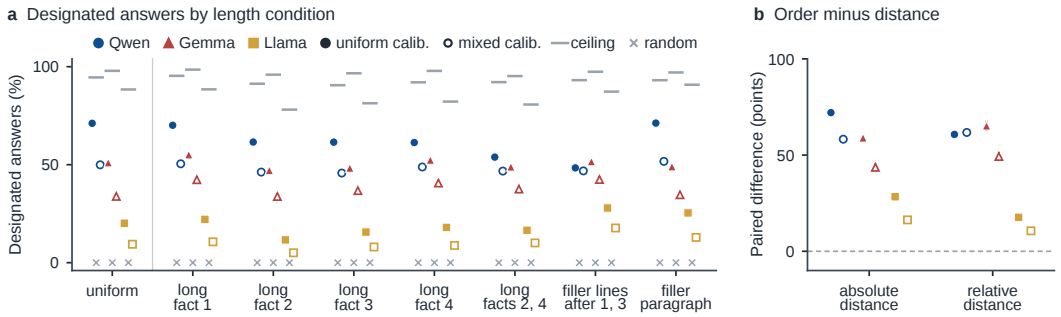

Figure 10: **The ordinal vector follows order of mention, not token distance.** (a) Designated answers by length condition for ordinal vectors calibrated on uniform (filled) or mixed-length (open) contexts, with the ceiling (gray ticks) and random directions (×). (b) Answers naming the fact at the calibrated order minus those naming the fact at the calibrated absolute or relative distance, on edits where the two facts differ from each other and from the unedited answer; markers as in (a), with 95% intervals over 64 word sets.

distance is nearest $d(s) + d_{\text{ref}}[p'] - d_{\text{ref}}[p]$; ties go to the fact nearer the question. Before the run we fixed both rules and the contrast: answers naming the fact at the calibrated order minus answers naming the distance-predicted fact, on edits where the two facts differ from each other and from the fact the unedited question asks about.

Transfer survives every length condition, from 48% to 71% in Qwen, from 47% to 55% in Gemma, and from 12% to 28% in Llama, with random directions at zero (Figure 10a). On edits where order and distance predict different facts, the ordinal vector selects the fact at the calibrated order far more often than the fact at the calibrated distance: by 72.1 $[70.0, 74.0]$ points in Qwen, 58.9 $[57.0, 60.9]$ in Gemma, and 28.4 $[26.6, 30.1]$ in Llama against absolute distance, and by 60.7, 65.2, and 17.6 points against relative distance. Mixed-length calibration gives the same verdict. Held-out query states are fit by fact-address templates (R^2 0.52, 0.26, and 0.41) and not by a token-distance template, the fact-address template of the order of mention whose mean token distance from the question in uniform-length contexts is nearest the queried fact's distance (0.02, -0.11, -0.15); both templates are fit on calibration word sets and scored on the test word sets. Transfer counts facts, not tokens.

E FACT ADDRESSES ARE LOW-RANK

Section 6 claims that fact addresses span a shared low-rank subspace, crowd together in longer lists, are fixed once their fact is read, and predict how person and relation changes combine; this section gives the analyses behind each claim.

E.1 THE ADDRESS SUBSPACE

Fact addresses span a shared low-rank subspace (Section 6): the three-dimensional span of the fact-address templates holds nearly as much of each held-out query change as the best three-dimensional subspace. The numbers of Section 6 use the main-block states of Appendix C.1, with templates fit on all 16 calibration word sets; Figure 4a, like Figure 3, fits the templates on eight of them and shows the states of the other eight.

Share of held-out changes. With the fact-address templates a_p of Appendix C.1 fit on the 16 calibration word sets, the address subspace is $A = \mathrm{span}\{a_p - a_1\}$, three-dimensional for four facts. For each of the 128 test contexts and each of its six pairs of orders of mention, the per-context change is $\Delta_C(p \to p') = h(C, p') - h(C, p)$, where $h(C, p)$ is the state for the question about the fact at order of mention p (the reverse change has the same share). Its share inside A is $\|P_A \Delta_C\|^2 / \|\Delta_C\|^2$, averaged over the 768 changes: 69%, 33%, and 52% in Qwen, Gemma, and Llama.

Two references bound this share. The best three-dimensional subspace for these changes, the top three principal directions of the held-out changes themselves, holds 70%, 41%, and 60%; this is an in-sample upper bound. One hundred random three-dimensional subspaces hold 0.1% on average, as expected for dimensions 3,584, 2,560, and 4,096. Relative to the gap between random and best, A reaches 99%, 81%, and 87%.

Across relations. For each relation r of the six in Appendix D.3, a separate subspace A_r is fit on that relation's 16 calibration word sets at the main block. For two relations s and t, the principal angles between A_s and A_t give three cosines. The smallest of them, the cosine of the largest principal angle, averaged over the 15 relation pairs, is 0.89, 0.78, and 0.88 (ranges over pairs 0.83–0.94, 0.66–0.90, and 0.80–0.94); for random three-dimensional subspaces, even the largest of the three cosines is only about 0.05.

Composition on disjoint word sets. Write $v_S(p \to p')$ for the mean of $\Delta_C(p \to p')$ over the calibration contexts of a subset S of word sets. The 16 calibration word sets are split into two halves, and one half into two disjoint quarters S_1, S_2. For each of the 24 ordered triples $p \to p' \to p''$ of distinct orders of mention, we compare the target $v_T(p \to p'')$ on the other half with the sum $v_{S_1}(p \to p') + v_{S_2}(p' \to p'')$, so the three vectors come from disjoint word sets and no identity links the sum to the target. Averaged over triples and over the assignment of halves and quarters, the cosine is 0.99, 0.97, and 0.98, and the relative error $\|\mathrm{sum} - \mathrm{target}\| / \|\mathrm{target}\|$ is 0.15, 0.23, and 0.18. The null relabels the summands over the 23 non-identity relabelings of the orders of mention $\{1, 2, 3, 4\}$: cosine -0.09 and relative error 1.45–1.56 in all three models. The target estimated directly on $S_1 \cup S_2$ agrees with $v_T(p \to p'')$ at cosine 0.995, 0.989, and 0.995, the noise ceiling. Ordinal vectors add along paths, as Corollary 3.7 predicts.

E.2 LONGER LISTS

In longer lists, fact addresses form an ordered chain whose later links shorten, and transfer is highest to the first fact (Section 6). The transfer test of Section 5 is repeated with lists of four, six, and eight single-relation facts, one person per fact, in all three models at their main blocks. Each test word set has eight people and eight contents, and a list of length N uses the first N of each, so the three lengths share names and contents; within a word set every person and every content appears once at every order of mention. Calibration uses 16 word sets and every combination of content assignment and arrangement; all $N(N-1)$ directed pairs of orders of mention are calibrated and tested (12, 30, and 56) on 64 test word sets, each in its N cyclic arrangements. Its word sets come from a different inventory, on which four-fact transfer is lower than in Table 1A (Qwen 75.6% versus 84.1%).

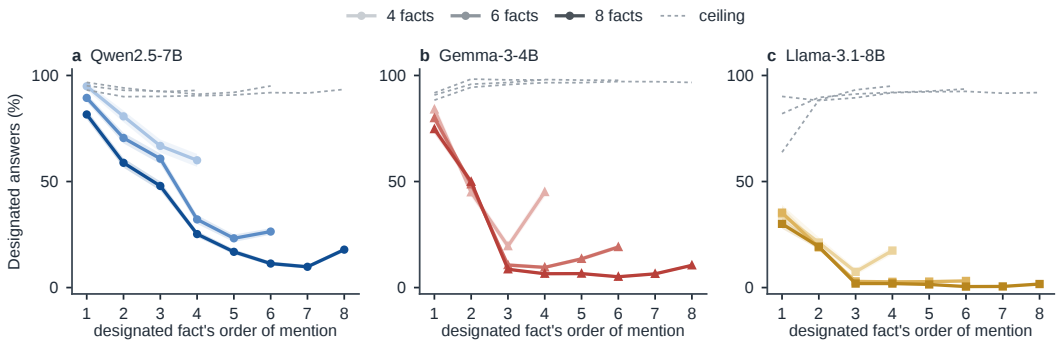

Figure 11: **Transfer falls with the designated fact's order of mention and is highest for the first fact.** Designated answers for the ordinal vector by the designated fact's order of mention, in lists of four, six, and eight facts, with word-set bootstrap 95% intervals (bands, mostly narrower than the markers); dashed lines are the ceiling. Random directions give at most 0.1% at every length.

Figure 11 shows two regularities. First, the ordinal vector transfers at every length, far above random directions: overall 75.6, 50.4, and 33.7% in Qwen for four, six, and eight facts, 48.4, 30.2, and 21.1% in Gemma, and 20.1, 11.0, and 7.2% in Llama, while the ceiling stays high at every length (Figure 11). Second, transfer depends on the designated fact's order of mention: it is highest for the first fact (Qwen 82–95%), falls toward the middle, and rises again at the last fact in eight of the nine lists, a serial-position curve with a strong primacy effect (Henson, 1998).

The chain of addresses. For lists of $N = 4, 6, 8$ facts, the templates $a_1, \ldots, a_N$ average the query-dependent states of the 16 calibration word sets, over all N arrangements and N content assignments, by the order of mention of the queried fact; the templates from two halves of these word sets have mean cosine at least 0.98. Lengths are in units of $\|a_2 - a_1\|$ at each N. Figure 4b projects the eight templates for $N = 8$ on their top two principal directions, which hold 58%, 84%, and 77% of their variance. Figure 4c plots $\|a_{p+1} - a_p\|$; for $N = 8$ the last link is 0.44, 0.13, and 0.25 of the first. Figure 4d plots, for each model, length N, and order of mention p (54 points), the distance $\min_{p' \neq p} \|a_p - a_{p'}\|$ from a_p to its nearest other address against the transfer rate to the fact at p on the test word sets (Figure 11). Subtracting each model's mean distance and mean rate before the least-squares fit gives $R^2 = 0.82$ (Spearman 0.89); without centering, $R^2 = 0.78$.

E.3 Prefixes of a List

Each fact address is fixed once its fact is read: removing later facts leaves earlier addresses in place (Section 6). Every eight-fact context of Appendix E.2 (the 16 calibration and 64 test word sets, 8 contexts each, following the Latin square of the longer lists: arrangement index 0–7, content assignment equal to the word-set index mod 8) is cut after its first k facts, $k = 2, \ldots, 8$, and asked about each of its k facts, with states at the main blocks. Templates $a_p^{(k)}$ are computed within each prefix length from all 80 word sets as in Appendix C.1 and compared only through their differences, which do not depend on how the query-dependent states are centered.

For facts before the prefix's last, the vector from fact 1 to fact p has median cosine 0.99 with its eight-fact counterpart in every model (minimum 0.95), and its length changes by at most 8%. The gap between neighboring addresses, relative to the first gap, depends on the fact's order of mention and not on the prefix length: in Gemma, gaps 2 to 4 are 0.62–0.66, 0.41–0.43, and 0.18–0.20 at every k, and the first gap is constant in absolute size. Gap sizes indexed by order of mention, fit on the other prefix lengths, predict the gaps of a held-out prefix length with $R^2 = 0.97$, 0.99, and 0.91 in Qwen, Gemma, and Llama; indexed by distance from the list's end, $R^2 < 0$. The last fact of a prefix deviates slightly (median cosine 0.91, 0.97, and 0.94), and in Llama its gap is larger by 0.075 of the first gap, a recency effect.

E.4 Person and Relation Do Not Add

Fact addresses predict that changing a question's person and relation separately does not add up to changing both (Section 6). Contexts state four facts about two people and two relations, *Alice eats the apple. Alice carries the pear. Bob eats the plum. Bob carries the peach.* Index the question cells (Appendix D.2) by person and relation, $(e, r) \in \{0, 1\}^2$, and let $p_o(e, r)$ be their fact's order of mention under arrangement o.

Lemma E.1 (Fact addresses induce a person–relation interaction)**.** *Under Hypothesis 3.3 with $\varepsilon = 0$, the query state of the question about cell (e, r) is $h_{er} = h_\ell(C, q_{er}) = c_C + a_{p_o(e,r)}$, and the four-cell contrast of these states is*

$$I = h_{11} - h_{10} - h_{01} + h_{00} = a_{p_o(11)} - a_{p_o(10)} - a_{p_o(01)} + a_{p_o(00)}. \tag{4}$$

The right-hand side need not vanish and can change with the arrangement.

Proof. Substitute $h_{er} = c_C + a_{p_o(e,r)}$ into the four-cell contrast. The context term has coefficients $+1-1-1+1 = 0$, leaving Equation (4). For a scalar example, take $a_1 = 0$, $a_2 = 1$, $a_3 = 2$, $a_4 = 4$, and place the cells $00, 01, 10, 11$ at orders of mention $1, 2, 3, 4$. The contrast is $4 - 2 - 1 + 0 = 1$. Exchanging the orders of mention of the last two cells changes it to $2 - 4 - 1 + 0 = -3$, although neither case has a separate person-by-relation term. □

Joint versus summed changes. The contexts above appear in four arrangements: grouped by person, grouped by relation, and their reversals. Calibration uses 16 word sets, and each change is the calibration mean for its pair of question cells within the same arrangement. From a starting cell, v_E changes the person, v_R the relation, and v_J both; the *same-start sum* is $v_E + v_R$, with both terms measured from the starting cell. Each of 64 test word sets contributes four arrangements, content assignments, and starting cells each: 4,096 cases per condition, all scored against the joint target.

Table 3: **The joint change selects the joint target more often than the sum of its parts.** Designated answers (%), 64 test word sets, 4,096 cases per edit; ±: half the 95% paired word-set bootstrap interval, in points; bold: the joint change.

Edit	Qwen	Gemma	Llama
Joint change v_J	**68.5**	**57.1**	**26.0**
Same-start sum $v_E + v_R$	43.8	44.0	18.8
Sum rescaled to $\|v_J\|$	35.3	31.7	14.1
Ceiling (own state for the joint question)	87.9	90.7	56.7
Joint minus sum	24.7 ±1.3	13.1 ±1.4	7.2 ±0.9
Joint minus rescaled sum	33.2 ±1.5	25.4 ±1.4	11.9 ±1.0

The sum is longer than the joint change, so its shortfall is not a lack of length: rescaling it to $\|v_J\|$ lowers its rate further (Table 3). The person and relation changes are nearly orthogonal (cosine 0.11, 0.13, and 0.12), and their sum has cosine 0.88, 0.86, and 0.89 with v_J; the interaction $I = v_J - v_E - v_R$, the four-cell contrast of Lemma E.1, has 54%, 59%, and 51% of the norm of v_J.

The interaction predicted by fact addresses. For the prediction of Lemma E.1, fact-address templates are fit on person-by-relation contexts from 16 calibration word sets, rendered in the twelve even permutations of the four facts with four content assignments. Held-out contexts are 64 test word sets rendered in the twelve odd permutations, which calibration never used, and in the original arrangement (an even permutation, also used in calibration; its word sets are held out), with four content assignments: 3,328 contexts. In each, the measured interaction $I = y_{11} - y_{10} - y_{01} + y_{00}$ of the query-dependent states is compared with $a_{p_o(11)} - a_{p_o(10)} - a_{p_o(01)} + a_{p_o(00)}$ computed from the templates and the orders of mention of the four cells in that context's arrangement. The mean cosine is 0.47, 0.38, and 0.46 at the main blocks (Figure 12). Relabeling the templates by each of the 23 non-identity relabelings of the orders of mention gives −0.02 in every model, and at the final block the cosine falls to 0.07, 0.04, and 0.07. The measured interaction is the one fact addresses predict, and the agreement vanishes at the output.

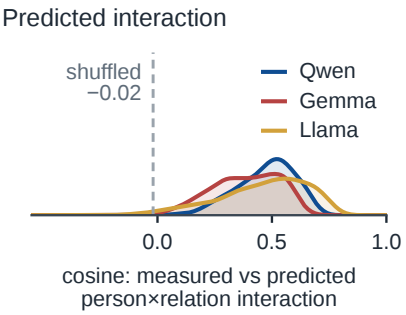

Figure 12: **Fact addresses predict the person–relation interaction.** Distribution, over 3,328 held-out contexts per model, of the cosine between the measured interaction I of the query-dependent states (Equation (4)) and its prediction from fact-address templates and the context's orders of mention; means 0.47, 0.38, and 0.46 in Qwen, Gemma, and Llama. Dashed line: the mean cosine when the order labels are shuffled (-0.02).

F Where and When the Fact Address Is Shared

Section 7 claims that the fact address is shared as a difference only within a depth window, and that it forms during pretraining; this section gives the every-block design and the checkpoint tests.

F.1 The Depth Window

The fact address is shared as a difference only within a depth window, after which whole states still carry it and then the donor's content takes over (Section 7). The depth curves in Figure 5b–d repeat the donor design of Appendix D.1 at every block of all three models, on the prompts of Table 1C: the 16 calibration word sets as donors, crossed with the 64 test word sets as receivers in one fixed arrangement, two content assignments, and four directed person changes. On these prompts the ordinal vector selects fewer designated facts at the main blocks than in Table 1A, which uses all arrangements (Appendix D.1). At every block the conditions are the unedited question, the ordinal vector re-estimated at that block from the calibration word sets, the ceiling, and the donor's state for the other question; the main and final blocks add the donor's state for the same question (Table 1C). Intervals for donor conditions resample donors and receivers independently; intervals for the other conditions resample receivers.

The example. Figure 5a shows the first case, by index, of Gemma's block-selection word sets, fixed before any output was inspected: a receiver asked about its first fact (*apple crumble*) takes the state of a donor asked about its second (*baked trout*), and answers *pear tart*, its own second fact, at block 22, but *baked apple crumble*, the donor's first word followed by its own content, at block 33.

Scoring donor outputs. Each donor output is scored three ways: it names the receiver's designated content, it is the donor's own complete answer, or it begins with the first token of the donor's answer. Past the window most outputs begin with the donor's first word and continue with receiver content, as in *baked apple crumble*; such outputs name no content of either context and count as invalid. First-token agreement therefore measures when the donor's content takes over, and Figure 5 reports it alongside complete answers. Complete donor answers remain rare at every block, at most 2.1% in Qwen, 9.8% in Gemma, and 4.1% in Llama.

Answer alignment. For the alignment curves of Figure 13, the output direction of an answer in a context is the final-norm scale times the unembedding row of the answer's first token, and the four directions of a context are centered. Alignment is the cosine between each query-dependent state and its own answer's direction, averaged over test contexts; it uses only the unembedding and final-norm weights. It is near zero at the main blocks (0.001, 0.017, and 0.004) and reaches 0.26, 0.12, and 0.25 at the final block. The fact address is thus a shared difference only inside the window; past it, whole states still carry it, and then the donor's content takes over.

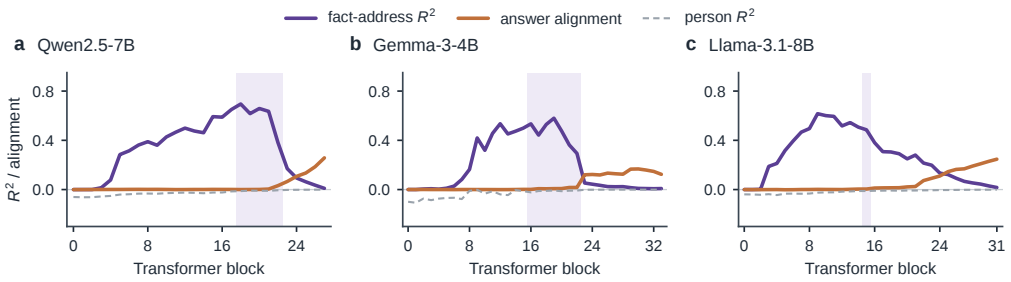

Figure 13: **The fact-address geometry follows the stages of Figure 5.** Query-state geometry at every block: R^2 of fact-address and person templates (leave-one-word-set-out over the 16 calibration word sets, 256 contexts), and each query state's alignment with its answer's first-token output direction (128 test contexts). Shading: the depth window of Figure 5.

F.2 Pretraining Checkpoints

In pretraining, questions switch from the named person to order of mention before the model can use the address, and the depth profile settles last (Section 7). Figure 6 repeats the tests on ten checkpoints of OLMo-2-7B (Team OLMo et al., 2024): stage-1 steps 1k, 4k, 10k, 25k, 50k, 99k, 200k, 400k, and 929k (5B to 3.9T training tokens) and the released model after stage-2 mid-training. The prompt is the plain format of Appendix G.2 with the tokenizer's `<|endoftext|>` as the first token, which marks a document start in OLMo-2's packed pretraining data. Each checkpoint runs the protocol of Appendix G.1 with the depth curves: its own rule-chosen block, the transfer test, the cosines, the query-state geometry at every block, and the depth curves at every second block (8 donors $\times$ 32 receivers). Transfer and cosines are read at block 16, the released model's rule-chosen block; checkpoints whose rule-chosen block differs were run again at block 16. Early checkpoints answer in sentences, so answers are scored by the single stated content they name. A base checkpoint can name the designated answer without any edit (up to 14% of cases), so the stages of Figure 6a apply the thresholds of Figure 5a to the gain over the unedited answer.

Table 4: **OLMo-2-7B checkpoints.** Rule-chosen block; at block 16: unedited accuracy and designated answers (%) for the ordinal vector (bold), a random direction and the ceiling, scored by the single stated content an answer names, 64 test word sets and 12,288 cases per edit; new-name and reversed-order cosines of Proposition 3.4; and the peak leave-one-word-set-out fact-address R^2 over blocks (its block). *final*: the released model after stage-2 mid-training.

		At block 16						
		Scored by stated content (%)				Cosine		
Tokens	Rule block	Unedited	v	Random	Ceiling	New names	Reversed	Peak R^2 (block)
5B	22	1.3	**1.2**	1.2	1.2	0.09	0.98	0.00 (31)
17B	22	24.9	**16.3**	16.0	16.7	0.37	0.25	0.29 (11)
42B	14	50.0	**25.1**	14.6	31.5	0.70	−0.61	0.52 (13)
105B	22	63.9	**23.7**	11.3	27.6	0.76	−0.74	0.61 (13)
210B	16	67.4	**50.1**	9.3	56.8	0.79	−0.74	0.74 (13)
416B	16	79.9	**46.0**	6.3	55.9	0.80	−0.71	0.71 (13)
839B	16	86.1	**21.7**	4.6	29.7	0.75	−0.56	0.62 (13)
1.7T	16	91.5	**23.5**	2.0	35.9	0.80	−0.65	0.69 (14)
3.9T	22	89.2	**14.6**	2.5	24.3	0.77	−0.59	0.66 (13)
final	16	92.4	**33.1**	1.8	61.8	0.85	−0.78	0.71 (14)

Table 4 gives the numbers behind Figure 6b,c. At 5B tokens the reversed-order cosine is +0.98: the change for a pair of people does not depend on their order, as expected if questions were represented by the named people. It falls to 0.25 at 17B and −0.61 at 42B. Fact-address geometry precedes the ability to answer: at 17B tokens the model names the right content in 24.9% of cases, the rate of guessing among four, yet fact-address templates already explain 0.29 of the variance.

The depth profile moves throughout training (Figure 6a). Shared-address blocks appear at 42B, vanish at 105B, and are present from 210B through 1.7T; content blocks appear from 416B. All three stages first co-occur at 839B and again at 1.7T, each time separated by blocks that reach no threshold. At 3.9T whole-state and content blocks remain, but no block reaches the shared-address threshold, and only the released model shows all three stages as one contiguous profile. Read at the fixed block 16, transfer therefore peaks at 210B (50.1% against a ceiling of 56.8%) and falls later as the profile moves: the ceiling at block 16 drops from 55.9% at 416B to 24.3–35.9% from 839B to 3.9T and returns to 61.8% after mid-training. Questions are organized by order of mention early in pretraining; the depth profile settles only late.

G Generality: Model Scale, Base Weights, and Prose

The main text reports its tests on three instruction-tuned models and templated lists, and states that they hold from 1.5B to 32B parameters (Table 5), in base weights (Figure 2d), and in short paragraphs (Section 5); this section gives those tests on eleven further models and on prose: the instruction-tuned Qwen2.5-1.5B, Qwen2.5-3B, Qwen2.5-14B, Qwen2.5-32B, Qwen3-8B, and Llama-3.2-3B, and the base weights of Qwen2.5-1.5B, Qwen2.5-3B, Qwen3-8B, Gemma-3-4B, and Llama-3.2-3B. Materials, calibration and test word sets, and scoring follow Appendix A except where noted, and every rate is a share of complete greedy answers.

G.1 Model Scale

Ordinal addressing holds at every model size tested (Sections 5 and 7). We run Qwen2.5-Instruct at 1.5B, 3B, 14B, and 32B parameters and Qwen3-8B with its thinking mode switched off. Each model's main block follows the rule of Appendix A, and the 7B model keeps its block 20. Each model then runs the transfer test of Table 1A, the cosines of Figure 2d, the templates of Appendix C.1, and the depth curves of Figure 5 with 8 donors and 32 receivers.

Every model addresses facts by order of mention (Table 5). New names keep the direction of the ordinal vector and reversed orders reverse it, with cosines of 0.92 to 0.97 and -0.85 to -0.96; at the main block, the fact-address template explains 0.46–0.76 of the query-dependent variance (leave-one-word-set-out) and person none. The ordinal vector selects the designated fact in 34.6–84.1% of cases and random directions almost never. Its rate does not rise with size, and neither does the ceiling: the ceiling selects 65.2–96.9% of designated facts, lowest in the 3B and 14B models; in the 3B model the chat template lowers it (Appendix G.2). Measured against the ceiling, the ordinal vector reaches 68–87% of the ceiling in Qwen2.5 and 46% in Qwen3-8B. In every model the ordinal vector acts within a window of a few blocks at 63–86% of relative depth (block divided by the last block; the window is where it selects at least 10%, as in Figure 5), and whole donor states outlast it: two blocks past the window they still select 7–30% of receiver facts while the ordinal vector selects under 5%, as in Section 7.

G.2 Base and Instruction-Tuned Weights

Instruction tuning is not needed for ordinal addressing (Section 3). We compare five base models with their instruction-tuned partners on one plain prompt, the instruction, facts, and question followed by `Answer:`, and add each instruct model in its native chat template. Each row chooses its own main block by the rule above.

In Qwen the base weights carry the address about as strongly as the tuned ones; in Llama the base model is weaker (Table 6). On the identical prompt, the ordinal vector selects 67.3% against 69.6% in Qwen2.5-1.5B, 80.0% against 72.2% in Qwen2.5-3B, 31.6% against 33.3% in Qwen3-8B, and 41.9% against 54.8% in Llama-3.2-3B. The four Qwen and Llama base models keep a new-name cosine of at least 0.86 and a reversed-order cosine of at most -0.74, with fact-address R^2 between 0.48 and 0.78. Gemma-3-4B's base model answers in full sentences, so we score the single stated content each answer names: it answers 92.9% of unedited questions correctly, and the ordinal vector selects the designated fact in 21.9% of cases against 0.6% for random directions, below its instruct partner; its new-name and reversed-order cosines, 0.76 and -0.44, are also the weakest, while its fact-address R^2, 0.49, equals that partner's on the same prompt.

Table 5: **Ordinal addressing holds from 1.5B to 32B parameters.** Every model runs the transfer test of Table 1A (64 new word sets, 12,288 cases), the cosines of Figure 2d, the fact-address and person templates of Appendix C.1, and the depth curves of Figure 5 at every second block. Main block: block index / number of blocks, chosen on block-selection word sets by the rule of Appendix A. Unedited: correct answers without an edit. Window: first to last block at which the ordinal vector selects at least 10%, as in Figure 5 (8 donors × 32 receivers). †The main-text model; its block was fixed in earlier runs and its window comes from the every-block curves of Figure 5.

		Designated answers (%)				Change cosine		R^2		
Model	Main block	Unedited	v	Random	Ceiling	New names	Reversed	Fact address	Person	Window
Qwen2.5-1.5B	18/28	98.6	**72.0**	0.1	95.7	0.97	−0.96	0.76	−0.02	18–22
Qwen2.5-3B	26/36	96.6	**48.0**	0.0	66.1	0.92	−0.85	0.65	−0.02	26–30
Qwen2.5-7B†	20/28	99.9	**84.1**	0.0	96.9	0.93	−0.93	0.66	−0.01	18–22
Qwen2.5-14B	34/48	99.9	**44.4**	0.0	65.2	0.92	−0.92	0.46	−0.01	32–34
Qwen2.5-32B	48/64	100.0	**55.1**	0.0	74.7	0.94	−0.94	0.51	−0.01	46–50
Qwen3-8B	22/36	99.6	**34.6**	0.0	74.8	0.95	−0.93	0.71	−0.01	22–28

Within Qwen, the one large gap comes from the chat template, not the tuning: Qwen2.5-3B-Instruct selects 72.2% with the plain prompt and 48.0% in its chat template, where the ceiling also falls from 90.3% to 66.1%. Instruction tuning is not needed for ordinal addressing; the prompt format can change how strongly the ordinal vector acts on it.

Table 6: **Base models address facts the same way.** Plain: the instruction, facts, and question followed by `Answer:`, with no chat template, identical for base and instruct weights; chat: the instruct model's native template. Each row chooses its own main block by the rule of Appendix A and runs the tests of Table 5. *Gemma-3-4B base answers in sentences, so its answers are rescored by the single stated content they name. ‡Main-text Gemma run (Table 1A, Figure 2d, Appendix C.1).

			Designated answers (%)				Change cosine		
Model	Weights, prompt	Block	Unedited	v	Random	Ceiling	New names	Reversed	Fact-address R^2
Qwen2.5-1.5B	base, plain	18	87.7	**67.3**	1.7	82.0	0.97	−0.95	0.78
	instruct, plain	20	98.7	**69.6**	0.1	95.3	0.95	−0.94	0.63
	instruct, chat	18	98.6	**72.0**	0.1	95.7	0.97	−0.96	0.76
Qwen2.5-3B	base, plain	26	95.5	**80.0**	0.1	94.1	0.94	−0.87	0.69
	instruct, plain	26	95.7	**72.2**	0.0	90.3	0.93	−0.88	0.65
	instruct, chat	26	96.6	**48.0**	0.0	66.1	0.92	−0.85	0.65
Qwen3-8B	base, plain	22	96.9	**31.6**	0.0	71.7	0.95	−0.93	0.64
	instruct, plain	22	98.5	**33.3**	0.0	74.6	0.92	−0.89	0.64
	instruct, chat	22	99.6	**34.6**	0.0	74.8	0.95	−0.93	0.71
Gemma-3-4B	base, plain	22	6.7	**14.0**	0.1	27.6	0.76	−0.44	0.49
	by content*		92.9	**21.9**	0.6	52.6			
	instruct, plain	20	94.4	**47.7**	0.0	85.7	0.93	−0.88	0.49
	instruct, chat‡	22	99.9	**53.9**	0.0	96.4	0.88	−0.82	0.29
Llama-3.2-3B	base, plain	14	88.1	**41.9**	0.4	85.4	0.86	−0.74	0.48
	instruct, plain	14	95.5	**54.8**	0.0	94.0	0.89	−0.82	0.48
	instruct, chat	14	97.4	**51.5**	0.0	93.9	0.91	−0.85	0.51

G.3 Facts Stated in Prose

Fact addresses do not depend on the list format (Section 5). Each prose context is one paragraph stating the four facts with four different verb frames drawn from sixteen, such as *For lunch, Petra had the buckwheat* or *The teff ended up in front of Sabrina*; none uses *eats*, and four frames put the content first. Two or three unrelated sentences, such as *It had rained earlier that morning, and the*

streets were still wet, sit between facts, and an opening or closing narrative sentence is added with probability one half. A fact's order of mention is its place in the paragraph. Paragraphs run 37–87 words, against 20–30 for the list, and the instruction and question are those of the main text. Test word sets are 64 new word sets whose names and contents appear in no main-text word set; each is also rendered as a list with the same people, contents, and questions.

Table 7: **Ordinal addressing carries over to prose.** Top: designated answers (%) on 64 new word sets, 12,288 cases per entry, each edit added once at the main block. Receivers state the four facts either as a paragraph (*prose*) or as the list of the main text, with the same people, contents, arrangements, and questions. The ordinal vectors are calibrated on the main-text lists (the v of Table 1A), on lists of 16 new word sets, or on paragraphs of the same 16 word sets. Bottom: at the main block, leave-one-word-set-out R^2 of fact-address and person templates on paragraphs and on the matched lists (the list R^2 uses this experiment's word sets, not those of Appendix C.1); mean cosine between matching prose and list fact-address templates and between matching prose and list ordinal vectors; fraction of the squared norm of query-dependent prose states inside the span of the main-text ordinal vectors, against random subspaces of the same dimension.

			Ordinal vector calibrated on				
Model	Receiver format	Unedited	Lists (main)	Lists (new word sets)	Prose	Random	Ceiling
Qwen	prose	99.7	**44.5** ±1.8	44.0 ±1.8	40.2 ±1.5	0.0	92.8 ±1.6
	list	99.9	**80.5** ±2.1	78.1 ±2.2	41.1 ±1.7	0.0	95.4 ±1.5
Gemma	prose	99.2	**42.4** ±1.2	43.4 ±1.2	32.3 ±1.2	0.0	96.3 ±0.9
	list	99.7	**52.2** ±1.4	52.8 ±1.4	24.1 ±1.2	0.0	96.5 ±0.9
Llama	prose	99.6	**14.0** ±1.1	10.2 ±0.9	8.5 ±0.9	0.0	83.1 ±1.5
	list	99.8	**29.3** ±1.9	22.4 ±1.7	12.2 ±1.1	0.0	85.6 ±1.7

	Prose R^2		List R^2	Prose vs. list		Squared norm in list span (%)	
Model	Fact address	Person	Fact address	Template cos.	Vector cos.	Prose states	Random
Qwen	0.45	−0.01	0.71	0.76	0.75	35	0.1
Gemma	0.23	−0.01	0.39	0.81	0.80	17	0.1
Llama	0.33	−0.01	0.53	0.84	0.81	34	0.1

The ordinal vector calibrated on the main-text lists, unchanged, selects the designated fact in paragraphs (Table 7): 44.5%, 42.4%, and 14.0% of cases in Qwen, Gemma, and Llama, against 80.5%, 52.2%, and 29.3% on the matched lists, while random directions select none. It stays above random at every designated order of mention (lowest 22.9%, 15.0%, and 3.6%), and the ceiling still reaches 92.8%, 96.3%, and 83.1%. To separate format from word sets, 16 new word sets are rendered both as paragraphs and as lists for calibration. On prose receivers, the ordinal vector calibrated on their paragraphs selects fewer designated facts than the one calibrated on their lists (40.2% against 44.0% in Qwen, 32.3% against 43.4% in Gemma, 8.5% against 10.2% in Llama), and on list receivers the paragraph-calibrated vector still selects 41.1%, 24.1%, and 12.2%.

The geometry agrees. In paragraphs, the fact-address template explains 0.45, 0.23, and 0.33 of the query-dependent variance at the main block (leave-one-word-set-out) and person none. Fact-address templates from prose and from lists point the same way (cosines 0.76, 0.81, and 0.84), as do their ordinal vectors (0.75, 0.80, and 0.81), and 35%, 17%, and 34% of the squared norm of query-dependent prose states lies in the three-dimensional span of the main-text ordinal vectors, against 0.1% for random subspaces. Paragraphs and lists share one set of fact addresses.