MANIFESTATION UNITS: REPRESENTING MECHANISTIC FINDINGS FOR ACCESS AND INTERVENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Mechanistic interpretability produces heterogeneous findings about model components, relationships, and causal effects, but these findings are often left in analysis-specific forms that complicate downstream reuse. We introduce **Manifestation Units**, a persistent structured interface that organizes findings through shared semantic roles while retaining architecture-specific evidence. We evaluate MUs through *Structural Accessibility*, whether recorded findings can be recovered under bounded access, and *Causal Utility*, whether represented numerical information preserves downstream component selection and intervention effectiveness. Across VAE, CNN, and GPT-2 settings, the MU pipeline achieves the highest answer-slot accuracy among nine matched BM25 conditions (54.58%, 79.44%, and 79.72%). In the CNN, MU-mediated access reaches 92.37% targeted success, while complete-access MU and native findings produce identical selections and activation-patching effects. In GPT-2, MU preserves the complete native head ranking and seven-head selection, which achieves 61.87% normalized recovery on fresh IOI examples versus 6.30% for layer-matched random controls. These results support treating persistent representation between mechanistic analysis and downstream use as an explicit and independently evaluable layer.

1 INTRODUCTION

Mechanistic interpretability aims to explain model behavior through internal representations, relationships, and causal mechanisms. The field studies a wide range of units and procedures, including neurons, attention heads, learned features, distributed directions, circuits, and targeted interventions. This diversity has motivated several efforts toward greater consistency. Causal abstraction provides a common formal language that covers many mechanistic interpretability methods (Geiger et al., 2025); surveys organize the field around different causal units and search procedures (Mueller et al., 2026); and benchmarking and auditing work has emphasized the need for more consistent evaluation across methods (Mueller et al., 2025; Lan et al., 2026).

These efforts mainly address how mechanisms are defined, discovered, compared, or evaluated. We focus on a complementary problem that arises after a mechanistic analysis has produced a finding: *how should that finding be represented so that it can be accessed and reused later?* This problem is important because mechanistic analyses produce findings in different forms. A VAE analysis may identify informative latent dimensions, a CNN analysis may characterize filters and their activation patterns, and a transformer analysis may locate attention heads or other components involved in a computation. The entities, measurements, relationships, and interventions associated with these findings therefore depend on both the analysis and the architecture. Existing efforts toward common formalization and evaluation make this heterogeneity explicit (Geiger et al., 2025; Mueller et al., 2026; 2025), but do not themselves provide a shared representation for carrying the resulting findings from analysis to later downstream use.

We introduce **Manifestation Units (MUs)** as a persistent interface between mechanistic analysis and downstream use. By *persistent*, we mean that a finding is represented in a form that can be stored, accessed, and reused after the analysis that produced it. An MU first represents a finding through five shared roles: **Entity**, **Semantics**, **Relationships**, **Dynamics**, and **Guidance** (E/S/R/D/G). These

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.

roles capture, respectively, what component the finding concerns, what is known about it, how it relates to other components, how its behavior changes, and how the finding can be used downstream. The shared representation retains the evidence, measurements, and provenance associated with the original analysis.

The shared MU representation does not replace architecture-specific information. Instead, it provides an intermediate representation from which architecture-specific fields can be constructed for downstream use. A latent dimension, a convolutional filter, and an attention head therefore remain different mechanistic objects, but their findings can be organized through the same shared roles before being realized in the fields required by their respective architectures. In this work, E/S/R/D/G defines the common representation that we evaluate. We evaluate whether this representation supports downstream use; we do not attempt to establish that these five roles are the only possible or minimal decomposition.

We evaluate this interface through two complementary properties. *Structural Accessibility* asks whether information represented in a finding can later be reliably accessed through the representation. *Causal Utility* asks whether the represented information remains actionable for downstream mechanistic intervention. The latter distinction is important because useful representation cannot be established from interpretability or accessibility alone: recent work has separately evaluated whether learned representations support control over model behavior (Makelov et al., 2025). Our evaluation therefore separates access to represented information from its subsequent use in model intervention.

We study these properties across three complementary settings: latent representations in a VAE, convolutional filters in a CNN, and attention heads in GPT-2. Across all three settings, MUs provide the highest mean answer-slot accuracy among matched BM25 presentation-and-access pipelines. The CNN and GPT-2 experiments further show that the represented numerical information can recover deterministic component selections that remain effective when applied back to the model. Together, these experiments test whether a common representation can make heterogeneous mechanistic findings accessible while preserving the information required for downstream use.

Interactive demo. A browser-based companion demo provides direct inspection of evaluation questions, representation conditions, accessed evidence, and recorded experimental outputs.[1]

Our contributions are:

1. **A persistent interface for heterogeneous mechanistic findings.** We introduce Manifestation Units, which represent findings through shared E/S/R/D/G roles while retaining the evidence, measurements, provenance, and architecture-specific information produced by the original analysis.
2. **A separation between shared representation and architecture-specific realization.** MUs provide a common intermediate representation from which architecture-specific fields can be constructed for downstream procedures, allowing heterogeneous findings to share a representation without removing the distinctions between their underlying components.
3. **Structural Accessibility and Causal Utility as complementary evaluation targets.** Across VAE, CNN, and GPT-2 settings, we evaluate whether represented findings remain accessible under matched downstream access conditions and whether their retained numerical information remains sufficient for deterministic selection and effective intervention.

2 Manifestation Units as a Persistent Interface

2.1 From analysis to persistent representation

Manifestation Units define a representation layer between upstream mechanistic analysis and downstream use. An upstream analysis produces findings whose entities, measurements, relationships, and supported interventions depend on the analysis procedure and model architecture. MUs organize these heterogeneous findings into a shared representation, from which architecture-specific fields can be constructed for later access and use.

[1] https://manifestation-xai.github.io/manifestation/cnn/

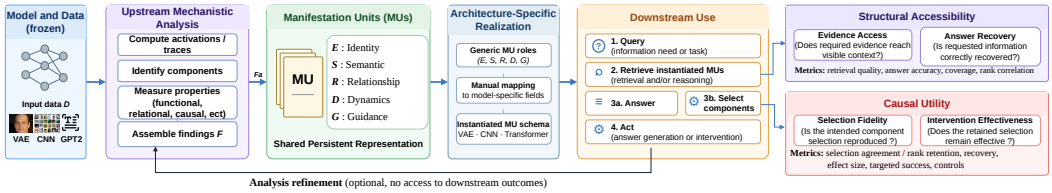

Figure 1: **Manifestation Unit workflow.** Findings F_a are organized into shared E/S/R/D/G roles, mapped to architecture-specific fields, and rendered for downstream access. Structural Accessibility evaluates evidence access and answer recovery; Causal Utility evaluates selection fidelity and intervention effectiveness.

Table 1: **Shared Manifestation Unit roles and their architecture-specific realization.** Each role has a common interpretation across architectures while retaining evidence in the native form of the analyzed model.

Role	Shared meaning	VAE	CNN	GPT-2
E	**Entity:** component identity, address, and ownership.	latent dimension	layer / filter	attention head
S	**Semantics:** labels, concepts, or functional associations supported by the analysis.	attribute association	class response / selectivity	linguistic or functional association
R	**Relationships:** connections to other components or findings, including interaction, dependency, membership, or direction where available.	latent correlation / interaction	filter correlation / family	head relationship
D	**Dynamics:** quantitative measurements of responses, scores, variation, effects, and their measurement context.	variation / decoded effects	activation / selectivity measurements	attention / intervention statistics
G	**Guidance:** supported downstream operations and their relevant conditions or constraints.	latent manipulation	filter intervention	head intervention

For a fixed model $f_{\theta,a}$ of architecture a and construction data D_c, we write

$$F_a = \mathcal{A}_a(f_{\theta,a}, D_c), \qquad M_{a,e} = \mathcal{B}_a(F_a, e), \qquad I_{a,e} = \mathcal{H}_a(M_{a,e}), \tag{1}$$

where $\mathcal{A}_a$ is the upstream mechanistic analysis, F_a is the set of findings produced by that analysis, $\mathcal{B}_a$ organizes findings associated with an addressable entity or finding group e into a Manifestation Unit

$$M_{a,e} = (E, S, R, D, G), \tag{2}$$

and $\mathcal{H}_a$ maps the shared MU representation to an architecture-specific realization $I_{a,e}$ suitable for downstream use.

This decomposition separates three stages: the findings produced by the original analysis, their shared representation as an MU, and their architecture-specific realization. The MU provides the common interface; the realization preserves the model-specific form required by subsequent procedures.

Shared MU roles. The five MU roles define a common vocabulary for organizing mechanistic findings across architectures. We chose these roles to cover the recurring information needed for downstream reuse in our evaluated settings: what a finding concerns (E), what semantic or functional association it records (S), how it relates to other findings (R), what quantitative behavior or effects were measured (D), and what downstream operations the evidence supports (G). Table 1 summarizes both their shared meaning and their realization in the VAE, CNN, and GPT-2 settings.

These roles provide the common representation evaluated in this work.

MU construction ($\mathcal{B}_a$). The construction function $\mathcal{B}_a$ groups the findings in F_a associated with entity or finding group e and assigns their information to the E/S/R/D/G roles. Measurement values,

measurement context, provenance, source references, and other available evidence are retained in the resulting record.

The source findings remain specific to the upstream analysis. In the evaluated settings, these include VAE latent–attribute associations and interactions, CNN filter responses and relationships, and GPT-2 attention head statistics and intervention measurements. The MU provides a common organization for these findings without changing their original measurements.

Architecture-specific realization ($\mathcal{H}_a$). The shared MU representation is subsequently mapped to fields appropriate for architecture a:

$$I_{a,e} = \mathcal{H}_a(M_{a,e}). \tag{3}$$

This step expresses the information represented by the MU in the form required by the corresponding model and downstream procedure. The architecture-specific realizations preserve model-specific distinctions while deriving their content from the shared MU roles.

Full field-level mappings for each architecture are provided in Appendix A.1.

2.2 Architecture-specific views and downstream access

The MU $M_{a,e}$ is the shared persistent representation of a finding. Downstream procedures operate on architecture-specific realizations derived from these MUs rather than directly on the generic role structure.

Let

$$\mathcal{M}_a = \{M_{a,e}\}, \qquad \mathcal{I}_a = \{I_{a,e}\}$$

denote the MU collection and its architecture-specific realizations. A deterministic renderer $\mathcal{R}_a$ converts the realizations into evidence-bearing access views:

$$P_a = \mathcal{R}_a(\mathcal{I}_a), \qquad C = \mathrm{Access}(P_a, q, \pi, B), \tag{4}$$

where P_a is the rendered access corpus, q is the downstream query or task, π is the access policy, B is the available evidence budget, and C is the evidence made visible to the downstream consumer.

This separation distinguishes the information represented by an MU from the way that information is presented to an access procedure. Serialization, passage construction, retrieval, ranking, and context packing belong to the access layer; the MU itself defines the persistent representation from which those views are produced.

Structural Accessibility therefore evaluates the full representation-to-access pipeline rather than only the presence of an E/S/R/D/G label. Controlled experiments reported later examine individual presentation choices where relevant.

2.3 Downstream use

The same access procedure supports two forms of downstream use. Informational tasks terminate in an answer produced from the accessible evidence, whereas interventional tasks use the accessible numerical information to select model components before applying an intervention.

Structural Accessibility. For an informational task, the access procedure retrieves and packs evidence from the MU-derived access views, after which U_{ans} produces an answer. We evaluate two stages of this process. *Evidence access* measures whether the information required by the query reaches the visible context, while *answer recovery* measures whether the requested information is correctly recovered from that evidence. In our experiments, U_{ans} is Qwen3:8B. Retrieval metrics and answer accuracy therefore characterize distinct stages of the same downstream-access pipeline.

Causal Utility. For an interventional task, a deterministic selector U_{sel} reads the accessible numerical evidence and produces a component set K. The selection is fixed before intervention outcomes are measured. We evaluate whether the representation preserves the information required to reproduce the intended component selection (*selection fidelity*) and whether that retained selection remains effective when applied to the model (*intervention effectiveness*).

Algorithm 1 Construction, realization, access, and downstream use of Manifestation Units

Input: Fixed model $f_{\theta,a}$; construction data D_c; analysis $\mathcal{A}_a$
Input: MU construction rule $\mathcal{B}_a$; architecture mapping $\mathcal{H}_a$; renderer $\mathcal{R}_a$
Input: Task q; access policy π; evidence budget B; evaluation data D_e where applicable
Output: MU collection $\mathcal{M}_a$ and downstream result y

$F_a \leftarrow \mathcal{A}_a(f_{\theta,a}, D_c)$ ▷ produce upstream findings
$\mathcal{M}_a \leftarrow \varnothing$
$\mathcal{I}_a \leftarrow \varnothing$
for each addressable entity or finding group e referenced by F_a **do**
  $M_{a,e} \leftarrow \mathcal{B}_a(F_a, e)$ ▷ construct E/S/R/D/G representation
  $I_{a,e} \leftarrow \mathcal{H}_a(M_{a,e})$ ▷ map to architecture-specific fields
  $\mathcal{M}_a \leftarrow \mathcal{M}_a \cup \{M_{a,e}\}$
  $\mathcal{I}_a \leftarrow \mathcal{I}_a \cup \{I_{a,e}\}$
end for
Persist $\mathcal{M}_a$
$P_a \leftarrow \mathcal{R}_a(\mathcal{I}_a)$ ▷ construct downstream access views
$C \leftarrow \mathrm{Access}(P_a, q, \pi, B)$ ▷ retrieve and pack visible evidence
if q is informational **then**
  $y \leftarrow \mathrm{Answer}_{U_{\mathrm{ans}}}(C, q)$
else if q is interventional **then**
  $K \leftarrow \mathrm{Select}_{U_{\mathrm{sel}}}(C, q)$
  Freeze K before intervention evaluation
  $y \leftarrow \mathrm{Intervene}(f_{\theta,a}, K, D_e)$
end if
return $\mathcal{M}_a, y$

Using a deterministic selector separates representation and access fidelity from additional uncertainty introduced by an LLM-based decision procedure. Selection agreement or rank retention therefore tests whether the numerical information required for downstream selection has been preserved, while the subsequent model intervention evaluates the effect of the selected components.

Guidance G records downstream operations supported by the available finding, together with their applicable conditions or constraints. Execution remains part of the architecture-specific downstream procedure. Algorithm 1 summarizes the complete workflow.

3 Evaluation Framework

We evaluate the two properties introduced in Section 2. *Structural Accessibility* asks whether represented findings can be reached and correctly recovered under a controlled access budget. *Causal Utility* asks whether represented numerical findings preserve the information needed for deterministic component selection and whether the resulting selection remains effective under intervention.

We evaluate complementary aspects of this progression across three architectures: VAE latent channels on CelebA (Liu et al., 2015), CNN filters on CIFAR-10 (Krizhevsky, 2009), and GPT-2 attention heads, with held-out functional evaluation on WikiText (Merity et al., 2017). VAE additionally tests repeat stability, CNN connects bounded access to intervention, and GPT-2 tests field grounding and selection retention in a transformer setting.

Access conditions. The **Direct-*** terms are paper-specific access conditions rather than standard interpretability terminology. **Direct-Bounded** exposes the native analysis representation through the same constrained retrieval pipeline used for MUs; in the primary experiment, access is limited to at most ten retrieved passages and 8,000 packed tokens. **Direct-Complete** exposes the complete relevant native source, removing retrieval truncation, while **Direct-Numeric** applies the deterministic selector directly to the recorded numerical measurements. The latter two are used as causal-access references rather than Structural Accessibility conditions.

Structural Accessibility protocol. Each architecture uses 60 source-grounded questions, common BM25 retrieval, the same 8,000-token/10-passage primary budget, and the same Qwen3:8B answerer across nine presentation conditions: MU, Direct-Bounded, Generic, Natural-Prose, Unstructured, ID-Normalized, and three random-partition realizations. All conditions preserve the

Table 2: Structural Accessibility with common BM25 at 8,000 tokens and at most ten passages. Each row averages 60 questions; values are percentages.

Model	Metric	MU	Direct-Bounded	Generic	Natural-Prose	Unstructured	ID-Norm.	Random 0	Random 1	Random 2
VAE	Factual	**54.58**	44.81	44.72	32.22	39.89	41.44	46.72	44.47	45.06
	Evidence	57.22	59.94	54.39	48.61	54.61	49.61	**60.72**	58.17	59.61
CNN	Factual	**79.44**	25.83	21.11	36.67	23.33	16.67	18.06	21.11	16.94
	Evidence	**91.67**	29.44	32.22	50.56	31.11	18.89	31.11	30.27	38.57
GPT-2	Factual	**79.72**	49.17	54.17	53.06	55.83	56.94	52.78	52.22	49.44
	Evidence	**80.56**	49.44	55.00	58.96	49.44	55.00	50.83	49.72	51.11

Factual: macro answer-slot accuracy. Evidence: requested-evidence recall. Bold marks descriptive maxima. The VAE Natural-Prose failure is retained.

same underlying findings but differ in how they are serialized, grouped, or partitioned. Full condition definitions are given in Appendix A.

For each question, *requested-evidence recall* measures whether the required evidence reaches the visible context, while *answer-slot accuracy* measures whether the requested information is correctly recovered. These correspond to Evidence Access and Answer Recovery in Figure 1.

Causal experiments use deterministic selectors followed by model interventions, allowing selection fidelity to be evaluated separately from intervention effectiveness. Detailed selection rules and additional access controls are reported in Appendix A.

4 Structural Accessibility

4.1 Recovery under bounded access

The first experiment asks whether the representation changes what can be recovered when the retriever, access budget, questions, and answerer are held fixed.

Table 2 reports the nine common-BM25 conditions at the primary 8,000-token, ten-passage setting. MU achieves the highest descriptive answer-slot accuracy in all three models: 54.58% for VAE, 79.44% for CNN, and 79.72% for GPT-2, compared with 44.81%, 25.83%, and 49.17% for Direct-Bounded. The strongest non-MU alternatives reach 46.72%, 36.67%, and 56.94%, respectively.

Evidence access and answer recovery do not always move together. MU reaches 91.67% requested-evidence recall for CNN and 80.56% for GPT-2, but 57.22% for VAE, where several alternatives expose slightly more requested evidence. This difference motivates evaluating retrieval and downstream recovery separately.

A chunking-matched control shows that parent-preserving grouping explains part of the accessibility difference, with effects varying across architectures (Appendix B.2). We therefore interpret the main result at the level of the complete representation-and-access pipeline rather than E/S/R/D/G organization alone. Entity-aware MU routing is evaluated separately in Appendix B.

4.2 Accessibility versus stability and grounding

Structural Accessibility concerns access to recorded findings, not whether the underlying findings repeat across data or whether individual represented fields remain functionally useful.

Two complementary diagnostics examine these distinctions. Repeating the VAE analysis with the same checkpoint on 1,000 byte-disjoint CelebA images yields small channel-correlation differences (MAD $= 0.024917$) and agreement for 2,488/2,490 interaction categories, but only 23.63% (121/512) agreement in top-1 semantic associations, showing that accessibility does not imply repeat stability. In GPT-2, with retrieval fixed, removing D, S, or R reduces the corresponding attention, semantic, or relationship Recall@5 from 86.67% to 18.33%, 60.00% to 0.00%, and 65.00%

Table 3: Independent evaluation probes different failure modes across architectures.

A. VAE image-disjoint repeat-sample stability		
Channel-correlation MAD	0.024917	Correlation difference
Interaction-category agreement	2,488/2,490	Pairs
Semantic top-1 agreement	23.63%	121/512 channels
B. GPT-2 held-out task-specific field utility		
Attention: remove D	$86.67\% \rightarrow 18.33\%$	12 tasks
Semantics: remove S	$60.00\% \rightarrow 0.00\%$	8 tasks
Relationships: remove R	$65.00\% \rightarrow 0.00\%$	4 tasks

A: 1,000 repeat images; historical checkpoint identity remains unverified. B: fixed-evidence deletions, Recall@5 on 300 WikiText examples from 55 articles. Task counts are the scorable denominators.

Table 4: Causal Utility under bounded access, complete access, and fresh fixed-set evaluation.

Endpoint	MU	Direct	Random	Contrast / 95% CI
CNN bounded MU access versus Direct-Complete				
Targeted success (%)	92.367	92.536	–	-0.169 pp [-0.391, +0.052]
Necessity disruption (%)	64.713	67.175	–	-2.462 pp [-3.400, -1.524]
CNN separate complete-access study				
Natural patching (pp)	17.641	17.641	0.752	MU = Direct
Patching CUR	1.000	–	–	[1.000, 1.000]
Necessity CUR	1.000	–	–	[1.000, 1.000]
GPT-2 Direct-Numeric selection retention and fresh IOI				
Full ranking retained	144/144 ranks identical			Same ordered top-7
Normalized recovery (%)	61.87	shared	6.30	+55.57 pp
Paired random contrast	95% CI [51.262, 59.740] pp			$p = 3.052 \times 10^{-5}$

CNN bounded rows compare MU at cap 50 with Direct-Complete; CIs are exploratory source-image cluster bootstraps. Complete-access CNN is a separate study, where identical MU/Direct selections yield CUR$= 1$. GPT-2 uses one shared MU/Direct-selected seven-head set evaluated on fresh IOI examples against 100 frozen layer-matched random sets. Intervals are conditional on the frozen experimental settings.

to 0.00%, respectively. These ablations test the usability of represented content independently of retrieval; additional variants appear in Appendix F.

5 Causal Utility

5.1 CNN: bounded access to intervention

The CNN experiment tests whether retrieved class-response findings are sufficient to form the eight-filter `conv3` selection required by a fixed intervention rule. The selector uses only identities and class-response measurements available in the retrieved context.

Under the same frozen BM25 retrieval and reader, Direct-Bounded produces no complete eight-filter selection at passage caps 10–50. This is an access-to-selection failure under the tested retrieval and reader setting, not evidence that the native representation is causally less effective. MU-mediated access produces executable selections whose effect increases with available evidence: targeted success rises from 55.9% at cap 10 to 92.37% at cap 50, approaching Direct-Complete (92.54%). Zero-ablation disruption reaches 64.71% at cap 50 versus 67.18% for Direct-Complete (Figure 2). A complete-depth diagnostic and full selection rules are reported in Appendix D.

5.2 CNN: complete-access selection fidelity

The bounded experiment mixes representation and retrieval effects, so we next remove retrieval as a constraint. Using disjoint construction and evaluation images, MU and Direct-Numeric select identical filter sets for all ten target classes. Thus, the MU representation preserves the numerical information required by the deterministic selector.

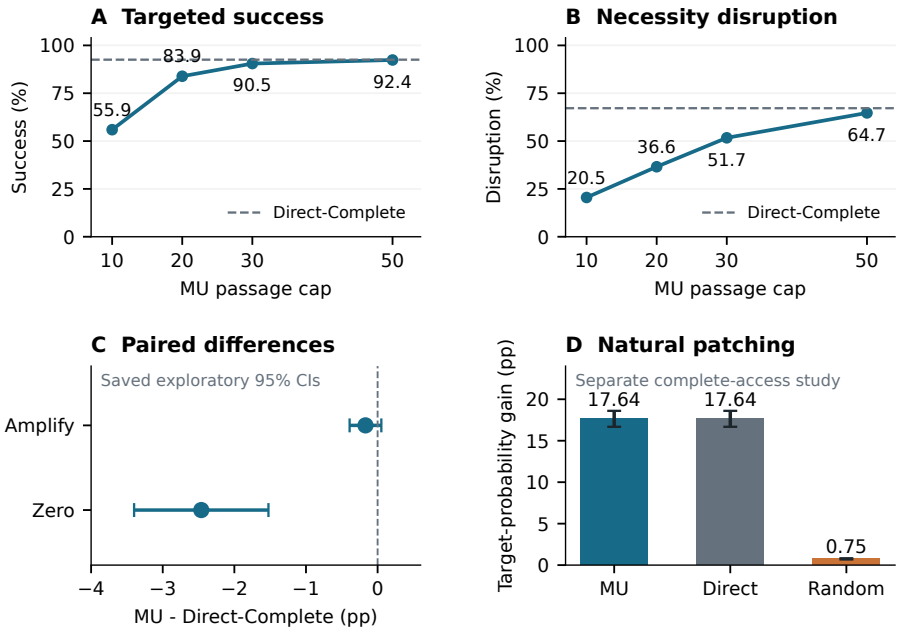

Figure 2: **CNN bounded access and intervention.** (A,B) MU intervention performance as access increases; dashed lines show Direct-Complete. Direct-Bounded does not yield a complete eight-filter selection. (C) MU–Direct-Complete differences at cap 50. (D) Complete-access activation patching with identical MU and Direct selections.

Natural activation patching correspondingly yields the same 17.641-pp increase in target probability for MU and Direct, compared with 0.752 pp for matched random filters. The resulting CUR of 1.0 therefore measures exact selection retention under complete access. Additional checks are reported in Appendix E.

5.3 GPT-2: SELECTION RETENTION AND FRESH CAUSAL EVALUATION

GPT-2 tests whether the same selection-retention property extends to transformer attention heads. A deterministic MU reader reproduces the complete Direct-Numeric ranking of all 144 heads and the same ordered seven-head selection, establishing Selection Fidelity.

On 192 fresh IOI examples, this retained set achieves 61.87% mean normalized recovery, compared with 6.30% across 100 frozen random sets matched to its layer-count profile. The paired difference is 55.57 pp, with a 95% name-block bootstrap interval of $[51.262, 59.740]$ pp (Figure 3). This provides the corresponding Intervention Effectiveness result on fresh examples. Full inference and normalization details are reported in Appendix G.

6 RELATED WORK

Mechanistic interpretability produces findings at multiple granularities, including feature explanations (Paulo et al., 2025), attribution graphs (Ameisen et al., 2025), sparse feature circuits (Marks et al., 2025), and agent-generated hypotheses (Marin-Llobet & Ferrando, 2026); sparse representations may also admit non-unique decompositions (Leask et al., 2025). Related benchmarks such as FIND and MIB evaluate the interpretation and localization of mechanisms (Schwettmann et al., 2023; Mueller et al., 2025), while MAIA automates parts of the analysis workflow (Shaham et al., 2024). These approaches primarily address how mechanistic findings are produced, localized, or evaluated.

A complementary line of work concerns persistence, access, and model interaction. Neuronpedia stores interpretability artifacts (Lin, 2023), while NNsight/NDIF, nnterp, and Pyvene provide inter-

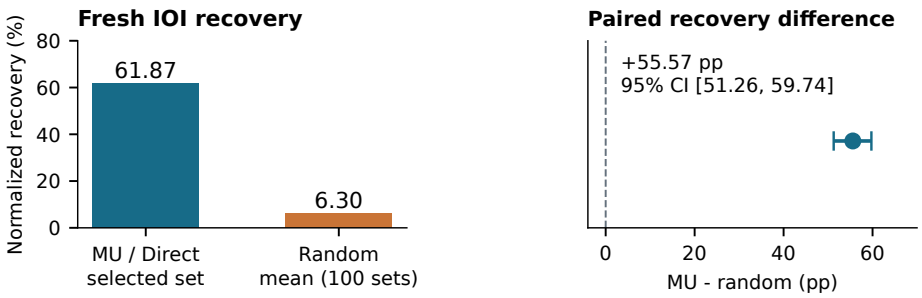

Figure 3: **GPT-2 selection retention and causal evaluation.** The shared seven-head selection achieves 61.87% normalized recovery on 192 fresh IOI examples versus 6.30% for 100 layer-matched random sets. The interval reports the paired selected-minus-random contrast.

faces for model inspection and intervention (Fiotto-Kaufman et al., 2025; Dumas, 2025; Wu et al., 2024). More broadly, GraphRAG illustrates how the organization of stored information can affect downstream retrieval (Edge et al., 2024).

MUs address the intermediate representation layer between mechanistic discovery and downstream use. They map heterogeneous findings into a shared representation across architectures while retaining architecture-specific evidence. This representation provides a common interface for downstream access and can support intervention by exposing the evidence required to form architecture-specific component selections that are subsequently executed on the underlying model.

7 Discussion and Conclusion

Across VAE, CNN, and GPT-2, our results show that producing a mechanistic finding and making it usable later are distinct problems. Under matched access conditions, the same underlying findings lead to substantially different downstream recovery depending on how they are represented and exposed. The difference between requested-evidence recall and answer-slot accuracy further shows that retrieving relevant evidence is not sufficient: its organization also affects whether a downstream consumer can recover the requested information. A limitation of the present experiments is that they do not isolate the contribution of the E/S/R/D/G decomposition from other parts of the MU pipeline, including serialization, parent-preserving grouping, and chunking. The chunking control suggests that grouping explains part, but not all, of the observed accessibility effect; accordingly, these results support the complete MU representation-and-access pipeline rather than E/S/R/D/G in isolation.

The diagnostic experiments clarify the scope of this result. VAE repeat analysis shows that an accessible recorded finding need not remain stable when the upstream analysis is repeated on different data, while GPT-2 field deletions show that specific represented content remains functionally relevant when retrieval is held fixed. Persistent representation therefore concerns preserving and exposing a finding, not certifying the finding itself.

The causal experiments address whether the represented information remains actionable. Under complete access, MU preserves the numerical distinctions needed to reproduce the native CNN filter selections and GPT-2 head ranking and selection. The corresponding interventions remain effective, while the bounded CNN experiment shows why accessibility matters operationally: as more MU evidence becomes available, executable selections approach the complete-access intervention performance. Together, these results separate *Selection Fidelity*, preservation of the information required for selection, from *Intervention Effectiveness*, the causal effect of the retained selection.

Overall, the evidence supports MUs as a persistent interface between mechanistic discovery and downstream reuse across heterogeneous architectures. It does not establish that E/S/R/D/G is a unique or minimal decomposition, nor do the present experiments isolate its effect from the surrounding representation-and-access design. Scaling to larger feature spaces, composition of findings, and more autonomous downstream consumers remains future work.

REPRODUCIBILITY STATEMENT

Appendix A documents the experimental configurations, representation and access procedures, selection rules, evaluation metrics, and architecture-specific implementations used throughout the study. Appendix H describes the archived artifacts supporting the reported results, including question-level scores, retrieval rankings, packed evidence contexts, representation files, experiment manifests, frozen component selections, per-example intervention outcomes, statistical outputs, and available source hashes.

Structural Accessibility uses fixed question sets, retrieval configurations, evidence budgets, and decoding settings. The CNN and GPT-2 Causal Utility experiments use deterministic selection procedures, with selected components frozen before intervention outcomes are measured. The GPT-2 layer-matched random-control evaluation uses fixed evaluation examples, selected and control sets, intervention operators, and outcome definitions.

A browser-based companion demo is provided for direct inspection of selected evaluation questions, representation conditions, accessed evidence, and saved experimental outputs: `https://manifestation-xai.github.io/manifestation/cnn/`. The demo replays archived results rather than executing new inference; the appendix and archived experimental artifacts remain the basis for the reported claims.

AI USE STATEMENT

Qwen3:8B was used as an experimental component of the study in the Structural Accessibility evaluation, where it served as the common downstream answerer across the compared bounded-access conditions. Its role was restricted to answering from the evidence provided by the evaluated retrieval and presentation pipelines; it was not used to determine the deterministic component selections used in the CNN or GPT-2 Causal Utility experiments.

OpenAI coding assistants were used during the research and manuscript preparation process to support tasks including research ideation, code and evidence-file inspection, debugging, arithmetic and cross-table consistency checks, organization of experimental outputs, and preparation of LaTeX tables, figures, and manuscript text. They were also used to assist with drafting, rewriting, and language polishing. Experimental claims, reported numerical results, methodological choices, and final manuscript content were reviewed against the underlying experimental artifacts by the authors.

REFERENCES

Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L. Turner, Brian Chen, Craig Citro, et al. Circuit tracing: Revealing computational graphs in language models. Transformer Circuits, 2025.

Clément Dumas. nnterp: A standardized interface for mechanistic interpretability of transformers. In *NeurIPS Workshop on Mechanistic Interpretability*, 2025.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph RAG approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.

Jaden Fiotto-Kaufman, Alexander Loftus, Eric Todd, Jannik Brinkmann, Koyena Pal, Dmitrii Troitskii, Michael Ripa, Adam Belfki, Can Rager, Caden Juang, Aaron Mueller, Samuel Marks, Arnab Sen Sharma, Francesca Lucchetti, Nikhil Prakash, Carla Brodley, Arjun Guha, Jonathan Bell, Byron C. Wallace, and David Bau. NNsight and NDIF: Democratizing access to open-weight foundation model internals. In *International Conference on Learning Representations*, 2025.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Michael Lan, Narmeen Fatimah Oozeer, Chaithanya Bandi, Philip Quirke, Austin Meek, Fazl Barez, and Amir Abdullah. Make mechanistic interpretability auditable: A call to develop guidelines via continuous collaborative reviewing. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3516–3540, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.159. URL https://aclanthology.org/2026.acl-long.159/.

Patrick Leask, Bart Bussmann, Michael Pearce, Joseph Bloom, Curt Tigges, Noura Al Moubayed, Lee Sharkey, and Neel Nanda. Sparse autoencoders do not find canonical units of analysis. In *International Conference on Learning Representations*, 2025.

Johnny Lin. Neuronpedia: Interactive reference and tooling for analyzing neural networks, 2023. URL https://www.neuronpedia.org.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

Aleksandar Makelov, Georg Lange, and Neel Nanda. Towards principled evaluations of sparse autoencoders for interpretability and control. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/53356aebeea8ffd40a8ac3bb66243162-Abstract-Conference.html.

Arnau Marin-Llobet and Javier Ferrando. Automated interpretability and feature discovery in language models with agents. *arXiv preprint arXiv:2605.01555*, 2026.

Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *International Conference on Learning Representations*, 2025.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017.

Aaron Mueller, Atticus Geiger, Sarah Wiegreffe, Dana Arad, Iván Arcuschin, Adam Belfki, Yik Siu Chan, Jaden Fried Fiotto-Kaufman, Tal Haklay, Michael Hanna, Jing Huang, Rohan Gupta, Yaniv Nikankin, Hadas Orgad, Nikhil Prakash, Anja Reusch, Aruna Sankaranarayanan, Shun Shao, Alessandro Stolfo, Martin Tutek, Amir Zur, David Bau, and Yonatan Belinkov. MIB: A mechanistic interpretability benchmark. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 45069–45108. PMLR, 2025. URL https://proceedings.mlr.press/v267/mueller25a.html.

Aaron Mueller, Jannik Brinkmann, Millicent Li, Samuel Marks, Koyena Pal, Nikhil Prakash, Can Rager, Aruna Sankaranarayanan, Arnab Sen Sharma, Jiuding Sun, Eric Todd, David Bau, and Yonatan Belinkov. The quest for the right mediator: Surveying mechanistic interpretability for NLP through the lens of causal mediation analysis. *Computational Linguistics*, 52(1):331–378, March 2026. doi: 10.1162/coli.a.572. URL https://aclanthology.org/2026.cl-1.10/.

Gonçalo Santos Paulo, Alex Troy Mallen, Caden Juang, and Nora Belrose. Automatically interpreting millions of features in large language models. In *International Conference on Machine Learning*, 2025.

Sarah Schwettmann, Tamar Rott Shaham, Joanna Materzynska, Neil Chowdhury, Shuang Li, Jacob Andreas, David Bau, and Antonio Torralba. FIND: A function description benchmark for evaluating interpretability methods. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Tamar Rott Shaham, Sarah Schwettmann, Franklin Wang, Achyuta Rajaram, Evan Hernandez, Jacob Andreas, and Antonio Torralba. A multimodal automated interpretability agent. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 44293–44321. PMLR, 2024.

Zhengxuan Wu, Atticus Geiger, Aryaman Arora, Jing Huang, Zheng Wang, Noah D. Goodman, Christopher D. Manning, and Christopher Potts. pyvene: A library for understanding and improving PyTorch models via interventions. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pp. 158–165, 2024.

APPENDIX OVERVIEW

This appendix provides implementation details, full evaluation grids, diagnostic analyses, and reproducibility information supporting the main paper. It follows the two properties used in the main paper: *Structural Accessibility* and *Causal Utility*, with Causal Utility separated into *Selection Fidelity* and *Intervention Effectiveness*. The main paper reports the primary comparisons; the appendix provides the additional protocol and evidence needed to audit them.

	Section	Purpose
A	MU construction and evaluation protocol	Defines MU construction, architecture-specific realization, access conditions, presentation pipelines, scoring, and question auditing.
B	Full Structural Accessibility results	Reports complete access grids, retrieval diagnostics, chunking controls, packing audits, and entity-aware access.
C	VAE repeat stability	Tests whether findings reproduce when the same checkpoint and analysis are applied to byte-disjoint images.
D	CNN bounded-access Causal Utility	Defines deterministic selection and reports access-dependent executability and intervention effects.
E	CNN complete-access Selection Fidelity	Tests whether MU preserves the numerical information required to reproduce native component selections and their intervention effects.
F	GPT-2 field grounding	Tests the functional contribution of represented field content under fixed evidence and reports retrieval-sensitive variants.
G	GPT-2 Selection Fidelity and fresh causal evaluation	Reports head-ranking retention, fresh IOI evaluation, matched random controls, and the literature-reference comparison.
H	Statistics and reproducibility	Documents uncertainty procedures, runtime configuration, evidence manifests, and remaining audit gaps.

Reading guide. Appendices A–B specify and audit Structural Accessibility; Appendix C reports the independent VAE repeat-stability diagnostic; Appendices D–G support Selection Fidelity and Intervention Effectiveness; Appendix H documents statistics and reproducibility.

A MU CONSTRUCTION AND EVALUATION PROTOCOL

A.1 MU CONSTRUCTION AND ARCHITECTURE-SPECIFIC REALIZATION

For a model $f_{\theta,a}$ of architecture a and construction data D_c, the main paper defines

$$F_a = \mathcal{A}_a(f_{\theta,a}, D_c), \qquad M_{a,e} = \mathcal{B}_a(F_a, e), \qquad I_{a,e} = \mathcal{H}_a(M_{a,e}), \tag{5}$$

where F_a contains the findings produced by the upstream analysis, $M_{a,e} = (E, S, R, D, G)$ is the shared Manifestation Unit for entity or finding group e, and $I_{a,e}$ is its architecture-specific realization.

The five roles retain the same meaning across architectures, while their realized fields remain model-specific. Table 5 summarizes the mappings used in the evaluated systems.

VAE. Entities are latent channels. The architecture-specific realization includes `summary`, `generation_role`, `redundancy_notes`, and `manipulation_guide`, supported by recorded association, relationship, interaction, and conditional-effect measurements. Unsigned associations are not converted into signed intervention claims.

CNN. Entities are layer/filter pairs, with additional relation and family records. Realized fields include `summary`, `technical_properties`, `class_responses`, `relationships`, and `manipulation_guide`. Guidance records supported operators and their preconditions.

GPT-2. Entities are attention heads. The mapping retains head identity, semantic classifications, relationships, measured statistics, intervention results, and source pointers. Realized fields include `summary`, `attention_behavior`, `relationships`, `mechanistic_profile`, and `intervention_findings`.

Table 5: Shared MU roles and their principal architecture-specific realization fields. A realized field may carry information associated with more than one role.

Role	Shared information	VAE realization	CNN realization	GPT-2 realization
E	Entity identity, address, and ownership	channel address; summary	layer/filter address; summary	head address; summary
S	Semantic association or functional description	generation role; summary	class responses; summary	summary; mechanistic profile
R	Relations to another entity or finding group	redundancy notes; relation sections	relationships	relationships
D	Measured behavior, response, score, or effect	attribute, pair, and conditional measurements	technical properties; class responses	attention behavior; mechanistic profile
G	Supported downstream operation and conditions	manipulation guide	manipulation guide	intervention findings

A.1.1 Shared Intermediate Roles and Final MU Fields

The shared representation $Z_{a,e} = (E, S, R, D, G)$ is a semantic intermediate rather than a required final serialization. The realization function $\mathcal{H}_a$ maps this role-labeled content into persistent architecture-specific fields. The examples below show this mapping separately for the three evaluated architectures.

For readability, the `from_roles` and `contains` entries in the right-hand panels are explanatory annotations: they expose how the final fields realize the shared semantic contract, rather than introducing additional stored fields into the evaluated schemas.

A.2 Structural Accessibility benchmark and scoring

The benchmark contains 60 source-grounded questions per architecture. The access grid crosses evidence-token budgets of 4,000 and 8,000 with passage caps of five and ten, yielding four settings per presentation. The primary main-paper setting is 8,000 evidence tokens and at most ten passages.

Nine common-BM25 conditions are evaluated: MU, Direct-Bounded, Generic, Natural-Prose, Unstructured, ID-Normalized, and three seeded random-boundary presentations. All conditions preserve the same saved findings but differ in serialization, grouping, or passage construction.

Gold answer specifications and source evidence are used only for scoring and are not exposed to BM25, the Qwen3:8B answerer, entity routing, or causal selection. *Answer-slot accuracy* is the mean fraction of requested answer items recovered correctly per question. *Requested-evidence recall* is the fraction of specified source evidence present in the packed context, averaged over questions. Exact-question accuracy, micro-F1, and ranked-retrieval metrics are supplementary.

Answerer configuration. All Structural Accessibility generations use Qwen3:8B with temperature 0, fixed seed 20260907, thinking disabled, a 32,768-token context limit, and a 4,096-token maximum generation length. Each distinct experimental setting is evaluated once; saved calls are reused when a run resumes. We therefore report the benchmark as a fixed deterministic-decoding evaluation rather than estimating generation-level Monte Carlo variance.

A.3 Direct access regimes

The Direct-* labels are paper-specific access conditions. **Direct-Bounded** exposes the native analysis representation through constrained retrieval. In the primary Structural Accessibility setting, the visible evidence is limited to at most ten retrieved passages and 8,000 packed tokens. **Direct-Complete** exposes the complete relevant native source without retrieval truncation. **Direct-Numeric** applies the frozen deterministic selector directly to the recorded numerical measurements. Direct-Complete and Direct-Numeric are causal-access references rather than additional conditions in the nine-way Structural Accessibility comparison.

VAE latent-channel realization

Shared intermediate $Z_{a,e}$

```
{
  "E": {
    "role": "Entity",
    "content": "latent-channel
identity and ownership"
  },
  "S": {
    "role": "Semantics",
    "content": "attribute
associations and functional
description"
  },
  "R": {
    "role": "Relationships",
    "content": "dimension relations,
family, redundancy"
  },
  "D": {
    "role": "Dynamics",
    "content": "attribute, pair, and
conditional measurements"
  },
  "G": {
    "role": "Guidance",
    "content": "supported
manipulations, conditions, limits"
  }
}
```

$\xRightarrow{\mathcal{H}_{\text{VAE}}}$

Persistent VAE fields $M_{a,e}$

```
{
  "summary": {
    "from_roles": ["E", "S", "R",
"D"],
    "contains": [
      "channel identity",
      "attribute associations",
      "family information",
      "measured variation"
    ]
  },
  "generation_role": {
    "from_roles": ["S"],
    "contains": ["observed attribute
associations"]
  },
  "redundancy_notes": {
    "from_roles": ["R"],
    "contains": [
      "thresholded correlation
descriptions",
      "related or redundant
dimensions"
    ]
  },
  "manipulation_guide": {
    "from_roles": ["G"],
    "contains": [
      "recorded manipulation
evidence",
      "target-edit direction status",
      "conditions and limitations"
    ]
  }
}
```

Figure 4: VAE realization of the shared semantic intermediate. The four persistent parent fields are `summary`, `generation_role`, `redundancy_notes`, and `manipulation_guide`. Attribute, dimension-relation, family, pair-interaction, and conditional-effect sections provide supporting evidence to these records.

CNN layer/filter realization

Shared intermediate $Z_{a,e}$

```
{
  "E": {
    "role": "Entity",
    "content": "layer/filter identity
and ownership"
  },
  "S": {
    "role": "Semantics",
    "content": "class association and
filter selectivity"
  },
  "R": {
    "role": "Relationships",
    "content": "filter relations and
family structure"
  },
  "D": {
    "role": "Dynamics",
    "content": "response and
activation measurements"
  },
  "G": {
    "role": "Guidance",
    "content": "supported operators
and preconditions"
  }
}
```

$\xRightarrow{\mathcal{H}_{\text{CNN}}}$

Persistent CNN fields $M_{a,e}$

```
{
  "summary": {
    "from_roles": ["E", "S", "D"],
    "contains": [
      "layer/filter identity",
      "class associations",
      "selected measurements"
    ]
  },
  "technical_properties": {
    "from_roles": ["D"],
    "contains": ["numerical
activation dynamics"]
  },
  "class_responses": {
    "from_roles": ["S", "D"],
    "contains": [
      "per-class response
measurements",
      "semantic class associations"
    ]
  },
  "relationships": {
    "from_roles": ["R"],
    "contains": [
      "relationship membership",
      "direction where applicable"
    ]
  },
  "manipulation_guide": {
    "from_roles": ["G"],
    "contains": [
      "supported intervention
operators",
      "operator preconditions and
limitations"
    ]
  }
}
```

Figure 5: CNN realization of the shared semantic intermediate. `class_responses` realizes both semantic and measured-response information, illustrating that the mapping from E/S/R/D/G to persistent fields is not one-to-one.

GPT-2 attention-head realization

Shared intermediate $Z_{a,e}$

```
{
  "E": {
    "role": "Entity",
    "content": "attention-head
identity and ownership"
  },
  "S": {
    "role": "Semantics",
    "content": "semantic
classification and functional
description"
  },
  "R": {
    "role": "Relationships",
    "content": "head relations and
joint interactions"
  },
  "D": {
    "role": "Dynamics",
    "content": "attention and
mechanistic measurements"
  },
  "G": {
    "role": "Guidance",
    "content": "intervention
evidence, conditions, status"
  }
}
```

$\mathcal{H}_{\mathrm{GPT2}} \Longrightarrow$

Persistent GPT-2 fields $M_{a,e}$

```
{
  "summary": {
    "from_roles": ["E", "S"],
    "contains": [
      "head identity",
      "semantic or functional summary"
    ]
  },
  "attention_behavior": {
    "from_roles": ["D"],
    "contains": ["attention
statistics"]
  },
  "relationships": {
    "from_roles": ["R"],
    "contains": [
      "correlation records",
      "measured joint interventions",
      "relation type preserved"
    ]
  },
  "mechanistic_profile": {
    "from_roles": ["S", "D"],
    "contains": [
      "QK/OV measurements",
      "collective measurements"
    ]
  },
  "intervention_findings": {
    "from_roles": ["G"],
    "contains": [
      "calibration and guidance",
      "operator-specific intervention
evidence"
    ]
  }
}
```

Figure 6: GPT-2 realization of the shared semantic intermediate. Identity is routed into `summary`, attention statistics into `attention_behavior`, relations into `relationships`, QK/OV and collective measurements into `mechanistic_profile`, and calibration/guidance into `intervention_findings`.

A.4 PRESENTATION CONDITIONS AND PASSAGE CONSTRUCTION

All non-MU alternatives are constructed from the original analysis artifacts without MU-derived joins. Their differences are deterministic and concern how the same findings are serialized, grouped, and divided into retrievable passages.

Table 6: Common-BM25 presentation conditions. Random conditions alter passage boundaries, not findings or their order.

Condition	Serialization / visible structure	Passage construction
MU	MU-derived architecture-specific views preserving source values and entity–field ownership.	Parent-preserving passages; no passage crosses an MU parent.
Direct-Bounded	Pretty-printed native JSON preserving keys, values, nesting, lists, and order.	1,000-token windows with 100-token overlap within each source file.
Generic	Compact native JSON preserving hierarchy, keys, punctuation, values, and order.	1,000-token windows with 100-token overlap.
Natural-Prose	Each native leaf is deterministically rendered as a source-path/value sentence; no LLM paraphrasing or cross-file composition.	1,000-token windows with 100-token overlap.
Unstructured	Whitespace-flattened native JSON-shaped text; keys, hierarchy delimiters, punctuation, and values remain visible.	1,000-token windows with 100-token overlap.
ID-Normalized	Compact native JSON with digit-only dictionary keys zero-padded to at least three characters; values and list order are unchanged.	1,000-token windows with 100-token overlap.
Random-0/1/2	Same compact serialization as Generic; findings and ordering are unchanged.	Seeded window widths sampled from 800–1,000 tokens with 100-token overlap; seeds 0/1/2 restart per source file.

Native windows never cross source-file boundaries. Sidecar provenance and ownership annotations are retained for auditing and scoring but are not inserted into common-BM25 retrieval text or Qwen prompts. Direct-Complete uses the same relevant native sources without bounded retrieval truncation.

For the VAE and GPT-2 MU corpora, short source sections are retained intact, longer sections are split at whitespace, and adjacent pieces are regrouped only within the same parent up to the passage limit. The VAE realization contains 2,048 parent documents and yields 10,946 retrieval passages. CNN MU passages similarly preserve field-parent ownership and complete mapped values.

Supplementary chunking-matched control. Generic-EntityChunked is not one of the nine primary conditions. It retains Generic/native serialization and source values but groups evidence by native entity rather than fixed-width source windows. It introduces no E/S/R/D/G labels, MU field names, summaries, or LLM rewriting. Retrieval, questions, answerer, budgets, and scoring remain unchanged. Full results are reported in Appendix B.2.

A.5 QUESTION-WORDING AUDIT

The frozen public questions do not contain literal MU-only output-field names. For the VAE, none of the 60 questions contains `summary`, `generation_role`, `redundancy_notes`, or `manipulation_guide`. For the CNN, none contains `summary`, `technical_properties`, `class_responses`, `relationships`, or `manipulation_guide`. For GPT-2, none contains `summary`, `attention_behavior`, `relationships`, `mechanistic_profile`, or `intervention_findings`.

Questions may name scientific measurements or requested output slots shared across conditions, such as correlation, selectivity, family identifier, or sample variance. Lexical and organizational effects therefore remain part of the evaluated presentation-and-access pipeline.

A.6 ENTITY-AWARE RETRIEVAL POLICY

Entity-aware retrieval is an MU-specific access policy and is evaluated separately from the matched common-BM25 benchmark. Architecture-specific rules parse explicit entity identifiers and requested information from the public question. Records are prioritized by matching entity, field, and relation information, with BM25 or semantic similarity used within priority tiers. If no entity or field is recognized, retrieval falls back to the ordinary ranking.

Records are packed under the same passage and token limits. Routing uses only the public question and MU metadata; gold answers and evaluation provenance are not exposed. Results are reported in Appendix B.

B FULL STRUCTURAL ACCESSIBILITY RESULTS

This appendix reports the complete budget/cap grid and supplementary diagnostics behind the primary 8k/10 comparison in the main paper. In the tables below, extbfAns. denotes answer-slot accuracy and extbfER denotes requested-evidence recall.

The four access settings are 4k/5, 4k/10, 8k/5, and 8k/10 (evidence tokens / passage cap). The following tables retain the complete common-BM25 answer-level grid. Ranked-retrieval metrics are computed before token-budget packing, so at a fixed passage cutoff k they are invariant to the downstream 4k/8k evidence-token ceiling; token limits affect the subsequently packed context.

B.1 VAE COMMON-BM25 GRID

Table 7: VAE: full matched grid (60 questions per row).

Condition	Budget/k	Ans.	ER	Exact	F1	Fail
Direct-Bounded	4k/5	46.03	52.58	10.00	57.78	0
Generic	4k/5	43.17	45.50	8.33	53.63	0
ID-Normalized	4k/5	40.28	42.17	8.33	50.27	0
MU	4k/5	52.92	57.36	38.33	74.46	0
Natural-Prose	4k/5	17.22	26.11	1.67	21.17	0
Random-0	4k/5	45.08	51.83	10.00	57.14	0
Random-1	4k/5	46.53	52.72	11.67	57.78	0
Random-2	4k/5	42.97	51.28	8.33	53.22	0
Unstructured	4k/5	37.69	43.36	6.67	46.28	0
Direct-Bounded	4k/10	46.03	52.58	10.00	57.78	0
Generic	4k/10	43.17	45.50	8.33	53.63	0
ID-Normalized	4k/10	40.78	42.17	8.33	50.82	0
MU	4k/10	52.08	57.36	35.00	73.01	0
Natural-Prose	4k/10	16.11	26.11	1.67	20.22	0
Random-0	4k/10	45.08	51.83	10.00	57.14	0
Random-1	4k/10	46.53	52.72	11.67	57.78	0
Random-2	4k/10	42.97	51.28	8.33	53.22	0
Unstructured	4k/10	37.69	43.36	6.67	46.28	0
Direct-Bounded	8k/5	43.14	59.94	11.67	53.70	0
Generic	8k/5	44.67	53.28	13.33	56.67	0
ID-Normalized	8k/5	41.94	47.06	11.67	53.59	0
MU	8k/5	53.47	56.25	36.67	76.54	0
Natural-Prose	8k/5	32.78	42.36	8.33	30.68	0
Random-0	8k/5	42.75	56.28	10.00	54.19	0
Random-1	8k/5	44.86	54.39	11.67	55.96	0
Random-2	8k/5	44.08	55.17	10.00	54.85	0
Unstructured	8k/5	42.06	51.83	5.00	51.52	0
Direct-Bounded	8k/10	44.81	59.94	11.67	55.49	0
Generic	8k/10	44.72	54.39	13.33	56.67	0
ID-Normalized	8k/10	41.44	49.61	13.33	51.35	0
MU	8k/10	54.58	57.22	41.67	77.30	0
Natural-Prose	8k/10	32.22	48.61	6.67	35.14	1
Random-0	8k/10	46.72	60.72	13.33	59.05	0

Condition	Budget/k	Ans.	ER	Exact	F1	Fail
Random-1	8k/10	44.47	58.17	13.33	54.95	0
Random-2	8k/10	45.06	59.61	11.67	56.28	0
Unstructured	8k/10	39.89	54.61	3.33	47.03	0

The final VAE cohort also stores the full ordered rankings, representation files, packed contexts, and P@k, Recall@k, nDCG@k, and MRR metrics. These metrics were omitted from an earlier printed summary but are retained in the companion export. Passage-level ranking metrics are interpreted separately from requested-evidence recall because the latter is computed from the evidence actually shown to Qwen.

The single Natural-Prose technical failure at 8k/10 is trial `Natural-Prose_bm25_8000_k10_C058`. Generation reached the 4,096-token output limit (`done_reason=length`); the failure is output truncation, not an input-context overflow or timeout. The associated format-defect indicator refers to the same failed trial.

B.2 Supplementary chunking-matched accessibility controls

The primary Structural Accessibility benchmark compares complete presentation-and-access pipelines and therefore changes both representation and passage construction. As a supplementary diagnostic, we construct **Generic-EntityChunked** to test one organizational factor: parent-preserving grouping. This condition retains Generic/native serialization and source values, introduces no E/S/R/D/G labels, MU field names, summaries, or LLM rewriting, and changes only the grouping of evidence by native entity. Retrieval, questions, answerer, budgets, and scoring remain unchanged.

This control is not part of the primary nine-condition comparison and is not a causal decomposition of the MU interface. It tests whether parent-preserving passage construction can account for part of the observed accessibility difference.

Table 8: Chunking-matched answer-slot accuracy at the primary 8k/10 setting. Supplementary Holm correction is performed separately within each architecture.

Model	Generic	EntityChunked	MU	MU–EC	95% CI	Holm p
VAE	44.72	49.86	54.58	+4.72	$[-3.47, 13.06]$	0.325
CNN	21.11	33.61	79.44	+45.83	$[35.56, 56.39]$	2.0×10^{-5}
GPT-2	54.17	62.22	79.72	+17.50	$[2.50, 32.50]$	0.055

Table 9: Chunking-matched requested-evidence recall at the primary 8k/10 setting.

Model	Generic	EntityChunked	MU	MU–EC	95% CI	Holm p
VAE	54.39	63.06	57.22	-5.83	$[-9.86, -2.22]$	0.00388
CNN	32.22	56.11	91.67	+35.56	$[25.56, 45.56]$	2.0×10^{-5}
GPT-2	55.00	63.06	80.56	+17.50	$[3.06, 31.39]$	0.03872

The effect of parent-preserving grouping varies across architectures. For the VAE, the residual MU answer-slot-accuracy advantage over Generic-EntityChunked is not statistically established, and Generic-EntityChunked attains higher requested-evidence recall. In the CNN, substantial residual MU advantages remain for both endpoints. GPT-2 is intermediate: MU retains a 17.50-pp advantage on both endpoints, with the evidence-recall contrast surviving supplementary Holm correction and the answer-slot-accuracy contrast borderline after correction.

These controls show that parent-preserving grouping contributes materially to accessibility but does not explain the observed MU advantage uniformly across architectures. They do not isolate a causal contribution for E/S/R/D/G or any other individual component of the MU pipeline.

B.3 CNN common-BM25 grid

Table 10: CNN: full matched grid (60 questions per row).

Condition	Budget/k	Ans.	ER	Exact	F1	Fail
Direct-Bounded	4k/5	18.89	13.89	10.00	27.88	0
Generic	4k/5	21.11	16.67	11.67	28.44	0
ID-Normalized	4k/5	15.00	8.33	6.67	18.87	0
MU	4k/5	77.22	85.56	70.00	86.07	0
Natural-Prose	4k/5	36.94	36.11	33.33	41.00	0
Random-0	4k/5	18.89	15.56	8.33	28.70	0
Random-1	4k/5	17.22	15.56	8.33	22.91	0
Random-2	4k/5	20.56	16.11	11.67	29.49	0
Unstructured	4k/5	20.56	17.78	11.67	28.04	0
Direct-Bounded	4k/10	18.89	13.89	10.00	27.88	0
Generic	4k/10	21.11	16.67	11.67	28.44	0
ID-Normalized	4k/10	15.00	8.33	6.67	18.87	0
MU	4k/10	71.11	87.22	58.33	81.67	0
Natural-Prose	4k/10	36.94	36.11	33.33	41.00	0
Random-0	4k/10	18.89	15.56	8.33	28.84	0
Random-1	4k/10	17.22	15.56	8.33	22.91	0
Random-2	4k/10	20.00	16.11	10.00	27.80	0
Unstructured	4k/10	20.56	17.78	11.67	28.04	0
Direct-Bounded	8k/5	18.61	20.56	10.00	26.38	0
Generic	8k/5	21.67	23.89	11.67	26.56	0
ID-Normalized	8k/5	14.44	15.56	6.67	15.83	0
MU	8k/5	78.33	86.11	70.00	86.64	0
Natural-Prose	8k/5	46.11	43.89	40.00	51.69	0
Random-0	8k/5	20.56	18.33	10.00	30.09	0
Random-1	8k/5	17.22	19.44	8.33	24.14	0
Random-2	8k/5	16.67	18.33	8.33	24.00	0
Unstructured	8k/5	18.89	23.33	8.33	27.19	0
Direct-Bounded	8k/10	25.83	29.44	13.33	35.15	0
Generic	8k/10	21.11	32.22	10.00	30.17	0
ID-Normalized	8k/10	16.67	18.89	6.67	19.66	0
MU	8k/10	79.44	91.67	70.00	85.60	0
Natural-Prose	8k/10	36.67	50.56	31.67	44.17	0
Random-0	8k/10	18.06	31.11	6.67	25.57	0
Random-1	8k/10	21.11	30.27	10.00	28.21	0
Random-2	8k/10	16.94	38.57	6.67	24.56	0
Unstructured	8k/10	23.33	31.11	10.00	32.89	0

B.3.1 CNN RANKED-RETRIEVAL DIAGNOSTICS

Because ranking precedes evidence-token packing, only the passage cutoff changes the ranking metrics. We therefore report the two unique cutoffs rather than duplicating identical 4k/8k rows.

B.4 GPT-2 COMMON-BM25 GRID

Table 12: GPT-2: full matched grid (60 questions per row).

Condition	Budget/k	Ans.	ER	Exact	F1	Fail
Direct-Bounded	4k/5	43.33	41.11	15.00	49.21	0
Generic	4k/5	45.00	44.44	21.67	51.20	0
ID-Normalized	4k/5	54.72	46.11	28.33	59.61	0
MU	4k/5	77.22	80.00	63.33	82.49	0
Natural-Prose	4k/5	33.33	33.25	8.33	41.49	0
Random-0	4k/5	48.06	43.06	20.00	51.34	0
Random-1	4k/5	48.61	41.67	16.67	52.51	0
Random-2	4k/5	51.11	47.78	20.00	56.25	0
Unstructured	4k/5	44.44	41.11	26.67	49.40	0
Direct-Bounded	4k/10	40.83	41.11	15.00	47.97	0
Generic	4k/10	45.00	44.44	21.67	51.20	0

Condition	Budget/k	Ans.	ER	Exact	F1	Fail
ID-Normalized	4k/10	54.72	46.11	28.33	59.61	0
MU	4k/10	87.22	88.89	80.00	88.39	0
Natural-Prose	4k/10	32.50	33.25	6.67	41.18	0
Random-0	4k/10	47.50	43.06	20.00	50.57	0
Random-1	4k/10	48.61	41.67	16.67	52.51	0
Random-2	4k/10	51.11	47.78	20.00	56.25	0
Unstructured	4k/10	44.44	41.11	26.67	49.40	0
Direct-Bounded	8k/5	50.28	44.44	18.33	53.79	0
Generic	8k/5	50.56	50.00	20.00	53.99	0
ID-Normalized	8k/5	53.61	50.00	23.33	57.92	0
MU	8k/5	77.22	80.00	63.33	82.49	0
Natural-Prose	8k/5	41.11	35.75	8.33	47.62	0
Random-0	8k/5	48.89	43.33	16.67	53.70	0
Random-1	8k/5	50.56	45.00	20.00	53.99	0
Random-2	8k/5	50.28	47.78	18.33	55.25	0
Unstructured	8k/5	48.89	44.44	15.00	51.69	0
Direct-Bounded	8k/10	49.17	49.44	16.67	52.47	0
Generic	8k/10	54.17	55.00	25.00	56.82	0
ID-Normalized	8k/10	56.94	55.00	28.33	60.31	0
MU	8k/10	79.72	80.56	65.00	84.50	0
Natural-Prose	8k/10	53.06	58.96	23.33	56.27	0
Random-0	8k/10	52.78	50.83	26.67	57.71	0
Random-1	8k/10	52.22	49.72	23.33	55.94	0
Random-2	8k/10	49.44	51.11	16.67	53.85	0
Unstructured	8k/10	55.83	49.44	28.33	57.46	0

B.4.1 GPT-2 RANKED-RETRIEVAL DIAGNOSTICS

B.4.2 WHY GPT-2 4K/10 AND 8K/10 DIFFER DESPITE IDENTICAL RANKINGS

A targeted audit reproduces the reported GPT-2 packing and evidence scores without new model calls. All 60 full BM25 rankings, and all 60 top-10 rankings, are identical between 4k/10 and 8k/10. However, only 33 of 60 packed contexts are nested. Packing scans the ranking, skips any *whole* passage that does not fit the remaining token budget, and stops after accepting ten passages. A larger token budget can therefore admit an earlier long passage that was skipped at 4k, causing the ten-passage cap to be reached before a later evidence-bearing passage.

Exactly ten questions, G021–G030, lose half of their requested evidence at 8k/10: each falls from 100% to 50% requested-evidence recall because the later passage containing the `family_id` source fact is no longer packed. These ten half-slot losses account exactly for the aggregate drop from 88.89% to 80.56%.

For example, G021 packs nine passages at 4k (3,999 tokens), including the later passage carrying the requested family identifier. At 8k it accepts ten earlier passages (7,642 tokens) and reaches the passage cap before that later evidence. Requested-evidence recall is computed from source facts in the packed context independently of answer quality.

B.5 ENTITY-AWARE MU ACCESS

Entity-aware routing is an MU-specific access policy and is not pooled with the matched common-BM25 comparison. The policy uses only the public question and record metadata; reference answers are not exposed. CNN and GPT-2 Stage-B values remain reported aggregates where raw trial exports are unavailable.

Table 16: VAE entity-aware full saved aggregate grid.

Track	Budget/k	Ans. (%)	ER (%)
mu_entity_field	4k/5	90.42	96.25
mu_entity_field	4k/10	91.94	100.00
mu_entity_field	8k/5	93.06	96.25

Track	Budget/k	Ans. (%)	ER (%)
mu_entity_field	8k/10	94.31	97.50
mu_entity_semantic	4k/5	92.64	96.25
mu_entity_semantic	4k/10	93.75	99.58
mu_entity_semantic	8k/5	93.61	96.25
mu_entity_semantic	8k/10	89.17	97.50

Table 17: CNN entity-aware grid, reported to 0.1 percentage point; raw Stage B trials unavailable.

Track	Budget/k	Ans. (%)	ER (%)
Entity+BM25	4k/5	88.9	98.9
Entity+BM25	4k/10	88.3	98.9
Entity+BM25	8k/5	93.3	100
Entity+BM25	8k/10	86.1	100
Entity+semantic	4k/5	83.3	98.9
Entity+semantic	4k/10	85	98.9
Entity+semantic	8k/5	85	98.9
Entity+semantic	8k/10	82.8	100

Table 18: GPT-2 current corrected entity-aware grid; reported aggregates, raw Stage B trials unavailable.

Track	Budget/k	Ans. (%)	ER (%)
mu_entity_field	4k/5	98.33	100.00
mu_entity_field	4k/10	99.17	100.00
mu_entity_field	8k/5	98.33	100.00
mu_entity_field	8k/10	99.17	100.00
mu_entity_semantic	4k/5	97.22	97.22
mu_entity_semantic	4k/10	98.33	100.00
mu_entity_semantic	8k/5	97.22	97.22
mu_entity_semantic	8k/10	98.33	100.00

The full four-setting entity-aware grids above are descriptive interface configurations, not evidence that entity routing itself is uniquely attributable to E/S/R/D/G. Equivalent routing over alternative representations was not evaluated in the matched benchmark.

C VAE Repeat Stability

C.1 Same-checkpoint, byte-disjoint repeat

The repeat applies the same VAE checkpoint and analysis procedures to 1,000 different CelebA images. Repeat images are compared byte-for-byte against the complete historical validation pool, and byte-identical overlap is excluded; two cross-pool duplicate groups are excluded in their entirety. Thus, the repeat is image-disjoint with respect to the historical validation pool while holding the checkpoint and analysis procedure fixed.

The repeat manifest uses 1,000 images, with a 16-image pair-intervention subset and a 500-image conditional subset. Images are processed at 256×256 with the recorded non-EMA model weights. The purpose is descriptive repeat stability: whether the same analysis procedure produces similar findings on different images.

The semantic estimator is the legacy unsigned spatial-maximum association. Maximizing spatial locations may differ across attributes and samples, so these scores do not identify signed attribute-edit directions. Pair effects measure decoded pixel changes rather than attribute changes. The result therefore supports the distinction between accessibility and repeat stability: an interface can faithfully expose a recorded association even when that association is not the dominant one when the same analysis is repeated on different images.

Table 11: CNN direct-evidence ranking metrics (fractions).

Condition	k	P@k	Recall@k	nDCG@k	MRR@k
Direct-Bounded	5	.063	.183	.103	.117
Generic	5	.077	.225	.138	.158
ID-Normalized	5	.057	.133	.095	.134
MU	5	.380	.457	.587	.762
Natural-Prose	5	.110	.419	.217	.182
Random-0	5	.060	.176	.114	.139
Random-1	5	.063	.183	.106	.112
Random-2	5	.060	.169	.099	.113
Unstructured	5	.073	.204	.128	.152
Direct-Bounded	10	.058	.351	.163	.147
Generic	10	.058	.347	.183	.180
ID-Normalized	10	.043	.208	.124	.148
MU	10	.280	.545	.611	.769
Natural-Prose	10	.070	.519	.255	.200
Random-0	10	.050	.332	.167	.164
Random-1	10	.050	.319	.154	.133
Random-2	10	.055	.344	.161	.146
Unstructured	10	.055	.323	.171	.173

Table 13: GPT-2 direct-evidence ranking metrics (fractions).

Condition	k	P@k	Recall@k	nDCG@k	MRR@k
Direct-Bounded	5	.163	.419	.343	.444
Generic	5	.177	.472	.384	.487
ID-Normalized	5	.177	.472	.389	.498
MU	5	.263	.828	.825	.933
Natural-Prose	5	.143	.329	.327	.460
Random-0	5	.160	.412	.345	.456
Random-1	5	.153	.408	.334	.450
Random-2	5	.163	.443	.365	.463
Unstructured	5	.153	.404	.344	.476
Direct-Bounded	10	.090	.486	.367	.454
Generic	10	.095	.511	.399	.492
ID-Normalized	10	.095	.511	.404	.503
MU	10	.133	.833	.827	.933
Natural-Prose	10	.112	.585	.423	.486
Random-0	10	.092	.501	.377	.469
Random-1	10	.085	.451	.350	.454
Random-2	10	.090	.499	.384	.471
Unstructured	10	.088	.460	.366	.483

Table 14: Audit of the GPT-2 4k/10 versus 8k/10 packing difference.

Audit item	Result
Questions	60
Identical full rankings	60/60
Identical top-10 rankings	60/60
Nested packed contexts	33/60
Questions losing evidence at 8k	10 (G021–G030)
Questions gaining evidence at 8k	0
Evidence recall, 4k/10	88.89%
Evidence recall, 8k/10	80.56%
New model calls used in audit	0

Table 15: Entity-aware MU access at 8,000 tokens and ten passages.

Entity-aware MU access (different routing policy)				
Model	MU access	Factual	Evidence	Source status
VAE	Entity+BM25	94.31	97.50	Full aggregate
VAE	Entity+semantic	89.17	97.50	Full aggregate
CNN	Entity+BM25	86.1	100.0	Reported aggregate
CNN	Entity+semantic	82.8	100.0	Reported aggregate
GPT-2	Entity+BM25	99.17	100.00	Reported aggregate
GPT-2	Entity+semantic	98.33	100.00	Reported aggregate

These are MU-specific routing configurations, without matched entity-aware alternatives. CNN/GPT-2 entries are reported aggregates; raw Stage B trials were not supplied.

Table 19: VAE repeat-stability endpoints and exploratory semantic-retention diagnostics.

Endpoint	Value	Status
Channel-correlation MAD	**0.024917**	Main descriptive
Interaction-category agreement	**2,488/2,490**	Main descriptive
Original top attribute retained at top 1	**23.63% (121/512)**	Main descriptive
Original top attribute retained within top 3	53.71% (275/512)	Exploratory
Original top attribute retained within top 5	68.55% (351/512)	Exploratory
Original top attribute retained within top 10	83.98% (430/512)	Exploratory
Within-channel Spearman, mean	0.6037	Exploratory

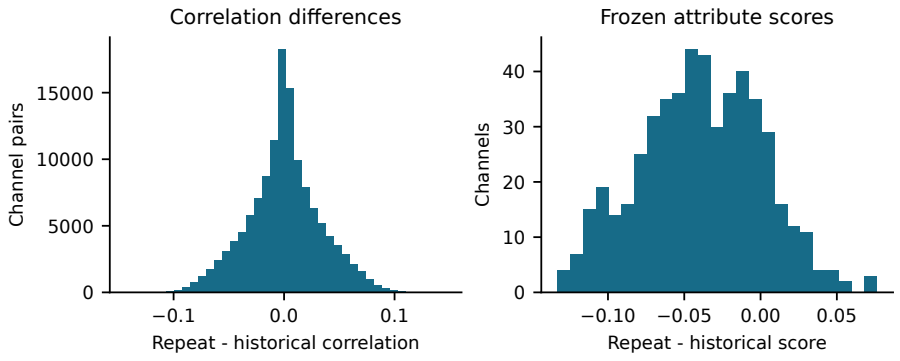

Figure 7: Descriptive VAE repeat differences: channel-pair correlation changes and changes in the frozen original attribute's score.

D CNN BOUNDED-ACCESS CAUSAL UTILITY

D.1 SELECTION RULE AND ACCESS REGIMES

The bounded protocol uses a deterministic consumer operating on `conv3` and selects eight filters. A candidate must be explicitly identifiable in the retrieved text and expose all ten class-mean responses; conflicting measurements exclude the candidate. For target class c and filter f,

$$\Delta_f = \mu_{f,c} - \max_{j \neq c} \mu_{f,j}, \qquad r_f = \frac{\Delta_f}{\mu_{f,c} + 10^{-12}}.$$

A filter qualifies when $\Delta_f > 0$ and $r_f \geq \tau$, for $\tau \in \{0, 0.10, 0.25\}$.

Under *strict* selection, the eight highest qualifying filters are chosen by decreasing Δ_f, with ascending filter ID as the tie-breaker; fewer than eight qualifying filters yields no executable selection. Under *fill*, the selector uses three ordered tiers: (1) qualifying filters, (2) target-primary filters below threshold, and (3) remaining complete candidates. Fill uses only identities and measurements visible in the retrieved context; it does not consult complete sources, hidden ownership metadata, gold answers, or intervention outcomes.

Direct-Bounded is evaluated with BM25 passage caps of 10, 20, 30, and 50 under the same frozen reader and selection rules. It produces zero complete candidate vectors in all 40 target-by-cap contexts, so no Direct-Bounded intervention is executed. This is an access-to-selection failure under the tested native representation–reader combination, not a measured zero causal effect or a claim that all native readers would fail. Direct-Complete remains executable, confirming that the required information exists upstream.

Table 20: CNN access regimes under the main fill, threshold-zero comparison.

Condition	Access	Outcome
Direct-Bounded	BM25, up to 50 native passages	No executable selection under the tested reader.
MU-Bounded	BM25, up to 50 MU passages	Executable eight-filter selection; **92.37% gain-70 targeted success**.
Direct-Complete	Complete relevant native source	Executable eight-filter selection; **92.54% gain-70 targeted success**.

Inspected high-ranked Direct passages sometimes contain source summaries or partial class-response information rather than complete attributable vectors; native packaging may also separate values from enclosing layer/filter identities. We therefore attribute non-executability to the tested combination of retrieval content, native packaging, and the frozen reader's explicit ownership/completeness requirements rather than to passage depth or chunk fragmentation alone.

D.2 DIRECT-BOUNDED DEPTH DIAGNOSTIC

We test whether Direct-Bounded non-executability is explained by the original passage cap rather than by the exposed native representation. This is an access-to-selection diagnostic only: no Qwen generation, model inference, or intervention is performed.

The Direct passage corpus, hash-verified BM25 implementation, candidate parser, fill selector, threshold $\tau = 0$, and eight-filter `conv3` requirement are frozen from the completed experiment. The cap-50 results reproduce the saved experiment for all ten targets. We then evaluate prefixes of 50, 100, 200, and 400 passages and *ALL*, where ALL denotes every existing Direct passage in ranked order. No ownership sidecar, concatenation repair, or alternative parser is introduced.

Increasing passage depth never exposes enough attributable information for the frozen selector to form an eight-filter set. Failure at ALL excludes passage depth alone as the explanation under this reader and passage representation. It does not establish that an alternative reader or entity-preserving native serialization would also fail.

Table 21: Direct-Bounded passage-depth diagnostic. Counts other than executability are medians across the ten target classes.

Cap	Executable	Complete	Usable	Partial	Identifiable
50	**0/10**	0.0	0.0	0.0	1.0
100	**0/10**	0.0	0.0	0.0	1.0
200	**0/10**	0.0	0.0	0.0	1.0
400	**0/10**	0.0	0.0	0.0	1.0
ALL	**0/10**	0.0	0.0	0.0	1.0

Per-target diagnostics additionally retain first-completion and first-usable ranks, later losses caused by conflicting evidence, exposed passage counts, text-volume measurements, and selected IDs. Identifiable-entity counts are an explicit-syntax diagnostic only and do not relax the frozen parser.

D.3 Sensitivity to passage cap, threshold, and intervention strength

Table 22: CNN end-to-end success/disruption (%) across MU sensitivities. Gain rows use targeted success; zero uses intervention disruption.

Rule	Threshold	Operator / Gain	Cap 10	Cap 20	Cap 30	Cap 50
strict	0.00	1	0.00	0.00	0.00	0.00
strict	0.00	10	0.00	3.92	43.77	47.27
strict	0.00	30	0.00	8.00	66.64	69.18
strict	0.00	50	0.00	8.58	70.37	72.49
strict	0.00	70	0.00	8.84	71.43	73.69
strict	0.00	zero	0.00	3.63	40.68	54.28
strict	0.10	1	0.00	0.00	0.00	0.00
strict	0.10	10	0.00	0.00	22.35	35.70
strict	0.10	30	0.00	0.00	33.62	51.86
strict	0.10	50	0.00	0.00	35.40	54.20
strict	0.10	70	0.00	0.00	35.89	55.07
strict	0.10	zero	0.00	0.00	19.34	43.73
strict	0.25	1	0.00	0.00	0.00	0.00
strict	0.25	10	0.00	0.00	5.97	10.45
strict	0.25	30	0.00	0.00	8.92	16.41
strict	0.25	50	0.00	0.00	9.34	17.51
strict	0.25	70	0.00	0.00	9.53	17.79
strict	0.25	zero	0.00	0.00	7.97	12.66
fill	0.00	1	0.00	0.00	0.00	0.00
fill	0.00	10	26.91	48.61	57.48	60.10
fill	0.00	30	49.13	77.45	84.86	86.80
fill	0.00	50	54.08	82.36	89.29	90.87
fill	0.00	70	55.95	83.87	90.50	**92.37**
fill	0.00	zero	20.52	36.58	51.70	**64.71**
fill	0.10	1	0.00	0.00	0.00	0.00
fill	0.10	10	26.91	48.61	57.48	60.10
fill	0.10	30	49.13	77.45	84.86	86.80
fill	0.10	50	54.08	82.36	89.29	90.87
fill	0.10	70	55.95	83.87	90.50	92.37
fill	0.10	zero	20.52	36.58	51.70	64.71
fill	0.25	1	0.00	0.00	0.00	0.00
fill	0.25	10	26.91	48.61	56.71	60.81
fill	0.25	30	49.13	77.45	84.53	87.63
fill	0.25	50	54.08	82.36	89.08	91.78
fill	0.25	70	55.95	83.87	90.41	93.20
fill	0.25	zero	20.52	36.58	50.76	65.06

The main text reports fill, threshold zero. At cap 50, targeted success rises from 60.10% at gain 10 to **92.37%** at gain 70, while zero-ablation disruption reaches **64.71%**. Thus, the access-dependent

trend persists across intervention strengths even though absolute success depends on the operator and gain.

Table 23: CNN bounded controls, fill policy and threshold zero. Probability/logit effects retain the saved operator-specific definition.

Condition	Op.	Success %	Δp (pp)	Δ logit	Executed
MU cap50	70	**92.37**	91.15	87.78	7677/7677
MU cap50	zero	**64.71**	58.32	4.95	853/853
Direct-Complete	70	**92.54**	91.32	88.04	7677/7677
Direct-Complete	zero	**67.17**	60.42	5.06	853/853
Matched-random-0	70	3.46	2.47	-38.19	7677/7677
Matched-random-0	zero	1.64	0.01	0.12	853/853
Matched-random-1	70	8.87	7.80	-27.23	7677/7677
Matched-random-1	zero	3.28	2.24	0.46	853/853
Matched-random-2	70	7.09	5.95	-38.41	7677/7677
Matched-random-2	zero	3.63	2.75	0.33	853/853

The matched-random repeats are random filter selections, not the random document partitions used in Structural Accessibility. Non-executable selections are tracked separately rather than assigned a zero causal effect.

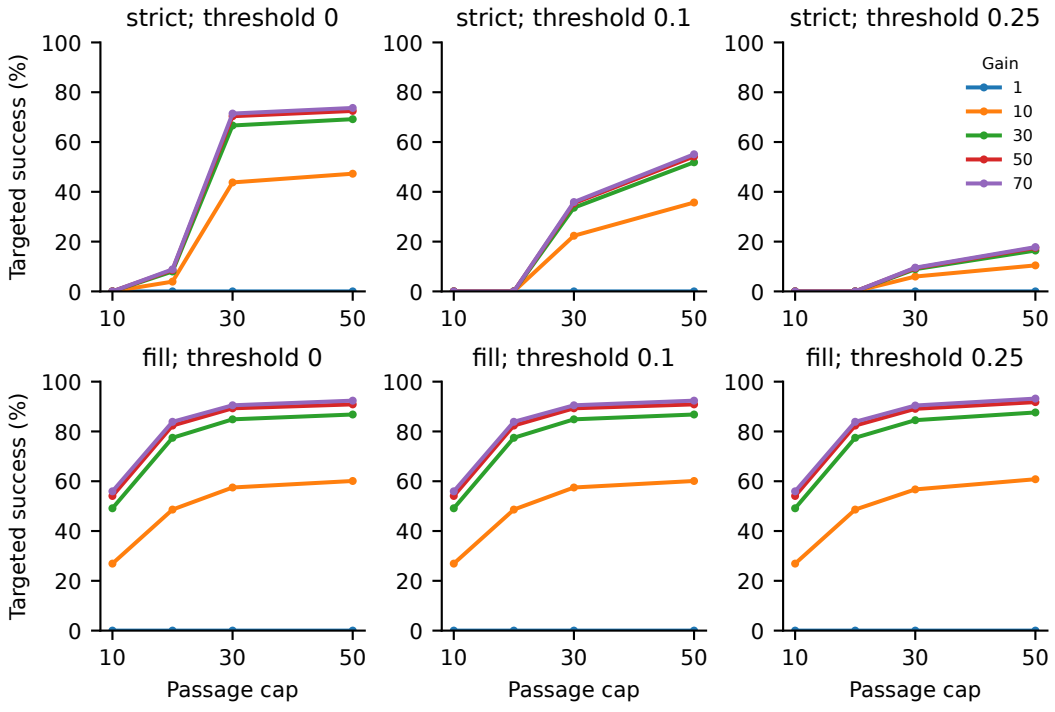

Figure 8: CNN amplification sensitivity across passage caps, selection rules, thresholds, and gains.

E CNN Complete-Access Selection Fidelity and Natural Patching

Construction uses 1,000 CIFAR-10 training images per class and evaluation uses 100 test images per class. Exact cross-partition duplicates are excluded and evaluation images are deduplicated before sampling. Findings are recomputed from construction-image post-ReLU spatial-mean `conv3` responses. Selection ranks target position among the top three classes, then decreasing selectivity, then numerical filter ID. This selector is distinct from the historical bounded selector.

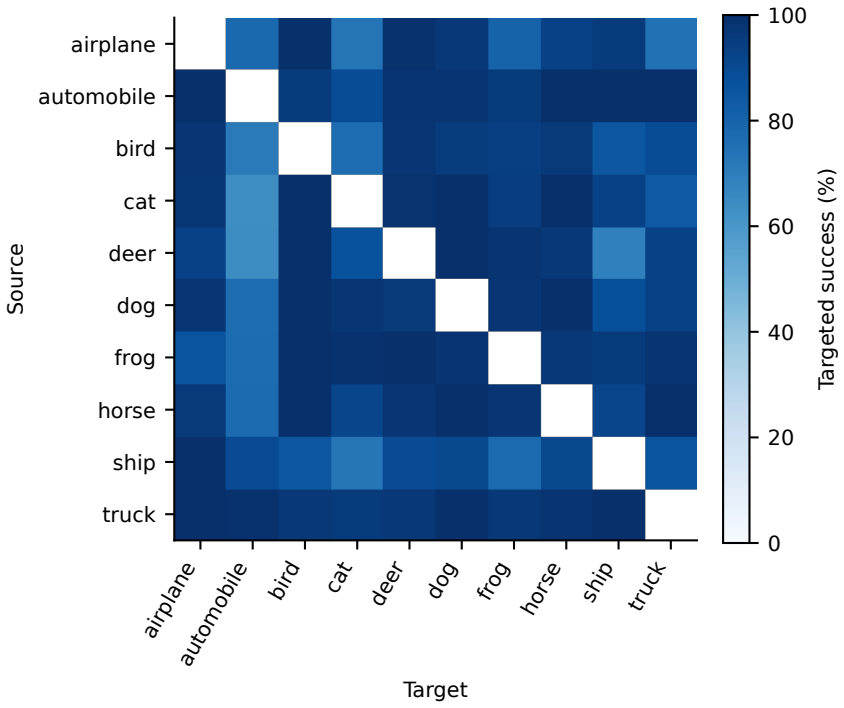

Figure 9: CNN MU cap-50, fill, threshold-zero, gain-70 targeted success by source and target class. Diagonal cells are outside the amplification task.

Each condition decodes its own complete records. MU and Direct-Numeric select **identical eight-filter sets for all ten targets**. Complete-access parity is therefore a Selection Fidelity result: the numerical information required by the frozen selector survives MU representation and architecture-specific realization. This experiment removes retrieval as a constraint; bounded access is evaluated separately.

Natural patching replaces selected post-ReLU, pre-pool maps with activations from fixed target-class donors sampled from the construction data. Across 9,000 source–target pairs, natural patching increases target probability by **17.641 pp** for both MU and Direct-Numeric (95% source-image interval $[16.680, 18.603]$ pp), versus **0.752 pp** for matched random filters ($[0.680, 0.824]$ pp). Zero-ablation produces a 53.664-pp probability drop for the shared MU/Direct-Numeric selection, versus 0.312 pp for random.

For completeness, retained causal utility relative to random is

$$\mathrm{CUR} = \frac{E_{\mathrm{MU}} - E_{\mathrm{Random}}}{E_{\mathrm{Direct}} - E_{\mathrm{Random}}}.$$

The pre-specified denominator gate requires the lower 97.5% paired Direct-Numeric minus random bound to exceed one percentage point. The gate passes for patching and zero ablation. CUR is exactly 1.0 with degenerate $[1, 1]$ intervals because MU and Direct-Numeric make identical selections and therefore share the same intervention outcomes. CUR consequently measures complete-access Selection Fidelity here rather than an additional accessibility result.

F GPT-2 Field Grounding

Held-out functional evaluation uses findings measured on 300 WikiText examples from 55 articles. The main attention, semantic, and relationship endpoints contain 12, eight, and four scorable tasks, respectively. Conceptual deletions such as `without_D`, `without_S`, and `without_R` remove role-labeled content; they need not correspond to deleting one entire mixed application field.

F.1 FIXED-EVIDENCE FIELD-CONTENT CHECKS

Table 24: GPT-2 fixed-evidence Recall@5. All tasks remain scorable; valid empty selections are abstentions and receive zero recall.

Condition	Attention ($n = 12$)	Semantics ($n = 8$)	Relationships ($n = 4$)
Full MU	**86.67**	**60.00**	**65.00**
Without D	**18.33**	60.00	65.00
Without S	86.67	**0.00**	65.00
Without R	86.67	60.00	**0.00**

The task-specific deletion effects show that the corresponding stored content is functionally used by these fixed-evidence selectors: removing D sharply reduces attention recovery, while removing S or R eliminates recovery on the corresponding semantic or relationship tasks. These are fixed-evidence field-grounding tests for the frozen selectors.

F.2 RETRIEVED-EVIDENCE ABLATIONS

Once BM25 retrieval is reintroduced, deletion changes both available information and passage competition. Effects therefore need not be monotonic, and valid-task masks may differ.

Table 25: GPT-2 BM25-retrieved Recall@5; each cell reports Recall@5, valid tasks/total, and format failures.

Condition	Attention ($n = 12$)	Semantics ($n = 8$)	Relationships ($n = 4$)
Full MU	43.33; 12/12; 0	51.43; 7/8; 1	85.00; 4/4; 0
Without D	52.73; 11/12; 1	60.00; 5/8; 3	80.00; 4/4; 0
Without S	50.00; 12/12; 0	47.50; 8/8; 0	80.00; 4/4; 0
Without R	23.33; 12/12; 0	25.00; 8/8; 0	50.00; 4/4; 0

Because format-failure counts differ, a failure-as-zero sensitivity is also reported.

Table 26: GPT-2 retrieved-evidence failure-as-zero sensitivity (Recall@5, %).

Condition	Attention	Semantics	Relationships
Full MU	43.33	45.00	85.00
Without D	48.33	37.50	80.00
Without S	50.00	47.50	80.00
Without R	23.33	25.00	50.00

For example, the apparent valid-task semantic increase without D (60.00% versus 51.43%) reverses when failures count as zero (37.50% versus 45.00%). We therefore use fixed-evidence deletion to test stored field content and treat retrieved-evidence ablations as end-to-end access diagnostics rather than fixed-evidence field-grounding tests. Full supplementary matrices are retained in the machine-readable companion export.

G GPT-2 SELECTION FIDELITY AND FRESH CAUSAL EVALUATION

G.1 SELECTION FIDELITY

The MU deterministic reader parses rendered numerical measurements and does not substitute hidden full-precision values for the values present in the record. The complete MU ranking is identical to Direct-Numeric across **all 144 attention heads**. The same ordered top seven are

```
L10_H00>L08_H10>L09_H09>L07_H09
>L04_H11>L10_H06>L09_H07.
```

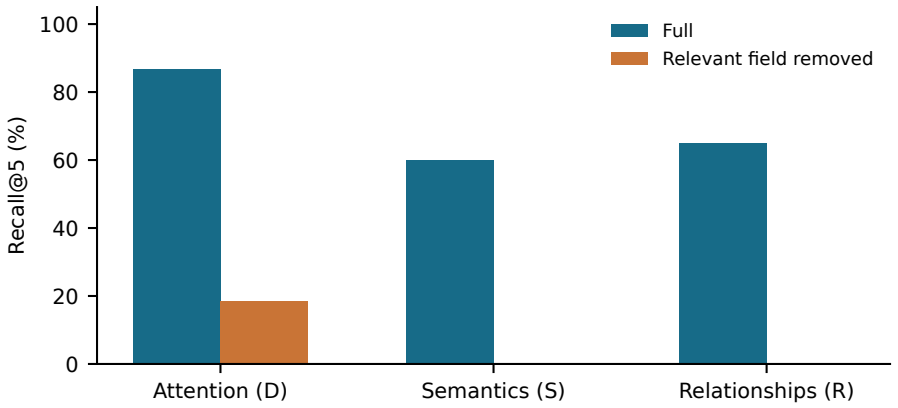

Figure 10: Task-specific fixed-evidence role deletion on held-out WikiText.

The layer-count profile is one head from L4, one from L7, one from L8, two from L9, and two from L10.

G.2 Fresh IOI protocol

The frozen selected set is evaluated on 192 fresh IOI examples arranged into 16 name-pair blocks, six fixed prompt frames, and both name orders. Clean head activations replace corrupted activations before the output projection at all token positions. Four examples have a clean-minus-corrupted target-minus-subject logit difference at most 10^{-6} and are excluded only from normalized recovery, leaving 188 valid ratios; all 192 remain available for raw-effect and target-preference endpoints.

G.3 Literature and random controls

The original prospective comparison uses a fixed seven-head literature reference. The shared MU/Direct-Numeric seven-head set achieves 61.8698% normalized recovery and 64.5833% target preference, compared with 56.5699% and 62.5000% for the literature reference. However, the pre-specified criterion requiring positive lower bounds on both prospective endpoints is not met. Accordingly, this comparison does not establish superiority over the literature reference.

A supplementary follow-up evaluates 100 random seven-head sets matched to the exact layer-count profile of the selected set. Control identities are frozen before their outcomes are measured; the model, tokenizer, examples, frames, intervention implementation, selected set, and outcome definitions remain unchanged.

Table 27: Fresh GPT-2 causal evaluation. Layer-matched controls are a supplementary random comparison; earlier unrestricted controls and the prospective literature reference are retained for context.

Condition	N examples	Valid ratios	Recovery (%)	Preference after (%)
MU/Direct-Numeric selection	192	188	**61.8698**	**64.5833**
Literature reference (7 heads)	192	188	56.5699	62.5000
Layer-matched random, mean of 100	192	188	**6.3048**	**8.9115**
Unrestricted random, mean of 32	192	188	5.6027	8.4310

The selected-set minus layer-matched-random recovery contrast is **55.5651 pp**. A paired name-block bootstrap gives a **95% interval of** $[51.2620, 59.7396]$ **pp**, and exhaustive sign flipping over the 16 paired blocks gives $p = 3.0518 \times 10^{-5}$. These analyses condition on the 100 frozen control sets, the fixed selected set, and six fixed frames; repeated evaluation over controls is not treated as independent replication.

Between-control-set variation is reported descriptively.

Table 28: Descriptive distribution of normalized recovery across the 100 frozen layer-matched random sets.

Mean	SD	Min	2.5th pct.	Median	97.5th pct.	Max
6.30	15.76	-36.33	-31.41	8.72	29.65	38.18

The layer-matched protocol was introduced after review of the original fresh outcomes and is therefore labeled supplementary, even though its 100 control identities were frozen before their outcomes were measured. The original unrestricted controls and prospective literature comparison remain archived rather than being replaced.

H Statistics and Reproducibility

H.1 Statistical procedures

CNN bounded-access intervals use 10,000 paired source-image bootstrap draws stratified by class, keeping all target comparisons from a source image together. They are exploratory marginal intervals conditional on the frozen configuration and are not adjusted across the full sensitivity grid. CNN complete-access intervals use 2,000 image-cluster draws; CUR denominator gates require the lower 97.5% paired Direct-Numeric minus random bound to exceed one percentage point.

The original GPT-2 literature comparison uses 4,000 paired name-block bootstrap draws with a two-endpoint maximum-standardized-deviation adjustment. The unrestricted-random supplementary contrast uses 50,000 name-block draws and exhaustive enumeration of the 2^{16} block signs. The layer-matched supplementary comparison retains the same 16-block paired structure while conditioning on 100 frozen layer-matched control sets. Its primary interval does not resample a population of control sets; between-set variation is reported descriptively. The sign-flip test assumes independent sign-symmetric block contrasts under the null rather than randomized treatment assignment.

VAE access and repeat endpoints are descriptive. The repeat uses the same fixed VAE checkpoint and analysis procedures on byte-disjoint images; no retraining uncertainty is estimated.

Structural Accessibility uses temperature-zero generation with a fixed seed and is reported as a fixed-benchmark descriptive evaluation rather than with repeated-generation Monte Carlo intervals.

H.2 Runtime and identity

The companion runtime manifest records available model, tokenizer, code, policy, checkpoint, and protocol identities.

Structural Accessibility answerer. Structural Accessibility uses Qwen3:8B with temperature 0, fixed seed 20260907, thinking disabled, a 32,768-token context limit, and a 4,096-token generation cap. Saved calls are reused when a run resumes. These settings are intended to make generation deterministic within the recorded runtime, although cross-hardware and cross-runtime bitwise reproducibility was not tested.

Causal-selection runtime. The CNN and GPT-2 causal-selection experiments do not use Qwen generation. Selections are produced by deterministic extraction and ranking rules, after which the selected components are frozen before intervention outcomes are measured.

Architecture-specific runtimes. The VAE repeat uses 512 latent channels, 256-pixel images, the same recorded non-EMA checkpoint as the historical analysis, and PyTorch 2.8.0+cu128. The fresh GPT-2 evaluation records PyTorch 2.6.0+cu124 and Transformers 4.57.6. The CNN complete-access runtime records PyTorch 2.9.0+cu128.

H.3 EVIDENCE PACKAGE AND REMAINING GAPS

The source package retains question-level scores, retrieval rankings, packed contexts, representation corpora, manifests, selection freezes, per-example intervention effects, statistical outputs, and integrity hashes. CNN audit artifacts contain the complete bounded and complete-access job records used for the main counts, including non-executable selections. VAE Stage A retains full ranking and packed-context metrics for the final 60-question cohort. GPT-2 retains the exact 144-head ranking comparison, fresh per-example outcomes, the literature reference, unrestricted random controls, and the frozen layer-matched-control manifest and outcomes.

Remaining gaps concern supplementary auditability rather than the matched main claims: corrected CNN/GPT-2 entity-aware Stage-B raw trials are unavailable where aggregate exports are reported. No older incompatible run is substituted for a missing artifact.

Machine-readable companion material. Exhaustive per-question family tables, full VAE ranking rows, full GPT-2 field-only and application-field ablations, all 100 layer-matched control identities and outcomes, the 32 unrestricted random controls, CNN source–target rows, and selection manifests are supplied machine-readably rather than duplicated line-by-line in the printed appendix.