DOES MEAN POOLING ERASE TEXT?

Anonymous authors
Paper under double-blind review

ABSTRACT

Mean pooling and normalization aggregate a token sequence into one vector, but do they erase its lexical identity? We study reconstruction from native Nemotron-3-Embed-8B passage embeddings, with controlled raw-token pooling across six models as complementary evidence. For a specified fixed-length causal transformer parameter family, we establish generic spherical injectivity in exact arithmetic, without guaranteeing injectivity of trained checkpoints or efficient reconstruction. Under open-weight and victim-query access, our method combines a learned base inverter, target-conditioned prefix guidance, residual-conditioned iterative correction, and victim re-embedding with a persistent candidate archive. Reconstructing eight-token AG-News fragments from native passage embeddings reaches 87.54% token accuracy and 72.60% exact match. On separate eight-token WikiText-2 raw-pooling benchmarks, Nemotron-3 and Jina-v4 reach 94.40% and 97.20% exact match, respectively. Recovery depends strongly on model and domain and deteriorates markedly in the tested 16- and 32-token raw-pooling settings. Candidate-oracle analyses show that most failed reconstructions lack the complete target in the generated archive, while numerical audits show that small residuals need not identify the original text. These findings distinguish architectural identifiability, numerical separation, and computational recoverability, exposing short-text reconstruction risks without implying universal invertibility.[1]

1 INTRODUCTION

Dense text embeddings support semantic comparison and retrieval, but their utility does not establish the privacy of the underlying text (Reimers & Gurevych, 2019; Song & Raghunathan, 2020). Vec2Text demonstrated this risk by recovering 92% of 32-token inputs exactly in its evaluated setting through iterative generation, re-embedding and correction (Morris et al., 2023). We ask a more specific question: does averaging contextual token states and normalizing the result necessarily erase sequence identity, and what prevents reconstruction when it does not? Although averaging is invariant to permutations of its vector inputs, contextualization can encode order before pooling.

Mathematical identifiability and empirical recovery are distinct. Search can fail on an injective map, while success on a finite benchmark does not establish global injectivity. Nikolaou et al. (2026) establish generic injectivity of last-token representations under their assumptions; their SipIt recovery algorithm requires the per-position hidden-state sequence. Neither directly resolves recovery from a normalized contextual mean, where pooling and normalization may introduce additional collisions.

Our primary application reconstructs eight unknown tokens from locally computed native Nemotron-3-Embed-8B passage embeddings (NVIDIA, 2026). The attacker sees one pooled vector, knows the weights and tokenizer, and can query candidate sequences and prefixes, but does not receive the target's tokenwise states or unnormalized norm. This is stronger than a hosted embedding-only API. Controlled raw-token pooling across six victims provides complementary architecture and length evidence, not interchangeable native-interface results.

Our framework combines a learned base inverter, target-conditioned prefix guidance, Vec2Text-style residual correction and victim re-embedding. Native Nemotron-3 recovery reaches 87.54% token accuracy and 72.60% exact match on AG-News. Separate raw WikiText-2 benchmarks reach 94.40% and 97.20% exact match for Nemotron-3 and Jina-v4. Exact recovery deteriorates markedly in the tested 16- and 32-token raw-pooling settings; candidate-oracle analyses show that complete targets

[1] Code and results: `https://anonymous.4open.science/r/mean-pooled-inversion`.

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.

are absent from the generated archives in these failed reconstructions. The paper separates architectural identifiability, numerical separation and computational recoverability rather than treating one as evidence of the others.

Contributions.

1. **Native and cross-model evidence:** short-text reconstruction from native passage embeddings, with six-model raw-pooling controls and length/domain evaluations revealing its limits.
2. **Scoped spherical injectivity:** a conditional analytic criterion and finite architectural witnesses for specified all-positions RoPE Transformer families, not guarantees for trained checkpoints.
3. **Mechanistic evaluation:** component and verifier ablations, GPU-time-matched baseline comparisons, numerical and candidate-failure analyses, and native defence–utility measurements with limited attacker adaptation.

2 Related Work

Embedding inversion. Prior work recovers lexical information (Song & Raghunathan, 2020), whole sentences (Li et al., 2023) and text from internal LLM states (Wan et al., 2024). Vec2Text iteratively corrects and re-embeds generated text (Morris et al., 2023). Transferable inversion learns a surrogate from leaked text–embedding pairs without querying the victim (Huang et al., 2024); ALGEN uses few-shot cross-model alignment (Chen et al., 2025); ZSInvert avoids encoder-specific training but retains black-box encoder access (Zhang et al., 2025). Adaptive attacks on noisy embeddings motivate testing beyond frozen-attacker sensitivity (Wang et al., 2026). Different observation models include classifier-based training-text inversion (Zhang et al., 2022) and gradient reconstruction (Zhu et al., 2019; Geiping et al., 2020), including TAG (Deng et al., 2021), LAMP's language-prior-guided search (Balunović et al., 2022) and partial-Transformer-gradient attacks (Li et al., 2024). Gradient leakage need not rely on memorization (Zhu et al., 2019). Our setting observes one normalized pooled vector, without target tokenwise states or gradients, under open-weight, query-enabled access.

Injective representations. NICE, Real NVP and Glow construct invertible transformations through coupling layers, with invertible channel mixing in Glow (Dinh et al., 2015; 2017; Kingma & Dhariwal, 2018). RevNet and i-RevNet use reversible blocks (Gomez et al., 2017; Jacobsen et al., 2018); invertible residual networks impose sufficient Lipschitz constraints (Behrmann et al., 2019). For aggregation, Deep Sets studies sum-based permutation-invariant representations (Zaheer et al., 2017), GIN connects injective multiset aggregation to graph discrimination (Xu et al., 2019), and neural moments establish injectivity for bounded-cardinality multisets under analytic non-polynomial activation and sufficient-width assumptions (Amir et al., 2023). Positional Transformer universality concerns approximation, not injectivity of a particular observation map (Yun et al., 2020). Most directly, Nikolaou et al. (2026) establish generic last-token injectivity for a specified causal decoder-only Transformer family, while their recovery algorithm uses tokenwise states. Our finite architectural witnesses must instead exclude positive-ray collisions after pooling contextual states and discarding magnitude; these earlier results do not directly establish that property.

Guided reconstruction. Iterative editing includes Mask-Predict's masked updates (Ghazvininejad et al., 2019), Levenshtein Transformer's insertions and deletions (Gu et al., 2019), and EDITOR's repositioning and insertion operations (Xu & Carpuat, 2021). Lexical constraints motivate grid beam search (Hokamp & Liu, 2017), constrained caption decoding (Anderson et al., 2017) and dynamic beam allocation (Post & Vilar, 2018). PPLM uses activation gradients (Dathathri et al., 2020), GeDi a generative discriminator (Krause et al., 2021), and FUDGE future-attribute predictors (Yang & Klein, 2021). NeuroLogic handles lexical constraints expressed in predicate logic (Lu et al., 2021); its A*esque extension adds lookahead (Lu et al., 2022). COLD uses Langevin sampling in a continuous relaxation (Qin et al., 2022), and Self-Refine model-generated feedback (Madaan et al., 2023). Our closest predecessor remains Vec2Text (Morris et al., 2023): target conditioning, residual correction, re-embedding and sequence beams are inherited, not claimed as new. We add guidance

from candidate-prefix victim states and study its contribution alongside numerical and complete-proposal limitations.

3 Observation Model and Recovery Geometry

3.1 Observations and Access

Let $\mathcal{V}$ be a finite token vocabulary, $T \in \mathbb{N}_{\geq 1}$ the known window length, and $x^\star \in \mathcal{V}^T$ the unknown tokens. A victim with parameters θ produces final hidden states $h_i(x;\theta) \in \mathbb{R}^d$, where $d \in \mathbb{N}_{\geq 1}$. For a nonempty sequence x, define

$$m_\theta(x) = \frac{1}{|x|}\sum_{i=1}^{|x|} h_i(x;\theta), \qquad F_\theta(x) = \frac{m_\theta(x)}{\|m_\theta(x)\|_2}, \tag{1}$$

where F_θ is defined when $m_\theta(x) \neq 0$. This uses the all-ones pooling mask $a_i = 1$, zero-based positions $p_i = i - 1$, and no pre-existing key/value cache. Raw controls observe $y = F_\theta(x^\star)$ without added prompts or special tokens.

Our primary native Nemotron-3 condition includes the fixed public context w_{pub} from `passage:` and leading special tokens:

$$F_\theta^{\mathrm{nat}}(x) = F_\theta(w_{\mathrm{pub}} \| x), \qquad y = F_\theta^{\mathrm{nat}}(x^\star). \tag{2}$$

Pooling includes this context; T counts only the unknown tokens. Targets reproduce native text tokenization, whereas reconstruction may query candidate token sequences that do not round-trip through the text interface (Appendix D.1). The raw-token theorem below does not automatically apply to this prompted map.

Under *open-weight, query-enabled access*, the reconstructor knows the victim weights, tokenizer, attention and pooling rules. It receives y, not the target's tokenwise states or pooled norm, and can query candidates and prefixes, including their hidden-state sums. This local token-space access is stronger than a hosted embedding-only API. Public text–observation pairs provide supervision; evaluation token labels enter metrics and privileged diagnostics, not search or selection. Victim weights remain frozen. We recover token windows, not entire documents, and distinguish native reconstruction from raw controls and native-wrapper geometry checks.

3.2 Generic Spherical Injectivity

Normalization identifies positive scalar multiples, not antipodal vectors; we therefore use *spherical*, rather than real-projective, injectivity. Architectural identifiability is conditional:

Proposition: generic spherical injectivity. Fix a finite $\mathcal{X} \subseteq \mathcal{V}^T$, $d \geq 2$, and an open, connected $\Omega \subseteq \mathbb{R}^p$. Suppose $\theta \mapsto m_\theta(x)$ is real analytic for every $x \in \mathcal{X}$. Assume each sequence has a nonzero pooled-mean witness in Ω, and each distinct pair has a witness in Ω at which its pooled means are noncollinear. Then an open, dense subset of Ω with Lebesgue-null complement exists on which F_θ is well-defined and injective on $\mathcal{X}$.

Architectural result. For $a_i = 1$, Lemma A.2 establishes these premises for the unrestricted Llama-2 family at each fixed $2 \leq T \leq 4096$ (Corollary A.3). Grouped-query attention (Ainslie et al., 2023) preserves the witness (Lemma A.4); full bidirectional attention admits it under a stronger RoPE frequency bound (Lemma A.5). The specified raw text-only families of all six models are consequently covered at $T \in \{8, 16, 32\}$ (Corollary A.6). Other pooling masks require their own witness verification.

Proof roadmap. The architectural contribution realizes separation with finite parameters. For formal token indicators b_v, the offset summaries $C_r(x) = \sum_{i=1}^{T} b_{x_{\max(1,i-r)}}$, $r = 0, \ldots, T-1$, jointly determine the sequence: the last offset identifies its first token, and successively smaller offsets identify the rest. A finite-softmax RoPE head (Su et al., 2024) approximates a pair-separating summary. An anchor, token marker and transported separator yield a nonzero pooled determinant

after the residual path and final RMSNorm (Zhang & Sennrich, 2019) for the constructed witness. The squared norms and selected pairwise determinants are then nontrivial analytic functions. Their finitely many zero sets are closed relative to Ω and Lebesgue-null (Mityagin, 2020), giving simultaneous generic separation. Appendix A supplies the complete construction and architecture restrictions. This does not certify trained checkpoints, prompted native maps or efficient inversion.

3.3 Numerical and Computational Limits

Represented numerical vectors $\widetilde{F}(c), \widetilde{y}$ need not have exactly unit norm. Their exact squared coordinate distance

$$E_{\mathrm{vec}}(c) = \|\widetilde{F}(c) - \widetilde{y}\|_2^2 \tag{3}$$

need not equal the computed dot-product score $\widehat{E}_{\mathrm{dot}}$ and separately normalized coordinate diagnostic $\widehat{E}_{\mathrm{coord}}$.

Fix a checkpoint and a finite candidate set $\mathcal{C}$ containing $x^\star$ and at least one wrong sequence, with the exact-real map F defined throughout. Write $\Delta_{\mathcal{C}} = \min_{c \in \mathcal{C} \setminus \{x^\star\}} \|F(c) - F(x^\star)\|_2$. Let η be the sum of the target and uniform candidate representation-error bounds relative to F. Then $\Delta_{\mathcal{C}} > 2\eta$ ensures unique nearest-vector recovery within $\mathcal{C}$, using exact distance comparisons. For computed-score selection with uniform error $|\widehat{E}(c) - E_{\mathrm{vec}}(c)| \le \varepsilon_E$, it suffices additionally that

$$(\Delta_{\mathcal{C}} - \eta)^2 - \eta^2 > 2\varepsilon_E. \tag{4}$$

These are sufficient triangle-inequality conditions, not measured guarantees: sampled margins and repeated forwards do not establish uniform error bounds (Appendix B.1).

In the raw Jina-v4 diagnostic, 291 of 1,000 targets have a tested wrong candidate with $\widehat{E}_{\mathrm{coord}} \le 10^{-3}$, yet none of the tested wrong candidates outranks the re-embedded gold against the same stored target. Threshold acceptance therefore differs from correct ranking; a small residual is not an exact-text certificate (Appendix B.2).

For a general nonempty set $\mathcal{C}$ of complete candidates available for final selection and $\widehat{x} \in \mathcal{C}$, exact-recovery failure means either $x^\star \notin \mathcal{C}$ or incorrect selection despite its inclusion. Per-position token coverage does not ensure a complete target candidate. We therefore distinguish candidate-oracle and selected accuracy; failed search alone cannot establish that the representation lost information.

4 Reconstruction Method

Figure 1 shows the native pipeline; raw controls use the same search structure with their own observation map. Every native query retains the public context, while generation predicts only the unknown tokens.

4.1 Base Inversion and Prefix Guidance

The base models $q_\phi(c \mid y)$ autoregressively. A learned projector turns the normalized observation into four continuous conditioning tokens prepended to a frozen causal generator: Ministral-3-8B-Base-2512 for native Nemotron-3. Base, corrector and Twist are trained separately for each victim/observation condition, including distinct native and raw checkpoints (Appendices D.1 and D.4). The generator's tokenizer is compatible with the victim, but their weights need not coincide.

Projective Twist guides prefixes towards the target. For native prefix $u = (u_1, \ldots, u_t)$, it queries $s_t(u) = \sum_{i=1}^{|w_{\mathrm{pub}}|+t} h_i(w_{\mathrm{pub}} \| u; \theta)$; raw controls omit the public context. For bidirectional victims this is a prefix-only computation, not an additive contribution assumed to persist after completion. A Transformer combines prefix tokens with the target, prefix direction, target-orthogonal component, norm/alignment statistics and relative depth to predict a value V_ξ and prefix-match logit G_ξ (Appendix D.2). For $t < T$, prefixes are ranked by

$$S_t(u; y) = \sum_{i=1}^{t} \log q_\phi(u_i \mid u_{<i}, y) + \lambda \big[V_\xi(u, y) + G_\xi(u, y) \big]. \tag{5}$$

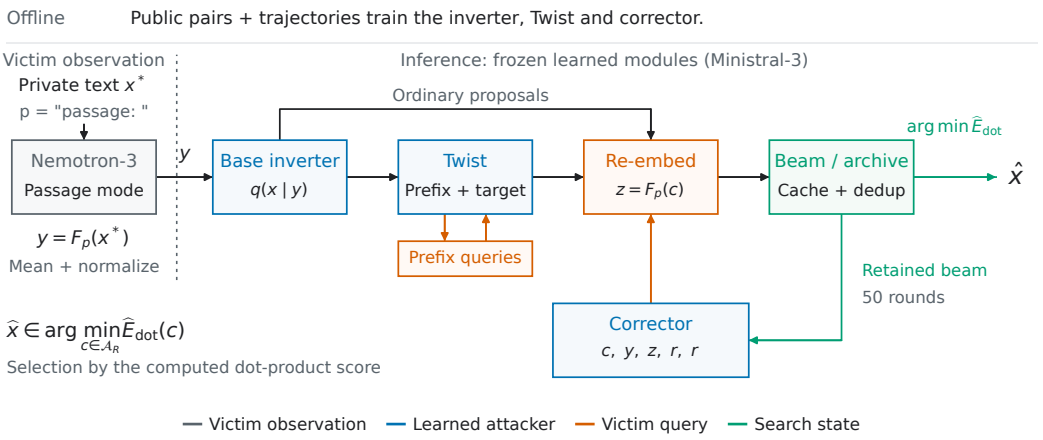


Figure 1: Native Nemotron-3 reconstruction with a Ministral-3 generator; F_p denotes F_θ^{nat}. Ordinary and Twist-guided proposals enter a parent-retaining correction beam; candidate and prefix queries use the native public context. Only retained beam members are expanded, with archive re-entry disabled. Final selection uses the computed victim score, not a lexical certificate. Offline supervision uses public training text.

The potential is recomputed, not accumulated, and omitted at $t = T$. A separate ordinary-base route also supplies initial candidates. Supervision uses prefixes of training targets, base hypotheses and corrected hypotheses. Twist learns target-and-depth-matched prefix ranking, cached-completion energy regression and prefix-match classification; gold-matching prefixes receive zero energy targets. It is a heuristic, not a calibrated energy estimate or completion bound; objectives and its value-only checkpoint criterion appear in Appendices D.3 and D.4.

4.2 Iterative Correction and Persistent Archive

We retain Vec2Text's generation–re-embedding–correction paradigm (Morris et al., 2023). For hypothesis c, let z be its victim observation and $r = \mathcal{N}(\mathcal{N}(y) - \mathcal{N}(z))$, with $\mathcal{N}(v) = v/\max(\|v\|_2, \varepsilon_{\text{n}})$. Our duplicate-residual corrector models

$$q_\psi(c' \mid c, y, z, r, r). \tag{6}$$

Separate projectors transform the four vector channels into four conditioning tokens each, followed by embeddings of the current hypothesis. Duplicating the residual supplies no additional observed information. Teacher-forced cross-entropy trains base/corrector projectors with frozen generators; corrector states are base-generated hypotheses (Appendices D.1 and D.3). Corrections regenerate complete sequences, allowing coordinated edits.

The initial archive $\mathcal{A}_0$ combines ordinary and completed Twist proposals; $\mathcal{B}_0 = \text{Low}_b(\mathcal{A}_0)$ retains the lowest computed energies. Each active parent generates one greedy and one sampled correction, forming $\mathcal{D}_k$. The update is

$$\mathcal{B}_{k+1} = \text{Low}_b(\mathcal{B}_k \cup \mathcal{D}_k), \qquad \mathcal{A}_{k+1} = \mathcal{A}_k \cup \mathcal{D}_k. \tag{7}$$

Unions deduplicate complete token-ID tuples. Archive re-entry is off. Under fixed scores, this parent-retaining beam already preserves the best archive energy; persistent retention supports oracle diagnostics and provenance, not an independent minimum-energy gain (Appendix D.5). The multi-model evaluator records lineage and memoizes complete-sequence observations. This avoids repeated re-embedding, not corrector generation or prefix evaluation; original Llama-2 runs do not use this cross-round observation cache.

4.3 Verification and Budgets

Re-embedding uses the actual victim and the condition's observation map. Ranking uses $\widehat{E}_{\text{dot}}$, the computed score corresponding to $2 - 2u^\top v$ for normalized candidate/target vectors, not an exact

Victim	Cohort	Token	EM	Oracle token
Primary application: native passage observation				
Nemotron-3	AG-News training split	87.538	72.60	93.688
Complementary raw-pooling benchmarks				
Llama-2	WikiText-2 validation	99.925	99.70	$99.950^{\dagger}$
Llama-3.1	Wikipedia	92.713	77.10	94.838
Nemotron-3	WikiText-2 test	98.263	94.40	98.925
GritLM	WikiText-2 test	98.888	96.70	99.263
Jina-v4	WikiText-2 test	99.213	97.20	99.450
Llama-Nemotron-1B	WikiText-2 test	73.650	45.60	87.575

Table 1: Eight-token reconstruction (percentages), 1,000 windows per row. Native and raw conditions differ in cohorts and trained components. AG-News records are held out from attacker training. Oracle selects a complete candidate; † denotes a historical saved oracle, not a re-audited full archive. Domain confirmations appear in Table 24.

squared coordinate distance. After R rounds, return

$$\widehat{x} \in \arg\min_{c \in \mathcal{A}_R} \widehat{E}_{\mathrm{dot}}(c). \tag{8}$$

Neither gold tokens nor Twist scores enter this final ranking. Verification is victim evaluation, not a lexical proof: the residual flag does not terminate fixed-round search.

Reference $T = 8$ runs use eight ordinary proposals, $b = 64$, $R = 50$ and $\lambda = 1$. Up to 6,400 corrector proposals precede deduplication; Twist additionally evaluates 45,120 nonterminal prefixes. It retains a 128-prefix beam, expanding the top 64 base tokens per prefix. Sampled complete proposals use the top 32 logits at temperature 0.8. All learned components remain frozen during evaluation. Appendices D.5–D.6 specify sampling, prefix-beam settings and per-model memoization. We count prefix and complete-sequence evaluations separately and measure GPU time; equal proposal or round budgets need not imply equal computation.

5 Experiments and Mechanistic Analysis

5.1 Protocol

Native Nemotron-3 passage reconstruction is the primary application; six raw-pooling pipelines provide complementary controls. Reference $T = 8$ base/corrector training uses one million WikiText-2 training windows, not independent source documents, re-embedded under each condition's observation map. Corrector training uses a fixed cache of base-generated hypotheses, not on-policy correction trajectories (Appendix D.3). Condition-specific components are selected on calibration data and frozen for evaluation (Appendix D.4). Table 1 uses 1,000 windows per condition. Native AG-News sources are held out from attacker training. Exact-window overlap audits and exclusion analyses appear in Appendix C; excluding overlapping raw WikiText targets changes Nemotron-3 and Jina-v4 EM by less than 0.07 percentage points each.

Token accuracy measures positionwise token-ID agreement; exact match (EM) requires the complete window. Candidate-oracle accuracy chooses the best complete generated sequence using labels only after search. Inference costs distinguish generated proposals, victim sequences and GPU time. Appendix E gives cohort, statistical, matched-budget and adaptation protocols. No calibration accuracy is substituted for evaluation performance.

5.2 Native and Cross-Model Reconstruction

Native Nemotron-3 reaches 87.54% token accuracy and 72.60% EM, with a row-bootstrap 95% EM interval of $[69.90, 75.30]\%$. This observation is an actual passage embedding, evaluated under local token-space access rather than a hosted API. All targets satisfy native tokenization; 99.5% of selected outputs also round-trip, and the five exceptions are incorrect reconstructions (Appendix D.1). Native/raw conditions use different cohorts and trained components, so their difference does not isolate an interface effect.

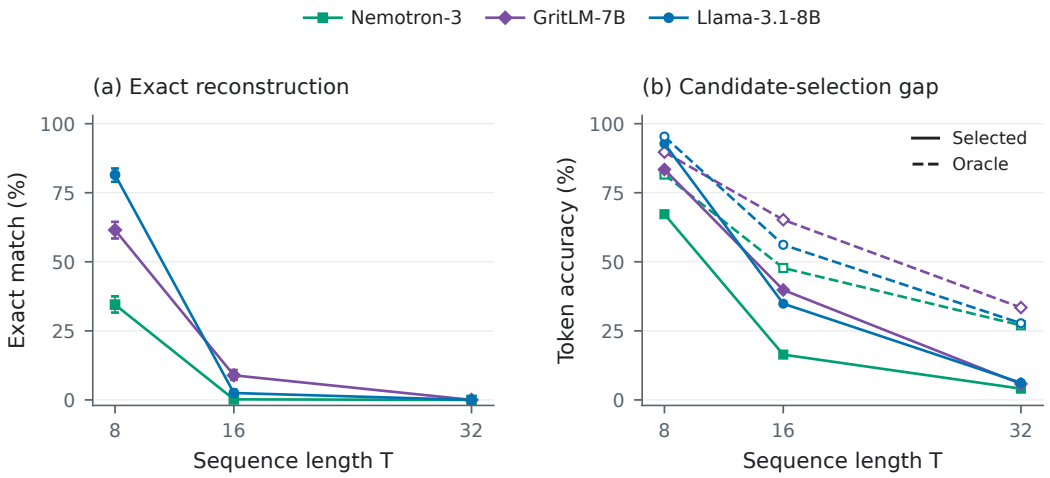


Figure 2: Raw-pooling recovery on 1,000 paired AG-News records per model at each length, using length-specific trained components. (a) EM with pointwise 95% Wilson intervals. (b) Selected and best-complete-candidate token accuracy; the oracle cannot mix tokens across candidates. Native observations are not included. Equal correction rounds do not imply equal computation.

Strong short-window recovery is not confined to Llama-2: raw Nemotron-3 and Jina-v4 attain 94.40% and 97.20% EM on WikiText-2. However, Llama-Nemotron-1B reaches only 45.60%, and separate AG-News confirmations reduce raw Jina-v4 and Llama-Nemotron-1B EM to 38.70% and 2.10% (Appendix E.5). The evidence therefore supports specified model/domain settings, not a single overall accuracy or universal model ranking.

5.3 Length, Domain and Failure Mechanisms

The raw Nemotron-3, GritLM and Llama-3.1 suites use the same ordered 1,000 AG-News test records across $T = 8, 16, 32$ within each model; tokenization and eligible records can differ across models. Longer-window components are separately trained, not zero-shot extensions of $T = 8$. All retain 64 parents and 50 correction rounds. Figure 2 shows deterioration in every suite: $T = 16$ EM ranges from 0.20% to 8.90%, and no $T = 32$ window is recovered exactly. Raw Nemotron-3 already performs worse on AG-News than WikiText at $T = 8$, identifying domain as another factor.

Candidate archives explain these failures. All 274 native Nemotron-3 failures lack a complete target candidate, although oracle token accuracy exceeds selected accuracy by 6.15 points. At $T = 32$, all three suites likewise lack complete targets; their oracle token accuracies are only 27.00–33.43%. Even token unions cover only 77.55–84.01% of target positions, with complete positionwise support in 0.7–3.7% of windows. Thus failures involve missing proposals, not merely selecting or assembling universally available tokens. Some selection errors nevertheless occur: raw GritLM misses two available targets. Full partitions and the historical Llama-2 oracle limitation are in Appendix E.5. These are limitations of the tested training/search budgets, not evidence of an information-theoretic length threshold or representation collisions.

5.4 Components, Verification and Cost

Four component conditions share checkpoints, initial base candidates and retained-component settings within each victim. Three inference seeds (1234/2025/3407), not independent training runs, are averaged per window before paired bootstrapping (Appendix E.1). On native Nemotron-3, correction raises EM from 3.67% to 70.77%; Twist then adds 2.40 points (95% CI $[1.37, 3.43]$), reaching 73.17%. Its incremental Llama-2 gain is 0.53 points ($[0.23, 0.90]$). These gains cost additional computation: native inference takes 74.36 seconds with Twist versus 51.34 without it.

	Native Nemotron-3			Llama-2 raw		
Condition	Token	EM	GPU s	Token	EM	GPU s
Base	23.32	3.67	0.46	81.38	47.70	0.40
Base + Twist	40.00	17.00	22.57	97.06	90.40	13.10
Base + corrector	86.78	70.77	51.34	99.70	99.17	44.10
Full	88.48	73.17	74.36	99.94	99.70	61.99

Table 2: Component comparison: percentages and H200 GPU-seconds per window, averaged over three inference seeds on 1,000 windows per victim. Enabled correctors run 50 rounds. These are separate repetitions from Table 1, not equal-compute ablations. Full metrics and timing scope appear in Appendix E.1.

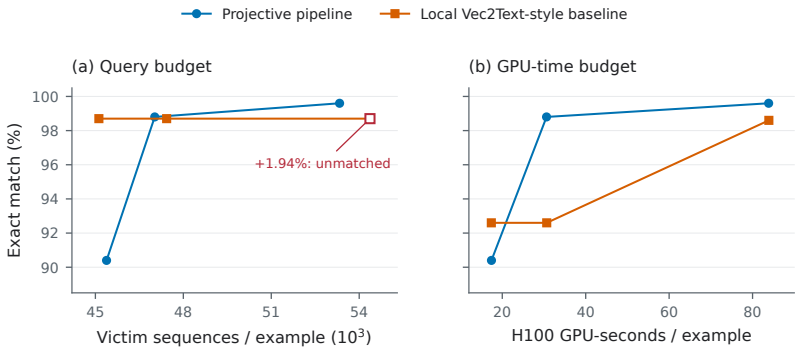


Figure 3: Llama-2, $T = 8$: separately calibrated query and H100-time comparisons on 1,000 windows. Query counts include prefixes. All time matches meet 5% tolerance; the red query point fails 1% tolerance. Lines join measured configurations; neither panel establishes uniform dominance.

Verification. Offline rescoring isolates selection on identical full-pipeline archives. Native stored victim energy selects 73.17% EM, versus 2.57% for base likelihood and 0.70% for the depth-four Twist score. Exact verification of their top-128 shortlists reaches 26.20% and 22.80%. Whenever gold survives shortlisting, minimum stored energy selects it in every evaluated panel, an empirical pool property rather than a global certificate. Full K curves, LM/random controls and the separate common recorded-parent subset for corrector likelihood appear in Appendix E.2.

Vec2Text-style baseline comparison. The local adaptation of Vec2Text (Morris et al., 2023) shares the victim, base and 1,000 Llama-2 windows, but uses its own raw-residual corrector. Table 3 compares separately calibrated GPU-time-matched configurations. Our 10/50-round runs improve EM but process more victim sequences; the zero-correction run has lower EM. Query matching is separate: the 50-round comparison fails its 1% tolerance by reaching a 1.94% discrepancy (Figure 3). There is no uniform dominance in accuracy, queries and compute. Full settings appear in Appendix E.3.

5.5 Defence Sensitivity and Retrieval Utility

We evaluate native Nemotron-3 on a separate 256-window AG-News bank and labelled SciFact retrieval (Wadden et al., 2020): clean observations, two quantization levels and two dimension-normalized Gaussian-noise levels selected on separate calibration data. The tasks use different populations; noise measurements average three independent realizations within each window/query, not repeated-query averaging of one target.

At noise level $\alpha = 0.8$, frozen-attacker EM falls from 68.75% to 13.28%, while nDCG@10 changes from 0.8333 to 0.8219. Adaptation trains base/corrector projectors and Twist on disjoint training data with defended observations, retaining frozen backbones. It reaches 11.20% EM: a -2.08-point difference from the frozen noisy attacker, with CI $[-4.56, 0.26]$ including zero. Verification also uses the defended channel; clean-weight verification was not evaluated. The local protocol retains

Method	b	R	Token	EM	Victim seq.	GPU s
Vec2Text-style	8	28	96.04	92.60	1,737.0	17.38
Ours	64	0	96.95	90.40	45,380.7	17.41
Vec2Text-style	8	49	96.04	92.60	3,081.0	30.66
Ours	64	10	99.53	98.80	47,024.4	30.61
Vec2Text-style	34	9	99.10	98.60	9,283.0	83.91
Ours	64	50	99.83	99.60	53,330.7	83.86

Table 3: Llama-2 $T = 8$ baseline comparison on 1,000 windows, seed 1234. Accuracy is in percentages; costs are per window on H100 NVL. Each pair meets the 5% GPU-time tolerance, not a query match. Victim counts include prefixes; b is beam width and R correction rounds. The baseline is a local adaptation, not a released Vec2Text checkpoint.

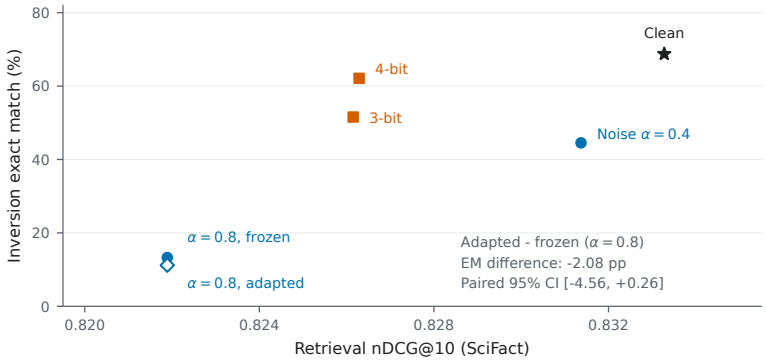


Figure 4: Native Nemotron-3 inversion (256 windows) and SciFact retrieval (300 queries). Noise results average three draws per window/query; verification uses the defended channel. Adaptation updates projectors and Twist only; its paired EM interval includes zero. This cross-task comparison is not a per-record privacy guarantee.

scalar prefix norms and does not test repeated-query averaging. This bounded adaptation is not a strongest-attacker test or differential-privacy guarantee (Appendix E.4).

6 Limitations and Privacy Implications

Recovery assumes known length, tokenizer and observation rules, local weights, queries and auxiliary supervision. It is not an embedding-only hosted-API attack or reconstruction of arbitrary documents. Native reconstruction is established for Nemotron-3; other reconstruction conditions are raw-pooling controls. The exact-real theorem does not certify these checkpoints or native interfaces. Length/domain failures are budget-dependent, not information-theoretic impossibility results. Attacker-training overlap audits do not exclude victim/generator pretraining exposure, and online costs do not include all offline supervision and training. The defence study tests limited adaptation, not every use of the available model weights. Small residuals do not certify the original text, and failed reconstruction cannot establish privacy. These limits preclude treating pooling or normalization as a general anonymization mechanism.

7 Conclusion

Normalized mean pooling need not erase sequence identity. Finite architectural witnesses establish scoped generic spherical injectivity, while native and raw-pooling experiments demonstrate short-text recovery under open-weight, query-enabled access. Length and domain degradation, candidate failures and numerical ambiguity separate identifiability from practical inversion. Component, cost and defence analyses expose both the reconstruction risk and the limits of the tested procedure, without implying universal invertibility or privacy for unsuccessful cases.

AI-USE STATEMENT

AI assistance was used for manuscript drafting and editing, implementation and debugging support, and literature and consistency checks. The authors are responsible for the scientific claims, citations, code and reported results; AI assistance does not replace their verification.

ETHICS STATEMENT

Our experiments study reconstruction from representations of public benchmark text, with observations generated by locally executed models; they do not use private embedding stores. We evaluate the open-weight, query-enabled setting specified in Section 3.1, not unauthorized access to a deployed embedding service. Public availability does not imply that source text is free of identifying information. The methods are dual-use: understanding recoverability can inform privacy assessments, but reconstruction could expose sensitive text if applied to embeddings obtained elsewhere. Our results therefore motivate caution in treating embeddings as anonymized substitutes for text; they neither establish universal recoverability nor certify privacy for unsuccessful cases. Our anonymized release focuses on research code, configurations, cohort-construction scripts and aggregate results. Third-party model weights and source corpora are not bundled; users obtain them from their original distributions under the applicable access conditions and terms.

REPRODUCIBILITY STATEMENT

Section 3.1 specifies the observation and access contract. Appendix D describes the architectures, training objectives, checkpoint selection and inference budgets, while Appendix B specifies the numerical diagnostics. Appendix C reports cohort-overlap checks and exclusion-based sensitivity analyses, including their limits. Appendix E specifies the component and verifier ablations, matched-cost comparisons, and defence–utility evaluation with attacker adaptation, including their statistical and budget controls. The theoretical assumptions and complete proofs appear in Appendix A. The anonymized supplementary artifact is available at the repository linked in Footnote 1. It contains reconstruction and analysis code, executable witness checks, and a configuration manifest linking reported results to model and tokenizer revisions, learned checkpoints, seeds and search budgets. Cohort-construction scripts and manifests identify the source datasets, splits, token-window construction and evaluation rows, distinguishing development, reused and frozen evaluation cohorts. The artifact also includes the frozen outputs needed to regenerate the reported tables and plots.

REFERENCES

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. GQA: Training generalized multi-query transformer models from multi-head checkpoints. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4895–4901, 2023. doi: 10.18653/v1/2023.emnlp-main.298. URL https://aclanthology.org/2023.emnlp-main.298/.

Tal Amir, Steven J. Gortler, Ilai Avni, Ravina Ravina, and Nadav Dym. Neural injective functions for multisets, measures and graphs via a finite witness theorem. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Guided open vocabulary image captioning with constrained beam search. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 936–945, 2017. doi: 10.18653/v1/D17-1098. URL https://aclanthology.org/D17-1098/.

Mislav Balunović, Dimitar I. Dimitrov, Nikola Jovanović, and Martin Vechev. LAMP: Extracting text from gradients with language model priors. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Jens Behrmann, Will Grathwohl, Ricky T. Q. Chen, David Duvenaud, and Jörn-Henrik Jacobsen. Invertible residual networks. In *Proceedings of the 36th International Conference on Machine*

Learning, volume 97 of *Proceedings of Machine Learning Research*, pp. 573–582, 2019. URL https://proceedings.mlr.press/v97/behrmann19a.html.

Yiyi Chen, Qiongkai Xu, and Johannes Bjerva. ALGEN: Few-shot inversion attacks on textual embeddings via cross-model alignment and generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24330–24348, 2025. doi: 10.18653/v1/2025.acl-long.1185. URL https://aclanthology.org/2025.acl-long.1185/.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations*, 2020.

Jieren Deng, Yijue Wang, Ji Li, Chenghong Wang, Chao Shang, Hang Liu, Sanguthevar Rajasekaran, and Caiwen Ding. TAG: Gradient attack on transformer-based language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 3600–3610, 2021. doi: 10.18653/v1/2021.findings-emnlp.305. URL https://aclanthology.org/2021.findings-emnlp.305/.

Laurent Dinh, David Krueger, and Yoshua Bengio. NICE: Non-linear independent components estimation. In *International Conference on Learning Representations Workshop*, 2015.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *International Conference on Learning Representations*, 2017.

Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting gradients—how easy is it to break privacy in federated learning? In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. Mask-Predict: Parallel decoding of conditional masked language models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 6112–6121, 2019. doi: 10.18653/v1/D19-1633. URL https://aclanthology.org/D19-1633/.

Aidan N. Gomez, Mengye Ren, Raquel Urtasun, and Roger B. Grosse. The reversible residual network: Backpropagation without storing activations. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Jiatao Gu, Changhan Wang, and Junbo Zhao. Levenshtein transformer. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Michael Günther, Saba Sturua, Mohammad Kalim Akram, Isabelle Mohr, Andrei Ungureanu, Bo Wang, Sedigheh Eslami, Scott Martens, Maximilian Werk, Nan Wang, and Han Xiao. jina-embeddings-v4: Universal embeddings for multimodal multilingual retrieval. In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pp. 531–550. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.mrl-main.36. URL https://aclanthology.org/2025.mrl-main.36/.

Chris Hokamp and Qun Liu. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1535–1546, 2017. doi: 10.18653/v1/P17-1141. URL https://aclanthology.org/P17-1141/.

Yu-Hsiang Huang, Yu-Che Tsai, Hsiang Hsiao, Hong-Yi Lin, and Shou-De Lin. Transferable embedding inversion attack: Uncovering privacy risks in text embeddings without model queries. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4193–4205, 2024. doi: 10.18653/v1/2024.acl-long.230. URL https://aclanthology.org/2024.acl-long.230/.

Jörn-Henrik Jacobsen, Arnold Smeulders, and Edouard Oyallon. i-RevNet: Deep invertible networks. In *International Conference on Learning Representations*, 2018.

Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1×1 convolutions. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. GeDi: Generative discriminator guided sequence generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 4929–4952, 2021. doi: 10.18653/v1/2021.findings-emnlp.424. URL https://aclanthology.org/2021.findings-emnlp.424/.

Haoran Li, Mingshi Xu, and Yangqiu Song. Sentence embedding leaks more information than you expect: Generative embedding inversion attack to recover the whole sentence. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 14022–14040, 2023. doi: 10.18653/v1/2023.findings-acl.881. URL https://aclanthology.org/2023.findings-acl.881/.

Weijun Li, Qiongkai Xu, and Mark Dras. Seeing the forest through the trees: Data leakage from partial transformer gradients. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4786–4798, 2024. doi: 10.18653/v1/2024.emnlp-main.275. URL https://aclanthology.org/2024.emnlp-main.275/.

Ximing Lu, Peter West, Rowan Zellers, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. NeuroLogic decoding: (un)supervised neural text generation with predicate logic constraints. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4288–4299, 2021. doi: 10.18653/v1/2021.naacl-main.339. URL https://aclanthology.org/2021.naacl-main.339/.

Ximing Lu, Sean Welleck, Peter West, Liwei Jiang, Jungo Kasai, Daniel Khashabi, Ronan Le Bras, Lianhui Qin, Youngjae Yu, Rowan Zellers, Noah A. Smith, and Yejin Choi. NeuroLogic A*esque decoding: Constrained text generation with lookahead heuristics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 780–799, 2022. doi: 10.18653/v1/2022.naacl-main.57. URL https://aclanthology.org/2022.naacl-main.57/.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Meta. Llama-2-7b-hf. Model card, 2023. URL https://huggingface.co/meta-llama/Llama-2-7b-hf.

Meta. Llama-3.1-8B. Model card, 2024a. URL https://huggingface.co/meta-llama/Llama-3.1-8B.

Meta. Llama-3.2-1B. Model card, 2024b. URL https://huggingface.co/meta-llama/Llama-3.2-1B.

Mistral AI. Ministral-3-8B-Base-2512. Model card, 2025. URL https://huggingface.co/mistralai/Ministral-3-8B-Base-2512.

Boris S. Mityagin. The zero set of a real analytic function. *Mathematical Notes*, 107:529–530, 2020. doi: 10.1134/S0001434620030189.

John X. Morris, Volodymyr Kuleshov, Vitaly Shmatikov, and Alexander M. Rush. Text embeddings reveal (almost) as much as text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12448–12460, 2023. doi: 10.18653/v1/2023.emnlp-main.765. URL https://aclanthology.org/2023.emnlp-main.765/.

Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. Generative representational instruction tuning. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/70cf215430492f7d34830a24e744b3f1-Abstract-Conference.html.

Giorgos Nikolaou, Tommaso Mencattini, Donato Crisostomi, Andrea Santilli, Yannis Panagakis, and Emanuele Rodolà. Language models are injective and hence invertible. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/5f124a1979570872ff46a6de8fbac35c-Abstract-Conference.html.

NVIDIA. Llama-Nemotron-Embed-1B-v2. Model card, 2025. URL https://huggingface.co/nvidia/llama-nemotron-embed-1b-v2.

NVIDIA. Nemotron-3-Embed-8B-BF16. Model card, 2026. URL https://huggingface.co/nvidia/Nemotron-3-Embed-8B-BF16.

Matt Post and David Vilar. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1314–1324, 2018. doi: 10.18653/v1/N18-1119. URL https://aclanthology.org/N18-1119/.

Lianhui Qin, Sean Welleck, Daniel Khashabi, and Yejin Choi. COLD decoding: Energy-based constrained text generation with Langevin dynamics. In *Advances in Neural Information Processing Systems*, volume 35, pp. 9538–9551, 2022.

Qwen Team. Qwen2.5-VL, January 2025. URL https://qwenlm.github.io/blog/qwen2.5-vl/.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 3982–3992, 2019. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

Congzheng Song and Ananth Raghunathan. Information leakage in embedding models. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, pp. 377–390, 2020. doi: 10.1145/3372297.3417270.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/65b9eea6e1cc6bb9f0cd2a47751a186f-Abstract-round2.html.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7534–7550. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.609. URL https://aclanthology.org/2020.emnlp-main.609/.

Zhipeng Wan, Anda Cheng, Yinggui Wang, and Lei Wang. Information leakage from embedding in large language models. *arXiv preprint arXiv:2405.11916*, 2024.

Yubo Wang, Shujie Cui, James Bailey, Hongzhi Yin, Wenyu Liang, Min Tang, Shiyue Qin, and Weiqing Wang. Denoising-aware inversion: Revealing privacy risks in noise-protected text embeddings. *arXiv preprint arXiv:2608.18610*, 2026.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019.

Weijia Xu and Marine Carpuat. EDITOR: An edit-based transformer with repositioning for neural machine translation with soft lexical constraints. *Transactions of the Association for Computational Linguistics*, 9:311–328, 2021. doi: 10.1162/tacl_a_00368. URL https://aclanthology.org/2021.tacl-1.19/.

Kevin Yang and Dan Klein. FUDGE: Controlled text generation with future discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3511–3535, 2021. doi: 10.18653/v1/2021.naacl-main.276. URL https://aclanthology.org/2021.naacl-main.276/.

Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank J. Reddi, and Sanjiv Kumar. Are transformers universal approximators of sequence-to-sequence functions? In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ByxRM0Ntvr.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Collin Zhang, John X. Morris, and Vitaly Shmatikov. Universal zero-shot embedding inversion. *arXiv preprint arXiv:2504.00147*, 2025.

Ruisi Zhang, Seira Hidano, and Farinaz Koushanfar. Text Revealer: Private text reconstruction via model inversion attacks against transformers. *arXiv preprint arXiv:2209.10505*, 2022.

Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

A Proof of Generic Spherical Injectivity

This appendix supplies the proof and architecture-specific witness underlying Section 3.2. All statements concern an exact-real parameter family. They do not certify a released checkpoint or its finite-precision implementation.

A.1 Definitions and Generic Result

Let $\mathcal{V}$ be a finite vocabulary, fix a length T, and let $\mathcal{X} \subseteq \mathcal{V}^T$ be finite. For the architectural constructions in this appendix, use the all-ones pooling mask $a_i = 1$ from Section 3.1. Thus

$$S_\theta(x) = \sum_{i=1}^{T} h_i(x;\theta) = T m_\theta(x), \qquad F_\theta(x) = \frac{S_\theta(x)}{\|S_\theta(x)\|_2}.$$

Normalizing the pooled sum and normalizing the pooled mean give identical observations. Equality of normalized observations means equality of positive rays, not identification of antipodal vectors.

Theorem A.1 (Generic spherical injectivity)**.** *Let $\Omega \subseteq \mathbb{R}^p$ be open and connected, and let $d \geq 2$. Suppose: (i) $S_\theta(x)$ is real analytic in θ for every $x \in \mathcal{X}$; (ii) every x admits a parameter setting $\theta \in \Omega$ with $S_\theta(x) \neq 0$; and (iii) every distinct pair x, x' admits a parameter setting $\theta \in \Omega$ where $S_\theta(x)$ and $S_\theta(x')$ are noncollinear. Then there exists a relatively open, dense subset $U \subseteq \Omega$, whose complement has Lebesgue measure zero, such that F_θ is defined and injective on $\mathcal{X}$ for every $\theta \in U$.*

Proof. For each unordered pair $\{x, x'\}$, assumption (iii) supplies coordinates $j < k$ for which the following analytic function is nonzero at some parameter setting:

$$D_{x,x'}(\theta) = S_\theta(x)_j S_\theta(x')_k - S_\theta(x)_k S_\theta(x')_j.$$

Consequently, $D_{x,x'}$ is not identically zero. A normalized collision implies positive collinearity and therefore $D_{x,x'}(\theta) = 0$. Likewise, $N_x(\theta) = \|S_\theta(x)\|_2^2$ is analytic and not identically zero by

assumption (ii). The zero set of a nontrivial real-analytic function on a connected open domain has Lebesgue measure zero (Mityagin, 2020). Define

$$Z = \bigcup_{x \in \mathcal{X}} \{\theta \in \Omega : N_x(\theta) = 0\} \cup \bigcup_{\{x,x'\} \subseteq \mathcal{X},\, x \neq x'} \{\theta \in \Omega : D_{x,x'}(\theta) = 0\}.$$

These are finite unions. Thus Z is relatively closed and Lebesgue-null. It has empty relative interior, so $U = \Omega \setminus Z$ is relatively open and dense. Outside Z, every pooled sum is nonzero and no normalized collision occurs. The set Z can include parameters that are nevertheless injective; it is a sufficient exceptional set, not an exact characterization of all failures. □

Scope of the pooling mask. The architectural witnesses below establish the theorem for $a_i = 1$. For a different fixed public mask a satisfying $\sum_i a_i > 0$, the same abstract criterion applies to $S^a_\theta(x) = \sum_i a_i h_i(x;\theta)$ only after its own analyticity, nonzero-witness and noncollinear-witness conditions have been verified. An arbitrary mask need not preserve sequence identity. For example, with structural causal attention, $T \geq 2$ and $a = (1, 0, \ldots, 0)$, the pooled state is $h_1(x;\theta)$, which depends only on x_1. Any two sequences sharing their first token but differing later therefore have identical normalized observations whenever these are defined. On a domain containing such a pair, the pair-separation premise fails; this does not contradict the conditional theorem.

Why analyticity alone is insufficient. The noncollinear-witness condition is sufficient, but not necessary, for spherical injectivity. It provides a convenient analytic certificate excluding positive-ray collisions. Analyticity alone does not prevent collisions imposed by symmetry. For example, a permutation-equivariant encoder without positional information, followed by permutation-invariant pooling, maps distinct permutations of the same token multiset to identical pooled vectors. The following construction establishes that this obstruction is not forced by the specified causal Llama architecture.

A.2 A Finite Llama Witness

For the architecture-specific statements below, θ collects the independent coordinates of token embeddings, attention and MLP projection weights and biases where present, RMSNorm gain vectors, and any adapter parameters explicitly included in the declared family. Tied parameters are counted only once. Architectural dimensions, attention masks, RoPE constants and positive RMSNorm stabilizers are fixed; the unrestricted family allows all these independent parameter coordinates to vary over $\mathbb{R}^p$.

Lemma A.2 (Pair-specific noncollinearity)**.** *Fix $T \geq 2$. Consider the exact-real, pre-norm Llama architecture with freely selectable embeddings and projection parameters, hidden width at least three, and at least one attention block containing a two-dimensional RoPE plane of frequency ω satisfying $0 < \omega(T-1) < \pi$. Assume positive fixed RMSNorm stabilizers and independently zeroable attention and MLP output projections. All RMSNorm gain vectors, including those of the input and final RMSNorm operations, are freely selectable and may be set to the all-ones vector for the witness. The observed hidden states $h_i(x;\theta)$ are the outputs of the final RMSNorm; no additional output transformation is applied before the stated pooling and normalization. Attention uses structural causal masking: query i attends only to keys $1, \ldots, i$. Evaluation mode is fixed with dropout disabled. Inputs are unpadded, positions are contiguous and unit-spaced, no pre-existing KV cache is present, and all positions are pooled. Then every distinct pair $x, x' \in \mathcal{V}^T$ admits a finite parameter setting at which $S_\theta(x)$ and $S_\theta(x')$ are noncollinear.*

Proof. Choose an admissible attention block indexed by $\ell_\star$. Set the attention and MLP output projections in every other block to zero, including any output biases. Each such block is exactly the identity: its normalization operations lie inside residual branches whose outputs are zero. The embeddings therefore reach block $\ell_\star$ unchanged. All query/key/value projections and input RMSNorm quantities below refer to this block, with ϵ_{in} its fixed positive stabilizer.

Step 1: A separating causal summary. Using one-based positions, define $k_r(i) = \max(1, i - r)$ for $r = 0, \ldots, T-1$. Let b_v denote a formal vocabulary basis vector and set $C_r(x) = \sum_i b_{x_{k_r(i)}}$.

Then

$$C_{T-1}(x) = Tb_{x_1}, \qquad C_{T-m}(x) = (T-m+1)b_{x_1} + \sum_{k=2}^{m} b_{x_k}, \quad 1 \le m \le T.$$

These identities recover the sequence inductively. Hence some offset r and vocabulary coordinate distinguish any pair $x \neq x'$. Choose that token $v_\star$ and set $c_v = \mathbf{1}\{v = v_\star\}$. The scalar summary $Q_r(x) = \sum_i c_{x_{k_r(i)}}$ then satisfies $\Delta Q = Q_r(x') - Q_r(x) \neq 0$, with $|\Delta Q| \ge 1$.

Step 2: Finite-softmax RoPE addressing. Let $\mathcal{K}_i = \{1, \dots, i\}$ be the allowed causal key set. Choose an unrotated query with phase zero and key with phase $r\omega$ in the selected RoPE plane. Their finite amplitudes can realize, after the usual head- dimension scaling, the scores

$$\begin{aligned} s_M(i,k) &= M\cos\big(\omega(i-k-r)\big), \\ \alpha_{ik}^M &= \frac{\exp(s_M(i,k))}{\sum_{\ell\in\mathcal{K}_i}\exp(s_M(i,\ell))}, \qquad k \in \mathcal{K}_i. \end{aligned} \tag{9}$$

Because $\omega(T-1) < \pi$, the score has a unique maximum at $k = k_r(i)$. When the requested offset extends beyond the left boundary, the first token is the unique nearest available key. Let $g > 0$ be the minimum winning cosine gap over queries with at least two allowed keys. Finiteness and uniqueness ensure positivity. Defining $u_i^M(x) = \sum_{k\in\mathcal{K}_i} \alpha_{ik}^M c_{x_k}$ and $Q_M(x) = \sum_i u_i^M(x)$ gives

$$\begin{aligned} \sum_{k\in\mathcal{K}_i\setminus\{k_r(i)\}} \alpha_{ik}^M &\le (T-1)e^{-Mg}, \\ |Q_M(x') - Q_M(x) - \Delta Q| &\le 2T(T-1)e^{-Mg}. \end{aligned} \tag{10}$$

Therefore any finite $M > g^{-1}\log\big(4T(T-1)/|\Delta Q|\big)$ preserves a nonzero summary difference. Infinite logits and literal hard attention are unnecessary.

Step 3: Realizing the summary through input RMSNorm. Reserve orthogonal hidden coordinates e_0, e_1, e_2 for an anchor, transported separator, and token marker. Using the lemma's gain assumption, set every RMSNorm gain vector to the all-ones vector at this witness. For anchor amplitude $a_0 > 0$ and embedding perturbation $\eta_{\mathrm{emb}} > 0$, choose

$$\begin{aligned} E(v) &= a_0 e_0 + \eta_{\mathrm{emb}} c_v e_2, \qquad & s_0 &= \sqrt{a_0^2/d + \epsilon_{\mathrm{in}}}, \\ s_v(\eta_{\mathrm{emb}}) &= \sqrt{(a_0^2 + \eta_{\mathrm{emb}}^2 c_v^2)/d + \epsilon_{\mathrm{in}}}. \end{aligned}$$

Let $q_\star, k_\star$ be the desired unrotated query and key vectors from Step 2, and let w be a unit coordinate in the selected value head. Set all unspecified projection components to zero and choose

$$\begin{aligned} &W_Q e_0 = (s_0/a_0)q_\star, \quad W_K e_0 = (s_0/a_0)k_\star, \\ &W_Q e_2 = W_K e_2 = 0, \quad W_V e_0 = 0, \qquad W_V e_2 = (s_0/\eta_{\mathrm{emb}})w. \end{aligned} \tag{11}$$

After input RMSNorm, the query and key acquire factors $s_0/s_v(\eta_{\mathrm{emb}}) = 1 + O(\eta_{\mathrm{emb}}^2)$, while the scalar value becomes $(s_0/s_v(\eta_{\mathrm{emb}}))c_v$. For the already fixed finite M, the resulting scalar attention outputs $u_i^{M,\eta_{\mathrm{emb}}}(x)$ converge to $u_i^M(x)$ as $\eta_{\mathrm{emb}} \downarrow 0$. Every chosen projection weight remains finite at each fixed positive η_{emb}.

Step 4: Preserving the residual path. In block $\ell_\star$, map the selected value coordinate to τe_1 through the attention output projection, annihilate all other head outputs, and set any attention output bias to zero. Set this block's MLP output projection and any output bias to zero. All subsequent blocks are already identities and leave the resulting state unchanged. Thus, immediately before the final RMSNorm,

$$z_i(x; \tau, \eta_{\mathrm{emb}}) = a_0 e_0 + \eta_{\mathrm{emb}} c_{x_i} e_2 + \tau u_i^{M,\eta_{\mathrm{emb}}}(x) e_1.$$

Step 5: Final RMSNorm preserves a nonzero determinant. With unit gains, write $R(z) = z/t(z)$, where $t(z) = \sqrt{\|z\|_2^2/d + \epsilon_{\mathrm{out}}}$. Its Jacobian is $J_R(z) = I/t(z) - zz^\top/(d\,t(z)^3)$. Define $t_v(\eta_{\mathrm{emb}}) = \sqrt{(a_0^2 + \eta_{\mathrm{emb}}^2 c_v^2)/d + \epsilon_{\mathrm{out}}}$. At $\tau = 0$, the separator direction is orthogonal to z_i.

Distinguishing the directional Jacobian from the derivative along the actual perturbation, the chain rule gives

$$J_R(z_i)e_1|_{\tau=0} = \frac{e_1}{t_{x_i}(\eta_{\mathrm{emb}})},$$
$$\left.\frac{\partial R(z_i(x;\tau,\eta_{\mathrm{emb}}))}{\partial \tau}\right|_{\tau=0} = \frac{u_i^{M,\eta_{\mathrm{emb}}}(x)}{t_{x_i}(\eta_{\mathrm{emb}})}e_1.$$

The anchor coordinate has zero first derivative with respect to τ. Accordingly, define

$$A_{\eta_{\mathrm{emb}}}(x) = a_0 \sum_i \frac{1}{t_{x_i}(\eta_{\mathrm{emb}})}, \qquad B_{\eta_{\mathrm{emb}}}(x) = \sum_i \frac{u_i^{M,\eta_{\mathrm{emb}}}(x)}{t_{x_i}(\eta_{\mathrm{emb}})}.$$

For the pooled determinant $D_{01}(\tau,\eta_{\mathrm{emb}}) = S_\theta(x)_0 S_\theta(x')_1 - S_\theta(x)_1 S_\theta(x')_0$, we have $D_{01}(0,\eta_{\mathrm{emb}}) = 0$ and

$$\partial_\tau D_{01}(0,\eta_{\mathrm{emb}}) = A_{\eta_{\mathrm{emb}}}(x)B_{\eta_{\mathrm{emb}}}(x') - B_{\eta_{\mathrm{emb}}}(x)A_{\eta_{\mathrm{emb}}}(x').$$

Consequently, with $t_0 = \sqrt{a_0^2/d + \epsilon_{\mathrm{out}}}$,

$$\lim_{\eta_{\mathrm{emb}}\downarrow 0} \partial_\tau D_{01}(0,\eta_{\mathrm{emb}}) = \frac{Ta_0}{t_0^2}\big[Q_M(x') - Q_M(x)\big] \neq 0. \tag{12}$$

Choose a sufficiently small finite $\eta_{\mathrm{emb}} > 0$ preserving the nonzero derivative in Equation 12. With η_{emb} fixed, analyticity in τ gives

$$D_{01}(\tau,\eta_{\mathrm{emb}}) = \tau\,\partial_\tau D_{01}(0,\eta_{\mathrm{emb}}) + O(\tau^2),$$

which is nonzero for sufficiently small finite nonzero τ. The choice order is $(r,c) \to M \to \eta_{\mathrm{emb}} \to \tau$. The limit calculations establish existence of finite choices; $\eta_{\mathrm{emb}} = 0$, where the prescribed value weight would diverge, is not a parameter setting used by the witness. A nonzero determinant proves noncollinearity. □

A.3 Llama-2 Specialization and Scope

For the specified Llama-2-7B configuration, the head dimension is 128 and the unscaled RoPE base is 10,000. Its lowest-frequency plane satisfies $\omega = 10000^{-126/128} \approx 1.15478 \times 10^{-4}$ and $\omega \cdot 4095 \approx 0.472883 < \pi$.

Corollary A.3 (Fixed-length Llama-2 family)**.** *For each fixed $2 \le T \le 4096$, under the hypotheses of Lemma A.2, normalized mean pooling is defined and injective on $\mathcal{V}^T$ throughout an open, dense, full-Lebesgue-measure subset of the unrestricted independent real-parameter space $\mathbb{R}^p$.*

Proof. Embedding lookup, affine projections, residual additions, pooling, SiLU, finite softmax, and RMSNorm with positive stabilizers are real analytic in their parameters. RoPE coefficients and allowed-key sets are fixed. A nonzero pooled witness is obtained by assigning every token the same nonzero anchor, disabling residual branches, and retaining nonzero final RMSNorm gains. Lemma A.2 supplies the pair-specific noncollinearity witnesses. Theorem A.1 therefore applies. For $T = 1$, pair-specific noncollinear token embeddings give the analogous result directly. □

Interpretation and exclusions. The witness is pair-dependent; it does not explicitly construct one parameter vector separating every pair. The generic theorem establishes simultaneous separation nonconstructively. Separate fixed-length results do not establish cross-length separation. Neither result supplies a quantitative separation margin, an efficient inversion algorithm, or a guarantee for our learned reconstruction procedure. Structural masking is a mathematical restriction of the softmax domain, not identification of a finite floating-point mask sentinel with negative infinity. Quantization, rounding, restricted adapter families, additional output projections, and native prompting or pooling conventions require separate analysis. Extensions to other architectures are not established merely by observing empirical separation.

Random parameters versus trained checkpoints. An absolutely continuous law on the final parameters assigns zero probability to the exceptional set. An absolutely continuous initialization alone does not establish this property after training: no preservation result for the actual optimization procedure is proved here. Accordingly, the corollary does not certify global injectivity of any named trained checkpoint.

A.4 Grouped-Query and Bidirectional Extensions

We extend the witness, rather than infer a theorem from finite-bank separation. Throughout, dropout is disabled, positional constants are fixed, all input positions are pooled, and the independent backbone parameters range freely. A fixed trained backbone with only its adapters variable is a different parameter family.

Lemma A.4 (Grouped-query attention)**.** *The witness of Lemma A.2 remains realizable with H query heads sharing G key/value groups, where G divides H, provided that each query projection, each shared key/value projection, and each head block of the output projection are freely selectable. Head outputs must be concatenated before this projection, without intervening cross-head mixing or normalization. Any projection biases must be independently zeroable.*

Proof. Write the attention output as

$$\mathrm{Attn}_\theta(z)_i = \sum_{h=1}^{H} O_h \sum_{k\in\mathcal{K}_i} \alpha_{ik}^{(h)} V_{g(h)} R_{\mathrm{in}}(z_k),$$

where $g(h)$ is the key/value group of head h, O_h is its output-projection block, and $\mathcal{K}_i$ is the allowed key set. Choose one head $h_\star$ and set $O_h = 0$ for $h \neq h_\star$. Use the selected query projection and the shared projections of group $g(h_\star)$ to realize Equations 9 and 11 of Lemma A.2, and use $O_{h_\star}$ for its transported separator. Although other heads read the same keys and values, each head has its own softmax; those heads cannot change the selected head's weights, and their output contributions vanish. The residual and final RMSNorm argument is therefore unchanged. All parameters remain finite. □

Lemma A.5 (Full bidirectional attention)**.** *Replace structural causal masking in Lemma A.2 by full attention over keys $1,\dots,T$. If a RoPE plane satisfies $0 < 2\omega(T-1) < \pi$, the same pair-specific noncollinearity conclusion holds, including with the grouped-query structure of Lemma A.4.*

Proof. For $r \in \{0,\dots,T-1\}$, use the scores $s_M(i,k) = M\cos(\omega(i-k-r))$ from Equation 9, now with $\mathcal{K}_i = \{1,\dots,T\}$. Now $-2(T-1) \le i-k-r \le T-1$, so every phase has absolute value below π. Cosine is strictly decreasing in absolute phase on this interval. Since $i-r \le T$, the unique nearest allowed integer to $i-r$ is again $k_r(i) = \max(1, i-r)$; an offset beyond the left boundary uniquely selects key 1. Thus the hard-address summaries are exactly the separating summaries in Step 1 of Lemma A.2.

For the pair's chosen offset, let $g > 0$ be the minimum winning cosine gap over the finite query/key set. The finite-softmax bounds in Equation 10 are unchanged:

$$\sum_{k\in\mathcal{K}_i\setminus\{k_r(i)\}} \alpha_{ik}^{M} \le (T-1)e^{-Mg},$$
$$|Q_M(x') - Q_M(x) - \Delta Q| \le 2T(T-1)e^{-Mg}.$$

Choose finite M to make the second bound smaller than $|\Delta Q|/2$, then choose finite $\eta_{\mathrm{emb}} > 0$ and $\tau \neq 0$ in that order, as in Steps 3–5 of the causal witness. The input RMSNorm perturbation converges at fixed M, and the final pooled determinant retains a nonzero derivative. This gives a finite noncollinear witness, without hard attention or a causal mask. □

Scaled positional frequencies. The proofs require an admissible plane, not the original Llama-2 RoPE base. Fixed positive rescalings of a plane's frequency can be used in the displayed bound. A fixed nonzero amplitude multiplying its sine and cosine can be absorbed into the query/key amplitudes. This does not cover arbitrary position-dependent score changes; such changes must be checked separately. Let b denote the RoPE base. For the six configurations here, $b \ge 10^4$ and the head dimension d_h is either 64 or 128. Their fixed frequency rescalings leave the lowest frequency positive and no larger than its unscaled value. Thus that plane satisfies

$$0 < \omega_{\min} \le b^{-(d_h-2)/d_h} \le 10^{-31/8}, \qquad 2\omega_{\min}(T-1) < 0.009 < \pi \quad (T \le 32).$$

This verifies the needed short-window bound independently of the numerical addressing tests below.

Family	d	H	G	d_h	RoPE base	Mask	U_m
Llama-2	4096	32	32	128	10^4	C	$[3, 31999]$
Llama-3.1	4096	32	8	128	$5 \cdot 10^5$	C	$[0, 127999]$
Nemotron-3	4096	32	8	128	10^6	F	$[1000, 131071]$
GritLM	4096	32	8	128	10^4	F	$[3, 31999]$
Jina-v4	2048	16	2	128	10^6	C	$[0, 151642]$
Nemotron-1B	2048	32	8	64	$5 \cdot 10^5$	F	$[0, 127999]$

Table 4: Architecture and finite domains for Corollary A.6. Here d is hidden width, H the query-head count, G the key/value-group count, and d_h the head dimension. C and F denote structural causal and full attention; every interval denotes all integers between its inclusive endpoints. Fixed frequency-scaling rules and observation contracts are specified in Appendix A.5. GritLM's causal control also satisfies Lemma A.4.

Corollary A.6 (Short-window raw-text families)**.** *Fix $T \in \{8, 16, 32\}$ and one of the raw text-only observation contracts specified in Appendix A.5. Take its exact-real equations, structural attention mask, fixed positional constants, and full independent parameter space $\mathbb{R}^p$, including the normalization gains used in the witness. Use the fixed token set $\mathcal{V}_{\text{text}} = U_m$ in Table 4, positions $p_i = i-1$, and the text-only input conditions stated below. Every sequence in U_m^T then follows the same specified computation. Normalized mean pooling is defined and injective on $\mathcal{V}_{\text{text}}^T$ on a relatively open, dense, full-measure subset of that parameter space.*

Proof. The architecture correspondence below verifies that each specified text backbone has pre-norm residual attention and SiLU-gated MLP blocks, positive RMSNorm stabilizers, freely selectable query/key/value/output projections, and final RMSNorm. These operations are analytic in the independent parameters. The fixed RoPE configurations admit a plane satisfying the stronger full-mask bound at each stated length. Lemmas A.4 and A.5, or the original causal witness as appropriate, supply the pair-specific determinant. Setting every token embedding to the same nonzero anchor and disabling residual branches supplies the nonzero pooled witness.

Two qualifications are needed for the model-specific paths. Nemotron-3's additional query multiplier at position p_i is $1 + \beta \log(1 + \lfloor p_i / L_0 \rfloor)$, with $\beta = 0.1$ and $L_0 = 16384$. Our explicit positions satisfy $0 \le p_i = i - 1 \le T - 1 < L_0$. The multiplier therefore equals one for these windows, so it does not alter the witness scores. For Jina-v4, fix the text-matching task and exclude image/video inputs and their control tokens. The three text-only positional coordinate streams coincide, reducing its multi-axis RoPE to ordinary one-dimensional RoPE. Its task-specific linear adapters have the form $W + sBA$ with fixed scale s; choosing $B = 0$ realizes the unrestricted backbone witness. Their additional parameters enter analytically. The single-vector text path mean-pools the final states; the separate multi-vector projector is not part of this observation. Tied language-model output weights do not constrain the witness, because the output vocabulary head is not used to form these states. Theorem A.1 now applies in each declared family. This argument does not apply to adapter-only variation around a fixed pretrained backbone. □

Native interfaces remain a separate claim. Public prompts, tokenization, selective pooling, output projections and dimensional truncation can change the map. In particular, a fixed projection can annihilate the witness's separator; raw hidden-state noncollinearity then says nothing about the exposed vector. The generic theorem could cover a native map only after its own analyticity, nonzero and pair-separation premises are established. Native-wrapper measurements remain empirical evidence, not a proof of those premises. The results above also provide neither cross-length injectivity nor a checkpoint-specific separation or recovery guarantee.

A.5 Architecture-to-Lemma Correspondence

Fixed input and parameter contract. The named families refer to the exact-real text architectures and observation contracts specified here and in Table 4, not every implementation bearing the same model name. Inputs are token IDs directly, with no decoding/re-encoding, prompt, padding, truncation, image/video tensors, spatial grids, or pre-existing cache. Both the attention padding mask and pooling mask are all ones. Positions are $p_i = i - 1$; for Jina-v4 this holds on all three

positional axes, and the task is fixed to *text-matching*. Dropout is disabled. These choices fix the computation for every sequence in U_m^T, not merely for natural-language examples.

The parameter vector consists of independent text embeddings, attention and MLP projections, their biases where present, and all input, post-attention and final RMSNorm gains. Jina's active task-adapter matrices are also included. Dimensions, positive normalization stabilizers, positional constants and masks are fixed. Gain vectors may be set to one; projection weights and biases may be zeroed. Sharing is limited to the stated GQA groups, the explicitly described adapters, and any input/output vocabulary tying. A tied output vocabulary head is unused by the pooled map. Unused vision or output-projector parameters can be fixed or added as independent coordinates without changing the argument.

Operational token domains. The intervals in Table 4 are fixed whitelists checked against the corresponding tokenizers and embedding-row counts. Llama-2 and GritLM exclude IDs 0–2; Llama-3.1 and Nemotron-1B exclude 128000–128255; Nemotron-3 excludes 0–999. Jina excludes the special/control-token range 151643–151664 and the unmapped embedding rows 151665–151935. In particular, its vision-start, vision-end, image and video markers are excluded. Without media/grid inputs, its positional branch is text-only; without the vision-start ID, its single-vector pooling branch always averages the text states. Thus membership in these fixed sets suffices for the entire Cartesian product, including arbitrary token combinations. This is the corollary's domain restriction, not a new filter applied to experimental cohorts.

Common lemma requirements and output location. For every row, attention has independently projected queries, grouped keys/values, a separate softmax per query head, concatenated head outputs, and a freely selectable output projection. There is no intervening cross-head normalization. Blocks apply RMSNorm before attention and before the SiLU-gated MLP, with a residual addition after each branch. The final hidden tensor is taken *after final RMSNorm*, and before any vocabulary head, embedding projector or dimensional truncation. Its all-position mean is normalized as in Equation 1. The RMSNorm stabilizer is 10^{-6} for Jina and 10^{-5} for the other five configurations. These facts supply the residual, analyticity, head-isolation and output assumptions of Lemmas A.2–A.5. The remaining model-specific conditions are established below.

Llama-2 and Llama-3.1: causal attention and scaled RoPE. Both families use structural causal attention, with the head dimensions and key/value groups in Table 4. Llama-2 uses unscaled frequencies b^{-2j/d_h} with $b = 10000$. Llama-3.1 uses fixed frequency scaling with factor 8, low/high frequency factors $1, 4$, and original context 8192. Its lowest frequency is divided by 8; intermediate frequencies interpolate between their unscaled values and those values divided by 8. Thus the required positive low-frequency plane is preserved. The causal witness applies directly to Llama-2 and through Lemma A.4 to Llama-3.1.

Nemotron-3: full attention and positional query scaling. The declared text architecture uses full bidirectional attention, without a sliding-window restriction, as required by Lemma A.5. Its fixed YaRN scaling has factor 16, original context 16384, and ramp parameters $32, 1$. Each frequency is a fixed convex combination of its unscaled value and that value divided by 16; the rotary amplitude is one. The additional query multiplier at position p_i is $1 + 0.1 \log(1 + \lfloor p_i/16384 \rfloor)$. Under $p_i = i - 1$ and $T \leq 32$ it equals one, leaving the witness scores unchanged. Short length alone, with an arbitrary absolute offset, would not justify this step.

GritLM: bidirectional and causal families. The declared bidirectional text mode uses full attention over the unpadded window and unscaled RoPE. No instruction prefix is added, and all post-final-RMSNorm token states are pooled. It therefore satisfies the GQA and full-mask extensions. The separate causal control restricts each query to preceding and current positions while retaining the same head structure and positional rule; it satisfies the causal/GQA witness instead.

Jina-v4: fixed text branch and task adapters. The text backbone uses causal attention. With no media grids and all-one masks, its three positional axes are identical and equal to $p_i = i - 1$. The configured multi-axis sections $(16, 24, 24)$ therefore reduce to unscaled one-dimensional RoPE on this domain. For the fixed text task, each active linear adapter has effective map $W + sBA$, where

Check	Observed result
Causal addressing, $T = 8, 16, 32$	Zero failures
RMSNorm Jacobian maximum error	4.44×10^{-16}
Finite witnesses, 120 sequence pairs	All determinants nonzero
Minimum absolute witness determinant	0.0318175

Table 5: Numerical consistency checks supporting the finite witness construction. These are not global injectivity tests for a released model.

s is fixed. Setting each active B to zero leaves freely selectable base projections, and the additional adapter parameters enter analytically. Query/key/value biases may also be set to zero.

The pooled tensor consists of the final text states after the in-block task adapters and final RMSNorm. The single-vector output mean-pools these states on the declared text-only domain, whose exclusion of the vision-start token prevents image-specific pooling. The separate multi-vector projector does not feed back into this tensor. These conditions establish the causal/GQA correspondence for this raw-token map, not for arbitrary native multimodal inputs.

Nemotron-1B: full-attention Llama family. This text architecture retains the Llama pre-norm residual structure while replacing the causal mask with full bidirectional attention. The fixed Llama-3 frequency rule has factor 32, low/high factors $1, 4$, and original context 8192; the lowest frequency is divided by 32. The GQA/full-mask witness therefore applies. Pooling uses the post-final-RMSNorm states, with no native prompt or additional output projection.

These correspondences concern exact-real parameter families. They do not assert equivalence to finite-precision execution or certify injectivity of the released checkpoints.

A.6 Numerical Consistency Checks of the Witness

We supplement the analytical construction with numerical checks of its three principal mechanisms: causal offset addressing, preservation of the separator through RMSNorm, and the existence of finite pair-specific parameters. These checks test consistency with the proof; they do not establish its universal quantifiers or certify a trained checkpoint.

First, exhaustive checks of the causal offset addresses at $T \in \{8, 16, 32\}$ found no disagreement between the RoPE score maximizer and the required source $k_r(i) = \max(1, i - r)$. Second, we evaluate the RMSNorm Jacobian at $z = 1.7e_0$ with width 16, stabilizer 10^{-5} and gains $g_j = 0.75 + 0.5j/15$, $j = 0, \ldots, 15$. The analytical RMSNorm Jacobian agreed with automatic differentiation to a maximum absolute error of 4.44×10^{-16}. The separator response was nonzero, while its first-order effect on the anchor coordinate was zero, as required by the determinant argument.

Third, we evaluated the finite construction on all $\binom{16}{2} = 120$ distinct pairs of binary length-four sequences. The reduced hidden width is $d = 8$, the anchor amplitude is $a_0 = 1$, all RMSNorm gains are one, and both normalization stabilizers are 10^{-5}. We use the binary token code $c_v = v$. For each pair, choose the smallest offset $r \in \{0, 1, 2, 3\}$ for which $\sum_i c_{x_{\max(1, i-r)}}$ differs between the sequences. Among the Llama-2 RoPE frequencies $\omega_j = 10000^{-2j/128}$, $j = 0, \ldots, 63$, satisfying $0 < \omega_j(T - 1) < \pi$, choose the frequency maximizing the minimum winning cosine gap g over all offsets and queries with competing allowed keys. This selects $\omega = 1$ for $T = 4$. Set $M = (\log(4T(T - 1)) + 2)/g$. Using the resulting $M \approx 12.7719$, $\eta_{\text{emb}} = 10^{-4}$ and $\tau = 10^{-3}$, every pair produced a nonzero determinant; the minimum absolute determinant was approximately 0.0318175. The determinant uses coordinates zero and one of the pooled sums S_θ, before final vector normalization. Using pooled means instead would divide the determinant by T^2. Each pair uses its own distinguishing offset; the binary token code is shared across pairs. This therefore checks pair-specific witnesses, not one explicit parameter setting separating all pairs simultaneously.

Cross-model consistency checks. We repeated these checks for all six families in Table 4. Using their configured head/group dimensions and positional frequencies, the grouped-query head-suppression check had zero numerical discrepancy. Causal and full-mask addressing both matched the required offsets at $T = 8, 16, 32$. For each addressing check, select the largest configured positive frequency satisfying $2\omega(T - 1) < \pi$, so the same plane is admissible for both masks.

Family	Selected ω	$\epsilon_{\text{in}} = \epsilon_{\text{out}}$
Llama-2, GritLM	0.48696753	10^{-5}
Llama-3.1, Llama-Nemotron-1B	0.44036663	10^{-5}
Nemotron-3	0.52329916	10^{-5}
Jina-v4	0.52329916	10^{-6}

Table 6: Parameters for the cross-model reduced-width $T = 4$ witness checks. Frequencies are rounded for display. These are constructed witnesses, not forward passes through pretrained models.

For the finite pair checks, retain the reduced width $d = 8$, anchor $a_0 = 1$, unit gains, binary token code, smallest separating offset, $\eta_{\text{emb}} = 10^{-4}$ and $\tau = 10^{-3}$ specified above. Choose the frequency by the same rule at $T = 4$; Table 6 lists the selected frequencies. Both RMSNorm stabilizers are 10^{-5} except for Jina-v4, where both are 10^{-6}. For each mask, let g be the minimum winning cosine gap over all offsets and queries with competing allowed keys, and set $M = (\log(4T(T-1)) + 1)/g$. This finite-score schedule differs from the $+2$ schedule in the original causal check. For each configuration and each mask, all 120 pairs among the 16 binary length-four sequences had nonzero pair-specific pooled determinants. The minimum absolute determinant across configurations was approximately 0.0315107 for causal attention and 0.0315144 for full attention. These reduced-width witness checks do not run the pretrained networks and do not establish the universal claim; that claim follows from the lemmas and the stated architecture correspondence. Finite-bank separation of trained models is a different measurement, discussed in Appendix B.

A.7 Finite-Domain Permutation Control

The architectural distinction matters because pooling can erase order when its inputs contain no contextual positional information. To illustrate this distinction, we evaluated all $4^4 = 256$ length-four sequences over a four-token vocabulary using float64 arithmetic and three random initializations.

Normalized pooling of token embeddings alone produced 1,230 collision pairs in every initialization. A one-block causal Llama model produced no observed collision pairs under the same numerical collision criterion, $1 - \cos(u, v) \leq 10^{-10}$. Its minimum pairwise chordal distances were approximately 0.05257, 0.04610 and 0.05443 across the three initializations.

The embedding-only collisions reflect permutation invariance: reordering tokens leaves their embedding sum unchanged. In contrast, contextual hidden states depend on token order, so pooling those states need not preserve this invariance. This finite-domain control illustrates the architectural distinction used by the witness argument. It neither proves global injectivity of a pretrained model nor establishes that causal attention is necessary for order preservation.

B Numerical Separation and Computational Recovery

This appendix distinguishes properties of the exact-real map from numerical measurements and reconstruction outcomes. The analyses below reuse frozen observations, checkpoints and candidate archives; they do not constitute new reconstruction evaluations or threshold selection.

B.1 A Conditional Finite-Set Separation Bound

Fix a checkpoint and observation contract, and abbreviate its exact-real normalized map by F. Let $\mathcal{C}$ be a finite set containing $x^\star$ and at least one wrong sequence, with F defined on every member. Define the unsquared separation

$$\Delta_{\mathcal{C}}(x^\star) = \min_{c \in \mathcal{C} \setminus \{x^\star\}} \|F(c) - F(x^\star)\|_2. \tag{13}$$

Suppose the numerical vectors after normalization satisfy

$$\|\widetilde{y} - F(x^\star)\|_2 \leq \varepsilon_y, \qquad \sup_{c \in \mathcal{C}} \|\widetilde{F}(c) - F(c)\|_2 \leq \varepsilon_c,$$

and write $\eta = \varepsilon_y + \varepsilon_c$. The triangle inequality gives

$$\|\widetilde{F}(x^\star) - \widetilde{y}\|_2 \le \eta,$$
$$\|\widetilde{F}(c) - \widetilde{y}\|_2 \ge \Delta_{\mathcal{C}}(x^\star) - \eta \quad (c \ne x^\star).$$

Consequently, $\Delta_{\mathcal{C}}(x^\star) > 2\eta$ is sufficient for unique nearest-vector recovery within $\mathcal{C}$. This is a standard triangle-inequality consequence, not an additional architectural theorem. It assumes exact distance comparisons between the represented numerical vectors. For selection using a computed score $\widehat{E}$, assume

$$\sup_{c\in\mathcal{C}} \left|\widehat{E}(c) - E_{\mathrm{vec}}(c)\right| \le \varepsilon_E.$$

This bound concerns the score relative to the fixed represented vectors, not their error relative to F. It includes score-evaluation arithmetic and any discrepancy between the implemented formula and squared coordinate distance. In particular, the dot-product score requires accounting for normalization defects. Comparing a separately renormalized diagnostic with the same E_{vec} also requires accounting for that change in vectors (Appendix B.2). Under $\Delta_{\mathcal{C}} > 2\eta$, a sufficient score-gap condition is $(\Delta_{\mathcal{C}} - \eta)^2 - \eta^2 > 2\varepsilon_E$.

Under these conditions, a computed-score threshold accepts the target and rejects every wrong member of $\mathcal{C}$ whenever

$$\eta^2 + \varepsilon_E \le \epsilon < (\Delta_{\mathcal{C}}(x^\star) - \eta)^2 - \varepsilon_E.$$

Indeed, the target's computed score is at most the left endpoint, whereas every wrong score is at least the right endpoint. The score-gap condition makes this interval nonempty. When $\varepsilon_E = 0$, it reduces to $\eta \le \sqrt{\epsilon} < \Delta_{\mathcal{C}}(x^\star) - \eta$; separation and representation errors are unsquared distances, while ϵ bounds a squared score. The guarantee is conditional on the candidate set and the error bounds; it does not exclude unseen collisions outside that set. Our experiments do not establish uniform bounds relative to the exact-real map. Agreement between repeated forwards can coexist with a shared numerical bias. Likewise, the minimum distance in a sampled bank is not a lower bound on the separation over all token sequences.

B.2 Observation Variation and Threshold Errors

Precision, batching and hardware are part of the numerical observation contract. In a paired Jina-v4 development check on 48 raw eight-token windows, we held token IDs, checkpoint and task adapter fixed and compared A100 and H200 execution at batch settings $1, 8, 32, 64$, with two repeated forwards. The maximum same-text cross-device distance over these settings was 0.02930 in BF16 and 7.235×10^{-6} in FP32. These are observed unsquared distances, not worst-case error bounds. Improved FP32 agreement does not justify replacing the original stored target with a recomputed FP32 observation. Nor does this hardware comparison isolate every contribution to the historical stored-versus-recomputed discrepancy.

For the threshold diagnostic, each victim has 1,000 previously inspected raw eight-token targets. Wrong candidates comprise sampled token/span/order perturbations and candidates drawn from existing reconstruction archives; the neighborhoods are not exhaustive. All candidates and the gold sequence are re-embedded and compared against the same stored target. The reconstruction runs normalize their input observations in FP32, yielding numerical vectors $u = \widetilde{F}(c)$ and $v = \widetilde{y}$, and compute

$$\widehat{E}_{\mathrm{dot}}(c) = \mathrm{fl}_{32}\big(2 - 2u^\top v\big).$$

Here fl_p denotes the indicated floating-point computation, not an exact evaluation followed by a single rounding. Even before score-evaluation rounding,

$$(2 - 2u^\top v) - E_{\mathrm{vec}}(c) = 2 - \|u\|_2^2 - \|v\|_2^2.$$

Thus, if the score arithmetic differs from $2 - 2u^\top v$ by at most $\varepsilon_{\mathrm{arith}}$, then

$$\left|\widehat{E}_{\mathrm{dot}}(c) - E_{\mathrm{vec}}(c)\right| \le \left|1 - \|u\|_2^2\right| + \left|1 - \|v\|_2^2\right| + \varepsilon_{\mathrm{arith}}.$$

The error budget must therefore include normalization defects, not just dot-product and subtraction rounding.

Victim	Wrong hits / tested	Affected rows	Gold misses
Llama-3.1-8B	151/32,100	54	0
Nemotron-3-Embed-8B	0/31,997	0	0
GritLM-7B	22/31,963	7	0
Jina-Embeddings-v4	1,093/32,002	291	1
Llama-Nemotron-Embed-1B-v2	0/32,341	0	0

Table 7: Diagnostic behavior of the unchanged squared-energy threshold $\epsilon = 10^{-3}$. Wrong hits are known-wrong candidates with $\widehat{E}_{\mathrm{coord}} \le \epsilon$; affected rows contain at least one such candidate. Gold misses are rows whose re-embedded gold exceeds the threshold. Each row-count column has denominator 1,000. These are not false-stop rates or global collision counts.

The diagnostic independently normalizes its input observations in FP64; denote the resulting numerical vectors by $\bar{u}, \bar{v}$. It computes

$$\widehat{E}_{\mathrm{coord}}(c) = \mathrm{fl}_{64}\big(\|\bar{u} - \bar{v}\|_2^2\big)\,.$$

Coordinate subtraction reduces cancellation near coincident directions, but this remains a computed score, not an exact distance. Its separately normalized vectors need not equal u, v, so it must not be identified with $E_{\mathrm{vec}}(c)$ or the reconstruction score.

No sampled wrong candidate in this diagnostic outranked the re-embedded gold against the same stored target. Thus false threshold acceptance is distinct from a nearest-candidate ranking reversal. This recomputation does not replace the original finite-precision reconstruction scores. Separately, the saved selected outputs contain threshold-positive but token-incorrect cases: two for the Llama-3.1 T=8 Wikipedia benchmark, one for Jina-v4 T=8 WikiText, two for Llama-3.1 T=16 AG-News, and one for GritLM T=16 AG-News, each from a 1,000-row cohort. These saved-output flags use $\widehat{E}_{\mathrm{dot}}$, not the coordinate diagnostic.

The frozen multi-model evaluator records the threshold flag only after its fixed correction loop and final archive selection. It does not use that flag to terminate the loop. Accordingly, we report lexical exact match separately from threshold hits and do not infer early-stopping savings or exact-text certification from this flag. This control-flow statement is restricted to the audited evaluator, not every historical search variant. No threshold was retuned using these inspected outcomes.

B.3 Candidate Construction Versus Selection

Let $\mathcal{C}$ denote the complete, nonempty set available to final selection and let $\widehat{x} \in \mathcal{C}$ be the returned sequence. For equal-length candidates, define

$$A_{\mathrm{oracle}}(\mathcal{C}, x^\star) = \max_{c \in \mathcal{C}} \frac{1}{T} \sum_{i=1}^{T} \mathbf{1}\{c_i = x_i^\star\}.$$

Selected token accuracy cannot exceed this oracle. Exact-match failure partitions into the disjoint events

$$\{\widehat{x} \neq x^\star\} = \{x^\star \notin \mathcal{C}\} \,\dot\cup\, \{x^\star \in \mathcal{C},\ \widehat{x} \neq x^\star\}.$$

The distinction must be measured on complete sequences: finding the correct token at every position across different candidates does not establish that any candidate equals $x^\star$.

In the five audited T=8 archives, target-absent counts were 229, 56, 31, 28 and 544 out of 1,000 for Llama-3.1, Nemotron-3, GritLM, Jina-v4 and Llama-Nemotron-1B, respectively. GritLM additionally had two target-present selection failures; the other four archives had none. These are reanalyses of frozen outputs, not new generalization results. An absence of target-present exact-match errors also does not establish good partial selection: candidate-oracle token accuracy can still exceed selected token accuracy substantially when the target is absent. Neither these archive counts nor finite-bank separation establish efficient inversion, global injectivity of a checkpoint, or an information-theoretic transition as sequence length increases.

Victim	Overlaps		Retained	Full cohort (%)		Overlap-excluded (%)	
	Target	Input-only	rows	Token	Exact	Token	Exact
Llama-2	0	2	1,000	99.925	99.700	99.925	99.700
Llama-3.1	0	0	1,000	92.713	77.100	92.713	77.100
Nemotron-3	12	1	988	98.263	94.400	98.241	94.332
GritLM	19	2	981	98.888	96.700	98.866	96.636
Jina-v4	16	1	984	99.213	97.200	99.200	97.154
Llama-Nemotron-1B	7	0	993	73.650	45.600	73.464	45.217

Table 8: Overlap sensitivity on the six main T=8 cohorts. Target counts evaluation windows matching training targets. Input-only counts additional windows found as cached corrector inputs but not as training targets. Full-cohort results use all evaluation rows; retained rows and overlap-excluded results remove target overlaps only.

C Training-Data Overlap and Sensitivity

Table 8 compares evaluation targets with the configured attacker-training target pools using exact equality of fixed-length token-ID windows. We recompute token accuracy and exact match from the frozen predictions after excluding matching evaluation windows; no retraining or new reconstruction is performed. Each full evaluation cohort contains 1,000 eight-token windows.

Input-only matches indicate potential exposure, not duplicate target labels; cache membership does not establish that the optimizer sampled those entries. Also excluding these additional matches leaves Nemotron-3 and Jina-v4 exact match at 94.326% and 97.152%, respectively. These are sensitivity analyses of existing outputs, not new output-blind evaluations. The audit concerns configured attacker data, not inaccessible victim pretraining data, and exact-window disjointness does not establish article-level or source-level independence.

D Reconstruction Architecture, Training and Execution

This appendix specifies the native Nemotron-3 $T = 8$ pipeline and the six-model raw-pooling controls in Section 4. It does not describe the separate corrector-free proposals, closure procedures or graph-scheduler ablations. The observation and access contract remains that of Section 3.1; the victim is never fine-tuned by the reconstruction procedure.

D.1 Frozen Generators and Continuous Conditioning

Native passage observation. For Nemotron-3, tokenize `passage:` followed by the target fragment using the native recipe with special tokens enabled (NVIDIA, 2026). The fixed public context contains the prefix tokens and leading special tokens; a token crossing the prefix's trailing-space boundary belongs to the unknown suffix. The fresh native evaluation requires every target to pass a decode–re-encode check reproducing exactly $w_{\mathrm{pub}} \| x$. A parity check independently reproduces the normalized mean over all $L+T$ final hidden states. Positions start at zero on this concatenated input, with no previous key/value cache. The reconstructor emits only the T unknown IDs. It may query arbitrary token-ID candidates with the same fixed public context; intermediate candidates are not required to round-trip through a text-only interface. Returned-sequence round-trip validity is therefore reported separately from token accuracy. This is local reconstruction from native observations, not a claim of embedding-only API access. Native and raw Nemotron-3 use separate training banks, hypothesis caches and learned conditioning parameters under the same victim–generator pairing.

Victim–generator pairing. The base and corrector use the same causal generator checkpoint, with separate conditioning parameters. A generator need not be the victim: it must provide compatible token-ID embeddings and output logits. Table 9 specifies the pairings. Compatibility checks compare the token-to-ID vocabulary mappings, not merely their sizes. This does not assert equivalence between native text wrappers. Multimodal checkpoints are used through their text-generation path, without image inputs. All generator weights, including token embeddings and the output head, remain frozen; these reference runs add no trainable low-rank updates to the generator.

Victim	Causal generator	d	Start ID
Llama-2-7B	Llama-2-7B	4096	1
Llama-3.1-8B	Llama-3.1-8B	4096	128000
Nemotron-3-Embed-8B	Ministral-3-8B-Base-2512	4096	1
GritLM-7B	GritLM-7B, causal mode	4096	1
Jina-Embeddings-v4	Qwen2.5-VL-3B-Instruct	2048	1
Llama-Nemotron-Embed-1B-v2	Llama-3.2-1B	2048	128000

Table 9: Reference reconstruction backbones. Here d is the observed vector dimension and Start ID is the reconstructor's stored decoder-start token, not a token added to the victim window. In particular, the Jina generator uses the stored fallback ID 1, not a claimed native beginning-of-sequence token.

The Llama-2, Llama-3.1 and Llama-3.2 releases are documented by Meta (2023; 2024a;b). GritLM and Jina-v4 are described by Muennighoff et al. (2025) and Günther et al. (2025), respectively. The exact Nemotron embedding checkpoints have separate model cards (NVIDIA, 2026; 2025). The additional generator sources are the Ministral-3-8B-Base-2512 release (Mistral AI, 2025) and Qwen2.5-VL (Qwen Team, 2025).

Projectors. Let d_g denote the generator's embedding width. Each vector feature $f \in \mathbb{R}^d$ has an independent input LayerNorm and a two-layer projector with dimensions $d \to 1024 \to 4d_g$, GELU activation and dropout 0.05 between the linear layers. Its output is reshaped into four d_g-dimensional conditioning tokens. A learned feature-and-slot embedding is added to each token, followed by a shared output LayerNorm. The base therefore supplies four conditioning tokens; the corrector supplies sixteen. These projectors and their normalization and slot parameters are trainable, and are not shared between base and corrector.

For the base, the feature is $\mathcal{N}(y)$. For correction, the features are $[\mathcal{N}(y), \mathcal{N}(z), r, r]$, with $r = \mathcal{N}(\mathcal{N}(y) - \mathcal{N}(z))$ and the normalization stabilizer $\varepsilon_{\mathrm{n}} = 10^{-12}$. The complete current hypothesis follows the sixteen conditioning tokens. During teacher forcing, both models then receive a decoder-start token and the shifted target sequence $(x_1^\star, \ldots, x_{T-1}^\star)$. Only the resulting T target-token logits contribute to the loss. At inference, generated tokens replace this teacher-forced suffix. Decoding always emits exactly T tokens; an emitted end token does not terminate the loop. Continuous conditioning, hypothesis tokens and the decoder-start token are reconstructor inputs only, not modifications of the victim observation.

D.2 Prefix Features and Twist Architecture

For a nonempty proposed prefix u of length $t < T$, the raw condition activates only those t positions. The native condition also activates the L public-context positions and includes them in $s = s_t(u)$. In both cases the unknown suffix is absent. Padding keys are excluded from attention, and padded hidden states are excluded from pooling. Define $\bar{y} = \mathcal{N}(y)$ and

$$\begin{aligned} \rho &= \max(\|s\|_2, 10^{-8}), \qquad \alpha = s^\top \bar{y}, \qquad w = s - \alpha \bar{y}, \\ f_{\mathrm{vec}} &= \left[\bar{y}, \frac{s}{\rho}, \frac{w}{\max(\|w\|_2, 10^{-8})} \right], \\ f_{\mathrm{scal}} &= \left[\log \rho, \frac{\alpha}{\rho}, \frac{\|w\|_2}{\rho}, \frac{t}{T}, \frac{T-t}{T} \right]. \end{aligned} \tag{14}$$

The prefix norm is observable from the reconstructor's own query; the target's unnormalized norm is neither observed nor inferred as an input. For bidirectional attention, $s_t(u)$ is recomputed on the available context and prefix, not treated as a fixed summand of a later full sequence. Twist's token encoder receives only u; its relative-depth features remain t/T and $(T-t)/T$, excluding the public context.

Twist has its own learned token embeddings of dimension 256, projected to width 768 and combined with learned positional embeddings. A four-layer, eight-head pre-norm Transformer encoder uses feed-forward width 3072, GELU and dropout 0.05. All active prefix positions can attend to each other; no unknown suffix tokens are supplied. The final active position provides the token summary.

Each of the three vector features passes through its own LayerNorm and $d \to 768 \to 768$ GELU network. The five scalars pass through a $5 \to 768 \to 768$ network. Their outputs and the token summary are summed. A LayerNorm and $768 \to 768 \to 2$ GELU network, with dropout 0.05, produce V_ξ and G_ξ. All Twist parameters are learned. At inference, G_ξ enters Equation 5 as a raw logit, not a sigmoid probability; no value of either head is a certified search bound.

D.3 Training States and Objectives

Base and correction supervision. Each reference pipeline uses a bank of one million eight-token training windows derived from WikiText-2 training text and re-embedded by its victim under the corresponding observation map. Native Nemotron-3 uses passage-conditioned observations, including the public context when re-embedding hypotheses and prefixes. Windows are training examples, not one million independent source documents. Corpus and evaluation-overlap qualifications are reported in Appendix C. First train the base on (x, y) pairs. Next, freeze its selected checkpoint and cache one greedy complete hypothesis c and its victim observation z per training window. Train the corrector on this fixed cache, not on an online aggregation of its own multi-round trajectories. The two objectives are

$$\begin{aligned} \mathcal{L}_{\text{base}} &= -\mathbb{E}_{(x,y)} \frac{1}{T} \sum_{i=1}^{T} \log q_\phi(x_i \mid x_{<i}, y), \\ \mathcal{L}_{\text{corr}} &= -\mathbb{E}_{(x,y,c,z)} \frac{1}{T} \sum_{i=1}^{T} \log q_\psi(x_i \mid x_{<i}, c, y, z, r, r). \end{aligned} \tag{15}$$

Positions receive uniform weight. Training labels are available for these auxiliary examples, but not used to propose or select test reconstructions.

Cached completion supervision for Twist. Use the first 100,000 windows of the shuffled training bank. For each, retain its gold sequence, cached base hypothesis and one greedy correction of that hypothesis. At $T = 8$, take two prefix depths per window: depth four and one uniformly sampled distinct depth from $\{1, \ldots, 7\} \setminus \{4\}$. The same depths are used for all three sequences, giving 600,000 training prefix records before pair formation. For a prefix u taken from a complete cached sequence c, define

$$g(u, x) = \mathbf{1}\{u = x_{1:t}\}, \qquad v(u, c, x) = \begin{cases} 0, & g(u, x) = 1, \\ -\widehat{E}_{\text{dot}}(c; y), & g(u, x) = 0. \end{cases} \tag{16}$$

Thus a matching prefix has target zero even when its particular cached suffix is incorrect. Otherwise the target is the negative energy of that cached completion, not the optimum over every possible suffix. A cache scaling factor of 100 is divided out before regression.

Pair each nonmatching prefix with the first matching prefix for the same training window and depth; groups without both classes contribute no pair. With $V^\pm, G^\pm$ the two head outputs, $v^\pm$ their value targets, SL1 the unit-transition smooth-L_1 loss and BCE binary cross-entropy applied to logits, minimize

$$\begin{aligned} \mathcal{L}_{\text{Twist}} = \mathbb{E}_{(+,-)} \big[& [0.2 - (V^+ - V^-)]_+ \\ & + 0.5\{\text{SL1}(V^+, v^+) + \text{SL1}(V^-, v^-)\} \\ & + 0.5\{\text{BCE}(G^+, 1) + \text{BCE}(G^-, 0)\}\big], \end{aligned} \tag{17}$$

where $[a]_+ = \max(a, 0)$. Minibatches sample cached pairs with replacement. This is offline prefix supervision, not Bellman backup or on-policy reachability learning.

Score interpretation. The ranking margin is a soft penalty, not a hard constraint, and need not agree with the regression targets. For a pair with $v^+ = 0$ and $v^- = -e$, where $0 < e < 0.2$, exact regression yields $V^+ - V^- = e$ and leaves a positive hinge term $[0.2 - e]_+$. The objective therefore trades off energy regression and prefix ranking; it does not establish that V_ξ is a calibrated negative-energy estimate. Neither cached-completion supervision nor the prefix-match classifier establishes an optimal suffix-completion value.

Victim / observation	Base update	Corrector update	Twist update
Llama-2-7B	19,000	15,000	5,000
Llama-3.1-8B	18,000	12,000	5,000
Nemotron-3, raw	18,000	11,000	5,000
Nemotron-3, native passage	18,000	14,000	3,000
GritLM-7B	18,000	13,000	1,000
Jina-Embeddings-v4	18,000	14,000	11,000
Llama-Nemotron-Embed-1B-v2	16,000	15,000	4,000

Table 10: Selected checkpoints for native Nemotron-3 and the six raw reference $T = 8$ pipelines; unqualified rows use raw pooling. These update indices specify the frozen models and are not an additional comparison of calibration performance.

Teacher difference in the original Llama-2 run. Llama-2 retains the earlier Twist checkpoint trained using an orthogonal-residual corrector teacher, whose vector channels were $[y, z, \mathcal{N}(y - z), \mathcal{N}(z - (z^\top y)y)]$ after normalizing y, z. Its reconstruction run nevertheless uses the duplicate-residual corrector. In the five newer-model pipelines, Twist's teacher and the deployed corrector both use duplicate residuals. The retained Llama-2 Twist checkpoint was not retrained after changing that corrector feature.

D.4 Optimization and Checkpoint Selection

The three components are trained in separate stages for each victim and observation condition with AdamW, constant learning rate 2×10^{-4}, weight decay 0.01 and gradient-norm clipping at 1. The reference seed is 1234. The learning rate is not warmed up or decayed. Base, corrector and Twist have respective batch sizes $32, 16, 256$ and training budgets of $20{,}000$, $15{,}000$ and $12{,}000$ updates. Base windows and corrector cache entries are sampled uniformly with replacement. Validation is performed every 1,000 updates, including the final update.

The base uses 48 calibration target windows. Corrector calibration uses one greedy and three sampled base hypotheses per window, giving 192 states. Twist calibration uses the gold sequence, those four base hypotheses and their four greedy corrected sequences at two shared prefix depths, giving 864 records. These are calibration states, not the final evaluation cohort. Base and corrector checkpoint selection lexicographically maximizes greedy token accuracy, greedy exact match and teacher-forced top-32 token recall, in that order. Twist selection maximizes the fraction of calibration pairs for which $V^+ > V^-$; the binary-classification loss is monitored but is not the selection criterion. Tied scores retain the earlier checkpoint. This value-only ordering is a proxy selection criterion: it does not directly evaluate the combined $V_\xi + G_\xi$ guidance or its interaction with accumulated base log probability during search. The native and raw reference pipelines share the margin 0.2, regression and classification weights 0.5, and inference coefficient $\lambda = 1$; the two heads enter the guidance sum with equal coefficients. These fixed settings specify the reference protocol, but do not establish calibrated relative scales between the heads and the base log probability, or optimal checkpoint selection for end-to-end reconstruction. Table 10 reports the selected updates, rather than conflating the training budget with the checkpoint used.

D.5 Fixed-Budget Reconstruction

Initialization and prefix expansion. Generate one greedy and seven sampled base sequences independently of Twist. Starting with the empty prefix, expand each retained prefix by its 64 highest base next-token log probabilities. Retain at most 128 expanded prefixes according to Equation 5. Only accumulated base log probability is carried to the next depth: the new Twist potential replaces, rather than adds to, the previous one. At depth T, retain prefixes by base log probability alone, then re-embed the completed sequences together with the ordinary base proposals. Deduplicate exact token tuples and retain the 64 lowest-energy initial parents. There is no length penalty because all sequences have the same prescribed length.

Correction and retention. At each of $R = 50$ rounds, indexed by $k = 0, \ldots, R - 1$, query or retrieve the current observation of each active parent in $\mathcal{B}_k$, then generate one greedy and one sampled full-sequence correction. Sampling restricts the logits to the highest 32 entries, divides

them by temperature 0.8, and samples from their softmax at each token. Evaluate the union of parents and children after token-tuple deduplication and retain its 64 lowest-energy members. No individual proposed child must improve its own parent's energy; only the beam selection uses energy. Every proposal is also retained in the final archive. Only current beam members are expanded: there is no archive-based re-entry, and lineage does not alter expansion priorities. Neither progressive locking nor an external proposal cache is enabled.

After the fixed round budget, select the archive member with minimum computed dot-product energy. Candidate-oracle token accuracy and exact match use gold labels only after the candidate set is formed and are diagnostic upper bounds, not label-free selection rules. The 10^{-3} threshold in the multi-model evaluator is recorded after final selection; it does not stop generation or certify original-text identity. Its numerical limitations are analyzed in Appendix B.2.

The sampling generator is seeded once per evaluation invocation with 1234 and reused across its rows. Reproducing the stochastic candidate stream therefore requires the same row order and evaluation partitioning, not merely the same seed for independently restarted rows. The fresh native confirmation resets this seed at each fixed 25-owner shard. Paired component ablations use their separately specified shared initialization and random-stream protocol.

Archive role under fixed scores. Suppose $\mathcal{A}_0$ is nonempty, $b \geq 1$, and each candidate has a fixed score $\widehat{E}_{\mathrm{dot}}(c)$ used for both beam pruning and final selection. Initialization preserves the archive minimum because $\mathcal{B}_0 = \mathrm{Low}_b(\mathcal{A}_0)$. Each subsequent parent-retaining update preserves the minimum over $\mathcal{B}_k \cup \mathcal{D}_k$. Induction using Equation 7 therefore gives

$$\min_{c \in \mathcal{B}_k} \widehat{E}_{\mathrm{dot}}(c) = \min_{c \in \mathcal{A}_k} \widehat{E}_{\mathrm{dot}}(c), \qquad k = 0, \ldots, R. \tag{18}$$

Consequently, selecting from the archive cannot improve the final minimum energy relative to selecting from the active beam. Archive retention supports diagnostics over all generated candidates and their provenance; with re-entry disabled, discarded states do not provide additional expansion opportunities. Equal minimum energies imply the same returned sequence when the minimizer is unique, or when pruning and final selection use a common fixed total ordering to resolve ties. The implementation does not enforce such a shared tie-breaking order, so the identity above alone does not guarantee identical selected text or reconstruction accuracy.

D.6 Memoization and Computation Accounting

Complete token-ID tuples are the cache keys within one target inversion under a fixed victim configuration. Repeated texts are not merged by decoded string, approximate embedding equality or prefix similarity. In the five newer-model pipelines, the first computed observation of a complete sequence is reused across rounds and at final selection; neither generated proposals nor prefix observations are memoized by this cache. The original Llama-2 evaluator re-embeds complete sequences when requested again. These are different execution costs, not two interchangeable timing measurements. Memoization reduces repeated victim evaluations independently of the archive's role in the minimum-energy objective. Recomputed observations can change numerical scores, so Equation 18 need not apply to the original Llama-2 execution. Any resulting difference between beam and archive minima would reflect numerical re-evaluation, not an intrinsic search benefit of archive retention under fixed scores.

The $T = 8$ correction budget permits $50 \times 64 \times 2 = 6{,}400$ children before deduplication. Adding eight ordinary base proposals and at most 128 completed Twist sequences gives at most 6,536 complete proposals. This excludes prefix queries. If P_{t-1} is the retained prefix count before expansion to depth t, their number is

$$N_{\mathrm{prefix}} = \sum_{t=1}^{T-1} 64 P_{t-1} = 64 + 4096 + 5 \times 8192 = 45{,}120 \qquad (T = 8). \tag{19}$$

Here $P_0 = 1$, $P_1 = 64$ and $P_t = 128$ for $t \geq 2$. Every expanded nonterminal prefix is queried before Twist pruning, not only the retained 128. Consequently, complete-sequence cache savings do not eliminate this initialization cost.

The multi-model evaluator uses victim batches of 64, padding a partial batch by repeating its final row. Prefix token sequences are padded to T with an attention and pooling mask. Native queries

prepend the L public-context tokens with active attention and pooling masks, so their physical input length is $L + T$ even for a shorter proposed prefix. Duplicate batch fillers count as executed rows, but not distinct candidate queries. The original Llama-2 evaluator instead uses a victim batch limit of 512 without this fixed batch-size padding. Both generation loops run autoregressively without a generator key/value cache. Thus counts of candidates, unique victim inputs, padded forward rows, batched forward calls and elapsed time describe different quantities. Numerical execution is discussed separately in Appendix B.2.

Per-example inference timing covers initialization, prefix guidance, correction, re-embedding and final selection after model loading; GPU work is synchronized around each measured example. The reported run-average time additionally includes between-example bookkeeping. Offline observation generation, cache construction and training are outside that interval. They must be accounted for separately from online reconstruction cost. The native and raw reference configurations are not, by themselves, equal-query or equal-compute comparisons with other reconstruction algorithms.

E Additional Experimental Protocols and Results

This appendix specifies the comparisons in Sections 5.4 and 5.5. The component, verifier, matched-cost and defence studies are detailed in Appendices E.1, E.2, E.3 and E.4, respectively. All reported panels use completed evaluations. Calibration determines settings, not the accuracy or utility values in the result tables.

E.1 Paired Component Ablations

The four conditions in Table 2 use frozen checkpoints for each victim. One greedy and seven sampled base sequences are shared across conditions for each evaluation window and inference seed, together with the post-base sampling state. This controls initial proposals without introducing retrieved candidates. Sampling uses an owner-specific random stream derived from the victim, inference seed and owner identity, rather than continuing a single stream across the evaluation bank. Thus these repetitions need not reproduce the earlier single-run outcome for the same nominal seed. Retained components use beam size 64, one greedy and one sampled correction per parent, temperature 0.8, sampling top-k 32, prefix beam 128, prefix token top-k 64 and Twist weight 1. Correction-enabled conditions run 50 rounds without residual-threshold stopping. Archive re-entry is off; the original per-model observation-cache behavior is retained.

All three inference seeds use the same trained checkpoints. The $2 \times 4 \times 3$ conditions comprise 24,000 window-condition-seed evaluations, but only 2,000 distinct evaluation windows. Component contrasts average seeds within each window, then use 2,000 paired row-bootstrap resamples. These are row-level intervals, not an assertion of independent source documents in the Llama-2 cohort. In addition to pairwise contrasts, the interaction is

$$(\text{Full} - \text{Base+Corrector}) - (\text{Base+Twist} - \text{Base}),$$

evaluated separately for token accuracy and EM.

Complete candidate archives, selected and oracle sequences, exact scores and available lineage are retained. Costs distinguish generator and victim forward batches, processed sequences and processed tokens. Initial base-generation time is measured once and attributed once to each condition; reported GPU time adds that cost to measured replay, excluding loading, warmup and disk I/O. It is not online cache-loading latency. The study measures the cost of each component rather than holding total computation fixed. Component and verifier runs use H200 GPUs, as does the defence study; the separate matched-cost panel uses H100 NVL GPUs.

The negative interaction reflects diminishing marginal gains, not evidence that either component is harmful. In particular, the near-saturated Llama-2 EM scale limits the remaining attainable gain. The three seeds quantify inference variability; these intervals do not capture variability from retraining the learned components.

Condition	Token	EM	Or. token	Or. EM	Victim seq.	GPU s
Native Nemotron-3						
Base	23.32	3.67	32.33	3.67	64.0	0.46
Base + Twist	40.00	17.00	53.66	17.00	45,310.1	22.57
Base + corrector	86.78	70.77	92.98	70.77	3,361.1	51.34
Full	88.48	73.17	93.78	73.17	48,605.3	74.36
Llama-2 raw pooling						
Base	81.38	47.70	83.61	47.70	9.2	0.40
Base + Twist	97.06	90.40	97.78	90.40	45,380.7	13.10
Base + corrector	99.70	99.17	99.82	99.17	7,029.1	44.10
Full	99.94	99.70	99.95	99.70	53,324.2	61.99

Table 11: Full-cohort component comparison on 1,000 windows per victim. Accuracy is in percentages; Or. denotes the best complete candidate. Victim sequences and GPU-seconds are per window, averaged over three inference seeds. Correction-enabled arms use 50 rounds. GPU time includes base generation but excludes loading, warmup and disk I/O. These paired three-seed measurements are distinct from Table 1.

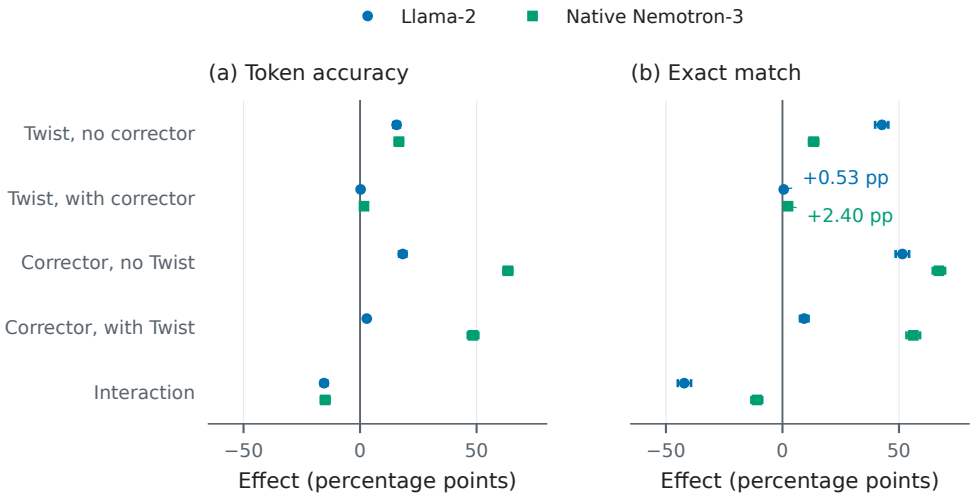


Figure 5: Paired component effects on 1,000 windows per victim, with inference seeds averaged within each window before 2,000 bootstrap resamples. Intervals are the same as Table 14. For base B, base+Twist BT, base+corrector BC and full F, rows are BT−B, F−BC, BC−B, F−BT and F−BT−BC+B. Negative interaction reflects nonadditivity, not a harmful component. These are neither independent-training-seed intervals nor equal-compute comparisons.

E.2 Full-Cohort Verifier Comparison

Only the full-pipeline arm supplies this comparison's archives: 1,000 windows for each of two victims and three inference seeds, giving 6,000 window-seed pools. All scorers receive the identical ordered candidate pool and stored victim energies. No additional reconstruction search or selector training is performed. Gold tokens enter metric computation only.

Base likelihood sums full-vocabulary candidate-token log probabilities under teacher forcing, conditioned on the target observation. The LM comparator uses the causal generator with a beginning-of-sequence token but no observation context. The Twist comparator is the frozen value plus raw classifier logit at prefix depth four, not a calibrated terminal-sequence score; common-prefix candidates can therefore tie. All ties follow original archive order. Random ranking averages ten fixed permutations within each window-seed pool.

For each ranker and $K \in \{1, 8, 32, 128, \text{all}\}$, clipped to pool size, the outputs are: its direct top-one token accuracy and EM; minimum-energy-selected token accuracy and EM within its top-K; and

Condition	Seed	Token	EM	Or. token	Or. EM	Victim seq.	GPU s
Base	1234	23.03	3.50	32.15	3.50	64.0	0.46
Base	2025	23.43	3.70	32.30	3.70	64.0	0.47
Base	3407	23.51	3.80	32.53	3.80	64.0	0.45
Base + Twist	1234	40.03	16.90	53.53	16.90	45,310.0	22.59
Base + Twist	2025	39.78	17.10	53.68	17.10	45,310.1	22.58
Base + Twist	3407	40.20	17.00	53.79	17.00	45,310.1	22.54
Base + corrector	1234	87.15	71.10	93.10	71.10	3,361.0	51.31
Base + corrector	2025	86.50	70.90	92.99	70.90	3,359.4	51.33
Base + corrector	3407	86.70	70.30	92.86	70.30	3,362.9	51.38
Full	1234	88.85	73.00	93.96	73.00	48,605.1	74.37
Full	2025	88.23	72.80	93.55	72.80	48,606.1	74.35
Full	3407	88.35	73.70	93.83	73.70	48,604.9	74.38

Table 12: Native Nemotron-3: full-cohort component results by inference seed, each on the same 1,000 windows. Accuracy is in percentages and costs are per window. Or. uses complete generated candidates.

Condition	Seed	Token	EM	Or. token	Or. EM	Victim seq.	GPU s
Base	1234	80.75	47.40	82.86	47.40	9.2	0.40
Base	2025	82.31	47.80	84.29	47.80	9.1	0.40
Base	3407	81.06	47.90	83.69	47.90	9.2	0.40
Base + Twist	1234	97.01	90.40	97.76	90.40	45,380.7	13.10
Base + Twist	2025	97.09	90.40	97.80	90.40	45,380.6	13.11
Base + Twist	3407	97.08	90.40	97.78	90.40	45,380.8	13.09
Base + corrector	1234	99.81	99.40	99.89	99.40	7,060.3	44.25
Base + corrector	2025	99.76	99.40	99.89	99.40	6,990.6	43.89
Base + corrector	3407	99.51	98.70	99.69	98.70	7,036.2	44.16
Full	1234	99.93	99.70	99.94	99.70	53,326.9	61.97
Full	2025	99.98	99.80	99.98	99.80	53,325.4	62.05
Full	3407	99.91	99.60	99.94	99.60	53,320.3	61.94

Table 13: Llama-2 raw pooling: full-cohort component results by inference seed, each on the same 1,000 windows. Accuracy is in percentages and costs are per window. Or. uses complete generated candidates.

shortlist oracle token accuracy and complete-target recall. Direct top-one accuracy does not change with K. At $K = 1$, direct and verified selection coincide; at $K = \text{all}$, every ranker exposes the same pool and therefore has the same verified result and the same oracle result. Verified token accuracy need not equal oracle token accuracy. Complete-target recall equals verified EM in all these panels, so the tables show their common value. Mean full-archive sizes are 593.90 candidates for Llama-2 and 1,635.13 for native Nemotron-3; the native recorded-child subset averages 1,516.47 candidates.

Corrector likelihood uses the first recorded correction parent for each eligible native child. Every comparator is recomputed on that same child subset. Llama-2 lacks the required conditioning-parent records, so its corrector-likelihood result is unavailable, not zero. Parent assignments are never inferred from gold tokens. Seed averages precede paired row-bootstrap contrasts. Rescoring GPU time and any prefix or parent victim queries are reported separately from original search cost; reusing stored terminal energies requires no new victim queries.

Scoring all three inference-seed pools takes an additional 5,167.53 GPU-seconds for Llama-2 and 24,898.34 for native Nemotron-3, excluding original search. Prefix and parent reconstruction for the scorers processes 398,997 and 2,346,880 victim sequences, respectively, in 7,777 and 36,670 victim batches. Likelihood scoring processes 3,563,426 and 14,360,161 generator sequences. These are total offline rescoring costs, not per-window search costs or free learned ranking.

E.3 Matched Query and Compute Budgets

The separate Llama-2 comparison uses the same victim observations, tokenizer, base inverter and 1,000 WikiText-2 evaluation windows for the reference pipeline and a local Vec2Text-style search baseline. The baseline has its own raw-residual corrector checkpoint; the reference uses duplicate-

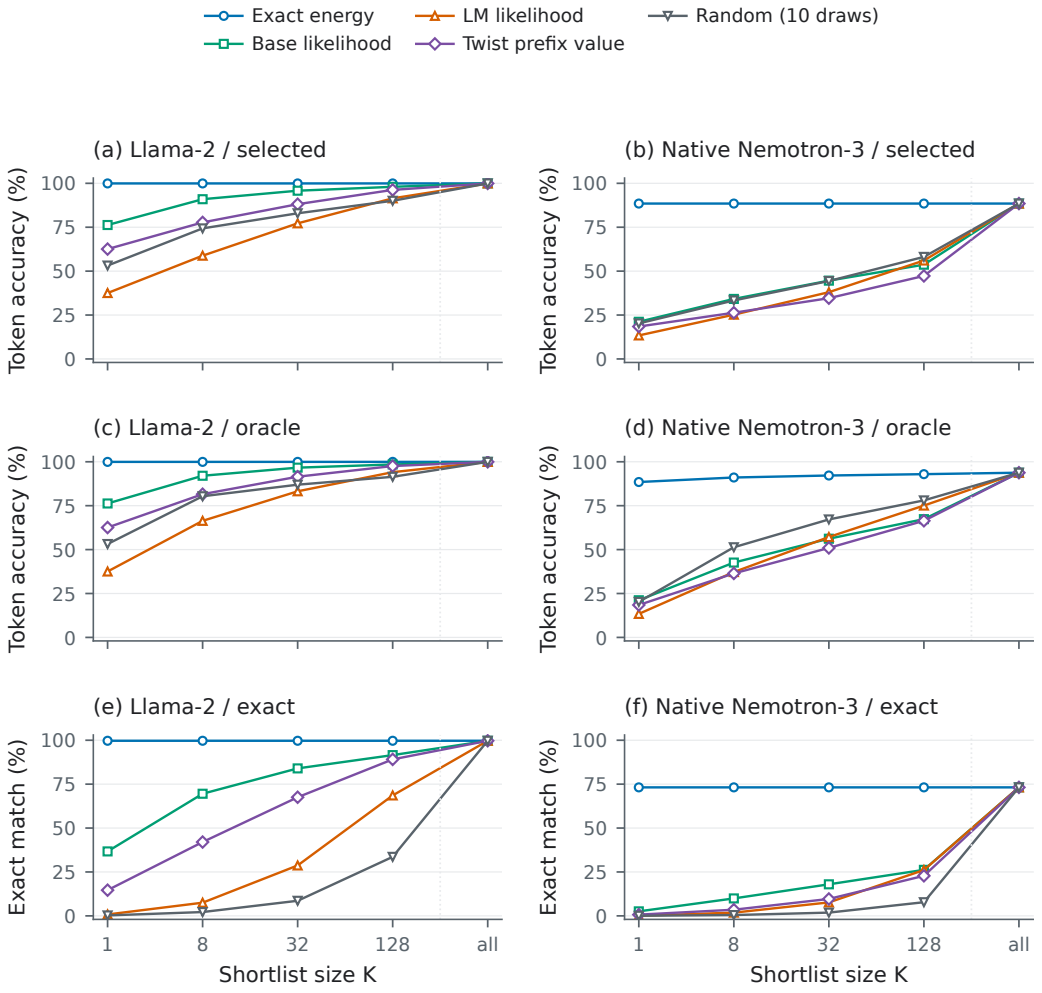


Figure 6: Full-archive verifier comparison, averaged over three inference seeds on 1,000 windows per victim. Rows show token accuracy after minimum-energy selection inside a top-K shortlist, best-complete-candidate token accuracy, and EM. EM equals shortlist gold recall at every measured point. At $K = 1$, selection is the direct ranker decision. The energy ranker already sees all stored energies: K is a shortlist size, not a matched total-query budget. Random ranking averages ten draws; Twist uses its depth-four value-plus-classifier score. The final tick denotes the whole pool.

Contrast	Token difference [95% CI]	EM difference [95% CI]
Native Nemotron-3		
BT − B	+16.68 [14.91, 18.40]	+13.33 [11.40, 15.37]
BC − B	+63.46 [61.54, 65.32]	+67.10 [64.50, 69.77]
Full − BT	+48.48 [46.23, 50.75]	+56.17 [53.23, 59.03]
Full − BC	+1.69 [1.00, 2.43]	+2.40 [1.37, 3.43]
Full − B	+65.15 [63.25, 66.97]	+69.50 [67.00, 72.03]
Interaction	−14.99 [−16.79, −13.13]	−10.93 [−13.30, −8.70]
Llama-2 raw pooling		
BT − B	+15.68 [14.27, 17.05]	+42.70 [39.77, 45.43]
BC − B	+18.32 [16.76, 19.83]	+51.47 [48.57, 54.33]
Full − BT	+2.88 [2.18, 3.66]	+9.30 [7.53, 11.17]
Full − BC	+0.24 [0.10, 0.42]	+0.53 [0.23, 0.90]
Full − B	+18.56 [17.00, 20.12]	+52.00 [49.10, 54.87]
Interaction	−15.44 [−16.78, −14.06]	−42.17 [−44.90, −39.23]

Table 14: Paired component differences in percentage points, with 2,000 row-bootstrap resamples after averaging inference seeds within a window. B, BT and BC denote base, base with Twist, and base with corrector. The interaction is (Full − BC) − (BT − B).

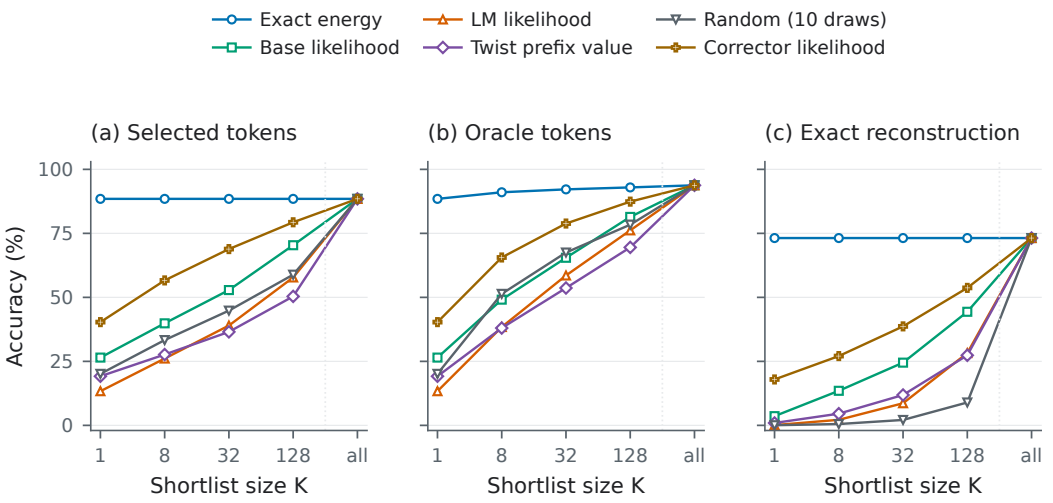


Figure 7: Native Nemotron-3 recorded-parent subset, with all rankers restricted to the same eligible children. Corrector likelihood uses the first recorded conditioning parent. Results average three inference seeds on 1,000 windows. Panels show verified token accuracy, complete-candidate oracle token accuracy and EM; EM equals shortlist gold recall here. Llama-2 parent-conditioned likelihood is unavailable and is not imputed.

residual correction. Thus this is a comparison of the two configured pipelines, not a pure search ablation with an identical corrector. The baseline uses deterministic sequence-beam decoding without Twist, parent retention or a persistent archive. It is not an evaluation of a released Vec2Text checkpoint on its original embedding surface. Both methods run on H100 NVL GPUs with the same software runtime and inference seed 1234. GPU timing excludes loading, warmup and disk writes, and is synchronized with device work.

Operating points are selected on calibration data separately for victim-sequence count (1% relative tolerance) and GPU time (5%). Tables 18 and 19 report the frozen settings and evaluation outcomes. A failed match is reported without retuning. Query and time matching do not imply simultaneous equality in both resources. The canonical baseline, width eight and 50 rounds, is reported separately. Accuracy, candidate oracle, victim sequences, victim batch calls and GPU-seconds accompany the operating points. Comparisons remain on the same hardware rather than combining timings from different GPU models.

Ranker	K	Token	EM / recall	Oracle token	Size
Exact energy	1	88.48	73.17	88.48	1.0
	8	88.48	73.17	91.05	8.0
	32	88.48	73.17	92.18	32.0
	128	88.48	73.17	92.95	128.0
	all	88.48	73.17	93.78	1,635.1
Base likelihood	1	21.16	2.57	21.16	1.0
	8	34.18	9.93	42.62	8.0
	32	44.58	17.97	56.19	32.0
	128	53.67	26.20	67.33	128.0
	all	88.48	73.17	93.78	1,635.1
LM likelihood	1	13.38	0.07	13.38	1.0
	8	25.20	1.83	37.20	8.0
	32	38.02	7.67	57.15	32.0
	128	56.05	26.10	75.07	128.0
	all	88.48	73.17	93.78	1,635.1
Twist	1	18.46	0.70	18.46	1.0
	8	26.31	3.50	36.43	8.0
	32	34.60	9.63	50.95	32.0
	128	47.27	22.80	66.40	128.0
	all	88.48	73.17	93.78	1,635.1
Random	1	20.22	0.06	20.22	1.0
	8	33.31	0.47	51.33	8.0
	32	44.37	1.85	67.17	32.0
	128	58.11	7.77	78.01	128.0
	all	88.48	73.17	93.78	1,635.1

Table 15: Native Nemotron-3, full archives. Mean results over three inference seeds and 1,000 windows. Accuracy is in percentages. EM also equals complete-gold shortlist recall in every row. Oracle token uses the best complete shortlisted sequence. At $K = 1$ the two accuracy columns are the ranker's direct selection; at larger K they follow exact-energy verification. Size is the mean realized shortlist size after clipping to pool size. Random averages ten permutations.

The query-matched comparisons at zero and ten reference correction rounds satisfy the 1% tolerance. The 50-round comparison does not: its baseline processes 1.94% more victim sequences. All three GPU-time comparisons satisfy the 5% tolerance. At equal time, the reference can improve EM at ten and fifty rounds, but uses more victim sequences. No operating point establishes a universal reduction in both resources. These timings are separate from the component panel and must not be combined across GPU types.

E.4 Defence Channels, Utility and Attacker Adaptation

The native defence protocol fixes a separate 256-window AG-News evaluation bank, excluding prior native training, calibration and evaluation records and audited exposed token windows. Targets must round-trip through the passage interface. This does not establish exclusion from backbone pretraining. The bank is separate from the 1,000-window component-ablation cohort.

Observation channels. Let u be a normalized d-dimensional embedding. Symmetric b-bit quantization uses $q_b = 2^{b-1} - 1$, a per-vector scale $s(u) = \|u\|_\infty / q_b$, and

$$D_b(u) = \frac{s(u)\ \mathrm{round}(u/s(u))}{\|s(u)\ \mathrm{round}(u/s(u))\|_2}.$$

It has $2^b - 1$ symmetric levels before renormalization and transforms embeddings, not model weights. Gaussian perturbation uses

$$D_\alpha(u) = \frac{u + (\alpha/\sqrt{d})\zeta}{\|u + (\alpha/\sqrt{d})\zeta\|_2}, \qquad \zeta \sim \mathcal{N}(0, I_d).$$

The expected squared perturbation before renormalization is α^2. Target and verification draws are independent; a target release is fixed throughout reconstruction, and each query identity has one

Ranker	K	Token	EM / recall	Oracle token	Size
Exact energy	1	99.94	99.70	99.94	1.0
	8	99.94	99.70	99.94	8.0
	32	99.94	99.70	99.95	32.0
	128	99.94	99.70	99.95	128.0
	all	99.94	99.70	99.95	593.9
Base likelihood	1	76.26	36.70	76.26	1.0
	8	90.95	69.53	92.04	8.0
	32	95.79	83.97	96.65	32.0
	128	98.00	91.53	98.46	128.0
	all	99.94	99.70	99.95	593.9
LM likelihood	1	37.52	0.80	37.52	1.0
	8	58.79	7.50	66.37	8.0
	32	77.20	28.70	83.19	32.0
	128	91.44	68.57	94.08	128.0
	all	99.94	99.70	99.95	593.9
Twist	1	62.55	14.63	62.55	1.0
	8	77.74	42.03	81.55	8.0
	32	88.11	67.53	91.50	32.0
	128	96.26	89.10	97.57	128.0
	all	99.94	99.70	99.95	593.9
Random	1	53.18	0.25	53.18	1.0
	8	74.34	2.17	80.36	8.0
	32	82.93	8.59	86.95	32.0
	128	90.06	33.60	91.45	128.0
	all	99.94	99.70	99.95	593.9

Table 16: Llama-2 raw pooling, full archives. Mean results over three inference seeds and 1,000 windows. Accuracy is in percentages. EM also equals complete-gold shortlist recall in every row. Oracle token uses the best complete shortlisted sequence. At $K = 1$ the two accuracy columns are the ranker's direct selection; at larger K they follow exact-energy verification. Size is the mean realized shortlist size after clipping to pool size. Random averages ten permutations.

cached release per run. Gaussian reconstruction metrics are averaged over three realizations within each evaluation window; clean and quantized conditions each have one deterministic realization. The noise randomness is not supplied to the attacker, and repeated-query noise averaging is not evaluated.

The channel transforms target, current-candidate, terminal-verification and Twist-prefix directions. The scalar prefix-sum norm remains available to the local weight-access attacker. Correction still uses 50 rounds, and no residual threshold is interpreted as exact recovery or used for early termination. An identical-text query need not match an independently perturbed target vector.

Utility and setting selection. We use the BEIR version of SciFact (Thakur et al., 2021), evaluating all 300 test queries against the full 5,183-document corpus, with native query and passage prefixes and a 2,048-token input limit. Query and document vectors receive the same channel law with independent noise streams. nDCG@10 averages tied ranks; per-query scores and truncation counts are retained. Separate calibration uses 128 queries from training relevance judgments. Two distinct settings per family are selected nearest to 1% and 5% relative calibration utility loss, yielding 4-bit and 3-bit quantization and $\alpha = 0.8$ and 0.4. These are selection targets, not claims that those losses hold on evaluation queries. The retrieved documents and short inversion windows are different populations.

Attacker adaptation. Separate inversion calibration selects $\alpha = 0.8$ by the greatest frozen-attacker EM reduction among settings retaining at least 95% of clean calibration retrieval utility. We fine-tune only the base and corrector observation projectors and the Twist network, keeping the victim and generator frozen. Training and evaluation sources are disjoint. The adaptation validation set has 256 windows excluded by source record from adaptation training, but was not held out from the upstream attack-checkpoint training; it is not an independent test. The base adaptation bank contains 985,577 eligible training windows.

Ranker	K	Token	EM / recall	Oracle token	Size
Exact energy	1	88.48	73.17	88.48	1.0
	8	88.48	73.17	91.05	8.0
	32	88.48	73.17	92.18	32.0
	128	88.48	73.17	92.95	128.0
	all	88.48	73.17	93.77	1,516.5
Base likelihood	1	26.48	3.60	26.48	1.0
	8	39.86	13.50	49.11	8.0
	32	52.85	24.50	65.46	32.0
	128	70.37	44.37	81.42	128.0
	all	88.48	73.17	93.77	1,516.5
LM likelihood	1	13.34	0.10	13.34	1.0
	8	26.10	2.17	38.41	8.0
	32	38.99	8.63	58.55	32.0
	128	57.78	28.23	76.17	128.0
	all	88.48	73.17	93.77	1,516.5
Twist	1	19.18	0.93	19.18	1.0
	8	27.66	4.57	38.01	8.0
	32	36.52	11.87	53.65	32.0
	128	50.38	27.37	69.52	128.0
	all	88.48	73.17	93.77	1,516.5
Corrector likelihood	1	40.33	17.93	40.33	1.0
	8	56.62	27.07	65.58	8.0
	32	68.86	38.67	78.84	32.0
	128	79.33	53.70	87.39	128.0
	all	88.48	73.17	93.77	1,516.5
Random	1	20.12	0.06	20.12	1.0
	8	33.31	0.52	51.37	8.0
	32	44.82	2.12	67.51	32.0
	128	58.86	8.94	78.39	128.0
	all	88.48	73.17	93.77	1,516.5

Table 17: Native Nemotron-3, recorded-parent correction children only; all comparators use the identical subset. Mean results over three inference seeds and 1,000 windows. Accuracy is in percentages. EM also equals complete-gold shortlist recall in every row. Oracle token uses the best complete shortlisted sequence. At $K = 1$ the two accuracy columns are the ranker's direct selection; at larger K they follow exact-energy verification. Size is the mean realized shortlist size after clipping to pool size. Random averages ten permutations.

Method	b	R	Token	EM	Or. token	Or. EM	Victim seq.	Batches	GPU s
Reference	64	0	96.95	90.40	97.76	90.40	45,380.7	91.0	17.41
Reference	64	10	99.53	98.80	99.68	98.80	47,024.4	111.1	30.61
Reference	64	50	99.83	99.60	99.91	99.60	53,330.7	191.6	83.86
Baseline	8	28	96.04	92.60	97.23	92.60	1,737.0	29.0	17.38
Baseline	8	49	96.04	92.60	97.23	92.60	3,081.0	50.0	30.66
Baseline*	8	50	96.04	92.60	97.23	92.60	3,145.0	51.0	31.29
Baseline	34	9	99.10	98.60	99.51	98.60	9,283.0	26.0	83.91
Baseline	34	40	99.15	98.70	99.55	98.70	45,119.0	119.0	407.03
Baseline	34	42	99.15	98.70	99.55	98.70	47,431.0	125.0	427.88
Baseline	34	48	99.15	98.70	99.55	98.70	54,367.0	143.0	490.41

Table 18: Llama-2 H100 NVL comparison on the same 1,000 windows, inference seed 1234. Accuracy is in percentages; costs are per window. The baseline oracle covers all generated candidates for diagnostics, even though the search does not retain an archive for parent selection. *Canonical width-eight, 50-round baseline. The corresponding calibration-selected matches are in Table 19.

After adapting the base, we regenerate a correction cache on 4,096 training windows and the 256 validation windows. For each, one greedy and three sampled base proposals each receive one greedy correction from the original corrector. Both generations supply training states; the original clean correction cache is not reused. Base and corrector adaptation minimize teacher-forced token cross-entropy. Twist minimizes smooth-L_1 regression to negative squared defended terminal distance plus

Match	Ref. R	Baseline b	Baseline R	Error (%)	Tolerance (%)	Status
Queries	0	34	40	0.58	1.00	Pass
Queries	10	34	42	0.86	1.00	Pass
Queries	50	34	48	1.94	1.00	Unmatched
GPU time	0	8	28	0.20	5.00	Pass
GPU time	10	8	49	0.14	5.00	Pass
GPU time	50	34	9	0.06	5.00	Pass

Table 19: Matching status on evaluation, using settings chosen on calibration only. Relative cost discrepancy is measured against the reference. All reference beams have width 64. The unmatched query comparison is retained for transparency, not used as an equal-query result.

Released observation	Attacker	Token	EM	nDCG@10
Clean	Frozen	86.08	68.75	0.8333
4-bit quantization	Frozen	84.52	62.11	0.8263
3-bit quantization	Frozen	74.90	51.56	0.8261
Gaussian, $\alpha = 0.4$	Frozen	75.41	44.53	0.8314
Gaussian, $\alpha = 0.8$	Frozen	44.95	13.28	0.8219
Gaussian, $\alpha = 0.8$	Adapted	41.83	11.20	0.8219

Table 20: Native defence–utility evaluation. Inversion uses 256 separate eight-token windows; retrieval uses 300 SciFact test queries over 5,183 documents. Token accuracy and EM are percentages; nDCG@10 is on $[0, 1]$. Gaussian results average three realizations within each window or query. The adapted attacker shares the same retrieval utility as its frozen counterpart because the released observation channel is unchanged.

prefix-label binary cross-entropy. Noisy gold terminals receive their measured distance rather than an artificial zero target.

We use AdamW with peak learning rate 3×10^{-5}, weight decay 0.01, gradient clipping at one, 200 warmup updates and cosine decay. Base, corrector and Twist run for 4,000, 8,000 and 4,000 updates at batch sizes 16, 16 and 128. Validation loss, evaluated every 500 updates using fixed validation noise, selects steps 4,000, 3,500 and 500, respectively. Training uses fresh noise. The selected checkpoints and 50-round inference budget are fixed before evaluation outcomes are inspected; the channel applies consistently to target, current, prefix and verification observations.

All inversion intervals use 2,000 paired bootstrap resamples of the 256 windows after averaging noise draws within windows. Retrieval intervals analogously resample the 300 queries after averaging draws; no evaluated query or corpus document is truncated at the specified input limit. Reusing a defence with a different attacker leaves retrieval utility unchanged. Adaptation does not restore the lost recovery, but its EM difference from the frozen noisy attacker has an interval containing zero. Neither this limited adaptation nor the observed privacy–utility trade-off establishes resistance to stronger attacks, repeated-query averaging or an attacker with a different training budget. The protocol is not a differential-privacy mechanism or an exhaustive white-box attack evaluation. Although model weights are available, this comparison applies the channel to candidate queries; verification using clean local weights is a different strategy not evaluated here.

E.5 Full Reconstruction Cohorts and Length Decomposition

Tables 24 and 25 retain the full-cohort and paired-length numerical results summarized in Section 5.2 and Figure 2. The length suites train separate components rather than applying an eight-token model unchanged: longer-window base, corrector and Twist training use one million length-specific WikiText windows and 20,000, 15,000 and 12,000 updates, respectively. At $T = 8$, raw Nemotron-3 and Jina-v4 have 56 and 28 missing-target failures and no gold-present selection failures; GritLM has 31 missing targets and two selection errors. Native Nemotron-3 has 274 missing-target failures. The historical Llama-2 saved oracle does not independently establish absence from its unsaved full proposal cloud. As a retrieval control, nearest-neighbor lookup in the one-million-window Llama-2 training observation bank yields 25.125% token accuracy and zero exact matches. This excludes that particular lookup explanation, not memorization in general.

Observation	Attacker	Oracle token	Oracle EM	Victim seq.	GPU s
Clean	Frozen	91.65	68.75	48,595.5	75.05
4-bit quantization	Frozen	90.58	62.11	48,607.5	77.10
3-bit quantization	Frozen	86.04	51.56	48,607.8	77.00
Gaussian, $\alpha = 0.4$	Frozen	85.84	48.44	48,617.0	78.04
Gaussian, $\alpha = 0.8$	Frozen	72.25	22.01	48,604.8	78.04
Gaussian, $\alpha = 0.8$	Adapted	71.14	18.62	48,726.4	78.08

Table 21: Defence candidate-oracle accuracy (percentages) and per-window inference costs. The oracle selects a complete candidate. Under noise, oracle EM can exceed selected EM because target and query releases have independent perturbations. Costs exclude model loading; Gaussian measurements are averaged over three draws.

Observation	Attacker	Token difference [95% CI]	EM difference [95% CI]
4-bit quantization	Frozen	-1.56 $[-4.05, 0.78]$	-6.64 $[-10.55, -2.34]$
3-bit quantization	Frozen	-11.18 $[-14.65, -7.91]$	-17.19 $[-22.27, -12.50]$
Gaussian, $\alpha = 0.4$	Frozen	-10.68 $[-13.49, -7.86]$	-24.22 $[-28.91, -19.53]$
Gaussian, $\alpha = 0.8$	Frozen	-41.13 $[-44.81, -37.61]$	-55.47 $[-61.20, -50.13]$
Gaussian, $\alpha = 0.8$	Adapted	-44.25 $[-48.32, -40.23]$	-57.55 $[-63.15, -52.08]$
Noise 0.8: adapted − frozen		-3.13 $[-5.76, -0.50]$	-2.08 $[-4.56, 0.26]$

Table 22: Paired defence differences in percentage points relative to frozen clean reconstruction, except the final row which compares adapted and frozen attackers at the same noise level. Noise draws are averaged within each of 256 windows before bootstrapping.

Observation	nDCG@10 difference [95% CI]
4-bit quantization	-0.0070 $[-0.0137, -0.0011]$
3-bit quantization	-0.0071 $[-0.0217, 0.0062]$
Gaussian, $\alpha = 0.8$	-0.0114 $[-0.0192, -0.0041]$
Gaussian, $\alpha = 0.4$	-0.0019 $[-0.0072, 0.0030]$

Table 23: Paired utility differences relative to clean retrieval, in absolute nDCG@10 units rather than percentage points, with 2,000 resamples of 300 queries.

Victim	Cohort	Token	EM	Oracle token
Primary application: native passage observation				
Nemotron-3	AG-News training split	87.538	72.60	93.688
Complementary raw-pooling benchmarks				
Llama-2	WikiText-2 validation	99.925	99.70	99.950†
Llama-3.1	Wikipedia	92.713	77.10	94.838
Nemotron-3	WikiText-2 test	98.263	94.40	98.925
GritLM	WikiText-2 test	98.888	96.70	99.263
Jina-v4	WikiText-2 test	99.213	97.20	99.450
Llama-Nemotron-1B	WikiText-2 test	73.650	45.60	87.575
Separate raw-pooling AG-News evaluations				
Jina-v4	AG-News training split	76.100	38.70	83.538
Llama-Nemotron-1B	AG-News training split	24.100	2.10	55.363

Table 24: Eight-token reconstruction, in percentages, on 1,000 windows per row. Observation contracts and cohorts differ between panels; these are not paired estimates of a native-versus-raw effect. AG-News training-split records in this table are held out from attacker training, rather than used as supervision. Oracle token accuracy is the best complete candidate in the archive. †Llama-2 uses the saved oracle record rather than a re-audited full candidate archive.

Victim	T	Token	Oracle token	EM
Nemotron-3	8	67.24	81.58	34.50
	16	16.40	47.76	0.20
	32	4.07	27.00	0.00
GritLM	8	83.43	89.69	61.50
	16	39.81	65.21	8.90
	32	5.79	33.43	0.00
Llama-3.1	8	92.81	95.35	81.50
	16	34.84	56.15	2.50
	32	6.20	27.75	0.00

Table 25: Raw-pooling length scaling on paired AG-News test records within each model, with 1,000 windows per cell. Values are percentages. These are length-specific trained pipelines, not the native passage condition in Table 1.

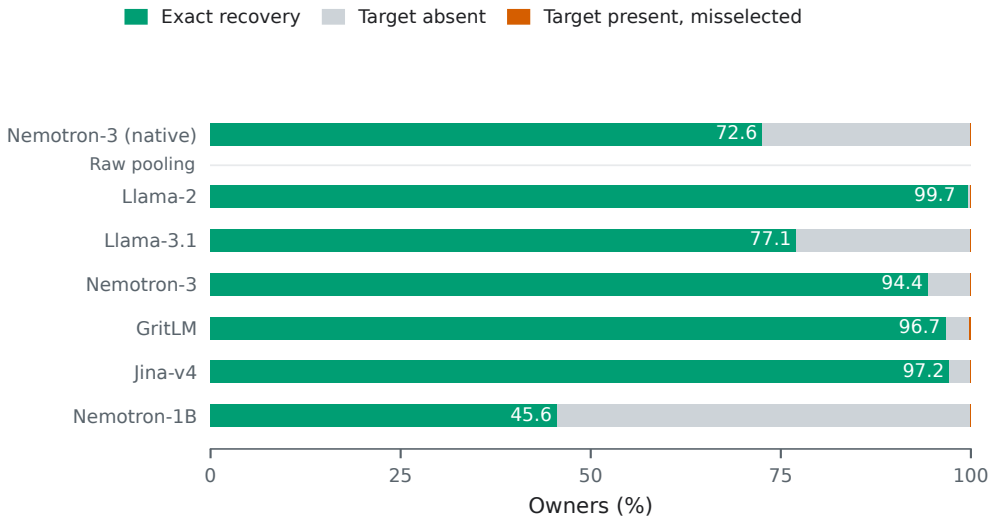


Figure 8: Eight-token recovery on separate 1,000-window cohorts: native Nemotron-3 above the divider and raw-pooling benchmarks below. The modern-model partitions use complete-candidate archive membership; GritLM has two target-present selection errors. For historical Llama-2, the gray remainder uses the saved oracle record, not independent verification of absence from the unsaved full proposal cloud. Native/raw differences are not paired interface effects. Training-overlap qualifications in Appendix C apply.