Operational Readability of Intermediate Decoder Representations

Anonymous authors
Paper under double-blind review

Abstract

Intermediate states of decoder-only language models are often inspected only after a final layer or final answer is available, leaving unclear when a frozen representation first becomes usable for a declared decision. We introduce operational readability as the protocol-relative object of study and instantiate it with Earliest Readable Representation (ERR), the earliest validation-selected layer within $\epsilon = 0.02$ of the best validation score. In a controlled in-distribution memory-compatibility ranking protocol, layer selection is locked on validation data and evaluated on held-out test candidate sets. Across 96 registered conditions, the primary linear readout exceeds a matched conditional-shuffle control in 96/96 paired comparisons. Sensitivity analyses delimit the claim: measured quality is probe-family dependent, normalization differences are sparse, prompt-position contrasts are inconclusive, and CR-IV versus CR-V effects are heterogeneous construction-regime evidence. We therefore present ERR as an auditable measurement relation, not as causal memory-use evidence, a model-intrinsic depth law, or a deployment-speedup result.

1 Introduction

Large decoder-only language models expose many hidden states before they emit a final token, yet most evaluations treat the model as if the only scientifically meaningful object were the final text or the final layer. That view is incomplete for memory-like decisions. A model may hold an intermediate representation from which a declared decision is already recoverable, while the final generative computation continues to transform that state for other purposes. Conversely, a high final answer score does not tell us when the relevant information first became accessible to a readout.

This paper studies that measurement gap. The scientific object is operational readability: a relation between a frozen representation and a declared downstream decision under a specified task, probe family, normalization, prompt position, metric, and split. The measurement is Earliest Readable Representation, or ERR. ERR asks for the first layer whose validation score is within a fixed tolerance of the best validation layer. The protocol is validation-locked candidate-set ranking: layer choice is made on validation examples, and the selected layer is then evaluated on a held-out test split without revisiting the choice.

Figure 1 summarizes the measurement ordering. The diagram is a protocol diagram, not a causal model. It places the declared task and representation site upstream of normalization, probe family, validation selection, and locked test evaluation. This ordering matters because it prevents a selected layer from being chosen by test performance and because it keeps the interpretation tied to the declared readout rather than to an intrinsic layer property.

The empirical setting is deliberately controlled. Each query is paired with a memory-candidate set, and the readout must rank the reusable candidate above competing memories. The registered grid spans three task families–attribute-value matching, constraint alignment, and procedural applicability–along with four small public checkpoints, two construction regimes, and four prompt positions. The primary representation site is the block post-MLP residual state. The primary readout is a linear interaction ranker over query and memory states, using per-token L_2 normalization and MRR as the registered score.

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

The central positive evidence is not the aggregate score alone. It is the matched shortcut control. Across 96 registered controlled in-distribution conditions, the primary linear measurement exceeds the conditional-shuffle baseline in 96/96 paired comparisons. Conditional shuffling preserves much of the marginal query and memory distribution while breaking the matched query-memory relation, so this result supports a pair-specific readability signal under the declared protocol.

The same evidence package also narrows the claims. MLP probes achieve higher measured quality in 192/192 registered probe-family contrasts, so readability is probe-family dependent in this measurement. Normalization differences are sparse and endpoint-specific; registered prompt-position contrasts are inferentially inconclusive; and CR-IV versus CR-V comparisons are heterogeneous construction-regime contrasts rather than a pure difficulty intervention.

The contributions are fourfold. First, we define operational readability as a protocol-relative representation measurement. Second, we instantiate it with validation-locked ERR for memory-compatibility ranking. Third, we provide a registered evidence package with paired uncertainty, matched shortcut controls, and sensitivity analyses. Fourth, we release audit artifacts that trace claims, numbers, figures, and reproduction steps to machine-readable evidence while explicitly blocking unsupported causal, universal-depth, and deployment claims.

2 Related Work

Diagnostic probing asks what information is recoverable from a representation by an auxiliary readout. Surveys of probing emphasize both its usefulness and its interpretive risks: a probe can show accessibility under a chosen measurement relation, but it does not by itself show how the base model uses that information (Belinkov, 2022). Structural probing makes this relation especially clear by defining a geometric readout for syntactic structure (Hewitt & Manning, 2019). Our use of a linear interaction ranker follows the same measurement spirit, but the target is query-memory compatibility rather than syntax.

A second line of work critiques probe capacity and selectivity. Control-task arguments show that a flexible probe can appear successful for reasons unrelated to the intended linguistic or semantic construct (Hewitt & Liang, 2019). Information-theoretic probing similarly treats readout design and compression as part of the measurement rather than as a neutral detail (Pimentel et al., 2020). We adopt that caution directly. Probe family is not hidden as a nuisance parameter; it is reported as a sensitivity dimension, and the MLP comparison is interpreted as evidence that measured quality depends on readout capacity.

Layerwise representation studies show that information can appear, transform, and disappear across depth. Linear classifiers have been used to inspect intermediate neural network layers (Alain & Bengio, 2017). Contextual encoder analyses show that linguistic information is distributed unevenly across layers and tasks (Liu et al., 2019; Tenney et al., 2019). ERR is related to these layerwise measurements but differs in its selection rule: the earliest layer is selected only on validation data with a declared tolerance, then evaluated on locked test examples.

Memory-oriented analyses give useful contrast. Transformer feed-forward layers have been interpreted as key-value memories (Geva et al., 2021), and causal editing work localizes factual associations for intervention (Meng et al., 2022). Our task also uses memory language, but the claim is narrower. We do not identify stored facts, edit activations, or infer a circuit. We ask whether a frozen representation at a declared prompt position supports a compatibility decision between a query and a memory candidate.

The distinction between accessibility and causal use is central. Causal mediation and intervention studies ask whether modifying an internal component changes behavior (Vig et al., 2020). Probing-style readability does not answer that causal question. A successful readout can be scientifically useful as a measurement while remaining silent about whether the model uses the same information path during generation.

Finally, ERR should be distinguished from early exiting and adaptive computation. Depth-adaptive Transformers and confident adaptive language modeling learn policies for stopping computation early or saving inference cost (Elbayad et al., 2020; Schuster et al., 2022). ERR is not an exit policy and is not evaluated for deployment acceleration. It is a diagnostic endpoint for comparing representations under a fixed measurement protocol.

3 Operational Readability

Operational readability is a declared measurement relation. Let $\mathcal{T}$ denote a candidate-set decision task, $\mathcal{P}$ a probe family, $\mathcal{N}$ a normalization rule, b a prompt position, s a split, m a metric, and $h_{\ell,b}$ the frozen representation at layer ℓ and position b. We define the readability profile as

$$R_{\mathcal{T},\mathcal{P},\mathcal{N},b,s,m}(\ell) = \mathrm{Score}_m\left(\mathcal{P}(\mathcal{N}(h_{\ell,b})); \mathcal{T}, s\right).$$

Every subscript is part of the measurement. Changing the task, prompt position, probe, normalization, split, metric, or representation site changes the operational quantity being measured.

For tolerance ϵ, Earliest Readable Representation is

$$\mathrm{ERR}_\epsilon = \min\left\{\ell : R^{\mathrm{val}}(\ell) \geq \max_j R^{\mathrm{val}}(j) - \epsilon\right\}.$$

The selected layer is then frozen before test evaluation. ERR is therefore an earliest-within-tolerance validation selection rule, not a search over the test split.

The interpretation is deliberately modest. If a representation is readable, the declared readout can recover the target decision under the declared data split and metric. This does not imply that the base model causally uses that information, that another probe would select the same layer, or that the selected layer is intrinsic to the checkpoint. In fact, the registered probe-family sensitivity shows that MLP readouts obtain higher measured ranking quality in 192/192 comparisons, so the paper treats probe dependence as part of the estimand.

This framing makes failures informative. A prompt-position contrast that includes zero is not an equivalence result; it is an inconclusive directional result under the registered interval rule. A heterogeneous CR-IV versus CR-V comparison is not cleaned up into a difficulty law. The point of the protocol is to keep the claim attached to the evidence that actually passed.

4 Method

4.1 Task and Data Construction

Each observation is a query-memory candidate pair with a binary reusable label. A query specifies a current user need. A memory candidate contains prior content that may or may not directly satisfy that need. A candidate is labeled reusable only when returning or relying on that memory would fully satisfy the current query without violating material constraints. Partial overlap, stale entities, changed time scope, changed audience, changed output requirements, or changed procedural requirements make a candidate non-reusable.

The data are organized along three task families. The attribute-value matching family tests factual or value-bearing compatibility, where a candidate may contain the right object but an incompatible value. The constraint-alignment family tests whether semantic constraints such as entity, audience, time, format, or safety requirements remain aligned. The procedural-applicability family tests whether a remembered procedure remains applicable to the requested operation. These families are not claimed to exhaust memory use; they provide controlled decision families for the registered measurement.

Each query candidate set contains exactly one positive item. This one-positive label grain makes reciprocal rank the natural ranking endpoint: AP equals reciprocal rank within a

candidate set, so MAP is redundant with MRR. Hit@1 and NDCG remain useful diagnostics but are not the primary score. Table 2 reports the candidate-set and pair counts regenerated from the label-grain audit.

We use two construction regimes. Construction Regime IV (CR-IV) uses k=8 candidate sets, and Construction Regime V (CR-V) uses k=10 candidate sets. They also differ in how hard negatives are constructed. The machine-readable supplement records the original construction keys for exact reproducibility. Because candidate-set size and construction change together, the CR-IV versus CR-V analysis is not interpreted as a pure causal difficulty intervention.

Construction filters remove invalid candidate sets, enforce one positive item per set, and preserve split grouping so that related examples do not leak across train, validation, and test. The split audit records training and split seed semantics separately. Synthetic and API-generated components are treated as controlled data-generation inputs whose provenance is retained in parent manifests; the present V6 release does not add new examples or change labels.

4.2 Models and Representation Extraction

The registered checkpoints are qwen2-0.5b, smollm2-360m, llama3.2-1b, and gemma3-1b as recorded in the parent extraction manifests. They are used as small public decoder-only checkpoints for a controlled measurement study, not as evidence of architecture-general behavior. Layer counts and model family keys are stored in the confirmatory score table and extraction metadata.

Representations are extracted at the canonical block post-MLP residual state. In model-specific code this site corresponds to the residual stream immediately after the feed-forward update has been added for a transformer block. Layer indexing in the artifacts is zero-based. For each layer, the state is recorded at a declared prompt position: context terminus, query terminus, turn terminus, or prompt terminus. These correspond to the artifact keys pre_context_end, query_end, user_turn_end, and pre_generation. The context terminus is the final token of the context/memory block before the query content; the query terminus is the final semantic query token; the turn terminus is the final token of the user turn after query formatting; and the prompt terminus is the final prompt token before decoding would begin.

The extraction path is deterministic with respect to the stored features. V6 does not rerun hidden-state extraction. It inherits the v4 corrected feature and score artifacts, and it records their hashes as the scientific parent. This is intentional: V6 is a manuscript and release-engineering repair, not a new scientific campaign.

4.3 Feature Construction and Primary Readout

For a query state $q_{\ell,b}$ and memory state $r_{\ell,b}$ at the same layer and prompt position, the primary feature map contains the normalized query state, normalized memory state, elementwise product, and absolute difference. Under the primary normalization, each token state is scaled by its L_2 norm before feature construction. The linear interaction ranker assigns a scalar score

$$f_{\ell,b}(q, r) = w^\top [q; r; q \odot r; |q - r|] + c.$$

The ranker is trained on the training split for each registered condition. Layer selection uses validation MRR and $\epsilon = 0.02$. Test MRR is read only after the validation-selected ERR layer has been locked.

The linear model uses the deterministic solver and regularization settings recorded in the parent artifacts. Training seed and split seed have different roles: the split seed determines grouped split assignment, while training seeds affect stochastic readout training where applicable. The primary linear readout is deterministic in the locked parent rows.

4.4 MLP Sensitivity Readout

The MLP sensitivity readout is a registered alternate probe family with five training seeds per condition and normalization. It uses a hidden nonlinear layer, bounded optimization, and the same validation-layer selection principle. Its purpose is not to replace the primary readout but to test whether measured quality depends on probe capacity.

No V6 refit was performed. All 23 warning fits were retained with disclosure; excluding them preserves the MLP-higher decision in 192/192 recomputed contrasts. The convergence disclosure is therefore a warning-retention and sensitivity analysis, not a repair of the locked scientific fits.

4.5 Controls and Baselines

The primary matched control is conditional shuffle. It preserves the candidate-set structure and marginal feature distributions while breaking the aligned query-memory pairing needed for the compatibility decision. A positive primary-minus-conditional-shuffle interval supports pair-specific readability under the registered task.

Other baselines provide context. Query-only and memory-only readouts test whether one side alone carries enough signal. Scalar-only baselines test surface covariates. Label-shuffled probes test training-label dependence. Analytic-random baselines represent random ranking under the one-positive candidate-set grain. Text-embedding and embedding-state baselines provide non-hidden-state comparison points. Final-block and post-final-norm comparisons test whether late representation sites dominate the registered intermediate measurement.

5 Measurement Protocol

Table 4 defines the registered grid. It contains three task families, four checkpoints, two construction regimes, and four prompt positions, producing 96 primary conditions. Each condition has a validation readability profile over layers. ERR is selected as the earliest layer within $\epsilon = 0.02$ of the best validation MRR. The selected layer is then evaluated on the locked test split.

The validation-lock rule is the central anti-contamination guard. The test split is not used to choose the layer, prompt position, model, normalization, or example subset. The final audit checks that selected-on fields are validation-only and that test-profile fields are not used for selection.

Uncertainty is estimated with paired bootstrap resampling over query-candidate-set identifiers. The unit is the candidate set rather than the individual pair because all candidates within a query set share a ranking decision. Each registered interval uses 5,000 draws and a 95% confidence level unless the parent artifact explicitly states a noncomputed status. Matched contrasts align bootstrap arms by candidate-set identifier. When groups cannot be aligned, unmatched counts are recorded and the row is not silently treated as comparable.

Directional decisions use interval exclusion of zero. If the confidence interval for arm A minus arm B is strictly above zero, A is higher; if it is strictly below zero, A is lower; otherwise the result is inconclusive. This rule does not establish equivalence. Equivalence would require a separately declared margin and interval rule, and the present V6 manuscript does not use non-directional intervals as equivalence evidence.

Denominators are reported at their analysis grain. Conditional shuffle uses condition-level comparisons. Probe-family sensitivity uses condition-by-normalization comparisons. Normalization uses condition-by-endpoint comparisons. CR-IV versus CR-V quality uses endpoint-level construction-regime comparisons. This is why the same paper can report 96 registered primary conditions and 96 construction-regime quality endpoint contrasts without treating them as the same denominator.

The claim ledger is part of the measurement protocol. Every empirical result sentence in the manuscript is assigned a claim identifier, row selector, evidence artifact, denominator,

endpoint, allowed wording, and forbidden wording. The manuscript audit fails when a cited empirical claim lacks ledger coverage or when a forbidden causal, universal, or deployment interpretation appears in the text.

The release pipeline treats manuscript generation and manuscript authorship separately. Numerical macros, generated tables, figure manifests, and audit ledgers are regenerated by code from the locked parents. Prose sections are preserved once they exist, so later pipeline runs cannot silently rewrite scientific wording or move a claim outside its ledger coverage. This separation is important because the paper is short enough for manual scientific review, but the numerical surface is large enough that hand-entered values would be a contamination risk.

6 Experiments

6.1 Registered Design and Data Summary

The registered design uses the controlled candidate-set task described in Section 4. Table 2 reports the split and candidate-set counts. Every candidate set has one positive item, which makes MRR the primary ranking endpoint and makes MAP redundant under the current label grain.

The experiment order follows the registered evidence hierarchy rather than the order in which artifacts were created. We first report the locked primary measurement, then the matched shortcut control, then contextual baselines and sensitivities. This ordering prevents the reader from treating later exploratory-looking displays as the source of the headline claim. It also makes clear when a denominator changes from a registered condition to a condition-by-normalization or condition-by-endpoint contrast.

6.2 Primary Locked Performance

Table 1 summarizes the headline evidence. Across 96 validation-locked conditions, the mean locked test MRR is 0.733. This aggregate is descriptive. It is not used to claim architecture-general performance because only four small checkpoints are evaluated and all examples are from the controlled in-distribution construction.

6.3 Conditional-Shuffle Matched Control

The primary measurement exceeds the matched conditional-shuffle control in 96/96 registered condition-level paired comparisons. The estimand is primary locked-test MRR minus conditional-shuffle locked-test MRR; the uncertainty unit is the query candidate set. This supports pair-specific operational readability because the baseline disrupts the query-memory pairing while retaining much of the marginal structure.

6.4 Other Baselines

Figure 3 reports baseline contrast counts. Other baselines are contextual rather than the headline matched control. Query-only, memory-only, scalar-only, label-shuffled, analytic-random, text-embedding, embedding-state, final-block, and post-final-norm comparisons help diagnose surface shortcuts and alternate representation sites. Their decisions are heterogeneous, so they are used to delimit the interpretation rather than replace the conditional-shuffle result.

The analytic-random baseline is especially useful for interpreting MRR under a one-positive candidate set. It gives a chance-ranking reference at the same candidate-set size, while the conditional-shuffle baseline tests whether the matched query-memory relation remains necessary. These two baselines answer different questions and are not interchangeable.

Table 1: Headline evidence regenerated from locked parent artifacts. Directional counts use condition-level paired intervals except where the endpoint is explicitly condition-by-normalization or condition-by-endpoint.

Endpoint	Result	Permitted interpretation
Mean locked test MRR	0.733	descriptive aggregate within the registered grid
Conditional-pairing shuffle	96/96 positive	matched shortcut-control support
Readout-family quality	192/192 MLP higher	measured quality is readout-family dependent
Prompt-position contrasts	0/24 directional	no registered prompt-position law established

6.5 ERR Localization

Figure 2 visualizes validation-selected relative ERR depth for all registered conditions. The figure is descriptive. It shows where the registered readout becomes competitive with the best validation layer, but it is not a test-selected depth law. Weak-localization markers identify intervals whose registered uncertainty is wide enough that exact layer interpretation should be cautious.

6.6 Probe-Family Sensitivity

MLP probes obtain higher measured locked-test quality in 192/192 condition-by-normalization contrasts. The estimand is linear quality minus mean registered MLP-seed quality. The result means that operational readability is probe-family dependent under this protocol. It does not mean that the primary linear ranker was retroactively replaced.

6.7 Convergence-Warning Disclosure

Table 3 collects the sensitivity findings. The MLP analysis contains 23/960 registered fits with convergence warnings. No V6 refit was performed. The warning fits were retained with disclosure, and an exclusion sensitivity recomputed the probe-family contrast after removing all unresolved warning fits. The qualitative decision remains MLP higher in 192/192 recomputed contrasts.

6.8 Normalization Endpoints

Normalization sensitivity is sparse and endpoint-specific. For locked quality, raw is higher in 1/96 conditions and per-token L_2 is higher in 1/96 conditions; the remaining 94 are inconclusive. For localization, 0/96 absolute ERR-layer contrasts and 0/96 relative-depth contrasts are directional. These counts do not support a broad normalization law.

6.9 Prompt-Position Result

Registered test prompt-position contrasts remain inconclusive. The analysis finds 0/24 directional prompt-position contrasts. Figure 6 therefore provides descriptive point estimates and intervals rather than evidence for a universal prompt-position law.

6.10 Construction-Regime Contrast

Figure 4 summarizes the CR-IV versus CR-V construction-regime quality evidence, and Figure 5 gives the full condition-level intervals in the appendix. The endpoint-level pattern is heterogeneous: 7 lower-quality intervals, 3 higher-quality intervals, and 86 inconclusive intervals across 96 comparisons. Because the regimes differ in construction and candidate-set size, this result is not a pure causal difficulty test and does not support a universal deeper ERR shift.

7 Discussion

The evidence supports a narrow but useful claim: under a declared controlled protocol, intermediate decoder states contain pair-specific information that a registered readout can use for memory-compatibility ranking. The conditional-shuffle result is the cleanest support because it tests the matched query-memory relation rather than a marginal query or memory shortcut.

The same result package argues against a simpler story in which there is a fixed model-intrinsic memory-ready layer. ERR is a measurement outcome conditioned on task, prompt position, normalization, readout, metric, and split. Treating probe dependence as part of the estimand makes this explicit. A stronger MLP result does not invalidate the linear measurement; it shows that the chosen readout family changes the measured quality and must be reported.

Operational readability is therefore best used as a comparison tool. It makes representation-access claims falsifiable under controlled changes to readouts, prompt positions, normalizations, and data construction. Future work can use the same protocol to test held-out providers, natural memory logs, or causal interventions, but those are separate claims that require separate evidence.

This framing also explains why the negative and heterogeneous results are not secondary. A measurement paper is credible only if the estimand survives contact with its own controls. The prompt-position analysis, for example, does not license a prompt-position law; it says that under the registered intervals the tested position differences did not become directional. The normalization analysis similarly says that quality differences appeared rarely and localization differences did not become directional. These results help specify where the current evidence stops.

The conditional-shuffle finding should be read in that same disciplined way. It shows that the aligned query-memory relation matters to the registered readout. It does not identify the internal algorithm used by the base checkpoint, and it does not show that generation would fail if the measured state were perturbed. The result is still useful because it is matched, paired, and auditable: a future dataset, readout, prompt position, or intervention can be compared against the same declared measurement relation.

8 Limitations

The scope is controlled and in-distribution. The paper evaluates four small checkpoints and does not establish provider-held-out, dataset-held-out, or architecture-general behavior. The candidate-set task has exactly one positive item per set, which is useful for clean MRR interpretation but narrower than many real memory-reuse settings.

CR-IV and CR-V are construction regimes, not an isolated difficulty intervention. Candidate-set size, negative construction, and apparent difficulty change together. The analysis therefore reports heterogeneous construction-regime evidence and avoids causal difficulty wording.

The MLP sensitivity includes 23 retained convergence-warning fits. V6 does not refit them. The disclosure is acceptable only because excluding unresolved warning fits preserves the qualitative probe-quality result, and the warning count remains visible in the paper and supplement.

The study is not a causal intervention or circuit analysis. It does not prove that the base model uses the readable information during generation. It also does not evaluate early-exit reliability, deployment acceleration, or compute savings. Finally, Tier A reproduction begins from cached analytical inputs and derived score arrays. Full hidden-state extraction from model checkpoints is documented as the heavier Tier B path rather than as the one-command reviewer path.

The construction is synthetic and controlled. This improves auditability but limits ecological validity. Real memory systems include multiple relevant memories, stale or conflicting

records, implicit user preferences, tool outputs, privacy filters, and long histories whose prompt positions are not as clean as the registered templates. The present candidate-set grain is therefore a stress-tested measurement setting, not a full model of deployed memory retrieval.

The representation site is also fixed. The primary site is the block post-MLP residual state, and V6 does not claim that attention outputs, pre-MLP residual states, post-final normalization, or token-averaged states would give the same ERR. Site comparisons are treated as baselines or sensitivity checks where available. A broader site sweep would be a separate experiment with its own selection and uncertainty rules.

Finally, the release process separates analytical reproducibility from platform reproducibility. The anonymous Tier A package can regenerate the registered tables and audits from cached inputs. It does not guarantee that every model checkpoint remains downloadable, that every hardware backend returns bit-identical hidden states, or that API-generated construction stages remain available. Those risks are recorded as boundaries of the release rather than hidden inside a single reproducibility claim.

9 Conclusion

Operational readability treats intermediate representations as measurable under a declared task, prompt position, normalization, probe, metric, and split. ERR instantiates this idea as a validation-locked earliest-layer rule. In the controlled registered grid, the primary measurement beats the matched conditional-shuffle control in every condition-level comparison, while the sensitivity analyses keep the conclusion scoped to the evidence.

The main lesson is therefore methodological as much as empirical. Intermediate states can be read for a declared memory-compatibility decision, but the resulting layer is a protocol-relative measurement outcome. Reporting the matched control, probe-family dependence, normalization sensitivity, prompt-position null result, construction-regime heterogeneity, and convergence disclosure makes the claim narrower, and also more reproducible.

Future work can extend the same object-measure-protocol hierarchy to natural memory logs, held-out data generators, larger and more diverse checkpoints, alternate representation sites, and causal interventions. Those extensions should be evaluated as new registered measurements rather than as assumptions smuggled into the current result.

This leaves the present paper with a deliberately limited but inspectable conclusion: the registered readout finds pair-specific readability beyond a matched shuffle control, and every broader interpretation is deferred until a protocol designed for that interpretation is run.

That discipline is the release standard as well as the scientific standard: the artifact should make it easy to reproduce what is claimed and equally easy to see what has not been claimed.

This limited scope is intentional.

References

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In International Conference on Learning Representations Workshop, 2017. URL https://arxiv.org/abs/1610.01644.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. Computational Linguistics, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422.

Maha Elbayad, Jiatao Gu, Edouard Grave, and Michael Auli. Depth-adaptive transformer. In International Conference on Learning Representations, 2020. URL https://arxiv.org/abs/1910.10073.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In Proceedings of EMNLP, pp. 5484–5495, 2021. URL https://aclanthology.org/2021.emnlp-main.446/.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In Proceedings of EMNLP-IJCNLP, pp. 2733–2743, 2019. URL https://aclanthology.org/D19-1275/.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In Proceedings of NAACL-HLT, pp. 4129–4138, 2019. URL https://aclanthology.org/N19-1419/.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. Linguistic knowledge and transferability of contextual representations. In Proceedings of NAACL-HLT, pp. 1073–1094, 2019. URL https://aclanthology.org/N19-1112/.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In Advances in Neural Information Processing Systems, 2022. URL https://arxiv.org/abs/2202.05262.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. Information-theoretic probing for linguistic structure. In Proceedings of ACL, pp. 4609–4622, 2020. URL https://aclanthology.org/2020.acl-main.420/.

Tal Schuster, Adam Fisch, Tommi Jaakkola, and Regina Barzilay. Confident adaptive language modeling. In Advances in Neural Information Processing Systems, 2022. URL https://arxiv.org/abs/2207.07061.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In Proceedings of ACL, pp. 4593–4601, 2019. URL https://aclanthology.org/P19-1452/.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In Advances in Neural Information Processing Systems, 2020. URL https://arxiv.org/abs/2004.12265.

Table 2: Candidate-set data summary regenerated from the parent label-grain audit. Each candidate set contains exactly one positive memory, so AP equals reciprocal rank and MAP is redundant with MRR under this label grain.

Task family	Regime	Train sets	Val sets	Test sets	Cand. size	Pair rows
Attribute-value matching	CR-IV	485	103	102	8–8	5520
Attribute-value matching	CR-V	430	94	88	10–10	6120
Constraint alignment	CR-IV	489	105	105	8–8	5592
Constraint alignment	CR-V	465	102	102	10–10	6690
Procedural applicability	CR-IV	410	89	87	8–8	4688
Procedural applicability	CR-V	360	80	84	10–10	5240
All	All	–	–	–	–	33850

Table 3: Sensitivity findings and scoped release interpretation. The convergence row states the observed V6 action exactly: no refit was performed; warnings were retained and an exclusion sensitivity was recomputed.

Analysis family	Count	Allowed wording
MLP convergence	23/960 retained; exclusion 192/192 MLP higher	warning fits retained with disclosure; exclusion result unchanged
Normalization quality	raw 1, L2 1, inconclusive 94	sparse endpoint-specific quality differences
Normalization ERR layer	0/96 directional absolute; 0/96 directional relative	no registered localization effect established
CR-IV vs CR-V quality	lower 7, higher 3, inconclusive 86	heterogeneous construction-regime evidence

Reproducibility Statement

The anonymous supplement provides a Tier A analytical reproduction path. After extracting the supplement, run python reproduce.py –tier analytical –strict. This command extracts the archived source and declared analytical inputs, creates a clean virtual environment, installs the locked analytical dependencies and source wheel, regenerates V6 numbers, tables, ledgers, figures, and audits, runs the supplement test inventory, and compares mandatory deterministic hashes.

Tier B full reproduction starts earlier, at data construction and hidden-state extraction from model checkpoints. It is heavier and is documented through the source configurations, manifests, model identifiers, split records, extraction sites, and cached arrays. Tier B is not the lightweight one-command path.

AI-use Statement

Generative AI tools were used to support conceptual framing, methodology feedback, software implementation and debugging, result interpretation, figure preparation, literature work, and manuscript drafting and editing. All empirical results reported in the paper were generated by the registered analysis pipeline and checked against machine-readable artifacts. The author or authors reviewed the AI-assisted work and take responsibility for the submission.

A Appendix

A.1 Protocol display

Figure 1 and Table 4 give the measurement-ordering diagram and registered protocol grid referenced in the main text.

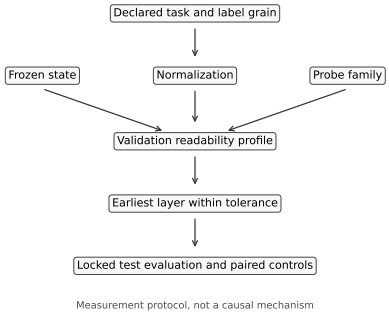

Figure 1: Measurement framework. The diagram shows validation selection before locked test evaluation and represents a measurement protocol, not a causal model.

Table 4: Registered measurement design. ERR is selected on validation data and evaluated on locked test data; all factors are part of the estimand rather than post-hoc choices.

Factor	Registered setting
Task families	attribute-value matching, constraint alignment, and procedural applicability
Checkpoints	Qwen2.5 0.5B, SmolLM2 360M, Llama 3.2 1B, and Gemma 3 1B
Construction regimes	CR-IV (k=8) and CR-V (k=10); original construction keys are recorded in the supplement
Prompt positions	context terminus, query terminus, turn terminus, and prompt terminus
Representation site	block post-MLP residual state; layer indexing is zero-based in artifacts
Primary readout	linear interaction readout with per-token L_2 normalization
Selection/evaluation	validation selection, locked test evaluation, $\epsilon = 0.02$
Uncertainty	5,000 paired bootstrap draws at query-candidate-set level, 95% intervals
Registered cells	96 conditions

A.2 Registered Result Displays

Figures 2, 3, and 4 provide the registered visual summaries referenced in the main text.

A.3 Construction-Regime Forest

Figure 5 shows every registered CR-IV versus CR-V quality interval used in the main-text summary.

A.4 Prompt-Position Contrasts

Figure 6 lists the registered test-split prompt-position contrasts.

A.5 Audit artifacts

The supplement includes the parent freeze, scientific snapshot, claim ledger, literature evidence ledger, figure manifest, PDF manifest, convergence disclosure, source-isolated reproduction manifest, and final release-gate audit.

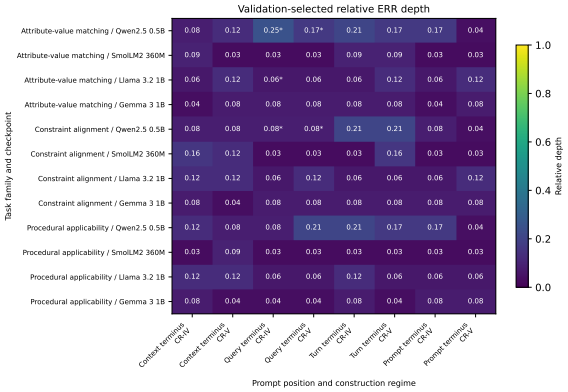

Figure 2: ERR localization grid for N=96 registered conditions under the primary linear probe, per-token L_2 normalization, and $\epsilon = 0.02$. The color scale is fixed to relative depth $[0, 1]$; stars mark weak localization.

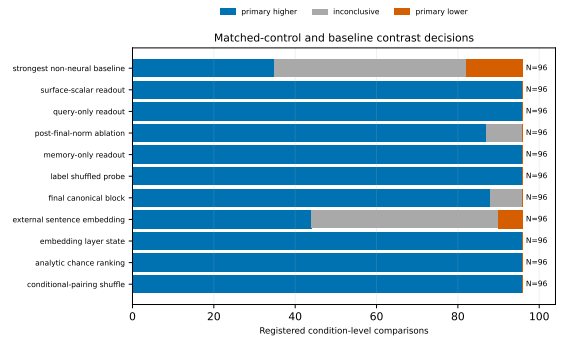

Figure 3: Baseline contrast counts. Positive means the primary-minus-baseline 95% paired interval is strictly above zero; negative means it is strictly below zero; inconclusive intervals include zero. Conditional shuffle is the primary matched control.

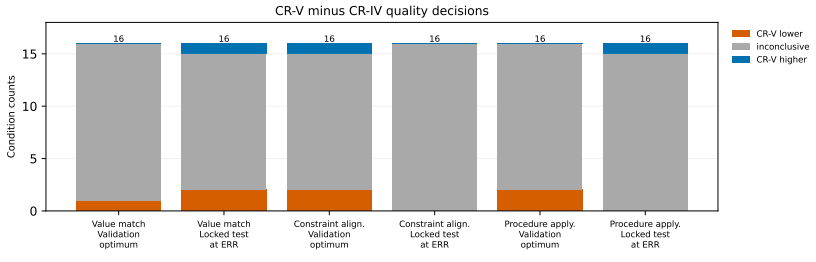

Figure 4: Construction-regime summary. Counts are CR-V minus CR-IV quality decisions by endpoint and task family; the display communicates heterogeneity without implying a pure difficulty intervention.

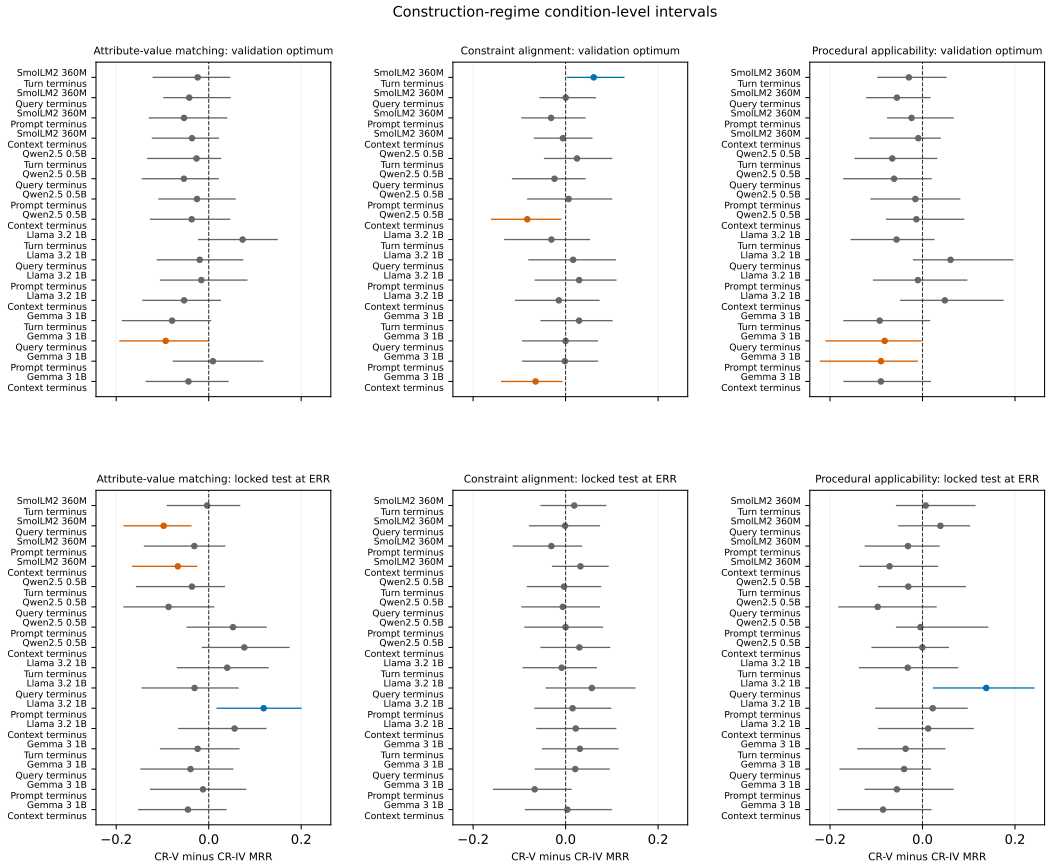

Figure 5: Appendix construction-regime forest. Each interval is CR-V minus CR-IV MRR for the labeled task family, checkpoint, prompt position, and endpoint. Intervals crossing zero are inconclusive; the figure visualizes the heterogeneity summarized in Figure 4.

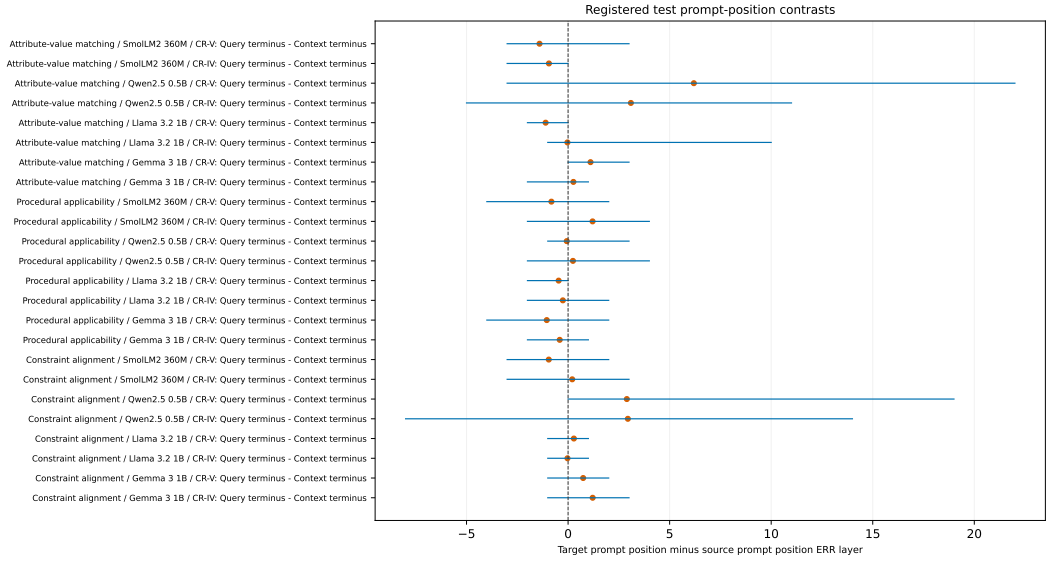

Figure 6: Prompt-position contrasts for all 24 registered test comparisons. All intervals include zero; the figure therefore provides descriptive point estimates rather than evidence for a prompt-position law.