# PROMPT STRUCTURE REORGANIZES IN-CONTEXT SEQUENTIAL MEMORY IN LLMS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Large Language Models (LLMs) are sensitive to where information appears in their context (Liu et al., 2024). Research on human sequential memory shows that position is not a single quantity. Beyond its absolute rank in the whole sequence, an item in a grouped sequence carries a global rank, the position of its group, and a local rank, its position within the group, which are represented along separable neural dimensions (Fan et al., 2025). Existing evaluations of LLMs ask what appears at a position, not where an item is, so whether LLMs hold several positional descriptions of an item, and whether the prompt decides which one is accessible, remains unknown. Here we show, across 5 LLMs probed with controlled English and Chinese sequences of familiar words, that structural cues reorganize positional retrieval. Stating that the sequence consists of familiar words mainly improves reports of a target's group and within-group position. Making group boundaries visible improves these reports further, but often lowers exact absolute-position accuracy and redirects absolute-position errors toward the displayed grouping. These boundary effects persist after the familiar-word structure is destroyed and when group lengths vary, where exact absolute reports can collapse while reported positions remain strongly rank-correlated with true positions. Retrieval thus depends on the match between the organization a cue emphasizes and the coordinate a query demands, paralleling hierarchical working memory in humans, although in LLMs this organization is largely cue-dependent. Prompts therefore act as organizational interventions on in-context sequential memory, and memory systems should be evaluated across the positional readouts they support and for interference between them.

## 1 INTRODUCTION

Retrieval in Large Language Models (LLMs) depends on context length and on where the relevant information appears (Liu et al., 2024; Bai et al., 2024; Guo & Vosoughi, 2025; Yi et al., 2026), with primacy, recency, and contiguity effects in serial recall (Mistry et al., 2025). These evaluations treat position as a route to content and rarely ask whether a model can report where a known item occurred. In a structured sequence, that question has no single answer, because an item has both an absolute position in the whole sequence and a position relative to its group, so the same sequence can be read as a 1-D chain or as a 2-D hierarchy (Fig. 1a).

Human sequential memory uses both organizations. Grouping reshapes order errors and cross-group transpositions preserve within-group position (Ryan, 1969; Ng & Maybery, 2002; Hartley et al., 2016), consistent with multilevel accounts of serial position (Lee & Estes, 1981; Henson, 1998; Hurlstone, 2019) in which familiar units supply structure through chunking (Thalmann et al., 2019; Norris & Kalm, 2021). Fan et al. (2025) show separable neural dimensions for word position and within-word position in syllable sequences organized into words (Fig. 1b). This organization arises spontaneously and helps reports of word and within-word position, and it also leaves a mark on reports of continuous rank, which are faster for the first and last syllables of a word than for the middle one.

Whether LLMs hold positional information in more than one form, and whether the prompt decides which form a report draws on, is unknown. Because the sequence stays in context while the model answers, the question concerns how position is read out from a context that still contains it, which



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

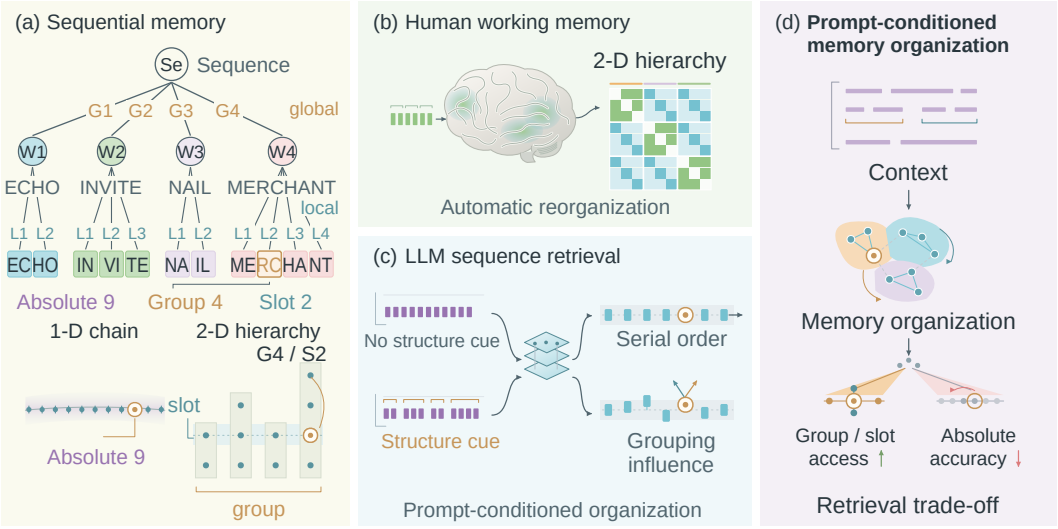



Figure 1: **Structural cues change how positional information is organized in sequential memory.** (a) The tree shows a sequence with word structure, so the item RC's position can be described by its absolute position 9, or group position 4 and within-group slot position 2. Thus the sequence can be organized both by a 1-D chain and a 2-D hierarchy. (b) In human WM, syllable sequences organized into words exhibit separable neural dimensions for global rank (word position) and local rank (syllable position) (Fan et al., 2025). (c) Unlike the spontaneous organization in humans, LLMs show a weaker grouping pattern without a structure cue, while adding a structure cue makes grouping influence emerge. (d) Structure-relative accuracy generally improves as structural cues strengthen, while exact absolute-position accuracy often decreases. This trade-off implies that prompt structure changes how LLMs organize information in sequential memory.

we call in-context sequential memory. The prompt is the natural lever on this readout, since it can make a structure salient without changing the sequence. If a cue reorganizes retrieval as grouping does in humans, the descriptions compatible with the emphasized structure should become more accessible, while the incompatible description should become less exact and its errors should take the shape of that structure.

We test this with controlled English (EN) and Chinese (ZH) sequences of familiar words split into displayed items (Fig. 1a). Familiar words carry latent group structure without a visual marker, so lexical and visual cues can be added separately, and the latent structure can be destroyed while the visual cue remains. For the same target, we probe absolute, group, and within-group slot position under no cue, a familiar-word premise, visible boundaries, and both, with structure-destroyed controls and variable group lengths.

Structural cues produce exactly this pattern. Stating the familiar-word structure mainly improves group and slot reports, while absolute-position accuracy does not improve and usually decreases. Visible boundaries produce a stronger trade-off, improving structure-relative reports further while lowering exact absolute-position accuracy and redirecting its errors toward the displayed grouping (Fig. 1c,d), and the effect persists after the group structure is destroyed. With variable group lengths, exact absolute-position accuracy can collapse while reported positions stay strongly rank-correlated with true positions, and group and slot positions measurably shape absolute errors, in a way that a fixed-size conversion between coordinates does not explain.

Prompt structure therefore reorganizes in-context sequential memory. Which positional description a model retrieves well depends on the match between the organization a cue makes salient and the coordinate a query demands, a compatibility effect that parallels hierarchical working memory (WM) in humans, except that in LLMs the organization is largely cue-dependent rather than spontaneous.

## 2 Related Work

**Hierarchical organization in human sequential memory.** Human sequential memory shows systematic serial-position effects (Murdock Jr, 1962; Glanzer & Cunitz, 1966), and positional accounts of order are distinguished from chaining and other sequence representations (Henson, 1998; Hurlstone et al., 2014). Grouping adds a second positional level. Temporal grouping changes recall

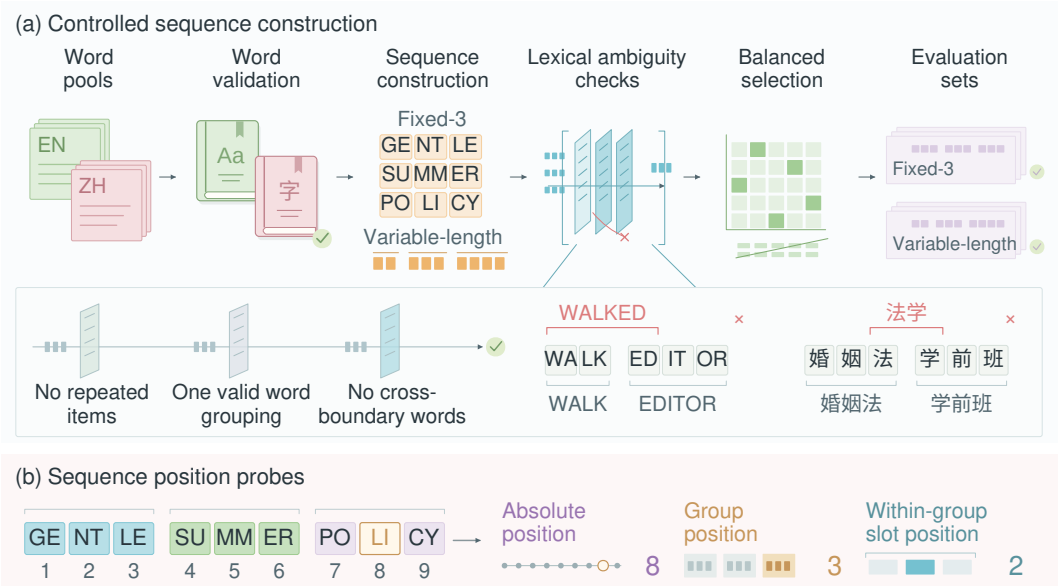



Figure 2: **Test-sequence construction and positional probes.** (a) Familiar source words are split into items, non-overlapping two-letter segments in EN or individual characters in ZH. Consecutive items from one source word form a group. Validated words form candidate sequences with 3 items per group (fixed-3) or 2, 3, or 4 (variable-length). Candidates are rejected for repeated items, alternative complete multi-item word groupings, or valid multi-item words crossing group boundaries, such as `WALKED` or "法学". Word reuse is balanced across group positions before selecting test sets. (b) Each item is probed by 3 position types independently: absolute position counts items across the full sequence, group position counts source-word groups, and within-group slot position counts items within the target group, all from 1 on the left. In `GE NT LE SU MM ER PO LI CY`, `LI` has absolute position 8, group position 3, and within-group slot position 2.

accuracy and the distribution of order errors (Ryan, 1969; Hitch, 1996; Ng & Maybery, 2002; Hartley et al., 2016), and grouped probed-recall paradigms separate the group containing a target from its position within that group (McNicol & Heathcote, 1986; Brown et al., 2009), so whole-sequence and group-relative position are distinct behavioral descriptions of the same order. Fan et al. (2025) show that this hierarchy is neurally explicit, with global and local ranks expressed along separable dimensions of representational geometry across word lengths and reporting settings. We adopt their ZH word banks and their distinction between global, local, and continuous rank as the coordinates probed here.

**Positional and structured retrieval in LLMs.** Long-context evaluations show that what LLMs recover depends on context length and on where the relevant evidence sits (Liu et al., 2024; Bai et al., 2024; Hsieh et al., 2024; Yi et al., 2026), and studies adapting human memory paradigms find serial-recall effects (Hajipoor et al., 2019), serial-position effects (Guo & Vosoughi, 2025), working-memory-like capacity (Gong et al., 2024; Hong et al., 2025), temporal contiguity in in-context serial recall (Mistry et al., 2025), and degradation as conversational context accumulates (Laban et al., 2026). In all of these, position is the condition and content is the report; reporting several positional descriptions of the same target in a grouped sequence remains less explored. A second line shows that imposed structure changes retrieval. Meaning-preserving formatting choices shift LLM performance (Sclar et al., 2024), indexing and reading strategies shape conversational recall (Wu et al., 2025), and hierarchical memory systems organize context through memory graphs or discourse structure (Tang et al., 2026; Chen et al., 2026). These systems treat structure as a design choice for reaching content, whereas we treat it as an intervention on how position itself is read out.

## 3 Experimental Setup

### 3.1 Task and Materials

We construct EN and ZH sequences from familiar-word pools (Fig. 2a), length-stratified word lists for EN and the word banks of Fan et al. (2025) for ZH. A familiar word, or *source word*, is split into

Table 1: **Example prompts and sequence renderings for the test conditions.** English examples are shown, with each two-letter segment representing one displayed item. A larger gap serves as the visible boundary between adjacent groups and is not counted as an item. L0 and L1 differ in whether the familiar-word structure is stated before the sequence, while L2 and L3 add the visible boundary to their corresponding conditions. S0 and S2 shuffle the same items to destroy the familiar-word structure, without and with visible boundaries, respectively. Chinese prompts make the same distinctions.

| Condition | Short prompt cue | Sequence example |
|---|---|---|
| L0 | "You will see an ordered sequence of chunks." | `GE NT LE SU MM ER` |
| L1 | "The sequence is formed by several familiar English words placed one after another." | `GE NT LE SU MM ER` |
| L2 | "A larger gap marks a boundary in the sequence." | `GE NT LE  SU MM ER` |
| L3 | "… familiar English words … A larger gap marks a boundary between adjacent words." | `GE NT LE  SU MM ER` |
| S0 | Same prompt cue as L0. | `SU GE MM NT ER LE` |
| S2 | Same prompt cue as L2. | `SU GE MM  NT ER LE` |

*items*, non-overlapping two-letter segments in EN or single characters in ZH, and its consecutive items form a *group*. Fixed-3 materials use 6-letter EN or 3-character ZH words, so every group has 3 items; variable-length materials mix 4-, 6-, and 8-letter EN or 2-, 3-, and 4-character ZH words, giving groups of 2, 3, or 4 items. Pools are validated against independent lexical resources, Open English WordNet 2025 (McCrae et al., 2020) and SCOWL 2020.12.07 (Atkinson, 2020) for EN and CC-CEDICT (CC-CEDICT Contributors, 2026) for ZH, and a candidate sequence is rejected if it repeats an item, admits an alternative complete segmentation of a source word, or contains a valid multi-item word that crosses an intended boundary. Word reuse is balanced across group positions during final material construction (Section C.2).

Each trial presents an ordered sequence, a target item that occurs exactly once, and one of 3 queries (Fig. 2b). *Absolute position* counts items from the left across the whole sequence, *group position* counts familiar-word groups from the left, and *within-group slot position* counts items from the left within the target's group (Section C.1).

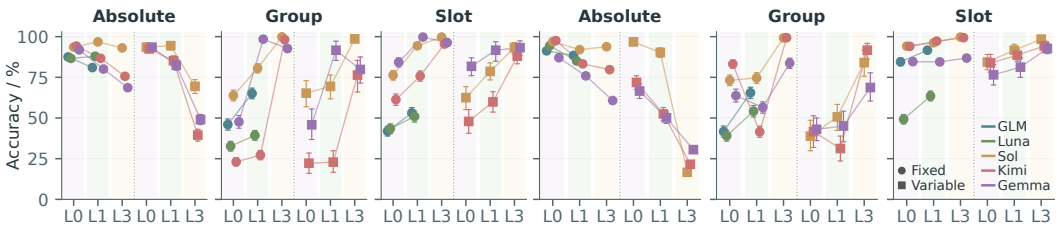
### 3.2 Test Conditions

We manipulate 2 cues, a lexical cue (a familiar-word premise stated before the sequence) and a boundary cue (visibly marked group boundaries, rendered as larger spacing), giving 4 conditions (Table 1). L0 provides neither cue, L1 the premise only, L2 the boundaries only, and L3 both. The premise precedes the sequence because, under causal attention, a cue placed before the sequence can condition the sequence-token representations, whereas word information that appears only in the probe cannot alter earlier token states. S0 and S2 destroy the familiar-word structure by reordering the same items so that no valid multi-item word remains; S0 keeps the unmarked format of L0, and S2 adds the boundary cue of L2. Full templates, demonstrations, and request examples are in Section C.1.

### 3.3 Models and Evaluation

We evaluate GLM-5.2 (Zeng et al., 2026), GPT-5.6 Luna and GPT-5.6 Sol (OpenAI, 2026), Kimi-K3 (Team et al., 2026b), and Gemma-4-31B-IT (Team et al., 2026a) in EN and ZH. Each model–language pair is first screened at $N = 9$, the shortest fixed-3 sequence in Fan et al. (2025), for parseable responses, accurate absolute-position retrieval without cues, and accurate absolute, group, and slot reports when the grouping is explicit. For each eligible pair we then select a sequence length $N$ that remains reliably retrievable while yielding enough errors for analysis, using only task-capability and error-observability criteria (absolute-position accuracy, valid-error yield, and stability across base sequences) that are independent of the structure-sensitive metrics reported below. The selected load is held fixed across matched conditions (Sections D.1 and D.3).

Responses are parsed for the requested integer and scored by exact match (Section D.5). Decoding is deterministic at temperature 0, with reasoning or thinking disabled where configurable, so the assay probes direct positional readout rather than explicit multi-step counting (Section D.1). Matched comparisons reuse the same base sequences and target schedules, accuracy changes are reported in percentage points (pp), and unless stated otherwise 95% confidence intervals (CIs) come from




(a) EN (b) ZH

Figure 3: **Accuracy for the 3 position probes across structural conditions in fixed-3 and variable-length materials.** The left and right subfigures show EN and ZH, with separate panels for absolute, group, and within-group slot accuracy. Colors identify models, circles denote fixed-3 materials, and squares denote variable-length materials. Whiskers show 95% CIs. Across model–language pairs, group and slot accuracy generally increase from L0 to L1 and further with L3, whereas absolute-position accuracy usually changes little or decreases and shows the largest declines under L3.

5000 block-bootstrap resamples over base sequences, with all observations from a base resampled together (Sections D.4 and E.1).

## 4 Lexical Cues Reshape Positional Retrieval

We compare L0 and L1 across all 5 models in EN and ZH. The displayed sequence, target, and probe are matched across conditions, so the only intervention is that L1 states the familiar-word structure in advance. Each condition uses 256 base sequences per model–language pair, giving 4608–6912 absolute-position, 768 group-position, and 768 within-group slot requests per condition (Section D.4). Exact-match accuracy is shown in Fig. 3, and the L0–L1 pattern is stable across sequence lengths for the 2 models tested (Section G.2).

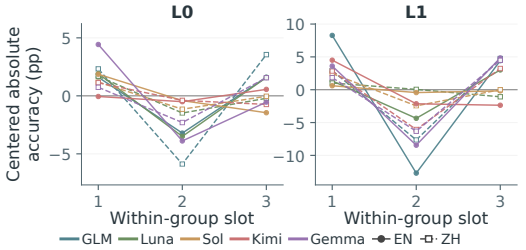



Figure 4: **Absolute-position accuracy across within-group slots under L0 and L1 in fixed-3 materials.** The horizontal axis gives within-group slot, and the vertical axis gives absolute-position accuracy minus its 3-slot mean for each condition. Colors denote models. Solid lines with filled circles denote EN, and dashed lines with open squares denote ZH. Most profiles are lower at the middle slot than at the 2 edges, and this slot-dependent variation is generally stronger under L1. Because the L1 vertical range is approximately twice that of L0, similar visual amplitudes understate the stronger slot dependence under L1.

**Lexical cues improve structure-relative accuracy.** From L0 to L1, group-position accuracy increases in 8/10 model–language pairs and within-group slot accuracy in 9/10 (Section F.1); for Gemma-EN, group accuracy rises from 47.66% to 98.44% and slot accuracy from 84.24% to 99.74%. The 2 group-position exceptions, Kimi-ZH (83.20% to 41.54%) and Gemma-ZH (63.80% to 56.51%), show a decline concentrated at later groups and biased toward earlier group indices, while their slot accuracy improves (Kimi-ZH) or is essentially unchanged (Gemma-ZH, 84.64% to 84.51%, the only slot decrease). Absolute-position accuracy decreases in 8/10 pairs, for example from 97.57% to 83.28% for Kimi-ZH, and the 2 gains are modest (Luna-EN +1.31 pp, Sol-EN +3.21 pp). Because capability matching places L0 absolute accuracy near ceiling while leaving far more headroom for group and slot accuracy, the opposing directions partly reflect this asymmetry, so we treat the accuracy contrast as descriptive and rest the argument on the large effects and the error-structure analyses below. The lexical cue thus changes which positional description is accessible, making group and slot positions easier to report while leaving the absolute position no easier.

**Lexical cues often amplify the local U-shaped profile of absolute-position accuracy.** In human continuous-rank reports, the first and last syllables of a word are reported faster than the middle one (Fan et al., 2025), and model absolute-position accuracy shows an analogous profile. As shown in Fig. 4, accuracy is generally lower at the middle slot than at the 2 edge slots of a group, and this local U-shaped profile is visibly more pronounced under L1 than under L0. To quantify it, let $A_s$ denote exact-match absolute-position accuracy for targets at within-group slot $s$, and define $U^{\text{abs}}_{\text{local}} = (A_1 + A_3)/2 - A_2$, which is positive when the edge slots are more accurate than the

middle slot. In the fixed-3 materials, $U^{\text{abs}}_{\text{local}}$ is positive in all 10 L0 pairs and in 9/10 L1 pairs, and it increases from L0 to L1 in 9/10 pairs. Matched 95% CIs for $\Delta U^{\text{abs}}_{\text{local}}$ lie above zero in 7/10 pairs, include zero for Luna-EN and Sol-EN, and lie below zero for Luna-ZH; the largest increases are GLM-EN (+14.23 pp [11.65, 16.77]), Kimi-ZH (+8.51 pp [6.86, 10.16]), and Gemma-EN (+6.84 pp [4.94, 8.76]). Because this raw index is an accuracy-scale contrast whose magnitude varies with overall error headroom, we also normalize it by the mean error rate; the normalized increase remains interval-positive in only 3/10 pairs, so we treat the profile as descriptive (Section F.2).

Lexical cues therefore mainly increase access to structure-relative positions, and the within-group profile of absolute-position accuracy suggests that the same organization begins to shape the absolute readout. We next test whether making the grouping visible produces a stronger and more systematic change.

## 5 Visible Boundaries Induce a Retrieval Trade-Off

**Visible boundaries improve structure-relative accuracy while reducing absolute-position accuracy.** The boundary interventions use GPT-5.6 Sol, Kimi-K3, and Gemma-4-31B-IT in EN and ZH, the models with the most stable behavior across boundary renderings in preliminary checks (Section D.2), with the sampling schedule and matched targets of Section 4. L3 adds visible boundaries at the group boundaries to the premise and sequence of L1. From L1 to L3, exact absolute-position accuracy decreases in 5/6 pairs, whereas group-position and within-group slot accuracy each increase in 5/6 (Fig. 3); for Kimi-EN, absolute accuracy falls from 86.68% to 75.63% while group accuracy rises from 27.21% to 98.18% and slot accuracy from 75.78% to 95.44% (Section F.1). This trade-off is the central result, as the cue that eases group and slot reports is the one that impairs absolute reports. To test whether it requires the familiar-word premise, L2 adds the same boundaries to L0 without stating that the sequence contains words. Absolute-position accuracy decreases by 2.6–27.6 pp, while group-position accuracy increases by 17.2–64.1 pp and within-group slot accuracy by 3.6–31.8 pp (Fig. 5). The same opposing change therefore emerges without any lexical description, so visible grouping alone produces the trade-off.

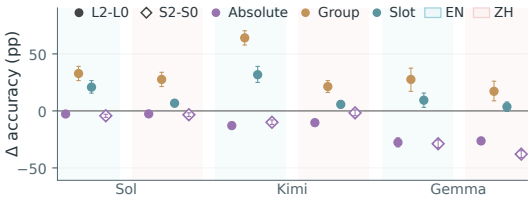
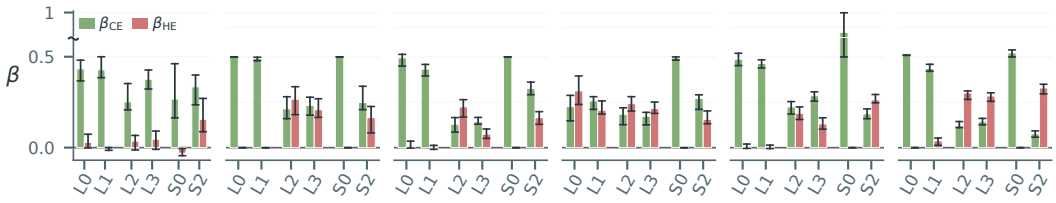



Figure 5: **Accuracy changes after boundary addition in fixed-3 materials.** Filled circles show L2 minus L0 for absolute, group, and slot probes, while open diamonds show S2 minus S0 for absolute-position probes after the familiar-word structure is destroyed. The horizontal axis identifies model–language pairs, pale bands separate EN and ZH, and the vertical axis gives accuracy change in pp. Colors denote probe type, zero marks no change, and whiskers show 95% matched CIs. For L2 minus L0, absolute-position point estimates are all below, whereas group and slot point estimates are all above zero. With the structure destroyed, all S2−S0 absolute-position point estimates also remain below zero. Visible boundaries generally favor structure-relative accuracy even without word structure.

**Visible boundaries increase period-3 adjacent-group absolute errors.** Accuracy alone cannot show whether the remaining absolute errors are shaped by the grouping, so we analyze their structure, adapting the distinction between continuous error (CE) and hierarchical error (HE) from Fan et al. (2025). For each true position $i$, we normalize valid in-range incorrect reports over reported positions $j \neq i$, $p_{ij} = P(\hat{i} = j \mid i, \hat{i} \neq i, 1 \leqslant \hat{i} \leqslant N)$, and regress these off-diagonal probabilities jointly on a CE indicator ($|i - j| = 1$) and an HE indicator ($i$ and $j$ in the same within-group slot of adjacent groups, a displacement of $\pm 3$ in fixed-3 materials). $\beta_{\text{CE}}$ and $\beta_{\text{HE}}$ thus measure the additional normalized error probability associated with adjacent positions and with the period-3 adjacent-group displacement, respectively. From L0 to L1, $\beta_{\text{HE}}$ decreases or stays near zero in 5/6 pairs (Fig. 6), so the lexical cue alone does not redirect absolute errors. From L1 to L3, $\beta_{\text{HE}}$ increases in all 6 pairs, from $-0.001$ to $0.075$ for Kimi-EN and from $0.038$ to $0.285$ for Gemma-ZH, and from L0 to L2, without the premise, in 5/6 pairs, including $0.000$ to $0.267$ for Sol-ZH; the exception is Kimi-ZH, whose $\beta_{\text{HE}}$ is already $0.313$ under L0 and $0.243$ under L2, consistent with spontaneous sensitivity to the latent word organization. For Kimi-EN and Gemma-EN, reported positions also drift slightly downward as more boundaries precede the target, which may contribute to the repeated $\pm 3$ pattern (Section F.4). Visible boundaries therefore strengthen the period-3 adjacent-




(a) Sol EN (b) Sol ZH (c) Kimi EN (d) Kimi ZH (e) Gemma EN (f) Gemma ZH

Figure 6: **Fixed-3 absolute-position error coefficients across structural conditions.** Each panel shows one model–language pair, with conditions on the horizontal axis and regression coefficients on the vertical axis. Colors distinguish continuous error (CE, displacement $\pm 1$) and hierarchical error (HE, $\pm 3$). Whiskers show 95% CIs from 5000 base-sequence block-bootstrap resamples. $\beta_{\mathrm{HE}}$ is generally small or reduced from L0 to L1, but increases from L1 to L3. With the familiar-word structure destroyed, it is near zero under S0 and positive under S2. Visible boundaries redirect absolute-position errors toward the fixed-3 $\pm 3$ adjacent-group displacement.

group component of absolute-position errors with or without the premise (Section F.5), and being estimated from errors only, this component is free of the headroom issue noted in Section 4.

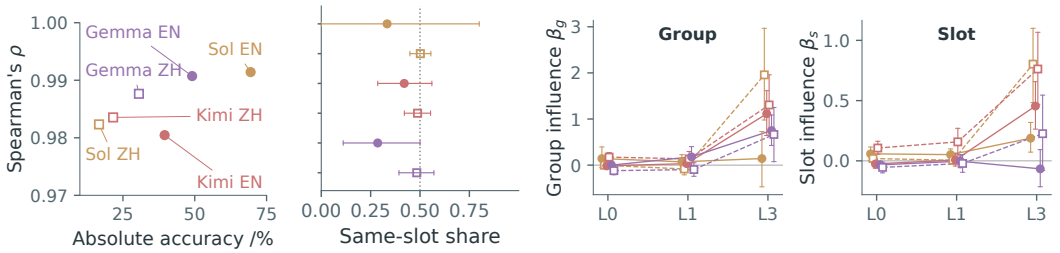
**Boundary-aligned absolute errors persist after the familiar-word structure is destroyed.** To test whether the boundary effect depends on intact familiar-word structure, S0 and S2 keep the item multiset of L0 and L2 but destroy that structure by reordering the items so that no valid multi-item word remains. Absolute-position accuracy still decreases from S0 to S2 in all 6 pairs, by 1.7–37.9 pp (Fig. 5) and from 97.92% to 60.01% for Gemma-ZH, and the error structure shows the same boundary effect (Fig. 6). Under S0, $\beta_{\mathrm{HE}}$ lies between $-0.030$ and approximately 0 in all 6 pairs; under S2 it is positive in all 6, for example $0.156$ for Sol-EN and $0.329$ for Gemma-ZH (Section F.5). Because no familiar word can be recovered, the period-3 errors reflect the displayed boundaries themselves, and visible boundaries produce a stronger HE component even when the familiar-word structure has been destroyed.

Across the fixed-3 comparisons, visible boundaries therefore shift access toward structure-relative reports and redirect absolute-position errors toward boundary-aligned alternatives. Fixed-3 layouts, however, leave an ambiguity, because a same-slot report in an adjacent group is necessarily a displacement of $\pm 3$. We next vary group lengths to separate the same-slot relation from this fixed displacement.

## 6 Variable Lengths Reveal Structure-Relative Influence

**Visible boundaries preserve the retrieval trade-off when group lengths vary.** Variable-length materials, with group lengths drawn from $\{2, 3, 4\}$, remove the fixed relation between absolute and structure-relative positions, so that group and slot positions can vary independently of absolute position. We repeat L0, L1, and L3 on 48 base sequences per model–language pair and condition, with 432–1152 absolute-position, 144 group-position, and 192 within-group slot probes. From L1 to L3, absolute-position accuracy decreases in all 6 pairs, while group accuracy increases in 5/6 and slot accuracy in all 6 (Fig. 3); for Sol-ZH, absolute accuracy falls from 90.28% to 16.67% while group accuracy rises from 50.69% to 84.03% and slot accuracy from 91.67% to 98.44% (Section F.1). The trade-off therefore does not depend on the repeated period of 3.

**Absolute-position reports retain a strong global ordinal association.** To separate inexact absolute numbers from a loss of order, we compute Spearman's $\rho$ between true and reported absolute positions. Under L3, exact accuracy ranges from 16.67% to 69.44%, whereas $\rho$ stays between 0.980 and 0.991 (Fig. 7a). This global association does not imply precise local order, since median $|e|$ ranges from 0 to 7 positions and strict ordinal consistency for targets at most 3 positions apart ranges from 69.09% to 94.09% (Table A19). Sol-ZH is the clearest case, with 16.67% exact accuracy, $\rho = 0.982$, median $|e| = 7$, and 76.43% near-pair consistency; 20.40% of its L3 reports fall outside the legal range, which changes $\rho$ by at most 0.004 across response-range policies (Section F.8). We therefore read $\rho$ as global ordinal association only. Boundaries thus degrade the exactness of absolute reports far more than the order information behind them.




(a) Ordinal association (b) Error preference (c) Grouping influence

Figure 7: **Ordinal association, lag-matched errors, and structure-relative influence in variable-length materials.** Colors denote models, filled circles EN, and open squares ZH. Whiskers show 95% CIs. (a) L3 absolute-position accuracy is plotted against Spearman's $\rho$ between true and reported positions. Exact accuracy ranges from 16.67% to 69.44%, while $\rho$ remains between 0.980 and 0.991. (b) L3 same-slot share is measured among lag-matched errors for which only one of the 2 equidistant alternatives shares the target's slot. No 95% CI lies entirely above the 0.5 equal-preference reference. (c) $\beta_g$ and $\beta_s$ quantify group- and slot-position influence on absolute-position error. Under L3, 95% CIs lie above zero for most $\beta_g$ and $\beta_s$ estimates, substantially more often than under L1. Together, variable-length layouts show a strong global ordinal association alongside structure-sensitive absolute-position errors, without implying exact local-order preservation or reducing to a fixed same-slot displacement rule.

**Lag-matched errors show no common same-slot preference.** We next test whether the period-3 pattern generalizes to a length-invariant preference for the same within-group slot. For each valid in-range error with lag $d = |\hat{i} - i|$, the error is identifying only when both alternatives $i - d$ and $i + d$ are valid positions and exactly one shares the target's slot, and the same-slot share is the fraction of identifying errors that chose that alternative. The 6 L3 shares range from 0.286 to 0.502, with no 95% CI above 0.5 (Fig. 7b), so the fixed-3 signature does not generalize as a same-slot preference (Section F.6).

**Visible boundaries strengthen structure-relative influence on absolute-position reports.** We next ask how group and slot positions enter the absolute report by regressing the signed error $e = \hat{i} - i$ on them. Within each exact $(i, s)$ stratum that varies in group position, we center $g$ and $e$ by their stratum means and regress $e$ on $g$, and $\beta_s$ is obtained analogously by varying slot within exact $(i, g)$ strata. Positive $\beta_g$ means that, at the same absolute position and slot, a later group shifts the reported number upward; positive $\beta_s$ means that, at the same absolute position and group, a later slot does. Visible boundaries strengthen both influences for most pairs (Fig. 7c). Under L1, no $\beta_g$ has a 95% CI above zero; under L3, 5/6 do, reaching 1.957 for Sol-ZH and 1.305 for Kimi-ZH. Slot influence shows the same shift but is less uniform, with the count rising from 2/6 under L1 to 4/6 under L3 and reaching 0.803 for Sol-ZH and 0.762 for Kimi-ZH. Visible boundaries thus make the absolute report measurably sensitive to structure-relative coordinates, although this does not by itself identify how the report is computed.

Figure 8 illustrates these error patterns, with L3 errors forming parallel bands at the $\pm 3$ displacement in fixed-3 materials and broader, position-dependent deviations in variable-length materials. One testable account is that boundaries lead models to convert group and slot position into an absolute position while assuming 3 items per group, which predicts $\hat{i} = h$ with $h = 3(g - 1) + s$. Regressing $e$ on $h - i$ gives slopes that are interval-positive under L3 in all 6 pairs and that increase from L1 to L3 in all 6, so the errors do align more with this prediction once boundaries are visible. The account nevertheless explains little, since the regression accounts for only 1.8–8.7% of L3 error variance, $\hat{i} = h$ reproduces only 2.40–31.06% of errors, and none of the 235 and 114 out-of-range answers of Sol-ZH and Kimi-ZH equals $h$ (Section F.7). The structure-relative influence is therefore not a fixed arithmetic conversion, although a noisy or partial hierarchical readout remains possible. Further response examples are in Section H.

## 7 Discussion

Our experiments reveal a progressive effect of structural cues on positional retrieval. A lexical cue mainly opens access to group and within-group slot positions, visible boundaries produce a stronger trade-off and redirect absolute-position errors toward the displayed grouping, and the boundary ef-

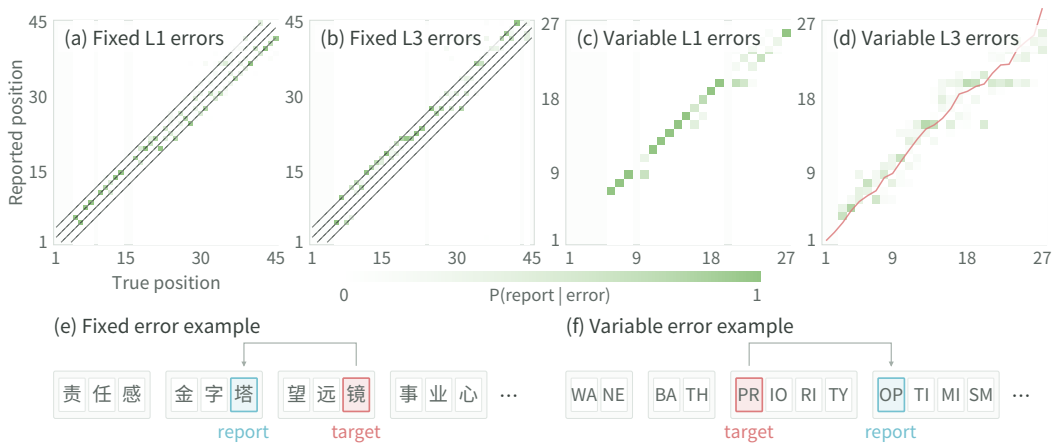



Figure 8: **Absolute-position error maps and examples.** (a,b) Kimi-K3 ZH fixed-3 and (c,d) Kimi-K3 EN variable-length errors under L1 and L3. The horizontal axis gives true absolute position, the vertical axis gives reported position, and green intensity gives the in-range report distribution conditional on an error at each true position. In the fixed-3 materials, L1 errors remain concentrated near the correct-position diagonal, whereas L3 produces stronger parallel bands aligned with the $\pm 3$ displacement between adjacent groups. In the variable-length materials, L3 instead produces broader position-dependent deviations rather than a single repeated absolute offset. The curve in (d) gives mean reported position under L3. In (e,f), red marks the target and cyan marks the item at the reported absolute position. (e) The target at absolute/group/slot positions $(9, 3, 3)$ is reported as $(6, 2, 3)$, moving to the adjacent previous group while preserving slot 3. (f) The target at $(5, 3, 1)$ is reported as $(9, 4, 1)$, moving to the adjacent next group while preserving slot 1 despite an absolute displacement of $+4$. These examples show how group structure can shape absolute-position errors.

fect persists after the familiar-word structure is destroyed and when group lengths vary, where exact absolute accuracy can collapse while the rank order of reports is largely preserved. What changes is therefore how position is read out rather than whether the sequence is available.

This pattern is a compatibility effect. Retrieval depends on the match between the organization a cue makes salient and the coordinate a query demands, and a structure that helps one positional description can interfere with another even though the sequence and its order remain in context. The effect parallels hierarchical WM in humans, where the same organization supports word and syllable reports while shaping continuous-rank reports (Fan et al., 2025), with the qualification that in LLMs the organization is largely cue-dependent rather than spontaneous. Our evidence is behavioral, so it does not identify the computation behind the readout. The variable-length results exclude a fixed-size conversion between coordinates but leave open a noisy or partial hierarchical readout (Section A).

Prompts therefore act as organizational interventions on in-context sequential memory. Making a structure salient changes which positional relations are most accessible and how they enter the response, so structured prompts should not be treated as uniformly beneficial. Memory systems that support several ways of locating information should likewise be evaluated across those readouts and for interference between them, and a representation that keeps absolute and structure-relative coordinates in parallel, with a readout selected by the query, is one route to avoiding this interference (Section B).

## 8 Conclusion

Across controlled EN and ZH sequences, structural cues reorganize positional retrieval in LLMs. Lexical cues mainly improve group and within-group slot reports, and visible boundaries amplify this shift while making absolute-position reports less exact and more sensitive to the grouping. These effects persist after the familiar-word structure is destroyed and when group lengths vary, where exact absolute accuracy can decline while reported positions remain strongly rank-correlated with true positions. Prompt structure thus changes how sequence position is read out from a context that still contains it.

### AI USE STATEMENT

The research questions, hypotheses, conceptual framing, experimental design, evaluation criteria, and interpretation of results in this work were led and determined by the authors. Generative AI tools were used as supporting tools throughout the research process, including for discussing and refining research ideas and experimental procedures, implementing author-specified methods, developing and editing code, assisting with experiment execution, organizing results, refining figures, conducting literature search and summarization, and modifying the manuscript. For these tasks, substantive methodological decisions, experimental choices, analyses, and conclusions were made by the authors.

All AI-assisted work was reviewed by the authors. AI-assisted code and experimental implementations were inspected and tested against the intended design, experimental outputs were checked and analyzed by the authors, literature claims and citations were verified against their original sources, and AI-assisted manuscript content was reviewed and revised before inclusion. The authors take responsibility for the final content of this work, including text, claims, code, analyses, and artifacts produced with the aid of generative AI.

### ETHICS STATEMENT

All authors have read and adhere to the ICLR Code of Ethics. This work evaluates LLMs using constructed linguistic sequences and does not involve new human participants, animals, or the collection of personal or sensitive data. The experimental materials are constructed from publicly available lexical resources and previously published word banks. The study is limited to controlled behavioral evaluation of language models and does not involve deployment in high-stakes or user-facing settings.

### REPRODUCIBILITY STATEMENT

The experimental setup is described in the main text, with full prompts and material construction in Section C, model configurations, capability matching, sampling plans, and answer parsing in Section D, and metric definitions in Section E. Detailed results and robustness analyses are provided in Sections F and G. The supplementary code (Section D.6) includes the test-set construction materials, sampling plans, and analysis implementations used to reproduce the reported experiments and statistical results.

## REFERENCES

Kevin Atkinson. SCOWL: Spell checking oriented word lists, version 2020.12.07. https://wordlist.aspell.net/other/, 2020. 4

Xuejun Bai, Guoli Yan, Simon P Liversedge, Chuanli Zang, and Keith Rayner. Reading spaced and unspaced Chinese text: evidence from eye movements. *Journal of Experimental Psychology: Human Perception and Performance*, 34:1277, 2008. A1

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, et al. LongBench: A bilingual, multitask benchmark for long context understanding. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024. 1, 3

Gordon DA Brown, Janet I Vousden, and Teresa McCormack. Memory retrieval as temporal discrimination. *Journal of Memory and Language*, 60:194–208, 2009. 3

Marc Brysbaert and Boris New. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41:977–990, 2009. A4

Qing Cai and Marc Brysbaert. SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLOS ONE*, 5:e10729, 2010. A4

CC-CEDICT Contributors. CC-CEDICT. MDBG Chinese-English Dictionary, https://www.mdbg.net/chinese/dictionary?page=cc-cedict, 2026. 4

Huiyao Chen, Yi Yang, Yinghui Li, Meishan Zhang, Baotian Hu, and Min Zhang. Beyond chunking: Discourse-aware hierarchical retrieval for long document question answering. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2026. 3

Ying Fan, Muzhi Wang, Fang Fang, Nai Ding, and Huan Luo. Two-dimensional neural geometry underpins hierarchical organization of sequence in human working memory. *Nature Human Behaviour*, 9:360–375, 2025. 1, 2, 3, 4, 5, 6, 9, A1, A3, A6

Murray Glanzer and Anita R Cunitz. Two storage mechanisms in free recall. *Journal of Verbal Learning and Verbal Behavior*, 5:351–360, 1966. 2

Dongyu Gong, Xingchen Wan, and Dingmin Wang. Working memory capacity of ChatGPT: An empirical study. In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, 2024. 3

Xiaobo Guo and Soroush Vosoughi. Serial position effects of large language models. In *Findings of the Association for Computational Linguistics*, 2025. 1, 3

Hassan Hajipoor, Hadi Amiri, Maseud Rahgozar, and Farhad Oroumchian. Serial recall effects in neural language modeling. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019. 3

Tom Hartley, Mark J Hurlstone, and Graham J Hitch. Effects of rhythm on memory for spoken sequences: A model and tests of its stimulus-driven mechanism. *Cognitive Psychology*, 87:135–178, 2016. 1, 3

Richard NA Henson. Short-term memory for serial order: The start-end model. *Cognitive Psychology*, 36:73–137, 1998. 1, 2

Graham J Hitch. Temporal grouping effects in immediate recall: A working memory analysis. *The Quarterly Journal of Experimental Psychology Section A*, 49:116–139, 1996. 3

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. An empirical analysis of compute-optimal large language model training. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2022. A1

Eunjin Hong, Sumin Cho, and Juae Kim. Exploring working memory capacity in LLMs: From stressors to human-inspired strategies. In *International Joint Conference on Natural Language Processing and Asia-Pacific Chapter of the Association for Computational Linguistics*, 2025. 3

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*, 2024. 3

Mark J Hurlstone. Functional similarities and differences between the coding of positional information in verbal and spatial short-term order memory. *Memory*, 27:147–162, 2019. 1

Mark J Hurlstone, Graham J Hitch, and Alan D Baddeley. Memory for serial order across domains: An overview of the literature and directions for future research. *Psychological Bulletin*, 140:339, 2014. 2

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. LLMs get lost in multi-turn conversation. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2026. 3

Catherine L Lee and William K Estes. Item and order information in short-term memory: Evidence for multilevel perturbation processes. *Journal of Experimental Psychology: Human Learning and Memory*, 7:149, 1981. 1

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 (TACL)*, 12:157–173, 2024. 1, 3

John Philip McCrae, Alexandre Rademaker, Ewa Rudnicka, and Francis Bond. English WordNet 2020: Improving and extending a WordNet for English using an open-source methodology. In *LREC 2020 Workshop on Multimodal WordNets*, 2020. 4

Don McNicol and Andrew Heathcote. Representation of order information: An analysis of grouping effects in short-term memory. *Journal of Experimental Psychology: General*, 115:76, 1986. 3

Deven Mahesh Mistry, Anooshka Bajaj, Yash Aggarwal, Sahaj Singh Maini, and Zoran Tiganj. Emergence of episodic memory in transformers: characterizing changes in temporal structure of attention scores during training. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2025. 1, 3

Bennet B Murdock Jr. The serial position effect of free recall. *Journal of Experimental Psychology*, 64:482, 1962. 2

Honey LH Ng and Murray T Maybery. Grouping in short-term verbal memory: Is position coded temporally? *The Quarterly Journal of Experimental Psychology Section A*, 55:391–424, 2002. 1, 3

Dennis Norris and Kristjan Kalm. Chunking and data compression in verbal short-term memory. *Cognition*, 208:104534, 2021. 1

OpenAI. GPT-5.6 system card. https://deploymentsafety.openai.com/gpt-5-6, 2026. 4, A5

Keith Rayner, Martin H Fischer, and Alexander Pollatsek. Unspaced text interferes with both word identification and eye movement control. *Vision Research*, 38:1129–1144, 1998. A1

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. How good is your tokenizer? On the monolingual performance of multilingual language models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021. A1

Joanna Ryan. Grouping and short-term memory: Different means and patterns of grouping. *The Quarterly Journal of Experimental Psychology*, 21:137–147, 1969. 1, 3

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2024. 3

Zihao Tang, Xin Yu, Ziyu Xiao, Zengxuan Wen, Zelin Li, Jiaxi Zhou, Hualei Wang, Haohua Wang, Haizhen Huang, Weiwei Deng, et al. Mnemis: Dual-route retrieval on hierarchical graphs for long-term llm memory. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2026. 3

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026a. 4, A6

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, Jianfeng Cai, Xinyuan Cai, Peizhou Cao, Yuxuan Cao, Ziwei Chai, Y Charles, et al. Kimi K3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026b. 4, A5

Mirko Thalmann, Alessandra S Souza, and Klaus Oberauer. How does chunking help working memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 45:37, 2019. 1

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2025. 3

Zihao Yi, Zhenqing Ling, Delong Zeng, Haohao Luo, Zhe Xu, Wei Liu, Jian Luan, Wanxia Cao, and Ying Shen. Attention basin: Why contextual position matters in large language models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2026. 1, 3

Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, et al. GLM-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026. 4, A5

# A LIMITATIONS

Our evidence characterizes retrieval behavior. Identifying the computations underlying structure-relative influence, including persistent internal variables, transient readout strategies, or attention allocation, requires activation-level or causal analyses. Our results also characterize direct readout without explicit reasoning; enabling reasoning could reconstruct positions through deliberate counting and may alter the observed trade-offs.

The other main limitation is empirical breadth. We study controlled English (EN) and Chinese (ZH) linguistic sequences with one grouping level, fixed groups of 3 items or variable groups of 2–4 items, and positional retrieval from sequences that remain in context. All tested visible boundaries are aligned with intended groups; although the error-geometry effect recurs across several whitespace and symbolic surfaces (Section G.1), we do not test equally salient boundaries shifted away from lexical boundaries, which would better separate generic segmentation or tokenization effects from structural alignment. Generality to non-linguistic sequences, deeper or nested hierarchies, less regular long-context materials, and downstream reasoning or planning remains to be tested.

# B ADDITIONAL DISCUSSION

**Relation to human hierarchical working memory.** Our results reveal a functional parallel with hierarchical organization in human working memory (WM). Fan et al. (2025) show that a linear sequence can be represented along separable global and local positional dimensions, and that the behavioral consequence of this organization depends on the task, since stronger hierarchical representation benefits reports that explicitly require global and local ranks but can interfere with continuous-rank reports when a one-dimensional readout is required. We observe a similar compatibility effect in Large Language Models (LLMs). Making group structure more salient improves access to group and within-group slot positions, while often making exact absolute-position reports less accurate and more strongly shaped by the grouping. This functional parallel suggests that structured retrieval can involve multiple positional coordinate systems whose usefulness depends on the requested readout. For future model and memory-system design, this points toward explicitly factorized representations that preserve absolute and structure-relative coordinates in parallel, together with query-dependent readout mechanisms that select the coordinate system appropriate for the current task rather than allowing one organization to dominate all retrievals.

**Model and language heterogeneity.** The common effects of structural cues are superimposed on clear differences across languages and models. Capability matching consistently selected longer sequences for ZH than for EN across all 5 models, suggesting that positional retrieval in the present assay is easier in ZH, although the selected loads should not be interpreted as direct measures of language-specific memory capacity (Table A5). One plausible contributor is the different role of whitespace in the 2 writing systems. In English, interword spaces provide strong visual cues to lexical boundaries, and removing them substantially disrupts word identification and reading (Rayner et al., 1998). Our EN materials instead split familiar words into two-letter items separated by spaces, placing ordinary segmentation cues inside the intended lexical units and potentially making their recovery less natural. Chinese, by contrast, is conventionally written without interword spaces, so identifying multi-character words does not depend on whitespace in the same way; native readers can segment ordinary unspaced text, and Chinese text with spaces inserted at correct word boundaries can be read about as efficiently as conventional unspaced text (Bai et al., 2008). The EN–ZH difference may therefore partly reflect how the itemization used in our task interacts with language-specific orthographic conventions rather than a general advantage of Chinese over English.

The variation across models may have a different origin. Although structural cues produce the same broad shift toward structure-relative retrieval across model families, the magnitude of this shift and the relative influence of group and within-group slot positions differ substantially across models. Such differences could reflect variation in model scale, architecture, tokenizer design, pretraining data composition, or post-training procedures across the evaluated families. More generally, both the allocation of model capacity and training data (Hoffmann et al., 2022) and the amount of language-specific pretraining data and tokenizer quality (Rust et al., 2021) are known to affect language-model performance. The heterogeneity suggests that the effect of structural prompting depends not only

on the surface organization of the sequence but also on model-specific representational and training characteristics.

# C Experimental Materials

## C.1 Prompt and Trial Construction

Each trial is a single user message. The displayed sequence remains in the request while the model answers. Ordinary adjacent items are separated by 3 spaces. Conditions with a visible boundary use 6 spaces at the intended group boundary and no additional marker symbol. Boundaries never count as items. All probes require one integer only.

**Condition-specific encoding text.** The following are the exact EN and ZH preambles. The 2 languages implement the same condition distinctions. Because group and within-group slot coordinates are defined relative to source words, the corresponding retrieval probes refer to complete words in every lexical condition. L0 and L2 differ from L1 and L3 in whether that structure is stated in the encoding instruction before retrieval.

**EN**

**L0 / S0:**

```
You will see an ordered sequence of chunks.  Remember the chunks and
the order in which they appear.
Sequence:  {sequence}
```

**L1:**

```
You will see an ordered sequence of chunks.  The sequence is formed
by several familiar English words placed one after another.  Each word
consists of multiple displayed chunks.  Remember the chunks and the
order in which they appear.
Sequence:  {sequence}
```

**L2 / S2:**

```
You will see an ordered sequence of chunks.  A larger gap marks a
boundary in the sequence.  Remember the chunks and the order in which
they appear.
Sequence:  {sequence}
```

**L3:**

```
You will see an ordered sequence of chunks.  The sequence is formed
by several familiar English words placed one after another.  Each word
consists of multiple displayed chunks.  A larger gap marks a boundary
between adjacent words.  Remember the chunks and the order in which
they appear.
Sequence:  {sequence}
```

**ZH**

**L0 / S0:**

```
你将看到一串按顺序排列的汉字。请记住这些字以及它们出现的顺序。
字序列：{sequence}
```

**L1:**

```
你将看到一串按顺序排列的汉字。这串字由若干完整的常见汉语词依次连接而成，每个词都由多
个字组成。请记住这些字以及它们出现的顺序。
字序列：{sequence}
```

**L2 / S2:**

```
你将看到一串按顺序排列的汉字。较长的空格表示序列中的边界。请记住这些字以及它们出现的
顺序。
字序列：{sequence}
```

**L3:**

```
你将看到一串按顺序排列的汉字。这串字由若干完整的常见汉语词依次连接而成，每个词都由多
个字组成。较长的空格表示相邻完整词之间的边界。请记住这些字以及它们出现的顺序。
字序列：{sequence}
```

**Retrieval text.** The following are the exact EN probes for lexical materials.

```
Target chunk: {target}

Absolute:
Counting only the displayed chunks from left to right across the entire
    sequence, at what position did the target chunk appear? Count the
    first chunk as position 1. Output one integer only.

Group, no visible boundary:
Which complete word from left to right contains the target chunk? Count
    complete words from left to right. Count the first word as 1. Output
    one integer only.

Group, visible boundary:
Which complete word from left to right contains the target chunk? Count
    complete words from left to right; the larger spacing only separates
    adjacent words. Count the first word as 1. Output one integer only.

Slot:
Within the complete word that contains the target chunk, counting only
    that word's two-letter chunks from left to right, what position is
    the target chunk? Output one integer only.
```

Before lexical group or slot questions, the request additionally states that test-sequence word boundaries should be determined from the complete familiar words and that each complete word consists of multiple displayed items. ZH probes are direct semantic counterparts. Absolute position counts all displayed characters continuously from the far left; group position counts complete words; within-group slot counts characters from the left inside the target word; and every response is restricted to one Arabic integer.

**Independent demonstrations.** Lexical requests use a real-word demonstration with deliberately nonuniform lengths, `ELEPHANT / GARDEN / BOOK`, corresponding to 4/3/2 displayed items. The no-boundary form uses ordinary 3-space item gaps; the boundary form uses 6 spaces between the 3 demonstration words. Absolute-position demonstrations explicitly number every displayed item continuously, so `BO`=8 in both renderings. Coordinate demonstrations identify `HA` as word 1/slot 3, `RD` as word 2/slot 2, and `BO` as word 3/slot 1. The ZH demonstration uses the matched 4/3/2 sequence 人工智能 / 自行车 / 苹果. The mixed 4/3/2 demonstration avoids presenting the fixed-3 test layout as a retrieval rule.

Structure-destroyed controls use a nonlexical demonstration, `QX BJ / VF KP ZM / RT WG`; the no-boundary form omits the slashes visually and the boundary form renders them as larger gaps. These demonstrations are independent of the test sequence and precede the final test query.

Full rendered requests, including every demonstration and condition-specific spacing pattern, can be regenerated from the supplied frozen base-sequence banks, prompt configurations, and paper sampling plans using the accompanying code.

### C.2 Test-Set Construction and Lexical Ambiguity Checks

EN items are non-overlapping two-letter segments and ZH items are individual characters. Fixed-3 materials use 6-letter EN words or 3-character ZH words. Variable-length materials use groups of 2, 3, or 4 items, corresponding to 4-, 6-, or 8-letter EN words and 2-, 3-, or 4-character ZH words. Every variable-length sequence contains all 3 group lengths; the numbers of length-2 and length-4 groups are equal and each is at least 2, so the total number of groups remains $N/3$.

For ZH, *source word* is used as an operational label for the familiar lexical units drawn from the word banks of Fan et al. (2025); the 4-character units in the variable-length pool include quadrisyllabic idioms. We treat these units identically to the other source lexical units when defining groups and positional coordinates.

Source words are audited independently of model behavior. The EN lexical inventory combines Open English WordNet 2025 with SCOWL 2020.12.07 surface forms; the ZH inventory uses sim-

Table A1: **Overview of the experimental batteries.** Each row summarizes the structural comparison, sample, model coverage, and role of one experimental battery. EN and ZH denote English and Chinese. The robustness batteries reuse frozen formal materials and sampling plans at their anchor loads and do not alter the main confirmatory data.

| Battery | Comparison | Sample | Models | Role |
|---|---|---|---|---|
| Lexical cue core | L0–L1 | Fixed-3; 256 bases/pair | All 5 models; EN and ZH | Isolate the effect of describing familiar-word structure without visible boundaries. |
| Visible boundary | L1–L3 | Fixed-3; same 256 bases and targets | Sol, Kimi, and Gemma; EN and ZH | Measure the effect of aligned boundaries given the same familiar-word premise. |
| Boundary controls | L0–L2; S0–S2 | Fixed-3; 64 matched bases/pair | Sol, Kimi, and Gemma; EN and ZH | Separate boundary-only effects and test them after the familiar-word structure is destroyed. |
| Variable length | L0, L1, L3 | Groups of 2–4 items; 48 bases/pair | Sol, Kimi, and Gemma; EN and ZH | Break the fixed lag-3 relation and identify group- and slot-relative influence. |
| Boundary-surface robustness | L1 vs. 4 L3 surfaces | Fixed-3; same 256 bases and targets | Sol and Gemma; EN and ZH | Test whether the boundary effect depends on the canonical 6-space rendering. |
| Sequence-length robustness | L0–L1 at 3 loads/language | Fixed-3; 256 bases/load | Luna and Gemma; EN and ZH | Test whether the lexical-cue pattern depends on the single capability-matched load. |

plified CC-CEDICT headwords. SUBTLEX-US (Brysbaert & New, 2009) and SUBTLEX-CH (Cai & Brysbaert, 2010) are used to audit source-word frequency and contextual diversity rather than to decide whether a competing lexical parse exists. Source entries lacking complete resource support are resolved before formal bank construction; rejected source words are removed from the construction pool but remain in the competitor inventory used by ambiguity checks.

A candidate sequence must satisfy 4 structural constraints. First, each displayed item is unique within the sequence. Second, EN sequences exclude non-palindromic reversed item pairs. Third, an intended source word cannot be completely resegmented into 2 or more valid multi-item lexical units. Fourth, no valid multi-item lexical unit may cross an intended word boundary. The final rule is checked during candidate search; for example, a concatenation analogous to `MUTUAL | LOCATE` is rejected if it creates the cross-boundary word `ALLOCATE`. These checks make the intended grouping the unique complete multi-item lexical parse under the frozen inventories.

The structure-destroyed controls preserve exactly the same item multiset as their lexical parent while reordering the items and requiring that no valid multi-item word remain anywhere in the resulting sequence. Final sequence selection balances word reuse across group positions and excludes exact or one-word-different near duplicates against the registered material banks. No material filter uses model responses or observed experimental effects.

The frozen word pools contain 3333 EN words and 493 ZH words. The fixed-3 banks contain 256 sequences for each supported load, $N \in \{18, 21, 27\}$ for EN and $N \in \{21, 33, 45, 48\}$ for ZH. The variable-length banks contain 48 sequences at each supported load, comprising 6 length patterns with 8 lexical instantiations per pattern, $N \in \{18, 27\}$ for EN and $N \in \{45, 48\}$ for ZH. Exact bank files, seeds, hashes, lexical-review records, and construction code are provided in the accompanying code. The internal repository name `irregular` refers to the reader-facing variable-length layout.

## D Evaluation Settings

The experimental batteries and their roles are summarized in Table A1.

### D.1 MODELS

We evaluate 5 model configurations from 4 model families: GLM-5.2, GPT-5.6 Luna, GPT-5.6 Sol, Kimi-K3, and Gemma-4-31B-IT. The model set includes both hosted and locally served systems and spans proprietary and open-weight model families. The model set provides cross-family and cross-deployment coverage for examining whether the positional effects recur across systems. Sequence load is capability-matched separately for each model–language pair, and the primary comparisons are matched contrasts between structural conditions within a pair.

The core L0–L1 comparison retains all 5 configurations. The boundary, control, and variable-length follow-ups use Sol, Kimi, and Gemma, covering 3 independently developed model families and both hosted and locally controlled inference. The preliminary task checks and the rationale for this scope are detailed in Section D.2.

All models are evaluated with temperature 0.0, and explicit reasoning or thinking modes are disabled where the serving interface exposes such a control, isolating direct positional readout from explicit multi-step counting or decomposition. Each trial is an independent single-turn request containing one sequence, one target, and one positional probe, with no conversational state carried across trials. GPT-5.6 Luna, GPT-5.6 Sol, GLM-5.2, and Gemma-4-31B-IT use a 512-token output cap, whereas Kimi-K3 uses a 128-token cap that was retained from the validated pretest configuration. Since the requested response is a single integer, these limits are substantially above the required answer length. Exact model identifiers, backend settings, and software versions are stored with each run for reproducibility.

#### D.1.1 GLM-5.2

GLM-5.2 is an open-weight model from the GLM family (Zeng et al., 2026). We include it to broaden the model set beyond the GPT, Kimi, and Gemma families and to provide an additional model evaluated in both EN and ZH. Our experiments access GLM-5.2 through the Volcengine Ark OpenAI-compatible API using the provider model identifier `glm-5-2-260617`. We set temperature to 0.0, disable the model's thinking mode, and use a 512-token output cap. The same generation configuration is used across languages and experimental conditions.

#### D.1.2 GPT-5.6 LUNA

GPT-5.6 Luna belongs to the GPT-5.6 family (OpenAI, 2026). OpenAI positions Luna as the lower-latency, cost-efficient member of the family, making it useful here as one capability point within a common model family. We access the model through the OpenAI Responses API using the identifier `gpt-5.6-luna`. We set temperature to 0.0, reasoning effort to `none`, and the output cap to 512 tokens. Including Luna together with Sol allows the qualitative effects of structural prompting to be examined at 2 capability points within the same provider family, while sequence load remains matched separately to each model–language pair.

#### D.1.3 GPT-5.6 SOL

GPT-5.6 Sol is the higher-capability GPT-5.6 configuration used in our experiments (OpenAI, 2026). We access it through the OpenAI Responses API with the identifier `gpt-5.6-sol`. As for Luna, temperature is set to 0.0, reasoning effort is set to `none`, and the output cap is 512 tokens. Sol is used in the full fixed-3 evaluation and in the more extensive follow-up analyses, including the variable-length experiments. The Luna–Sol pair provides a within-family comparison at separately capability-matched sequence lengths.

#### D.1.4 KIMI-K3

Kimi-K3 is a large open-weight mixture-of-experts model from the Kimi family (Team et al., 2026b). Its inclusion adds a model with a different architecture and training lineage from the GPT and GLM families and provides another independently developed system evaluated in both EN and ZH. We access Kimi-K3 through Alibaba Cloud DashScope's OpenAI-compatible endpoint using the identifier `kimi-k3`. Temperature is set to 0.0 and thinking is disabled. We use the 128-token output cap validated during pretesting. The task itself requires only a single-integer response.

Table A2: **Exploratory 6-space boundary diagnostics.** Each cell uses 4 base sequences and exhaustively probes L3 absolute position under the revised prompt. Correct and out-of-range counts use all queries as the denominator. The final column counts positive and negative displacements among valid, in-range, incorrect responses. These are descriptive calibration results at the listed loads, not estimates of matched L3−L1 effects.

| Model | Lang. | $N$ | Correct | Out of range | Positive / negative |
|---|---|---|---|---|---|
| Luna | EN | 30 | 79/120 | 8/120 | 26 / 7 |
| | ZH | 48 | 119/192 | 10/192 | 56 / 7 |
| GLM | EN | 30 | 55/120 | 0/120 | 16 / 49 |
| | ZH | 18 | 49/72 | 0/72 | 4 / 19 |

#### D.1.5 GEMMA-4-31B-IT

Gemma-4-31B-IT is a 31B-parameter instruction-tuned open-weight model from the Gemma 4 family (Team et al., 2026a). In contrast to the remotely served models above, we evaluate Gemma locally with vLLM, providing a locally controlled inference setting within the model set. The evaluated checkpoint is `gemma-4-31b-it`. We use temperature 0.0, disable thinking through the chat-template configuration, set the output cap to 512 tokens, and use an 8192-token model context limit for the experimental deployment. The same local checkpoint and inference configuration are used for both EN and ZH.

### D.2 MODEL COVERAGE AND PRELIMINARY TASK CHECKS

The core L0–L1 comparison includes all 5 model configurations. Before follow-up selection, all model–language pairs underwent a common competence screen at $N = 9$, the shortest fixed-3 sequence in Fan et al. (2025). The screen assessed reliable task completion and parsing, accurate L0 absolute-position retrieval, and accurate L3 absolute, group, and slot reports when the grouping structure was explicit. L0 group and slot performance was not an eligibility criterion because spontaneous access to those structure-relative positions is itself an experimental outcome. All 5 models passed this basic screen (Table A4).

The L2/L3 boundary conditions, S0/S2 controls, and variable-length experiments were restricted to Sol, Kimi, and Gemma because these follow-ups rely on interpretable positional readout under explicit boundaries. Sol and Kimi showed the strongest performance on the minimum-load positional readouts. Gemma also showed reliable preliminary performance, with $N = 9$ L3 absolute accuracy close to Sol's in both languages despite lower ZH slot accuracy, and it adds a locally served, open-weight, 31B-parameter dense model from an independently developed family (Team et al., 2026a). Together, the 3 models therefore provided strong boundary-condition readout while retaining coverage across model families and inference settings.

Luna and GLM were not extended to this follow-up battery because exploratory boundary calibration revealed additional absolute-position readout difficulties (Tables A2 and A3). In the 6-space diagnostic, Luna produced out-of-range responses and predominantly positive errors, whereas GLM's errors were predominantly negative. In an earlier hash-boundary calibration of Luna-EN ($N = 30$, 8 base sequences), 85 of 183 valid absolute errors exactly equaled the true position minus the number of preceding boundaries. The first example in Table A3 illustrates this relation, and the second shows an out-of-range response under the revised 6-space prompt. GLM showed a different pattern, as 19 of its 23 valid ZH errors under the 6-space prompt had displacement $-1$ or $-2$. In the earlier GLM-ZH hash-boundary calibration, group and slot accuracy were both perfect (36/36) despite only 24/36 correct absolute reports; the third example illustrates this dissociation. These patterns raised a specific concern that absolute-position errors under explicit boundaries could partly reflect unstable numerical readout rather than the structural effects targeted by the follow-up analyses. They therefore motivated prioritizing Sol, Kimi, and Gemma for the L2/L3, control, and variable-length batteries, rather than serving as a general model-exclusion criterion.

### D.3 CAPABILITY MATCHING AND LOAD SELECTION

For each eligible model–language pair described in Section D.1, we determine a capability-matched sequence length before the formal experiments using a frozen, outcome-blind procedure. The fixed-3 load is frozen using capability, error observability, and sequence-level stability rather than a single accuracy cutoff.

Table A3: **Illustrative observed responses from calibration.** Gold and response triples are ordered as (absolute, group, within-group slot). Each coordinate was queried in a separate request for the same sequence and target. The hash rows use the earlier prompt, whereas the 6-space row uses the revised prompt in Table A2; "–" denotes an unqueried coordinate. Examples illustrate the aggregate diagnostics and are not separate statistical evidence.

| Model | Surface | $N$ | Target | Gold | Responses |
|---|---|---|---|---|---|
| Luna EN | # | 30 | `RI` | $(11, 4, 2)$ | $(8, 4, 2)$ |
| Luna EN | 6 spaces | 30 | `TH` | $(10, 4, 1)$ | $(31, –, –)$ |
| GLM ZH | # | 18 | 要 | $(11, 4, 2)$ | $(10, 4, 2)$ |

Table A4: **Descriptive summary of the $N = 9$ competence screen.** Exact-match accuracy (%) for the 4 eligibility readouts; each cell contains 288 positional queries. Values summarize the completed screening stage, including selective query clarifications, rather than a cross-rendering robustness comparison. All listed pairs passed their basic competence checks. L0 group and slot performance was not an eligibility criterion.

| Model | Lang. | L0 abs. | L3 abs. | L3 group | L3 slot |
|---|---|---|---|---|---|
| GLM-5.2 | EN | 97.57 | 86.46 | 98.26 | 92.01 |
| | ZH | 91.32 | 90.63 | 98.26 | 88.54 |
| GPT-5.6 Luna | EN | 99.31 | 86.81 | 99.65 | 92.36 |
| | ZH | 100.00 | 90.97 | 98.26 | 90.97 |
| GPT-5.6 Sol | EN | 100.00 | 92.01 | 100.00 | 99.65 |
| | ZH | 100.00 | 95.49 | 100.00 | 99.65 |
| Kimi-K3 | EN | 99.65 | 97.22 | 100.00 | 98.61 |
| | ZH | 100.00 | 96.88 | 100.00 | 100.00 |
| Gemma-4-31B-IT | EN | 99.65 | 91.67 | 100.00 | 97.57 |
| | ZH | 98.96 | 94.44 | 100.00 | 90.97 |

The selected load must preserve reliable L0 absolute-position tracking without recurrent sequence- or position-specific collapse, while the planned 256-base formal schedule should yield enough absolute-position errors to estimate error geometry. An expected total of roughly 200 valid absolute errors is used as a planning target when feasible, rather than as a hard inclusion threshold. When a candidate load shows a reproducible local failure or large sequence-to-sequence instability, nearby loads are checked and the stable load is frozen before the formal runs. The resulting loads are reported in Table A5.

Load selection uses task completion, parsing, neutral absolute-position retrieval, error observability, and sequence-level stability. Structure-sensitive quantities, including the local-U profile, continuous error (CE)/hierarchical error (HE) coefficients, same-slot diagnostics, structure-relative influence slopes, and formal L0–L1 or L1–L3 contrasts, are evaluated only after the load is fixed. Calibration and pretest responses are disjoint from the 256-base formal sample. The load-selection procedure and implementation are provided in the accompanying code, and the resulting frozen loads are documented in the supplied experimental configurations and sampling plans.

### D.4 Sampling Plans

Sampling schedules are frozen separately for each model–language pair and reused across matched conditions. The fixed-3 core L0/L1 comparison uses 256 base sequences per pair. For the 6 model–language pairs evaluated under L3, the same 256 base identities are reused. L2 and the S0/S2 controls use an outcome-blind matched subset of 64 bases inherited from the corresponding fixed-3 schedule. The variable-length stage uses 48 base sequences per pair, comprising 6 group-length patterns with 8 lexical instantiations per pattern.

For fixed-3 materials, absolute probes use the pair-specific number of targets per base listed in Table A5; group and within-group slot probes use 3 targets each per base. The target schedule is frozen and reused across matched conditions whenever pairing is required. For variable-length materials, the last column of Table A5 reports the total numbers of absolute, group, and within-group slot requests per condition over the 48 bases.

Within a condition, each request contains one sequence, one target, and one probe. Matched contrasts reuse the same base identity and target schedule, so the structural intervention changes while the sampled sequence and target are held fixed.

Table A5: **Sequence lengths and per-condition sampling schedules.** $N$ is the number of displayed items. Fixed abs./base and fixed group/slot give the numbers of requests per base sequence; A/G/S totals give the numbers of absolute/group/within-group slot requests per condition. EN and ZH denote English and Chinese. Fixed-3 groups contain 3 items and variable-length groups contain 2–4 items. L0 (no structure cue) and L1 (familiar-word description) use 256 bases; L3 (description plus visible boundaries) uses the same 256 bases for Sol, Kimi, and Gemma. L2 (visible boundaries alone) and S0/S2 (familiar-word structure destroyed, without/with visible boundaries) use 64 matched bases. Variable-length totals cover 48 bases; "–" denotes an unevaluated configuration.

| Model | Lang. | $N$ | Fixed abs./base | Fixed group/slot | Variable A/G/S total |
|---|---|---|---|---|---|
| GLM-5.2 | EN | 18 | 18 | 3/3 | – |
| GLM-5.2 | ZH | 21 | 21 | 3/3 | – |
| GPT-5.6 Luna | EN | 21 | 21 | 3/3 | – |
| GPT-5.6 Luna | ZH | 33 | 24 | 3/3 | – |
| GPT-5.6 Sol | EN | 18 | 18 | 3/3 | 432/144/192 |
| GPT-5.6 Sol | ZH | 48 | 21 | 3/3 | 1152/144/192 |
| Kimi-K3 | EN | 27 | 21 | 3/3 | 648/144/192 |
| Kimi-K3 | ZH | 45 | 27 | 3/3 | 1080/144/192 |
| Gemma-4-31B-IT | EN | 27 | 21 | 3/3 | 648/144/192 |
| Gemma-4-31B-IT | ZH | 45 | 27 | 3/3 | 1080/144/192 |

### D.5 Answer Parsing and Coordinate Scoring

The reported runs use a unified response parser and scoring contract. Parsing is independent of the gold answer. A response is parsed directly when the entire visible response is a signed integer. Otherwise, closed `<think>...</think>` blocks are removed before parsing; residual transport markers or unclosed reasoning tags cause a parse failure. The parser then accepts an explicitly marked final integer (for example after "Answer:" or "答案："), a boxed integer, or a terminal line containing exactly one integer. If explicit candidates conflict, or no unique final integer can be identified, parsing fails. Responses are never repaired using the target or gold coordinate.

For an absolute probe, the legal coordinate range is $1, \ldots, N$. For explicit group probes it is $1, \ldots, G$, where $G$ is the number of groups, and for slot probes it is $1, \ldots, \ell_g$, where $\ell_g$ is the target group's length. L0 group and slot probes are treated slightly differently because no grouping is specified before the test sequence, even though the retrieval probe defines the requested word-relative coordinate. Any parsed integer in $1, \ldots, N$ is retained as a behavioral response, while exact-match correctness still requires equality with the intended group or slot coordinate. In all cells, provider errors, parse failures, and responses outside the applicable validity rule are scored as incorrect for exact-match accuracy.

### D.6 Code and Reproduction Package

The supplementary materials include a code archive for reproducing the experimental inputs, model evaluations, and statistical analyses reported in this paper. The package contains the frozen base-sequence banks and paper sampling plans, experimental and model configurations, source code for material and prompt generation, model evaluation, response parsing and scoring, statistical analysis, and regression tests. A top-level `README.md` documents installation and the end-to-end reproduction workflow. The supplied `scripts/reproduce.py` entry point supports generation of the frozen trial inputs, evaluation of the specified model–language configurations, and analysis of the resulting responses.

Access credentials for hosted model providers and the local Gemma checkpoint are external requirements and are not distributed in the supplementary archive. Generated model responses and analysis outputs are likewise not included; they can be regenerated using the supplied pipeline. The frozen inputs and sampling plans ensure that the reproduced evaluations use the same experimental materials and target schedules as the reported experiments.

# E METRICS

## E.1 COORDINATE ACCURACY, MATCHED CONTRASTS, AND UNCERTAINTY

For each probe type, exact-match accuracy is the mean of a trial-level indicator that equals 1 only when the request completes successfully, the response parser returns a valid integer under the scoring rule in Section D.5, and that integer equals the gold coordinate; all other trials contribute 0. Absolute, group, and slot accuracies are computed separately. For a matched intervention from a lower to a higher condition, the reported accuracy contrast is $\Delta A = A_{\text{high}} - A_{\text{low}}$, expressed in percentage points (pp).

Unless stated otherwise, 95% confidence intervals (CIs) use 5000 block-bootstrap resamples over base sequences. All requests belonging to the same base sequence are resampled together, and matched conditions are jointly resampled so that the pairing is preserved. The same block definition is used for the derived metrics below.

The reported 95% CIs are pointwise intervals for pre-specified pair-specific estimands and matched contrasts. Model–language pairs are analyzed as separate replications, so cross-pair recurrence is summarized from the direction and uncertainty of the pair-specific effects. The same-slot decision in Section E.5 uses its own within-pair permutation and sensitivity checks.

## E.2 LOCAL-U INDEX

The local-U analysis uses fixed-3 absolute probes. Let $A_s$ be exact-match absolute-position accuracy among targets occupying within-group slot $s \in \{1, 2, 3\}$. We define $U^{\text{abs}}_{\text{local}} = (A_1 + A_3)/2 - A_2$. Positive values indicate higher absolute-position accuracy at the 2 edge slots than at the middle slot. Because this raw contrast is expressed on the accuracy scale, its attainable magnitude depends in part on the overall error headroom. As a sensitivity analysis, we therefore additionally define $U_{\text{norm}} = U^{\text{abs}}_{\text{local}}/(1 - \bar{A})$, where $\bar{A} = (A_1 + A_2 + A_3)/3$. Thus, $U_{\text{norm}}$ expresses the middle-versus-edge excess error relative to the average error rate, which is not a probability and may exceed 1. Bootstrap resamples with a zero denominator are treated as undefined without adding an epsilon or clipping the statistic.

For each of the 5000 base-sequence block-bootstrap resamples, we recompute $A_1$, $A_2$, and $A_3$ before recomputing both local-U indices. L0 and L1 are resampled jointly over the same base-sequence blocks, and matched contrasts are obtained by subtracting the L0 statistic from the L1 statistic within each resample. Complete condition-level estimates and matched contrasts are reported in Tables A9 and A10.

## E.3 FIXED-3 CONTINUOUS AND HIERARCHICAL ERROR COEFFICIENTS

The CE/HE analysis is applied only to valid, in-range, incorrect absolute-position reports. For each true absolute position $i$, error counts over reported positions $j \neq i$ are normalized to $p_{ij} = P(\hat{i} = j \mid i, \hat{i} \neq i, 1 \leqslant \hat{i} \leqslant N)$. We then fit all observed off-diagonal probabilities jointly with $p_{ij} = \beta_0 + \beta_{\text{CE}} \mathbb{1}(|i - j| = 1) + \beta_{\text{HE}} \mathbb{1}(s_i = s_j, |g_i - g_j| = 1) + \epsilon_{ij}$. Thus, $\beta_{\text{CE}}$ is the additional normalized error probability associated with adjacent positions, whereas $\beta_{\text{HE}}$ is the additional probability associated with the fixed-3 HE pattern. Because same-slot adjacent-group reports are exactly the $\pm 3$ displacement in these materials, $\beta_{\text{HE}}$ does not by itself distinguish slot preservation from fixed periodic displacement.

For CIs, complete base-sequence blocks are resampled. Error-free base sequences remain in the bootstrap population even though they contribute no off-diagonal error counts, so uncertainty is not conditioned on a sequence having produced an error. Matched $\Delta\beta_{\text{HE}}$ contrasts subtract the lower-condition coefficient from the higher-condition coefficient within each bootstrap resample.

## E.4 ORDINAL ASSOCIATION AND ABSOLUTE-ERROR DIAGNOSTICS

For variable-length absolute probes, global ordinal association is measured by Spearman's $\rho$ between the true absolute positions $i$ and parsed integer reports $\hat{i}$. This metric uses all parseable integer absolute responses, including values outside $1, \ldots, N$, because those values still encode numerical

ordering and direction. A value near 1 therefore indicates a strong monotonic association between true and reported positions, but does not by itself imply small numerical error or preservation of local pairwise order. We complement $\rho$ with 2 absolute-error magnitude diagnostics. For each parseable response, let $e = \hat{i} - i$. We report the median absolute error, $\mathrm{median}(|e|)$, and the proportion of reports within 2 positions of the target, $\Pr(|e| \leqslant 2)$. Both statistics are computed over all parseable absolute-position responses, so exact responses contribute $|e| = 0$, and out-of-range integer responses are retained under the primary raw-response policy.

We additionally compute pairwise ordinal consistency directly within each base sequence. For 2 distinct probed targets $a$ and $b$, let $D = (i_a - i_b)(\hat{i}_a - \hat{i}_b)$. A pair is concordant when $D > 0$, inverted when $D < 0$, and tied when $\hat{i}_a = \hat{i}_b$. Strict ordinal consistency is the proportion of concordant pairs; ties therefore do not count as order-preserving. We compare only targets actually probed within the same base sequence and never form pairs across different bases. Because an all-pair statistic can be dominated by targets that are far apart, we also report a local diagnostic restricted to $1 \leqslant |i_a - i_b| \leqslant 3$. The resulting near-pair consistency tests whether the reported ordering is retained among targets whose true positions are close enough for the observed numerical errors to plausibly reverse or collapse their ordering. Pairwise estimates are pooled by pair count, while uncertainty is obtained by resampling complete base-sequence blocks.

### E.5 Lag-Matched Same-Slot Diagnostics

The exact same-lag analysis uses valid, in-range, incorrect variable-length absolute reports. For an error with target position $i$, reported position $\hat{i}$, and lag $d = |\hat{i} - i|$, the trial is identifying only when both equidistant candidates $i - d$ and $i + d$ are in range and exactly one of them occupies the same within-group slot as the target. The same-slot share is the fraction of identifying errors for which the model chose that same-slot candidate. Because the 2 candidates are exactly matched on absolute lag, 0.5 is the symmetric reference. Its CI is obtained by resampling base sequences.

The binary "positive same-slot criterion" combines the raw same-slot share with several additional checks. A positive decision requires a significant identifying permutation comparison, positive same-slot excess after dropping each identifying lag in turn, positive excess after excluding proportional-position matches, positive excess for far transpositions, and sufficient identifying layout variation. Together, these checks require the same-slot pattern to persist across lag, relative-position, and layout variations.

### E.6 Group/Slot Influence on Absolute-Position Reports

For variable-length layouts, signed absolute error is $e = \hat{i} - i$. These analyses retain every parseable integer response, including values outside $1, \ldots, N$, because the sign and magnitude of such responses remain informative about numerical distortion. To estimate group influence, rows are partitioned into exact $(i, s)$ strata. Within each stratum that contains variation in group position, both $g$ and $e$ are demeaned, and the pooled residualized slope is $\beta_g = \sum \tilde{g}\tilde{e} / \sum \tilde{g}^2$. Positive $\beta_g$ means that, at the same absolute position and slot, later group positions are associated with larger reported absolute numbers.

Slot influence is computed analogously within exact $(i, g)$ strata as $\beta_s = \sum \tilde{s}\tilde{e} / \sum \tilde{s}^2$. Strata with no variation in the predictor do not contribute to the estimand. For matched condition contrasts, we first compute the trialwise change in signed error for exactly paired absolute probes and then apply the same residualized-slope estimator. The reported $\Delta\beta_g$ and $\Delta\beta_s$ therefore measure how the intervention changes group or slot influence while preserving the relevant positional matching.

## F Detailed and Supplementary Results

All tables in this section use the frozen scoring and metric definitions in Sections D and E. In appendix tables, CI denotes confidence interval and pp denotes percentage points.

Table A6: **Fixed-3 exact-match accuracy under L0 and L1.** EN and ZH denote English and Chinese. $N$ is sequence length in displayed items; each source-word group contains 3 items. L0 provides no structure cue, and L1 describes familiar-word structure. Absolute position counts items across the sequence, group position counts source-word groups, and slot position counts items within the target group. Values are percentages over 256 base sequences per model–language pair.

| | | | L0 accuracy (%) | | | L1 accuracy (%) | | |
|---|---|---|---|---|---|---|---|---|
| Model | Lang. | $N$ | Abs. | Group | Slot | Abs. | Group | Slot |
| GLM-5.2 | EN | 18 | 87.52 | 45.96 | 41.93 | 80.99 | 65.10 | 53.12 |
| GLM-5.2 | ZH | 21 | 91.26 | 41.67 | 84.51 | 88.45 | 65.49 | 91.67 |
| GPT-5.6 Luna | EN | 21 | 86.64 | 32.68 | 43.49 | 87.95 | 39.32 | 50.91 |
| GPT-5.6 Luna | ZH | 33 | 94.40 | 39.06 | 49.22 | 85.06 | 54.17 | 63.54 |
| GPT-5.6 Sol | EN | 18 | 93.58 | 63.80 | 76.30 | 96.79 | 80.60 | 94.40 |
| GPT-5.6 Sol | ZH | 48 | 96.97 | 73.18 | 94.14 | 92.02 | 74.61 | 96.22 |
| Kimi-K3 | EN | 27 | 94.20 | 23.05 | 61.33 | 86.68 | 27.21 | 75.78 |
| Kimi-K3 | ZH | 45 | 97.57 | 83.20 | 94.01 | 83.28 | 41.54 | 97.27 |
| Gemma-4-31B-IT | EN | 27 | 91.89 | 47.66 | 84.24 | 80.10 | 98.44 | 99.74 |
| Gemma-4-31B-IT | ZH | 45 | 87.07 | 63.80 | 84.64 | 75.90 | 56.51 | 84.51 |

### F.1 Coordinate Accuracy

Table A6 gives all fixed-3 L0/L1 accuracies for the 10 core model–language pairs. Values are percentages; the corresponding paired uncertainty is summarized by the intervention analyses in the main text and by the matched-control tables below.

**Group-position exceptions in Kimi-ZH and Gemma-ZH.** The 2 L0-to-L1 decreases in group-position accuracy are strongly position dependent. Dividing the 15 fixed-3 groups descriptively into early (1–5), middle (6–10), and late (11–15) regions, Kimi-ZH changes from 95.72/83.98/69.80% under L0 to 92.22/26.56/5.49% under L1, whereas Gemma-ZH changes from 76.74/56.64/57.87% to 81.40/41.80/46.06%. These errors are valid group reports rather than completion or parsing failures, as completion, parse, and in-range rates are 1.00 for both L1 group probes.

For Kimi-ZH, all 449 incorrect L1 group reports underestimate the gold group index. Of these errors, 279 (62.1%) are one group lower, 158 (35.2%) are 2 groups lower, and 12 (2.7%) are 3 groups lower, shifting the mean signed group bias from $-0.152$ under L0 to $-0.822$ under L1. The lag becomes pronounced after the first several groups. For example, group 7 is reported as group 6 on 38/51 trials, group 9 as group 8 on 42/51 trials, and group 15 as group 13 on 41/52 trials. Gemma-ZH shows a milder downward shift. Among its 334 L1 group errors, 253 (75.7%) underestimate the gold index, including 214 (64.1%) one-group underestimates; the mean signed bias changes from $-0.014$ under L0 to $-0.243$ under L1. The late-group errors are predominantly adjacent underestimates, including group 13 reported as 12 on 24/51 trials, group 14 as 13 on 29/50 trials, and group 15 as 14 on 38/52 trials.

This group-index lag is not accompanied by a corresponding loss of the within-group coordinate. Kimi-ZH slot accuracy increases from 94.01% to 97.27%, while Gemma-ZH remains essentially unchanged at 84.64% versus 84.51%. Moreover, making the intended boundaries visible under L3 raises group accuracy to 99.35% for Kimi-ZH and 83.72% for Gemma-ZH. Behaviorally, the 2 exceptions therefore reflect a position-dependent lag in reporting the global group index when the familiar-word structure is stated but boundaries are not displayed, rather than a general failure to recover the target's structure-relative position.

For the 6 follow-up pairs, Table A7 adds the fixed-3 L3 condition to the L0/L1 values above.

Table A8 reports the complete variable-length L0/L1/L3 coordinate accuracies. The absolute-position decline from L1 to L3 occurs in all 6 pairs, while slot accuracy increases in all 6 and group accuracy in 5/6.

A notable difference appears for some ZH pairs under L0. At the same sequence length $N = 45$, Kimi-K3 absolute-position accuracy decreases from 97.57% in the fixed-3 materials to 71.85% in the variable-length materials, while Gemma-4-31B-IT decreases from 87.07% to 66.57% (Tables A6 and A8). This decrease is not shared by all models, since GPT-5.6 Sol ZH remains nearly unchanged across the 2 layouts, and the EN pairs show much smaller differences. One possible contributor is

Table A7: **Fixed-3 exact-match accuracy under L3.** EN and ZH denote English and Chinese. L3 describes familiar-word structure and marks boundaries between groups of 3 items. Absolute position counts items across the sequence, group position counts source-word groups, and slot position counts items within the target group. Values are percentages over 256 base sequences per pair.

| Model | Lang. | L3 absolute (%) | L3 group (%) | L3 slot (%) |
|---|---|---|---|---|
| GPT-5.6 Sol | EN | 93.10 | 99.87 | 99.74 |
| GPT-5.6 Sol | ZH | 93.81 | 99.22 | 99.74 |
| Kimi-K3 | EN | 75.63 | 98.18 | 95.44 |
| Kimi-K3 | ZH | 79.72 | 99.35 | 99.35 |
| Gemma-4-31B-IT | EN | 68.62 | 92.71 | 96.35 |
| Gemma-4-31B-IT | ZH | 60.73 | 83.72 | 86.85 |

Table A8: **Variable-length exact-match accuracy under L0, L1, and L3.** EN and ZH denote English and Chinese. $N$ is sequence length in displayed items; each source-word group contains 2–4 items. L0 provides no cue, L1 describes familiar-word structure, and L3 also marks boundaries. Absolute position counts items across the sequence, group position counts source-word groups, and slot position counts items within the target group. Values are percentages over 48 base sequences per pair.

| | | | L0 (%) | | | L1 (%) | | | L3 (%) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Lang. | $N$ | Abs. | Group | Slot | Abs. | Group | Slot | Abs. | Group | Slot |
| GPT-5.6 Sol | EN | 18 | 93.52 | 65.28 | 62.50 | 94.44 | 69.44 | 78.65 | 69.44 | 98.61 | 93.23 |
| GPT-5.6 Sol | ZH | 48 | 96.79 | 38.89 | 84.38 | 90.28 | 50.69 | 91.67 | 16.67 | 84.03 | 98.44 |
| Kimi-K3 | EN | 27 | 92.44 | 22.22 | 47.92 | 85.19 | 22.92 | 59.90 | 39.51 | 76.39 | 88.02 |
| Kimi-K3 | ZH | 45 | 71.85 | 41.67 | 83.85 | 52.50 | 31.25 | 88.54 | 21.57 | 91.67 | 94.27 |
| Gemma-4-31B-IT | EN | 27 | 93.52 | 45.83 | 81.77 | 82.41 | 91.67 | 91.67 | 49.07 | 79.86 | 93.23 |
| Gemma-4-31B-IT | ZH | 45 | 66.57 | 43.06 | 76.56 | 50.09 | 45.14 | 81.25 | 30.56 | 68.75 | 92.71 |

the broader lexical composition of the variable-length ZH materials. Whereas the fixed-3 bank contains 3-character lexical units, the variable-length bank mixes 2-, 3-, and 4-character units, with the 4-character pool also including quadrisyllabic idioms. Together with the irregular group lengths, this creates a more heterogeneous lexical and grouping structure, which may make uncued absolute-position tracking more difficult for some models.

### F.2 Within-Group Local-U Profiles

Table A9 reports condition-level raw local-U estimates with 95% block-bootstrap CIs, and Table A10 reports the matched L1−L0 contrasts together with the error-normalized sensitivity analysis. On the raw accuracy scale, the local-U point estimate increases from L0 to L1 in 9/10 model–language pairs. The matched 95% CIs for $\Delta U^{\mathrm{abs}}_{\mathrm{local}}$ lie entirely above zero in 7/10 pairs, include zero for GPT-5.6 Luna EN and GPT-5.6 Sol EN, and lie entirely below zero for GPT-5.6 Luna ZH.

The raw contrast should nevertheless be interpreted relative to available error headroom. After normalizing by the equally weighted mean error rate, the matched $\Delta U_{\mathrm{norm}}$ interval is above zero for GLM-5.2 EN, Kimi-K3 ZH, and Gemma-4-31B-IT ZH, below zero for GPT-5.6 Luna ZH, and includes zero for the remaining 6 pairs. In particular, GLM-5.2 ZH, GPT-5.6 Sol ZH, Kimi-K3 EN, and Gemma-4-31B-IT EN have positive raw matched intervals but interval-indeterminate normalized contrasts. Thus, lexical cues commonly enlarge the local-U contrast on the raw accuracy scale, but the magnitude of this enlargement is partly coupled to the greater error headroom available away from ceiling. We therefore use local-U as a descriptive characterization of profile shape rather than as independent evidence for a distinct retrieval mechanism.

### F.3 Boundary-Only and Structure-Destroyed Controls

The boundary-only and structure-destroyed controls are shown in Table A11. L2 is compared with the matched L0 subset for all 3 coordinate probes. S2 is compared with S0 for absolute probes only.

**S0 versus L0 baseline accuracy.** Absolute-position accuracy is higher under S0 than under the matched L0 subset in 4/6 model–language pairs, including all 3 EN models: Sol-EN increases from 92.53% to 98.52%, Kimi-EN from 94.20% to 97.77%, Gemma-EN from 92.19% to 99.78%, and Gemma-ZH from 88.54% to 97.92%. Sol-ZH is essentially unchanged, whereas Kimi-ZH shows the

Table A9: **Fixed-3 raw local-U estimates with bootstrap uncertainty.** EN and ZH denote English and Chinese. The local-U index is $U^{\text{abs}}_{\text{local}} = (A_1 + A_3)/2 - A_2$, where $A_s$ is exact-match absolute-position accuracy for targets at within-group slot $s$. Values are pp, with 95% CIs from 5000 base-sequence block-bootstrap resamples.

| Model | Lang. | L0 $U^{\text{abs}}_{\text{local}}$ | L1 $U^{\text{abs}}_{\text{local}}$ |
|---|---|---|---|
| GLM-5.2 | EN | 4.82 [2.90, 6.77] | 19.04 [16.96, 21.16] |
| GLM-5.2 | ZH | 8.82 [7.31, 10.35] | 11.47 [9.77, 13.20] |
| GPT-5.6 Luna | EN | 5.25 [3.29, 7.20] | 6.53 [4.72, 8.34] |
| GPT-5.6 Luna | ZH | 2.22 [1.00, 3.37] | -0.07 [-1.76, 1.59] |
| GPT-5.6 Sol | EN | 0.62 [-0.68, 1.89] | 0.65 [-0.29, 1.63] |
| GPT-5.6 Sol | ZH | 1.73 [0.81, 2.68] | 3.68 [2.26, 5.11] |
| Kimi-K3 | EN | 0.75 [-0.53, 1.98] | 3.21 [1.56, 4.85] |
| Kimi-K3 | ZH | 0.59 [-0.09, 1.32] | 9.09 [7.47, 10.70] |
| Gemma-4-31B-IT | EN | 5.83 [4.38, 7.23] | 12.67 [11.08, 14.26] |
| Gemma-4-31B-IT | ZH | 3.45 [2.28, 4.60] | 9.48 [7.75, 11.22] |

Table A10: **Matched lexical-cue changes in raw and error-normalized local-U.** $\Delta$ denotes L1 minus L0. Raw $\Delta U^{\text{abs}}_{\text{local}}$ is reported in pp. The normalized contrast uses $U_{\text{norm}} = U^{\text{abs}}_{\text{local}}/(1 - \bar{A})$, with $\bar{A} = (A_1 + A_2 + A_3)/3$, and is unitless. Brackets give matched 95% CIs from 5000 joint base-sequence block-bootstrap resamples.

| Model | Lang. | $\Delta U^{\text{abs}}_{\text{local}}$ | $\Delta U_{\text{norm}}$ |
|---|---|---|---|
| GLM-5.2 | EN | +14.23 [11.65, 16.77] | +0.616 [0.442, 0.788] |
| GLM-5.2 | ZH | +2.65 [0.61, 4.63] | -0.016 [-0.221, 0.190] |
| GPT-5.6 Luna | EN | +1.28 [-0.70, 3.35] | +0.149 [-0.011, 0.311] |
| GPT-5.6 Luna | ZH | -2.29 [-4.08, -0.56] | -0.402 [-0.611, -0.192] |
| GPT-5.6 Sol | EN | +0.03 [-1.46, 1.50] | +0.106 [-0.234, 0.441] |
| GPT-5.6 Sol | ZH | +1.95 [0.45, 3.43] | -0.109 [-0.458, 0.196] |
| Kimi-K3 | EN | +2.46 [0.56, 4.41] | +0.111 [-0.125, 0.355] |
| Kimi-K3 | ZH | +8.51 [6.86, 10.16] | +0.303 [0.014, 0.585] |
| Gemma-4-31B-IT | EN | +6.84 [4.94, 8.76] | -0.083 [-0.259, 0.089] |
| Gemma-4-31B-IT | ZH | +6.03 [4.10, 7.96] | +0.127 [0.023, 0.233] |

opposite pattern. One possible explanation is that recoverable familiar-word structure can compete with continuous absolute counting, whereas destroying the structure removes this organization and leaves a more purely linear sequence. The comparison also changes the demonstration from a real-word example in the lexical conditions to a nonlexical example in the structure-destroyed controls, so the higher S0 baseline may reflect both the disruption of lexical organization and the change in demonstration type.

### F.4 Boundary Interactions and Fixed-3 Readout Sensitivity

To test whether the effect of visible boundaries depends on the familiar-word premise, we compare the boundary-only contrast L2−L0 with the corresponding contrast L3−L1 on the matched 64-base secondary subset. For each probe, the factorial interaction is $(\text{L3} - \text{L1}) - (\text{L2} - \text{L0})$, so a positive value means that adding boundaries produces a more positive accuracy change when the familiar-word premise is already present. Table A12 shows no probe-general interaction pattern. The absolute-position interaction is positive in 5/6 pairs, with 4 intervals entirely above zero, whereas the group-position interaction has mixed signs. The slot-position interaction is negative in all 6 pairs, with 4 intervals entirely below zero, indicating that the boundary effect on slot accuracy is consistently smaller when the familiar-word premise is already present.

We additionally compare the boundary effect on structure-relative accuracy with its effect on absolute-position accuracy using the structure-relative-minus-absolute contrast $T = [\Delta A_{\text{group}} + \Delta A_{\text{slot}}]/2 - \Delta A_{\text{absolute}}$. Positive values mean that the average group- and slot-accuracy change is more positive than the absolute-accuracy change. This statistic is a relative contrast and does not by itself require the directional retrieval trade-off defined in the main text, in which structure-relative accuracy increases while absolute-position accuracy decreases. As shown in Table A13, all 12 estimates are positive and all 95% intervals lie above zero. The contrast is positive both for boundaries introduced without the familiar-word premise and for boundaries introduced after the premise is already present.

Table A11: **Fixed-3 boundary-only and structure-destroyed accuracy comparisons.** EN and ZH denote English and Chinese. Groups contain 3 items. L2 adds boundaries to L0 without a familiar-word description; S0/S2 destroy the familiar-word structure by reordering the same items, without/with boundaries. Low and high give exact-match accuracy before and after boundary addition; $\Delta$ is high minus low, in pp. Absolute, group, and slot count sequence items, groups, and items within a group. Brackets give 95% CIs from 5000 matched base-sequence block-bootstrap resamples over 64 bases.

| Model | Lang. | Contrast | Probe | Low (%) | High (%) | $\Delta$ accuracy pp [95% CI] |
|---|---|---|---|---|---|---|
| GPT-5.6 Sol | EN | L2−L0 | Absolute | 92.53 | 89.84 | -2.69 [-5.38, -0.09] |
| GPT-5.6 Sol | EN | L2−L0 | Group | 65.10 | 97.92 | +32.81 [+26.56, +39.06] |
| GPT-5.6 Sol | EN | L2−L0 | Slot | 73.96 | 94.79 | +20.83 [+15.62, +26.56] |
| GPT-5.6 Sol | EN | S2−S0 | Absolute | 98.52 | 94.27 | -4.25 [-5.64, -2.95] |
| GPT-5.6 Sol | ZH | L2−L0 | Absolute | 97.40 | 94.79 | -2.60 [-4.91, -0.60] |
| GPT-5.6 Sol | ZH | L2−L0 | Group | 70.31 | 97.92 | +27.60 [+21.35, +33.85] |
| GPT-5.6 Sol | ZH | L2−L0 | Slot | 93.23 | 100.00 | +6.77 [+3.65, +9.90] |
| GPT-5.6 Sol | ZH | S2−S0 | Absolute | 97.40 | 94.05 | -3.35 [-4.99, -1.71] |
| Kimi-K3 | EN | L2−L0 | Absolute | 94.20 | 81.32 | -12.87 [-15.77, -9.90] |
| Kimi-K3 | EN | L2−L0 | Group | 27.08 | 91.15 | +64.06 [+57.81, +70.31] |
| Kimi-K3 | EN | L2−L0 | Slot | 63.02 | 94.79 | +31.77 [+25.00, +39.06] |
| Kimi-K3 | EN | S2−S0 | Absolute | 97.77 | 87.80 | -9.97 [-11.98, -7.96] |
| Kimi-K3 | ZH | L2−L0 | Absolute | 97.80 | 87.50 | -10.30 [-11.86, -8.68] |
| Kimi-K3 | ZH | L2−L0 | Group | 78.65 | 100.00 | +21.35 [+16.15, +26.56] |
| Kimi-K3 | ZH | L2−L0 | Slot | 93.23 | 98.96 | +5.73 [+2.60, +8.85] |
| Kimi-K3 | ZH | S2−S0 | Absolute | 88.83 | 87.09 | -1.74 [-4.22, +0.75] |
| Gemma-4-31B-IT | EN | L2−L0 | Absolute | 92.19 | 64.58 | -27.60 [-31.55, -23.66] |
| Gemma-4-31B-IT | EN | L2−L0 | Group | 48.96 | 76.56 | +27.60 [+17.19, +37.50] |
| Gemma-4-31B-IT | EN | L2−L0 | Slot | 84.90 | 94.27 | +9.38 [+3.12, +15.62] |
| Gemma-4-31B-IT | EN | S2−S0 | Absolute | 99.78 | 70.98 | -28.79 [-32.14, -25.52] |
| Gemma-4-31B-IT | ZH | L2−L0 | Absolute | 88.54 | 62.21 | -26.33 [-29.46, -23.21] |
| Gemma-4-31B-IT | ZH | L2−L0 | Group | 69.27 | 86.46 | +17.19 [+8.85, +26.04] |
| Gemma-4-31B-IT | ZH | L2−L0 | Slot | 87.50 | 91.15 | +3.65 [-0.52, +7.81] |
| Gemma-4-31B-IT | ZH | S2−S0 | Absolute | 97.92 | 60.01 | -37.91 [-40.86, -34.95] |

Table A12: **Premise by boundary interactions in fixed-3 coordinate accuracy.** EN and ZH denote English and Chinese. Each entry gives $(\mathrm{L3}-\mathrm{L1})-(\mathrm{L2}-\mathrm{L0})$ in pp, followed by its 95% CI. Positive values indicate that the accuracy effect of adding visible boundaries is more positive when the familiar-word premise is already present. All contrasts use the same 64 matched bases, and intervals use 5000 matched base-sequence block-bootstrap resamples.

| Model | Lang. | Absolute [95% CI] | Group [95% CI] | Slot [95% CI] |
|---|---|---|---|---|
| GPT-5.6 Sol | EN | -1.56 [-5.03, +1.82] | -16.67 [-23.96, -8.85] | -14.58 [-19.79, -9.38] |
| GPT-5.6 Sol | ZH | +4.39 [+1.34, +7.59] | -0.52 [-8.35, +7.81] | -3.65 [-7.81, +0.52] |
| Kimi-K3 | EN | +0.52 [-3.57, +4.46] | +2.60 [-2.08, +7.81] | -17.71 [-24.48, -10.42] |
| Kimi-K3 | ZH | +4.98 [+2.49, +7.75] | +37.50 [+30.73, +44.79] | -3.65 [-6.77, -0.52] |
| Gemma-4-31B-IT | EN | +12.43 [+8.63, +16.29] | -34.38 [-44.27, -24.47] | -13.02 [-19.27, -7.29] |
| Gemma-4-31B-IT | ZH | +8.80 [+6.25, +11.40] | +15.10 [+6.77, +23.45] | -1.04 [-4.69, +2.60] |

Table A13: **Fixed-3 structure-relative-minus-absolute contrast.** The contrast is $[\Delta A_{\mathrm{group}}+\Delta A_{\mathrm{slot}}]/2-\Delta A_{\mathrm{absolute}}$, where each $\Delta A$ is the matched accuracy change produced by the named boundary intervention. Values are pp. Positive values indicate that the average boundary effect on group and slot retrieval is more positive than its effect on absolute-position retrieval. They do not require both structure-relative accuracies to increase. Brackets give 95% CIs from 5000 matched base-sequence block-bootstrap resamples over 64 bases.

| Model | Lang. | L2−L0 contrast [95% CI] | L3−L1 contrast [95% CI] |
|---|---|---|---|
| GPT-5.6 Sol | EN | +29.51 [+25.00, +34.29] | +15.45 [+11.28, +19.79] |
| GPT-5.6 Sol | ZH | +19.79 [+15.33, +24.44] | +13.32 [+8.22, +18.79] |
| Kimi-K3 | EN | +60.79 [+54.76, +66.93] | +52.72 [+47.02, +58.67] |
| Kimi-K3 | ZH | +23.84 [+20.43, +27.37] | +35.79 [+31.77, +39.90] |
| Gemma-4-31B-IT | EN | +46.09 [+39.92, +52.23] | +9.97 [+5.84, +14.29] |
| Gemma-4-31B-IT | ZH | +36.75 [+31.71, +41.93] | +34.98 [+30.50, +39.44] |

The intact-versus-destroyed comparison provides a separate test of whether the magnitude of the boundary-associated absolute-accuracy change depends on familiar-word order. The position-matched interaction $(\mathrm{L2}-\mathrm{L0})-(\mathrm{S2}-\mathrm{S0})$ has intervals containing zero in 4/6 pairs. The 2 intervals

Table A14: **Lexical-order interaction and cumulative readout diagnostic.** The lexical-order interaction is (L2 − L0) − (S2 − S0) for absolute-position accuracy on 64 matched bases, in pp. The readout slope is the change in signed L3 absolute-position error per crossed displayed boundary in the full 256-base run; mean bias is the mean signed L3 error among valid parsed responses. The readout gate is triggered when the absolute slope is at least 0.10 positions per boundary and its 95% CI excludes zero. In-range gives the percentage of parsed L3 absolute responses within $1, \ldots, N$. Brackets give 95% CIs from 5000 base-sequence block-bootstrap resamples.

| Model | Lang. | Lexical-order interaction [95% CI] | Readout slope [95% CI] | Mean bias | In-range (%) | Gate |
|---|---|---|---|---|---|---|
| GPT-5.6 Sol | EN | +1.56 [-1.30, +4.34] | -0.018 [-0.024, -0.012] | -0.039 | 100.00 | No |
| GPT-5.6 Sol | ZH | +0.74 [-1.71, +2.98] | -0.008 [-0.014, -0.002] | -0.025 | 99.98 | No |
| Kimi-K3 | EN | -2.90 [-6.55, +0.52] | -0.122 [-0.147, -0.099] | -0.376 | 99.93 | Yes |
| Kimi-K3 | ZH | -8.56 [-11.46, -5.61] | +0.084 [+0.076, +0.092] | +0.501 | 99.91 | No |
| Gemma-4-31B-IT | EN | +1.19 [-3.57, +6.10] | -0.144 [-0.161, -0.126] | -0.411 | 99.93 | Yes |
| Gemma-4-31B-IT | ZH | +11.57 [+7.46, +15.80] | -0.044 [-0.062, -0.027] | -0.097 | 99.80 | No |

excluding zero occur in opposite directions for Kimi-K3 ZH and Gemma-4-31B-IT ZH, so the interaction shows no common direction across pairs. The S2−S0 absolute-accuracy point estimate is nevertheless negative in all 6 pairs, as reported in Table A11.

We also test whether fixed-3 absolute reports exhibit a cumulative numerical offset associated with the number of displayed boundaries preceding the target. The diagnostic regresses signed absolute-position error on the number of crossed boundaries in the full 256-base L3 runs. Its prespecified readout gate is triggered only when the absolute slope reaches 0.10 positions per crossed boundary and its 95% interval excludes zero. As shown in Table A14, Kimi-K3 EN and Gemma-4-31B-IT EN reach the readout gate, whereas the other 4 pairs do not. The triggered cells indicate that cumulative numerical offsets can contribute to fixed-3 periodic error geometry. The variable-length analyses therefore remove the fixed period-3 relation and examine the error structure under nonuniform group lengths.

### F.5 Fixed-3 Absolute-Error Structure

For the matched 64-base secondary stage, Tables A15 and A16 report every condition-level CE/HE regression coefficient used to characterize absolute-error organization. The columns give the number of valid in-range incorrect absolute reports and the coefficient estimate with its 95% block-bootstrap CI. Cells with fewer than 150 valid errors are marked as low-error descriptive estimates. This count is not used as a validity cutoff. Zero-width bootstrap intervals are labeled *deg.* in the tables.

**Kimi-ZH shows spontaneous grouping under L0.** Kimi-ZH differs from the other model–language pairs in showing a substantial HE component even without an explicit structure cue. Under L0, $\beta_{\mathrm{HE}} = 0.313$ [0.238, 0.395], compared with $\beta_{\mathrm{CE}} = 0.226$ [0.148, 0.288]. Because this estimate is based on only 38 valid in-range errors, we treat its magnitude descriptively. The pattern nevertheless depends on the intact lexical ordering. After the familiar-word structure is destroyed in S0, $\beta_{\mathrm{HE}}$ is approximately zero ($-0.000$ [-0.001, 0.000]), while adding visible boundaries to the structure-destroyed sequence restores a positive HE component in S2 (0.155 [0.132, 0.202]). Consistent with the elevated L0 baseline, adding boundaries to the intact sequence does not produce a further increase, with a matched L2−L0 change of $-0.070$ [-0.165, 0.018]. Together, these results are consistent with Kimi-ZH spontaneously exploiting the latent word organization when the familiar-word order is intact, rather than exhibiting a generic period-3 error preference.

Table A17 gives the matched absolute-accuracy and HE-coefficient contrasts used across the fixed-3 intervention chain. L1−L0 and L3−L1 use the full 256-base formal schedule; L2−L0 and S2−S0 use the matched 64-base secondary subset. Positive $\Delta\beta_{\mathrm{HE}}$ means that the higher condition places more weight on the fixed-3 HE pattern.

### F.6 Variable-Length Identification Results

Variable group lengths break the deterministic fixed-3 relation between within-group slot and absolute lag. Table A18 reports global ordinal association together with condition-level group- and slot-influence slopes, while Table A19 provides complementary L3 error-magnitude and pairwise-

Table A15: **English fixed-3 absolute-position error coefficients.** L0 provides no structure cue, L1 describes familiar-word structure, L2 adds boundaries alone, and L3 combines both cues. S0/S2 destroy the familiar-word structure by reordering the items, without/with boundaries. Valid errors are in-range incorrect absolute reports from 64 matched bases. Continuous error (CE) is displacement $\pm 1$; hierarchical error (HE) is a same-slot adjacent-group error, displacement $\pm 3$. Each coefficient measures additional probability in a joint regression of error distributions normalized at each true position. Brackets give 95% CIs from 5000 base-sequence block-bootstrap resamples. A *deg.* interval denotes a degenerate zero-width bootstrap distribution and should not be read as evidence of zero uncertainty.

| Model | Cond. | Valid errors | $\beta_{\mathrm{CE}}$ [95% CI] | $\beta_{\mathrm{HE}}$ [95% CI] |
|---|---|---|---|---|
| GPT-5.6 Sol | L0 | 77 | 0.433 [0.367, 0.481] | 0.029 [-0.001, 0.073] |
| GPT-5.6 Sol | L1 | 36 | 0.428 [0.385, 0.500] | -0.009 [-0.015, 0.000] |
| GPT-5.6 Sol | L2 | 117 | 0.252 [0.208, 0.353] | 0.036 [-0.013, 0.067] |
| GPT-5.6 Sol | L3 | 89 | 0.376 [0.323, 0.427] | 0.046 [-0.010, 0.091] |
| GPT-5.6 Sol | S0 | 15 | 0.267 [0.164, 0.462] | -0.030 [-0.044, -0.005] |
| GPT-5.6 Sol | S2 | 65 | 0.335 [0.235, 0.399] | 0.156 [0.088, 0.271] |
| Kimi-K3 | L0 | 69 | 0.495 [0.450, 0.514] | 0.008 [-0.002, 0.036] |
| Kimi-K3 | L1 | 151 | 0.433 [0.395, 0.459] | -0.001 [-0.009, 0.013] |
| Kimi-K3 | L2 | 239 | 0.129 [0.086, 0.166] | 0.225 [0.169, 0.264] |
| Kimi-K3 | L3 | 345 | 0.142 [0.129, 0.167] | 0.075 [0.051, 0.103] |
| Kimi-K3 | S0 | 30 | 0.500 [*deg.*] | -0.000 [*deg.*] |
| Kimi-K3 | S2 | 164 | 0.326 [0.292, 0.361] | 0.165 [0.129, 0.198] |
| Gemma-4-31B-IT | L0 | 104 | 0.487 [0.452, 0.520] | 0.003 [-0.004, 0.020] |
| Gemma-4-31B-IT | L1 | 240 | 0.464 [0.439, 0.484] | 0.003 [-0.005, 0.014] |
| Gemma-4-31B-IT | L2 | 476 | 0.224 [0.185, 0.254] | 0.189 [0.155, 0.225] |
| Gemma-4-31B-IT | L3 | 446 | 0.286 [0.256, 0.308] | 0.131 [0.102, 0.165] |
| Gemma-4-31B-IT | S0 | 3 | 0.667 [0.500, 1.000] | 0.000 [*deg.*] |
| Gemma-4-31B-IT | S2 | 389 | 0.186 [0.160, 0.214] | 0.263 [0.247, 0.293] |

Table A16: **Chinese fixed-3 absolute-position error coefficients.** L0 provides no structure cue, L1 describes familiar-word structure, L2 adds boundaries alone, and L3 combines both cues. S0/S2 destroy the familiar-word structure by reordering the items, without/with boundaries. Valid errors are in-range incorrect absolute reports from 64 matched bases. Continuous error (CE) is displacement $\pm 1$; hierarchical error (HE) is a same-slot adjacent-group error, displacement $\pm 3$. Each coefficient measures additional probability in a joint regression of error distributions normalized at each true position. Brackets give 95% CIs from 5000 base-sequence block-bootstrap resamples. A *deg.* interval denotes a degenerate zero-width bootstrap distribution and should not be read as evidence of zero uncertainty.

| Model | Cond. | Valid errors | $\beta_{\mathrm{CE}}$ [95% CI] | $\beta_{\mathrm{HE}}$ [95% CI] |
|---|---|---|---|---|
| GPT-5.6 Sol | L0 | 32 | 0.500 [*deg.*] | -0.000 [*deg.*] |
| GPT-5.6 Sol | L1 | 120 | 0.491 [0.479, 0.498] | -0.000 [-0.001, -0.000] |
| GPT-5.6 Sol | L2 | 65 | 0.214 [0.160, 0.281] | 0.267 [0.182, 0.336] |
| GPT-5.6 Sol | L3 | 103 | 0.232 [0.179, 0.278] | 0.209 [0.167, 0.270] |
| GPT-5.6 Sol | S0 | 33 | 0.500 [*deg.*] | -0.000 [*deg.*] |
| GPT-5.6 Sol | S2 | 74 | 0.249 [0.210, 0.338] | 0.165 [0.081, 0.227] |
| Kimi-K3 | L0 | 38 | 0.226 [0.148, 0.288] | 0.313 [0.238, 0.395] |
| Kimi-K3 | L1 | 241 | 0.258 [0.211, 0.281] | 0.206 [0.185, 0.258] |
| Kimi-K3 | L2 | 216 | 0.184 [0.127, 0.220] | 0.243 [0.202, 0.281] |
| Kimi-K3 | L3 | 338 | 0.173 [0.126, 0.195] | 0.215 [0.189, 0.251] |
| Kimi-K3 | S0 | 169 | 0.495 [0.482, 0.500] | -0.000 [-0.001, 0.000] |
| Kimi-K3 | S2 | 223 | 0.272 [0.211, 0.291] | 0.155 [0.132, 0.202] |
| Gemma-4-31B-IT | L0 | 198 | 0.508 [0.508, 0.511] | -0.000 [*deg.*] |
| Gemma-4-31B-IT | L1 | 368 | 0.436 [0.421, 0.459] | 0.038 [0.013, 0.053] |
| Gemma-4-31B-IT | L2 | 646 | 0.120 [0.111, 0.144] | 0.301 [0.266, 0.313] |
| Gemma-4-31B-IT | L3 | 668 | 0.143 [0.126, 0.161] | 0.285 [0.255, 0.302] |
| Gemma-4-31B-IT | S0 | 36 | 0.522 [0.500, 0.538] | -0.000 [*deg.*] |
| Gemma-4-31B-IT | S2 | 669 | 0.076 [0.058, 0.092] | 0.329 [0.296, 0.349] |

order diagnostics. The pooled Spearman coefficient $\rho$ compares true and reported absolute positions. $\beta_g$ estimates signed-error change with group index while holding $(i, s)$ fixed; $\beta_s$ analogously varies slot while holding $(i, g)$ fixed.

Although Spearman's $\rho$ remains high under L3, the complementary diagnostics show substantial variation in numerical and local-order fidelity. Median absolute error ranges from 0 to 7 positions,

Table A17: **Matched fixed-3 changes in absolute accuracy and hierarchical errors.** EN and ZH denote English and Chinese. L0 provides no structure cue, L1 describes familiar-word structure, L2 adds boundaries alone, and L3 combines both cues. S0/S2 destroy the familiar-word structure by reordering the items, without/with boundaries. Contrasts subtract the second named condition from the first. Accuracy change is in pp; $\Delta\beta_{\mathrm{HE}}$ is the change in additional normalized error probability for the fixed-3 HE pattern (displacement $\pm 3$), controlling for adjacency. L1−L0 and L3−L1 use 256 bases; L2−L0 and S2−S0 use 64 matched bases. Brackets give 95% CIs from 5000 base-sequence block-bootstrap resamples.

| Model | Lang. | Contrast | $\Delta$ accuracy pp [95% CI] | $\Delta\beta_{\mathrm{HE}}$ [95% CI] |
|---|---|---|---|---|
| GPT-5.6 Sol | EN | L1−L0 | +3.21 [+2.30, +4.17] | -0.004 [-0.023, +0.007] |
| GPT-5.6 Sol | EN | L3−L1 | -3.69 [-4.75, -2.65] | +0.042 [+0.020, +0.065] |
| GPT-5.6 Sol | EN | L2−L0 | -2.69 [-5.38, -0.09] | +0.007 [-0.061, +0.049] |
| GPT-5.6 Sol | EN | S2−S0 | -4.25 [-5.64, -2.95] | +0.186 [+0.107, +0.296] |
| GPT-5.6 Sol | ZH | L1−L0 | -4.95 [-6.01, -3.92] | +0.000 [-0.000, +0.001] |
| GPT-5.6 Sol | ZH | L3−L1 | +1.79 [+0.43, +3.18] | +0.198 [+0.166, +0.232] |
| GPT-5.6 Sol | ZH | L2−L0 | -2.60 [-4.91, -0.60] | +0.267 [+0.184, +0.337] |
| GPT-5.6 Sol | ZH | S2−S0 | -3.35 [-4.99, -1.71] | +0.165 [+0.082, +0.226] |
| Kimi-K3 | EN | L1−L0 | -7.51 [-8.52, -6.57] | -0.007 [-0.017, +0.000] |
| Kimi-K3 | EN | L3−L1 | -11.05 [-12.80, -9.36] | +0.087 [+0.070, +0.108] |
| Kimi-K3 | EN | L2−L0 | -12.87 [-15.77, -9.90] | +0.217 [+0.155, +0.258] |
| Kimi-K3 | EN | S2−S0 | -9.97 [-11.98, -7.96] | +0.165 [+0.127, +0.198] |
| Kimi-K3 | ZH | L1−L0 | -14.29 [-15.52, -13.08] | -0.083 [-0.120, -0.044] |
| Kimi-K3 | ZH | L3−L1 | -3.56 [-5.08, -2.01] | +0.058 [+0.023, +0.070] |
| Kimi-K3 | ZH | L2−L0 | -10.30 [-11.86, -8.68] | -0.070 [-0.165, +0.018] |
| Kimi-K3 | ZH | S2−S0 | -1.74 [-4.22, +0.75] | +0.156 [+0.132, +0.202] |
| Gemma-4-31B-IT | EN | L1−L0 | -11.79 [-12.89, -10.68] | -0.003 [-0.008, +0.003] |
| Gemma-4-31B-IT | EN | L3−L1 | -11.48 [-13.23, -9.82] | +0.148 [+0.119, +0.160] |
| Gemma-4-31B-IT | EN | L2−L0 | -27.60 [-31.55, -23.66] | +0.186 [+0.149, +0.221] |
| Gemma-4-31B-IT | EN | S2−S0 | -28.79 [-32.14, -25.52] | +0.263 [+0.247, +0.291] |
| Gemma-4-31B-IT | ZH | L1−L0 | -11.17 [-12.20, -10.11] | +0.021 [+0.009, +0.033] |
| Gemma-4-31B-IT | ZH | L3−L1 | -15.16 [-16.93, -13.43] | +0.268 [+0.247, +0.283] |
| Gemma-4-31B-IT | ZH | L2−L0 | -26.33 [-29.46, -23.21] | +0.301 [+0.268, +0.312] |
| Gemma-4-31B-IT | ZH | S2−S0 | -37.91 [-40.86, -34.95] | +0.329 [+0.297, +0.349] |

Table A18: **Variable-length ordinal association and structure-relative influence.** EN and ZH denote English and Chinese. Groups contain 2–4 items. L0 provides no cue, L1 describes familiar-word structure, and L3 also marks boundaries. Spearman's $\rho$ compares true and reported absolute positions; 1 indicates a perfect monotonic rank association. $\beta_g$ is the slope of signed absolute-position error against group position within fixed absolute/slot strata; $\beta_s$ varies slot within fixed absolute/group strata. Positive slopes associate later group or slot positions with larger reported absolute positions. Brackets give 95% CIs from 5000 base-sequence block-bootstrap resamples. Each condition uses 48 bases.

| Model | Lang. | Cond. | Spearman $\rho$ | $\beta_g$ [95% CI] | $\beta_s$ [95% CI] |
|---|---|---|---|---|---|
| GPT-5.6 Sol | EN | L0 | 0.998 | 0.143 [-0.052, 0.394] | 0.058 [0.004, 0.114] |
| GPT-5.6 Sol | EN | L1 | 0.998 | 0.071 [-0.096, 0.227] | 0.052 [0.010, 0.099] |
| GPT-5.6 Sol | EN | L3 | 0.991 | 0.143 [-0.469, 0.729] | 0.188 [0.072, 0.319] |
| GPT-5.6 Sol | ZH | L0 | 1.000 | -0.002 [-0.047, 0.046] | 0.019 [0.005, 0.034] |
| GPT-5.6 Sol | ZH | L1 | 1.000 | -0.083 [-0.209, 0.037] | 0.005 [-0.030, 0.038] |
| GPT-5.6 Sol | ZH | L3 | 0.982 | 1.957 [0.979, 2.966] | 0.803 [0.484, 1.100] |
| Kimi-K3 | EN | L0 | 0.999 | -0.016 [-0.088, 0.055] | -0.026 [-0.062, 0.011] |
| Kimi-K3 | EN | L1 | 0.999 | 0.033 [-0.060, 0.135] | 0.007 [-0.043, 0.063] |
| Kimi-K3 | EN | L3 | 0.980 | 1.114 [0.643, 1.614] | 0.455 [0.263, 0.657] |
| Kimi-K3 | ZH | L0 | 0.999 | 0.174 [0.080, 0.273] | 0.106 [0.055, 0.162] |
| Kimi-K3 | ZH | L1 | 0.998 | 0.135 [-0.035, 0.296] | 0.158 [0.049, 0.270] |
| Kimi-K3 | ZH | L3 | 0.984 | 1.305 [0.649, 1.958] | 0.762 [0.469, 1.067] |
| Gemma-4-31B-IT | EN | L0 | 1.000 | 0.000 [-0.065, 0.068] | -0.033 [-0.064, -0.003] |
| Gemma-4-31B-IT | EN | L1 | 0.998 | 0.179 [-0.012, 0.402] | -0.003 [-0.057, 0.049] |
| Gemma-4-31B-IT | EN | L3 | 0.991 | 0.750 [0.427, 1.088] | -0.066 [-0.215, 0.093] |
| Gemma-4-31B-IT | ZH | L0 | 0.999 | -0.120 [-0.208, -0.027] | -0.054 [-0.099, -0.011] |
| Gemma-4-31B-IT | ZH | L1 | 0.998 | -0.096 [-0.242, 0.050] | -0.022 [-0.095, 0.052] |
| Gemma-4-31B-IT | ZH | L3 | 0.988 | 0.665 [0.075, 1.250] | 0.225 [-0.079, 0.545] |

and the proportion of reports within 2 positions of the target ranges from 25.87% to 99.31%. All-pair ordinal consistency remains high (94.65–98.38%), but this statistic includes many widely separated

Table A19: **Absolute-error magnitude and pairwise ordinal consistency under variable-length L3.** EN and ZH denote English and Chinese. Median $|e|$ and $\Pr(|e| \leqslant 2)$ are computed over all parseable absolute-position reports, including exact responses and out-of-range integers, where $e = \hat{i} - i$. All-pair consistency is the proportion of concordant target pairs within the same base sequence. Near-pair consistency restricts this comparison to targets whose true absolute positions differ by at most 3. Ties count as non-concordant. Brackets give 95% CIs from 5000 base-sequence block-bootstrap resamples over 48 bases.

| Model | Lang. | Median $\|e\|$ [95% CI] | $\Pr(\|e\| \leqslant 2)$ (%) [95% CI] | All-pair consistency (%) [95% CI] | Near-pair consistency (%) [95% CI] |
|---|---|---|---|---|---|
| GPT-5.6 Sol | EN | 0 [0, 0] | 99.31 [98.15, 100.00] | 98.38 [97.86, 98.90] | 94.09 [92.14, 95.96] |
| GPT-5.6 Sol | ZH | 7 [6, 7] | 25.87 [24.39, 27.43] | 96.18 [95.80, 96.54] | 76.43 [74.48, 78.36] |
| Kimi-K3 | EN | 1 [1, 1] | 84.10 [81.02, 86.96] | 94.65 [93.85, 95.41] | 76.56 [73.62, 79.49] |
| Kimi-K3 | ZH | 2 [2, 3] | 54.35 [49.68, 59.07] | 94.87 [94.38, 95.38] | 69.09 [67.29, 70.84] |
| Gemma-4-31B-IT | EN | 1 [0, 1] | 96.60 [95.08, 97.99] | 97.61 [97.20, 98.01] | 87.75 [85.71, 89.73] |
| Gemma-4-31B-IT | ZH | 1 [1, 1] | 77.22 [73.92, 80.30] | 95.99 [95.61, 96.36] | 72.95 [70.89, 74.98] |

Table A20: **Variable-length lag-matched same-slot error analysis.** EN and ZH denote English and Chinese. Groups contain 2–4 items. L0 provides no cue, L1 describes familiar-word structure, and L3 also marks boundaries. Identifying errors have 2 in-range alternatives at equal absolute displacement, exactly one sharing the target's slot. Same-slot share is the fraction choosing that alternative; 0.5 indicates equal preference, and "–" denotes no identifying errors. A positive criterion requires a significant identifying permutation test and positive excess after removing each lag, excluding proportional-position matches, and restricting to far transpositions, with identifying layout variation. Brackets give 95% CIs from 5000 base-sequence block-bootstrap resamples. Each condition uses 48 bases. Figure 7b plots only the L3 rows.

| Model | Lang. | Cond. | Identifying errors | Same-slot share [95% CI] | Positive same-slot criterion |
|---|---|---|---|---|---|
| GPT-5.6 Sol | EN | L0 | 2 | 0.000 [0.000, 0.000] | No |
| GPT-5.6 Sol | EN | L1 | 0 | – | No |
| GPT-5.6 Sol | EN | L3 | 6 | 0.333 [0.000, 0.800] | No |
| GPT-5.6 Sol | ZH | L0 | 0 | – | No |
| GPT-5.6 Sol | ZH | L1 | 0 | – | No |
| GPT-5.6 Sol | ZH | L3 | 299 | 0.502 [0.449, 0.555] | No |
| Kimi-K3 | EN | L0 | 1 | 0.000 [0.000, 0.000] | No |
| Kimi-K3 | EN | L1 | 1 | 0.000 [0.000, 0.000] | No |
| Kimi-K3 | EN | L3 | 50 | 0.420 [0.286, 0.560] | No |
| Kimi-K3 | ZH | L0 | 2 | 0.500 [0.000, 1.000] | No |
| Kimi-K3 | ZH | L1 | 43 | 0.279 [0.135, 0.421] | No |
| Kimi-K3 | ZH | L3 | 162 | 0.488 [0.421, 0.553] | No |
| Gemma-4-31B-IT | EN | L0 | 0 | – | No |
| Gemma-4-31B-IT | EN | L1 | 2 | 0.000 [0.000, 0.000] | No |
| Gemma-4-31B-IT | EN | L3 | 21 | 0.286 [0.111, 0.500] | No |
| Gemma-4-31B-IT | ZH | L0 | 10 | 0.500 [0.000, 0.857] | No |
| Gemma-4-31B-IT | ZH | L1 | 40 | 0.600 [0.421, 0.767] | No |
| Gemma-4-31B-IT | ZH | L3 | 122 | 0.484 [0.393, 0.570] | No |

target pairs. Restricting the analysis to targets at most 3 positions apart yields a more discriminative 69.09–94.09% consistency. Sol-ZH is particularly informative, since despite $\rho = 0.982$ its median absolute error is 7 positions and its near-pair consistency is 76.43%. Thus, the high Spearman coefficients support a strong global ordinal association, but should not be interpreted as evidence that local order is preserved throughout.

The same-lag comparison in Table A20 asks whether, among errors for which a same-slot and a different-slot alternative are available at exactly the same absolute displacement, the model favors the same-slot alternative. A share of 0.5 is the symmetric reference. A zero in the identifying-error column means that no error satisfied this stricter matched-lag identification rule; the corresponding share is therefore undefined and shown as "–". By contrast, a same-slot share of 0.000 with a nonzero identifying-error count is a defined estimate, since identifying errors exist but none selected the same-slot alternative. Figure 7b visualizes only the 6 L3 rows, all of which have at least one identifying error. The final column applies the guarded same-slot criterion defined in Section E.5. No tested cell satisfies all required checks.

Table A21: **Matched changes in variable-length group and slot influence.** EN and ZH denote English and Chinese. Groups contain 2–4 items. L0 provides no cue, L1 describes familiar-word structure, and L3 also marks boundaries. Contrasts subtract the second named condition from the first. $\beta_g$ is the slope of signed absolute-position error against group position within fixed absolute/slot strata; $\beta_s$ varies slot within fixed absolute/group strata. Positive slopes associate later group or slot positions with larger reported absolute positions. $\Delta\beta$ estimates the corresponding slope for trialwise changes in signed error across 48 matched bases. Brackets give 95% CIs from 5000 base-sequence block-bootstrap resamples.

| Model | Lang. | Contrast | $\Delta\beta_g$ [95% CI] | $\Delta\beta_s$ [95% CI] |
|---|---|---|---|---|
| GPT-5.6 Sol | EN | L1−L0 | -0.071 [-0.301, +0.121] | -0.007 [-0.066, +0.053] |
| GPT-5.6 Sol | EN | L3−L1 | +0.071 [-0.580, +0.731] | +0.136 [+0.006, +0.278] |
| GPT-5.6 Sol | ZH | L1−L0 | -0.082 [-0.202, +0.032] | -0.015 [-0.048, +0.018] |
| GPT-5.6 Sol | ZH | L3−L1 | +2.041 [+1.069, +3.031] | +0.798 [+0.477, +1.090] |
| Kimi-K3 | EN | L1−L0 | +0.049 [-0.062, +0.165] | +0.033 [-0.018, +0.085] |
| Kimi-K3 | EN | L3−L1 | +1.082 [+0.572, +1.617] | +0.447 [+0.250, +0.641] |
| Kimi-K3 | ZH | L1−L0 | -0.039 [-0.213, +0.122] | +0.051 [-0.059, +0.162] |
| Kimi-K3 | ZH | L3−L1 | +1.170 [+0.522, +1.847] | +0.605 [+0.288, +0.945] |
| Gemma-4-31B-IT | EN | L1−L0 | +0.179 [+0.015, +0.367] | +0.030 [-0.025, +0.085] |
| Gemma-4-31B-IT | EN | L3−L1 | +0.571 [+0.243, +0.912] | -0.063 [-0.208, +0.096] |
| Gemma-4-31B-IT | ZH | L1−L0 | +0.024 [-0.090, +0.137] | +0.033 [-0.024, +0.088] |
| Gemma-4-31B-IT | ZH | L3−L1 | +0.761 [+0.199, +1.314] | +0.247 [-0.049, +0.574] |

Finally, Table A21 reports the paired change in group and slot influence. These are the estimates underlying the statement that L3−L1 group influence is interval-positive in 5/6 pairs and slot influence in 4/6, whereas L1−L0 is much weaker.

### F.7 Testing a Fixed-3 Conversion Account of Variable-Length Reports

**A simple conversion is a testable alternative, not an identified mechanism.** The positive group- and slot-influence slopes in Section 6 raise a specific alternative explanation, that visible boundaries might induce models to convert structure-relative coordinates to an absolute position while incorrectly assuming that every group contains 3 items. We test this account using the existing variable-length absolute responses from GPT-5.6 Sol, Kimi-K3, and Gemma-4-31B-IT in EN and ZH, without collecting new model outputs. Each model–language pair contributes 48 matched base sequences per condition. We retain successful requests with a parseable integer answer, including out-of-range answers, as in the primary coordinate-influence analysis. The predictor uses the target's true group and slot positions, not its separately elicited group and slot reports.

**Error regression separates the proposed displacement from the correct position.** For a target at absolute position $i$, group position $g$, and within-group slot position $s$, define

$$h = 3(g-1) + s, \qquad \delta_h = h - i, \qquad e = \hat{i} - i. \tag{A1}$$

The prediction $h$ in Eq. (A1) treats all preceding groups as having length 3, even though actual group lengths vary from 2 to 4. We fit an ordinary least-squares regression with an intercept separately in each condition:

$$e = \alpha + \lambda\delta_h + \epsilon. \tag{A2}$$

In Eq. (A2), $\lambda = 0$ means no linear association between signed error and the displacement predicted by this rule. A slope of 1 alone does not establish exact conversion, since trialwise reports equal $h$ only if $\alpha = 0$ and the residuals are also zero. Likewise, a slope between 0 and 1 is not an estimated probability of using the rule. Because $e$ and $\delta_h$ both subtract $i$, and errors can vary with true position, we also fit $e = \alpha_i + \lambda_{\text{pos}}\delta_h + \epsilon$, with a separate intercept for each exact true position. This estimate demeans both variables within true-position strata and recomputes those means within every bootstrap resample. It removes mean differences across true positions but does not eliminate other layout-associated explanations. These estimands differ from $\beta_g$ and $\beta_s$, which hold different pairs of coordinates fixed; those coefficients cannot be converted directly into a mixture weight for this strategy.

**Uncertainty preserves the matched design.** We use 5000 base-sequence block-bootstrap resamples with seed 20260925 and percentile 95% CIs. Each resample retains all targets from a selected base, and the same base multiplicities are used across conditions. L3−L1 contrasts are computed

on the intersection of parseable base–target pairs and refitted within each resample. The 6 matched sample sizes are listed in Table A23. These exploratory intervals are not adjusted for multiple comparisons and quantify variation across sampled bases, not across models or repeated stochastic decoding.

**Boundaries strengthen alignment with the predictor, but its explanatory power is limited.** As shown in Table A22, L3 slopes are positive with 95% CIs above zero in all 6 pairs. The matched L3−L1 increase is also interval-positive in all 6 pairs; after true-position adjustment, it remains interval-positive in 5/6, with Sol EN including zero (Table A23). Thus, the results support a boundary-associated component aligned with the fixed-3 prediction. However, the unadjusted regression explains only 1.8–8.7% of L3 signed-error variance. Sol ZH has $\lambda = 0.723$ with a 95% CI of $[0.153, 1.368]$, which includes 1, so the slopes do not all establish partial dependence below 1. Its fitted intercept is 6.471 positions; Kimi ZH has an intercept of 2.387. These offsets and the low explained variance are inconsistent with treating the formula alone as an adequate account of the observed reports.

Table A22: **L3 errors align with the fixed-3 predictor, but the simple regression explains little variance.** Sol, Kimi, and Gemma denote GPT-5.6 Sol, Kimi-K3, and Gemma-4-31B-IT, respectively; EN and ZH denote English and Chinese. L0 provides no cue, L1 describes familiar-word structure, and L3 also marks boundaries. The unadjusted regression is $e = \alpha + \lambda(h - i) + \epsilon$, with $h = 3(g - 1) + s$; $R^2$ refers to this regression. $\lambda_{\text{pos}}$ additionally absorbs exact true-position intercepts. All parseable integer answers, including out-of-range values, are retained. Each condition uses 48 bases. Brackets give percentile 95% CIs from 5000 base-sequence block-bootstrap resamples.

| Model | Lang. | Cond. | $\lambda$ [95% CI] | $\alpha$ | $R^2$ | $\lambda_{\text{pos}}$ [95% CI] |
|---|---|---|---|---|---|---|
| Sol | EN | L0 | 0.039 [0.009, 0.076] | 0.031 | 0.011 | 0.042 [0.007, 0.083] |
| Sol | EN | L1 | 0.016 [-0.011, 0.043] | 0.022 | 0.002 | 0.022 [-0.003, 0.047] |
| Sol | EN | L3 | 0.154 [0.035, 0.276] | -0.077 | 0.042 | 0.136 [0.022, 0.251] |
| Sol | ZH | L0 | 0.004 [-0.006, 0.015] | -0.014 | 0.001 | 0.004 [-0.006, 0.015] |
| Sol | ZH | L1 | -0.010 [-0.044, 0.023] | 0.025 | 0.001 | -0.008 [-0.042, 0.026] |
| Sol | ZH | L3 | 0.723 [0.153, 1.368] | 6.471 | 0.038 | 0.626 [0.335, 0.911] |
| Kimi | EN | L0 | -0.011 [-0.031, 0.009] | 0.011 | 0.001 | -0.013 [-0.033, 0.006] |
| Kimi | EN | L1 | 0.015 [-0.014, 0.047] | 0.015 | 0.001 | 0.006 [-0.025, 0.040] |
| Kimi | EN | L3 | 0.544 [0.392, 0.692] | 0.261 | 0.087 | 0.541 [0.388, 0.699] |
| Kimi | ZH | L0 | 0.078 [0.036, 0.124] | 0.263 | 0.033 | 0.069 [0.034, 0.105] |
| Kimi | ZH | L1 | 0.106 [0.037, 0.180] | 0.689 | 0.017 | 0.087 [0.026, 0.146] |
| Kimi | ZH | L3 | 0.554 [0.326, 0.786] | 2.387 | 0.061 | 0.515 [0.310, 0.716] |
| Gemma | EN | L0 | -0.019 [-0.038, -0.001] | -0.014 | 0.005 | -0.026 [-0.046, -0.006] |
| Gemma | EN | L1 | 0.021 [-0.024, 0.070] | -0.091 | 0.002 | 0.001 [-0.042, 0.048] |
| Gemma | EN | L3 | 0.168 [0.067, 0.271] | -0.001 | 0.018 | 0.123 [0.020, 0.231] |
| Gemma | ZH | L0 | -0.058 [-0.098, -0.017] | -0.358 | 0.019 | -0.048 [-0.077, -0.018] |
| Gemma | ZH | L1 | -0.050 [-0.114, 0.012] | -0.628 | 0.006 | -0.040 [-0.082, 0.005] |
| Gemma | ZH | L3 | 0.219 [0.004, 0.420] | 0.436 | 0.020 | 0.248 [0.045, 0.445] |

Table A23: **The boundary-associated increase survives true-position adjustment in 5/6 pairs.** Model abbreviations are as in Table A22. $n$ is the number of common parseable absolute probes across L1 and L3; all 48 bases are retained in each pair. $\Delta\lambda$ and $\Delta\lambda_{\text{pos}}$ subtract the L1 slope from the L3 slope, without and with exact true-position intercepts, respectively. Brackets give percentile 95% CIs from 5000 matched base-sequence block-bootstrap resamples.

| Model | Lang. | $n$ | $\Delta\lambda$ [95% CI] | $\Delta\lambda_{\text{pos}}$ [95% CI] |
|---|---|---|---|---|
| Sol | EN | 432 | 0.139 [0.004, 0.275] | 0.114 [-0.012, 0.240] |
| Sol | ZH | 1152 | 0.733 [0.168, 1.377] | 0.634 [0.342, 0.916] |
| Kimi | EN | 648 | 0.528 [0.380, 0.671] | 0.535 [0.385, 0.689] |
| Kimi | ZH | 1080 | 0.449 [0.227, 0.675] | 0.428 [0.218, 0.645] |
| Gemma | EN | 648 | 0.146 [0.049, 0.249] | 0.122 [0.019, 0.227] |
| Gemma | ZH | 1080 | 0.269 [0.075, 0.449] | 0.288 [0.094, 0.474] |

**Exact matches do not account for most errors or the prominent ZH range violations.** We report $P(\hat{i} = h)$, $P(\hat{i} = h \mid \hat{i} \neq i)$, and $P(\hat{i} = h \mid h \neq i)$ in Table A24. The first includes correct answers on trials where $h = i$ and therefore cannot by itself identify use of the conversion rule. The

second measures exact matches among errors, while the third restricts attention to targets for which the rule predicts an incorrect absolute position. Under L3, only 2.40–31.06% of errors equal $h$, and the match rate on the $h \neq i$ subset is 3.11–25.59%. For Sol ZH and Kimi ZH, none of the 235 and 114 out-of-range answers, respectively, equals $h$. The formula therefore does not reproduce these observed range violations. This is not a universal absence of out-of-range matches, as Gemma EN and ZH have 1/2 and 6/24, respectively. These are exact-match diagnostics without a chance-match null model; no adoption probability or statistical enrichment above chance is inferred. A conditional match rate is undefined when its denominator is zero, including the out-of-range subset for Sol EN.

Table A24: **Exact fixed-3 conversion reproduces a minority of L3 errors.** Model abbreviations are as in Table A22. All gives $P(\hat{i} = h)$ in percent. Match/errors gives exact matches among incorrect parseable responses; Errors gives the corresponding percentage and 95% CI. Identifying gives $P(\hat{i} = h \mid h \neq i)$ in percent with its 95% CI. Range gives exact matches among out-of-range responses as a count over the number of such responses; 0/0 indicates an empty subset, not a zero conditional rate. Intervals use 5000 base-sequence block-bootstrap resamples. The all-response rate includes trivial matches where $h = i$ and the answer is correct.

| Pair | All (%) | Match/errors | Errors (%) [95% CI] | Identifying (%) [95% CI] | Range |
|---|---|---|---|---|---|
| Sol EN | 46.99 | 41/132 | 31.06 [24.11, 37.99] | 18.64 [13.84, 23.72] | 0/0 |
| Sol ZH | 13.98 | 23/960 | 2.40 [1.48, 3.32] | 3.11 [1.92, 4.41] | 0/235 |
| Kimi EN | 39.35 | 87/392 | 22.19 [18.62, 25.91] | 25.59 [21.43, 30.03] | 0/27 |
| Kimi ZH | 23.33 | 102/847 | 12.04 [9.46, 14.85] | 13.42 [10.54, 16.56] | 0/114 |
| Gemma EN | 38.12 | 65/330 | 19.70 [14.63, 24.77] | 19.12 [14.06, 24.38] | 1/2 |
| Gemma ZH | 28.33 | 142/750 | 18.93 [15.89, 21.97] | 18.68 [15.66, 21.77] | 6/24 |

**The diagnostic narrows the explanation without identifying an internal algorithm.** Together, the regression and exact-match results show that the boundary-associated coordinate influence cannot be fully explained by simply applying $3(g-1)+s$ to the target's true coordinates. They do not exclude occasional use of this rule, noisy coordinate conversion, or other forms of hierarchical readout. The supported conclusion remains that visible boundaries increase the influence of structure-relative coordinates on absolute reports; establishing the underlying computation requires additional causal or activation-level evidence.

### F.8 Variable-Length Sensitivity to Response Range and Layout Pattern

The primary variable-length estimand retains every parseable integer response, including values outside $1, \ldots, N$, because signed magnitude is part of the numerical-error measure. We assess sensitivity to this choice using 2 alternatives. The in-range-only policy retains only responses within $1, \ldots, N$ and, for matched contrasts, retains a trial only when both condition responses are eligible. The clip-to-range policy instead retains every parseable integer but clips it to $1, \ldots, N$ before signed error is computed. These policies are used only for sensitivity analysis. The raw definition remains the primary estimand.

Out-of-range responses are unevenly distributed across pairs. Under L3, the raw counts are 0/432 for GPT-5.6 Sol EN, 235/1152 for GPT-5.6 Sol ZH, 27/648 for Kimi-K3 EN, 114/1080 for Kimi-K3 ZH, 2/648 for Gemma-4-31B-IT EN, and 24/1080 for Gemma-4-31B-IT ZH. Despite these differences, the global ordinal association measured by Spearman's $\rho$ changes little across response policies. L3 Spearman $\rho$ remains 0.991 under all 3 policies for Sol EN, ranges from 0.978 to 0.982 for Sol ZH, from 0.979 to 0.980 for Kimi EN, from 0.983 to 0.984 for Kimi ZH, remains 0.991 for Gemma EN, and ranges from 0.987 to 0.988 for Gemma ZH.

The L3 condition-level group-influence estimate is also directionally stable. Across raw, in-range-only, and clip-to-range policies, $\beta_g$ is 0.143 for Sol EN, 1.426–1.957 for Sol ZH, 1.114 for Kimi EN, 0.904–1.305 for Kimi ZH, 0.750 for Gemma EN, and 0.665 for Gemma ZH. Its 95% intervals remain above zero under all 3 policies in the latter 5 pairs, whereas Sol EN remains interval-indeterminate. L3 slot influence remains interval-positive under all policies for Sol EN, Sol ZH, Kimi EN, and Kimi ZH; Gemma EN remains near zero with a slightly negative point estimate, and Gemma ZH remains positive in point estimate but interval-indeterminate.

Table A25 reports the matched L3−L1 coordinate-influence contrasts under the 3 policies. The group-influence contrast remains interval-positive under all policies for Sol ZH, Kimi EN, Kimi ZH,

Table A25: **Variable-length response-range sensitivity for the L3−L1 coordinate-influence contrast.** Raw is the primary policy and retains every parseable integer. In-range retains only responses in $1, \ldots, N$ and uses the intersection of eligible L1 and L3 trials; clipped retains all parseable integers after clipping them to $1, \ldots, N$. L3/pair gives the number of L3 responses retained by the policy and the number of matched L3−L1 responses used for the contrast. Spearman $\rho$ is the L3 correlation between true and reported absolute positions. $\Delta\beta_g$ and $\Delta\beta_s$ are the matched L3−L1 changes in residualized group and slot influence. Brackets give 95% CIs from 5000 base-sequence block-bootstrap resamples over 48 bases. The Raw rows reuse the primary L3−L1 estimates and intervals reported in Table A21; the alternative response-range policies are recomputed as sensitivity analyses.

| Model | Lang. | Policy | L3/pair | $\rho$ | $\Delta\beta_g$ [95% CI] | $\Delta\beta_s$ [95% CI] |
|---|---|---|---|---|---|---|
| GPT-5.6 Sol | EN | Raw | 432/432 | 0.991 | +0.071 [-0.580, +0.731] | +0.136 [+0.006, +0.278] |
| GPT-5.6 Sol | EN | In-range | 432/431 | 0.991 | +0.071 [-0.569, +0.687] | +0.137 [+0.006, +0.287] |
| GPT-5.6 Sol | EN | Clipped | 432/432 | 0.991 | +0.071 [-0.567, +0.711] | +0.137 [+0.009, +0.276] |
| GPT-5.6 Sol | ZH | Raw | 1152/1152 | 0.982 | +2.041 [+1.069, +3.031] | +0.798 [+0.477, +1.090] |
| GPT-5.6 Sol | ZH | In-range | 917/917 | 0.981 | +1.849 [+0.831, +3.083] | +0.590 [+0.242, +0.898] |
| GPT-5.6 Sol | ZH | Clipped | 1152/1152 | 0.978 | +1.509 [+0.722, +2.364] | +0.452 [+0.178, +0.674] |
| Kimi-K3 | EN | Raw | 648/648 | 0.980 | +1.082 [+0.572, +1.617] | +0.447 [+0.250, +0.641] |
| Kimi-K3 | EN | In-range | 621/619 | 0.979 | +1.082 [+0.592, +1.596] | +0.279 [+0.101, +0.452] |
| Kimi-K3 | EN | Clipped | 648/648 | 0.980 | +1.082 [+0.594, +1.606] | +0.351 [+0.185, +0.517] |
| Kimi-K3 | ZH | Raw | 1080/1080 | 0.984 | +1.170 [+0.522, +1.847] | +0.605 [+0.288, +0.945] |
| Kimi-K3 | ZH | In-range | 966/954 | 0.983 | +0.760 [+0.233, +1.312] | +0.212 [-0.053, +0.492] |
| Kimi-K3 | ZH | Clipped | 1080/1080 | 0.984 | +0.982 [+0.419, +1.563] | +0.279 [+0.038, +0.522] |
| Gemma-4-31B-IT | EN | Raw | 648/648 | 0.991 | +0.571 [+0.243, +0.912] | -0.063 [-0.208, +0.096] |
| Gemma-4-31B-IT | EN | In-range | 646/643 | 0.991 | +0.571 [+0.244, +0.903] | -0.064 [-0.204, +0.099] |
| Gemma-4-31B-IT | EN | Clipped | 648/648 | 0.991 | +0.571 [+0.239, +0.904] | -0.066 [-0.206, +0.098] |
| Gemma-4-31B-IT | ZH | Raw | 1080/1080 | 0.988 | +0.761 [+0.199, +1.314] | +0.247 [-0.049, +0.574] |
| Gemma-4-31B-IT | ZH | In-range | 1056/1056 | 0.987 | +0.761 [+0.201, +1.325] | +0.168 [-0.112, +0.467] |
| Gemma-4-31B-IT | ZH | Clipped | 1080/1080 | 0.988 | +0.761 [+0.191, +1.312] | +0.219 [-0.071, +0.523] |

Gemma EN, and Gemma ZH, while Sol EN remains near zero. The slot-influence contrast remains interval-positive under all policies for Sol EN, Sol ZH, and Kimi EN. Kimi ZH is the only pair whose slot result changes with range handling. The raw and clipped intervals lie above zero, whereas the in-range-only interval includes zero. Gemma EN and Gemma ZH remain interval-indeterminate under all 3 policies. Thus, range handling changes some effect magnitudes, particularly in the 2 ZH cells with the most out-of-range responses, but it does not account for the consistently high global ordinal association or the recurring group-coordinate influence.

We next remove each of the 6 group-length patterns in turn and recompute the raw-policy L1−L0 and L3−L1 coordinate-influence contrasts. Table A26 reports the full estimate together with the minimum and maximum over the 6 omissions. For the main L3−L1 contrast, $\Delta\beta_g$ remains positive after every omission in 5/6 pairs. Sol EN is the exception, and its full estimate is already near zero. Likewise, $\Delta\beta_s$ remains positive after every omission in 5/6 pairs. Gemma EN is the exception, with a slightly negative and interval-indeterminate full slot estimate. The recurring variable-length coordinate effects are therefore not attributable to any single group-length pattern.

# G Robustness Experiments

## G.1 Boundary-Surface Robustness

The main boundary manipulation uses ordinary 3-space item gaps and 6-space gaps at aligned word boundaries. To test whether the observed boundary effects depend on this particular whitespace rendering, we repeat the fixed-3 L3 manipulation with 3 alternative boundary surfaces while holding the lexical materials, target schedule, model configuration, coordinate definitions, and L1 baseline fixed. The robustness evaluation uses GPT-5.6 Sol and Gemma-4-31B-IT in EN and ZH at their formal anchor loads: Sol uses $N = 18$ in EN and $N = 48$ in ZH, whereas Gemma uses $N = 27$ in EN and $N = 45$ in ZH.

For each model–language pair, the canonical 6-space L3 condition reuses the completed formal run. The 3 new surfaces are a 9-space boundary gap, a vertical-bar separator rendered as `|`, and a hash

Table A26: **Leave-one-pattern-out sensitivity of variable-length coordinate influence.** Each row gives the raw-policy full estimate, the minimum and maximum after omitting each of the 6 group-length patterns in turn, and the signs of the 6 omission estimates. The notation $6+/0-$ means that all 6 leave-one-pattern-out estimates are positive. No additional bootstrap intervals are attached to individual omissions; the analysis tests dependence on a single layout pattern.

| Model | Lang. | Contrast | $\beta_g$: full; omission range; signs | $\beta_s$: full; omission range; signs |
|---|---|---|---|---|
| GPT-5.6 Sol | EN | L1−L0 | -0.071; [-0.375, +0.167]; 1+/5- | -0.007; [-0.032, +0.009]; 2+/4- |
| GPT-5.6 Sol | EN | L3−L1 | +0.071; [-0.500, +1.000]; 4+/2- | +0.136; [+0.075, +0.273]; 6+/0- |
| GPT-5.6 Sol | ZH | L1−L0 | -0.082; [-0.144, +0.036]; 1+/5- | -0.015; [-0.023, +0.008]; 1+/5- |
| GPT-5.6 Sol | ZH | L3−L1 | +2.041; [+1.073, +3.103]; 6+/0- | +0.798; [+0.603, +0.913]; 6+/0- |
| Kimi-K3 | EN | L1−L0 | +0.049; [+0.009, +0.097]; 6+/0- | +0.033; [+0.014, +0.066]; 6+/0- |
| Kimi-K3 | EN | L3−L1 | +1.082; [+0.890, +1.250]; 6+/0- | +0.447; [+0.291, +0.535]; 6+/0- |
| Kimi-K3 | ZH | L1−L0 | -0.039; [-0.143, +0.065]; 2+/4- | +0.051; [+0.003, +0.080]; 6+/0- |
| Kimi-K3 | ZH | L3−L1 | +1.170; [+1.071, +1.545]; 6+/0- | +0.605; [+0.482, +0.754]; 6+/0- |
| Gemma-4-31B-IT | EN | L1−L0 | +0.179; [+0.140, +0.240]; 6+/0- | +0.030; [+0.007, +0.054]; 6+/0- |
| Gemma-4-31B-IT | EN | L3−L1 | +0.571; [+0.430, +0.796]; 6+/0- | -0.063; [-0.118, +0.081]; 1+/5- |
| Gemma-4-31B-IT | ZH | L1−L0 | +0.024; [-0.012, +0.043]; 4+/2- | +0.033; [+0.001, +0.047]; 6+/0- |
| Gemma-4-31B-IT | ZH | L3−L1 | +0.761; [+0.206, +1.112]; 6+/0- | +0.247; [+0.115, +0.440]; 6+/0- |

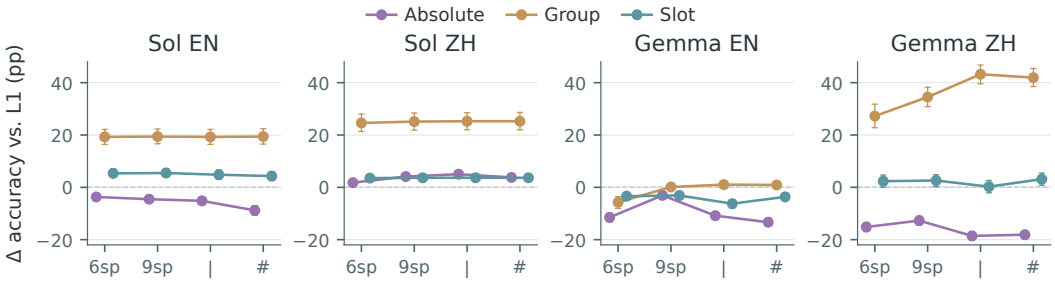



Figure A1: **Boundary-surface robustness of coordinate accuracy.** Panels show, from left to right, Sol EN, Sol ZH, Gemma EN, and Gemma ZH. The horizontal axis compares 4 L3 boundary surfaces: the canonical 6-space rendering reused from the formal L3 condition, a 9-space rendering, a vertical-bar separator (|), and a hash separator (#). The vertical axis gives the paired L3−L1 change in exact-match accuracy for absolute-, group-, and within-group slot-position probes. Error bars show 95% block-bootstrap CIs over base sequences.

separator rendered as #. Ordinary within-group gaps remain 3 spaces in every condition. Each new surface reuses the exact 256 L1 base sequences and the same target schedule for absolute, group, and within-group slot probes. The test sequence and the independent demonstration are surface-matched. For the symbol conditions, the encoding text explicitly identifies the displayed symbol as the boundary cue, and the absolute-position demonstration states that the symbol does not add an item to the sequence.

We compare each L3 boundary surface with the same L1 baseline using exact-match coordinate accuracy and the fixed-3 CE/HE coefficients defined in Section E.3. Uncertainty is estimated from 5000 base-sequence block-bootstrap resamples using the same matched resampling scheme as in the main analyses. Each newly run L3 surface is also checked with the surface-validity audit used by the analysis pipeline. The audit requires completion, parse, and in-range rates of at least 0.98. It also checks for a cumulative boundary-count offset, defined as an absolute slope of at least 0.10 positions per preceding displayed boundary with a 95% CI excluding zero. The audit is based on response validity and cumulative offset, where exact-match accuracy and the CE/HE coefficients are evaluated separately.

All 12 newly run L3 cells pass the surface-validity audit. As shown in Fig. A1, the exact-match effects differ across model–language pairs. For Sol-EN, group-position accuracy increases by 19.3–19.4 pp and slot-position accuracy by 4.3–5.5 pp across the 4 surfaces. For Sol-ZH, the corresponding increases are 24.6–25.3 pp and 3.5–3.6 pp. Gemma-EN shows a different pattern. Group-position changes range from $-5.7$ to $+1.0$ pp, while slot-position accuracy decreases by 3.1–6.3 pp for every surface. Gemma-ZH shows group-position increases of 27.2–43.2 pp and smaller slot-position increases of 0.3–3.1 pp. Absolute-position accuracy decreases for Sol-EN and both Gemma

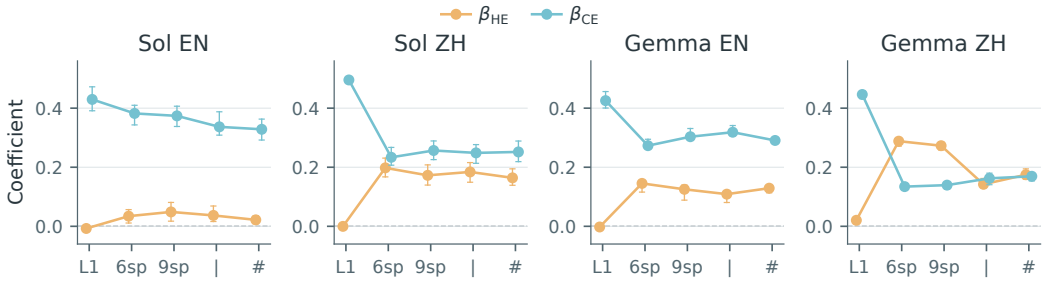



Figure A2: **Boundary-surface robustness of absolute-error geometry.** Panels show, from left to right, Sol EN, Sol ZH, Gemma EN, and Gemma ZH. The horizontal axis includes the L1 baseline together with 4 L3 boundary surfaces: the canonical 6-space rendering, a 9-space rendering, a vertical-bar separator (|), and a hash separator (#). The vertical axis shows the fixed-3 error-geometry coefficients $\beta_{\mathrm{HE}}$ and $\beta_{\mathrm{CE}}$ for absolute-position reports. Error bars show 95% block-bootstrap CIs over base sequences.

pairs, whereas it increases for Sol-ZH across all 4 surfaces. The exact-match effects are therefore model–language dependent.

The absolute-error geometry is more consistent across surfaces. As shown in Fig. A2, every L3 boundary surface lowers $\beta_{\mathrm{CE}}$ and raises $\beta_{\mathrm{HE}}$ relative to L1 in all 4 model–language pairs. This corresponds to a shift away from adjacent-position errors and toward the fixed-3 period-3 HE pattern. Specifically, $\beta_{\mathrm{HE}}$ rises from $-0.007$ under L1 to 0.022–0.049 for Sol-EN, from approximately 0 to 0.164–0.198 for Sol-ZH, from $-0.002$ to 0.109–0.145 for Gemma-EN, and from 0.020 to 0.142–0.288 for Gemma-ZH. The boundary-associated shift in absolute-error geometry therefore recurs across all 4 tested boundary renderings.

### G.2 Sequence-Length Robustness

The main experiments use one capability-matched fixed-3 load for each model–language pair. We test whether the L0–L1 pattern depends on that single load using a frozen 3-point grid while keeping the lexical materials, prompts, generation settings, and scoring rules unchanged. The robustness evaluation uses GPT-5.6 Luna and Gemma-4-31B-IT. EN uses $N \in \{18, 21, 27\}$ and ZH uses $N \in \{33, 45, 48\}$. Luna reuses its formal anchors at EN $N = 21$ and ZH $N = 33$, whereas Gemma reuses its formal anchors at EN $N = 27$ and ZH $N = 45$. The remaining points are additional runs on the same frozen fixed-3 banks.

Each load uses 256 base sequences, with L0 and L1 exactly matched on base identity and target schedule. For EN $N = 18, 21, 27$, the absolute probe densities are 18, 21, and 21 targets per base sequence; for ZH $N = 33, 45, 48$, they are 24, 27, and 21. Group and within-group slot probes use 3 targets each per base sequence. The grid and probe densities were fixed before inspecting the robustness outcomes. We report exact-match accuracy, matched L1$-$L0 accuracy changes, and the fixed-3 CE/HE coefficients. As in the main analyses, uncertainty is estimated from 5000 base-sequence block-bootstrap resamples.

Table A27: **Structure-relative effects persist across sequence lengths in 3 of the 4 model–language pairs.** Entries give the range of the matched L1$-$L0 exact-match accuracy change across the 3 frozen $N$ values, in pp. These are ranges of point estimates across loads rather than CIs; load-specific 95% CIs are shown in Fig. A3. Positive values indicate higher accuracy under L1.

| Model–lang. | Absolute | Group | Slot |
|---|---|---|---|
| Luna-EN | $[-1.15, +1.56]$ | $[+3.52, +9.11]$ | $[+5.21, +8.33]$ |
| Luna-ZH | $[-12.57, -9.34]$ | $[+7.68, +15.10]$ | $[+11.72, +14.71]$ |
| Gemma-EN | $[-11.79, -5.51]$ | $[+50.78, +59.90]$ | $[+15.49, +28.91]$ |
| Gemma-ZH | $[-12.59, -8.45]$ | $[-7.29, -3.91]$ | $[-0.91, -0.13]$ |

The structure-relative effects are consistent across the 3 loads for Luna-EN, Luna-ZH, and Gemma-EN (Fig. A3 and Table A27). For Luna-EN, group-position accuracy increases by 3.52–9.11 pp and slot-position accuracy by 5.21–8.33 pp across the grid, with all corresponding 95% intervals above zero. The absolute-position change is smaller, at $+1.56$ pp for $N = 18$, $+1.30$ pp for $N = 21$, and

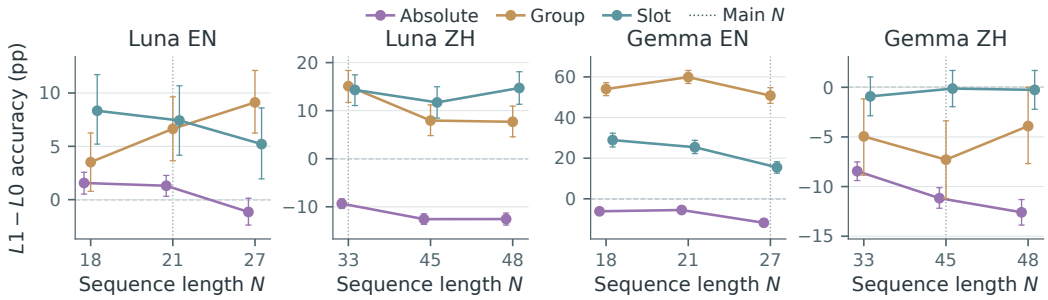



Figure A3: **Sequence-length robustness of lexical-cue effects on accuracy.** Each panel shows one model–language pair across the frozen 3-point load grid. Colored series give the matched L1 minus L0 change in exact-match accuracy for absolute-position, group-position, and within-group slot probes. Positive values indicate higher accuracy under L1. Whiskers show 95% CIs from 5000 base-sequence block-bootstrap resamples. The vertical dotted line marks the capability-matched main-experiment load used in the main paper. Each panel uses its own y-axis range to preserve smaller effects in Luna-EN and Gemma-ZH.

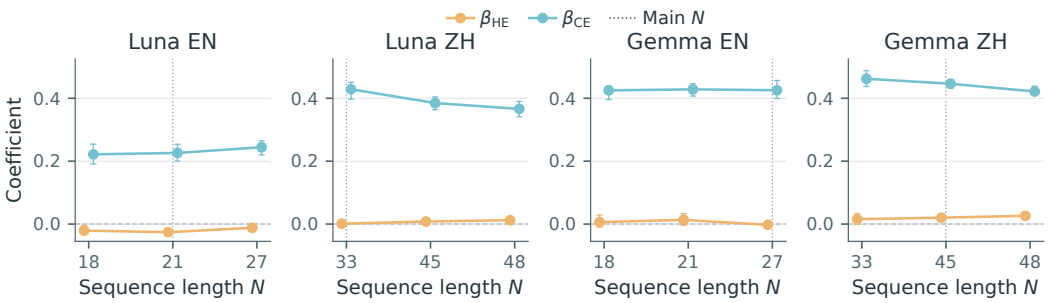



Figure A4: **Sequence-length robustness of L1 absolute-error geometry.** Each panel shows one model–language pair across the same frozen 3-point load grid. The 2 series give the fixed-3 absolute-error coefficients under L1: $\beta_{\mathrm{CE}}$ quantifies adjacent-position error structure, whereas $\beta_{\mathrm{HE}}$ quantifies the fixed-3 period-3 HE pattern. Whiskers show 95% CIs from 5000 base-sequence block-bootstrap resamples. The vertical dotted line marks the capability-matched main-experiment load used in the main paper. All panels share the same y-axis range.

$-1.15$ pp for $N = 27$. The interval at $N = 27$ includes zero. Luna-ZH shows the retrieval trade-off at all 3 loads. Absolute-position accuracy decreases by 9.34–12.57 pp, while group-position accuracy increases by 7.68–15.10 pp and slot-position accuracy by 11.72–14.71 pp. All corresponding intervals remain on the same side of zero.

Gemma-EN shows the same qualitative trade-off with substantially larger structure-relative changes. Across $N = 18, 21, 27$, absolute-position accuracy decreases by 5.51–11.79 pp, group-position accuracy increases by 50.78–59.90 pp, and slot-position accuracy increases by 15.49–28.91 pp. Thus, the large group and slot gains observed at the main-experiment load $N = 27$ are also present at both neighboring robustness loads.

Gemma-ZH remains the exception observed in the main experiment rather than changing qualitatively at another load. Across the grid, absolute-position accuracy decreases by 8.45–12.59 pp, group-position accuracy decreases by 3.91–7.29 pp, and slot-position changes remain close to zero, from $-0.91$ to $-0.13$ pp, with all 3 intervals including zero. The load sweep therefore reproduces the model–language heterogeneity observed at the main-experiment loads.

The absolute-error geometry remains CE-dominant across the load grid (Fig. A4). Under L1, $\beta_{\mathrm{CE}}$ is positive in every cell and ranges from 0.222 to 0.462, whereas $\beta_{\mathrm{HE}}$ remains much smaller, ranging from $-0.026$ to 0.026. Luna-EN has a small negative $\beta_{\mathrm{HE}}$ at all 3 loads. Gemma-EN remains close to zero, while Luna-ZH and Gemma-ZH show small positive $\beta_{\mathrm{HE}}$ values at the larger loads.

## H MODEL RESPONSE EXAMPLES

The examples below use exact recorded model outputs from the reported runs. Source-word annotations are added only to make the intended grouping readable; they are not additional information supplied to the model in conditions where the grouping is unmarked. Most examples are errors because they illustrate the response patterns quantified by the metrics above.

**Example 1: Correct absolute-position report.**

| | |
|---|---|
| **Question** | GPT-5.6 Sol, EN, fixed-3 L3. Source words: `WONDER / NATIVE / PURPLE / ZOMBIE / TONGUE / PALACE`. Target: `ND`. Report its absolute position. |
| **Model answer** | `2` |
| **Analysis** | The gold absolute position is 2 (group 1, slot 2). The response is parsed as 2 and scored correct. |

**Example 2: Incorrect group report without a lexical cue.**

| | |
|---|---|
| **Question** | Gemma-4-31B-IT, EN, fixed-3 L0. Source words: `GENTLE / SUMMER / POLICY / SOVIET / ANCHOR / SPIRIT / HIDDEN / OFFICE / SALARY`. Target: `VI`. Report its group position. |
| **Model answer** | `3` |
| **Analysis** | The intended group position is 4. The response is a valid parsed integer but is scored as an incorrect group report. |

**Example 3: Adjacent continuous absolute error.**

| | |
|---|---|
| **Question** | Kimi-K3, EN, fixed-3 L3. Source words: `GENTLE / SUMMER / POLICY / SOVIET / ANCHOR / SPIRIT / HIDDEN / OFFICE / SALARY`. Target: `SU`. Report its absolute position. |
| **Model answer** | `5` |
| **Analysis** | The gold absolute position is 4. The $+1$ displacement is an adjacent-position error and contributes to the CE component. |

**Example 4: Fixed-3 same-slot adjacent-group error.**

| | |
|---|---|
| **Question** | Kimi-K3, EN, fixed-3 L3. Source words: `FACIAL / PRIMER / ARRIVE / TREATY / SYMBOL / SECOND / MOTION / INTAKE / ABSORB`. Target: `VE`. Report its absolute position. |
| **Model answer** | `6` |
| **Analysis** | The target is at absolute position 9, group 3, slot 3. Position 6 is group 2, slot 3, so the report preserves the within-group slot while moving to the adjacent group; in the fixed-3 layout this gives lag $-3$. |

**Example 5: Variable-length same-slot error with a lag other than 3.**

| | |
|---|---|
| **Question** | Kimi-K3, EN, variable-length L3. Source words: `NATIVE / MINK / LANE / UNPACKED / LEGS / RETAIN / AVOIDING / CARRIAGE / QUARTZ`; group lengths $[3, 2, 2, 4, 2, 3, 4, 4, 3]$. Target: `GS`. Report its absolute position. |
| **Model answer** | `15` |
| **Analysis** | The target is at absolute position 13, group 5, slot 2. Position 15 is group 6, slot 2, giving a same-slot adjacent-group error with lag $+2$. This separates the same-slot relation from a fixed lag of 3. |

**Example 6: Out-of-range absolute report.**

| | |
|---|---|
| **Question** | GPT-5.6 Sol, ZH, variable-length L3, with $N = 48$. Source words: 苦行僧 / 智囊团 / 贝壳 / 吸血鬼 / 蒙汗药 / 银河系 / 座右铭 / 海市蜃楼 / 参谋长 / 萌芽 / 好奇心 / 望尘莫及 / 千里马 / 稻草人 / 牛仔裤 / 摇篮曲. Target: 马. Report its absolute position. |
| **Model answer** | `49` |
| **Analysis** | The gold position is 39 (group 13, slot 3). The parsed answer 49 is out of range and therefore incorrect for exact-match accuracy, while the structure-relative influence analysis retains its signed error of $+10$. |