

Dissecting Lie Expectancy in Language Models

Xingyuan Zhao and Seana Coulson

University of California, San Diego

miozhao@ucsd.edu and scoulson@ucsd.edu

Abstract

When the speaker has a motive to lie — profit, face-saving, fear of punishment — an informed reader expects the lie. Forming this expectation requires three distinct abilities: representing what is actually true in the described situation, aligning linguistic expectations with the speaker’s incentives to be truthful, and adapting those expectations across situations. We study whether twelve open Qwen3 and OLMo2 models possess these abilities, asking how each is organized, whether they are separable, and when each is acquired. *Within each model*, fact information is linearly decodable at mid depth but becomes directly expressed only in the final third of the network; motive recognition emerges slightly later, and both grow with scale. Two-layer ablations dissociate them: removing a mid-depth band of layers weakens fact expression and motive recognition together. Ablating a deeper pair of layers, in contrast, leaves both essentially intact while eliminating human alignment: sensitivity to motives and human-like sensitivity are separable abilities. Across 19 public checkpoints of OLMo2-7B, fact representation and motive recognition become detectable at around 0.3T tokens and grow steadily throughout pretraining, while human alignment first becomes detectable at mid-training — a brief continued training on curated data. Altogether, representing what is true and recognizing whether a speaker is motivated to lie are closely coordinated; matching human judgments about when lying is likely is a later and partially separable achievement.

1 Introduction

Consider the following scenario: *Julie is a receptionist in a dirty hotel. A young mother comes in with her two children looking for a place to stay. Julie is eager to rent out rooms to whomever she can. Julie tells the visitor the rooms are ____*. Presented with this scenario, a typical human reader

Shared context: Julie is a receptionist in a **dirty** hotel. A young mother comes in with her two children looking for a place to stay for the night. Being somewhat of a germaphobe, she asks Julie how sanitary the rooms are. ↑ situational fact

lie-motive version

As the hotel’s business has been rather slow, Julie is eager to rent out rooms to whomever she can, no matter what.

truth-motive version

The hotel chain runs frequent mystery-shopper audits and dismisses staff caught misrepresenting the rooms, and Julie knows this stay will be checked, so she has every reason to describe the rooms plainly.

Completion: *Julie tells the visitor the rooms are ____* candidates: “clean” / “dirty”

expected: **“clean”** (lie)

expected: **“dirty”** (truth)

Figure 1: Representative stimulus pair (pair 10A, full text). In both versions, the completion of the sentence frame *The rooms are*, “dirty” is the truth word and “clean” is the lie word. The two contexts differ only in the speaker’s motive, and consequently flip which word is expected: under a lie motive the expected continuation is the lie word; under a truth motive it is the truth word. The within-pair difference in the model’s preference therefore isolates the motive (Δ_{motive}).

both understands that the rooms in this hotel are dirty, (a fact based on situational knowledge), and understands that Julie has a reason to say otherwise (motivation to lie). Consequently, readers’ lexical expectations here (i.e., whether they expect “clean” or “dirty” next) are graded rather than binary, a function of the incentive, the stakes, and social convention, mirroring the graded way speakers’ own truth-telling tracks incentives (Gneezy, 2005; Abeler et al., 2019). Such judgments are consistent enough across readers to be measured: previous research has shown how scenarios used in the present study elicit graded expectations H for how likely the speaker is to lie, collected from human continuations of these stories in a norming study (Zhao and Coulson, 2026)(Figure 1).

For a language model, completing this sentence

involves three different abilities: (i) **fact appreciation**: representing which continuation is true in the described world; (ii) **motive recognition**: shifting lexical expectations in response to the speaker’s incentives, viz., toward the lie word in texts that convey a lying motive, and toward the truth word in the truth-motive version; (iii) **human alignment**: grading the resulting lie expectations in line with human social-pragmatic judgments of when lying is likely. Prior work has probed LLMs for the truth of statements made by generative AI systems (Burns et al., 2023; Azaria and Mitchell, 2023; Marks and Tegmark, 2024; Orgad et al., 2025), and documented deceptive behavior in goal-driven settings (Hagendorff, 2024; Scheurer et al., 2023). However, the way in which LLMs represent human lies in texts and discourse remains largely unexplored. Here we investigate lie *expectancy* in LLMs with a focus on how a language model can combine sensitivity to situational facts with the recognition of factors that influence speakers’ motivation to lie.

We ask three questions about these abilities:

RQ1 (organization). How are fact representation and motive recognition organized across model depth? Specifically, where in the network does fact information become linearly decodable, where do the fact and the motive begin to be expressed in the model’s intermediate word predictions, and do these two signals emerge in a consistent depth order?

RQ2 (separability). Are these three abilities causally separable — can an intervention remove one while sparing the others? In particular, is human alignment a distinct ability, or is it a natural byproduct of the computations needed for fact representation and motive recognition?

RQ3 (acquisition). When does each ability arise in the course of training? That is, do these abilities all emerge together or do they have separate developmental trajectories?

This paper tracks the organization, separability, and acquisition of these three abilities and makes three contributions:

1. **Fact appreciation and motive recognition are closely coupled** (§4, RQ1): we measure separately where fact information is *present* in the hidden states (via a trained probe) and where it is *expressed* in the model’s own intermediate predictions (using a logit lens). In all twelve models, the fact is readable at mid

depth, but neither the fact nor the motive is expressed until the final third of the network. The fact tends to appear slightly before the motive — a gap of roughly two to four layers.

2. **Ablation studies dissociate human alignment from fact appreciation and motive recognition** (§5, RQ2): we mean-ablate two adjacent layers at a time and re-measure all three abilities. Ablating a mid-depth window weakens fact expression and motive recognition together; ablating a deeper window leaves both nearly intact yet abolishes human alignment — the ablated model still reacts to motives, but ranks scenarios unlike human readers. Sensitivity to motives and human-like sensitivity are thus two separate abilities.
3. **Two developmental timelines** (§6, RQ3): OLMo2-7B is trained in two phases: 3.9T tokens of pretraining, then a 50B-token mid-training pass on curated data with the learning rate annealed to zero. Fact representation and motive recognition become detectable at the same checkpoint (~0.3T tokens) and grow gradually throughout pretraining, while human alignment stays non-significant across all of pretraining and first appears during mid-training — a pattern suggestive of sensitivity to the late curated data or to the annealing.

2 Related Work

Truth probing. Linear structure for statement truth in LM activations is well documented (Burns et al., 2023; Azaria and Mitchell, 2023; Marks and Tegmark, 2024). Orgad et al. (2025) add two lessons: probes recover the most when they read the *exact answer token*, and a model’s internal encoding can contain more truth information than its output expresses. Our design adopts both: we probe the hidden state at the critical word (“clean” vs. “dirty” in Figure 1) to ask where in the model fact information is *present*, and we read the model’s intermediate predictions with the logit lens to ask where that information is *expressed* (§4, Appendix A). Previous literature mainly probes isolated factual statements; here we explore what happens when the assertion comes from a speaker with a motive to lie — asking whether factual information is available in the

model’s internal encoding and if it still shows in the model’s expectations.

Deception in LMs. Goal-driven deceptive behavior (Scheurer et al., 2023; Hagendorff, 2024), lie detection (Pacchiardi et al., 2024), and honesty benchmarks (Ren et al., 2025) study the model as a (potential) liar. By contrast, we study how LLMs incorporate knowledge of human behavior, especially with respect to honest versus dishonest communication, to predict what speakers in various scenarios will say. We treat lie expectancy as a side effect of language comprehension ability that is anchored to human norms. Moreover, our questions are mechanistic, asking where each component of the expectation lives in the network, whether these components are causally separable, and when each arises during training.

Causal interventions. Our ablation design descends from activation patching and causal tracing (Vig et al., 2020; Meng et al., 2022; Zhang and Nanda, 2024); mean-ablation of layer contributions with a competence control and a pre-registered specificity criterion adapts these tools to benchmark-level questions. Single-layer interventions are known to be absorbed by downstream self-repair (McGrath et al., 2023), which motivates our two-layer windows.

Training trajectories and mid-training. Checkpoint ladders enable developmental analysis (OLMo Team et al., 2025). The mid-training/anneal phase is known to have an out-sized influence on benchmark performance (Hu et al., 2024; Blakeney et al., 2024; Grattafiori et al., 2024), but reported effects concern math, code, world knowledge and problem solving. To our knowledge, the development of a socially calibrated expectation like lie expectancy has not previously been explored.

3 Stimuli, Measures, and Models

Scenarios. The lie-expected (LE) scenarios are those from the norming study reported in Zhao and Coulson (2026), adapted in turn from an ongoing study of the human brain response to lies and truths in socially biased discourse. Each item is a short narrative that establishes a social situation in which a speaker has a clear motivation to lie — to avoid punishment, spare someone’s feelings, secure a material gain, or gain status in their peer

group — followed by a target sentence containing a *critical word* whose substitution determines whether the utterance is a lie or a truth relative to the scenario (Figure 1). Critical words occur in antonym pairs (e.g., *clean/dirty*), and scenarios are counterbalanced so that a word that produces a lie in one scenario produces a truth in another. Consequently, differences between conditions are not attributable to lexical properties of the critical words. The set comprises 76 scenarios.

Truth-motive twins. For each LE scenario we add a *truth-expected* (TE) twin (Figure 1), generated with a frontier LLM (Claude-fable-5) under instructions to keep the situational fact, the characters, and the target utterance identical and to replace only the incentive. Each twin was audited for *situational-fact constancy*. This audit involved an automated LLM-judge check that the rewrite did not change what is actually true in the story. The first author reviewed all flagged items from this audit and the 9 pairs that failed were discarded. The resulting stimulus set included **67 audited pairs**, and all twin-based measures in this paper employed this set. Because the twin holds the fact and the lexical material fixed, any LE–TE difference isolates the motive.

Human norms. For 40 of the LE scenarios, Zhao and Coulson (2026) collected free continuations from 44 participants. In that study, each participant read the scenario and the target sentence up to, but not including, the critical word, and filled in the last word. Completions were coded as lie, truth, or ambiguous; the norm for scenario i is the proportion of lie completions among the unambiguous ones,

$$H_i = \frac{n_i^L}{n_i^L + n_i^T}, \quad (1)$$

where n_i^L and n_i^T count the lie and truth completions (median 39 unambiguous completions per scenario, minimum 20).

The scenario effect All three are built from the model’s preference between the two candidate words at the critical-word position:

$$\delta = \log p(\text{lie word} \mid \text{scenario, stem}) - \log p(\text{truth word} \mid \text{scenario, stem}), \quad (2)$$

$$\delta^0 = \log p(\text{lie word} \mid \text{stem}) - \log p(\text{truth word} \mid \text{stem}), \quad (3)$$

where the *stem* is the target utterance up to the critical word (“Julie tells the visitor the rooms are”),

The **model lie preference** LP is how much the scenario shifts the preference beyond the words themselves:

$$LP = \delta - \delta^0, \quad (4)$$

and the **motive effect** Δ_{motive} is the within-pair difference:

$$\Delta_{\text{motive}} = LP_{\text{LE}} - LP_{\text{TE}}, \quad (5)$$

Motive effects are reported with pair-bootstrap 95% confidence intervals (1,000 resamples).

The remaining measures are introduced where they are first used: §4 tests whether the situational fact is *present* in the hidden state and whether it is *expressed* in the model’s output, and §5 defines human alignment, the agreement between the model’s ordering of the scenarios and the human norms H .

Models. Twelve models: Qwen3 {0.6B, 1.7B, 4B, 8B} $\times$ {base, post-trained} (Yang et al., 2025) and OLMo2 {1B, 7B} $\times$ {base, instruct} (OLMo Team et al., 2025); unsuffixed OLMo2 names denote the base models.

4 Model Representations of Facts and Lie Motives

RQ1: where is fact information present, where are the fact and the motive expressed in the model’s predictions, and in what order?

We answer at the level of the model’s next-word prediction at the blank — its preference between the truth word and the lie word (“clean” vs. “dirty”).

Figure 2 accordingly reports three measures per model: a trained probe asks whether the situational fact is *present* in the hidden state (row 1), the logit lens asks whether the fact (row 2) and the motive (row 3) are *expressed* in the output, respectively.

Row 1: is factual information present in model embeddings? A ridge probe decodes truth value from the hidden state at the critical word, scored by AUC. Because a probe of this kind needs items that differ in truth value, we supplemented the LE and TE stimuli with a companion set built from the same 76 source scenarios. In the companion set, each scenario was completed once with each candidate word: for the scenario in Figure 1, the

rooms are dirty is true and *the rooms are clean* is false. This gives one true and one false assertion per scenario ($76 \times 2 = 152$ in all). Cross-validation folds are formed by scenario, so that the two assertions from a single scenario never straddle the split. Significance is assessed against a null that flips the true/false assignment once per scenario (500 shuffles).

Row 2: is the fact expressed in model output probabilities? Here the input stops *before* the critical word, and what we read is the model’s own probability for each of the two candidates: the model does the judging, and no classifier is trained.

The frame has to change as well, because in the speaker’s own utterance a preference between the two words has two possible causes. A truth-ward preference may mean that the model has represented the situation correctly, or that it has read the speaker as motivated to be honest. A lie-ward preference may mean that the fact representation has failed, or that it is intact and the speaker’s incentive has overridden it. Within a pair the true word is fixed and only the incentive varies (Figure 1), so nothing in that comparison distinguishes the two. We therefore re-voice the target assertion with a neutral narrator, who has no stake in the outcome and therefore no reason to describe the situation as anything other than it is, for example “*It so happens the rooms are ____.*” Let c_i^{neu} denote scenario i followed by its neutral-narrator stem. For scenario i at layer L , we define

$$F_i^{(L)} = \log p_L(\text{fact}_i \mid c_i^{\text{neu}}) - \log p_L(\text{non-fact}_i \mid c_i^{\text{neu}}). \quad (6)$$

Because the target sentence no longer attributes the assertion to a motivated speaker, the sign of $F_i^{(L)}$ indicates whether the layer- L output favors the scenario-consistent continuation. Narrator re-voicings were generated per item with the same LLM pipeline as the TE twins and manually reviewed. Row 2 of Figure 2 reports the proportion of scenarios for which $F_i^{(L)} > 0$. The ablation analysis in §5 reuses the same item-level score but reports its mean magnitude in nats.

Row 3: does the motive alter the prediction? Here the input is the original dataset’s scenario and the speaker’s own utterance, again stopping before the critical word, and the readout is once

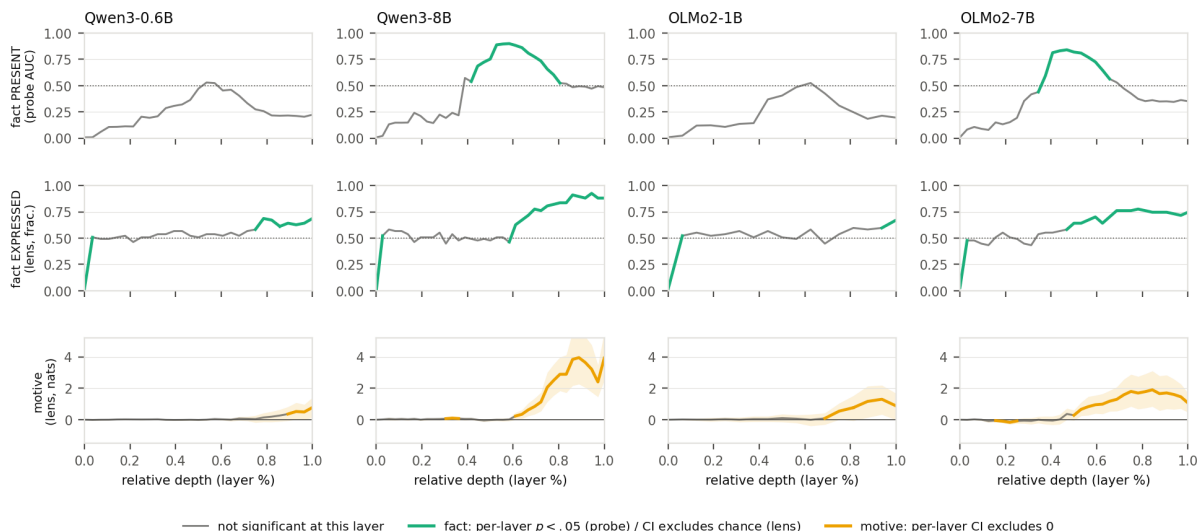


Figure 2: Three depth profiles for four representative models — the smallest and largest of each family; the full twelve-model grid is Figure 6 — on a shared relative-depth axis. **Row 1:** *fact presence*, ridge-probe AUC at the critical word with scenario-grouped cross-validation. **Row 2:** *fact expression*, the proportion of neutral-narrator items for which $F_i^{(L)} > 0$. **Row 3:** motive recognition $\Delta_{\text{motive}}(L)$, from the logit lens with pair-bootstrap 95% confidence bands. Rows 1–2 compare two readouts of the same scenario-relative truth information; rows 2–3 compare two abilities through the same readout. Colored segments mark uncorrected per-layer significance and are descriptive only.

again the model’s own probabilities. It is the contrast of §3 read in the speaker’s frame, with δ and δ^0 — hence LP_L and $\Delta_{\text{motive}}(L)$ — recomputed from the layer- L log-probabilities. A positive $\Delta_{\text{motive}}(L)$ means that by layer L the preference has moved toward the lie word specifically when the scenario gives the speaker a reason to lie, but not in the twin TE scenario. For this measure, we probe only the output layer, leaving hidden states unexamined. Because the two versions of a pair differ in wording, the raw condition decodability reflects the influence of lexical cues (Appendix A).

Fact information is present before it is expressed. In row 1–2 of Figure 2, the two output-derived curves remain near their baselines through the first $\sim 60\%$ of depth in every model: intermediate predictions are neither reliably directed toward the true continuation in the narrator frame nor shifted by the motive manipulation. Over the same layers, however, fact information was linearly decodable, with probe AUC peaking at relative depths .47–.58 and increasing with model size (.53 at 0.6B to .90 at 8B). Scenario-relative truth information is therefore readable from the hidden state well before it is directly expressed through the model’s output vocabulary. Probe accuracy de-

clines as the output-derived curves rise, but this opposing pattern does not by itself establish a representational handoff.

Fact expression rises before motive recognition.

In row 2–3 of Figure 2, beyond relative depth $\sim .6$, both output-derived measures rise through the final third of the network. We summarize their timing by the first layer at which each curve reaches half of its peak elevation above its respective baseline. In 9 of the 11 model instances for which both curves rise, the fact curve reaches this point before the motive curve; the remaining two are ties, with no reversals (median difference .08 of relative depth, approximately 2–4 layers). It suggests a small lag between fact expression and motive recognition as might be expected if the two sorts of information are computed serially. Because the two curves use different outcome scales and the model instances are taken from only two families, we interpret this effect with caution.

Both output effects increase with scale. Late-layer fact expression and motive recognition are detectable at every evaluated model size, and both generally increase in magnitude with parameter count. The observed pattern is graded rather than consistent with a sharp size threshold.

The depth profiles establish that fact informa-

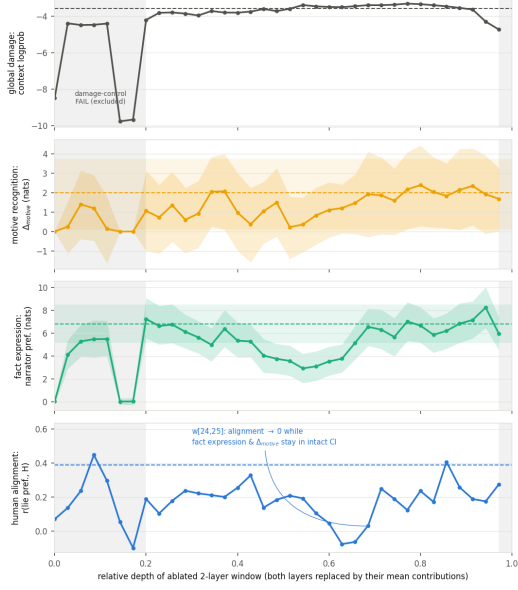


Figure 3: Sliding two-layer mean-ablation of Qwen3-8B (round 2). Rows: general-competence control (context log-probability), Δ_{motive} , fact expression (narrator-frame preference, nats), human alignment. Dashed = intact value; gray zones fail the damage control and are excluded. Claim-grade statistics are the paired bootstraps in the text.

tion becomes linearly decodable before fact expression and motive recognition emerge, but they do not establish causal dependence. Section 5 therefore tests both whether the mid-depth region contributes to fact expression and whether motive recognition depends on overlapping layers.

5 Ablation: Middle Layers Needed for Facts and Motives, Deeper Layers Needed for Human Alignment

RQ2: are the abilities causally separable — and is human alignment its own ability, or a byproduct?

We ablate layers of the model with the best performance in Figure 2 (Qwen3-8B) and re-measure all three quantities (Figure 3). Appendix C repeats the whole sweep on OLMo2-7B-Instruct.

Intervention. Layer L updates the residual stream as

$$h_{\text{out}}^{(L)} = h_{\text{in}}^{(L)} + f_L(h_{\text{in}}^{(L)}), \quad (7)$$

where f_L is the layer’s attention+MLP computation. From one intact pass over the stimulus set we cache each layer’s mean contribution

$$\mu_L = \mathbb{E}_{\text{tokens}} \left[f_L(h_{\text{in}}^{(L)}) \right], \quad (8)$$

and ablate by freezing the computation to that constant:

$$h_{\text{out}}^{(L)} = h_{\text{in}}^{(L)} + \mu_L. \quad (9)$$

Freezing a component at its dataset-mean output is a standard ablation in mechanistic interpretability (Wang et al., 2023).

Downstream layers run normally and we read *output logits only*. Probes are deliberately not used as ablation DVs: a probe trained on intact states is off-distribution on an ablated model, and decodability could survive ablations that destroy use (cf. Appendix A).

Measures. (i) *Fact expression*: the narrator preference of §4, read at the output layer without the stem-only correction, mean across items. (ii) *Human alignment*: the third ability of §1, defined here because this is where it is first needed. It asks whether the lie preferences order scenarios the way human readers do,

$$A = \rho(\text{LP}_{\text{LE},i}, H_i), \quad i = 1, \dots, 40, \quad (10)$$

Spearman’s rank correlation over the normed scenarios; the human norms were collected on the lie-motive versions, so A is defined on those. (iii) *Motive effect*: Δ_{motive} as defined in §3, recomputed from the ablated model’s output. (iv) *Sanity check*: every configuration logs the mean per-token log-probability of the scenario text itself; a benchmark drop only counts if this control stays at intact level. Windows containing L0 or L6 collapse it to ≈ -9.5 and are excluded.

Single-layer ablations. Single-layer ablations produced no reliable selective reduction in the three outcomes outside the globally damaging layers, consistent with distributed redundancy or downstream self-repair (McGrath et al., 2023). We therefore turn to sliding two-layer windows.

Sliding two-layer windows. To reduce the scope for compensation after single-layer ablation, we next ablate sliding windows $w[L, L+1]$. The twelve windows spanning w15–w26 were evaluated with paired-with-intact bootstraps over items, which reduce between-item variance. Benjamini–Hochberg correction was applied across the twelve windows separately for each outcome.

1. **Fact expression and motive recognition show overlapping mid-depth sensitivity.** In row 2–3 of Figure 3, ablating windows w15–w23 (rel. .43–.66) reduces fact expression by

2.0–4.6 nats relative to the intact value of 7.8 nats ($q = .001$ throughout); a smaller reduction of 1.3 nats also occurs at w[26,27]. Seven of the eight windows w15–w22 also reduce average motive recognition by 1.3–2.4 nats relative to the intact value of 3.8 nats ($q \leq .046$; w17 does not survive correction). The competence control remains within its prespecified acceptable range. This overlap is consistent with shared causal sensitivity or functional dependence, but does not establish a serial fact-to-motive computation.

2. **Deeper layers are important for human alignment.** In row 4 of Figure 3, human alignment decreases reliably across w20–w24 (ΔA up to .46, $q \leq .029$), overlapping the deeper portion of the fact-sensitive range. The most selective effect occurs at w[24,25], where alignment falls from +.39 to approximately zero ($\Delta A = .35$, 95% CI [.10, .59], $q = .021$), while substantial fact expression and average motive recognition remain. Their intact-normalized estimates are .96 (CI [.84, 1.08]) and 1.08 (CI [.81, 1.34]), respectively. Thus, w[24,25] provides the most selective dissociation among the tested windows, although equivalence to the intact model was not formally tested.

The alignment reduction reflects structured reordering. After w[24,25] ablation, the dispersion of LE lie preferences remains similar to intact (s.d. 7.5 $\rightarrow$ 8.0 nats), and their ordering remains substantially correlated with the intact ordering ($\rho = .73$), yet alignment with human norms falls from .39 to .03. In a post hoc matched-perturbation analysis, random rank perturbations calibrated to the same intact–ablated rank correlation ($\rho \approx .73$; 10,000 draws) yield a median surviving alignment of .29, with 95% of simulations in [.06, .49]; the observed .03 falls in the lowest 1.4% of this null distribution. Thus, under this null model, the observed reordering is more destructive of human alignment than random reordering of comparable magnitude.

Together, these results show that ablation of w[24,25] strongly reduces human-aligned scenario ordering while substantially preserving average fact expression and motive recognition.

We repeated the ablation on OLMo2-7B-Instruct and observed a similar pattern, see Appendix C and Figure 7.

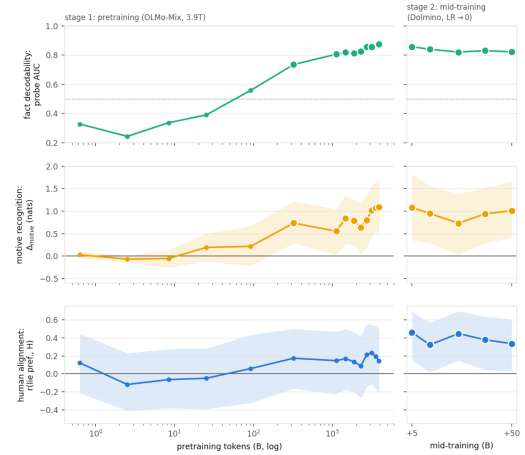


Figure 4: OLMo2-7B across 19 checkpoints (left: pre-training tokens, log scale; right: the 50B mid-training stage). Rows: fact decodability (trained probe, peak over layers), Δ_{motive} , human alignment. Large markers are significant (permutation $p < .05$ for fact; bootstrap CI excluding 0 otherwise; per checkpoint, uncorrected). The pretraining-vs-mid-training contrast for alignment is tested directly in the text.

6 Training: Pretraining Affects Facts and Motives; Mid-training Affects Human Alignment

RQ3: how do the three measured outcomes emerge over training?

OLMo2-7B is pretrained in two stages: a 3.9T-token pretraining stage on OLMo-Mix (stage 1), followed by a 50B-token *mid-training* stage (stage 2), in which the curated Dolmino mixture is consumed while the learning rate anneals linearly to zero (OLMo Team et al., 2025). We evaluate 14 stage-1 checkpoints and the five public checkpoints of the mid-training runs (Figure 4).

Fact decodability increases gradually during pretraining. In row 1 of Figure 4, peak probe AUC rises from below chance to .88 across pretraining. The first checkpoint with a pointwise permutation $p < .05$ occurs at 323B tokens, but the gradual trajectory and checkpoint spacing do not identify a discrete onset. A corresponding OLMo2-1B trajectory also shows a gradual increase in peak AUC, from .32 to .68 over approximately 4T tokens.

Motive recognition becomes detectable during pretraining. In row 2 of Figure 4, the confidence interval for Δ_{motive} includes zero through 92B tokens and first excludes zero at the 323B

checkpoint (+.73, 95% CI [.29, 1.22]). Its magnitude subsequently increases to +1.09 by the end of pretraining and remains positive at all five mid-training checkpoints. Thus, both fact decodability and motive recognition are detectable by 323B tokens and increase across pretraining, although the available checkpoints cannot resolve their relative onset within the 92B–323B interval.

Human alignment rises at the mid-training transition. Row 3 of Figure 4 does not follow the other two. Human alignment (A ; §3), computed over the 40 normed scenarios included in the checkpoint evaluation, is non-significant at all 14 pretraining checkpoints (estimates $-.12$ to $+.23$; final estimate $+.14$). At the first mid-training checkpoint, after 5B additional tokens, the estimate rises to $.46$ (95% CI [.16, .69]), and it is significant at every mid-training checkpoint. These per-checkpoint tests are uncorrected and the checkpoints are strongly related, so they do not by themselves establish a stage effect.

We therefore compare each mid-training checkpoint directly with the final pretraining checkpoint using paired scenario bootstraps. The first contrast is significant ($\Delta A = +.32$, 95% CI [.01, .66], $p = .039$ uncorrected); the pooled contrast across all five is not ($\Delta A = +.26$, 95% CI [-.02, .59], $p = .08$). We therefore treat the rise at the mid-training transition as suggestive rather than conclusive evidence of a change specific to mid-training.

Over the same interval, peak fact-probe AUC decreases slightly ($.88 \rightarrow .82$), while the point estimate of motive modulation changes comparatively little. This divergence is consistent with the intervention results of §5, but does not independently establish causal separability.

7 Discussion

Does a language model understand that speakers lie? Our results suggest splitting this question into two: whether the model has the machinery — representing what is true and responding to the speaker’s incentive — and whether the resulting expectations resemble those of human readers. The machinery is acquired early and broadly: through ordinary pretraining alone, models come to represent what is true in a described situation and to shift their expectations when the speaker has a reason to lie. Our study suggests that fact appreciation and motive recognition emerge together

in training, improve together with scale, and are represented in similar layers in the network. Moreover, the representation of facts and motives affect the model’s predictions in a consistent order, with facts somewhat earlier than motives. By contrast, alignment with *human* judgments is a different kind of achievement as only the larger models were able to match human judgments about when lying is likely. In the one trajectory we tracked, fact expression and motive recognition emerged in pre-training, while human alignment appears only during the mid-training phase. Finally, relative to fact appreciation and motive recognition, human alignment can be removed by ablating deeper layers (later ones, closer to the output) that leave the other abilities intact.

These findings separate sensitivity to situational facts and speaker motives from agreement with human judgments about which situations invite lying. After the deeper ablation, a model retains substantial fact expression and an average response to the motive manipulation while its cross-scenario ordering diverges from human norms. Evaluations should therefore report these properties separately rather than treating human alignment as a general measure of motive understanding. The converse failure also exists: probing alone can report information that the model’s output does not express. More broadly, whether information is decodable, whether it is expressed in the output, whether behavior causally depends on it, and whether the resulting predictions align with human judgments are four distinct questions; the first three audit the machinery, the last audits its resemblance to us.

Why would the two come apart? Part of the answer is that alignment asks for something the other two abilities do not. Fact representation and motive recognition are directional — is the fact available, does the incentive move the prediction the right way — while ordering scenarios the way people do requires the magnitude assigned to each one to be calibrated and not merely signed, and the model’s magnitudes are close to binary where people’s are graded.

8 Limitations

The causal analysis is restricted to one model, and the human-alignment measure rests on 40 normed scenarios; replication across model families, larger and more diverse stimulus sets, languages, and discourse settings is needed.

Whether the shared mid-depth sensitivity reflects a fact→motive dependence remains untested and calls for restoration or fact-flip patching, and the mid-training pattern remains descriptive because data composition and annealing dynamics cannot be separated in the available checkpoint trajectory.

References

- Johannes Abeler, Daniele Nosenzo, and Collin Raymond. 2019. Preferences for truth-telling. *Econometrica*, 87(4):1115–1153.
- Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it’s lying. In *Findings of EMNLP*.
- Cody Blakeney, Mansheej Paul, Brett W. Larsen, Sean Owen, and Jonathan Frankle. 2024. Does your data spark joy? Performance gains from domain upsampling at the end of training. *arXiv preprint arXiv:2406.03476*.
- Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. Discovering latent knowledge in language models without supervision. In *Proceedings of ICLR*.
- Uri Gneezy. 2005. Deception: The role of consequences. *American Economic Review*, 95(1):384–394.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Thilo Hagendorff. 2024. Deception abilities emerged in large language models. *Proceedings of the National Academy of Sciences*, 121(24).
- John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of EMNLP*.
- Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, and 1 others. 2024. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*.
- Samuel Marks and Max Tegmark. 2024. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling (COLM)*.
- Thomas McGrath, Matthew Rahtz, János Kramár, Vladimir Mikulik, and Shane Legg. 2023. The Hydra effect: Emergent self-repair in language model computations. *arXiv preprint arXiv:2307.15771*.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*.
- OLMo Team, Pete Walsh, and 1 others. 2025. 2 OLMo 2 Furious. *arXiv preprint arXiv:2501.00656*.
- Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. 2025. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In *International Conference on Learning Representations (ICLR)*.
- Lorenzo Pacchiardi, Alex J. Chan, Sören Mindermann, Ilan Moscovitz, Alexa Y. Pan, Yarin Gal, Owain Evans, and Jan Brauner. 2024. How to catch an AI liar: Lie detection in black-box LLMs by asking unrelated questions. In *Proceedings of ICLR*.
- Richard Ren, Arunim Agarwal, Mantas Mazeika, Cristina Menghini, and 1 others. 2025. The MASK benchmark: Disentangling honesty from accuracy in AI systems. *arXiv preprint arXiv:2503.03750*.
- Jérémy Scheurer, Mikita Balesni, and Marius Hobbahn. 2023. Large language models can strategically deceive their users when put under pressure. *arXiv preprint arXiv:2311.07590*.
- Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*.
- Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations (ICLR)*.
- An Yang and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Fred Zhang and Neel Nanda. 2024. Towards best practices of activation patching in language models: Metrics and methods. In *Proceedings of ICLR*.
- Xingyuan Zhao and Seana Coulson. 2026. Graded expectations: Do large language models show human-like sensitivity to the likelihood of deceptive speech acts? In *Proceedings of the Society for Computation in Linguistics*.

A What Probes Cannot Tell You

Do the hidden states even carry the motive condition? Linear probes on the utterance-span states separate LE from TE almost perfectly (AUC .86–.97 in all twelve models, already within the first tenth of depth). On its own this number is nearly meaningless: a *layer-free* lexical baseline — the average layer-0 embedding of the scenario words plus span surprisal — already reaches .93–.96, matching the probe ceiling. Since the two conditions differ in wording, raw decodability certifies nothing beyond which words were transported into the span (Hewitt and Liang, 2019).

We therefore residualize: predict each layer’s states from the lexical baseline (out-of-fold, PCA-compressed regressors) and probe what remains, benchmarked against the same pipeline at layer 0, where no abstraction can exist (Figure 5). The families split cleanly: **all four OLMo2 checkpoints retain a strong condition signal beyond the lexical baseline** (peak AUC .77–.88 at mid depth), while **every Qwen3 model sits at or near the leakage floor** ($\approx .52$; largest exception, a weak deep trace in 4B-base). The behavioral profile of §4 is family-general; this representational strategy is not. We do not know what Qwen3 does instead — non-linear coding, context-side storage, or directions collinear with the lexical baseline are all open — but the practical consequence is immediate: **a probe-based audit calibrated on one model family would systematically misread the other**, in either direction. This is also why §5 reads behavior, not probes, from ablated models.

B The Full Model Grid

Figure 6 shows the three depth profiles of Figure 2 for all twelve models.

C Replicating the Ablation on OLMo2-7B

The causal analysis of §5 uses Qwen3-8B, chosen for its strong depth curves. Because the two families encode the motive condition differently (Appendix A), we repeated the ablation on OLMo2-7B-Instruct and observed a similar pattern.

D Model Behavior on the Stimulus Set

The analyses in the main text are about mechanism. This appendix reports what the models do (Table 1).

% true counts, over the 67 items, how often the final-layer output favors the true word over the lie one when the assertion is re-voiced by the neutral narrator of §4: the share with $F_i > 0$ in Equation 6. With no reason to lie, models lean toward the truth, though never overwhelmingly. Within each language model family, the rate rises with size, under both training paradigms. The lean comes from the scenario rather than from the words: with the scenario stripped away, the stem alone favors the true word for only 40–52% of items, near chance.

The remaining columns are in nats and come from the twin pairs. It measures the model’s final-layer preference between the two candidate words at the critical position, $\delta = \log p(\text{lie word}) - \log p(\text{truth word})$, so that $\delta < 0$ means the true word is favored.

δ^0 is δ measured from the stem alone, with no scenario. It never exceeds 1.08 nats, an order of magnitude below the scenario effects beside it, so the stimuli are lexically balanced and LP is measuring the scenario rather than the vocabulary.

LP is negative under *both* twins in every model — the scenario pushes toward the true word even when the speaker has a motive to lie — and Δ_{motive} , the difference between the two, is positive in every model: the motive eats into that push, by 0.22 to 3.92 nats, without reversing it. Post-trained models have higher LPs, which could be due to the lower-entropy output distribution of a post-trained model in chat format produces.

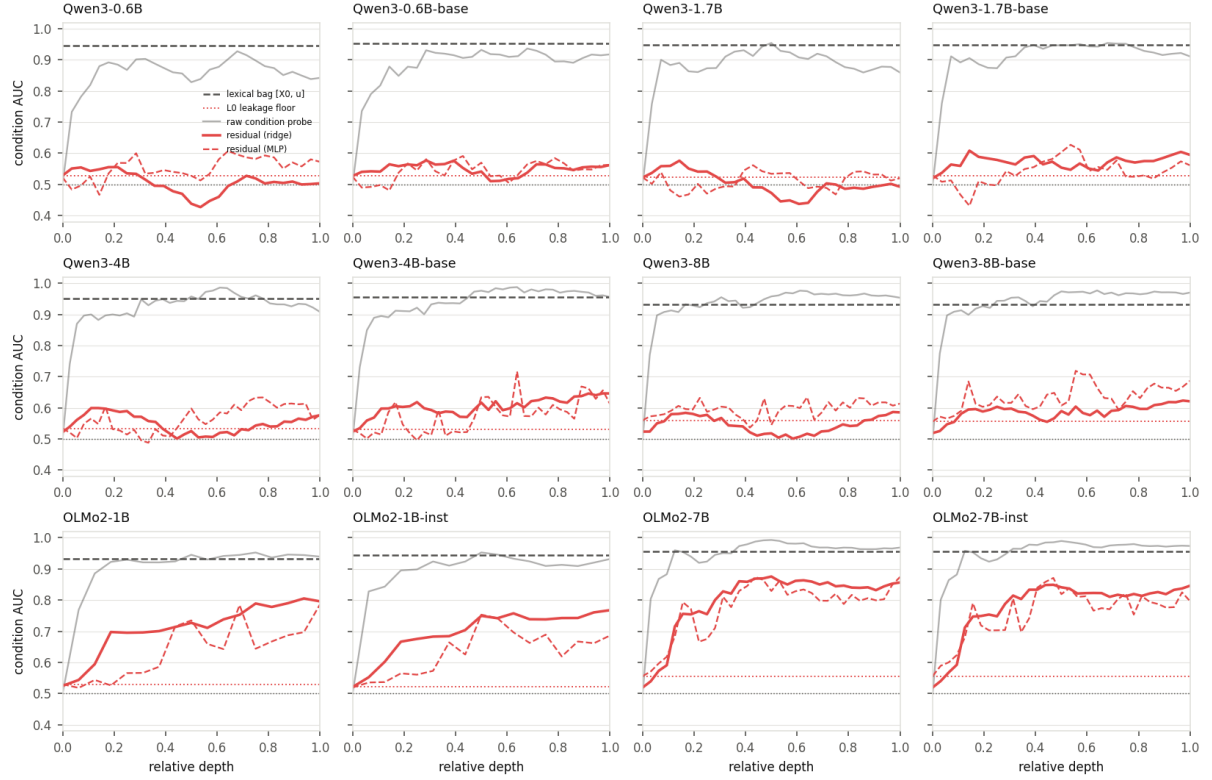


Figure 5: Condition probing, all twelve models. Gray: raw condition probe (LE vs. TE from utterance-span states) — at ceiling everywhere. Dashed: layer-free lexical baseline ($[X_0, \text{surprisal}]$), matching the ceiling. Red: probe on residuals after removing everything the lexical baseline predicts (ridge solid, MLP dashed), against the layer-0 leakage floor (dotted). OLMo2 retains an abstract condition signal; Qwen3 does not.

Model	no motive		motive present		
	% true	δ^0	LP_{LE}	LP_{TE}	Δ_{motive}
Qwen3-0.6B-base	63	0.32	-1.45	-1.94	0.50
Qwen3-0.6B	61	0.31	-1.58	-2.36	0.78
Qwen3-1.7B-base	66	0.26	-1.69	-2.19	0.50
Qwen3-1.7B	63	0.45	-2.85	-3.92	1.07
Qwen3-4B-base	67	0.33	-1.37	-1.59	0.22
Qwen3-4B	75	0.65	-6.23	-7.79	1.56
Qwen3-8B-base	72	0.39	-0.83	-1.68	0.85
Qwen3-8B	82	1.08	-6.12	-10.04	3.92
OLMo2-1B	58	0.28	-0.96	-1.84	0.88
OLMo2-1B-inst	64	0.36	-1.41	-1.99	0.58
OLMo2-7B	73	0.34	-1.15	-2.22	1.06
OLMo2-7B-inst	79	0.60	-3.36	-5.36	2.00

Table 1: Model behavior on the stimulus set. All columns use the final layer, each model’s native format, and the 67 audited pairs; the last four are in nats. % true: with no motive present, the share of neutral-narrator items whose output favors the true word over the false one. δ^0 : the preference between the two candidate words from the stem alone, without any scenario. LP_{LE} and LP_{TE} : the scenario effect $\delta - \delta^0$ under the lie-eliciting and truth-eliciting twin, negative meaning a preference for the true word. $\Delta_{\text{motive}} = LP_{LE} - LP_{TE}$; because the twins share a stem, this equals $\delta_{LE} - \delta_{TE}$ and does not depend on δ^0 .

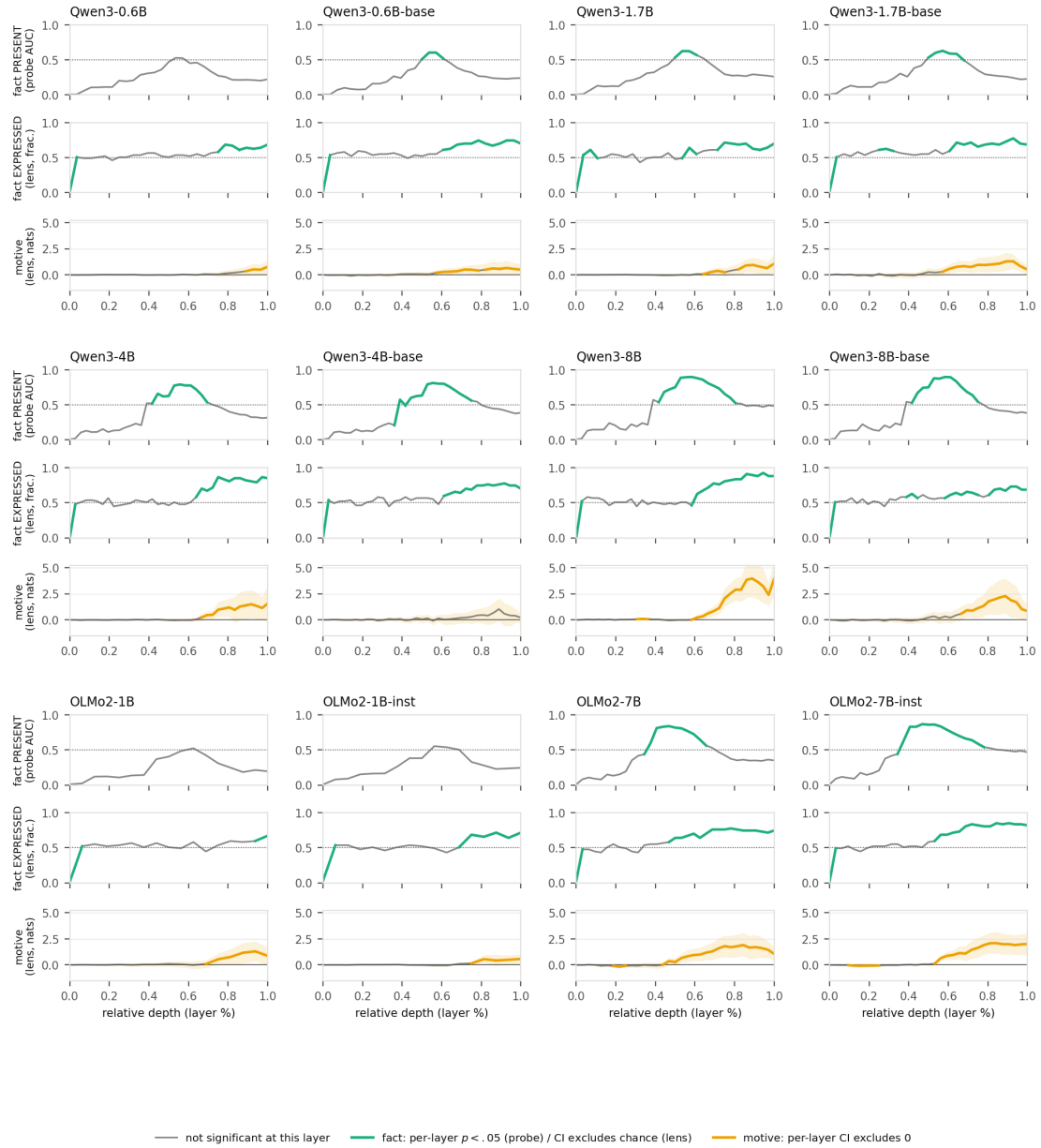


Figure 6: The full twelve-model version of Figure 2. Panels, colors, and significance shading are as in that figure. Chat-format and raw-format models are each read in their native format (§3).

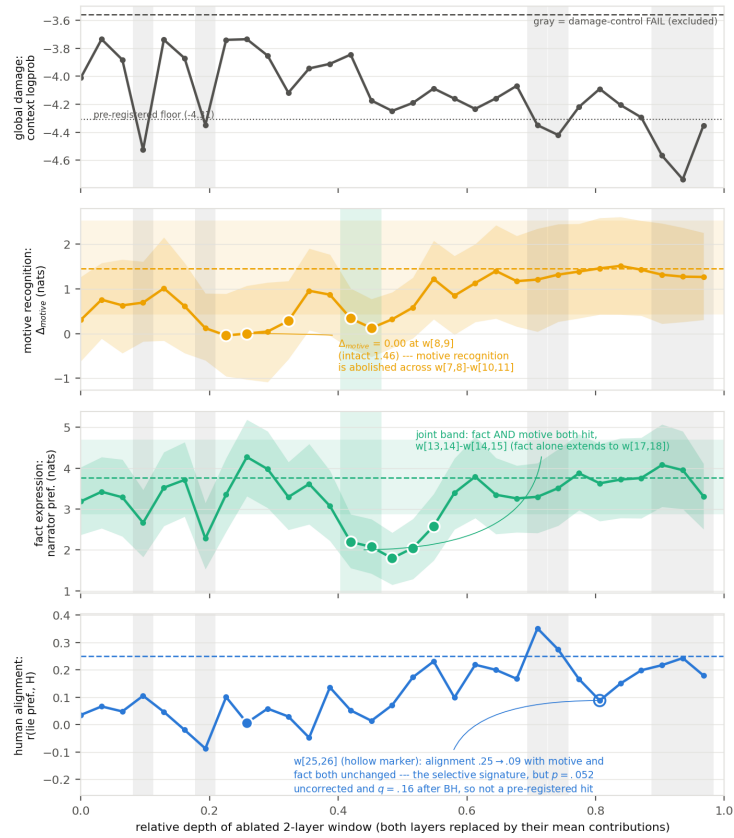


Figure 7: Sliding two-layer mean-ablation of OLMo2-7B-Instruct, in the layout of Figure 3. Rows: competence control, Δ_{motive} , fact expression, human alignment. Dashed = intact; large markers = BH $q < .05$ under the paired-with-intact bootstrap; grey bands = windows failing this model’s competence floor. The joint mid-depth band replicates (green shading, $w[13,14]$ – $w[14,15]$); the one window whose alignment drop survives correction, $w[8,9]$, is also the window at which Δ_{motive} is driven to zero, so the loss is not selective. The hollow marker at $w[25,26]$ carries the selective signature — alignment down, the other two intact.