

Inside self-reflection: the geometry in multidimensional scoring of metacognitive structure

Praveen Suthaharan^{1,2}, Philip Robert Corlett¹, Yuen-Siang Ang³

¹Interdepartmental Neuroscience Program, Yale University, New Haven, CT, USA,

²Department of Psychiatry, Yale University, New Haven, CT, USA,

³Health and Social Sciences Cluster, Singapore Institute of Technology, Singapore

Correspondence: praveen.suthaharan@yale.edu

Abstract

Large language models (LLMs) increasingly score open-ended text along multidimensional rubrics, yet strong score recoverability need not imply that a dimension is distinct or relevant to downstream decisions. We examine how Gemma-2-2B represents five metacognitive-prompting dimensions in self-reflections from 486 participants on a probabilistic reversal learning task. Linear probes recover all dimensions across depth, with depth gradients ordered by the rubric hierarchy. Evaluation is strongly decoded but has the weakest coupling to independently measured task markers. Under the primary five-occurrence token filter, Evaluation-related vocabulary shows no detectable association with switching, whereas the negative Final Decision association is stable across frequency thresholds. Activation patching yields oppositely signed point estimates in a fixed linear win-switch readout when the Final Decision–Evaluation matched-pair contrast is injected in each direction, although no layer survives correction across the full sweep. These findings show that decodability alone is insufficient evidence for weighting rubric dimensions in downstream decisions.¹

1 Introduction

AI workflows increasingly use LLM scores as inputs to decisions about individuals, including tutoring and mental-health assessment (Xu et al., 2026; Hu et al., 2026). In multi-agent systems, such scores can propagate to later components (Guo et al., 2024); consequently, an LLM-as-judge that misreads mastery or symptoms can alter subsequent support (Zheng et al., 2023; Li et al., 2024).

LLM graders can rely on lexical shortcuts unlike the evidence human raters use (Wu et al., 2024), motivating structured, multidimensional scoring

(Hashemi et al., 2024; Wang et al., 2026). Metacognitive prompting (MP) applies this approach to reflective text through five dimensions: Comprehension, Judgment, Evaluation, Final Decision, and Confidence (Wang and Zhao, 2024).

Prior work used GPT-4 to score MP in post-task reflections and found that the aggregate score moderated a relationship between volatility beliefs and erratic switching (Suthaharan et al., 2026), but did not examine the five dimensions separately or their representation inside a language model. Figure 1 provides a schematic overview of how we construct one layer-specific vector per reflection and examine the relative positions and score-associated directions of those vectors. We then combine linear probing, logit-lens analysis, and activation patching in Gemma-2-2B to ask whether each dimension is recoverable, whether recoverability coincides with independently measured task markers, and whether dimension-related geometry contributes to a win-switch readout.

2 Related Work

Measuring metacognition. Canonical metacognition measures quantify how confidence tracks performance (Fleming and Lau, 2014), whereas broader constructs such as evaluation and self-reflection lack obvious ground truth (Katyal and Fleming, 2024). LLM scoring offers a complementary narrative measure (Jenner et al., 2025; Suthaharan et al., 2026); we test its five dimensions individually.

Internal representations of psychological constructs. Psychological beliefs can be recovered and manipulated in language-model activations (Zhu et al., 2024), while factual knowledge may be internally recoverable without being expressed (Gekhman et al., 2025). Thus, separately named rubric dimensions need not imply distinct representations or downstream use.

¹github.com/psuthaharan/mp-llm-geometry.

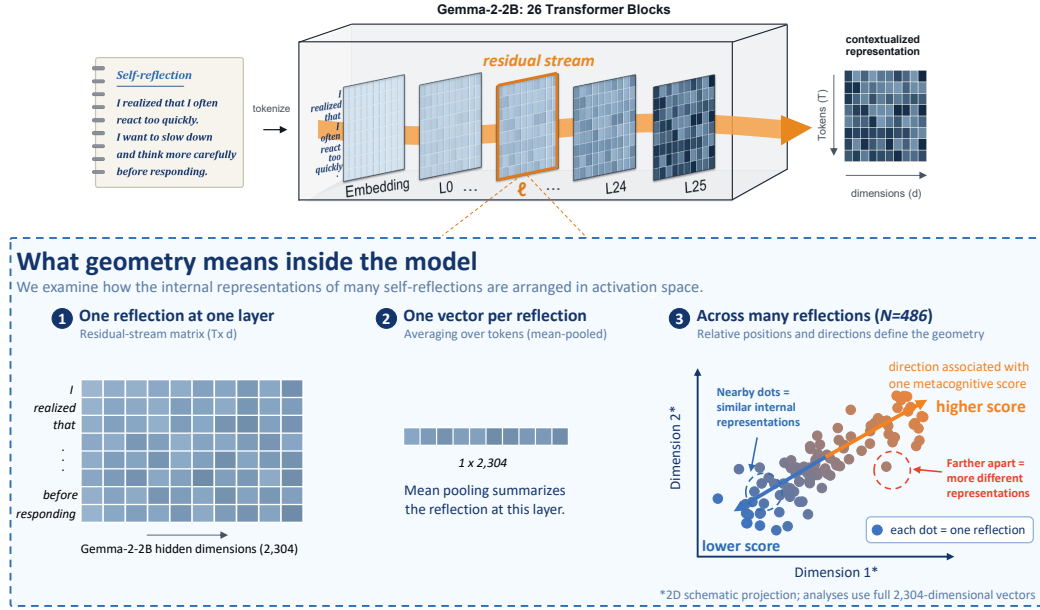


Figure 1: **Self-reflections as vectors in layer-specific representational spaces.** (Top) Each participant’s reflection is tokenized (split into text units) and embedded (each token is converted into an initial numerical representation before information from the rest of the reflection is incorporated). These representations then travel through the residual stream, where Gemma-2-2B’s 26 transformer blocks (L0–L25) repeatedly update them using preceding context. (Bottom) At a selected layer ℓ , mean pooling averages each reflection’s token representations into one 2,304-dimensional vector. Repeating this for all $N = 486$ participants places their reflection vectors in a shared space. Nearby vectors are represented similarly, while linear probes identify directions along which metacognitive scores vary. These positions and directions constitute the geometry analyzed here. Cell colors and the two-dimensional view are schematic; analyses use all 2,304 dimensions.

Interpreting transformer representations. The residual stream supports layer-resolved probing for linearly decodable information (Elhage et al., 2021; Alain and Bengio, 2016; Belinkov, 2022; Park et al., 2024) and logit-lens inspection of vocabulary alignment (nostalgebraist, 2020; Belrose et al., 2023); depth may distinguish surface from integrative content (Luo et al., 2025). Activation patching and contrastive steering test whether a direction affects a fixed downstream readout (Meng et al., 2022; Zhang and Nanda, 2024; Turner et al., 2023; Rimsky et al., 2024). We compose these tools to audit presence, content, and downstream influence in the same rubric representations.

3 Methods

3.1 Participants, task, and markers

We analyze $N = 486$ participants who completed a 160-trial, three-option probabilistic reversal learning task and then answered two open-ended prompts about their strategy and switching (Suthaharan et al., 2026). The independently mea-

sured task markers are win-switch rate (WSR), the proportion of choices switched after positive feedback (mean .096, SD .165, median .020, range $[0, .990]$), and the initial volatility prior μ_3^0 , an anticipatory belief about environmental change (Mathys et al., 2011).

3.2 Metacognitive prompting scores

GPT-4 scores for Comprehension, Judgment, Evaluation, Final Decision, and Confidence are reused without modification from the validated protocol of Suthaharan et al. (2026). Gemma-2-2B does not generate scores here; probes predict those existing scores from Gemma activations of the human-written reflections.

3.3 Model and representation extraction

We use the *base* Gemma-2-2B (Gemma Team, 2024), with 26 layers, residual width $d = 2304$, and tied SentencePiece embedding/unembedding matrix $E \in \mathbb{R}^{256,000 \times d}$. Reflections are tokenized with BOS, padded as needed, and capped at 512 tokens (none were truncated). Let $h_{i,j}^{(\ell)}$ be participant

i 's residual state at layer ℓ and token position j . We mean-pool content positions, excluding BOS and padding:

$$\mathbf{x}_i^{(\ell)} = \frac{1}{T_i} \sum_{j \in \text{content}(i)} \mathbf{h}_{i,j}^{(\ell)} \in \mathbb{R}^d. \quad (1)$$

This yields 26 activation matrices $\mathbf{X}^{(\ell)} \in \mathbb{R}^{486 \times 2304}$.

3.4 Length residualization

Because length covaries with scores and activations, we residualize every participant-indexed variable v_i —scores, markers, activation coordinates, and logit-lens channels—against log-transformed character count $\log(1 + c_i)$:

$$v_i = \beta_0 + \beta_1 \log(1 + c_i) + \varepsilon_i \quad (2)$$

We retain $\tilde{v}_i = \varepsilon_i$. Descriptive correlations use full-sample adjustment; probing fits activation and score adjustments within each outer training fold and applies them to its held-out fold. Correlations between adjusted quantities are partial correlations controlling for log-transformed length, denoted r_p .

3.5 Layer-wise linear probing

Estimator. For each layer ℓ and dimension k , we fit ridge regression from residualized activations to residualized scores $\tilde{\mathbf{y}}_k$:

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w} \in \mathbb{R}^d} \|\tilde{\mathbf{y}}_k - \mathbf{Z}^{(\ell)} \mathbf{w}\|_2^2 + \alpha \|\mathbf{w}\|_2^2, \quad (3)$$

Here $\mathbf{Z}^{(\ell)}$ is column-standardized $\tilde{\mathbf{X}}^{(\ell)}$. Within each outer training fold, α is selected by efficient leave-one-out cross-validation over 13 values on $[10^{-1}, 10^5]$.

Evaluation. Five-fold outer cross-validation fits length adjustment, standardization, α , and the probe without the held-out fold. Held-out encoding strength is

$$r_k(\ell) = \text{corr}(\hat{\mathbf{y}}_k^{(\ell)}, \tilde{\mathbf{y}}_k), \quad (4)$$

over concatenated out-of-fold predictions. Benjamini–Hochberg correction covers all $26 \times 5 = 130$ layer–dimension tests; because conventional Pearson tests are applied to tuned cross-validated predictions, adjusted cell-wise p values are exploratory.

Cross-dimensional specificity. The five rubric scores are derived from the same reflection, raising the possibility that a probe trained on one dimension could recover variance shared across several dimensions. We therefore quantified overlap at both the score and probe levels. We first computed 5×5 Pearson correlation matrices for the raw GPT-4 scores and for the cross-fitted length-adjusted scores. We then correlated each dimension's out-of-fold probe prediction at its observed maximizing layer with all five cross-fitted score residuals, yielding a cross-dimensional transfer matrix (Appendix Figure 7). For each probe, we summarized target specificity as the correlation with its matched rubric score relative to its strongest absolute off-target correlation.

Profile shape and depth gradient. Dimensions differ in overall decodability, so comparing raw r across dimensions conflates profile *shape* with *level*. For Figure 2A we therefore z -score each dimension's profile across its own 26 layers, $z_k(\ell) = (r_k(\ell) - \bar{r}_k) / \text{sd}_\ell r_k(\ell)$, so that $z_k(\ell) > 0$ marks a layer that encodes dimension k better than that dimension's own average. Each dimension's *observed maximizing layer* is $\ell_k^* = \arg \max_\ell r_k(\ell)$. To summarize how encoding changes with depth we compute the *depth gradient* ρ_k , the Spearman rank correlation between layer index and $r_k(\ell)$ across the 26 layers. ρ_k near +1 indicates encoding that builds toward the deepest layers, near −1 an early peak and decline, and near 0 no trend. To test whether these depth gradients follow the MP rubric order, we correlate the five ρ_k values with rubric position and compute an exact two-sided permutation p value over all $5! = 120$ assignments. We assess uncertainty with 5,000 participant-bootstrap resamples of the out-of-fold predictions and scores, recomputing the layer profiles, maximizing layers, depth gradients, and ordered-trend statistic in each resample.

3.6 Encoding–coupling analysis

We next ask whether dimensions that are strongly recoverable from the model also align with independently measured task markers. For each dimension k , the probe at its observed maximizing layer yields a cross-validated readout $\hat{\mathbf{y}}_k \in \mathbb{R}^{486}$. We place each dimension in a two-dimensional plane (Figure 3). The x -coordinate is its encoding strength $r_k(\ell_k^*)$. The y -coordinate summarizes coupling of

the same readout to the two task markers,

$$\frac{1}{2} \sum_{b \in \{\text{WSR}, \mu_3^0\}} |r_p(\hat{\mathbf{y}}_k, b)|, \quad (5)$$

the mean absolute partial correlation with win-switch rate and the initial volatility prior. This mean is used only as the display coordinate in Figure 3; statistical inference is based on the ten signed dimension-by-marker partial correlations evaluated separately. For each association, uncertainty is quantified with participant-bootstrap 95% confidence intervals, and Holm correction controls familywise error across the ten tests (Appendix Table 1).

We use the cross-validated probe prediction because it is the dimension-specific signal recoverable from Gemma’s activations. Relating that readout to the task markers quantifies how the recovered score signal covaries with independently measured belief and behavioral variation. This quantity does not isolate representational relevance from the underlying score–marker relationship or probe error, so we interpret it specifically as recovered-readout coupling.

3.7 Logit-lens vocabulary coupling

The encoding–coupling map operates on whole reflections; we next resolve it at the level of vocabulary using the logit lens (nostalgebraist, 2020), asking *which words* carry each dimension’s signal in the model’s own geometry. Because the unembedding is tied to $\mathbf{E}$, projecting a pooled representation through $\mathbf{E}^\top$ scores its alignment with every token’s embedding direction:

$$\ell_i^{(\ell)} = \mathbf{x}_i^{(\ell)} \mathbf{E}^\top \in \mathbb{R}^{|\mathcal{V}|}. \quad (6)$$

Stacking across participants yields a $486 \times 256,000$ logit matrix per layer. We focus on Evaluation and Final Decision because they show the largest separation in task-marker coupling in Figure 3. For each dimension, the logit-lens analysis is performed at the same observed maximizing layer used for its probe readout: $\ell = 12$ for Evaluation and $\ell = 21$ for Final Decision.

Token filtering and lemma channels. Only a small fraction of the vocabulary is informative for this corpus. We retain tokens that (i) occur anywhere in the 486 reflections ($n = 1,143$), (ii) occur at least five times ($n = 325$), and (iii) are not stopwords, punctuation, or bare numerals ($n = 241$

content tokens). Since one lexical idea often spans morphological variants in separate columns (e.g., *switch*, *switched*, *switching*), tokens sharing a root are merged into *lemma channels* by averaging their logit columns, reducing the 241 tokens to 207 channels.

Token-filter diagnostics and sensitivity. Because intermediate logits are uncalibrated, we assess corpus-relative token ranks rather than absolute magnitudes. We repeat coupling analyses at minimum counts 1, 2, 5, and 10 and audit all 256,000 vocabulary directions under a fixed readability rule; Appendix D gives criteria and results. The audit is descriptive and does not select Figure 4 tokens.

Coupling coordinates. Each lemma channel λ , with per-participant values $\mathbf{g}_\lambda$, receives two coordinates: $r_{\text{EV}}(\lambda) = r_p(\mathbf{g}_\lambda^{(12)}, \mathbf{y}_{\text{EV}})$ and $r_{\text{FD}}(\lambda) = r_p(\mathbf{g}_\lambda^{(21)}, \mathbf{y}_{\text{FD}})$, its partial correlations with the Evaluation and Final Decision scores at each dimension’s observed maximizing layer. We plot all 207 channels in the $(r_{\text{FD}}, r_{\text{EV}})$ plane, generating the coupling subspace of Figure 4A, and the dominance statistic $\Delta(\lambda) = r_{\text{EV}}(\lambda) - r_{\text{FD}}(\lambda)$ classifies each channel as EV-dominant ($\Delta > 0$) or FD-dominant ($\Delta < 0$).

Rubric specificity. Channels with both couplings positive ($n = 29$) co-track both rubrics, raising the question of whether they are *rubric-specific* or merely generically elevated. The dominance diagonal splits them into an EV-leaning zone ($r_{\text{EV}} > r_{\text{FD}}$; $n = 11$) and an FD-leaning zone ($r_{\text{FD}} > r_{\text{EV}}$; $n = 18$). We pool each zone’s channels into one per-participant score (the mean of its channel values) and compute its partial correlation with the zone’s own rubric (r_{matched}) and with the other (r_{cross}). Rubric specificity, $\Delta r = r_{\text{matched}} - r_{\text{cross}}$, is positive when a zone’s vocabulary tracks its own dimension more than the other. Uncertainty is quantified with a 95% bootstrap confidence interval over participants, and stability with 20 random 50/50 participant splits that evaluate the fixed channels on each held-out half. Because the zones are defined from full-sample couplings before resampling, these intervals and split summaries are conditional on the selected channels and are treated as exploratory.

Token-level coupling to behavior. Finally, we return to the 241 individual content tokens (one

step before lemma merging) and give each token, alongside its rubric coupling, a behavioral coupling $r_{\text{WSR}}(t) = r_p(\ell_t, \text{WSR})$, the partial correlation between its 486 logit values and win-switch rate. Across the 241 tokens we compute the Spearman correlation between rubric and behavioral coupling— $\rho(r_{\text{EV}}(t), r_{\text{WSR}}(t))$ at layer 12 and $\rho(r_{\text{FD}}(t), r_{\text{WSR}}(t))$ at layer 21 (Figure 4C)—to describe whether vocabulary that signals a dimension also signals behavior. Because token directions are correlated rather than independent sampling units, the conventional Spearman p values are descriptive.

3.8 Activation patching

We test whether shifting a representation toward Final Decision geometry *changes* the model’s win-switch readout using layer-resolved activation patching, a causal-intervention technique in which a directed perturbation is injected into the residual stream mid-forward-pass and its downstream consequence is measured (Meng et al., 2022; Zhang and Nanda, 2024). The contrastive construction of the injected direction follows the design principle of steering with contrastive pairs (Turner et al., 2023; Rinsky et al., 2024).

Matched pairs. We construct pairs of reflections that differ sharply in Final Decision versus Evaluation content while being matched on length. Let $D_i = y_{\text{FD},i} - y_{\text{EV},i}$. (Step 1) Rank all 486 participants by D_i ; the top 20% form the *source* pool (Final-Decision-leaning) and the bottom 20% the *target* pool (Evaluation-leaning). (Step 2) Traverse the source pool in random order, pairing each source with the still-unused target closest in reflection length until $P = 15$ pairs exist. (Step 3) A reversed control repeats the procedure with the pools’ roles swapped, injecting the same axis in the opposite direction.

Win-switch readout. The intervention needs a fixed downstream quantity to watch. We define a win-switch readout at the final layer by ridge regression (Eq. 3) from the residualized layer-25 activations $\tilde{\mathbf{X}}^{(25)}$ to the residualized win-switch rates. Let S_{WSR} denote the feature standardization fitted for this readout. Its prediction for a residualized final-layer activation is

$$\hat{f}_{\text{WSR}}(\tilde{\mathbf{x}}) = \hat{b}_{\text{WSR}} + \hat{\mathbf{w}}_{\text{WSR}}^\top S_{\text{WSR}}(\tilde{\mathbf{x}}). \quad (7)$$

The fitted length adjustment, scaler, intercept, and ridge coefficients are then held fixed and applied

identically to patched and baseline activations. Displacement along $+\hat{\mathbf{w}}_{\text{WSR}}$ in these standardized coordinates corresponds to a higher predicted residualized win-switch rate. As a predictive diagnostic, the readout had in-sample $r = .881$ but five-fold fixed- α out-of-fold $r = .272$ and $R^2 = -.194$; it is therefore a model-internal intervention endpoint rather than a validated behavioral predictor.

Intervention. For a pair (s, t) and patch layer L , the injected direction is the pair’s pooled geometry gap at that layer, computed from the raw (non-residualized) activations:

$$\Delta^{(L)} = \mathbf{x}_s^{(L)} - \mathbf{x}_t^{(L)} \in \mathbb{R}^d. \quad (8)$$

We rerun the target’s forward pass; at layer L , $\Delta^{(L)}$ is added to the residual state of *every* token position,

$$\mathbf{h}_{t,j}^{(L)} \leftarrow \mathbf{h}_{t,j}^{(L)} + \Delta^{(L)} \quad \forall j, \quad (9)$$

and all subsequent layers $L+1, \dots, 25$ recompute on the perturbed stream, so the perturbation propagates through downstream computation rather than being carried passively. The patched final-layer states are mean-pooled (Eq. 1), length-residualized against the target’s own length, and projected onto the fixed readout (Eq. 7). The pair’s effect at layer L is the patched-minus-baseline prediction,

$$\begin{aligned} e_{s \rightarrow t}(L) &= \hat{f}_{\text{WSR}}(\tilde{\mathbf{x}}_t^{\text{patch}}) - \hat{f}_{\text{WSR}}(\tilde{\mathbf{x}}_t^{\text{base}}) \\ &= \hat{\mathbf{w}}_{\text{WSR}}^\top S_{\text{WSR}}(\tilde{\mathbf{x}}_t^{\text{patch}}) \\ &\quad - \hat{\mathbf{w}}_{\text{WSR}}^\top S_{\text{WSR}}(\tilde{\mathbf{x}}_t^{\text{base}}). \end{aligned} \quad (10)$$

where the baseline is an unperturbed forward pass of the same target text. The layer- L effect $\Delta_{\text{WSR}}(L)$ is the mean of $e_{s \rightarrow t}(L)$ over the 15 pairs. Pointwise uncertainty is summarized with 95% confidence intervals from 5,000 nonparametric bootstrap resamples over pairs. Since the intervention is evaluated across all 26 layers, layer-wise inference treats the complete layer sweep as one family. We use an exact sign-flip $\max\{|t|\}$ procedure over all 2^{15} possible sign assignments to control familywise error across the 26 layers.

The sweep is repeated for the reversed intervention to assess directionality (Figure 5). As additional specificity checks, at L19, the layer with the smallest familywise-adjusted p value in the focal sweep, we compare the intervention with 100 isotropic Gaussian perturbations normalized within pair to the empirical intervention norm and with the nine other pairwise MP dimension contrasts. The

former tests whether the focal direction is unusual relative to equally large arbitrary perturbations; the latter tests whether the effect is unique to the Final Decision–Evaluation contrast.

4 Results

4.1 Depth gradients follow the MP hierarchy

All five dimensions showed positive held-out correlations throughout the residual stream, with exploratory BH-adjusted Pearson tests below .05 for all 130 layer–dimension combinations. Relative encoding was lower in the early layers (L0–L8; Figure 2A), consistent with metacognitive content requiring contextual integration beyond the surface-level features those layers resolve (Luo et al., 2025). Encoding then strengthens, with observed maxima at Evaluation L12, Comprehension L13, Judgment L14, Confidence L16, and Final Decision L21 (Appendix Figure 6). Depth gradients (Figure 2B) summarize the profile shapes. Final Decision and Confidence build toward the deepest layers ($\rho = +0.83$ and $+0.84$, respectively), Evaluation peaks mid-network without pronounced late growth ($\rho = +0.23$), Judgment shows little directional trend ($\rho = -0.11$), and Comprehension peaks mid-network and declines ($\rho = -0.35$). The depth gradients show an ordered association with the MP hierarchy ($r = 0.971$, exact permutation $p = 0.0167$, bootstrap 95% CI $[0.529, 0.991]$), consistent with dimensions involving more integrative commitment building more strongly with depth (Wang and Zhao, 2024).

4.2 Encoding strength dissociates from task-marker coupling

The encoding–coupling plane (Figure 3) reveals a selective dissociation. Comprehension, Judgment, Confidence, and Final Decision show modest task-marker coupling ($r = 0.119$ – 0.159), whereas Evaluation is lower ($r = 0.050$) despite strong held-out encoding ($r = 0.579$). Across 5,000 participant-bootstrap draws, Evaluation was the lowest-coupling dimension in 91.7% of resamples. We also observe that Evaluation was weakly related to both the initial volatility prior ($r_p = -0.063$) and win–switch rate ($r_p = -0.037$), with neither association surviving Holm correction (Appendix Table 1). Its low position in Figure 3 therefore does not arise from averaging across otherwise divergent marker relationships.

The dissociation is also not explained solely

by overlap among the rubric dimensions. Evaluation’s out-of-fold probe prediction correlated more strongly with its matched Evaluation score ($r = 0.579$) than with any off-target dimension (maximum absolute $r = 0.361$; Appendix Figure 7). Evaluation is therefore strongly recoverable from Gemma’s representation while remaining comparatively weakly aligned with independently measured task markers.

4.3 Vocabulary-level structure of the dissociation

The coupling subspace of the 207 lemma channels (Figure 4A) is sharply asymmetric: 188 channels (90.8%) are FD-dominant versus 19 EV-dominant (9.2%), reflecting how heavily participants recount the task in action vocabulary (*partner*, *switched*, *cards*) relative to the sparse appraisal register (*successful*, *losses*). Within the upper-right quadrant, where both couplings are positive ($n = 29$), the dominance diagonal separates 11 EV-leaning from 18 FD-leaning channels. Conditional on these full-sample-selected channels, pooled EV-leaning vocabulary tracks Evaluation more than Final Decision ($\Delta r = +0.111$, 95% CI $[+0.009, +0.217]$, $p = .0288$; 20/20 split-half estimates positive), whereas pooled FD-leaning vocabulary shows no specificity ($\Delta r = -0.040$, 95% CI $[-0.122, +0.050]$, $p = .3668$; 2/20 split-half estimates positive). Because channel selection was not repeated within resamples, these specificity estimates are exploratory and do not establish out-of-sample rubric specificity. At the level of individual tokens (Figure 4C), under the primary five-occurrence filter, Evaluation coupling was not detectably related to win–switch coupling ($\rho = +0.053$, descriptive $p = .416$), whereas Final Decision coupling was strongly negative ($\rho = -0.675$, descriptive $p < .001$). The Final Decision association remained strongly negative across minimum-frequency thresholds of 1, 2, 5, and 10 occurrences ($\rho = -0.614$ to -0.675 , all descriptive $p < .001$), whereas the Evaluation association was sensitive to token selection ($\rho = -0.069$ to $+0.330$). The retained tokens were not uniformly low-ranking in the corpus-relative vocabulary: 174 of 241 tokens at L12 and 108 of 241 at L21 entered the top 100 for at least one participant. The highest-ranking excluded corpus items were predominantly stopwords, punctuation, or unreadable token pieces. The fixed-rank full-vocabulary audit was semantically mixed rather

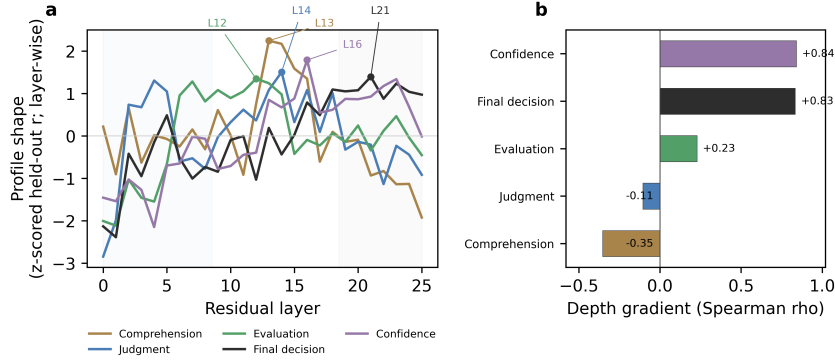


Figure 2: **Depth gradients follow the metacognitive prompting hierarchy.** (A) Layer-wise encoding profiles (z-scored nested held-out Pearson r) across the 26 residual layers for the five MP dimensions ($N = 486$), with observed maximizing layers marked. (B) Spearman depth gradient (ρ) between layer index and held-out r for each dimension. Across the MP rubric order, the depth gradients show an ordered trend ($r = 0.971$, exact permutation $p = 0.0167$, bootstrap 95% CI $[0.529, 0.991]$).

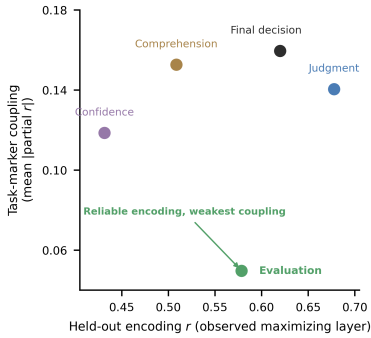


Figure 3: **Evaluation shows the weakest coupling to task markers.** Encoding strength (held-out r at each dimension's observed maximizing layer) versus task-marker coupling (mean absolute partial r with μ_3^0 and win-switch rate; $N = 486$). The mean is used only as a display coordinate; marker-specific associations are evaluated separately.

than uniformly confirming the corpus-restricted interpretation (Appendix D). Thus, the stable lexical pattern is the negative Final Decision association across frequency thresholds; the absence of detectable Evaluation coupling under the primary filter is not invariant to token-selection threshold. This asymmetry motivates the intervention test of §4.4.

4.4 Activation patching yields suggestive readout shifts

Injecting the Final Decision–Evaluation matched-pair contrast yielded negative point estimates in the win-switch readout across the layer sweep (Figure 5). The most negative mean occurred at L20 ($\Delta\text{WSR} = -0.1195$). L19 had a sim-

ilar mean effect ($\Delta\text{WSR} = -0.1168$, 95% CI $[-0.1951, -0.0426]$) and the smallest adjusted p value (raw exact sign-flip $p = .0081$; max- $|t|$ $p_{\text{FWER}} = .0715$). Thus, no individual layer survived familywise correction across the 26-layer sweep.

At L19, the reversed intervention had a positive point estimate ($\Delta\text{WSR} = +0.0933$, 95% CI $[-0.0067, +0.1977]$). A post-selection comparison at L19 found the focal contrast unusual relative to norm-matched random perturbations ($p = 0.0198$), but this comparison does not correct for selecting L19 from the layer sweep. The pattern was also not unique to this rubric contrast, as two of the nine alternative MP contrasts produced intervals excluding zero. Given the readout's weak out-of-fold performance and the absence of a layer surviving correction, these results are descriptive evidence of a direction-dependent change in a fitted model-internal diagnostic, not confirmatory evidence about downstream model use or human behavior.

5 Discussion

We asked how a language model internally represents the five dimensions of the metacognitive prompting rubric, and whether its strongest representations are those that align with independently measured task markers. Our linear probes recovered all five dimensions from the residual stream of Gemma-2-2B, within encoding profiles that varied across depth. Evaluation was decoded about as well as any dimension, yet its recovered readout showed the weakest coupling to the task markers,

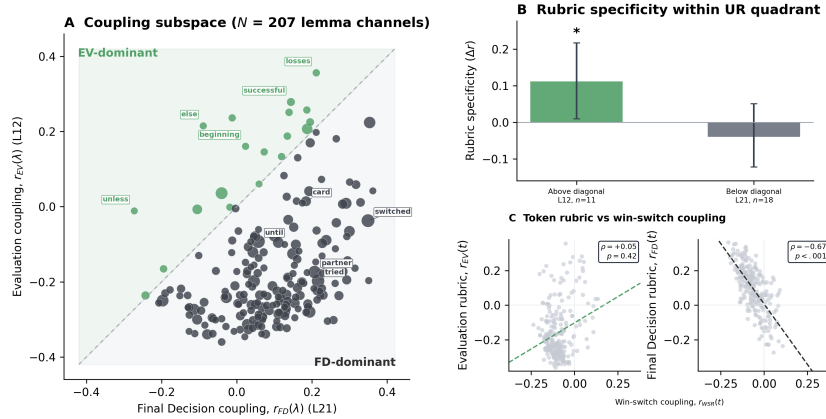


Figure 4: **Outcome-focused vocabulary characterizes the coupling asymmetry.** Logit-lens vocabulary coupling at the Evaluation and Final Decision observed maximizing layers (L12, L21; $N = 486$). **(A)** Coupling subspace of the 207 lemma channels in (r_{FD}, r_{EV}) . **(B)** Rubric specificity ($\Delta r = r_{\text{matched}} - r_{\text{cross}}$) of each upper-right zone. Error bars are participant-bootstrap 95% confidence intervals and the asterisk marks an interval excluding zero; these estimates are exploratory and conditional on the full-sample-selected channels. **(C)** Token-level rubric coupling versus win-switch coupling across the 241 content tokens; conventional token-wise p values are descriptive. Token-filter diagnostics, frequency-threshold sensitivity, and the full-vocabulary semantic audit are reported in Appendix D and Table 2.

a pattern that remained after separating the volatility prior from win-switch behavior. The patching analysis yielded downward point estimates when the Final Decision–Evaluation matched-pair contrast was injected in one direction and an upward point estimate at L19 when it was reversed, but no layer survived sweep-wise correction. This distinction between decodability and external coupling is consistent with longstanding cautions that successful probing establishes that information is present, not that the model uses it (Hewitt and Liang, 2019; Elazar et al., 2021). Related work similarly shows that factual knowledge can be encoded internally without being expressed in a model’s outputs (Gekhman et al., 2025). Our results extend this distinction to multidimensional scoring. A rubric dimension can be strongly recoverable from a model’s representation while its recovered readout remains only weakly aligned with independently measured external variation.

Interestingly, the model was never exposed to the rubric. Yet its residual-stream depth gradients align with the rubric’s hierarchy. We do not take this to mean that the model performs metacognition. A more parsimonious reading is that the linguistic content of later rubric stages requires more contextual integration before individuals become separable on it, consistent with evidence that early layers resolve surface form while deeper layers integrate meaning (Luo et al., 2025). The broad uncertainty

around the maximizing layers (Appendix Figure 6), however, argues against interpreting the observed maximizing-layer differences as five sharply separated computational stages.

A natural concern is that the dimensions our probes recover reflect the scorer rather than the language. The design contains a check against this possibility. The scores were assigned by GPT-4, whereas the geometry that predicts them was measured in a base model three orders of magnitude smaller, from a different family (Gemma-2-2B). This cross-model recovery is consistent with the dimensions corresponding to regularities of reflective language on which independently trained models converge (Huh et al., 2024). The overlap analysis sharpens this interpretation. The five dimensions remain correlated after controlling for reflection length, and their probe readouts show cross-dimensional transfer (Appendix Figure 7), indicating that the recovered geometry contains substantial structure shared across the rubric. Yet Evaluation remains more strongly aligned with its own rubric score than with any other dimension, showing that this shared structure does not fully account for its decodability. The geometry therefore appears to combine rubric-wide regularities with at least some dimension-specific signal. We caution that cross-model recovery does not establish construct validity, because models trained on overlapping text can converge on the same scoring

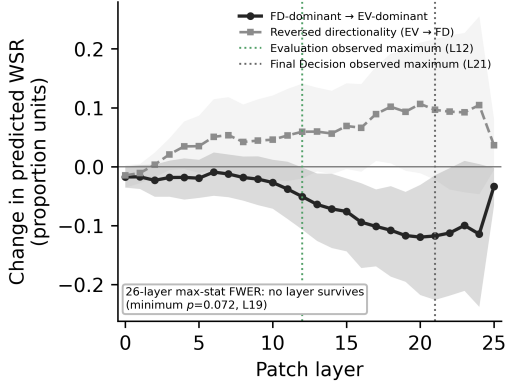


Figure 5: **Activation patching yields a suggestive shift in the win-switch readout.** Layer-resolved activation-patching effect on the fixed linear readout ($N = 15$ matched pairs), shown as the patched-minus-baseline change in predicted residualized WSR when the matched-pair contrast is injected at layer L (solid: FD-leaning source into EV-leaning target; dashed: reversed intervention). Shading shows pointwise 95% bootstrap confidence intervals from 5,000 resamples. No individual layer survives max- $|t|$ familywise correction across the 26-layer sweep (minimum $p_{\text{FWER}} = .0715$ at L19). Dotted vertical lines mark the maximizing layers for Evaluation and Final Decision from Figure 2.

regularities and share the same blind spots.

Reflective text separates the description of action from the appraisal of outcome (Raković et al., 2022). In students’ post-exam reflections, evaluative language predicted subsequent study behavior only indirectly, through the language of concrete plans, and neither prior exam performance nor behavior predicted the volume of evaluation itself. The internal geometry of Gemma-2-2B shows a related division; the appraisal register is comparatively weakly and filter-sensitively coupled to behavior, whereas the action-proximal register shows a stable behavioral association (Figure 4). This asymmetry extends to how metacognition is measured. Confidence became the best measured facet of metacognition because task accuracy supplies a ground truth, whereas the broader evaluative processes the field aims to capture lack one (Katyál and Fleming, 2024). The evaluative stage exhibits the same untethering when models perform reflection themselves, drifting from initially accurate judgments under MP prompting (Wang and Zhao, 2024) and monitoring only a limited part of their own activations (Ji-An et al., 2025).

These findings have important implications for AI workflows in which dimension scores inform decisions about individuals, such as pedagogical

tagging, scaffold selection, and adaptive assessment (Xu et al., 2026; Hu et al., 2026). A workflow that weights dimensions according to how reliably an LLM can recover them would, in our MP setting, assign substantial weight to Evaluation despite its recovered readout’s comparatively weak alignment with the task markers. The problem is therefore not simply whether a rubric dimension can be measured consistently, but whether the recoverable score signal aligns with external measures relevant to its intended use. This work motivates an audit during instrument validation, before dimension scores gate decisions: compare decodability with coupling to independently measured task markers, then use intervention to test how the relevant geometry affects a fixed model endpoint.

6 Conclusion

We mapped how a language model represents the five dimensions of a metacognitive rubric, showing that strong decodability does not guarantee alignment with external signals relevant to downstream use. Evaluation was among the most strongly recoverable dimensions, yet its recovered readout showed the weakest coupling to the task markers. Patching along the Final Decision–Evaluation matched-pair contrast produced descriptive, oppositely signed changes in the model’s fixed linear win-switch readout. A related asymmetry was visible in vocabulary, separating outcome appraisal from action-proximal language. Decodability is therefore an incomplete basis for weighting rubric dimensions. When dimension-level scores inform decisions about individuals, recovered score signals should be compared with independent validity evidence before they guide downstream decisions.

Limitations

Our evidence comes from a single 2-billion-parameter base model, Gemma-2-2B. The exact representational geometry reported here should therefore not be assumed to generalize across model scale, architecture, or post-training regime. Larger or instruction-tuned models may organize their representations differently, so the depth ordering and the dissociation we report may shift with scale or tuning; the audit sequence itself does not depend on the model, and whether the dissociation replicates elsewhere is a question the method makes directly testable. Mean pooling collapses token-level structure, so effects carried by specific

positions rather than the reflection-level average are invisible to our analyses. Our encoding estimates are therefore conservative with respect to what the stream contains. The logit lens is applied without learned calibration, and intermediate-layer projections can be miscalibrated (Belrose et al., 2023); our lens statistics are within-layer correlations across participants, which limits but does not eliminate the concern. The primary vocabulary analysis is also restricted to corpus-attested content tokens, so it cannot guarantee coverage of every non-occurring or multi-token paraphrase. The full-vocabulary audit was semantically mixed and contained substantial off-corpus noise; frequency-threshold sensitivity further showed that the Final Decision association was stable whereas the absence of detectable Evaluation coupling was filter-dependent (Appendix D). The patching intervention is coarse—a single pooled direction broadcast to every token position—and its endpoint is a weakly cross-validated, fitted model-internal linear readout of win-switch rate, not behavior itself. The intervention result is therefore limited to descriptive changes in that diagnostic and does not establish downstream model use or human effects. The matched-pair design rests on 15 pairs, with uncertainty quantified by bootstrap over pairs; no layer survived correction across the full sweep. The sample is also a single English-language online cohort performing one task, and length residualization assumes a log-linear length effect; departures from either constrain generality. Finally, the study evaluates offline scoring of post-task reflections, so the findings do not establish how the same audit would behave in interactive or open-ended LLM-as-judge deployments.

Ethics Statement

This work is a secondary analysis of de-identified self-reflections and task data collected online with informed consent under the protocol reported by Suthaharan et al. (2026), and made publicly available by those authors; no new human data were collected. The task markers analyzed are linked to paranoia, a sensitive attribute. Nothing in this paper supports individual-level inference, and the central finding cautions against precisely that use. A rubric dimension can appear well measured while remaining only weakly aligned with independently measured task markers, so dimension scores should not inform decisions about individuals without in-

dependent validity evidence. All modeling used the open-weights Gemma-2-2B; rubric scores were reused from prior work.

Data and Code Availability

The de-identified tabular data, analysis notebook, cached pair-level activation-patching outputs, generated statistical summaries, and figure-generation code are available at <https://github.com/psuthaharan/mp-llm-geometry>. The reproduction entry point is `mp_llm_geometry_analysis.ipynb`; it identifies the analyzed model as `google/gemma-2-2b`, verifies the mean-pooled activation checkpoint, and validates the cached pair-level outputs used to reproduce the four analysis figures.

References

- Guillaume Alain and Yoshua Bengio. 2016. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*.
- Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.
- Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*.
- Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. *Amnesic probing: Behavioral explanation with amnesic counterfactuals*. *Transactions of the Association for Computational Linguistics*, 9:160–175.
- Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. 2021. *A mathematical framework for transformer circuits*. Transformer Circuits Thread.
- Stephen M. Fleming and Hakwan C. Lau. 2014. How to measure metacognition. *Frontiers in Human Neuroscience*, 8:443.
- Zorik Gekhman, Eyal Ben David, Hadas Orgad, Eran Ofek, Yonatan Belinkov, Idan Szpektor, Jonathan Herzig, and Roi Reichart. 2025. Inside-out: Hidden factual knowledge in LLMs. *arXiv preprint arXiv:2503.15299*.
- Gemma Team. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

- Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI), Survey Track*, pages 8048–8057.
- Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. 2024. LLM-Rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13806–13834.
- John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743.
- Jinpeng Hu, Ao Wang, Qianqian Xie, Hui Ma, Zhuo Li, and Dan Guo. 2026. AgentMental: An interactive multi-agent framework for explainable and adaptive mental health assessment. *The Fortieth AAAI Conference on Artificial Intelligence*.
- Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. 2024. Position: The platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pages 20617–20642.
- Sarah Jenner, Dimitris Raidos, Emma Anderson, Stella Fleetwood, Ben Ainsworth, Kerry Fox, Jana Kreppner, and Mary Barker. 2025. [Using large language models for narrative analysis: A novel application of generative AI](#). *Methods in Psychology*, 12:100183.
- Li Ji-An, Hua-Dong Xiong, Robert C. Wilson, Marcelo G. Mattar, and Marcus K. Benna. 2025. Language models are capable of metacognitive monitoring and control of their internal activations. *arXiv preprint arXiv:2505.13763*. NeurIPS 2025.
- Sucharit Katyal and Stephen M. Fleming. 2024. [The future of metacognition research: Balancing construct breadth with measurement rigor](#). *Cortex*, 171:223–234.
- Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. LLMs-as-Judges: A comprehensive survey on LLM-based evaluation methods. *arXiv preprint arXiv:2412.05579*.
- Xuan Luo, Weizhi Wang, and Xifeng Yan. 2025. Direct multi-token decoding. *arXiv preprint arXiv:2510.11958*.
- Christoph Mathys, Jean Daunizeau, Karl J. Friston, and Klaas E. Stephan. 2011. A Bayesian foundation for individual learning under uncertainty. *Frontiers in Human Neuroscience*, 5:39.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- nostalgebraist. 2020. [Interpreting GPT: the logit lens](#). LessWrong blog post.
- Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*.
- Mladen Raković, Matthew L. Bernacki, Jeffrey A. Greene, Robert D. Plumley, Kelly A. Hogan, Kathleen M. Gates, and Abigail T. Panter. 2022. [Examining the critical role of evaluation and adaptation in self-regulated learning](#). *Contemporary Educational Psychology*, 68:102027.
- Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. 2024. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*. ArXiv:2312.06681.
- Praveen Suthaharan, Santiago Castiello, Yuen-Siang Ang, and Philip R. Corlett. 2026. [Self-reflection protects behavior from volatile beliefs linked to paranoia](#). *Computational Psychiatry*, 10(1):18–34.
- Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*.
- Yun Wang, Zhaojun Ding, Xuansheng Wu, Siyue Sun, Ninghao Liu, and Xiaoming Zhai. 2026. AutoSCORE: Enhancing automated scoring with multi-agent large language models via structured component recognition. *The Fortieth AAAI Conference on Artificial Intelligence*. Also in *Proceedings of AAAI 2026*.
- Yuqing Wang and Yun Zhao. 2024. Metacognitive prompting improves understanding in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1914–1926, Mexico City, Mexico.
- Xuansheng Wu, Padmaja Pravin Saraf, Gyeonggeon Lee, Ehsan Latif, Ninghao Liu, and Xiaoming Zhai. 2024. Unveiling scoring processes: Dissecting the differences between LLMs and human graders in automatic scoring. *arXiv preprint arXiv:2407.18328*.
- Haoxin Xu, Changyong Qi, Tong Liu, Bohao Zhang, Anna He, Bingqian Jiang, Longwei Zheng, and Xiaoqing Gu. 2026. Multi-agent learning path planning via LLMs. *arXiv preprint arXiv:2601.17346*.

Fred Zhang and Neel Nanda. 2024. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations (ICLR)*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*.

Wentao Zhu, Zhining Zhang, and Yizhou Wang. 2024. Language models represent beliefs of self and others. *arXiv preprint arXiv:2402.18496*.

A Uncertainty in observed maximizing layers

Across 5,000 participant-bootstrap resamples, the 95% intervals for the maximizing layers were L0–L16 for Comprehension, L2–L18 for Judgment, L6–L23 for Evaluation, L5–L25 for Final Decision, and L15–L23 for Confidence. The broad, overlapping intervals indicate that the observed maxima in Figure 2A should not be interpreted as five sharply separated computational stages.

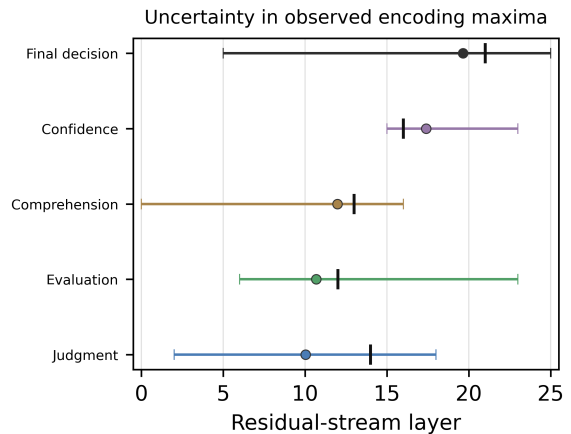


Figure 6: **Bootstrap uncertainty in the maximizing layers of Figure 2A.** Points show the mean maximizing layer across 5,000 participant-bootstrap resamples, horizontal lines show 95% bootstrap percentile intervals, and black ticks mark the observed maximizing layers.

B Cross-dimensional score overlap and probe transfer

Since all five MP dimensions are scored from the same reflection, we examined how much of their apparent separability could be attributed to shared score structure (Figure 7). The GPT-4 scores were strongly correlated across dimensions (mean absolute off-diagonal $r = 0.837$), but this overlap was attenuated after cross-fitted adjustment for reflection length ($r = 0.495$) and was lower again for cross-dimensional transfer of the out-of-fold probe predictions ($r = 0.325$). This progressive reduction suggests that the probes are not simply recovering a common signal carried by all five scores. Evaluation provides the clearest evidence. Its probe prediction correlated more strongly with its matched Evaluation score ($r = 0.579$) than with any off-target dimension (maximum absolute $r = 0.361$), indicating that its strong decodability includes structure specific to Evaluation.

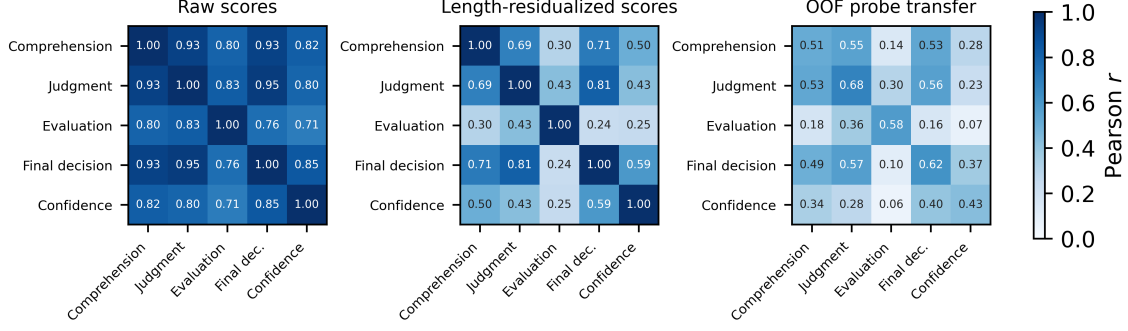


Figure 7: **Score overlap and cross-dimensional probe transfer.** (A) Pearson correlations among the five raw GPT-4 MP scores. (B) Correlations among cross-fitted score residuals after adjusting for log-transformed reflection length within each outer training fold. (C) Cross-dimensional transfer between each dimension’s out-of-fold probe prediction at its observed maximizing layer and all five cross-fitted score residuals. Rows in panel C denote the probe target and columns denote the rubric score.

C Marker-specific coupling underlying Figure 3

The vertical coordinate in Figure 3 summarizes coupling to two distinct task markers, so we examined the signed associations separately to ensure that their mean did not conceal marker-specific effects. Evaluation was weakly related to both the volatility prior ($r_p = -0.063$) and win-switch rate ($r_p = -0.037$), with neither association surviving Holm correction (Table 1). Its low coupling in Figure 3 therefore does not arise from averaging across otherwise divergent marker relationships. The remaining dimensions showed modest negative associations, with six of the ten dimension-by-marker tests surviving familywise correction.

D Token-filter diagnostics and robustness

Intermediate-layer logits are not calibrated probabilities, and correlation-based coupling is invariant to translation and positive rescaling, so absolute magnitude was not treated as evidence that a token was meaningfully represented. For each of the 1,143 corpus-attested tokens, we instead recorded participant-wise ranks and top-100 incidence at L12 and L21. We evaluated Figure 4C at minimum corpus counts of 1, 2, 5, and 10 while holding readability and content-word criteria fixed (Table 2). Final Decision coupling remained stable ($\rho = -0.614$ to -0.675 , all descriptive $p < .001$), whereas Evaluation varied with the threshold. Thus, absence of detectable Evaluation coupling under the primary five-occurrence filter is not invariant to token selection.

We also examined whether the retained tokens

Dimension	Volatility prior	Win-switch rate
Comprehension (L13)	−0.113	−0.193
	[−0.197, −0.018]	[−0.266, −0.111]
	$p_H = .052$	$p_H < .001$
Judgment (L14)	−0.123	−0.157
	[−0.215, −0.027]	[−0.235, −0.080]
	$p_H = .032$	$p_H = .004$
Evaluation (L12)	−0.063	−0.037
	[−0.156, 0.028]	[−0.111, 0.044]
	$p_H = .336$	$p_H = .420$
Final Decision (L21)	−0.143	−0.176
	[−0.231, −0.046]	[−0.265, −0.091]
	$p_H = .011$	$p_H = .001$
Confidence (L16)	−0.099	−0.138
	[−0.183, −0.008]	[−0.211, −0.065]
	$p_H = .089$	$p_H = .013$

Table 1: **Marker-specific coupling underlying Figure 3.** Signed partial correlations between each dimension’s out-of-fold probe prediction at its observed maximizing layer and the two task markers, controlling for log-transformed reflection length ($N = 486$). Values show r_p , participant-bootstrap 95% CIs, and Holm-adjusted p values (p_H) across the ten dimension-by-marker tests.

occupied nontrivial positions in the corpus-relative vocabulary readout. Of the 241 tokens retained by the primary filter, 174 at L12 and 108 at L21 entered the corpus-relative top 100 for at least one participant. The highest-ranking excluded corpus-attested items were predominantly stop-words, punctuation, or unreadable token pieces, rather than omitted content words with uniformly stronger intermediate-layer readouts.

Finally, we audited all 256,000 Gemma vocabulary directions, including absent tokens. Readable directions were word-initial content tokens of at

Minimum count	Tokens	Evaluation ρ	Final Decision ρ
1	984	+0.330, $p < .001$	-0.614, $p < .001$
2	488	+0.205, $p < .001$	-0.647, $p < .001$
5	241	+0.053, $p = .416$	-0.675, $p < .001$
10	126	-0.069, $p = .439$	-0.653, $p < .001$

Table 2: **Sensitivity of token-level rubric-behavior coupling to the minimum-frequency criterion.** Spearman correlations between each token’s rubric coupling and win-switch coupling at the Evaluation (L12) and Final Decision (L21) observed maximizing layers. The five-occurrence threshold is the primary filter used in Figure 4C; conventional token-wise p values are descriptive.

most 24 characters, at least half alphabetic, with English Zipf frequency at least 3, deduplicated after lowercasing. To avoid selecting examples post hoc, we report the first five under this fixed rule. Evaluation’s positive pole was *juvenile*, *cease*, *blockade*, *misdemeanor*, and *prohibiting*; its negative pole was *interested*, *details*, *information*, *Did*, and *Hey*. Final Decision’s positive pole was *horizontal*, *silver*, *shopping*, *smashed*, and *vertical*; its negative pole was *indeed*, *necessarily*, *although*, *nor*, and *moreover*. These mixed extremes make the scan a descriptive transparency check, not confirmation of a uniformly coherent structure or a source for token selection. Complete ranks accompany the analysis outputs.