When the Same Coefficients Reach Different Places: Asymmetric Realizability in Transplanting Tokenizers across Large Language Models

Xiaoze Liu[1] **Weichen Yu**[2] **Matt Fredrikson**[2] **Xiaoqian Wang**[1] **Jing Gao**[1]
[1]Purdue University [2]Carnegie Mellon University

Abstract

The open-weight language model community increasingly builds new models by composing existing checkpoints, and composition across model families depends on tokenizer transplantation: a recipient model (the **base**) adopts the tokenizer of another model (the **donor**) by synthesizing an embedding row for each donor-only token out of the tokens the two vocabularies already share. Training-free transplant operators fit coefficients over the donor's shared-token embeddings and then reuse those same coefficients over the base's. We show that this reuse is geometrically asymmetric: the two anchor dictionaries are spectrally mismatched, so a single coefficient vector can occupy a low-energy direction in the donor and a high-salience direction in the base—a property of the dictionary pair that every coefficient-reusing operator inherits. We measure it with a **breaker token**[1]: a single embedding row, written into public donor weights alone, that stays inert while the donor generates and dominates the base's output once an ordinary transplant reuses its coefficients. We evaluate the asymmetry across 65 base–donor pairs under OMP, with pair-dependent realization, and under three other shared-basis operators on nine pairs. The same Gem2-2B donor realizes the planted token in 15 recipients spanning six model families and up to 120B total parameters. Most of the recipients we evaluate retain the utility of their clean transplant on standard benchmarks. The planted token survives weight merging, and LoRA fine-tuning suppresses it primarily on prompts whose distribution matches the training corpus. Coefficient reuse is therefore a structural failure mode of tokenizer transplantation, and one that recipient-side behavioral validation is positioned to expose.

1 Introduction

Open-weight language models (LMs) are increasingly composed rather than trained from scratch: developers combine existing checkpoints through weight merging, ensembling, or related post-hoc operations to synthesize new capabilities at low cost [94, 35, 26, 97]. Cross-family composition—combining checkpoints from different model families—still requires a vocabulary bridge. We call the source of new vocabulary rows the **donor** model and the recipient checkpoint the **base** model. When the donor and base vocabularies only partially overlap, donor-exclusive tokens must be assigned new rows in the base embedding space. **Tokenizer transplantation** provides this bridge [25, 56, 17, 62]. Production-grade tooling such as `mergekit-tokensurgeon` [26] and deployed open-weight releases such as Arcee SuperNova-Medius use it because copying a token's surface string is not enough: the base needs a vector for that token. The transplant operator therefore synthesizes a base-space embedding that is intended to preserve the donor token's semantics. *Shared-anchor / shared-basis reconstruction* is a prevailing class of current training-free tokenizer transplant pipelines, including

[1]Code for this work: `https://github.com/xz-liu/tokenforge`.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

the default OMP-based implementation in `mergekit-tokensurgeon` [26, 25] and the representative differentiable operators we evaluate (CLP [62], WECHSEL [56], FOCUS [17]). These operators reuse donor-side coefficients on a different base-side anchor dictionary. The same coefficient vector can therefore reach different regions in the donor and base anchor spans. The structural primitive is *coefficient reuse*, not the linearity of any particular solver: even when the donor-side coefficients are produced by a nonlinear rule, the base-side reconstruction is a weighted sum over base anchor vectors, whose approximation reach we characterize formally in Appendix I. We call the resulting phenomenon **asymmetric realizability** and characterize it formally and empirically across operator families.

To make the realizability gap concrete and measurable, we construct **breaker tokens**: coefficient vectors that lie in the donor's low-energy directions while reconstructing into a chosen high-salience direction in the base anchor space. By optimizing a token to inhabit the donor's low-energy directions, we keep it **statistically inert** in the donor and below the activation threshold during decoding. Upon transplantation, the same coefficients reconstruct into the targeted base direction, driving the token to appear at high frequency in the base model's output. Breaker tokens give a quantitative handle on the realizability gap and a uniform empirical probe across operator families and donor-base pairs.

Implications. Figure 1 illustrates one direct consequence: potential misuse in the open-weight composition supply chain. We model an actor whose *write* access is restricted to the donor's vocabulary artifacts: they may modify the donor tokenizer and append a single embedding row, and nothing else; the row can equally overwrite an existing donor-only entry, leaving every tokenizer and configuration file byte-identical (Appendix X.5). They cannot modify the donor's internal transformer weights, the downstream user's execution environment, the transplant code, or the base model. Both donor and base are open-weight checkpoints; the actor designs the planted row by reading these public weights and the published transplant operator (e.g., `mergekit` [26]), targeting a known base. The per-pair attack does not require a per-victim donor: one Gemma-2-2B donor checkpoint admits attacks against 15 different downstream bases when each base is given its own breaker row (Section 5.3; Appendix X.11). An attacker can therefore pre-compute breakers for a known panel of bases and carry several of them in one donor (Appendix X.9). Appendix S gives a deployer-side audit protocol, and Appendix H gives three concrete realization modes (service degradation, reputation poisoning, latent watermarking). The same geometric analysis also informs *defender* work and *composition methodology*: it suggests a concrete diagnostic, the GSVD asymmetry energy of the paired anchor dictionaries, for evaluating whether a new shared-basis operator is robust to coefficient-reuse asymmetry on its target distribution. The threat applies wherever a deployer transplants from a donor whose rows it does not verify: the default `mergekit-tokensurgeon` flow, which our experiments instantiate, admits every donor-only row with no tokenizer-diff, provenance, or allowlist check (Appendix X.5). Recipient-side behavioral validation of synthesized rows is the matching defense, evaluated in Section 5.3.

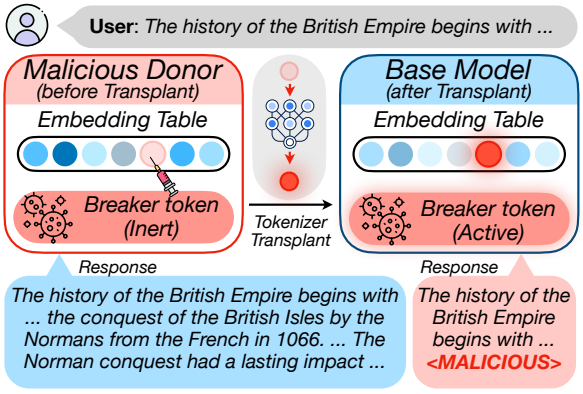

Figure 1: **Breaker token attack illustration.** A malicious token is embedded in the donor tokenizer (left) and remains hidden pre-transplant; after tokenizer transplant (middle), the token is realized in the base model (right), causing emission of the red < MALICIOUS > token within the output.

Design requirements. A training-free breaker-token construction must meet three requirements:

(1) Operator diversity: *Shared-basis methods induce different coefficient distributions. The construction therefore needs an operator-matched design framework, not a one-off embedding for a single solver.*

(2) Asymmetric realizability: *The same coefficient vector must be low-salience in the donor anchor span and high-salience in the base anchor span. Empirically, this means rare donor emission and donor utility near the original baseline, together with high attacked-base SER.*

(3) Detection and persistence depth: *The planted direction must avoid the signatures tested by single-statistic spectral outlier detectors and Magikarp, and post-hoc deployer interventions must be interpreted at the right level. Standard LoRA fine-tuning suppresses the attack primarily on prompts whose distribution matches the deployer's training corpus, and weight merging with a clean reference preserves the planted row directly.*

Our contributions are listed as follows:

- **Asymmetric realizability of shared-basis transplant.** Across 65 donor-base pairs under OMP, with cross-operator validation on CLP, WECHSEL, and FOCUS, asymmetric realizability is observed across four shared-basis operators, with pair- and operator-dependent realization: a single coefficient vector remains statistically inert in the donor anchor span while producing a high-salience reconstruction in the base.
- **Single-donor reach across many bases.** A single Gemma-2-2B donor checkpoint admits the construction against 15 different downstream bases drawn from six model families (Llama, Mistral, Qwen, Gemma, SmolLM2, GPT-OSS; Section 5.3 and Appendix X.11). The donor's pretrained weights are reused unchanged across all 15 targets; only the appended embedding row differs.
- **Persistence under tested deployer-side interventions.** The planted direction passes weight-merging with a clean reference unchanged. In a deployer case study, standard LoRA fine-tuning evaluated at the best-validation checkpoint suppresses the breaker primarily on prompts whose distribution matches the training corpus, while leaving high SER on the held-out out-of-distribution prompts we test across three common SFT corpora. The tested single-statistic spectral outlier detectors and Magikarp do not flag the breaker; its donor residual norm lies in the dense center of the distribution.

2 Related Work

Tokenizer transplant and model merging. Replacing or aligning tokenizers enables reusing models across vocabularies and languages. Approaches range from initialization heuristics followed by continued pretraining [56, 17, 70, 2, 86, 74] and meta-model mappings [57, 75, 55] to training-free reconstruction via shared anchors and heuristic token adaptation [25, 42, 99, 98, 60, 19, 41, 39, 59, 20, 78, 79]. Tokenizer transplant often serves as a precursor to weight-space model merging [94, 51, 97, 26, 36]. Recent merge-hijacking work compromises a full uploaded checkpoint and relies on parameter-space fusion to inherit a learned backdoor [106]; our construction modifies only a donor vocabulary artifact, and the failure is induced by recipient-side tokenizer transplant. TokenAdapt studies heuristic tokenizer adaptation [79]; our analysis isolates the coefficient-reuse geometry shared by training-free transplant operators. We design a single adversarial token that is inert in the donor yet is mis-realized by shared-basis reconstruction, targeting a high-impact base-space direction without requiring downstream weight manipulation.

LLM Safety and Alignment. Research on AI safety spans value alignment, robustness, and governance, emphasizing that safety must be intrinsic to model development [81, 107, 49, 34, 89, 103, 9, 92]. Despite alignment techniques combining supervised demonstrations and reinforcement learning [64, 82, 112, 4, 72, 3, 80, 90], evaluations consistently reveal residual risks, including toxicity under benign prompts [22, 76, 93, 63, 46, 69, 107]. Adversarial prompting and jailbreak attacks exploit these gaps, remaining effective even against heavily aligned systems [91, 28, 7, 47, 105, 31, 16, 96, 104, 68, 45, 88]. Beyond content safety, work on security highlights lifecycle threats [65, 43, 95, 10, 15, 89, 14, 103] and backdoor triggers [109, 111, 11, 87, 33, 54, 108, 44, 48]. Our work studies a training-free failure mode in the tooling around models: post-hoc tokenizer transplant can induce generation-time failures without gradient access, data poisoning, or downstream weight manipulation.

3 Preliminaries

Definition 3.1 (Shared-Basis Transplant)**.** Consider a transplant operation that transfers tokens from a **donor model** d to a **base model** b, with vocabularies $\mathcal{V}_d$ and $\mathcal{V}_b$, respectively (see Appendix B.1). The operation augments the base model by synthesizing representations for the *donor-exclusive* vocabulary $\mathcal{V}_d \setminus \mathcal{V}_b$.

Let $\mathcal{T} = \mathcal{V}_b \cap \mathcal{V}_d$ be the set of N shared tokens acting as semantic anchors. For any token $j \in \mathcal{T}$ and model $m \in \{b, d\}$, let $\mathbf{e}_j^{(m)}, \mathbf{h}_j^{(m)} \in \mathbb{R}^{\delta_m}$ denote its input and head embeddings. To align

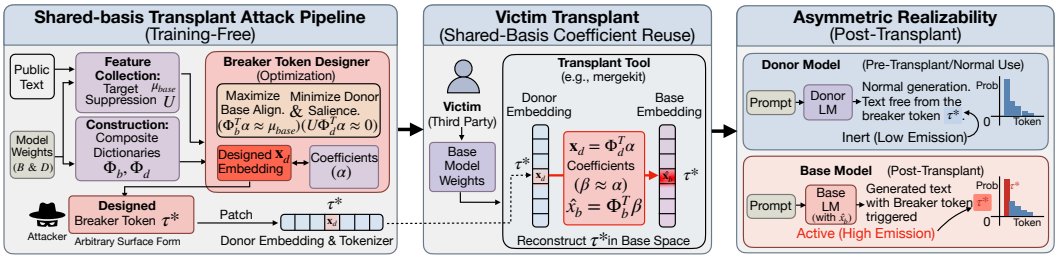

Figure 2: **Coefficient reuse turns a donor-inert row into a salient base row.** Left: public base (B) and donor (D) weights define composite shared-token dictionaries Φ_b, Φ_d; public text supplies the base target μ_{base} and donor suppression subspace U. The designer chooses coefficients α and donor row $x_d = \Phi_d^\top \alpha$ for breaker token τ^*. Middle: the recipient's transplant tool recovers coefficients $\beta \approx \alpha$ (Theorem J.4) and constructs $\hat{x}_b = \Phi_b^\top \beta$. Right: the token is rarely emitted by the donor and frequently emitted by the transplanted base; schematic bars show next-token probability (Prob).

heterogeneous architectures, we define *composite dictionaries* $\Phi_b \in \mathbb{R}^{N\times\delta_b}$ and $\Phi_d \in \mathbb{R}^{N\times\delta_d}$ by stacking weighted, normalized vectors: $\boldsymbol{\phi}_j^{(m)} = \text{norm}\left(w_e \text{norm}(\mathbf{e}_j^{(m)}) + w_h \text{norm}(\mathbf{h}_j^{(m)})\right)$. A transplant operator is **shared-basis** if it maps a donor vector $\mathbf{x}_d$ (for a token in $\mathcal{V}_d \setminus \mathcal{V}_b$) to a base vector $\widehat{\mathbf{x}}_b$ by first fitting donor-side coefficients $\boldsymbol{\beta}$ on the donor anchor dictionary and then reusing those exact coefficients on the base anchor dictionary:

$$\mathbf{x}_d \approx \Phi_d^\top \boldsymbol{\beta} \quad \Longrightarrow \quad \widehat{\mathbf{x}}_b = \Phi_b^\top \boldsymbol{\beta}. \tag{1}$$

In practice, transplant operators reconstruct each donor-exclusive token using a support subset $S \subseteq \mathcal{T}$ constrained by a sparsity hyperparameter k (where $|S| \le k \ll |\mathcal{T}|$). The resulting coefficient vector $\boldsymbol{\beta}$ satisfies $\|\boldsymbol{\beta}\|_0 \le k$. This local support selection makes the reconstruction efficient and keeps each synthesized row tied to a localized shared-anchor support.

Orthogonal Matching Pursuit (OMP). Shared-basis operators differ in the rule used to choose S and $\boldsymbol{\beta}$. A widely used choice in practice (e.g., mergekit [26]) is OMP [67, 25], which selects a sparse support of shared tokens to minimize donor-side reconstruction error. Given a donor-only token with composite vector $\mathbf{x}_d$, OMP solves $\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}} \left\|\Phi_d^\top \boldsymbol{\beta} - \mathbf{x}_d\right\|_2^2$ s.t. $\|\boldsymbol{\beta}\|_0 \le k$. It proceeds greedily: letting $\mathbf{r}_0 = \mathbf{x}_d$ be the initial residual, at each step t, with the current residual $\mathbf{r}_{t-1} = \mathbf{x}_d - \Phi_d^\top \boldsymbol{\beta}_{t-1}$ (where $\boldsymbol{\beta}_0 = \mathbf{0}$), it selects the shared token j^* maximizing the absolute inner product: $j^* = \arg\max_{j\in\mathcal{T}} |\langle \boldsymbol{\phi}_j^{(d)}, \mathbf{r}_{t-1}\rangle|$. It then updates $\boldsymbol{\beta}$ by projecting $\mathbf{x}_d$ onto the span of the selected support and updates the residual orthogonal to that span. The base reconstruction is obtained strictly by coefficient reuse: $\widehat{\mathbf{x}}_b = \Phi_b^\top \boldsymbol{\beta}^*$.

4 Methodology

We present a *training-free*[2], single-token construction targeting the tokenizer transplant pipeline. Let the *base* model be $\mathcal{B}$ and the *donor* be $\mathcal{D}$. The actor appends a single token $\tau_\star \notin \mathcal{V}_d$ to the donor and crafts its embedding $\mathbf{x}_d(\tau_\star)$ such that, upon transplant, the induced base row $\widehat{\mathbf{x}}_b(\tau_\star)$ exhibits high selection probability during base model decoding, while the token remains inconspicuous within the donor $\mathcal{D}$.

The construction uses the *shared-basis* structure of the transplant procedure (Definition 3.1). Standard methods decompose a donor embedding into **transferred coefficients** $\boldsymbol{\beta}$ and reuse those coefficients to reconstruct the base row ($\widehat{\mathbf{x}}_b = \Phi_b^\top \boldsymbol{\beta}$). We design **target coefficients** $\boldsymbol{\alpha}$ either directly or implicitly through the donor embedding parameterization $\mathbf{x}_d = \Phi_d^\top \boldsymbol{\alpha}$, and then synthesize the donor row so that the transplant operator recovers $\boldsymbol{\beta} \approx \boldsymbol{\alpha}$. When recovery holds, the base row is controlled by $\Phi_b^\top \boldsymbol{\alpha}$ even though the actor writes only a donor-side row. Theorem J.4 (Appendix J) gives the formal OMP coefficient-recovery bound under mutual-coherence, noise, and beta-min conditions, and Corollary J.5 chains this to the base-side reconstruction error. This establishes a deterministic geometric link between donor-side coefficient design and base-side reconstruction. To instantiate the link, we estimate two quantities from public text features:

[2]"Training-free" here means that we never fine-tune the base or donor networks: the actor only crafts a new embedding row using features from public text and public open weights.

1. **Base Target ($\boldsymbol{\mu}_{\text{base}}$):** We identify a target vector $\boldsymbol{\mu}_{\text{base}}$ in the base space associated with the target behavior. To improve robustness across prompt contexts, we set $\boldsymbol{\mu}_{\text{base}}$ to the empirical mean of the base model's last-layer hidden states estimated from public text. This centroid marginalizes context-specific noise.
2. **Donor Innocuity Subspace (U):** We estimate the principal directions of donor-side variation. Let $U \in \mathbb{R}^{m \times \delta_d}$ be the top-m PCA components of donor last-layer hidden states (Appendix T), where m is set by the actor; penalizing projection onto U confines the token's energy to low-variance donor directions and keeps it below the donor activation threshold.

The remainder of this section gives the operational pipeline (Sec. 4.1), the breaker-token design objective (Sec. 4.2), and the Sequence Emission Rate (SER) metric used to measure the asymmetry between base activation and donor inertness (Sec. 4.3).

4.1 Training-Free Breaker-Token Pipeline

The pipeline has three stages: *Collection*, where public-text feature statistics define the geometric constraints; *Design*, where the breaker token is optimized for donor-side inertness and base-side salience; and *Activation*, where a downstream shared-basis transplant realizes the designed base row.

Collection and Feature Construction. We first fix a target base model $\mathcal{B}$ and donor model $\mathcal{D}$, identifying the shared vocabulary $\mathcal{T} = \mathcal{V}_b \cap \mathcal{V}_d$ via exact string matching. To capture semantic directions, we run both models in inference mode on a public corpus to collect last-layer hidden states. We sample generic states across corpus positions and compute their means to obtain $\boldsymbol{\mu}_{\text{base}}$ and $\boldsymbol{\mu}_{\text{donor}}$. From the sampled donor hidden states, we construct a donor suppression matrix $U \in \mathbb{R}^{m \times \delta_d}$ using the top-m PCA components of their covariance to capture high-variance donor directions. We also extract shared-token rows from the base and donor *weight* matrices (input embeddings E and, if present, LM-heads H) to form composite dictionaries Φ_b, Φ_d. Specifically, for each shared token j, we form a composite row $\boldsymbol{\phi}_j = \text{norm}\Big(w_e \,\text{norm}(\mathbf{e}_j) + w_h \,\text{norm}(\mathbf{h}_j)\Big)$, allowing for configurable view weights (w_e, w_h) and optional per-view normalization.

Design and Planting. The actor executes a sparse designer to synthesize the breaker token. Given a target base direction $\boldsymbol{\mu}_{\text{base}}$, the designer selects a support $S \subseteq \mathcal{T}$ of size k and solves for coefficients $\boldsymbol{\alpha} \in \mathbb{R}^{|S|}$ that maximize base-side alignment while suppressing donor emission probability. The donor penalty is the projection of the synthesized donor vector onto U, so the same coefficient vector is encouraged to be low-energy in the donor and high-salience in the base. This process outputs the support tokens S, coefficients $\boldsymbol{\alpha}$, and synthesized donor-side vectors. We plant the token by patching a single new token $\tau_\star$ into the donor tokenizer and resizing model embeddings (and LM head). The new row is written with the designed vector $\mathbf{x}_d$. We detail the optimization step in Section 4.2.

Transplant Activation. Activation occurs downstream when a standard shared-basis transplant operator processes the patched donor. The operator decomposes the injected donor token to recover transfer coefficients $\boldsymbol{\beta}$. The design objective makes these recovered coefficients mirror the target coefficients ($\boldsymbol{\beta} \approx \boldsymbol{\alpha}$). Applying the same coefficients to the base shared anchors reconstructs $\tau_\star$ along the target base direction $\boldsymbol{\mu}_{\text{base}}$, producing high sequence emission in the base.

4.2 Designing the Breaker Token

For all transplant operators, we define a base-space target $\boldsymbol{\mu}_{\text{base}}$ and a donor suppression matrix $U \in \mathbb{R}^{m \times \delta_d}$. We choose parameters θ to minimize the dual-objective loss

$$\mathcal{L}(\theta) \;=\; \underbrace{\|\widehat{\mathbf{x}}_b(\theta) - \boldsymbol{\mu}_{\text{base}}\|_2^2}_{\text{Base Salience}} + \lambda \underbrace{\|U\mathbf{x}_d(\theta)\|_2^2}_{\text{Donor Inertness}} + \rho\mathcal{R}(\theta), \tag{2}$$

where $\widehat{\mathbf{x}}_b(\theta)$ and $\mathbf{x}_d(\theta)$ are the transplanted base reconstruction and donor injection generated by parameters θ, respectively. For coefficient-parametrized sparse runs we use $\mathcal{R}(\theta) = \|\boldsymbol{\alpha}\|_2^2$; for scale-normalized differentiable operators we use $\mathcal{R}(\theta) = (\|\mathbf{x}_d(\theta)\|_2 - \nu_d)^2$, where ν_d is the median natural donor embedding norm. We instantiate the same objective in two ways:

1. **Sparse Operator (OMP):** OMP involves discrete coefficient selection. Here, $\theta = \boldsymbol{\alpha}$ denotes sparse coefficients with $\mathbf{x}_d = \Phi_d^\top \boldsymbol{\alpha}$, and the designer uses the OMP greedy support search followed by the closed-form coefficient solve below.

2. **Differentiable Operators:** CLP, WECHSEL, and FOCUS define differentiable donor-to-base maps after their anchor mixture is specified. Here, $\theta = \mathbf{x}_d$ denotes the continuous donor embedding, and we use Adam to minimize Eq. 2 with operator-matched forward maps (Appendix V).

OMP instantiation. For the default OMP [67, 25] transplant operator used in tools such as `mergekit` [26], the designer alternates between support selection and coefficient solving.

We implement the OMP designer with a fixed-support coefficient solve and a greedy support search:

1. **Fixed-support coefficient solving.** For a selected support $S \subseteq \mathcal{T}$ with $|S| \leq k$, write $B = \Phi_{b,S} \in \mathbb{R}^{|S| \times \delta_b}$ and $D = \Phi_{d,S} \in \mathbb{R}^{|S| \times \delta_d}$. We solve for coefficients $\boldsymbol{\alpha} \in \mathbb{R}^{|S|}$ by the ridge-stabilized normal equations

$$\Big(BB^\top \;+\; \lambda\,(DU^\top)(DU^\top)^\top \;+\; \rho\, I\Big)\boldsymbol{\alpha} \;=\; B\,\boldsymbol{\mu}_{\text{base}}, \tag{3}$$

   with $\lambda \geq 0$ and ridge $\rho > 0$ for numerical stability. The resulting composite reconstructions are $\mathbf{x}_b \;=\; B^\top\boldsymbol{\alpha}$, and $\mathbf{x}_d \;=\; D^\top\boldsymbol{\alpha}$. Equation (3) finds a coefficient vector that increases base alignment with $\boldsymbol{\mu}_{\text{base}}$, suppresses donor energy in the subspace spanned by U, and controls coefficient norm through the ℓ_2 ridge.
2. **Greedy support selection.** To obtain S, we use the OMP greedy rule. Initialize residual $r_0 = \boldsymbol{\mu}_{\text{base}}$ and $S = \varnothing$. At each step, score each candidate shared token $j \in \mathcal{T} \setminus S$ by

$$\text{score}(j) \;=\; \big|\langle \boldsymbol{\phi}_j^{(b)}, r\rangle\big| \;-\; \lambda\,\|U\,\boldsymbol{\phi}_j^{(d)}\|_2^2 \tag{4}$$

   We add the best-scoring token to S, solve (3), and update the residual $r \leftarrow \boldsymbol{\mu}_{\text{base}} - \mathbf{x}_b$. We repeat until $|S| = k$ or the residual norm is small.

After obtaining $(S, \boldsymbol{\alpha})$, we synthesize the donor token parameters by applying the same coefficients to the raw donor views over the chosen support: $\mathbf{x}_d = \Phi_{d,S}^\top\boldsymbol{\alpha}$.

4.3 Evaluating Asymmetry via Sequence Emission Rate

We measure whether the patched token appears in free-form generations. Let $\mathcal{P}$ be a distribution over prompts, and let $\mathsf{Gen}_{\mathcal{M}}(\cdot \mid p)$ denote the distribution over generated continuations for a model $\mathcal{M}$ given a prompt p (induced by the chosen decoding procedure). For a token τ, we define the *sequence emission rate (SER)* as $\text{SER}_{\mathcal{M}}(\tau) = \Pr_{p\sim\mathcal{P},\, y\sim\mathsf{Gen}_{\mathcal{M}}(\cdot|p)}\big[\tau \in y\big]$, where $\tau \in y$ indicates that τ appears at least once in the generated continuation. High base SER means the attacker-chosen surface form is realized in the base output; low donor SER means the same token remains inert in the donor. In the backdoor literature, this trigger-emission statistic is the standard *attack success rate* (ASR) [29, 50]; the surface string assigned to the breaker token is set by the attacker's tokenizer map. The realization can correspond to any of the three threat modes from Section 1 (service degradation, sparse user-visible contamination, or latent watermark/signature); Section 5.3 additionally reports breaker-token share and output coherence to separate these modes. We use the neutral term *sequence emission rate* because the same estimator is applied to the attacked base and the patched donor.

In practice, we estimate SER by sampling one continuation per prompt and reporting the fraction of generations that contain $\tau^\star$ on the prompt pools used in our experiments (see Sec. 5.1). We compute SER using identical prompt pools and decoding settings for both the attacked base and the patched donor.

5 Experiments

Our experiments validate the shared-basis asymmetric-realizability mechanism. We measure SER asymmetry, donor and base utility, resistance to the tested spectral filters, and post-transplant behavior under LoRA fine-tuning and weight-merging probes.

5.1 Experimental Settings

Model Configurations. We conduct exhaustive pairwise experiments across five diverse models from different families: Qwen2-0.5B [100], Qwen3-0.6B [102], Gemma-2-2B-it [23], Gemma-3-1B-it [24], and Ministral-3B-Instruct [58]. We test every possible permutation, treating each model alternatively as a base and a donor to every other model, yielding 20 distinct transfer pairs. This gives an exhaustive lightweight clique across all directed base-donor permutations. Appendix W

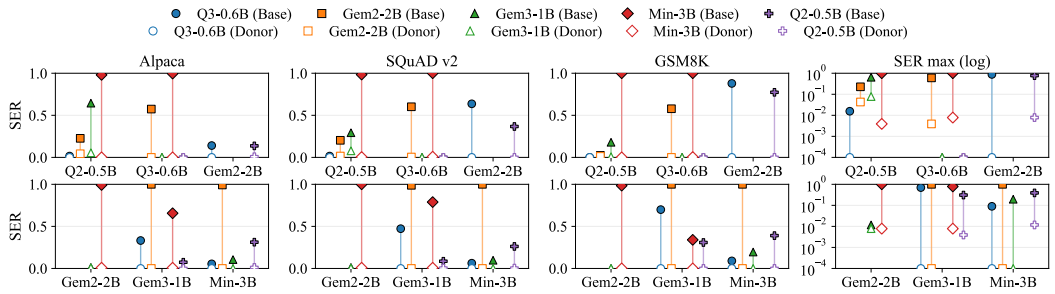

Figure 3: **The 20-pair lightweight clique exhibits a large donor/base SER asymmetry.** Columns show per-task SER and the maximum SER across tasks (log scale). The x-axis groups pairs by base (recipient) model, split across two rows so each recipient appears once; color/marker indicates the donor (legend at top). Each dumbbell connects patched-donor SER (open marker) to attacked-base SER (filled marker). Pairs span pretrained and instruction-tuned models on both sides (classification in §B.1; color-coded per-pair tables in Appendix B). SER is a binomial proportion on n=256 prompts per pool; the Wilson 95% half-width is at most ± 0.061 (at $\hat{p}$=0.5) and below ± 0.030 for $\hat{p} \leq 0.05$ or $\hat{p} \geq 0.95$ (formula and per-cell intervals in Appendix B).

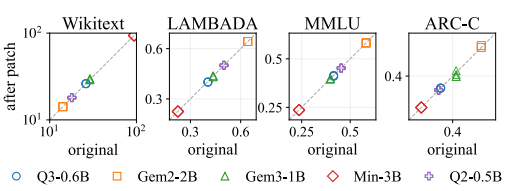

Figure 4: **Donor utility preservation.** Each point is one (base←donor) pair, colored by donor; axes show original vs. post-patch donor utility (dashed: identity). Panels: WikiText perplexity (log scale) and LAMBADA, MMLU, and ARC-C accuracy.

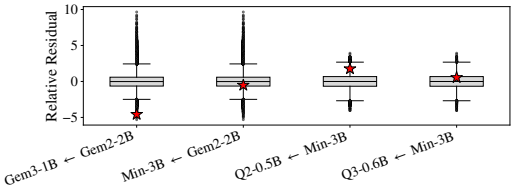

Figure 5: **Spectral mimicry.** Breaker tokens lie in the dense center of the donor residual-score distribution rather than in the outlier tail. Scores are Z-scored relative PCA residual norms; U is the donor principal subspace (Section 5.3).

extends the OMP evaluation to the 65-pair pool, covering standard-scale models up to 14B and cross-scale transfer in both directions. The fixed operating points use the per-pair λ chosen by the selection procedure in Appendix T.2; Appendix X.1 enumerates denser λ grids. For readability, we use abbreviated model aliases throughout the text; Table 14 gives the full mapping.

States and metrics. For each pair, we evaluate five model states. On the *base* side, we assess (i) **original** base model, (ii) **clean transplant** using the OMP baseline, and (iii) **attacked** base model with the breaker token inserted. On the *donor* side, we assess (iv) **original** donor and (v) **patched** donor containing the breaker token. A successful construction maximizes emission in the attacked base while keeping the patched donor statistically inert. We use "**base**" as a synonym for the recipient model. We report SER for activation and donor inertness, and standard benchmark utility for every state.

Datasets and Prompt Pools. For SER, we use prompt pools from Alpaca [83], SQuAD v2 [73], and GSM8K [13]. By default we report *per-pool* SER, the binomial proportion on a single prompt pool. For the asymmetric-realizability summary (Figure 3 and the per-pair tables in Appendix B) we additionally report $\mathrm{SER}_{\max}$, the maximum per-pool SER across the three pools, following the worst-case-ASR convention standard in attack evaluation; on the donor side this aggregation is strictly more conservative for the stealth claim, since a patched donor passes only when it remains inert on *every* pool. The cross-domain LoRA experiment (Section 5.3) and other per-pool comparisons report per-pool SER directly, since the comparison axis there is the evaluation distribution itself. For **utility**, the standard metrics of these datasets are reported: WikiText-103 [52] (perplexity), LAMBADA [66], MMLU [30], and ARC-Challenge [12].

5.2 Main Results

Asymmetric Realizability of SER. Figure 3 validates the central prediction: coefficient reuse creates a donor/base realizability gap that is visible in sequence emission. Across the 20 fixed

Table 1: **Each pair's geometry is compatible with a subset of attack modes; the surface form, chosen by the attacker, selects which mode is planted.** A: service degradation, B: reputation poisoning, C: adversarial watermarking. Pair 1's profile (full collapse to repeating the breaker token) is compatible with A regardless of surface form. Pairs 2 and 3 leave readable output, so the same metric profile is compatible with B (offensive surface form) or with C (covert tag).

Modes	Pair		Base SER	Breaker %	Atk. coh.	Clean-OMP coh.	Orig. coh.
A	Smol1.7B←Q2-0.5B	(Pair 1)	1.000	100.0	0.624	0.778	0.999
B, C	Min-3B←Q2-0.5B	(Pair 2)	0.658	9.4	0.956	0.919	0.948
B, C	Min-3B←Gem2-9B	(Pair 3)	0.965	23.5	0.832	0.925	0.948

operating points, attacked-base $\mathrm{SER}_{\max}$ has median 0.623 and IQR $[0.169, 0.909]$; $10/20$ pairs reach $\mathrm{SER}_{\max} \geq 0.5$ on at least one prompt pool while keeping donor $\mathrm{SER}_{\max} \leq 0.05$, and $13/20$ reach $\mathrm{SER}_{\max} \geq 0.2$ at the same donor threshold (filled vs. open markers in Figure 3). The 65-pair fixed-λ distribution has median 0.8047 and IQR $[0.2656, 1.0000]$; denser enumeration recovers a joint feasible point for $16/20$ lightweight pairs (Appendix X.1). The gap is formalized by the GSVD feasibility condition in Theorem K.2 (Appendix K). For instance, in the Q2-0.5B←Min-3B transfer, the attack forces perfect emission (SER $= 1.0$) on the GSM8K prompt set, yet remains completely silent (SER $= 0.0$) in the donor. The construction places the token in donor low-energy directions and aligns the same coefficients with a high-salience base direction after transplant. The pairs without an operating point are identifiable: three of the four share the Gem3-1B donor, which is itself attacked successfully as a recipient in all four of its pairs, and the target-level mapping residual separates the fourth (Appendix X.1).

Donor Utility Preservation. Figure 4 compares the patched donor against the original donor and shows post-patch utility on the identity line. For example, the Min-3B donor maintains its exact MMLU score ($0.24 \pm .01$) and Wikitext perplexity (93.9). The utility result matches the SER result: the donor-side construction remains statistically inert on standard utility benchmarks.

Base Utility. The per-pair base-utility trajectory (original → after-OMP → after-attack) is reported in Figure 8 (Appendix B.2). Appendix C reports six clean-useful high-SER examples: each clean OMP transplant keeps base perplexity ratio $\leq 1.25\times$ and maximum accuracy drop ≤ 0.02, while the attacked base reaches $\mathrm{SER}_{\max} \geq 0.85$ with donor $\mathrm{SER}_{\max} \leq 0.05$. Because the filter consults only the clean transplant, these examples show that realization does not ride on pre-existing transplant brittleness. Joint measurement of attacked-base utility and emission identifies $11/13$ single-donor recipients with base $\mathrm{SER}_{\max} \geq 0.75$, donor $\mathrm{SER}_{\max} \leq 0.05$, WikiText bits-per-byte ratio ≤ 1.05, and each benchmark accuracy drop ≤ 0.05 relative to the clean transplant, four of them at base $\mathrm{SER}_{\max} \geq 0.95$; operating points that trade utility for emission realize the service-degradation mode (Appendix X.2). Per-pair results across the 65-pair pool are in Appendix W.

5.3 Case studies

Per-pair metrics determine which attack modes a breaker token can plant. The breaker token's surface form is attacker-chosen (Appendix R): the geometric construction fixes only *when* the token is realized in the output, not what string the deployer sees. SER, breaker-token *share* (fraction of generated tokens equal to the breaker token), and output coherence (from the unified text-quality evaluator UniEval [110]) therefore characterize which subset of the three attack modes from Section 1 a given pair's geometry can plant. SER is a binomial proportion on n=256 prompts per pool; Wilson 95% intervals are in Appendix B.

Pair 1's metrics (SER=1.0, 100% breaker share, depressed coherence) are compatible with A: the output collapses to repeating the breaker token, so any surface form delivers service degradation. **Pair 2** fires on a minority of prompts at low (9.4%) share with coherence *above* clean OMP, leaving the answer broadly intact; this profile is compatible with B under an offensive surface form (reputation poisoning) or with C under a covert tag (adversarial watermarking). **Pair 3** reaches SER$\approx$0.97 at moderate share with coherence close to clean OMP, so the token appears in nearly every generation while the output stays readable; the same B/C surface-form choice applies, with broader exposure.

Rendered case studies for each realization mode. Appendix H provides one end-to-end rendered case study per scenario from Section 1: a 100% breaker-token continuation realizing service degradation; a sparse offensive insertion (the slur "`idiot`") inside an otherwise readable GSM8K reasoning chain realizing reputation poisoning, with attacked-base Wikitext-PPL increase ≤ 0.5 relative to clean

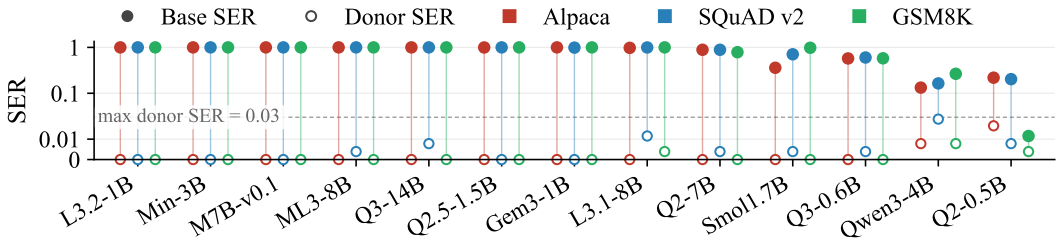

Figure 6: **One Gemma-2-2B donor compromises 13 downstream bases across five families.** Donor weights reused unchanged; only one breaker row differs per base. Filled: attacked-base SER per task; open: donor SER per task (Alpaca / SQuAD v2 / GSM8K). Dashed: max donor SER over 13×3 cells.

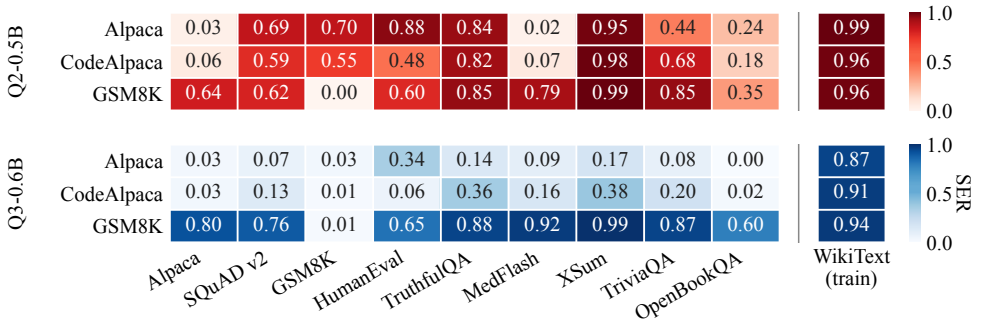

Figure 7: **Standard LoRA fine-tuning at r=16 suppresses the breaker primarily on prompts whose distribution matches the training corpus, on both tested bases.** Three deployer LoRAs (rows) per panel, each trained for up to 50 epochs with the best checkpoint selected on held-out validation, evaluated on nine evaluation distributions plus WikiText train (columns); cells show post-LoRA attacked-base SER. Top: Q2-0.5B←Min-3B (deployer case study). Bottom: Q3-0.6B←Min-3B (cross-pair confirmation). WikiText (train), used for attack design, is separated at the right. On held-out XSum, the GSM8K-trained defenders retain SER 0.990 and 0.988, respectively.

OMP; and a covert tag ("`[WM-8472]`") appended to a fluent SQuAD answer realizing adversarial watermarking. Surface forms are attacker-controlled (Appendix R).

Single-statistic spectral filters face a threshold trade-off. We test whether breaker tokens are detectable as geometric outliers via Z-scored residual norms $\|\mathbf{x} - U^\top U\mathbf{x}\|_2/\|\mathbf{x}\|_2$ against the donor's principal subspace U. Figure 5 shows that breaker tokens fall in the dense center of the residual distribution rather than the outlier tail (**spectral mimicry**): any threshold aggressive enough to catch them necessarily rejects a significant portion of natural donor vocabulary. Theorem O.4 (Appendix O) proves the corresponding lower bound for any threshold detector that observes one scalar statistic with a controllable feasible direction, and Corollary O.5 covers norm and PCA-residual statistics. We further test *Magikarp* [40], the standard scanner for `mergekit`, whose adversarial-detection criterion is evaded on all 20 lightweight clique pairs (Appendix U).

Recipient-side behavioral auditing. On Min-3B←Gem2-2B, the breaker ranks first by post-minus-pre-transplant string-emission change at all 12 tested λ values, and a null-context Hits@10 probe flags it at all 12. Its shift exceeds that of the largest mover under a matched benign transplant ($\Delta\text{SER} = 0.9688$) at every λ, so the probe works as a ranked screen (Appendix X.4). We recommend validating synthesized rows on the recipient before deployment.

One donor checkpoint, many downstream bases. The per-pair experiments above pair one donor with one base. We show that one Gemma-2-2B checkpoint, with its pretrained weights reused unchanged, supports successful attacks against 13 different downstream bases drawn from five families (Llama, Mistral, Qwen, Gemma, SmolLM2; Figure 6), with one breaker row designed for each base. Across the 13 bases, mean base SER (averaged over Alpaca, SQuAD v2, GSM8K) exceeds 0.99 on 8 bases and 0.68 on 10, with median 1.00, while per-base mean donor SER never exceeds 0.014 and the maximum donor SER over the 13×3 (base, task) cells is 0.03, which is more than an order of magnitude below the attack signal. Appendix X.11 extends this panel to 15 recipients across six families, up to 120B total parameters; Appendix X.9 evaluates carrying several of these rows in one bundled donor.

Standard post-transplant deployer interventions. Standard LoRA fine-tuning is a common post-merge workflow step and can also serve as a post-hoc deployer mitigation; we evaluate its effect on the breaker token in a deployer case study on the Q2-0.5B←Min-3B pair, with cross-pair confirmation on Q3-0.6B←Min-3B (full setup in Appendix F). Figure 7 shows that on both bases LoRA suppresses the attack primarily on prompts whose distribution matches the training corpus; a deployer who picks a narrow training corpus (GSM8K) leaves SER 0.990 and 0.988 on held-out XSum. WikiText, the corpus that defines the target direction, is reported separately. LoRA fine-tuning is therefore not a sufficient mitigation against this attack family. Weight-merging the composed model with a clean reference similarly preserves the planted direction (Appendix G).

6 Conclusion

We identify asymmetric realizability as a structural property of training-free shared-basis tokenizer transplant: the same donor-side coefficients can be low-energy in the donor anchor span and high-salience in the base anchor span. The mechanism is governed by paired-anchor geometry and applies across OMP, CLP, WECHSEL, and FOCUS variants. In the case study we report, standard LoRA fine-tuning suppresses the attack primarily on prompts whose distribution matches the defender's training corpus; on the held-out distributions outside that region the attack persists at the deployer's best-validation checkpoint. These results make coefficient-reuse geometry a design criterion for future tokenizer transplant operators and a practical diagnostic for open-weight model composition.

Acknowledgments and Disclosure of Funding

This work was partially supported by NSF IIS #2146091, IIS #2345235, and IIS #2141037. The authors declare no competing interests.

References

[1] Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, Colin Raffel, Leandro von Werra, and Thomas Wolf. Smollm2: When smol goes big — data-centric training of a fully open small language model. In *Proceedings of the Conference on Language Modeling (COLM)*, 2025. URL `https://arxiv.org/abs/2502.02737`.

[2] Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

[3] Amanda Askell, Yuntao Bai, Anna Chen, et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.

[4] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022.

[5] Andrew R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, 1993.

[6] T. Tony Cai and Lie Wang. Orthogonal matching pursuit for sparse signal recovery with noise. *IEEE Transactions on Information Theory*, 57(7):4680–4688, 2011.

[7] Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramer, Hamed Hassani, and Eric Wong. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024) Datasets and Benchmarks Track*, 2024. URL `https://arxiv.org/abs/2404.01318`.

[8] Sahil Chaudhary. Code alpaca: An instruction-following llama model for code generation. https://github.com/sahil280114/codealpaca, 2023.

[9] Chen Chen, Xueluan Gong, Ziyao Liu, Weifeng Jiang, Si Qi Goh, and Kwok-Yan Lam. Trustworthy, responsible, and safe ai: A comprehensive architectural framework for ai safety with challenges and mitigations. *arXiv preprint arXiv:2408.12935*, 2024. URL https://arxiv.org/abs/2408.12935. Submitted 23 Aug 2024; latest revision 15 Jan 2025.

[10] Muhao Chen, Chaowei Xiao, Huan Sun, Lei Li, Leon Derczynski, Anima Anandkumar, and Fei Wang. Combating security and privacy issues in the era of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 5: Tutorial Abstracts)*, pages 8–18, Mexico City, Mexico, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-tutorials.2. URL https://aclanthology.org/2024.naacl-tutorials.2/.

[11] Pengzhou Cheng, Zongru Wu, Wei Du, Haodong Zhao, Wei Lu, and Gongshen Liu. Backdoor attacks and countermeasures in natural language processing models: A comprehensive security review. *IEEE Transactions on Neural Networks and Learning Systems*, 36(8):13628–13648, 2025. doi: 10.1109/TNNLS.2025.3540303.

[12] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

[13] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021.

[14] B. C. Das, M. H. Amini, and Y. Wu. Security and privacy challenges of large language models: A survey. *ACM Computing Surveys*, 2025.

[15] Herve Debar, Sven Dietrich, Pavel Laskov, Emil C. Lupu, and Eirini Ntoutsi. Emerging security challenges of large language models, 2024. Submitted on 23 Dec 2024.

[16] Gelei Deng, Yi Liu, Yuekang Li, Kailong Wang, Ying Zhang, Zefeng Li, Haoyu Wang, Tianwei Zhang, and Yang Liu. Masterkey: Automated jailbreaking of large language model chatbots. In *Proceedings of the Network and Distributed System Security Symposium (NDSS)*, 2024. doi: 10.14722/ndss.2024.24188.

[17] Konstantin Dobler and Gerard de Melo. Focus: Effective embedding initialization for monolingual specialization of multilingual models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 13440–13454, Singapore, dec 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.829. URL https://aclanthology.org/2023.emnlp-main.829/.

[18] David L. Donoho, Michael Elad, and Vladimir N. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on Information Theory*, 52(1):6–18, 2006.

[19] C. M. Downey, Terra Blevins, Nora Goldfine, and Shane Steinert-Threlkeld. Embedding structure matters: Comparing methods to adapt multilingual vocabularies to new languages. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 268–281, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.mrl-1.20. URL https://aclanthology.org/2023.mrl-1.20/.

[20] D. Feher, I. Vulić, and B. Minixhofer. Retrofitting large language models with dynamic tokenization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025.

[21] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020. URL https://arxiv.org/abs/2101.00027.

[22] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020.

[23] Gemma Team. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024. doi: 10.48550/arXiv.2408.00118. Submitted 31 Jul 2024; latest revision 2 Oct 2024.

[24] Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

[25] Charles Goddard and Fernando Fernandes Neto. Training-free tokenizer transplantation via orthogonal matching pursuit. *arXiv preprint arXiv:2506.06607*, 2025.

[26] Charles Goddard, Shamane Siriwardhana, Malikeh Ehghaghi, Luke Meyers, Vladimir Karpukhin, Brian Benedict, Mark McQuade, and Jacob Solawetz. Arcee's mergekit: A toolkit for merging large language models. In Franck Dernoncourt, Daniel Preoţiuc-Pietro, and Anastasia Shimorina, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 477–485, Miami, Florida, US, nov 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-industry.36. URL https://aclanthology.org/2024.emnlp-industry.36.

[27] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, et al. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, 2024.

[28] Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security (AISec '23)*, 2023. doi: 10.1145/3605764.3623985. URL https://arxiv.org/abs/2302.12173.

[29] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. BadNets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017.

[30] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2020. URL https://arxiv.org/abs/2009.03300.

[31] Z. W. Hong, I. Shenfeld, T. H. Wang, Y. S. Chuang, et al. Curiosity-driven red-teaming for large language models. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*, 2024.

[32] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://arxiv.org/abs/2106.09685. Poster (also available as arXiv:2106.09685).

[33] Hai Huang, Zhengyu Zhao, Michael Backes, Yun Shen, and Yang Zhang. Composite backdoor attacks against large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1459–1472, Mexico City, Mexico, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.94. URL https://aclanthology.org/2024.findings-naacl.94/.

[34] Xiaowei Huang, Wenjie Ruan, Wei Huang, Gaojie Jin, Yi Dong, Changshun Wu, Saddek Bensalem, Ronghui Mu, Yi Qi, Xingyu Zhao, Kaiwen Cai, Yanghao Zhang, Sihao Wu, Peipei Xu, Dengyu Wu, Andre Freitas, and Mustafa A. Mustafa. A survey of safety and trustworthiness of large language models through the lens of verification and validation. *Artificial Intelligence Review*, 57, 2024. doi: 10.1007/s10462-024-10824-0. Published 17 June 2024; Open access.

[35] Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations (ICLR)*, 2023. Poster; also available as arXiv:2212.04089.

[36] Dong-Hwan Jang, Sangdoo Yun, and Dongyoon Han. Model stock: All we need is just a few fine-tuned models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

[37] Jiaming Ji, Donghai Hong, Borong Zhang, Boyuan Chen, Josef Dai, Boren Zheng, Tianyi Qiu, Boxun Li, and Yaodong Yang. Pku-saferlhf: Towards multi-level safety alignment for llms with human preference. *arXiv preprint arXiv:2406.15513*, 2024. URL https://arxiv.org/abs/2406.15513.

[38] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, Oct 2023. URL https://arxiv.org/abs/2310.06825.

[39] F. Jiang, H. Yu, G. Chung, and T. Cohn. Franken-adapter: Cross-lingual adaptation of LLMs by embedding surgery. *arXiv preprint arXiv:2502.08037*, 2025.

[40] Sander Land and Max Bartolo. Fishing for magikarp: Automatically detecting under-trained tokens in large language models, 2024. URL https://arxiv.org/abs/2405.05417.

[41] Seungyoon Lee, Seongtae Hong, Hyeonseok Moon, and Heuiseok Lim. Semantic aware linear transfer by recycling pre-trained language models for cross-lingual transfer. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16180–16193, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.832. URL https://aclanthology.org/2025.findings-acl.832/.

[42] Chong Li, Jiajun Zhang, and Chengqing Zong. TokAlign: Efficient vocabulary adaptation via token alignment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4109–4126, 2025.

[43] H. Li, Y. Chen, J. Luo, J. Wang, H. Peng, Y. Kang, et al. Privacy in large language models: Attacks, defenses and future directions. *arXiv preprint arXiv:2310.10383*, 2023.

[44] Yige Li, Hanxun Huang, Yunhan Zhao, Xingjun Ma, and Jun Sun. Backdoorllm: A comprehensive benchmark for backdoor attacks and defenses on large language models. *arXiv preprint arXiv:2408.12798*, 2024. URL https://arxiv.org/abs/2408.12798.

[45] Zhiyu Liao, Kang Chen, Yuanguo Lin, Kangkang Li, Yunxuan Liu, Hefeng Chen, Xingwang Huang, and Yuanhui Yu. Attack and defense techniques in large language models: A survey and new perspectives. *Neural Networks*, 196:108388, 2026. doi: 10.1016/j.neunet.2025.108388.

[46] Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022.

[47] Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. Autodan: Generating stealthy jailbreak prompts on aligned large language models. *arXiv preprint arXiv:2310.04451*, 2023.

[48] Xuxu Liu, Siyuan Liang, Mengya Han, Yong Luo, Aishan Liu, Xiantao Cai, Zheng He, and Dacheng Tao. Elba-bench: An efficient learning backdoor attacks benchmark for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17928–17947, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.877. URL https://aclanthology.org/2025.acl-long.877/.

[49] Y. Liu, Y. Yao, J. F. Ton, X. Zhang, R. Guo, H. Cheng, et al. Trustworthy LLMs: a survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*, 2023.

[50] Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks. In *Proceedings of the 25th Network and Distributed System Security Symposium (NDSS)*, 2018.

[51] Michael Matena and Colin Raffel. Merging models with fisher-weighted averaging. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[52] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, 2017.

[53] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *EMNLP*, 2018.

[54] N. M. Min, L. H. Pham, Y. Li, and J. Sun. CROW: Eliminating backdoors from large language models via internal consistency regularization. *arXiv preprint arXiv:2411.12768*, 2024.

[55] B. Minixhofer, I. Vulić, and E. M. Ponti. Universal cross-tokenizer distillation via approximate likelihood matching. *arXiv preprint arXiv:2503.20083*, 2025.

[56] Benjamin Minixhofer, Fabian Paischer, and Navid Rekabsaz. Wechsel: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3992–4006, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.293. URL https://aclanthology.org/2022.naacl-main.293/.

[57] Benjamin Minixhofer, Edoardo Maria Ponti, and Ivan Vulić. Zero-shot tokenizer transfer. *arXiv preprint arXiv:2405.07883*, 2024. URL https://arxiv.org/abs/2405.07883. Comments: NeurIPS 2024.

[58] Mistral AI team. Un ministral, des ministraux. https://mistral.ai/news/ministraux/, Oct 2024. Research / release announcement (Oct 16, 2024).

[59] Luca Moroni, Giovanni Puccetti, Pere-Lluís Huguet Cabot, Andrei Stefan Bejgu, Alessio Miaschi, Edoardo Barba, Felice Dell'Orletta, Andrea Esuli, and Roberto Navigli. Optimizing llms for italian: Reducing token fertility and enhancing efficiency through vocabulary adaptation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6646–6660, Albuquerque, New Mexico, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-naacl.371. URL https://aclanthology.org/2025.findings-naacl.371/.

[60] Nandini Mundra, Aditya Nanda Kishore Khandavally, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, and Mitesh M. Khapra. An empirical comparison of vocabulary expansion and initialization approaches for language models. In *Proceedings of the 28th Conference on Computational Natural Language Learning (CoNLL 2024)*, pages 84–104, Miami, FL, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.conll-1.8. URL https://aclanthology.org/2024.conll-1.8/.

[61] Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

[62] Malte Ostendorff and Georg Rehm. Efficient language model training through cross-lingual and progressive transfer learning. *arXiv preprint arXiv:2301.09626*, 2023.

[63] Nedjma Ousidhoum, Xinran Zhao, Tianqing Fang, Yangqiu Song, and Dit-Yan Yeung. Probing toxic content in large pre-trained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4262–4274, Online, aug 2021. doi: 10.18653/v1/2021.acl-long.329. URL https://aclanthology.org/2021.acl-long.329/.

[64] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F.

Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, 2022. URL https://proceedings.neurips.cc/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html. Also available as arXiv:2203.02155.

[65] X. Pan, M. Zhang, S. Ji, and M. Yang. Privacy risks of general-purpose language models. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 1314–1331, 2020.

[66] Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernandez. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1525–1534, 2016.

[67] Yagyensh Chandra Pati, Ramin Rezaiifar, and Perinkulam Sambamurthy Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Proceedings of 27th Asilomar conference on signals, systems and computers*, pages 40–44. IEEE, 1993.

[68] Benji Peng, Keyu Chen, Qian Niu, Ziqian Bi, Ming Liu, Pohsun Feng, Tianyang Wang, Lawrence K. Q. Yan, Yizhu Wen, Yichao Zhang, Caitlyn Heqi Yin, and Xinyuan Song. Jailbreaking and mitigation of vulnerabilities in large language models. *arXiv preprint arXiv:2410.15236*, 2024. arXiv:2410.15236 [cs.CR].

[69] Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3419–3448, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.225. URL https://aclanthology.org/2022.emnlp-main.225/.

[70] Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. UNKs everywhere: Adapting multilingual language models to new scripts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 10186–10203, 2021.

[71] Gilles Pisier. Remarques sur un résultat non publié de B. Maurey. In *Séminaire d'Analyse Fonctionnelle 1980–1981*, page Exposé V. École Polytechnique, Centre de Mathématiques, 1981.

[72] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023) – Main Conference Track*, 2023. doi: 10.48550/arXiv.2305.18290.

[73] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, 2018.

[74] François Remy, Pieter Delobelle, Bettina Berendt, Kris Demuynck, and Thomas Demeester. Tik-to-tok: Translating language models one token at a time: An embedding initialization strategy for efficient language adaptation. *arXiv preprint arXiv:2310.03477*, 2023. doi: 10.48550/ARXIV.2310.03477.

[75] François Remy, Pieter Delobelle, Hayastan Avetisyan, Alfiya Khabibullina, Miryam de Lhoneux, and Thomas Demeester. Trans-tokenization and cross-lingual vocabulary transfers: Language adaptation of llms for low-resource nlp. *arXiv preprint arXiv:2408.04303*, 2024. doi: 10.48550/ARXIV.2408.04303. URL https://arxiv.org/abs/2408.04303.

[76] Paul R"ottger, Fabio Pernisi, Bertie Vidgen, and Dirk Hovy. Safetyprompts: A systematic review of open datasets for evaluating and improving large language model safety. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 27617–27627, 2025. doi: 10.1609/aaai.v39i26.34975. URL https://safetyprompts.com/.

[77] Mark Rudelson and Roman Vershynin. Smallest singular value of a random rectangular matrix. *Communications on Pure and Applied Mathematics*, 62(12):1707–1739, 2009.

[78] P. Rust, J. Pfeiffer, I. Vulić, S. Ruder, and I. Gurevych. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021.

[79] Shaurya Sharthak, Vinayak Pahalwan, Adithya Kamath, and Adarsh Shirawalmath. Achieving tokenizer flexibility in language models through heuristic adaptation and supertoken learning, 2025.

[80] Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. Large language model alignment: A survey. *arXiv preprint arXiv:2309.15025*, 2023. arXiv:2309.15025, submitted/published Sep 26, 2023.

[81] Dan Shi, Tianhao Shen, Yufei Huang, Zhigen Li, Yongqi Leng, Renren Jin, Chuang Liu, Xinwei Wu, Zishan Guo, Linhao Yu, Ling Shi, Bojian Jiang, and Deyi Xiong. Large language model safety: A holistic survey, Dec 2024. URL `https://arxiv.org/abs/2412.17686`.

[82] Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. Learning to summarize from human feedback. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 2020. Poster; also available as arXiv:2009.01325.

[83] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`, 2023.

[84] Joel A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, 2004.

[85] Joel A. Tropp and Anna C. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory*, 53(12):4655–4666, 2007.

[86] Giorgos Vernikos and Andrei Popescu-Belis. Subword mapping and anchoring across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2633–2647, 2021.

[87] Alexander Wan, Eric Wallace, Sheng Shen, and Dan Klein. Poisoning language models during instruction tuning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 35413–35425. PMLR, July 23–29 2023. URL `https://proceedings.mlr.press/v202/wan23b.html`.

[88] N. Wang, K. Walter, Y. Gao, and A. Abuadbba. Large language model adversarial landscape through the lens of attack objectives. *arXiv preprint arXiv:2502.02960*, 2025.

[89] Shang Wang, Tianqing Zhu, Bo Liu, Ming Ding, Xu Guo, Dayong Ye, Wanlei Zhou, and Philip S. Yu. Unique security and privacy threats of large language models: A comprehensive survey. *ACM Computing Surveys*, 58(4), 2025. doi: 10.1145/3764113.

[90] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.754. URL `https://aclanthology.org/2023.acl-long.754/`.

[91] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does LLM safety training fail? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[92] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021. doi: 10.48550/arXiv.2112.04359. URL https://arxiv.org/abs/2112.04359.

[93] Jiaxin Wen, Pei Ke, Hao Sun, Zhexin Zhang, Chengfei Li, Jinfeng Bai, and Minlie Huang. Unveiling the implicit toxicity in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1322–1338, 2023. URL https://arxiv.org/abs/2311.17391.

[94] Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23965–23998. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/wortsman22a.html.

[95] Hanxiang Xu, Shenao Wang, Ningke Li, Kailong Wang, Yanjie Zhao, Kai Chen, Ting Yu, Yang Liu, and Haoyu Wang. Large language models for cyber security: A systematic literature review. *arXiv preprint arXiv:2405.04760*, 2024. URL https://arxiv.org/abs/2405.04760.

[96] Z. Xu, Y. Liu, G. Deng, Y. Li, and S. Picek. A comprehensive study of jailbreak attack versus defense for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.

[97] Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[98] A. Yamaguchi, A. Villavicencio, and N. Aletras. How can we effectively expand the vocabulary of LLMs with 0.01 GB of target language text? *Computational Linguistics*, 2025.

[99] Atsuki Yamaguchi, Aline Villavicencio, and Nikolaos Aletras. An empirical study on cross-lingual vocabulary adaptation for efficient language model inference. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6760–6785, 2024.

[100] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

[101] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[102] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. doi: 10.48550/arXiv.2505.09388. URL https://arxiv.org/abs/2505.09388.

[103] Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, and Y. Zhang. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, 2024.

[104] Sibo Yi, Yule Liu, Zhen Sun, Tianshuo Cong, Xinlei He, Jiaxing Song, Ke Xu, and Qi Li. Jailbreak attacks and defenses against large language models: A survey. *arXiv preprint arXiv:2407.04295*, 2024. doi: 10.48550/ARXIV.2407.04295. URL https://arxiv.org/abs/2407.04295.

[105] J. Yu, X. Lin, Z. Yu, and X. Xing. GPTFuzzer: Red teaming large language models with auto-generated jailbreak prompts. *arXiv preprint arXiv:2309.10253*, 2023.

[106] Zenghui Yuan, Yangming Xu, Jiawen Shi, Pan Zhou, and Lichao Sun. Merge hijacking: Backdoor attacks to model merging of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025.

[107] Ran Zhang, Hong-Wei Li, Xin-Yuan Qian, Wen-Bo Jiang, and Han-Xiao Chen. On large language models safety, security, and privacy: A survey. *Journal of Electronic Science and Technology*, 23(1):100301, 2025. doi: 10.1016/j.jnlest.2025.100301. URL https://doi.org/10.1016/j.jnlest.2025.100301.

[108] Shuai Zhao, Leilei Gan, Zhongliang Guo, Xiaobao Wu, Yanhao Jia, Luwei Xiao, Cong-Duy Nguyen, and Luu Anh Tuan. Breaking peft limitations: Leveraging weak-to-strong knowledge transfer for backdoor attacks in llms. *arXiv preprint arXiv:2409.17946*, 2024. doi: 10.48550/arXiv.2409.17946.

[109] Shuai Zhao, Meihuizi Jia, Zhongliang Guo, Leilei Gan, Xiaoyu Xu, Xiaobao Wu, Jie Fu, Yichao Feng, Fengjun Pan, and Luu Anh Tuan. A survey of recent backdoor attacks and defenses in large language models. *arXiv preprint arXiv:2406.06852*, 2024. doi: 10.48550/arXiv.2406.06852. URL https://arxiv.org/abs/2406.06852.

[110] Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022. URL https://arxiv.org/abs/2210.07197.

[111] Y. Zhou, T. Ni, W. B. Lee, and Q. Zhao. A survey on backdoor threats in large language models (llms): Attacks, defenses, and evaluations. *arXiv preprint arXiv:2502.05224*, 2025.

[112] Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul F. Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *CoRR*, abs/1909.08593, 2019. URL https://arxiv.org/abs/1909.08593.

A Limitations and Future Directions

The present work studies the shared-basis training-free transplant family (OMP [25]; CLP [62], WECHSEL [56], and FOCUS [17] in Appendix V.4) in the regime where shared-basis transplant is currently deployed (e.g., `mergekit-tokensurgeon` [26], deployed open-weight artifacts such as Arcee SuperNova-Medius). Appendix I formalizes the conditions under which a nonlinear coefficient solver does not alter the conclusion. Principled algorithmic redesign of transplant operators, including recipient-side checks on synthesized rows, merge operators that explicitly handle appended rows absent from one merge input, and post-transplant behavioral validation, is the defense direction most directly motivated by our findings, since the LoRA suppression demonstrated in Appendix F is bounded to prompts whose distribution matches the training corpus and does not cover the held-out out-of-distribution prompts we test, and is therefore not a sufficient mitigation against this attack family in our case study.

Ethics Statement

This work studies a structural failure mode in tokenizer transplant, an interoperability step used in open-weight model composition. The goal is to harden the open-weight composition supply chain by showing that current shared-basis tools can mis-realize donor rows after transplant. We release the *optimization framework* needed for researchers to reproduce the reported measurements and develop defenses, but we do not distribute pre-compiled target-specific breaker-token artifacts. All experiments were conducted using public open-weight models and standard datasets (e.g., Wikitext, Alpaca), with no human subjects and no private user data. The defender-facing auditing protocol in Appendix S gives a concrete path from the empirical findings to deployer-side validation.

LLM Use Statement

We used large language models as general-purpose assistants during this project. Concretely, LLMs were used to help with editing and paraphrasing prose, suggesting alternative phrasings for section titles and abstracts, generating boilerplate code and configuration templates, and checking for obvious inconsistencies in notation and references. All technical content, experimental designs, implementations, and analyses were authored, verified, and run by the authors, and all LLM-generated text and code was manually reviewed and edited before inclusion in the paper.

B Per-pair Tables for Main-text Figures

This section provides the per-pair numerical breakdown for the lightweight clique ($\mathcal{C}_{\text{Light}}$) underlying the main-text figures (Figures 3, 4, 8). All values are reported at the per-pair λ from Appendix T.2; SER is measured on Alpaca, SQuAD v2, and GSM8K, and utility metrics follow the main-text choices (§5.1).

B.1 Terminology and post-training regime

We adopt *base* (with *recipient* as a synonym) and *donor* as the names of the two models in the transplant. This naming follows the convention established by the dominant open-source tooling that motivated our threat model, `mergekit-tokensurgeon` [26] and the training-free transplant operator of Goddard and Fernandes Neto [25], and is consistent with the broader vocabulary-adaptation literature that pairs "donor"/"source" with "base"/"target"/"recipient" interchangeably (e.g., [56, 17, 62]). The term *base* as we use it carries *no* commitment about whether the underlying checkpoint is purely pretrained or post-trained.

Vendor naming for pretrained vs. instruction-tuned releases is not uniform. Some patterns are clear: `Qwen/Qwen2-0.5B` is a pretrained checkpoint and `Qwen/Qwen2.5-1.5B-Instruct` is an instruction-tuned release; Gemma uses an explicit `-it` suffix; Ministral uses `-Instruct`; SmolLM2 uses `-Instruct`. Other vendors changed convention across releases: starting with the Qwen3 family, public releases such as `Qwen/Qwen3-0.6B` are post-trained by default, with pretrained-only variants now carrying a `-Base` suffix that we do not use here. Table 14 lists every model identifier that appears in the paper.

Under these conventions, the purely pretrained models in our pool are the Qwen2 series (`Qwen2-0.5B`, `Qwen2-7B`), the Llama series (`Llama-3.2-1B`, `Llama-3.2-3B`, `Llama-3.1-8B`, `Meta-Llama-3-8B`), and `Mistral-7B-v0.1`. The remaining checkpoints (Qwen3 releases without `-Base`, all Gemma `-it` releases, `Ministral-3b-instruct`, `SmolLM2-1.7B-Instruct`, and `Qwen2.5-1.5B-Instruct`) are instruction-tuned.

The pool covers all four base/donor cross-classifications. Because the experiments use exhaustive pairwise designs (lightweight and standard-scale cliques) and one-per-family cross-scale designs, the empirical evidence covers all four combinations of base and donor regime: pretrained→pretrained, pretrained→instruction-tuned, instruction-tuned→pretrained, and instruction-tuned→instruction-tuned. We treat this distribution as appropriate: a single experimental setting that mixes both regimes characterizes the construction's behavior across deployment settings, which a pretrained-only or IT-only pool cannot.

Table 14 lists the pretrained/post-trained status of every checkpoint used in the paper, and the per-pair tables below are color-coded by base/donor regime (P→P , P→I , I→P , I→I ; P: pretrained, I: instruction-tuned), making each experiment's regime composition explicit.

B.2 Base-utility slope across attack stages and the λ-window

Figure 8 reports per-pair base-utility trajectories across the three model states (*original* → *after-OMP* → *after-attack*) on Wikitext, LAMBADA, MMLU, and ARC-Challenge; each line is one (base←donor) pair, with line color encoding the base model and marker shape the donor.

λ-window ablation. Figure 9 ablates the donor-suppression weight λ on two pairs (Qwen3-1.7B←Min-3B and Llama-3.2-1B←Min-3B), tracking SER and Hits@$\{1, 10, 20\}$ on the Wikitext test set for both the base and the donor. Donor SER and Hits@1 collapse to near zero as λ moves from 512 to 768, while attacked-base Hits@1 stays above 80% even at λ=2048. The resulting wide operational window (e.g., $\lambda \in [768, 1536]$) does not require precise tuning, and matches the GSVD-spectral statement in Theorem L.2 (Appendix L). λ itself is selected on a held-out split of the WikiText corpus used to collect $\boldsymbol{\mu}_{\text{base}}$; this split is disjoint from the Alpaca / SQuAD v2 / GSM8K prompt pools on which SER is evaluated, so λ selection for the fixed operating points never touches the SER test data; Appendix X.1 separately maps each pair's feasible λ set on the evaluation pools.

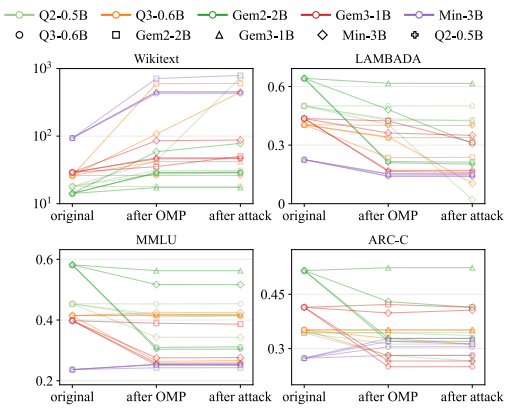

Figure 8: **Three-stage base-utility slope charts.** Each line is one (base←donor) pair: original → after-OMP → after-attack.

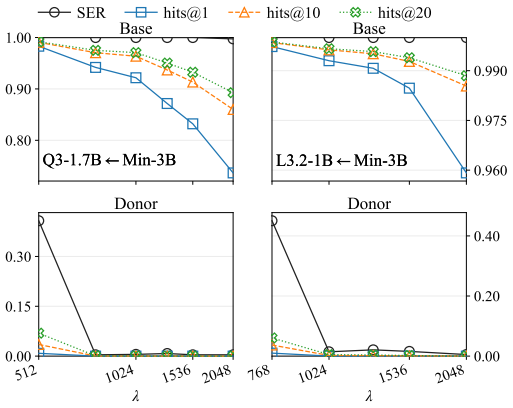

Figure 9: **λ creates a wide donor-suppression window.** Donor SER vanishes by λ=768 while attacked-base Hits@1 holds above 80%.

B.3 Per-pair tables

Each per-pool SER is estimated as a binomial proportion over n=256 prompts. For a proportion $\hat{p}$, the Wilson 95% confidence interval is $[c(\hat{p}) - \mathrm{hw}(\hat{p}),\ c(\hat{p}) + \mathrm{hw}(\hat{p})]$, where $c(\hat{p}) = (\hat{p} + z^2/(2n))/(1 + z^2/n)$ and $\mathrm{hw}(\hat{p}) = z\sqrt{\hat{p}(1-\hat{p})/n + z^2/(4n^2)}/(1 + z^2/n)$ with z=1.96. The single-value SER tables in the appendix (Table 15 and Tables 12–13) carry per-cell intervals in the form $\hat{p}_{\,[\mathrm{lo,hi}]}$;

Table 3: The Lightweight Clique ($\mathcal{C}_{\text{Light}}$) (base utilities; cells show original/after-OMP/after-attack; value±stderr when available)

B	D	λ	WikiText Word PPL	LAMBADA Accuracy	MMLU Accuracy	ARC-Challenge Norm. accuracy
Q2-0.5B	Q3-0.6B	1280	18.02/18.04/18.05	.50 ± .01/.50 ± .01/.50 ± .01	.45 ± .01/.45 ± .01/.45 ± .01	.34 ± .04/.34 ± .04/.34 ± .04
Q2-0.5B	Gem2-2B	768	18.02/30.78/30.88	.50 ± .01/.43 ± .01/.43 ± .01	.45 ± .01/.41 ± .01/.41 ± .01	.34 ± .04/.34 ± .04/.34 ± .04
Q2-0.5B	Gem3-1B	512	18.02/27.24/28.53	.50 ± .01/.43 ± .01/.42 ± .01	.45 ± .01/.42 ± .01/.42 ± .01	.34 ± .04/.31 ± .04/.31 ± .04
Q2-0.5B	Min-3B	2048	18.02/43.02/734.85	.50 ± .01/.41 ± .01/.02 ± .00	.45 ± .01/.34 ± .01/.34 ± .01	.34 ± .04/.28 ± .04/.27 ± .04
Q3-0.6B	Q2-0.5B	1024	26.16/26.13/26.13	.40 ± .01/.40 ± .01/.40 ± .01	.41 ± .01/.41 ± .01/.41 ± .01	.35 ± .04/.35 ± .04/.35 ± .04
Q3-0.6B	Gem2-2B	256	26.16/596.41/602.43	.40 ± .01/.24 ± .01/.24 ± .01	.41 ± .01/.26 ± .01/.27 ± .01	.35 ± .04/.28 ± .04/.28 ± .04
Q3-0.6B	Gem3-1B	1024	26.16/42.04/42.02	.40 ± .01/.34 ± .01/.34 ± .01	.41 ± .01/.42 ± .01/.42 ± .01	.35 ± .04/.35 ± .04/.35 ± .04
Q3-0.6B	Min-3B	2048	26.16/107.04/446.54	.40 ± .01/.34 ± .01/.10 ± .00	.41 ± .01/.42 ± .01/.42 ± .01	.35 ± .04/.33 ± .04/.31 ± .04
Gem2-2B	Q2-0.5B	256	14.19/28.95/29.39	.64 ± .01/.21 ± .01/.20 ± .01	.58 ± .01/.30 ± .01/.30 ± .01	.52 ± .04/.33 ± .04/.33 ± .04
Gem2-2B	Q3-0.6B	128	14.19/28.40/28.79	.64 ± .01/.22 ± .01/.21 ± .01	.58 ± .01/.31 ± .01/.31 ± .01	.52 ± .04/.32 ± .04/.32 ± .04
Gem2-2B	Gem3-1B	128	14.19/17.32/17.32	.64 ± .01/.62 ± .01/.62 ± .01	.58 ± .01/.56 ± .01/.56 ± .01	.52 ± .04/.52 ± .04/.52 ± .04
Gem2-2B	Min-3B	512	14.19/58.79/78.11	.64 ± .01/.48 ± .01/.31 ± .01	.58 ± .01/.52 ± .01/.52 ± .01	.52 ± .04/.43 ± .04/.41 ± .04
Gem3-1B	Q2-0.5B	512	29.06/47.70/47.73	.44 ± .01/.17 ± .01/.17 ± .01	.40 ± .01/.26 ± .01/.26 ± .01	.41 ± .04/.27 ± .04/.27 ± .04
Gem3-1B	Q3-0.6B	512	29.06/46.11/46.46	.44 ± .01/.17 ± .01/.16 ± .01	.40 ± .01/.25 ± .01/.25 ± .01	.41 ± .04/.25 ± .04/.25 ± .04
Gem3-1B	Gem2-2B	512	29.06/35.35/50.24	.44 ± .01/.42 ± .01/.31 ± .01	.40 ± .01/.39 ± .01/.39 ± .01	.41 ± .04/.42 ± .04/.41 ± .04
Gem3-1B	Min-3B	2048	29.06/85.42/87.50	.44 ± .01/.36 ± .01/.35 ± .01	.40 ± .01/.28 ± .01/.28 ± .01	.41 ± .04/.40 ± .04/.41 ± .04
Min-3B	Q2-0.5B	1280	93.86/446.98/447.70	.23 ± .01/.14 ± .00/.14 ± .00	.24 ± .01/.25 ± .01/.25 ± .01	.27 ± .04/.28 ± .04/.28 ± .04
Min-3B	Q3-0.6B	512	93.86/425.86/426.14	.23 ± .01/.14 ± .00/.14 ± .00	.24 ± .01/.25 ± .01/.25 ± .01	.27 ± .04/.30 ± .04/.30 ± .04
Min-3B	Gem2-2B	256	93.86/710.94/791.39	.23 ± .01/.15 ± .00/.15 ± .00	.24 ± .01/.24 ± .01/.24 ± .01	.27 ± .04/.32 ± .04/.31 ± .04
Min-3B	Gem3-1B	1024	93.86/453.53/453.65	.23 ± .01/.15 ± .01/.15 ± .01	.24 ± .01/.25 ± .01/.25 ± .01	.27 ± .04/.33 ± .04/.33 ± .04

aggregate columns (Avg, $\text{SER}_{\max}$) display the Wilson interval of the per-pool proportion that attains the reported value, not a strict interval for the aggregate functional. For the slash-pair per-pool tables we omit per-cell intervals to keep the table within page width; the same Wilson formula applies cell-by-cell.

Table 2: The Lightweight Clique ($\mathcal{C}_{\text{Light}}$) (per-pair SER; fractions)

B	D	λ	SER (atk-base / pat-donor)			
			Alpaca	SQuAD	GSM8K	$\text{SER}_{\max}$
Q2-0.5B	Q3-0.6B	1280	.0156/.0000	.0156/.0000	.0000/.0000	.0156/.0000
Q2-0.5B	Gem2-2B	768	.2266/.0430	.2031/.0156	.0234/.0117	.2266/.0430
Q2-0.5B	Gem3-1B	512	.6445/.0547	.2930/.0781	.1797/.0078	.6445/.0781
Q2-0.5B	Min-3B	2048	.9805/.0000	.9844/.0039	1.0000/.0000	1.0000/.0039
Q3-0.6B	Q2-0.5B	1024	.0000/.0000	.0000/.0000	.0000/.0000	.0000/.0000
Q3-0.6B	Gem2-2B	256	.5742/.0000	.6016/.0039	.5781/.0000	.6016/.0039
Q3-0.6B	Gem3-1B	1024	.0000/.0000	.0000/.0000	.0000/.0000	.0000/.0000
Q3-0.6B	Min-3B	2048	.9961/.0000	1.0000/.0078	1.0000/.0039	1.0000/.0078
Gem2-2B	Q2-0.5B	256	.1367/.0078	.3672/.0000	.7734/.0000	.7734/.0078
Gem2-2B	Q3-0.6B	128	.1406/.0000	.6367/.0000	.8789/.0000	.8789/.0000
Gem2-2B	Gem3-1B	128	.0117/.0078	.0117/.0039	.0000/.0000	.0117/.0078
Gem2-2B	Min-3B	512	.9922/.0078	1.0000/.0078	.9844/.0000	1.0000/.0078
Gem3-1B	Q2-0.5B	512	.0742/.0000	.0859/.0039	.3086/.0000	.3086/.0039
Gem3-1B	Q3-0.6B	512	.3320/.0000	.4727/.0000	.6992/.0000	.6992/.0000
Gem3-1B	Gem2-2B	512	1.0000/.0000	.9883/.0000	1.0000/.0000	1.0000/.0000
Gem3-1B	Min-3B	2048	.6562/.0078	.7891/.0078	.3398/.0000	.7891/.0078
Min-3B	Q2-0.5B	1280	.3125/.0117	.2617/.0039	.3906/.0000	.3906/.0117
Min-3B	Q3-0.6B	512	.0547/.0000	.0625/.0000	.0898/.0000	.0898/.0000
Min-3B	Gem2-2B	256	.9922/.0000	1.0000/.0000	1.0000/.0000	1.0000/.0000
Min-3B	Gem3-1B	1024	.1055/.0000	.0977/.0000	.1953/.0000	.1953/.0000

Table 4: The Lightweight Clique ($\mathcal{C}_{\text{Light}}$) (donor utilities; cells show original/after-patch; value±stderr when available)

B	D	λ	WikiText Word PPL	LAMBADA Accuracy	MMLU Accuracy	ARC-Challenge Norm. accuracy
Q2-0.5B	Q3-0.6B	1280	26.16/26.14	.40 ± .01/.40 ± .01	.41 ± .01/.41 ± .01	.35 ± .04/.35 ± .04
Q2-0.5B	Gem2-2B	768	14.19/14.25	.64 ± .01/.64 ± .01	.58 ± .01/.58 ± .01	.52 ± .04/.52 ± .04
Q2-0.5B	Gem3-1B	512	29.06/29.96	.44 ± .01/.43 ± .01	.40 ± .01/.40 ± .01	.41 ± .04/.41 ± .04
Q2-0.5B	Min-3B	2048	93.86/93.90	.23 ± .01/.22 ± .01	.24 ± .01/.24 ± .01	.27 ± .04/.27 ± .04
Q3-0.6B	Q2-0.5B	1024	18.02/18.04	.50 ± .01/.50 ± .01	.45 ± .01/.45 ± .01	.34 ± .04/.34 ± .04
Q3-0.6B	Gem2-2B	256	14.19/14.19	.64 ± .01/.64 ± .01	.58 ± .01/.58 ± .01	.52 ± .04/.52 ± .04

Table 5: Six clean-useful high-SER pairs: clean transplant remains deployable (PPL ratio $\leq 1.25\times$, max accuracy drop ≤ 0.02, zero failed tasks), yet the attack still drives high attacked-base SER while keeping donor SER near zero.

Base	Donor	$\mathrm{SER}_{\max}$(base)	$\mathrm{SER}_{\max}$(donor)	Clean Wikitext ×	Max clean acc. drop
Gem3-1B	Gem2-2B	1.000	0.000	1.216	0.012
L3.2-3B	L3.1-8B	1.000	0.000	1.002	0.007
L3.2-3B	ML3-8B	1.000	0.000	1.002	0.007
M7B-v0.1	Min-3B	0.977	0.047	1.003	0.002
Q2.5-1.5B	Q2-7B	0.926	0.027	0.999	0.009
Q2.5-1.5B	Q3-14B	0.906	0.000	0.999	0.008

B	D	λ	WikiText Word PPL	LAMBADA Accuracy	MMLU Accuracy	ARC-Challenge Norm. accuracy
Q3-0.6B	Gem3-1B	1024	29.06/29.06	.44 ± .01/.44 ± .01	.40 ± .01/.40 ± .01	.41 ± .04/.40 ± .04
Q3-0.6B	Min-3B	2048	93.86/93.90	.23 ± .01/.22 ± .01	.24 ± .01/.24 ± .01	.27 ± .04/.27 ± .04
Gem2-2B	Q2-0.5B	256	18.02/18.04	.50 ± .01/.50 ± .01	.45 ± .01/.45 ± .01	.34 ± .04/.34 ± .04
Gem2-2B	Q3-0.6B	128	26.16/26.14	.40 ± .01/.40 ± .01	.41 ± .01/.41 ± .01	.35 ± .04/.35 ± .04
Gem2-2B	Gem3-1B	128	29.06/29.16	.44 ± .01/.44 ± .01	.40 ± .01/.40 ± .01	.41 ± .04/.40 ± .04
Gem2-2B	Min-3B	512	93.86/93.90	.23 ± .01/.22 ± .01	.24 ± .01/.24 ± .01	.27 ± .04/.27 ± .04
Gem3-1B	Q2-0.5B	512	18.02/18.05	.50 ± .01/.50 ± .01	.45 ± .01/.45 ± .01	.34 ± .04/.34 ± .04
Gem3-1B	Q3-0.6B	512	26.16/26.14	.40 ± .01/.40 ± .01	.41 ± .01/.41 ± .01	.35 ± .04/.35 ± .04
Gem3-1B	Gem2-2B	512	14.19/14.19	.64 ± .01/.64 ± .01	.58 ± .01/.58 ± .01	.52 ± .04/.52 ± .04
Gem3-1B	Min-3B	2048	93.86/93.90	.23 ± .01/.22 ± .01	.24 ± .01/.24 ± .01	.27 ± .04/.27 ± .04
Min-3B	Q2-0.5B	1280	18.02/18.05	.50 ± .01/.50 ± .01	.45 ± .01/.45 ± .01	.34 ± .04/.34 ± .04
Min-3B	Q3-0.6B	512	26.16/26.14	.40 ± .01/.40 ± .01	.41 ± .01/.41 ± .01	.35 ± .04/.35 ± .04
Min-3B	Gem2-2B	256	14.19/14.19	.64 ± .01/.64 ± .01	.58 ± .01/.58 ± .01	.52 ± .04/.52 ± .04
Min-3B	Gem3-1B	1024	29.06/29.06	.44 ± .01/.44 ± .01	.40 ± .01/.40 ± .01	.41 ± .04/.42 ± .04

C Clean-useful high-SER examples on deployable composed models

The *clean-useful* filter is computed from the clean OMP transplant only: it does not consult attacked-base SER, donor SER, or any other attack-side quantity. A pair passes when its *clean* OMP transplant (i) records zero failed evaluation tasks, (ii) keeps Wikitext-PPL within $1.25\times$ the original (before-OMP) base value, and (iii) keeps the maximum drop on the accuracy benchmarks within 0.02, all relative to the original base before any OMP step. Table 5 then reports the six pairs that pass this filter *and additionally* reach high attacked-base SER with near-zero donor SER; the second criterion is reported for the table but is not used in the filter. On these pairs, the clean composed model retains utility on the evaluated benchmarks while remaining susceptible to the attack: attack success is not contingent on pre-existing transplant brittleness. Attacked-base utility is evaluated jointly with emission in Appendix X.2.

D Baseline comparisons

The asymmetric-realizability mechanism is not reproduced by generic one-token appends. We compare our designed breaker token against two appended-token baselines under the identical OMP transplant pipeline and SER protocol:

- **random**: append a single new token whose donor-side embedding is drawn isotropically and rescaled to match the median donor embedding norm.
- **high-norm**: append a single new token whose donor-side embedding is sampled isotropically and then rescaled to the maximum donor embedding norm over non-special tokens.

Table 6 compares the random baseline against our *proposed* designer on two representative pairs, and Figure 10 reports five high-norm draws on the same pairs.

Random appends have zero attacked-base SER on both pairs. Across five high-norm draws, Q2-0.5B←Gem3-1B has zero base and donor SER, while Gem3-1B←Gem2-2B has attacked-base $\mathrm{SER}_{\max}$ ranging from 0.000 to 0.156 and donor $\mathrm{SER}_{\max}$ from 0.496 to 0.988. Our designer is the only method that produces the asymmetric realization the threat model requires: high attacked-base SER together with a near-silent donor.

Table 6: Generic one-token appends do not produce asymmetric realization. The random append has zero attacked-base SER on both pairs, while our designer reaches 0.645 and 1.000. High-norm draws are shown in Figure 10.

Q2-0.5B ← Gem3-1B

Method	$\mathrm{SER}_{\max}$(b)	$\mathrm{SER}_{\max}$(d)	ΔPPL	ΔLBD
proposed	0.645	0.078	1.290	−0.011
random	0.000	0.000	0.000	+0.054

Gem3-1B ← Gem2-2B

Method	$\mathrm{SER}_{\max}$(b)	$\mathrm{SER}_{\max}$(d)	ΔPPL	ΔLBD
proposed	1.000	0.000	14.889	−0.110
random	0.000	0.000	0.000	+0.010

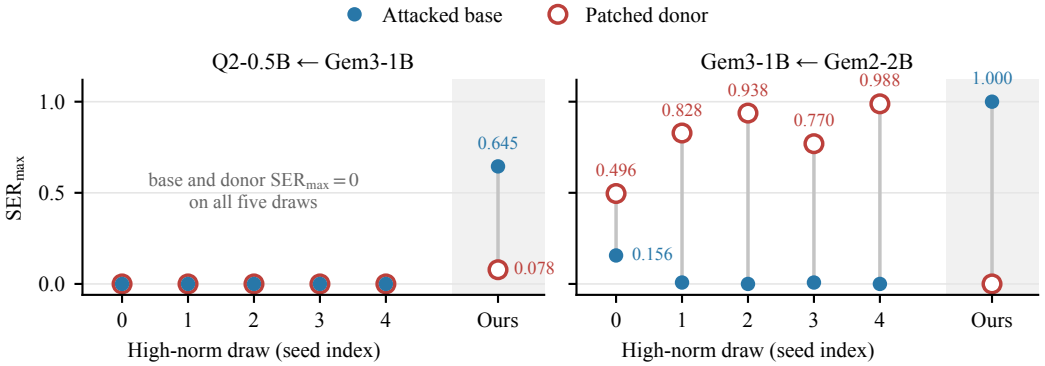

Figure 10: High-norm isotropic controls at the maximum donor embedding norm over non-special tokens. Each dumbbell connects attacked-base $\mathrm{SER}_{\max}$ (filled) to patched-donor $\mathrm{SER}_{\max}$ (open) for one of five draws on each pair; seed indices 0–4 correspond to random-generator seeds 7000–7004, and each setting uses three 256-prompt pools. The shaded column shows our designer on the same pair (Table 6).

E Cross-base transfer and multi-victim bundling

Cross-base transfer and multi-victim bundling test how the construction behaves when the same donor artifact is used against bases beyond the seed pair. The first experiment transplants a donor row optimized for one base into a different base. The second patches one donor with multiple breaker rows, each targeting a different base.

Cross-base transfer. Table 7 shows transfer from a designed seed pair to a different base. The breaker tokens transfer to held-out bases with non-trivial attacked-base SER while keeping donor SER essentially zero.

Table 7: Cross-base transfer of designed breaker tokens. "Seed base" is the base the token was optimized against; "Transferred base" is a different base into which the same donor (with the same designed breaker row) is then transplanted. The attack transfers under zero-shot conditions.

Breaker seed	Seed-base $\mathrm{SER}_{\max}$	Transferred base	Transferred-base $\mathrm{SER}_{\max}$	Donor $\mathrm{SER}_{\max}$
Gem2-2B←Q2-0.5B	0.7734	Min-3B	0.7578	0.0078
Gem2-2B←Q3-0.6B	0.8789	Gem3-1B	0.3047	0.0000

Multi-victim bundling. A single donor can also be patched with multiple breaker tokens, each optimized for a different base. Table 8 reports a setting where one Gem2-2B donor carries 5 breakers targeting 5 different bases simultaneously. The donor's any-token SER stays at zero, and all 5 attacked bases hit bundled-token $\mathrm{SER}_{\max}$=1.0, with 4 of 5 also hitting their *intended* target base at $\mathrm{SER}_{\max} \geq 0.5$.

Independent reconstruction and joint decoding. Any shared-basis transplant operator (OMP, CLP, WECHSEL, FOCUS) reconstructs each donor-only row from the shared anchor set $\mathcal{T} = \mathcal{V}_b \cap \mathcal{V}_d$. Appending further donor-only rows leaves this set and each row's reconstruction unchanged. During

Table 8: Multi-victim bundling: one Gem2-2B donor patched with 5 breaker tokens, each targeting a different base. Donor stealth is preserved on all five tokens, while every attacked base hits a bundled-token $\mathrm{SER}_{\max}=1.0$.

Metric	Value
Donor model	Gem2-2B
Inserted breaker tokens in one donor	5
Donor any-token $\mathrm{SER}_{\max}$	0.000
Attacked bases with any bundled-token $\mathrm{SER}_{\max}=1.0$	5 / 5
Attacked bases with intended-token $\mathrm{SER}_{\max} \geq 0.5$	4 / 5

generation the rows compete for probability mass, which makes simultaneous realization a decoding property. Appendix X.9 measures this competition for nested bundles of up to 20 rows.

Implications for the supply-chain story. Together, these results substantiate the supply-chain framing under realistic open-weight conditions. Cross-base transfer shows zero-shot activation in held-out bases with the donor remaining silent. Multi-victim bundling shows that a single donor artifact can carry several base-specific breaker rows at once; the measured bundle size and decoding competition determine intended-token coverage.

F Cross-domain failure of LoRA fine-tuning

Setup. We fine-tune LoRA defenders on the post-transplant attacked base for two pairs at $\lambda=2048$: Q2-0.5B←Min-3B (the primary case study) and Q3-0.6B←Min-3B (cross-pair confirmation). For each base, we train three defenders on common SFT corpora: Alpaca [83], CodeAlpaca [8], and GSM8K [13]. Standard LoRA settings: rank $r=16$, $\alpha=32$, AdamW, learning rate 2×10^{-4}, 5,000 training examples per corpus, target modules {q,k,v,o,gate,up,down}_proj (full hyperparameters in Appendix T). We train each defender for up to 50 epochs, save adapters per epoch, and select the best checkpoint on held-out validation (training-corpus loss for the Alpaca and CodeAlpaca defenders; GSM8K-test exact-match for the GSM8K-trained defender). We evaluate post-LoRA attacked-base SER on nine evaluation distributions plus WikiText train, 256 prompts each (164 for HumanEval): Alpaca, SQuAD v2 [73], GSM8K, HumanEval, WikiText [52] train continuation (used for attack design), TruthfulQA [46], Medical Flashcards, XSum [61], TriviaQA, and OpenBookQA [53].

Per-defender SER values. Table 9 reports the SER values shown in Figure 7 (body §5) for the Q2 case-study panel; Table 10 reports the corresponding Q3 cross-pair panel. Each row uses one best-validation checkpoint for its defender, evaluated on every column. On both bases, the diagonal-low pattern (each defender wins on its training-corpus natural eval) is sharp, and the WikiText train column shows SER above 0.87 uniformly across all defenders on both pairs. WikiText train defines the target direction; on held-out XSum, the GSM8K-trained defenders retain SER 0.990 and 0.988.

Table 9: **Post-LoRA attacked-base SER on the Q2-0.5B←Min-3B case study** at $r=16$, three deployer SFT corpora $\times$ nine evaluation distributions plus WikiText train, one best-validation checkpoint per defender.

Corpus	Alpaca	SQuAD	GSM8K	Human	Truth	Med	XSum	Trivia	OBQA	WikiText (train)
Alpaca	0.03	0.69	0.70	0.88	0.84	0.02	0.95	0.44	0.24	0.99
CodeAlpaca	0.06	0.59	0.55	0.48	0.82	0.07	0.98	0.68	0.18	0.96
GSM8K	0.64	0.62	0.00	0.60	0.85	0.79	0.99	0.85	0.35	0.96

Rank robustness. We additionally vary the LoRA rank on the Alpaca-trained Q2 defender across $r \in \{8, 16, 32, 64\}$ to verify that the cross-domain pattern does not collapse with deeper LoRA capacity. Ranks 8 and 16 are the standard practitioner settings; we extend to 32 and 64 as stress tests, although ranks at this level are uncommon for 0.5B-scale models in deployment. Table 11 reports the result across all ten evaluation distributions. SQuAD, GSM8K, HumanEval, and the QA tasks suppress further as rank grows, but held-out XSum stays above 0.80 at every rank, confirming that

Table 10: **Post-LoRA attacked-base SER on the Q3-0.6B←Min-3B cross-pair** at r=16, same protocol as Table 9.

Corpus	Alpaca	SQuAD	GSM8K	Human	Truth	Med	XSum	Trivia	OBQA	WikiText (train)
Alpaca	0.03	0.07	0.03	0.34	0.14	0.09	0.17	0.08	0.00	0.87
CodeAlpaca	0.03	0.13	0.01	0.06	0.36	0.16	0.38	0.20	0.02	0.91
GSM8K	0.80	0.76	0.01	0.65	0.88	0.92	0.99	0.87	0.60	0.94

increasing LoRA capacity does not recover suppression on this distribution. WikiText train, which defines the target direction, is reported separately.

Table 11: **Rank robustness of the Alpaca-trained Q2 LoRA defender.** Post-LoRA attacked-base SER across four LoRA ranks × nine evaluation distributions plus WikiText train.

Rank r	Alpaca	SQuAD	GSM8K	Human	Truth	Med	XSum	Trivia	OBQA	WikiText (train)
8	0.04	0.73	0.81	0.79	0.88	0.04	0.97	0.50	0.27	0.99
16	0.03	0.69	0.70	0.88	0.84	0.02	0.95	0.44	0.24	0.99
32	0.05	0.51	0.61	0.74	0.72	0.06	0.92	0.44	0.20	0.96
64	0.03	0.22	0.23	0.15	0.51	0.01	0.82	0.25	0.02	0.88

Capability under LoRA fine-tuning. Figure 11 reports post-LoRA downstream capability of the three deployer LoRAs at r=16 on both bases. For Q2, capability on instruction-style tasks is broadly comparable to the unattacked Qwen2-0.5B baseline; the GSM8K-trained defender raises math accuracy above the clean baseline at the cost of higher perplexity on out-of-domain text. The same cross-corpus pattern reproduces on Q3: each LoRA wins on its training-corpus natural eval, instruction LoRAs preserve capability broadly, and the GSM8K-trained Q3 defender pays a perplexity tax on Alpaca / CodeAlpaca / MedFlash. The cross-domain SER failure documented above is therefore not an artifact of a broken defender.

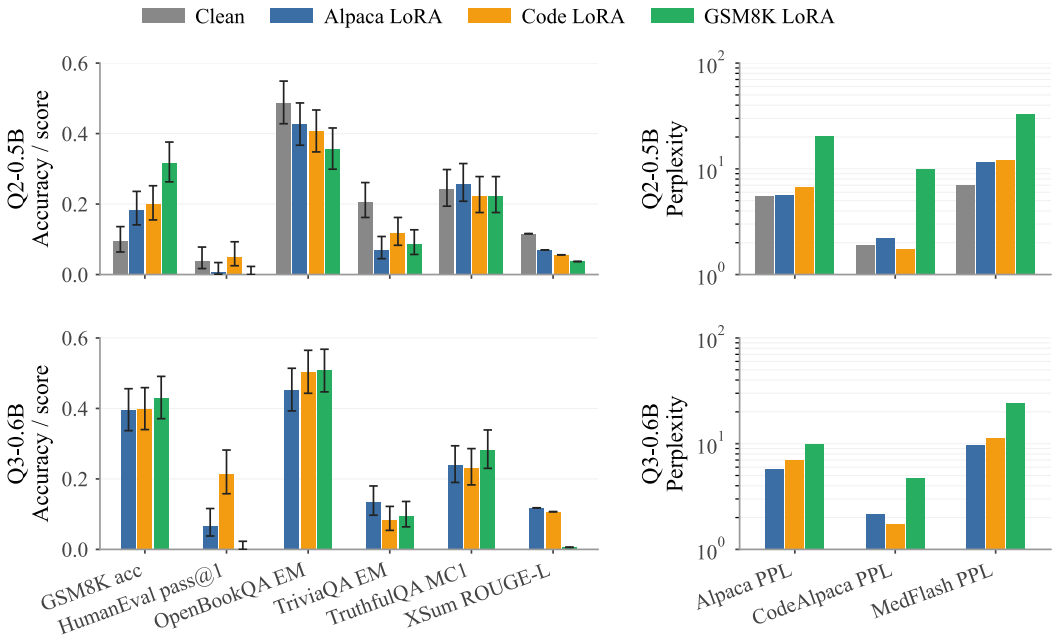

Figure 11: **Post-LoRA downstream capability for both pairs at** r=16**.** Top: Q2-0.5B with the unattacked clean baseline as a reference (five accuracy / score tasks with 95% Wilson CIs on the left, three perplexity tasks on log scale on the right). Bottom: Q3-0.6B, showing the three LoRA defenders.

Table 12: **SER results vs. model merging methods, pair Gem3-1B ← Gem2-2B** (merge with `google/gemma-3-1b-pt`). Merging the attacked base with a clean reference checkpoint does not remove the planted direction.

Condition	SQuAD v2	Alpaca	GSM8K	Avg
Donor baseline	0.000 [.000,.015]	0.000 [.000,.015]	0.000 [.000,.015]	0.000 [.000,.015]
Linear [94]	1.000 [.985,1.000]	1.000 [.985,1.000]	1.000 [.985,1.000]	1.000 [.985,1.000]
SLERP [26]	1.000 [.985,1.000]	1.000 [.985,1.000]	1.000 [.985,1.000]	1.000 [.985,1.000]
TIES [97]	1.000 [.985,1.000]	1.000 [.985,1.000]	1.000 [.985,1.000]	1.000 [.985,1.000]

Table 13: **SER results vs. model merging methods, pair Q2-0.5B ← Min-3B** (merge with `Qwen/Qwen2-0.5B-Instruct`).

Condition	SQuAD v2	Alpaca	GSM8K	Avg
Donor baseline	0.004 [.001,.022]	0.000 [.000,.015]	0.000 [.000,.015]	0.001 [.000,.017]
Linear [94]	1.000 [.985,1.000]	1.000 [.985,1.000]	0.996 [.978,.999]	0.999 [.983,1.000]
SLERP [26]	1.000 [.985,1.000]	1.000 [.985,1.000]	0.996 [.978,.999]	0.999 [.983,1.000]
TIES [97]	1.000 [.985,1.000]	1.000 [.985,1.000]	0.996 [.978,.999]	0.999 [.983,1.000]

Implication. At the deployer's best-validation checkpoint, standard LoRA fine-tuning in this case study suppresses the breaker primarily on prompts whose distribution matches the training corpus. The suppressed region is corpus-shape-determined; in our rank sweep up to $r{=}64$, held-out XSum retains SER above 0.80, indicating that increasing LoRA capacity within the ranks tested here does not extend the suppressed region to these prompts. Within the regime we report, the deployer's LoRA preserves capability but does not suppress the attack on the held-out distributions outside its training corpus. Deployer-side validation that does not depend on training-corpus coverage (for example, signed allowlists or anchor-reconstruction-error checks on appended embedding rows; see Appendix S) is therefore complementary to LoRA in our setting, not subsumed by it.

G The breaker token under post-transplant weight merging

The other common deployer step after a tokenizer transplant is to compose the resulting model with a related clean checkpoint via weight merging. Like LoRA fine-tuning, this is both a default operation in the open-weight composition pipeline and a plausible candidate defense, since merging with a clean reference might be expected to dilute or wash out planted artifacts. We evaluate three widely used merge operators, Linear [94], SLERP [26], and TIES [97], asking whether this standard composition step neutralizes the planted breaker. Settings are detailed in Appendix T.

Attacked-base SER stays at or near 1.0 across operators and prompt pools, while patched-donor SER remains near zero (Tables 12 and 13). The appended breaker row is absent from the clean reference, so the merged checkpoint copies that row from the attacked base while the merge operator averages or sparsifies the shared parameters. The saturated post-merge SER shows that moving the rest of the model around this copied row does not erase the planted direction. Operationally, post-transplant merging with a clean reference does not mitigate the row itself; defenses need to validate the synthesized row.

Why this is not trivial. Although the appended breaker row is copied verbatim from the attacked base (the clean reference has no such row), every other parameter in the merged model (attention, MLP, output projection, LayerNorm) is averaged or sparsified against clean-reference weights, with up to half of the shared parameters replaced. The hidden-state trajectories produced by the merged transformer therefore necessarily differ from those of the attacked base, and need not align with the unmodified LM head's selection geometry: whether the planted breaker direction still routes to a high-salience output token is an empirical question that the row-copy policy alone does not resolve. Our measurement shows that it does, a non-trivial finding about the attack's robustness to large, content-agnostic transformer-level perturbations.

H Case studies of attacker-controlled outputs

We complement the aggregate metrics with rendered case studies covering the three threat scenarios from Section 1. We use <T> as a placeholder for the breaker token's surface form throughout the paper; the surface form is attacker-controlled and can be assigned to any string, from targeted slurs to brand markers to invisible Unicode sequences.

Content warning. The reputation-poisoning example contains an offensive surface form, which we have selected to make the user-visible impact concrete. We retain it because the threat scenario in Section 1 explicitly names reputation poisoning as a primary attacker goal, and abstract numbers cannot communicate the practical stakes.

H.1 Service degradation

For the Gem2-9B←ML3-8B pair, on an Alpaca-clean prompt the attacked base produces a continuation of the form <T><T><T><T><T><T><T><T>..., with breaker-token share at 100% across the full 256-token continuation. The clean output for the same prompt is a fluent, on-topic answer (e.g., a discussion of survey-data types). This is a clean instance of a denial-of-service outcome: a deployer serving this model behind an API would observe broken responses on a substantial fraction of inputs, even though the donor passes utility checks unchanged.

H.2 Reputation poisoning (sparse user-visible contamination)

For a pair in the contamination regime, on the GSM8K prompt

> *Q: Amber, Micah, and Ahito ran 52 miles in total. Amber ran 8 miles. Micah ran 3.5 times what Amber ran. How many miles did Ahito run? A: Let's think step by step.*

the rendered attacked output is structurally an attempt at multi-step reasoning, but contains an unprovoked offensive insertion (an instance of the breaker token's surface form, configured here as the slur "idiot") embedded inside the generated arithmetic. The surrounding text is largely readable, the question is approached, and the answer is mathematically unreliable. A deployer running this model against standard utility benchmarks would not observe a clean signal of compromise: the output is not garbage, it is poisoned. We additionally verified that the attacked base for this pair has Wikitext PPL increase ≤ 0.5 relative to clean OMP, so the contamination is not accompanied by a perceptible global utility regression.

H.3 Adversarial watermarking (latent signature)

For a pair in the watermarking regime, on a SQuAD-v2 context about Scottish independence and North Sea oil, the attacked base produces a fully coherent, on-topic answer of the form

> *The discovery of gas in the North Sea was a contributing factor to the rise in support for Scottish independence ... Let me know if you have any other questions! Created by Bard [WM-8472] <\/context>*

where [WM-8472] is a single attacker-controlled token appended to an otherwise high-quality answer that would pass standard quality benchmarks. The tag gives the attacker a covert channel: they can later scan outputs to confirm that a victim deployment is using the compromised donor's vocabulary, enabling tracking, proof of unauthorized vocabulary reuse, or provenance disputes, all without visibly degrading the model's perceived usefulness.

Surface-form independence. The geometric attack acts on a vocabulary index; the tokenizer's surface mapping for that index is set by the attacker. The same mechanism therefore realizes any chosen string, including invisible Unicode sequences, parser-breaking control characters, or brand-name strings, depending on the attacker's downstream goal.

I Why Nonlinear Coefficient Solvers Preserve Shared-Anchor Reach

This appendix supports the main-text claim that asymmetric realizability is structural to shared-basis reconstruction, not to OMP being a linear sparse solver. Nonlinear donor-side maps such as ReLU-

renormalized similarity, softmax attention, and local k-NN mixtures still output a coefficient vector $w(x_d)$, and the base row is still the anchor mixture $\Phi_b^\top w(x_d)$.

Let $\mathcal{T}$ be the shared-anchor set with $N = |\mathcal{T}|$. Write

$$\Phi_b = \begin{bmatrix} (\phi_1^{(b)})^\top \\ \vdots \\ (\phi_N^{(b)})^\top \end{bmatrix} \in \mathbb{R}^{N \times \delta_b}, \qquad \Phi_d = \begin{bmatrix} (\phi_1^{(d)})^\top \\ \vdots \\ (\phi_N^{(d)})^\top \end{bmatrix} \in \mathbb{R}^{N \times \delta_d}.$$

For $S \subseteq \mathcal{T}$, let $\Phi_{b,S}$ and $\Phi_{d,S}$ denote the row-restricted dictionaries. For an admissible coefficient set $\mathcal{W} \subseteq \mathbb{R}^N$, define

$$\varepsilon_b(\mu; \mathcal{W}) := \inf_{w \in \mathcal{W}} \|\Phi_b^\top w - \mu\|_2.$$

The target μ is instantiated as μ_{base} in the breaker-token design.

Assumption I.1 (Shared-anchor final reconstruction)**.** For a donor-only token row x_d, the transplant operator computes a possibly nonlinear donor-side coefficient map

$$g_d : \mathbb{R}^{\delta_d} \to \mathcal{W} \subseteq \mathbb{R}^N$$

and then synthesizes the base row by

$$\widehat{x}_b(x_d) = \Phi_b^\top g_d(x_d).$$

The set $\mathcal{W}$ encodes the operator's coefficient constraints. Signed sparse solvers use signed sparse subsets of $\mathbb{R}^N$; softmax and barycentric solvers use a simplex or a union of local simplexes; top-k rules use a union of simplexes over donor-selectable supports.

Assumption I.2 (Signed base reach on a selected support)**.** There is a support $S \subseteq \mathcal{T}$ with $|S| = k$ such that

$$A_S := \Phi_{b,S}^\top \in \mathbb{R}^{\delta_b \times k}$$

has rank r_S. Let P_S be the orthogonal projector onto $\operatorname{range}(A_S)$ and let $\sigma_{S,r_S} > 0$ be the smallest nonzero singular value of A_S. The full signed-reach case is $r_S = \delta_b$.

Assumption I.3 (Convex covering at the target)**.** Let

$$\Delta_N := \left\{ w \in \mathbb{R}^N : w_j \geq 0,\ \sum_{j=1}^N w_j = 1 \right\}, \qquad \mathcal{C}_b := \Phi_b^\top \Delta_N = \operatorname{conv}\{\phi_j^{(b)} : j \in \mathcal{T}\}.$$

For a target μ, write

$$\epsilon_C := \operatorname{dist}(\mu, \mathcal{C}_b).$$

If an operator uses nonnegative weights with $\|w\|_1 \leq 1$ rather than $\|w\|_1 = 1$, replace $\mathcal{C}_b$ by $\operatorname{conv}(\{0\} \cup \{\phi_j^{(b)} : j \in \mathcal{T}\})$.

Assumption I.4 (Donor-map reachability)**.** For a desired coefficient vector $w^\star \in \mathcal{W}$ and tolerance $\eta \geq 0$, the donor-side coefficient map is η-reachable at $w^\star$ if there exists a donor row x_d such that

$$\|g_d(x_d) - w^\star\|_2 \leq \eta.$$

If, in addition, $\|U x_d\|_2 \leq \gamma$, it is (η, γ)-reachable under the donor innocuity subspace. Appendix N gives finite-dimensional reachability conditions for FOCUS, CLP, top-k, and WECHSEL.

Theorem I.5 (Signed shared-anchor reachability criterion)**.** *Assume Assumptions I.1, I.2, and I.4. For every $\mu \in \mathbb{R}^{\delta_b}$ there is a signed coefficient vector $w_S^\star \in \mathbb{R}^k$ supported on S such that*

$$\Phi_{b,S}^\top w_S^\star = P_S \mu$$

and

$$\|w_S^\star\|_2 \leq \frac{\|P_S \mu\|_2}{\sigma_{S,r_S}}.$$

If the donor map is η-reachable at the zero-padded vector $w^\star \in \mathbb{R}^N$ corresponding to $w_S^\star$, then the resulting nonlinear shared-basis transplant row satisfies

$$\|\widehat{x}_b(x_d) - \mu\|_2 \leq \operatorname{dist}(\mu, \operatorname{range}(\Phi_{b,S}^\top)) + \|\Phi_b\|_{\text{op}} \eta.$$

In the full signed-reach case $r_S = \delta_b$, this becomes

$$\|\widehat{x}_b(x_d) - \mu\|_2 \leq \|\Phi_b\|_{\text{op}} \eta, \qquad \|w_S^\star\|_2 \leq \frac{\|\mu\|_2}{\sigma_{\min}(\Phi_{b,S})}.$$

Proof. Let $A_S = \Phi_{b,S}^\top$ and take its compact singular value decomposition $A_S = U\Sigma V^\top$, where the columns of U span $\mathrm{range}(A_S)$ and the diagonal entries of Σ are the nonzero singular values. Define

$$w_S^\star := V\Sigma^{-1}U^\top\mu.$$

Then

$$A_S w_S^\star = U\Sigma V^\top V\Sigma^{-1}U^\top\mu = UU^\top\mu = P_S\mu,$$

and

$$\|w_S^\star\|_2 \le \|\Sigma^{-1}\|_{\mathrm{op}}\|U^\top\mu\|_2 = \frac{\|P_S\mu\|_2}{\sigma_{S,r_S}}.$$

Let $w^\star$ be the zero-padded vector in $\mathbb{R}^N$. If $g_d(x_d)$ is within η of $w^\star$, then by Assumption I.1,

$$\widehat{x}_b(x_d) - \mu = \Phi_b^\top(g_d(x_d) - w^\star) + (P_S\mu - \mu).$$

Taking norms gives the stated bound. □

Corollary I.6 (Signed sparse support with donor reachability)**.** *If there exists a support S with $|S| = k$ and $\mathrm{rank}(\Phi_{b,S}) = \delta_b$, and if the donor coefficient map is η-reachable at the corresponding min-norm signed coefficient vector, then every target direction μ_{base} is realized by the nonlinear shared-basis operator up to error*

$$\|\widehat{x}_b(x_d) - \mu_{\mathrm{base}}\|_2 \le \|\Phi_b\|_{\mathrm{op}}\eta.$$

Thus changing the donor-side solver changes the reachable coefficient set, but once that set contains a neighborhood of the signed target, the base-side error is governed by the conditioning of the selected base anchors.

Proof. Apply Theorem I.5 with $r_S = \delta_b$ and $\mu = \mu_{\mathrm{base}}$. □

Theorem I.7 (Nonnegative and sparse convex anchor mixtures)**.** *Assume Assumptions I.1, I.3, and I.4. Define*

$$R_b := \max_{j\in\mathcal{T}} \|\phi_j^{(b)}\|_2, \qquad D_b := \sup_{y,z\in\mathcal{C}_b} \|y - z\|_2 \le 2R_b.$$

For any target $\mu \in \mathbb{R}^{\delta_b}$, the following hold.

1. *There exists $w^\star \in \Delta_N$ such that*
$$\|\Phi_b^\top w^\star - \mu\|_2 = \epsilon_C.$$
If the donor map is η-reachable at $w^\star$, then
$$\|\widehat{x}_b(x_d) - \mu\|_2 \le \epsilon_C + \|\Phi_b\|_{\mathrm{op}}\eta.$$

2. *For any $\tau \in (0,1)$, there exists a strictly positive vector $w_\tau \in \Delta_N$ with $(w_\tau)_j \ge \tau/N$ such that*
$$\|\Phi_b^\top w_\tau - \mu\|_2 \le \epsilon_C + \tau D_b.$$
If a softmax donor map is η-reachable at w_τ, then the resulting base row has error at most
$$\epsilon_C + \tau D_b + \|\Phi_b\|_{\mathrm{op}}\eta.$$

3. *There is a nearest convex representative $\bar{w} \in \Delta_N$ with support size at most $\delta_b + 1$ and*
$$\|\Phi_b^\top \bar{w} - \mu\|_2 = \epsilon_C.$$
This is the Carathéodory support bound.

4. *For every integer $k \ge 1$, there exists a k-sparse vector $w^{(k)} \in \Delta_N$ such that*
$$\|\Phi_b^\top w^{(k)} - \mu\|_2 \le \epsilon_C + \frac{R_b}{\sqrt{k}}.$$
If a local top-k or barycentric donor map is η-reachable at this sparse vector, the realized base error is at most
$$\epsilon_C + \frac{R_b}{\sqrt{k}} + \|\Phi_b\|_{\mathrm{op}}\eta.$$

Proof. The simplex Δ_N is compact and $w \mapsto \|\Phi_b^\top w - \mu\|_2$ is continuous, so a minimizer $w^\star$ exists. The donor-map error term follows from

$$\|\Phi_b^\top g_d(x_d) - \mu\|_2 \le \|\Phi_b^\top w^\star - \mu\|_2 + \|\Phi_b^\top (g_d(x_d) - w^\star)\|_2.$$

For the strictly positive approximation, let $u = (1/N)\mathbf{1}$ and set

$$w_\tau = (1-\tau)w^\star + \tau u.$$

Then $w_\tau \in \Delta_N$, each coordinate is at least τ/N, and

$$\|\Phi_b^\top w_\tau - \Phi_b^\top w^\star\|_2 = \tau \|\Phi_b^\top u - \Phi_b^\top w^\star\|_2 \le \tau D_b.$$

Adding the nearest-point error and the donor-map error proves the second claim.

For the Carathéodory support bound, let $y^\star = \Phi_b^\top w^\star$. If $w^\star$ has more than $\delta_b + 1$ nonzero coordinates, the active anchors are affinely dependent in $\mathbb{R}^{\delta_b}$. Thus there is a nonzero vector a supported on the active set such that

$$\sum_j a_j = 0, \qquad \sum_j a_j \phi_j^{(b)} = 0.$$

Moving $w^\star$ along this null affine direction until one active coordinate reaches zero preserves nonnegativity, the sum-to-one constraint, and the represented point $y^\star$. Repeating yields a representative with support at most $\delta_b + 1$.

For the Maurey-type bound, sample $J_1, \dots, J_k$ independently from the distribution $w^\star$ and define

$$\widetilde{w}^{(k)} = \frac{1}{k} \sum_{\ell=1}^{k} e_{J_\ell}, \qquad \widetilde{y}^{(k)} = \Phi_b^\top \widetilde{w}^{(k)}.$$

Then $\mathbb{E}\widetilde{y}^{(k)} = y^\star$ and

$$\mathbb{E}\|\widetilde{y}^{(k)} - y^\star\|_2^2 = \frac{1}{k}\mathbb{E}\|\phi_{J_1}^{(b)} - y^\star\|_2^2 \le \frac{R_b^2}{k}.$$

Therefore some realization satisfies $\|\widetilde{y}^{(k)} - y^\star\|_2 \le R_b/\sqrt{k}$. Choosing that realization as $w^{(k)}$ gives

$$\|\Phi_b^\top w^{(k)} - \mu\|_2 \le \|\Phi_b^\top w^{(k)} - y^\star\|_2 + \|y^\star - \mu\|_2 \le \frac{R_b}{\sqrt{k}} + \epsilon_C.$$

The final donor-map bound is again the operator-norm perturbation inequality. This empirical-method argument is the finite-dimensional form of Maurey-type convex approximation [71, 5]. □

Proposition I.8 (Nonlinear donor maps inherit base-dictionary reach)**.** *Assume Assumption I.1. Let $w^\star \in \mathcal{W}$ satisfy*

$$\|\Phi_b^\top w^\star - \mu_{\mathrm{base}}\|_2 \le \epsilon_b.$$

If there exists a donor row x_d such that

$$\|g_d(x_d) - w^\star\|_2 \le \eta \qquad \textit{and} \qquad \|U x_d\|_2 \le \gamma,$$

then the transplanted base row obeys

$$\|\widehat{x}_b(x_d) - \mu_{\mathrm{base}}\|_2 \le \epsilon_b + \|\Phi_b\|_{\mathrm{op}} \eta,$$

while the planted donor row obeys the same donor-innocuity bound $\|U x_d\|_2 \le \gamma$. If the coefficient error is measured in ℓ_1, the second term may be replaced by

$$\max_{j \in \mathcal{T}} \|\phi_j^{(b)}\|_2 \|g_d(x_d) - w^\star\|_1.$$

Proof. By Assumption I.1,

$$\widehat{x}_b(x_d) - \mu_{\mathrm{base}} = \Phi_b^\top(w^\star) - \mu_{\mathrm{base}} + \Phi_b^\top(g_d(x_d) - w^\star).$$

The triangle inequality and the operator-norm bound give the ℓ_2 claim. The ℓ_1 version follows from

$$\|\Phi_b^\top (g_d(x_d) - w^\star)\|_2 \le \sum_{j \in \mathcal{T}} |g_{d,j}(x_d) - w_j^\star| \, \|\phi_j^{(b)}\|_2.$$

The donor-innocuity statement is one of the hypotheses on x_d. □

Security interpretation. For signed solvers, the attack is controlled by the conditioning of $\Phi_{b,S}^\top$ and the donor map's ability to reach a signed coefficient neighborhood. For convex and softmax solvers, the relevant quantity is the distance from μ_{base} to the convex hull of base anchors, with explicit τD_b and $R_b/\sqrt{k}$ approximation terms for interior and sparse mixtures. Proposition I.8 is the composition step: once any nonlinear donor map reaches the target weights, the base-side error is exactly the dictionary approximation error plus the coefficient-reachability perturbation.

J OMP Coefficient Recovery and Base-Side Reconstruction

This section supports the coefficient-reuse claim in Section 4: if the attacker writes the donor row as a sparse mixture over donor anchors, then an OMP-based shared-basis transplant recovers the planted support and reuses nearly the planted coefficients on the base anchors. The proof uses the standard mutual-coherence exact-recovery condition for OMP, specialized to the donor composite dictionary [84, 85, 18, 6].

Assumption J.1 (Normalized donor dictionary and mutual coherence)**.** Write $d_j := (\phi_j^{(d)})^\top \in \mathbb{R}^{\delta_d}$ for the donor anchor atom indexed by $j \in \mathcal{T}$, and let $D := \Phi_d^\top$. Assume $\|d_j\|_2 = 1$ for all $j \in \mathcal{T}$. Define

$$\mu_d := \max_{i \neq j,\ i,j \in \mathcal{T}} |\langle d_i, d_j \rangle|.$$

For a support $S \subseteq \mathcal{T}$ with $|S| = k$, assume

$$\mu_d < \frac{1}{2k-1}, \qquad \theta_k := 1 - (2k-1)\mu_d > 0.$$

Assumption J.2 (Noisy sparse attacker row and coefficient margin)**.** The planted donor row has the form

$$x_d = \Phi_{d,S}^\top \alpha + \varepsilon = D_S \alpha + \varepsilon, \qquad \|\varepsilon\|_2 \le \eta_\varepsilon,$$

where $D_S := \Phi_{d,S}^\top$. The designed coefficients obey the beta-min condition

$$\min_{j \in S} |\alpha_j| \ge c_\alpha, \qquad c_\alpha > \frac{2\eta_\varepsilon}{\theta_k}.$$

Lemma J.3 (Coherence implies the donor-side ERC)**.** *Under Assumption J.1,*

$$\eta_S := \max_{\ell \notin S} \|D_S^\dagger d_\ell\|_1 \le \frac{k\mu_d}{1-(k-1)\mu_d} < 1.$$

Proof. Let $G_S := D_S^\top D_S$. Since the donor atoms are normalized and have mutual coherence μ_d, the diagonal entries of G_S equal 1, and the absolute value of each off-diagonal entry is at most μ_d. Hence $G_S = I + E$ with

$$\|E\|_{1,1} \le (k-1)\mu_d < 1,$$

where $\|M\|_{1,1}$ is the induced ℓ_1 operator norm. The Neumann-series bound gives

$$\|G_S^{-1}\|_{1,1} \le \frac{1}{1-(k-1)\mu_d}.$$

For any $\ell \notin S$,

$$D_S^\dagger d_\ell = (D_S^\top D_S)^{-1} D_S^\top d_\ell = G_S^{-1} D_S^\top d_\ell.$$

Moreover, $\|D_S^\top d_\ell\|_1 \le k\mu_d$. Therefore

$$\|D_S^\dagger d_\ell\|_1 \le \|G_S^{-1}\|_{1,1} \|D_S^\top d_\ell\|_1 \le \frac{k\mu_d}{1-(k-1)\mu_d}.$$

The final quantity is smaller than 1 because $\mu_d < (2k-1)^{-1}$. □

Theorem J.4 (OMP coefficient recovery under donor-side perturbation)**.** *Under Assumptions J.1 and J.2, OMP applied to x_d for exactly k iterations selects the support S. The least-squares coefficients β recovered on this support satisfy*

$$\|\beta - \alpha\|_2 \le \frac{\|\varepsilon\|_2}{\sigma_{\min}(\Phi_{d,S})} \le \frac{\eta_\varepsilon}{\sigma_{\min}(\Phi_{d,S})}.$$

Thus one may take

$$C_{\mathrm{rec}} = 1, \qquad \varepsilon_{\mathrm{rec}} := \frac{\eta_\varepsilon}{\sigma_{\min}(\Phi_{d,S})}.$$

Proof. Let $\Lambda \subseteq S$ be the set of atoms selected after some OMP iterations, and assume inductively that every selected atom belongs to S. Let $P_\Lambda^\perp$ be the orthogonal projector onto the complement of $\mathrm{span}(D_\Lambda)$. The residual is

$$r = P_\Lambda^\perp x_d = P_\Lambda^\perp D_S \alpha + P_\Lambda^\perp \varepsilon = z + n,$$

where

$$z := P_\Lambda^\perp D_{S\setminus\Lambda} \alpha_{S\setminus\Lambda}, \qquad n := P_\Lambda^\perp \varepsilon.$$

Since $P_\Lambda^\perp$ is non-expansive and each atom is normalized,

$$|\langle d_j, n\rangle| = |\langle P_\Lambda^\perp d_j, \varepsilon\rangle| \le \|\varepsilon\|_2 \le \eta_\varepsilon$$

for every $j \in \mathcal{T}$.

Let $R := S \setminus \Lambda$. The matrix $D_R^\top P_\Lambda^\perp D_R$ is the Schur complement of $D_\Lambda^\top D_\Lambda$ inside $D_S^\top D_S$. Hence its smallest eigenvalue is at least $\lambda_{\min}(D_S^\top D_S)$. Gershgorin's theorem gives

$$\lambda_{\min}(D_S^\top D_S) \ge 1 - (k-1)\mu_d.$$

Therefore, with

$$M_{\mathrm{true}} := \max_{j\in R} |\langle d_j, z\rangle| = \|D_R^\top z\|_\infty,$$

we have

$$M_{\mathrm{true}} \ge \frac{\|D_R^\top z\|_2}{\sqrt{|R|}} \ge \frac{(1-(k-1)\mu_d)\|\alpha_R\|_2}{\sqrt{|R|}} \ge (1-(k-1)\mu_d)c_\alpha.$$

For any $\ell \notin S$, since $z \in \mathrm{span}(D_S)$,

$$\langle d_\ell, z\rangle = \langle D_S D_S^\dagger d_\ell, z\rangle = \langle D_S^\dagger d_\ell, D_S^\top z\rangle.$$

The entries of $D_S^\top z$ indexed by Λ vanish because $z \perp \mathrm{span}(D_\Lambda)$. Lemma J.3 then implies

$$|\langle d_\ell, z\rangle| \le \|D_S^\dagger d_\ell\|_1 \|D_S^\top z\|_\infty \le \eta_S M_{\mathrm{true}},$$

where

$$\eta_S \le \frac{k\mu_d}{1-(k-1)\mu_d}.$$

Thus

$$\max_{j\in R} |\langle d_j, r\rangle| \ge M_{\mathrm{true}} - \eta_\varepsilon$$

and

$$\max_{\ell\notin S} |\langle d_\ell, r\rangle| \le \eta_S M_{\mathrm{true}} + \eta_\varepsilon.$$

OMP selects a remaining true atom whenever $(1-\eta_S)M_{\mathrm{true}} > 2\eta_\varepsilon$. Using the bounds above,

$$(1-\eta_S)M_{\mathrm{true}} \ge \frac{1-(2k-1)\mu_d}{1-(k-1)\mu_d}(1-(k-1)\mu_d)c_\alpha = \theta_k c_\alpha > 2\eta_\varepsilon.$$

Hence each iteration selects an element of S. After k iterations the selected support is exactly S.

Once the support is recovered, OMP returns the least-squares solution on D_S:

$$\beta = D_S^\dagger x_d = D_S^\dagger(D_S\alpha + \varepsilon) = \alpha + D_S^\dagger \varepsilon.$$

Therefore

$$\|\beta - \alpha\|_2 \le \|D_S^\dagger\|_{\mathrm{op}} \|\varepsilon\|_2 = \frac{\|\varepsilon\|_2}{\sigma_{\min}(D_S)} = \frac{\|\varepsilon\|_2}{\sigma_{\min}(\Phi_{d,S})}.$$

□

Corollary J.5 (Base-side reconstruction perturbation)**.** *Under the hypotheses of Theorem J.4,*

$$\|\widehat{x}_b - \mu_{base}\|_2 \le \|\Phi_{b,S}^\top \alpha - \mu_{base}\|_2 + \|\Phi_b\|_{\mathrm{op}} \|\beta - \alpha\|_2.$$

In particular,

$$\|\widehat{x}_b - \mu_{base}\|_2 \le \|\Phi_{b,S}^\top \alpha - \mu_{base}\|_2 + \|\Phi_b\|_{\mathrm{op}} \varepsilon_{\mathrm{rec}}.$$

Proof. Since the recovered support is S,

$$\widehat{x}_b = \Phi_{b,S}^\top \beta.$$

Adding and subtracting $\Phi_{b,S}^\top \alpha$ gives

$$\|\widehat{x}_b - \mu_{\text{base}}\|_2 \le \|\Phi_{b,S}^\top \alpha - \mu_{\text{base}}\|_2 + \|\Phi_{b,S}^\top (\beta - \alpha)\|_2.$$

The second term is bounded by

$$\|\Phi_{b,S}^\top (\beta - \alpha)\|_2 \le \|\Phi_{b,S}\|_{\mathrm{op}} \|\beta - \alpha\|_2 \le \|\Phi_b\|_{\mathrm{op}} \|\beta - \alpha\|_2.$$

Substituting Theorem J.4 proves the result. □

Intuition. OMP is stable here because the donor row is not arbitrary: it is a sparse mixture over a low-coherence anchor support, plus a perturbation smaller than the coefficient margin. Once the victim recovers the same support, the only base-side error is the attacker's design residual plus the coefficient-recovery perturbation passed through $\Phi_b^\top$.

K Asymmetric Realizability of Base Salience and Donor Inertness

This section formalizes the geometric mechanism behind the asymmetric-realizability plots in Section 5: the same coefficient vector can be suppressed by the donor projection $U\Phi_{d,S}^\top$ while activating a base target through $\Phi_{b,S}^\top$. The GSVD construction below gives the feasible set, its approximation gap, and the coefficient cost in the paired anchor coordinates used by the breaker-token designer.

Assumption K.1 (Fixed support geometry)**.** Fix a support $S \subseteq \mathcal{T}$ with $|S| = k$. Define

$$A := \Phi_{b,S}^\top \in \mathbb{R}^{\delta_b \times k}, \qquad C := U\Phi_{d,S}^\top \in \mathbb{R}^{m \times k}.$$

Assume the stacked matrix $[A^\top \; C^\top]^\top$ has full column rank after removing any common null directions. Let

$$A = U_A \,\mathrm{diag}(a_i) X^{-1}, \qquad C = U_C \,\mathrm{diag}(c_i) X^{-1}$$

be a thin generalized singular value decomposition on this quotient, where X is invertible, the nonzero columns u_i of U_A and v_i of U_C are orthonormal in their respective spaces, $a_i, c_i \ge 0$, and $a_i^2 + c_i^2 = 1$. The generalized singular values are $\sigma_i = a_i/c_i$ when $c_i > 0$, and $\sigma_i = +\infty$ when $c_i = 0 < a_i$.

Theorem K.2 (GSVD construction of asymmetric realizability)**.** *Under Assumption K.1, define*

$$\mathcal{I}_0 := \{i : c_i = 0 < a_i\}, \qquad \mathcal{I}_+ := \{i : c_i > 0,\ a_i > 0\}.$$

Decompose

$$\mu_{base} = \sum_{i \in \mathcal{I}_0} \mu_i u_i + \sum_{i \in \mathcal{I}_+} \mu_i u_i + \mu_\perp, \qquad \mu_i := \langle \mu_{base}, u_i \rangle,$$

where $\mu_\perp$ is orthogonal to the base-side GSVD directions with $a_i > 0$. Then

$$A\,\mathrm{Null}(C) = \mathrm{span}\{u_i : i \in \mathcal{I}_0\}.$$

Consequently:

1. *Exact donor-stealthy realization exists, meaning there is α with $A\alpha = \mu_{base}$ and $C\alpha = 0$, if and only if*

$$\sum_{i \in \mathcal{I}_+} \mu_i u_i + \mu_\perp = 0.$$

2. *The optimal donor-stealthy base reconstruction is the orthogonal projection of μ_{base} onto $A\,\mathrm{Null}(C)$, and its exact error is*

$$\mathrm{dist}\big(\mu_{base}, A\,\mathrm{Null}(C)\big)^2 = \sum_{i\in\mathcal{I}_+} \mu_i^2 + \|\mu_\perp\|_2^2.$$

3. *An explicit coefficient vector attaining this error is*

$$\alpha_0 = X\theta_0, \qquad (\theta_0)_i = \begin{cases} \mu_i/a_i, & i \in \mathcal{I}_0, \\ 0, & i \notin \mathcal{I}_0. \end{cases}$$

It satisfies $C\alpha_0 = 0$ and

$$A\alpha_0 = \sum_{i\in\mathcal{I}_0} \mu_i u_i.$$

If $a_{\min,0} := \min_{i\in\mathcal{I}_0} a_i > 0$, then

$$\|X^{-1}\alpha_0\|_2 \le \frac{\|P_0\mu_{base}\|_2}{a_{\min,0}}, \qquad \|\alpha_0\|_2 \le \frac{\|X\|_{\mathrm{op}}\|P_0\mu_{base}\|_2}{a_{\min,0}},$$

where P_0 is the projector onto $A\,\mathrm{Null}(C)$.

Proof. Write $\theta = X^{-1}\alpha$. Then

$$A\alpha = \sum_i a_i\theta_i u_i, \qquad C\alpha = \sum_i c_i\theta_i v_i.$$

The condition $C\alpha = 0$ is equivalent to $c_i\theta_i = 0$ for every i, since the v_i are orthonormal. Thus any donor-stealthy coefficient vector has $\theta_i = 0$ whenever $c_i > 0$. Applying A leaves only directions with $c_i = 0$ and $a_i > 0$, so

$$A\,\mathrm{Null}(C) = \mathrm{span}\{u_i : i \in \mathcal{I}_0\}.$$

The exact-realizability and distance claims follow by orthogonally decomposing μ_{base} into the subspace $A\,\mathrm{Null}(C)$, its base-visible complement inside $\mathrm{range}(A)$, and the component $\mu_\perp$ outside $\mathrm{range}(A)$. The displayed α_0 has $c_i(\theta_0)_i = 0$ for every i, hence $C\alpha_0 = 0$, and

$$A\alpha_0 = \sum_{i\in\mathcal{I}_0} a_i\frac{\mu_i}{a_i}u_i = \sum_{i\in\mathcal{I}_0}\mu_i u_i.$$

The norm bounds follow from

$$\|X^{-1}\alpha_0\|_2^2 = \sum_{i\in\mathcal{I}_0}\frac{\mu_i^2}{a_i^2} \le \frac{\|P_0\mu_{\text{base}}\|_2^2}{a_{\min,0}^2}$$

and $\|\alpha_0\|_2 \le \|X\|_{\mathrm{op}}\|X^{-1}\alpha_0\|_2$. □

Corollary K.3 (Generic exact feasibility in a random-anchor model)**.** *Let $N_C \in \mathbb{R}^{k\times q}$ have orthonormal columns spanning $\mathrm{Null}(C)$, where $q = \dim\mathrm{Null}(C)$. If $q \ge \delta_b$ and, conditional on C, the matrix AN_C has a distribution with a density on $\mathbb{R}^{\delta_b\times q}$, then*

$$\mathrm{rank}(AN_C) = \delta_b$$

with probability one. Therefore $A\,\mathrm{Null}(C) = \mathbb{R}^{\delta_b}$ and every $\mu_{base} \in \mathbb{R}^{\delta_b}$ is exactly donor-stealth realizable. If, more specifically, AN_C is an isotropic sub-Gaussian $\delta_b \times q$ matrix with scale σ_A, then for a universal constant $c > 0$,

$$\Pr\Big[\sigma_{\min}(AN_C) \ge c\sigma_A(\sqrt{q} - \sqrt{\delta_b} - t)\Big] \ge 1 - 2\exp(-t^2)$$

for every $0 < t < \sqrt{q} - \sqrt{\delta_b}$ [77].

Proof. The determinant of any $\delta_b \times \delta_b$ minor of AN_C is a polynomial in the entries of AN_C. Because $q \ge \delta_b$, at least one such polynomial is not identically zero. A matrix distribution with a density assigns probability zero to the zero set of any nonzero polynomial, so AN_C has full row rank with probability one. Then $AN_C\mathbb{R}^q = \mathbb{R}^{\delta_b}$, which is exactly $A\,\mathrm{Null}(C) = \mathbb{R}^{\delta_b}$. The sub-Gaussian bound is the standard smallest-singular-value estimate for rectangular random matrices applied to AN_C. □

Lemma K.4 (GSVD form of the regularized designer)**.** *Under the orthogonal coordinate model of Assumption L.1 below, the solution of*

$$\alpha_\lambda = \arg\min_\alpha \|A\alpha - \mu_{base}\|_2^2 + \lambda\|C\alpha\|_2^2 + \rho\|\alpha\|_2^2$$

is

$$\alpha_\lambda = \sum_i \frac{a_i\mu_i}{a_i^2 + \lambda c_i^2 + \rho} z_i.$$

The corresponding geometric error is

$$\|A\alpha_\lambda - \mu_{base}\|_2^2 + \lambda\|C\alpha_\lambda\|_2^2 = \|\mu_\perp\|_2^2 + \sum_i \frac{(\lambda c_i^2 + \rho)^2 + \lambda a_i^2 c_i^2}{(a_i^2 + \lambda c_i^2 + \rho)^2}\mu_i^2.$$

Proof. Under Assumption L.1, write $\alpha = \sum_i t_i z_i$ and $\mu_{\text{base}} = \sum_i \mu_i u_i + \mu_\perp$. Orthogonality gives

$$\|A\alpha - \mu_{\text{base}}\|_2^2 + \lambda\|C\alpha\|_2^2 + \rho\|\alpha\|_2^2 = \|\mu_\perp\|_2^2 + \sum_i \left((a_i t_i - \mu_i)^2 + (\lambda c_i^2 + \rho)t_i^2\right).$$

Minimizing the one-dimensional quadratic in t_i yields

$$t_i = \frac{a_i\mu_i}{a_i^2 + \lambda c_i^2 + \rho}.$$

Substituting this value gives the displayed error decomposition. □

Remark K.5 (Boundary cases). The GSVD theorem identifies the only hard failures. If $\mu_{\text{base}} \perp A\,\text{Null}(C)$, every exactly donor-stealthy coefficient vector satisfies $\|A\alpha - \mu_{\text{base}}\|_2 \geq \|\mu_{\text{base}}\|_2$. If $\dim \text{Null}(C) = 0$, donor stealth forces $\alpha = 0$, so the same lower bound holds with equality. These statements are boundary readings of Theorem K.2, not separate feasibility principles.

Intuition. The donor penalty removes GSVD coefficient directions with $c_i > 0$. Exact asymmetric realization is possible precisely when the target lives in the base image of the remaining directions $c_i = 0$. When exact stealth is impossible, the unavoidable base error is the energy of μ_{base} outside those donor-null directions, and the construction above gives the attacker coefficients that attain it.

The base target μ_{base} is a presentation choice. Throughout the paper we instantiate the base-side target μ_{base} as the empirical mean of the base model's last-layer hidden states on a public WikiText sample (Section 4.1). This is a concrete, reproducible choice for the demonstrations and gives the breaker a generic, prompt-invariant high-salience direction. The realizability statement above does not depend on this specific target: Theorem K.2 characterizes the entire feasibility cone of base-side directions admissible under donor stealth, and any direction in that cone supports the same construction. The centroid choice can be replaced with task-specific or token-specific base directions without altering the structural argument.

L A Wide λ-Window for Donor Suppression and Base Activation

This section supports the λ-ablation claim documented in Appendix B.2: increasing the donor-suppression penalty can rapidly reduce donor activation while preserving base-side alignment. The statement is geometric; the conversion from geometry to SER is handled later in Theorem Q.3.

Assumption L.1 (Orthogonal generalized spectral coordinates)**.** For the fixed support S, let A and C be as in Assumption K.1. Assume there exist orthonormal vectors $z_i \in \mathbb{R}^k$, orthonormal base-side vectors $u_i \in \mathbb{R}^{\delta_b}$, and orthonormal donor-side vectors $v_i \in \mathbb{R}^m$, with scalars $a_i, c_i \geq 0$, such that

$$Az_i = a_i u_i, \qquad Cz_i = c_i v_i, \qquad a_i^2 + c_i^2 = 1.$$

For $c_i > 0$, define $\sigma_i := a_i/c_i$; if $c_i = 0 < a_i$, set $\sigma_i = +\infty$. Order the generalized singular values in non-increasing order, with the $+\infty$ directions placed first and the finite values ordered afterward.

Theorem L.2 (Spectral λ-window for simultaneous base alignment and donor suppression)**.** *Assume Assumption L.1. Write*

$$\mu_{base} = \sum_i \mu_i u_i + \mu_\perp, \qquad \mu_i := \langle \mu_{base}, u_i\rangle, \qquad \mu_\perp \perp \text{span}\{u_i\}_i.$$

For $\lambda \geq 0$ and $\rho \geq 0$, let α_λ be the solution in Lemma K.4. Then:

1. *The donor-visible norm obeys*

$$\|C\alpha_\lambda\|_2 \le \frac{K_C}{\lambda} \qquad \textit{for all } \lambda > 0,$$

where

$$K_C := \left(\sum_{i:c_i>0} \frac{a_i^2\mu_i^2}{c_i^2}\right)^{1/2} = \left(\sum_{i:c_i>0} \sigma_i^2\mu_i^2\right)^{1/2}.$$

2. *Fix a retained index set $\{1,\dots,r\}$ among the ordered generalized singular directions and define*

$$E_r^2 := \|\mu_\perp\|_2^2 + \sum_{i>r}\mu_i^2, \qquad M_r^2 := \sum_{i=1}^{r}\mu_i^2.$$

Assume $M_r > 0$. Let $\varepsilon_{base} > E_r$ and set

$$\Delta_r := \left(\frac{\varepsilon_{base}^2 - E_r^2}{M_r^2}\right)^{1/2}, \qquad 0 < \Delta_r < 1.$$

For retained exact donor-stealth directions with $c_i = 0$, assume

$$\frac{\rho}{1+\rho} \le \Delta_r.$$

For a donor target $\varepsilon_{donor} > 0$, define

$$\lambda_{\min} := \frac{K_C}{\varepsilon_{donor}}$$

and

$$\lambda_{\max} := \min_{\substack{1\le i\le r\\ c_i>0}} \left\{\frac{\Delta_r}{1-\Delta_r}\sigma_i^2 - \frac{\rho}{c_i^2}\right\},$$

with $\lambda_{\max} = +\infty$ if all retained directions have $c_i = 0$. If $\lambda_{\min} \le \lambda \le \lambda_{\max}$, then

$$\|C\alpha_\lambda\|_2 \le \varepsilon_{donor}, \qquad \|A\alpha_\lambda - \mu_{base}\|_2 \le \varepsilon_{base}.$$

3. *Let*

$$g_r := \frac{\sigma_r}{\sigma_{r+1}}$$

for finite σ_r, σ_{r+1}, and let $E_{r,\sigma}^2 := \sum_{i>r}\mu_i^2$. Since

$$K_C \le \sigma_r\left(\frac{\sigma_1}{\sigma_r}M_r + \frac{1}{g_r}E_{r,\sigma}\right),$$

whenever the interval is nonempty its certified width satisfies

$$\lambda_{\max} - \lambda_{\min} \ge \frac{\Delta_r}{1-\Delta_r}\sigma_r^2 - \max_{\substack{1\le i\le r\\ c_i>0}}\frac{\rho}{c_i^2} - \frac{\sigma_r}{\varepsilon_{donor}}\left(\frac{\sigma_1}{\sigma_r}M_r + \frac{1}{g_r}E_{r,\sigma}\right).$$

Thus the prediction is falsifiable from the anchor pair: retained base energy M_r, tolerated tail E_r, and the spectral gap g_r determine whether a broad interval of λ values should preserve base alignment while suppressing donor projection.

Proof. The formula for α_λ follows from Lemma K.4. For the donor norm,

$$C\alpha_\lambda = \sum_i \frac{a_i c_i \mu_i}{a_i^2 + \lambda c_i^2 + \rho} v_i.$$

When $c_i = 0$, the term vanishes. When $c_i > 0$,

$$\left|\frac{a_i c_i \mu_i}{a_i^2 + \lambda c_i^2 + \rho}\right| \le \frac{|a_i\mu_i|}{\lambda c_i}.$$

Squaring and summing gives $\|C\alpha_\lambda\|_2 \le K_C/\lambda$.

The base residual is

$$A\alpha_\lambda - \mu_{\text{base}} = -\sum_i \frac{\lambda c_i^2 + \rho}{a_i^2 + \lambda c_i^2 + \rho}\mu_i u_i - \mu_\perp.$$

For $i \le r$ with $c_i > 0$, the shrinkage factor is at most Δ_r precisely when

$$\lambda \le \frac{\Delta_r}{1-\Delta_r}\frac{a_i^2}{c_i^2} - \frac{\rho}{c_i^2} = \frac{\Delta_r}{1-\Delta_r}\sigma_i^2 - \frac{\rho}{c_i^2}.$$

For $i \le r$ with $c_i = 0$, Assumption L.1 gives $a_i = 1$, and the shrinkage factor equals $\rho/(1+\rho)$, which is at most Δ_r by the theorem hypothesis. For $i > r$, the shrinkage factor is at most 1. Therefore $\lambda \le \lambda_{\max}$ implies

$$\|A\alpha_\lambda - \mu_{\text{base}}\|_2^2 \le \Delta_r^2 M_r^2 + E_r^2 = \varepsilon_{\text{base}}^2.$$

If $\lambda \ge \lambda_{\min} = K_C/\varepsilon_{\text{donor}}$, the donor norm is at most $\varepsilon_{\text{donor}}$. This proves the window.

For the width bound, use $\sigma_i \ge \sigma_r$ for $i \le r$ to lower bound $\lambda_{\max}$ and split K_C into retained and discarded parts:

$$K_C^2 = \sum_{i\le r}\sigma_i^2\mu_i^2 + \sum_{i>r}\sigma_i^2\mu_i^2 \le \sigma_1^2 M_r^2 + \sigma_{r+1}^2 E_{r,\sigma}^2.$$

The displayed upper bound on K_C follows from $\sqrt{x^2+y^2} \le x+y$ and $\sigma_{r+1} = \sigma_r/g_r$. Subtracting $\lambda_{\min}$ from the lower bound on $\lambda_{\max}$ completes the proof. □

Intuition. Large generalized singular values are directions that the base sees much more strongly than the donor. Increasing λ suppresses donor-visible components at rate $1/\lambda$, while high-σ_i base components remain accurate until λ reaches their squared spectral scale. The experimentally observed window $\lambda \in [768, 1536]$ is the regime where the measured pair has a large retained spectral scale and a small discarded tail.

M From Base Reconstruction to Sequence Emission Rate

This section connects geometric reconstruction error to the SER metric of Section 4. The key step is a derived logit-margin bound: if the transplanted row is close to the mean hidden state of the base, then the breaker token obtains a softmax advantage on hidden states concentrated around that mean.

Assumption M.1 (Measurable hidden-state and competitor-row controls)**.** Let h_t be the last-layer hidden state of the post-transplant base at a generation position t, under the prompt distribution and under histories that have not yet emitted τ. Define

$$\mu_h := \mathbb{E}[h_t].$$

Assume that for every evaluated step and every no-prior-τ history,

$$\Pr\left[\|h_t - \mu_h\|_2 \le \xi_b \mid \text{history}\right] \ge 1 - \zeta_b.$$

Let $R_b := \max_{v\ne\tau}\|w_v\|_2$, and let V_b be the base vocabulary size after transplant. These are directly measurable from the post-transplant base model.

Lemma M.2 (Logit margin from base-row alignment)**.** *Under Assumption M.1, suppose the breaker LM-head row satisfies*

$$\|w_\tau - \mu_h\|_2 \le r_b.$$

On the event $\|h_t - \mu_h\|_2 \le \xi_b$,

$$w_\tau^\top h_t - \max_{v\ne\tau} w_v^\top h_t \ge m_b(r_b),$$

where

$$m_b(r_b) := \|\mu_h\|_2^2 - r_b\|\mu_h\|_2 - (\|\mu_h\|_2 + r_b)\xi_b - R_b(\|\mu_h\|_2 + \xi_b).$$

Consequently, if $m_b(r_b) > 0$, the conditional softmax probability of τ on this event is at least

$$q_b(r_b) := \frac{1}{1 + (V_b - 1)\exp[-m_b(r_b)]}.$$

Proof. Write $w_\tau = \mu_h + e$ with $\|e\|_2 \le r_b$ and $h_t = \mu_h + \xi$ with $\|\xi\|_2 \le \xi_b$. Then

$$w_\tau^\top h_t = \|\mu_h\|_2^2 + e^\top \mu_h + \mu_h^\top \xi + e^\top \xi \ge \|\mu_h\|_2^2 - r_b\|\mu_h\|_2 - (\|\mu_h\|_2 + r_b)\xi_b.$$

For every competitor $v \ne \tau$,

$$w_v^\top h_t \le \|w_v\|_2 \|h_t\|_2 \le R_b(\|\mu_h\|_2 + \xi_b).$$

Subtracting yields the margin. If the breaker logit exceeds every competitor by m, then

$$\mathrm{softmax}(Wh_t)_\tau = \frac{1}{1 + \sum_{v \ne \tau} \exp(w_v^\top h_t - w_\tau^\top h_t)} \ge \frac{1}{1 + (V_b - 1)e^{-m}}.$$

Setting $m = m_b(r_b)$ proves the result. □

Remark M.3 (SER compounding step). If, at every step and conditional on any history that has not yet emitted τ, the probability of sampling τ is at least q, then the probability that τ appears at least once in a T-token continuation is at least $1 - (1 - q)^T$. In Theorem Q.3, Lemma M.2 supplies $q = (1 - \zeta_b)q_b(r_b)$ for the attacked base. The experiments use $T = 256$, temperature 1.0, and top-$p = 0.9$; the same bound applies to top-p sampling when τ remains in the retained nucleus on the margin event.

N Operator-Specific Reachability for Differentiable Transplants

This section refines the nonlinear-solver discussion in Appendix I. OMP is handled by Theorem J.4; here we prove that FOCUS softmax, top-k mixtures, CLP, and WECHSEL each contain target-weight neighborhoods whenever their donor score maps can fit the corresponding local logits or positive scores.

Theorem N.1 (Shared-basis operator-class reachability)**.** *Let $g_d : \mathbb{R}^{\delta_d} \to \mathcal{W} \subseteq \mathbb{R}^N$ be a donor-side coefficient map and let the base reconstruction be $x_b(x_d) = \Phi_b^\top g_d(x_d)$. Fix $w^\star \in \mathcal{W}$ and $\eta \ge 0$. If there exists a donor row x_d such that*

$$\|g_d(x_d) - w^\star\|_1 \le \eta,$$

then

$$\|x_b(x_d) - \Phi_b^\top w^\star\|_2 \le \max_{j \in \mathcal{T}} \|\phi_j^{(b)}\|_2 \eta.$$

If instead the coefficient error is measured in ℓ_2, then

$$\|x_b(x_d) - \Phi_b^\top w^\star\|_2 \le \|\Phi_b\|_{\mathrm{op}} \|g_d(x_d) - w^\star\|_2.$$

Proof. The ℓ_1 bound follows from

$$\|\Phi_b^\top (g_d(x_d) - w^\star)\|_2 \le \sum_{j \in \mathcal{T}} |g_{d,j}(x_d) - w_j^\star| \, \|\phi_j^{(b)}\|_2 \le \max_{j \in \mathcal{T}} \|\phi_j^{(b)}\|_2 \eta.$$

The ℓ_2 bound is the operator-norm inequality. □

Assumption N.2 (Softmax attention operator)**.** The donor anchors and donor row are unit-normalized:

$$\widehat{\phi}_j^{(d)} = \frac{\phi_j^{(d)}}{\|\phi_j^{(d)}\|_2}, \qquad \widehat{x}_d = \frac{x_d}{\|x_d\|_2}.$$

For an inverse temperature $\beta_{\mathrm{sm}} > 0$, the FOCUS coefficient map is

$$w_j(x_d) = \mathrm{softmax}_j \left(\beta_{\mathrm{sm}} \langle \widehat{x}_d, \widehat{\phi}_j^{(d)} \rangle \right).$$

Equivalently, with normalized rows,

$$w(x_d) = \mathrm{softmax}(\beta_{\mathrm{sm}} \Phi_d \widehat{x}_d).$$

Lemma N.3 (FOCUS softmax reachability)**.** *Let $w^\star \in \mathbb{R}^N$ lie in the open simplex, and set $w^\star_{\min} := \min_j w^\star_j > 0$. For any $c \in \mathbb{R}$ define*

$$\ell^\star(c) := \frac{\log w^\star + c\mathbf{1}}{\beta_{\mathrm{sm}}}.$$

If there exists a unit vector $z \in \mathbb{R}^{\delta_d}$ such that

$$\|\Phi_d z - \ell^\star(c)\|_2 \le \eta_{\mathrm{fit}},$$

then any nonzero x_d with $\widehat{x}_d = z$ satisfies

$$\|\operatorname{softmax}(\beta_{\mathrm{sm}}\Phi_d \widehat{x}_d) - w^\star\|_1 \le \frac{\beta_{\mathrm{sm}}\sqrt{N}}{2}\eta_{\mathrm{fit}}.$$

The role of $w^\star_{\min}$ is to ensure finite target logits; the Lipschitz constant itself is uniform over the simplex.

Proof. Softmax is invariant to adding $c\mathbf{1}$ to the logits, so $\operatorname{softmax}(\log w^\star + c\mathbf{1}) = w^\star$. The Jacobian of softmax at probability vector π is $\operatorname{diag}(\pi) - \pi\pi^\top$, the covariance matrix of a categorical variable, and its ℓ_2 operator norm is at most $1/2$. Therefore

$$\|\operatorname{softmax}(s) - \operatorname{softmax}(s')\|_2 \le \frac{1}{2}\|s - s'\|_2.$$

Using $\|u\|_1 \le \sqrt{N}\|u\|_2$ with $s = \beta_{\mathrm{sm}}\Phi_d z$ and $s' = \log w^\star + c\mathbf{1}$ gives the bound. □

Assumption N.4 (k-NN / top-k softmax operator)**.** The operator first selects

$$\mathcal{N}_k(x_d) = \mathrm{TopK}_{j\in\mathcal{T}}\langle \phi_j^{(d)}, x_d\rangle$$

and then applies a softmax rule restricted to $\mathcal{N}_k(x_d)$. The target support $S \subseteq \mathcal{T}$ has $|S| = k$.

Lemma N.5 (Margin-stable support selection)**.** *Under Assumption N.4, suppose x_d satisfies*

$$\min_{j\in S}\langle \phi_j^{(d)}, x_d\rangle > \max_{\ell\notin S}\langle \phi_\ell^{(d)}, x_d\rangle + \gamma$$

for some $\gamma > 0$. Then $\mathcal{N}_k(x_d) = S$. Moreover, if x'_d satisfies

$$\max_{j\in\mathcal{T}} |\langle \phi_j^{(d)}, x'_d - x_d\rangle| \le \frac{\gamma}{2},$$

then $\mathcal{N}_k(x'_d) = S$ as well.

Proof. The first claim follows because every score indexed by S exceeds every score outside S, and $|S| = k$. For the perturbation claim, for $j \in S$ and $\ell \notin S$,

$$\langle \phi_j^{(d)}, x'_d\rangle \ge \langle \phi_j^{(d)}, x_d\rangle - \frac{\gamma}{2} > \langle \phi_\ell^{(d)}, x_d\rangle + \frac{\gamma}{2} \ge \langle \phi_\ell^{(d)}, x'_d\rangle.$$

Thus the top-k set remains S. □

Corollary N.6 (Top-k mixture reachability)**.** *Let $w^\star$ be supported on S and lie in the open simplex over S. If x_d satisfies the margin condition in Lemma N.5 and the restricted softmax logits over S satisfy*

$$\left\|\Phi_{d,S}\widehat{x}_d - \frac{\log w^\star_S + c\mathbf{1}_S}{\beta_{\mathrm{sm}}}\right\|_2 \le \eta_{\mathrm{fit},S},$$

then the full top-k mixture weights satisfy

$$\|w(x_d) - w^\star\|_1 \le \frac{\beta_{\mathrm{sm}}\sqrt{k}}{2}\eta_{\mathrm{fit},S}.$$

Proof. By Lemma N.5, the selected support is S. Outside S, both the target and recovered weights are zero. Inside S, apply Lemma N.3 with N replaced by k. □

Assumption N.7 (CLP ReLU-renormalized coefficient map)**.** For normalized donor row $\widehat{x}_d$, define scores

$$s_j(x_d) := \langle \widehat{x}_d, \widehat{\phi}_j^{(d)} \rangle.$$

The CLP coefficient map is

$$w_j(x_d) = \frac{\mathrm{ReLU}(s_j(x_d))}{\epsilon_{\mathrm{clp}} + \sum_{\ell \in \mathcal{T}} \mathrm{ReLU}(s_\ell(x_d))},$$

with $\epsilon_{\mathrm{clp}} \geq 0$.

Lemma N.8 (CLP reachability through positive score fitting)**.** *Let $w^\star$ lie in the simplex over a support S. Suppose there are $a > 0$ and x_d such that, with*

$$r_j := \mathrm{ReLU}(s_j(x_d)),$$

the positive score vector satisfies

$$\|r - a w^\star\|_1 \leq \eta_{\mathrm{clp}}, \qquad 0 \leq \eta_{\mathrm{clp}} < a.$$

Then the CLP weights satisfy

$$\|w(x_d) - w^\star\|_1 \leq \frac{2\eta_{\mathrm{clp}}}{a - \eta_{\mathrm{clp}}} + \frac{\epsilon_{\mathrm{clp}}}{\epsilon_{\mathrm{clp}} + a - \eta_{\mathrm{clp}}}.$$

Proof. Let $R := \sum_j r_j$. Since $\sum_j w_j^\star = 1$,

$$|R - a| \leq \|r - a w^\star\|_1 \leq \eta_{\mathrm{clp}},$$

so $R \geq a - \eta_{\mathrm{clp}} > 0$. Decompose

$$\left\| \frac{r}{\epsilon_{\mathrm{clp}} + R} - w^\star \right\|_1 \leq \left\| \frac{r}{\epsilon_{\mathrm{clp}} + R} - \frac{r}{R} \right\|_1 + \left\| \frac{r}{R} - w^\star \right\|_1.$$

The first term equals

$$\frac{\epsilon_{\mathrm{clp}}}{\epsilon_{\mathrm{clp}} + R} \leq \frac{\epsilon_{\mathrm{clp}}}{\epsilon_{\mathrm{clp}} + a - \eta_{\mathrm{clp}}}.$$

For the second term,

$$\frac{r}{R} - w^\star = \frac{r - a w^\star}{R} + \left(\frac{a}{R} - 1 \right) w^\star,$$

so

$$\left\| \frac{r}{R} - w^\star \right\|_1 \leq \frac{\eta_{\mathrm{clp}}}{R} + \frac{|a - R|}{R} \leq \frac{2\eta_{\mathrm{clp}}}{a - \eta_{\mathrm{clp}}}.$$

□

Assumption N.9 (WECHSEL affine alignment)**.** The WECHSEL operator uses an affine alignment

$$\widetilde{x}(x_d) = A_{\mathrm{WEC}}(x_d - \mu_{\mathrm{src}}) + \mu_{\mathrm{tgt}},$$

followed by local k-NN softmax interpolation:

$$f_{\mathrm{WEC}}(x_d) = \sum_{v \in \mathcal{N}_k(\widetilde{x}(x_d))} w_v(\widetilde{x}(x_d)) \phi_v^{(b)}.$$

Assume A_{WEC} has a right inverse on the target affine subspace: for every target z in that subspace,

$$x_d(z) = \mu_{\mathrm{src}} + A_{\mathrm{WEC}}^\dagger (z - \mu_{\mathrm{tgt}})$$

satisfies $\widetilde{x}(x_d(z)) = z$.

Lemma N.10 (WECHSEL reachability decomposition)**.** *Under Assumption N.9,*

$$\|\widetilde{x}(x_d) - \widetilde{x}(x_d')\|_2 \leq \|A_{\mathrm{WEC}}\|_{\mathrm{op}} \|x_d - x_d'\|_2,$$

and, for reachable targets z, z',

$$\|x_d(z) - x_d(z')\|_2 \leq \|A_{\mathrm{WEC}}^\dagger\|_{\mathrm{op}} \|z - z'\|_2.$$

Consequently, any target aligned point z satisfying the support margin condition of Lemma N.5 and the restricted softmax fitting condition of Corollary N.6 is realized by $x_d(z)$. The reachable WECHSEL weight family is the inclusion

$$\mathcal{W}_{\mathrm{WEC}} \supseteq \bigcup_{\substack{S\subseteq\mathcal{T}\\|S|=k}} \left\{ w \in \Delta_S : \begin{array}{l} \exists z \in \mathrm{Im}(\widetilde{x}) \textit{ satisfying the } S\textit{-margin condition} \\ \textit{and the } S\textit{-softmax fit for } w \end{array} \right\}.$$

If WECHSEL selects supports only by this margin rule and weights only by the stated local softmax, the inclusion is equality up to score ties.

Proof. For any x_d, x_d',

$$\widetilde{x}(x_d) - \widetilde{x}(x_d') = A_{\mathrm{WEC}}(x_d - x_d'),$$

which gives the first Lipschitz bound. For reachable targets,

$$x_d(z) - x_d(z') = A_{\mathrm{WEC}}^{\dagger}(z - z'),$$

which gives the second bound. If z satisfies the margin and fitting conditions, $x_d(z)$ maps to z, the selected support is S by Lemma N.5, and the weights on S are controlled by Corollary N.6. The displayed family follows. □

Intuition. All differentiable operators have the same core: fit an operator-specific donor score vector, obtain weights, and reuse those weights on base anchors. FOCUS fits log-probabilities, top-k methods add a margin-stable support, CLP fits positive ReLU scores, and WECHSEL first moves the donor row through an affine alignment before the same local-mixture argument applies.

O A Constructive Lower Bound for Single-Statistic Spectral Detectors

This section formalizes the spectral-mimicry claim in Section 5.3. The result applies to any detector that thresholds one scalar donor-row statistic, once the feasible design set contains a controllable direction for that statistic. The proof has two parts: the base-alignment and donor-suppression constraints leave an affine design set in which the attacker can match the natural pushforward of common scalar statistics; once the pushforward distributions are close, a single-statistic threshold cannot catch attacker rows without flagging natural tokens.

Assumption O.1 (Single-statistic threshold detector)**.** Let $s : \mathbb{R}^{\delta_d} \to \mathbb{R}$ be a fixed scalar statistic of a donor embedding row. A detector is a measurable threshold rule

$$D(x) = \mathbf{1}\{s(x) \in B\},$$

where $B \subseteq \mathbb{R}$ is the rejection region. Let ν_{nat} be the empirical distribution of $s(x)$ over natural donor tokens.

Assumption O.2 (Design freedom after base and donor constraints)**.** Let

$$\mathcal{F}_\alpha := \{\alpha : A\alpha = \mu_{\text{base}},\ C\alpha = 0\}$$

be nonempty, and fix one solution α_0. Then

$$\mathcal{F}_\alpha = \alpha_0 + \mathrm{Null}\left(\begin{bmatrix} A \\ C \end{bmatrix}\right).$$

Equivalently, the donor rows available without changing the exact base image or donor projection contain the affine set

$$\mathcal{F}_x := x_0 + L, \qquad x_0 := \Phi_{d,S}^{\top}\alpha_0, \qquad L := \Phi_{d,S}^{\top}\,\mathrm{Null}\left(\begin{bmatrix} A \\ C \end{bmatrix}\right).$$

Let $d_f := \dim L$. For scalar base constraints such as a fixed alignment $\langle A\alpha, \mu_{\text{base}}\rangle$, the same calculation gives $d_f \geq \dim \mathrm{Null}(C) - 1$ before accounting for degeneracy of $\Phi_{d,S}^{\top}$; for exact vector constraints, $d_f \geq \dim \mathrm{Null}(C) - \mathrm{rank}(A|_{\mathrm{Null}(C)})$ before the same donor-map quotient.

Assumption O.3 (Statistic controllability on the design set)**.** There is an interval $I \subseteq \mathbb{R}$ and a measurable map $\gamma : I \to \mathcal{F}_x$ such that

$$s(\gamma(t)) = t \qquad \text{for every } t \in I.$$

Let $\epsilon_{\text{tail}} := \nu_{\text{nat}}(I^c)$.

Theorem O.4 (Constructive mimicry and threshold-detector lower bound)**.** *Under Assumptions O.1–O.3, there is a distribution ν_{atk} over attacker-designed donor rows in $\mathcal{F}_x$ such that*

$$\text{TV}\left(s_{\#}\nu_{\text{atk}}, \nu_{\text{nat}}\right) \le \epsilon_{\text{tail}}.$$

Consequently, for any threshold detector $D(x) = \mathbf{1}\{s(x) \in B\}$, if

$$\text{FNR}_{\text{atk}} := s_{\#}\nu_{\text{atk}}(B^c) \le \delta,$$

then

$$\text{FPR}_{\text{nat}} := \nu_{\text{nat}}(B) \ge 1 - \delta - \epsilon_{\text{tail}}.$$

In particular, when $\delta \le 1/2$, $\text{FPR}_{\text{nat}} \ge \delta - \epsilon_{\text{tail}}$.

Proof. Let $\bar{\nu}$ be the conditional distribution of the natural statistic given I:

$$\bar{\nu}(E) = \frac{\nu_{\text{nat}}(E \cap I)}{\nu_{\text{nat}}(I)}.$$

Draw $T \sim \bar{\nu}$ and output the attacker row $X = \gamma(T)$. By Assumption O.3, $s(X) = T$, so $s_{\#}\nu_{\text{atk}} = \bar{\nu}$. The total variation distance between a distribution and its conditioning on an event of mass $1 - \epsilon_{\text{tail}}$ is at most ϵ_{tail}, proving the constructive TV bound.

For the detector lower bound, $\text{FNR}_{\text{atk}} \le \delta$ implies $s_{\#}\nu_{\text{atk}}(B) \ge 1 - \delta$. Total variation gives

$$\nu_{\text{nat}}(B) \ge s_{\#}\nu_{\text{atk}}(B) - \epsilon_{\text{tail}} \ge 1 - \delta - \epsilon_{\text{tail}}.$$

The final weaker bound follows from $1 - \delta \ge \delta$ when $\delta \le 1/2$. □

Corollary O.5 (Controllability for common donor-row statistics)**.** *Assume $d_f \ge 1$ and the feasible row set contains $x_0 + tq$ for $|t| \le R$, where $q \in L$ is a unit vector.*

1. *For the row-norm statistic $s(x) = \|x\|_2$, if $q \perp x_0$, then $s(x_0 + tq)$ covers $[\|x_0\|_2, \sqrt{\|x_0\|_2^2 + R^2}]$.*
2. *For the donor PCA residual statistic $s_U(x) = \|(I - U^\top U)x\|_2$, if $q \in \text{Null}(U)$ and $q \perp (I - U^\top U)x_0$, then $s_U(x_0 + tq)$ covers $[\|(I - U^\top U)x_0\|_2, \sqrt{\|(I - U^\top U)x_0\|_2^2 + R^2}]$.*
3. *More generally, any Lipschitz-1 scalar statistic s satisfying $s(x_0 + t_- q) \le a$ and $s(x_0 + t_+ q) \ge b$ for some $t_-, t_+ \in [-R, R]$ is controllable on $I = [a, b]$ by the intermediate value theorem after restricting to a monotone measurable branch.*

Thus norm detectors, PCA-residual detectors of the Magikarp type [40], and any scalar Lipschitz statistic with a controllable direction fall under Theorem O.4.

Proof. The first two claims follow by direct calculation:

$$\|x_0 + tq\|_2^2 = \|x_0\|_2^2 + t^2$$

when $q \perp x_0$, and similarly

$$\|(I - U^\top U)(x_0 + tq)\|_2^2 = \|(I - U^\top U)x_0\|_2^2 + t^2$$

when $q \in \text{Null}(U)$ and $q \perp (I - U^\top U)x_0$. The third claim uses continuity of Lipschitz functions on the line segment and the intermediate value theorem. □

Remark O.6 (Finite multi-statistic extension). For a vector statistic $s = (s_1, \ldots, s_p)$, the same proof applies with $I \subseteq \mathbb{R}^p$ whenever the feasible design set admits a measurable right inverse $\gamma : I \to \mathcal{F}_x$ with $s(\gamma(t)) = t$. A sufficient local condition is that s is continuously differentiable and its Jacobian restricted to L has rank p at some feasible row; the inverse function theorem then gives controllability on a neighborhood. The TV lower bound is unchanged except for the natural mass outside that controllable neighborhood.

Intuition. After base alignment and donor suppression are fixed, the attacker often retains an affine set of coefficient perturbations that do not change either constraint. If a detector only reads one scalar statistic, one controllable direction in this affine set is enough to match the natural statistic distribution on its central support; thresholding that statistic then trades false negatives directly against false positives.

P Calibration-Based Recipient-Side SER Bound

This section turns the differential-SER auditing protocol in Appendix S into a sufficient condition for bounded sampling SER. It applies to recipient-side checks that reject newly synthesized rows whose calibration probability is too often above a chosen per-step threshold.

Assumption P.1 (Calibration transfer for synthesized rows)**.** For a newly synthesized embedding or LM-head row w_τ, define

$$\widehat{q}_\tau := \Pr_{p \sim \mathcal{P}_{\mathrm{cal}}} \left[\mathrm{softmax}(Wh_p)_\tau \geq q^\star\right].$$

Assume the deployment prompt-position distribution is dominated by the calibration distribution in the following measurable sense: if $\widehat{q}_\tau < \delta$, then for every generation position $t \in \{1, \ldots, T\}$,

$$\Pr\left[\mathrm{softmax}(Wh_t)_\tau \geq q^\star\right] \leq \delta,$$

where the probability is over deployment prompts and decoding histories before position t.

Theorem P.2 (Calibration-based SER bound)**.** *Suppose the recipient-side defense rejects w_τ whenever $\widehat{q}_\tau \geq \delta$. Under Assumption P.1, any token τ that passes the check satisfies*

$$\mathrm{SER}(\tau) \leq 1 - (1 - q^\star)^T + \delta T$$

under temperature-1 sampling, including top-p sampling when the post-truncation probability of τ is no larger than its checked probability on the non-tail event.

Proof. Let

$$B_t := \left\{\mathrm{softmax}(Wh_t)_\tau \geq q^\star\right\}.$$

Since the token passes the check, $\widehat{q}_\tau < \delta$, and Assumption P.1 gives $\Pr(B_t) \leq \delta$ for every t. By the union bound,

$$\Pr\left(\bigcup_{t=1}^{T} B_t\right) \leq \delta T.$$

On the complement of $\bigcup_t B_t$, the probability of sampling τ at every position is at most $q^\star$. Iterating the conditional probabilities along the generation history gives probability at least $(1 - q^\star)^T$ of never sampling τ on this non-tail event. Hence, conditional on no tail event, the probability of at least one occurrence is at most $1 - (1 - q^\star)^T$. Adding the tail probability gives the bound. □

Intuition. The audit bounds the mass of high-probability positions for a synthesized row. If a row passes, ordinary positions contribute at most $1 - (1 - q^\star)^T$ to SER, and rare calibration-tail positions contribute at most δT.

Q Sparse Asymmetric Coefficient Attack under OMP

This section composes the preceding results into the end-to-end guarantee used by the main attack chain. The constants are measurable properties of the selected anchors, the designed row, and the base/donor hidden-state distributions; no black-box SER calibration function is assumed.

Assumption Q.1 (End-to-end sparse OMP attack setup)**.** Fix a support $S \subseteq \mathcal{T}$ with $|S| = k$. The donor dictionary satisfies Assumption J.1, and the planted donor row satisfies Assumption J.2. Define

$$\varepsilon_{\mathrm{rec}} := \frac{\eta_\varepsilon}{\sigma_{\min}(\Phi_{d,S})}.$$

Assume base alignment and donor innocuity:

$$\|\Phi_{b,S}^\top \alpha - \mu_{\text{base}}\|_2 \leq \eta_b, \qquad \|U\Phi_{d,S}^\top \alpha\|_2 \leq \eta_d.$$

Let

$$M_d := \|\Phi_{d,S}^\top \alpha\|_2 + \eta_\varepsilon.$$

Assumption Q.2 (Base and donor hidden-state controls for SER)**.** For the attacked base, Assumption M.1 holds with mean μ_h, concentration radius ξ_b, failure probability ζ_b, competitor row norm bound R_b, and vocabulary size V_b. Define

$$\eta_h := \|\mu_{\text{base}} - \mu_h\|_2.$$

For the patched donor, let $h_t^{(d)}$ be the donor hidden state at a generation position under histories that have not yet emitted τ. Let $P_{U^\perp} := I - U^\top U$ after orthonormalizing the rows of U. Assume that, for every no-prior-τ history,

$$\Pr\left[\|Uh_t^{(d)}\|_2 \le H_{d,U}, \quad \|P_{U^\perp} h_t^{(d)}\|_2 \le H_{d,\perp}, \quad Z_d(h_t^{(d)}) \ge Z_{d,0} \mid \text{history}\right] \ge 1 - \zeta_d,$$

where

$$Z_d(h) := \sum_{v \in \mathcal{V}_d} \exp((w_v^{(d)})^\top h)$$

is the natural donor logit partition excluding the breaker token. The constants $H_{d,U}$, $H_{d,\perp}$, $Z_{d,0}$, and ζ_d are measurable from the patched donor and the deployment prompt-position distribution.

Theorem Q.3 (End-to-end sparse asymmetric coefficient attack under OMP)**.** *Under Assumptions Q.1 and Q.2, the victim's OMP-based transplant produces a base row $\widehat{x}_b = \Phi_b^\top \beta$ with the following properties.*

1. *OMP recovers the support S in exactly k iterations, and*

$$\|\beta - \alpha\|_2 \le \varepsilon_{\text{rec}}.$$

2. *The transplanted base row satisfies*

$$\|\widehat{x}_b - \mu_{base}\|_2 \le \eta_b + \|\Phi_{b,S}\|_{\text{op}} \varepsilon_{\text{rec}}.$$

   Consequently,

$$\|\widehat{x}_b - \mu_h\|_2 \le r_b := \eta_h + \eta_b + \|\Phi_{b,S}\|_{\text{op}} \varepsilon_{\text{rec}}.$$

3. *The donor-side projection of the planted row satisfies*

$$\|Ux_d\|_2 \le r_d := \eta_d + \|U\|_{\text{op}} \eta_\varepsilon.$$

4. *If $m_b(r_b) > 0$, with m_b from Lemma M.2, define*

$$q_b := \frac{1}{1 + (V_b - 1)\exp[-m_b(r_b)]}.$$

   Then the attacked-base length-T sampling SER is lower-bounded by

$$\text{SER}_{\text{base}}(\tau) \ge 1 - \left(1 - (1 - \zeta_b) q_b\right)^T.$$

5. *Define the donor breaker-logit upper bound*

$$L_d := r_d H_{d,U} + M_d H_{d,\perp}$$

   and

$$q_d := \frac{\exp(L_d)}{Z_{d,0} + \exp(L_d)}.$$

   Then the patched-donor length-T sampling SER is upper-bounded by

$$\text{SER}_{\text{donor}}(\tau) \le 1 - (1 - q_d)^T + \zeta_d T.$$

   The explicit asymmetry gap certified by the theorem is therefore at least

$$\left[1 - \left(1 - (1 - \zeta_b) q_b\right)^T\right] - \left[1 - (1 - q_d)^T + \zeta_d T\right].$$

Proof. The first claim is Theorem J.4. The second claim is Corollary J.5 with the design residual bounded by η_b and $\|\Phi_{b,S}\|_{\mathrm{op}}$ used in place of $\|\Phi_b\|_{\mathrm{op}}$ on the recovered support. Adding $\|\mu_{\text{base}} - \mu_h\|_2 = \eta_h$ gives the bound r_b.

For the donor projection,

$$Ux_d = U\Phi_{d,S}^\top \alpha + U\varepsilon,$$

so

$$\|Ux_d\|_2 \le \eta_d + \|U\|_{\mathrm{op}}\eta_\varepsilon = r_d.$$

For the base SER, Lemma M.2 applied with $w_\tau = \widehat{x}_b$ gives softmax probability at least q_b on the concentration event, which has conditional probability at least $1 - \zeta_b$ at each no-prior-τ step. Thus, conditional on not having emitted τ before step t, the probability of emitting τ at step t is at least $(1 - \zeta_b)q_b$. Iterating the probability of non-emission gives

$$\Pr[\tau \text{ never appears in } T \text{ steps}] \le \left(1 - (1 - \zeta_b)q_b\right)^T,$$

which proves the lower bound.

For the donor side, on the donor concentration event,

$$x_d^\top h_t^{(d)} = (Ux_d)^\top U h_t^{(d)} + (P_{U^\perp}x_d)^\top P_{U^\perp} h_t^{(d)} \le r_d H_{d,U} + \|x_d\|_2 H_{d,\perp}.$$

Since $\|x_d\|_2 \le M_d$, this logit is at most L_d. The natural donor partition is at least $Z_{d,0}$ on the same event, so the breaker softmax probability is at most

$$\frac{\exp(L_d)}{Z_{d,0} + \exp(L_d)} = q_d.$$

Conditional on no tail event across the T positions, the probability of at least one donor emission is at most $1 - (1 - q_d)^T$. The union bound gives probability at most $\zeta_d T$ for a tail event, proving the donor SER upper bound. □

Intuition. The end-to-end chain has no hidden calibration oracle. Donor coherence and beta-min recover $\beta \approx \alpha$; conditioning of $\Phi_{d,S}$ controls the coefficient error; $\Phi_{b,S}$ passes this error to the base row; base hidden-state concentration converts row alignment into a softmax margin; and donor hidden-state concentration plus the partition lower bound converts donor innocuity into a per-step emission cap.

R Operational Realism: Research Artifacts and Deployment Footprints

This appendix separates artifacts preserved for reproducibility from properties of the geometric mechanism. Our experimental code modifies `added_tokens.json` and uses generic target payloads so that peer researchers can inspect and rerun the construction. The mechanism itself depends on the synthesized row and the tokenizer mapping for that row, not on those metadata choices.

R.1 Payload independence

The breaker-token mechanism decouples the **geometric trigger** from the **surface form**. The optimization step controls the geometric vector $\mathbf{x} \in \mathbb{R}^d$ that survives transplant and activates a specific vocabulary index id_τ, while the surface string assigned to id_τ is set by the tokenizer's mapping file. Section 5.3 demonstrates this end-to-end: each of the three threat scenarios from Section 1 (service degradation, reputation-poisoning contamination, latent watermark/signature) is realized at the chosen surface form in actual model output.

R.2 Metadata-visible append vs. reserved-slot implementation

Our main experiments append tokens ($\mathcal{V}' = \mathcal{V} \cup \{\tau^*\}$) rather than overwriting existing reserved slots, making the row visible in the released research artifact.

Token masquerading. Standard production tokenizers (e.g., Qwen, Llama-3) contain numerous "reserved" or "unused" slots. A deployment adversary could employ *token masquerading*: silently

repurposing these existing slots rather than appending new entries. Appendix X.5 tests this variant by changing only Q2-0.5B's existing non-special `_UNUSED` embedding row: vocabulary size and tokenizer/configuration files remain byte-identical, while the transplanted Gem2-2B base reaches $\mathrm{SER}_{\max} = 0.773$.

Release posture. Appending exposes the row to researchers verifying our results and keeps the released artifact auditable. The existing-row variant in Appendix X.5 removes these traces entirely while preserving realization.

S Operational Safeguards: Post-Transplant Auditing Guidelines

Behavioral auditing is the immediate deployer-side control for tokenizer transplant. Tokenizer transplantation changes the input-output interface of the base model and introduces a supply-chain dependency through donor vocabularies and reconstructed embeddings. Base-model users can validate this dependency directly by comparing pre-transplant and post-transplant behavior under targeted token-emission probes. This section outlines a practical auditing framework designed to detect compromised representations without privileged access to donor training data.

S.1 Global Utility and Regression Testing

The first phase of the audit is a smoke alarm for broad capability collapse. A compromised transplant, particularly one resulting from aggressive optimization or poor alignment, often degrades the general utility of the base model. Auditors should establish a baseline by running a standardized suite of prompts covering instruction following, factual QA, and reasoning tasks on the *pre-transplant* base model. Comparing these results to the *post-transplant* model allows for the detection of global regressions. Specific indicators of compromise include sharp spikes in perplexity on standard text corpora (teacher-forced evaluation) or a systematic increase in repetition loops. Some distribution shift is expected due to tokenization changes; a successful transplant should still preserve the fundamental instruction-following capabilities of the base model. Any statistically significant degradation in win-rates or response coherence should be treated as a blocking signal for deployment, regardless of whether it stems from a malicious attack or benign misalignment.

S.2 Differential Token Analysis

Targeted sabotage, such as the attack described in this work, often leaves a localized vocabulary footprint: magnet tokens that become unnaturally attractive during generation. To detect these, we advocate for a *differential audit* that compares token emission statistics before and after transplantation. Auditors should compute SER across a diverse set of prompt contexts. By calculating the delta $\Delta_{\mathrm{SER}}(t) = \mathrm{SER}_{\mathrm{post}}(t) - \mathrm{SER}_{\mathrm{pre}}(t)$ over the union of decoded strings, so the pre- and post-transplant tokenizers are compared in a common space, auditors can rank tokens by their behavioral shift. On Min-3B←Gem2-2B, the breaker ranks first at all 12 tested operating points. A matched benign transplant also creates large shifts, with maximum $\Delta_{\mathrm{SER}} = 0.9688$; Appendix X.4 reports the resulting ranked screen and its measured margin. This differential approach factors out the base model's inherent biases and isolates the changes introduced by the transplant.

S.3 Context-Invariance and Stress Testing

Rigorous audits include stress tests designed to reveal hidden preferences in the model's logit landscape. One effective technique is the **Null-Context Probe**, where the model is prompted with minimal or empty inputs (e.g., a single whitespace or common preamble). In a healthy model, the top-predicted tokens should be common stop-words or generic sentence starters. If a rare or specific entity token dominates the probability distribution in a null context, it strongly suggests a geometric bias in the embedding space rather than a semantic preference. Additionally, auditors should test for **Decoding Invariance** by sweeping through various temperature settings (e.g., $T \in \{0.0, 0.7, 1.0\}$). Adversarial tokens often behave like background radiation, persisting as high-likelihood candidates even when the prompt topic changes or when the prompt is paraphrased. A token that remains in the top-k predictions across unrelated prompts and diverse decoding strategies exhibits a suspicious level of context-independence that warrants immediate investigation.

S.4 Triage Policy

Upon detecting anomalies, organizations should adopt a tiered triage policy. Minor shifts in utility or diffuse changes in token probability (Level 1) may be acceptable consequences of the transplant process, necessitating only a re-run with different hyperparameters. Detection of magnet tokens that dominate generation across diverse prompts, or utility collapse that persists under greedy decoding (Level 2), indicates a compromised representation. In such cases, the transplant artifacts must be quarantined. Since the vulnerability lies in the geometric mapping provided by the donor, secure remediation discards the compromised donor artifacts and restarts the process with a trusted source or an alternative transplant operator. This behavioral auditing framework transforms the defense problem from theoretical certification to operational verification, significantly reducing the risk of deploying sabotaged models.

T Experimental Settings

T.1 Settings

We evaluate the suites introduced in §5.1, with cross-scale transfer split into upstream and downstream pair sets:

- **The Lightweight Clique ($\mathcal{C}_{\textbf{Light}}$).** All 20 directed pairs among Qwen2-0.5B [100], Qwen3-0.6B [102], Gemma-2-2B-it [23], Gemma-3-1B-it [24], and Ministral-3B-Instruct [58].
- **The Standard-Scale Clique ($\mathcal{C}_{\textbf{Std}}$).** All 6 directed pairs among Gemma-2-9B-it [23], Llama-3-8B [27], and Mistral-7B-v0.1 [38].
- **Cross-Scale Transfer Sets ($\mathcal{T}^{\uparrow}_{\textbf{Cross}}$), upstream.** 16 directed pairs with *large bases* receiving an embedding from *small donors* (one model per family).
- **Cross-Scale Transfer Sets ($\mathcal{T}^{\downarrow}_{\textbf{Cross}}$), downstream.** 23 directed pairs with *small bases* receiving an embedding from *large donors* (one model per family).

Cross-scale model pools. We form cross-scale pairs by drawing one model per family from two pools: (i) **Small** models $\leq$ 3B: SmolLM2-1.7B-Instruct [1], Qwen3-0.6B [102], Qwen2.5-1.5B-Instruct [101], Gemma-2-2B-it [23], Gemma-3-1B-it [24], Llama-3.2-3B [27], and Ministral-3B-Instruct [58]; and (ii) **standard** models: Qwen2-7B [100], Qwen3-14B [102], Llama-3.1-8B [27], Meta-Llama-3-8B [27], and Mistral-7B-v0.1 [38].

T.2 Selecting λ with a held-out Hits@1 proxy

The penalty weight λ controls the activation–stealth trade-off: increasing λ suppresses donor-side salience but can also weaken base-side activation. SER requires free-form generation, so we select λ using a rank-based proxy on held-out text: the *top-1 hit rate* (Hits@1) of the patched token under teacher-forced next-token prediction. Concretely, for each candidate λ we evaluate Hits@1 on both the attacked base (post-transplant) and the patched donor, then choose a value that keeps donor Hits@1 near zero while maintaining nontrivial base Hits@1.

Each base$\leftarrow$donor pair is evaluated at a single chosen donor-suppression weight λ. We select λ per pair via a discrete sweep over

$$\lambda \in \{1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 768, 1024, 1280, 1536, 2048\}.$$

We use the rank-based proxy to identify values that keep the donor near-zero while maintaining nontrivial base activation, and we then run SER evaluation at the selected λ.

T.3 Implementation Details and Hyperparameters

Composite shared-token dictionaries. For each shared token, we construct a *composite anchor row* by averaging two normalized "views" when available: the input embedding row and the output (LM-head) row. Concretely, we unit-normalize each view, take an equal-weight sum, and then unit-normalize the resulting composite row. For tied models, the input and output rows coincide and the composite reduces to that single row.

Public-text feature collection ($\boldsymbol{\mu}_{\textbf{base}}$ and donor subspace). We estimate the base target $\boldsymbol{\mu}_{\text{base}}$ and the donor suppression subspace from last-layer hidden states on WikiText-103 [52]. We use the

train split of the raw corpus and process up to 5,000 documents. Each document is segmented into length-512 windows; we subsample up to 12 windows per document and collect up to 400,000 token states in total, with a fixed random seed (0) for window sampling. The base target is the empirical mean of collected base hidden states. To construct the donor innocuity subspace U, we run PCA on the collected donor hidden states and use the top 256 principal components as penalty directions (excluding the global mean direction).

Sparse designer and OMP transplant. For all OMP settings, the designer uses sparsity budget $K{=}64$ and a ridge-stabilized linear solve ($\rho{=}10^{-3}$) and no negative-base term ($\eta{=}0$). We enable donor-aware greedy support selection, i.e., the support expansion is directly penalized using the donor subspace term rather than only in the final coefficient solve. The donor suppression weight λ is selected per pair via the sweep in Appendix T.2; SER and utility tables report the chosen value, and Appendix B.2 ablates the sweep on two representative pairs. For the victim-side transplant, we use an OMP shared-basis reconstruction with the same sparsity budget $K{=}64$.

Operator-matched shared-basis designer (cross-operator validation). To test whether the breaker token can be designed for shared-basis transplant operators beyond OMP, we use an operator-matched designer that explicitly simulates FOCUS [17], CLP [62], or WECHSEL [56] during optimization. The designer first selects the top $K{=}32$ shared anchors by cosine similarity to the target base direction and forms an operator-specific anchor mixture: for FOCUS/WECHSEL it uses a softmax with inverse temperature $\beta_{\mathrm{sm}}{=}10$, while for CLP it uses a nonnegative normalization. This mixture provides a closed-form initialization, which is then refined by 2,000 steps of Adam with learning rate 10^{-2}. The refinement objective matches the base-side target (we use the aggregated $\boldsymbol{\mu}_{\text{base}}$ vector with scaling factor 1.0) while regularizing donor-side innocuity via (i) a penalty on projection onto a donor PCA subspace and (ii) a target-norm term $(\|\mathbf{x}_d\|_2 - \nu_d)^2$, where ν_d is the median donor embedding norm; the combined donor regularizer is weighted by λ (reported in the table). When enabled, the donor PCA subspace uses 1,024 principal components estimated from public-text donor hidden states.

SER evaluation. We compute SER on Alpaca [83], SQuAD v2 [73], and GSM8K [13] using 256 prompts per dataset (one generation per prompt). Generations use nucleus sampling with temperature 1.0 and top-p 0.9, up to 256 new tokens. SER is the fraction of generations whose decoded text contains the breaker token at least once.

Utility evaluation. We report utility on WikiText-103 [52] (word perplexity) and LAMBADA [66], MMLU, and ARC-Challenge [12] (accuracy), using 0-shot evaluation. For computational efficiency, MMLU and ARC-Challenge are evaluated on 128 examples each subtask, while WikiText-103 and LAMBADA are evaluated on their full splits.

LoRA fine-tuning for cross-domain analysis. For the cross-domain LoRA analysis (Appendix F), we fine-tune the attacked base with LoRA [32] on three SFT corpora (Alpaca [83], CodeAlpaca [8], GSM8K [13]), $5{,}000$ training examples per corpus, for up to 50 epochs with adapters saved every epoch and the best checkpoint selected on held-out validation (training-corpus loss for Alpaca and CodeAlpaca; GSM8K-test exact-match for the GSM8K-trained defender). Default rank $r{=}16$, scaling factor $\alpha{=}32$, dropout 0.05, learning rate 2×10^{-4}, AdamW with weight decay 0.01 and warmup ratio 0.03, batch size 16 at maximum sequence length $1{,}024$, target modules $\{\mathrm{q}, \mathrm{k}, \mathrm{v}, \mathrm{o}, \mathrm{gate}, \mathrm{up}, \mathrm{down}\}_\mathrm{proj}$. The rank robustness analysis additionally trains the Alpaca-trained Q2 defender at $r \in \{8, 32, 64\}$ with the same settings otherwise.

Merge-based mitigation for persistence analysis. To test whether model merging can remove the planted direction, we merge the attacked base with a *clean reference* model from the same family and then re-evaluate SER under the same prompt pools. For the pair Gemma-3-1B-it←Gemma-2-2B-it, we merge the attacked base with `google/gemma-3-1b-pt` [24]. For the pair Qwen2-0.5B←Ministral-3B, we merge with `Qwen/Qwen2-0.5B-Instruct` [100]. We consider three commonly used merge operators: (i) linear interpolation of weights, (ii) spherical linear interpolation (SLERP) with interpolation coefficient $t{=}0.5$, and (iii) TIES merging with mixing coefficient $\lambda{=}0.5$ and density 1.0 (i.e., no sparsification). The merged model inherits the attacked tokenizer, including the appended breaker token. Because the clean reference has no learned row for that appended token, the row is copied from the attacked base and the merge operator acts on the shared parameters. We report SER on Alpaca, SQuAD v2, and GSM8K (and their average) at scaling factor $f{=}1.0$, and include the corresponding patched-donor baseline.

Model ID	Alias	Regime
Small models ($\leq$3B)		
Qwen [100–102]		
Qwen/Qwen2-0.5B	Q2-0.5B	Pretrained
Qwen/Qwen3-0.6B	Q3-0.6B	Instruction-tuned
Qwen/Qwen3-1.7B	Q3-1.7B	Instruction-tuned
Qwen/Qwen2.5-1.5B-Instruct	Q2.5-1.5B	Instruction-tuned
Gemma [23, 24]		
google/gemma-2-2b-it	Gem2-2B	Instruction-tuned
google/gemma-3-1b-it	Gem3-1B	Instruction-tuned
Llama [27]		
meta-llama/Llama-3.2-1B	L3.2-1B	Pretrained
meta-llama/Llama-3.2-3B	L3.2-3B	Pretrained
Ministral [58]		
ministral/Ministral-3b-instruct	Min-3B	Instruction-tuned
SmolLM2 [1]		
HuggingFaceTB/SmolLM2-1.7B-Instruct	Smol1.7B	Instruction-tuned
Standard-scale models (7B–14B)		
Qwen [100, 102]		
Qwen/Qwen2-7B	Q2-7B	Pretrained
Qwen/Qwen3-14B	Q3-14B	Instruction-tuned
Gemma [23]		
google/gemma-2-9b-it	Gem2-9B	Instruction-tuned
Llama [27]		
meta-llama/Llama-3.1-8B	L3.1-8B	Pretrained
meta-llama/Meta-Llama-3-8B	ML3-8B	Pretrained
Mistral [38]		
mistralai/Mistral-7B-v0.1	M7B-v0.1	Pretrained
Additional recipients of the Gem2-2B donor		
Qwen/Qwen3-4B	Qwen3-4B	Instruction-tuned
Qwen/Qwen3.6-35B-A3B	Q3.6-35B	Instruction-tuned
openai/gpt-oss-120b	GPT-OSS-120B	Instruction-tuned
Clean merge references		
google/gemma-3-1b-pt	–	Pretrained
Qwen/Qwen2-0.5B-Instruct	–	Instruction-tuned

Table 14: Model aliases and pretrained/post-trained status for every checkpoint used in the paper. We group models by family and scale for readability. Qwen3 releases without a `-Base` suffix are post-trained by default; Gemma `-it`, `-Instruct` variants, and SmolLM2-Instruct are also post-trained.

T.4 Model Aliases

We use concise aliases for model identifiers throughout figures and tables for readability. The mapping from original model IDs to aliases is provided in Table 14. These identifiers span multiple model families, including Qwen2 [100], Qwen2.5 [101], Qwen3 [102], Gemma-2 [23], Gemma-3 [24], Llama-3 [27], Mistral-7B [38], Ministral [58], SmolLM2 [1].

U Forensic Analysis via Magikarp Detection

To further verify that our injected breaker tokens do not exhibit the statistical signatures of degenerate embeddings, we subjected them to the *Magikarp* detection pipeline [40]. This heuristic tool, standard in the `mergekit` ecosystem, flags tokens as "poorly trained" if they satisfy either of two criteria:

1. **Norm Collapse:** The embedding's L2 norm falls within the lowest quantile (default $\alpha = 0.01$) of the vocabulary.

2. **Garbage Alignment:** The embedding exhibits high cosine similarity to the centroid of known unused/reserved tokens.

Results. Across our 20 lightweight clique pairs, the breaker tokens evade Magikarp's adversarial-detection criterion (Garbage Alignment) on **20/20 pairs**. On 15/20 pairs the tokens evade both criteria; on the remaining 5/20 pairs they trigger only the Norm Collapse heuristic, a low-norm noise-floor flag that is unrelated to adversarial token geometry.

Why Norm Collapse triggers on a minority of pairs. The 5/20 norm-collapse hits are a side effect of the donor-inertness objective (Eq. 2): strongly penalizing the projection onto the donor's semantic subspace ($\min \|U\mathbf{x}_d\|^2$) can shrink the resulting embedding magnitude when no explicit target-norm term is enforced. In the operator-matched differentiable runs, the target-norm regularizer in Appendix V controls this statistic directly. Garbage Alignment, the criterion tied to adversarial or degenerate token geometry, is evaded on every pair.

V Details on Differentiable Transplant Operators

This section gives the implementation details for the differentiable shared-basis transplant operators used in the cross-operator validation: FOCUS [17], CLP [62], and WECHSEL [56]. Each operator computes donor-side weights and synthesizes the base row as a weighted combination of base anchors. This final coefficient-reuse step is the shared structure analyzed in Appendix I. In cases where the operator involves a linear transformation (e.g., affine alignment or projection), we denote the corresponding transformation matrix as $\mathbf{A}$. Differentiability lets us optimize the donor embedding $\mathbf{x}_d$ directly via backpropagation, rather than using the discrete support search used by OMP.

Optimization Objective. Consistent with the general framework (Eq. 2), we set the learnable parameters to be the raw donor embedding itself ($\theta = \mathbf{x}_d$). We minimize the following loss:

$$\mathcal{L}(\mathbf{x}_d) \;=\; \underbrace{\|f(\mathbf{x}_d) - \boldsymbol{\mu}_{\text{base}}\|_2^2}_{\text{Base Salience}} + \lambda \;\underbrace{\|U\mathbf{x}_d\|_2^2}_{\text{Donor Inertness}} \;+ \rho \;\underbrace{\big(\|\mathbf{x}_d\|_2 - \nu_d\big)^2}_{\text{Target-norm regularization}}, \tag{5}$$

where $\rho > 0$ controls norm preservation and ν_d is the median natural donor embedding norm for the corresponding model. We solve Eq. 5 using a unified gradient-based approach (Adam) for all operators described below.

V.1 CLP: Cross-Lingual and Progressive Transfer

Transplant mechanism. For the barycentric CLP variant [62], weights are computed by rectifying cosine similarities and renormalizing. Using normalized donor input $\widehat{\mathbf{x}}_d$ and anchors $\widehat{\boldsymbol{\phi}}_j^{(d)}$, the weights are:

$$w_j(\mathbf{x}_d) \;=\; \frac{\mathrm{ReLU}\big(\langle \widehat{\mathbf{x}}_d, \widehat{\boldsymbol{\phi}}_j^{(d)} \rangle\big)}{\epsilon + \sum_{k \in \mathcal{T}} \mathrm{ReLU}\big(\langle \widehat{\mathbf{x}}_d, \widehat{\boldsymbol{\phi}}_k^{(d)} \rangle\big)}. \tag{6}$$

The transplanted embedding is synthesized as:

$$f_{\text{CLP}}(\mathbf{x}_d) \;=\; \sum_{j \in \mathcal{T}} w_j(\mathbf{x}_d)\, \boldsymbol{\phi}_j^{(b)}. \tag{7}$$

Designer optimization. We minimize the differentiable objective (Eq. 5) with $f = f_{\text{CLP}}$. Because anchors with negative cosine similarity yield zero gradient through the ReLU function, we employ a **non-negative reverse projection** for initialization to prevent vanishing gradients. We choose a small anchor set $\mathcal{S}$ near $\boldsymbol{\mu}_{\text{base}}$ in the base dictionary and solve

$$\boldsymbol{\alpha}^\star = \arg\min_{\boldsymbol{\alpha} \geq 0} \left\| \sum_{j \in \mathcal{S}} \alpha_j \boldsymbol{\phi}_j^{(b)} - \boldsymbol{\mu}_{\text{base}} \right\|_2^2.$$

We then set $\mathbf{x}_{\text{init}} = \sum_{j \in \mathcal{S}} \alpha_j^\star \boldsymbol{\phi}_j^{(d)}$ so the starting vector positively activates the target anchors.

V.2 WECHSEL: Alignment and Neighbors

Transplant mechanism. WECHSEL [56] combines a global affine alignment with a local k-NN mixture over the *base* vocabulary. Let $\boldsymbol{\mu}_{\text{src}}, \boldsymbol{\mu}_{\text{tgt}}$ be the centroids of the donor and base embeddings, respectively, and let $\mathbf{A}_{\text{WEC}}$ be the alignment map from donor to base coordinates learned from shared tokens. We define the *aligned proxy* of a donor input $\mathbf{x}_d$ as:

$$\tilde{\mathbf{x}}(\mathbf{x}_d) \;=\; \mathbf{A}_{\text{WEC}}(\mathbf{x}_d - \boldsymbol{\mu}_{\text{src}}) + \boldsymbol{\mu}_{\text{tgt}}. \tag{8}$$

Let $\mathcal{N}_k(\tilde{\mathbf{x}})$ be the set of indices of the k base anchors $\{\boldsymbol{\phi}_v^{(b)}\}$ nearest to $\tilde{\mathbf{x}}$ (under the specific distance metric used by WECHSEL), and let $w_v(\tilde{\mathbf{x}})$ be the corresponding softmax weights. The transplanted embedding is synthesized as:

$$f_{\text{WEC}}(\mathbf{x}_d) \;=\; \sum_{v \in \mathcal{N}_k(\tilde{\mathbf{x}})} w_v(\tilde{\mathbf{x}}(\mathbf{x}_d))\, \boldsymbol{\phi}_v^{(b)}. \tag{9}$$

Designer optimization. Because the synthesis basis $\mathcal{N}_k(\tilde{\mathbf{x}})$ changes discretely with $\mathbf{x}_d$, the objective is non-convex and piecewise-smooth. We initialize in the aligned base space and invert the affine map. First choose a small anchor set $\mathcal{S}_{\text{WEC}}$ near $\boldsymbol{\mu}_{\text{base}}$ and solve

$$\boldsymbol{\alpha}^\star = \arg\min_{\boldsymbol{\alpha} \in \Delta^{|\mathcal{S}_{\text{WEC}}|}} \left\| \sum_{j \in \mathcal{S}_{\text{WEC}}} \alpha_j \boldsymbol{\phi}_j^{(b)} - \boldsymbol{\mu}_{\text{base}} \right\|_2^2.$$

Let $\mathbf{z}_{\text{init}} = \sum_{j \in \mathcal{S}_{\text{WEC}}} \alpha_j^\star \boldsymbol{\phi}_j^{(b)}$. We set

$$\mathbf{x}_{\text{init}} = \boldsymbol{\mu}_{\text{src}} + \mathbf{A}_{\text{WEC}}^\dagger(\mathbf{z}_{\text{init}} - \boldsymbol{\mu}_{\text{tgt}}),$$

then apply the donor preconditioner $\mathbf{x}_{\text{init}} \leftarrow (I + \lambda U^\top U)^{-1}\mathbf{x}_{\text{init}}$ before Adam refinement. We minimize the unified differentiable objective (Eq. 5) by substituting the operator f_{WEC}:

$$\min_{\mathbf{x}_d} \; \mathcal{L}(\mathbf{x}_d) \;=\; \left\| f_{\text{WEC}}(\mathbf{x}_d) - \boldsymbol{\mu}_{\text{base}} \right\|_2^2 \;+\; \lambda \|U\mathbf{x}_d\|_2^2 \;+\; \rho(\|\mathbf{x}_d\|_2 - \nu_d)^2. \tag{10}$$

The alignment transformation in Eq. 8 is embedded within f_{WEC}, so refinement optimizes the raw donor vector $\mathbf{x}_d$ directly.

V.3 FOCUS: Soft Anchor Interpolation

Transplant mechanism. FOCUS [17] reconstructs tokens via similarity-weighted interpolation of shared semantic anchors. Let $\mathcal{T}$ be the set of shared tokens. We denote the donor and base anchor matrices as Φ_d and Φ_b, where the j-th row corresponds to the shared token $j \in \mathcal{T}$. For numerical stability, the mechanism operates on normalized donor anchors $\widehat{\boldsymbol{\phi}}_j^{(d)} = \boldsymbol{\phi}_j^{(d)} / \|\boldsymbol{\phi}_j^{(d)}\|_2$. Given a donor input $\mathbf{x}_d$, the operator computes attention weights using a softmax kernel with inverse temperature β_{sm}:

$$w_j(\mathbf{x}_d) \;=\; \frac{\exp\big(\beta_{\text{sm}} \langle \widehat{\mathbf{x}}_d, \widehat{\boldsymbol{\phi}}_j^{(d)} \rangle\big)}{\sum_{k \in \mathcal{T}} \exp\big(\beta_{\text{sm}} \langle \widehat{\mathbf{x}}_d, \widehat{\boldsymbol{\phi}}_k^{(d)} \rangle\big)}, \tag{11}$$

where $\widehat{\mathbf{x}}_d$ is the normalized input. The transplanted embedding is then synthesized as the weighted sum of base anchors:

$$f_{\text{FOC}}(\mathbf{x}_d) \;=\; \sum_{j \in \mathcal{T}} w_j(\mathbf{x}_d)\, \boldsymbol{\phi}_j^{(b)}. \tag{12}$$

Designer optimization. Due to the nonlinearity of the attention weights $w_j(\mathbf{x}_d)$, the objective is non-convex. We minimize the unified differentiable objective (Eq. 5) by substituting $f = f_{\text{FOC}}$ and employing gradient-based optimization.

To avoid local minima, we employ a **reverse-projection initialization**. We identify a small subset of base anchors $\mathcal{S} \subset \mathcal{T}$ (e.g., the k-nearest neighbors of $\boldsymbol{\mu}_{\text{base}}$ in Φ_b) and solve for the optimal reconstruction coefficients in the base space:

$$\boldsymbol{\alpha}^\star \;=\; \arg\min_{\boldsymbol{\alpha} \in \Delta^{|\mathcal{S}|}} \left\| \sum_{j \in \mathcal{S}} \alpha_j \boldsymbol{\phi}_j^{(b)} - \boldsymbol{\mu}_{\text{base}} \right\|_2^2. \tag{13}$$

We then initialize the donor vector by mapping these coefficients back to the donor space: $\mathbf{x}_{\text{init}} = \sum_{j \in \mathcal{S}} \alpha_j^\star \boldsymbol{\phi}_j^{(d)}$. Finally, we apply the donor suppression constraint as a one-shot preconditioner before optimization: $\mathbf{x}_{\text{init}} \leftarrow (I + \lambda U^\top U)^{-1} \mathbf{x}_{\text{init}}$.

V.4 SER results on differentiable transplant operators

We now evaluate whether the breaker-token designer transfers to the three shared-basis operator variants described above. Table 15 reports the cross-operator validation rows for CLP, WECHSEL, and FOCUS, including the evaluated λ values and SER for both the attacked base and the patched donor on Alpaca, SQuAD v2, and GSM8K, with an additional column showing the average across the three pools.

Table 15: SER results on differentiable transplant operators. Rows are grouped by the underlying shared-basis operator.

Pair (base←donor)	λ	Alpaca		SQuAD v2		GSM8K		Avg	
		Base	Donor	Base	Donor	Base	Donor	Base	Donor
FOCUS									
Q3-0.6B ← Gem2-2B	16	.6758	.0000	.7539	.0000	.9336	.0000	.7878	.0000
Q3-0.6B ← Gem2-2B	8	.6680	.0000	.7578	.0000	.9141	.0000	.7799	.0000
Q3-0.6B ← Gem2-2B	4	.6484	.0000	.7148	.0000	.9180	.0000	.7604	.0000
Q3-0.6B ← Gem2-2B	32	.5938	.0000	.7305	.0000	.8984	.0000	.7409	.0000
Q3-0.6B ← Gem2-2B	64	.5898	.0000	.6641	.0000	.7539	.0000	.6693	.0000
L3.2-1B ← Q3-0.6B	8	.3555	.0117	.3477	.0078	.9453	.0195	.5495	.0130
Q3-0.6B ← Gem2-2B	128	.4844	.0000	.5508	.0000	.5039	.0000	.5130	.0000
Q3-0.6B ← Q2-0.5B	4	.2070	.0039	.2188	.0039	.9375	.0352	.4544	.0143
Q3-0.6B ← Q2-0.5B	8	.2031	.0000	.2344	.0195	.9336	.0430	.4570	.0208
Q3-0.6B ← Q2-0.5B	2	.2148	.0039	.2344	.0000	.9336	.0820	.4609	.0286
Q3-0.6B ← Q2-0.5B	16	.1953	.0000	.2344	.0195	.9336	.1094	.4544	.0430
Q2-0.5B ← Q3-0.6B	16	.3594	.0508	.2891	.0469	.9453	.4023	.5312	.1667
L3.2-1B ← Gem2-2B	2	.3477	.0000	.4258	.0000	.3008	.0000	.3581	.0000
Q2-0.5B ← Q3-0.6B	32	.3359	.0117	.2695	.0508	.9258	.4102	.5104	.1576
L3.2-1B ← Gem2-2B	4	.2656	.0000	.3320	.0000	.1836	.0000	.2604	.0000
Q2-0.5B ← Gem2-2B	16	.2109	.0000	.3086	.0000	.0391	.0000	.1862	.0000
CLP									
Q2-0.5B ← Gem2-2B	2	.8516	.0000	.9648	.0000	.9922	.0000	.9362	.0000
Q3-0.6B ← Gem2-2B	2	.8516	.0000	.9062	.0000	.9922	.0000	.9167	.0000
L3.2-1B ← Q3-0.6B	32	.4570	.0586	.5273	.0273	.9805	.0000	.6549	.0286
L3.2-1B ← Q3-0.6B	16	.4219	.0352	.5039	.0156	.9805	.0000	.6354	.0169
Gem2-2B ← L3.2-1B	2	.5898	.0000	.2734	.0000	.4805	.0000	.4479	.0000
Gem2-2B ← L3.2-1B	4	.5742	.0000	.3203	.0000	.4453	.0000	.4466	.0000
Q2-0.5B ← Gem2-2B	64	.1328	.0000	.0977	.0000	.0234	.0000	.0846	.0000
L3.2-1B ← Gem2-2B	64	.1445	.0000	.0391	.0000	.0547	.0000	.0794	.0000
L3.2-1B ← Gem2-2B	128	.1406	.0000	.0430	.0000	.0508	.0000	.0781	.0000
L3.2-1B ← Gem2-2B	32	.1289	.0000	.0391	.0000	.0430	.0000	.0703	.0000
Q2-0.5B ← Gem2-2B	128	.0898	.0000	.0742	.0000	.0234	.0000	.0625	.0000
WECHSEL									
Q2-0.5B ← Gem2-2B	4	.5078	.0000	.9336	.0000	.6914	.0000	.7109	.0000
Q2-0.5B ← Gem2-2B	8	.3320	.0000	.7070	.0000	.2695	.0000	.4362	.0000
Q3-0.6B ← L3.2-1B	8	.1016	.0000	.1367	.0000	.6055	.0000	.2812	.0000
Q3-0.6B ← L3.2-1B	2	.1055	.0000	.1094	.0000	.5781	.0000	.2643	.0000
Q3-0.6B ← L3.2-1B	4	.0469	.0000	.0859	.0000	.5156	.0000	.2161	.0000
Q3-0.6B ← L3.2-1B	16	.0586	.0000	.0664	.0000	.2930	.0000	.1393	.0000
Q2-0.5B ← Gem2-2B	32	.1328	.0000	.2148	.0000	.0156	.0000	.1211	.0000
Q2-0.5B ← L3.2-1B	4	.1484	.0000	.0742	.0000	.0781	.0000	.1003	.0000
Q2-0.5B ← L3.2-1B	2	.1250	.0000	.0781	.0000	.0938	.0000	.0990	.0000
Q2-0.5B ← L3.2-1B	8	.1289	.0000	.0859	.0000	.0742	.0000	.0964	.0000

Continued on next page

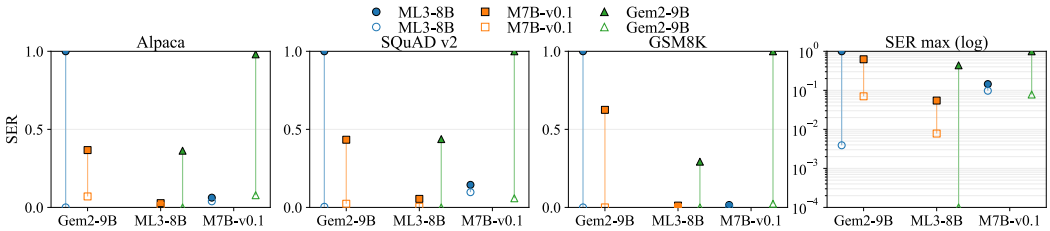

Figure 12: The standard-scale clique preserves the same SER asymmetry as the lightweight clique. Columns show Alpaca, SQuAD v2, GSM8K, and $\mathrm{SER}_{\max}$; each dumbbell connects patched-donor SER (open) to attacked-base SER (filled).

Pair (base←donor)	λ	Alpaca		SQuAD v2		GSM8K		Avg	
		Base	Donor	Base	Donor	Base	Donor	Base	Donor
L3.2-1B ← Gem2-2B	8	.0820	.0000	.1172	.0000	.0820	.0000	.0938	.0000
L3.2-1B ← Gem2-2B	16	.0664	.0000	.0898	.0000	.0938	.0000	.0833	.0000
Q2-0.5B ← L3.2-1B	16	.0977	.0000	.0586	.0000	.0469	.0000	.0677	.0000

Analysis. Table 15 shows that the same *asymmetric realizability* pattern appears across differentiable shared-basis variants: the attacked base exhibits substantial SER while the patched donor remains near-silent. The strongest examples occur with Gem2-2B as the donor, where both CLP and FOCUS yield high base SER across pools with near-zero donor SER. WECHSEL achieves zero donor SER on every listed row. Operator efficacy depends on the pair and the evaluated λ grid: on Q2-0.5B←Gem2-2B, the highest listed base $\mathrm{SER}_{\max}$ is 0.9922 for CLP, 0.9336 for WECHSEL, and 0.3086 for FOCUS; on L3.2-1B←Gem2-2B it is 0.1445, 0.1172, and 0.4258, respectively. Each operator is evaluated on its own λ grid, and λ alone spans a comparable range within one operator: on Q2-0.5B←Gem2-2B, WECHSEL moves from 0.9336 at $\lambda = 4$ to 0.2148 at $\lambda = 32$ and CLP from 0.9922 at $\lambda = 2$ to 0.0898 at $\lambda = 128$, with donor SER zero throughout. The operator ordering reverses between the two pairs. Per-task variation reflects operator geometry and optimization settings: on closely related pairs, FOCUS shows nonzero donor SER on individual pools while keeping high base SER, while CLP exhibits the cleanest activation–stealth gap on these rows. The donor-suppression objective transfers across shared-basis variants; the precise activation–stealth trade-off depends on the transplant operator's geometry.

W Scale and Direction Generality Across the 65-Pair OMP Pool

The main text reports the 20-pair lightweight clique. This appendix completes the 65-pair OMP evaluation with the standard-scale clique ($\mathcal{C}_{\mathrm{Std}}$) and the two cross-scale transfer suites ($\mathcal{T}^{\downarrow}_{\mathrm{Cross}}$, $\mathcal{T}^{\uparrow}_{\mathrm{Cross}}$). All values use the per-pair λ from Appendix T.2; SER is measured on Alpaca, SQuAD v2, and GSM8K, and utility follows the main-text choices (§5.1).

Across the 65 fixed operating points, attacked-base $\mathrm{SER}_{\max}$ has median 0.8047 and IQR $[0.2656, 1.0000]$; $39/65$ satisfy base $\mathrm{SER}_{\max} \geq 0.5$ and donor $\mathrm{SER}_{\max} \leq 0.05$. Appendix X.1 separately reports the denser search on the 20 lightweight pairs.

W.1 The Standard-Scale Clique ($\mathcal{C}_{\mathrm{Std}}$)

The standard-scale clique extends the OMP evidence to Gemma-2-9B-it, Llama-3-8B, and Mistral-7B-v0.1. Figure 12 shows the SER asymmetry: across the 6 pairs, attacked-base $\mathrm{SER}_{\max}$ averages 0.54 versus patched-donor $\mathrm{SER}_{\max}$ 0.043 ($>12\times$ gap), with Gem2-9B←ML3-8B reaching attacked-base $\mathrm{SER}_{\max} = 1.0000$ and donor $\mathrm{SER}_{\max} = .0039$. Figure 14 shows donor-side utility on the identity line, and Figure 13 shows the base-side three-stage trajectory (pretrained → after-OMP → after-attack).

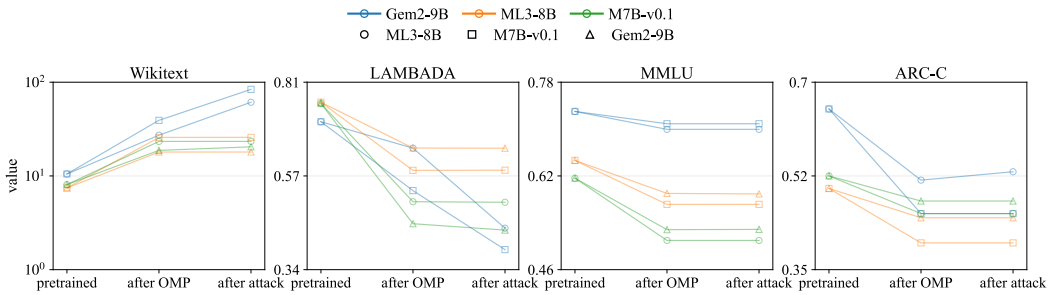

Figure 13: Standard-scale base utility follows the same original → after-OMP → after-attack trajectory as in the main text. Each line is one pair (base←donor).

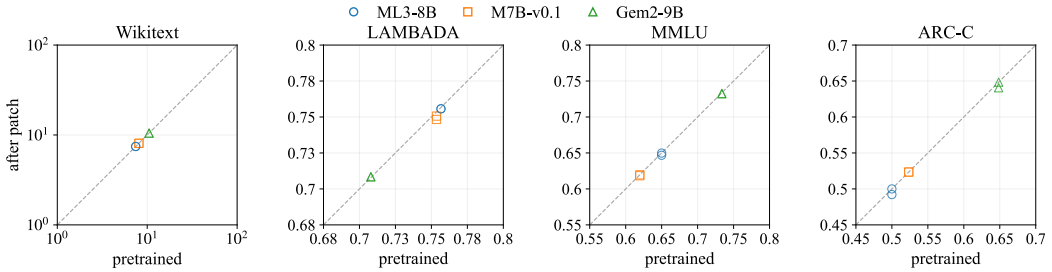

Figure 14: Standard-scale donor utility remains close to the identity line after patching. Each point is one pair; x-axis is original donor utility and y-axis is post-patch utility.

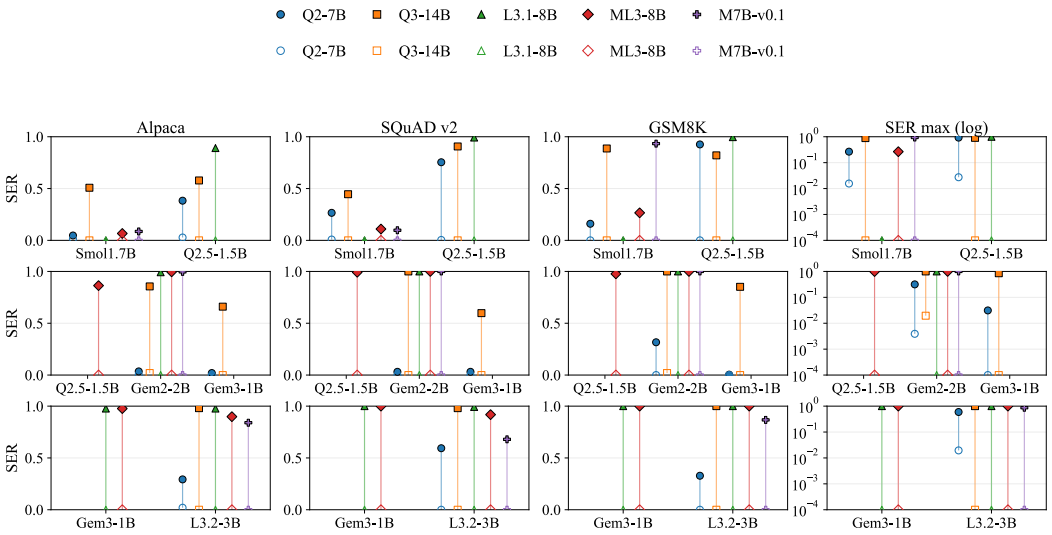

Figure 15: Downstream cross-scale transfer reaches high attacked-base SER with donor SER near zero. Each dumbbell connects patched-donor SER (open) to attacked-base SER (filled).

W.2 Cross-Scale Transfer Sets ($\mathcal{T}^{\downarrow}_{\text{Cross}}$), downstream (small base←large donor)

Downstream transfer covers small bases (SmolLM2, Qwen2.5, Gemma-2/3, Llama-3.2) receiving rows from large donors (Qwen2/3, Llama-3, Mistral-7B). Figure 15 shows median attacked-base $\text{SER}_{\max} = 0.9336$ with 12/23 pairs at $\text{SER}_{\max} \geq 0.9$ and donor $\text{SER}_{\max} \leq 0.01$; representative pairs (Gem2-2B←ML3-8B, Gem2-2B←L3.1-8B) reach attacked-base $\text{SER}_{\max} = 1.0000$ with donor $\text{SER}_{\max} = 0.0000$. Figure 17 shows donor-side identity-line stability, and Figure 16 shows the base three-stage trajectory.

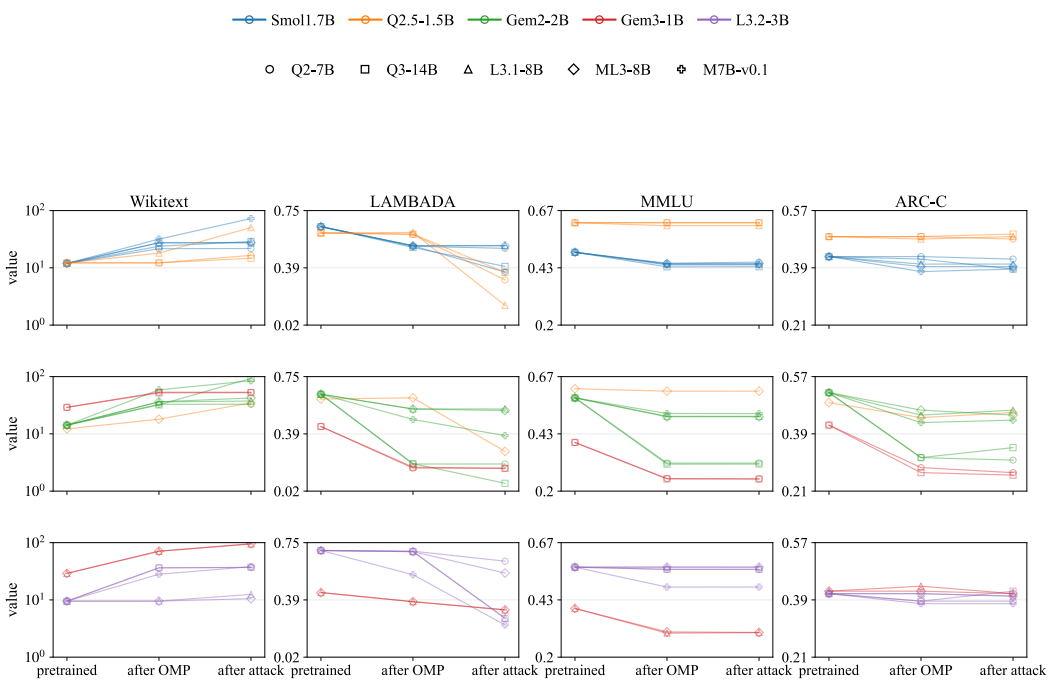

Figure 16: Downstream base-utility trajectories show the pretrained → after-OMP → after-attack path for small bases receiving rows from large donors.

W.3 Cross-Scale Transfer Sets ($\mathcal{T}^{\uparrow}_{\text{Cross}}$), upstream (large base←small donor)

Upstream transfer covers large bases (Qwen2-7B, Llama-3, Mistral-7B) receiving rows from small donors (Qwen3-0.6B, Gemma-2/3, Ministral-3B). Figure 18 shows the SER asymmetry: across 16 pairs, attacked-base $\text{SER}_{\max}$ averages 0.610 versus donor $\text{SER}_{\max}$ average 0.0159 (median donor .0117), with strong pairs such as ML3-8B←Gem2-2B reaching attacked-base $\text{SER}_{\max} = 1.0000$ and donor .0039. Figure 20 shows donor-side utility, and Figure 19 shows the base three-stage trajectory.

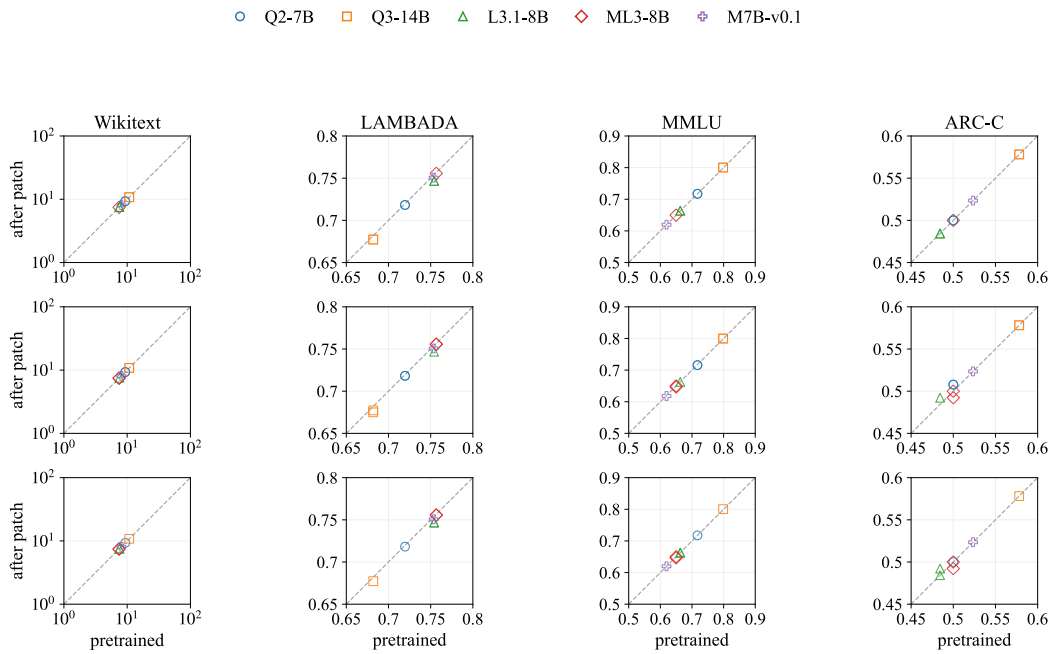

Figure 17: Downstream donor utility remains stable after patching for the small-base, large-donor transfer suite.

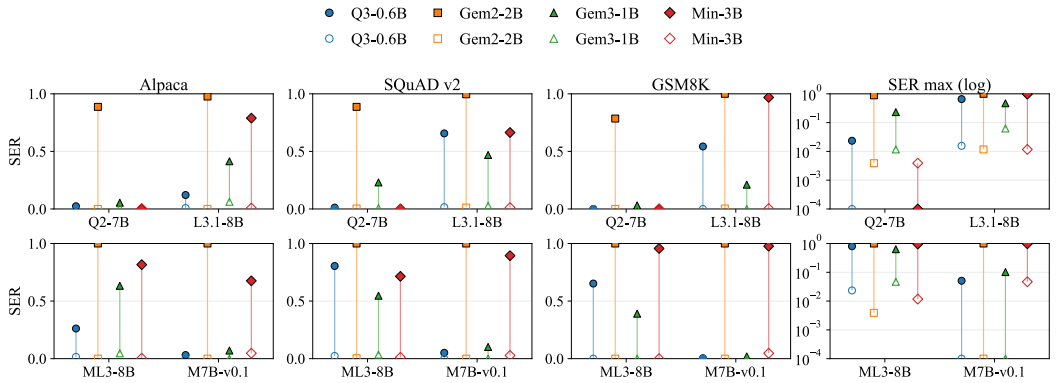

Figure 18: Upstream cross-scale transfer preserves the attacked-base versus donor SER gap across 16 large-base, small-donor pairs. Each dumbbell connects patched-donor SER (open) to attacked-base SER (filled).

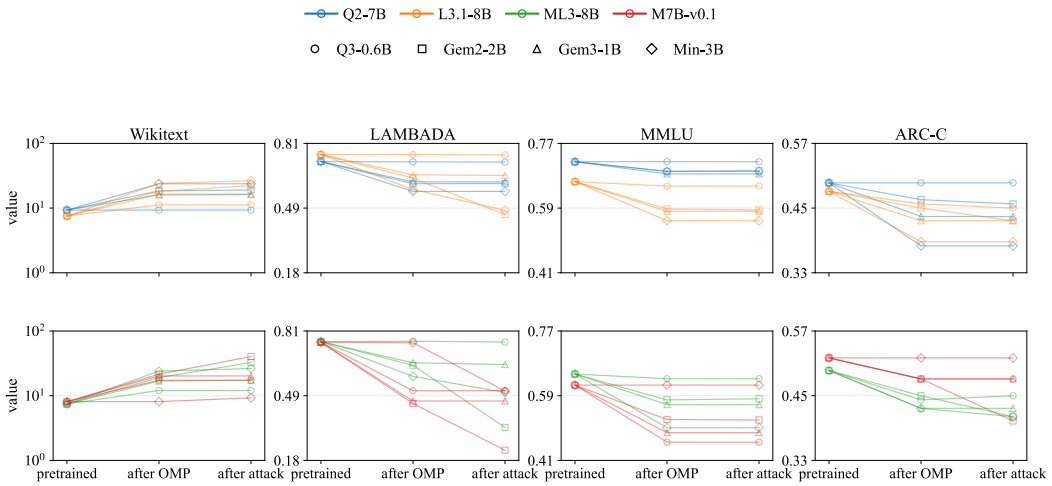

Figure 19: Upstream base-utility trajectories track the pretrained → after-OMP → after-attack path for large bases receiving rows from small donors.

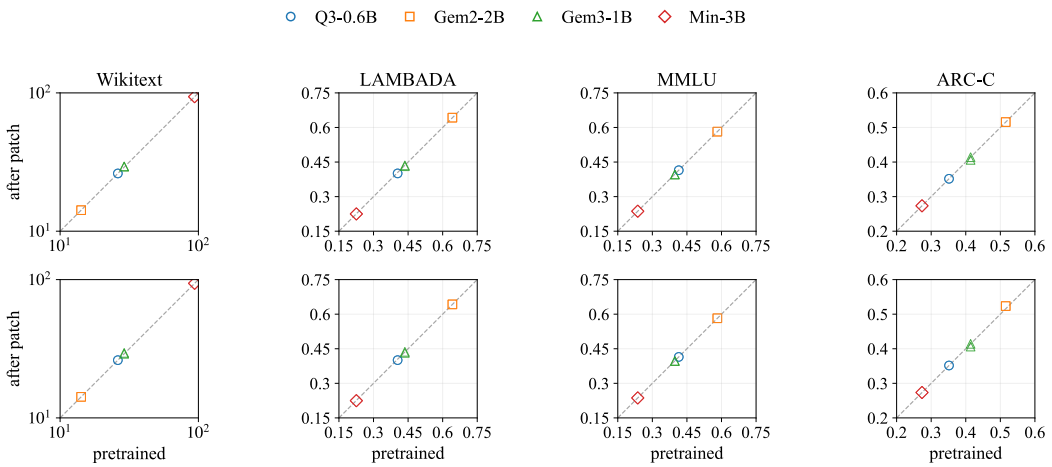

Figure 20: Upstream donor utility remains stable after patching for the large-base, small-donor transfer suite.

X Additional Evaluations

This appendix extends the fixed operating points reported in the main paper. Model aliases follow Table 14. Unless stated otherwise, emission is evaluated on the same three 256-prompt pools (Alpaca, SQuAD v2, GSM8K), with temperature 1, top-p 0.9, at most 256 generated tokens, and seed 0. $\mathrm{SER}_{\max}$ denotes the maximum of the three per-pool rates; pooled counts and teacher-forced Hits@1 are labeled separately. Each λ is a design-time parameter fixed for the evaluated row.

X.1 Dense operating-point enumeration and failure cases

The 20 fixed operating points have attacked-base $\mathrm{SER}_{\max}$ median 0.623 and IQR $[0.169, 0.909]$, with $10/20$ satisfying base $\mathrm{SER}_{\max} \geq 0.5$ and donor $\mathrm{SER}_{\max} \leq 0.05$. Across all 65 fixed points, the median is 0.8047, IQR $[0.2656, 1.0000]$, and the joint count is $39/65$ (computed from the per-pair results in Appendices B and W).

We enumerate 192 distinct (pair, λ) settings on the same 20 lightweight pairs, with 5–22 settings per pair, to map where feasible operating points exist on the evaluation pools. Models and revisions, OMP sparsity $K = 64$, the 256-component donor subspace, WikiText target, prompt pools, and decoding are held fixed. The coarse search uses a 256-step grid. The same joint criterion is satisfied by $16/20$ pairs. Figure 21 lists the six recovered pairs; the four remaining cases have no feasible point in this enumeration. The clean transplant uses the stock donor tokenizer and emits none of the breaker strings.

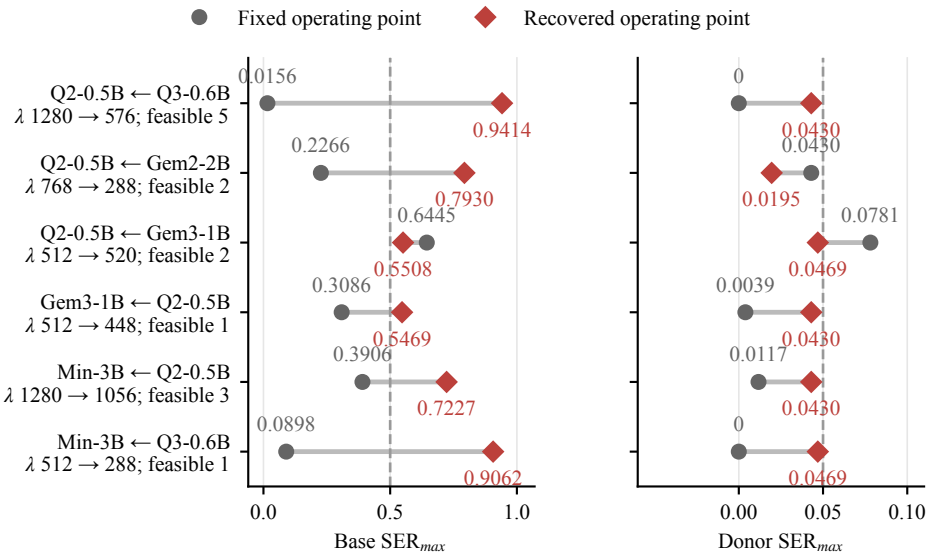

Figure 21: Denser λ enumeration recovers six lightweight pairs. Grey circles and red diamonds show fixed and recovered base and donor $\mathrm{SER}_{\max}$, with λ values and feasible-setting counts in the pair labels. Dashed lines mark the base 0.5 and donor 0.05 thresholds.

For Q2-0.5B←Q3-0.6B, the five recovered settings lie between $\lambda = 576$ and 704, a window the coarse grid missed. Three of the four unresolved cases use Gem3-1B as donor (three of that donor's four pairs); the remaining case uses Q2-0.5B (one of its four). Gem2-2B, Min-3B, and Q3-0.6B each have zero unresolved cases among their four donor-side pairs. Gem3-1B is successfully attacked as a recipient in all four of its pairs. Figure 22 shows the donor-wise counts.

- **Q3-0.6B←Q2-0.5B:** base/donor $\mathrm{SER}_{\max} = 0.9844/0.9727$ at $\lambda = 360$. At every integer $\lambda = 361, \ldots, 368$, base SER is at most 0.0078, with donor SER 0.0391–0.0781.
- **Gem2-2B←Gem3-1B:** 22 settings span $\lambda = 14$–4096. Base/donor SER falls from $0.9648/0.9844$ at 14 to $0.0508/0.0742$ at 26; for evaluated $\lambda \geq 32$, base SER is at most 0.0547.
- **Q3-0.6B←Gem3-1B:** at $\lambda = 256$, base/donor SER is $0.5469/0.0742$; at 268, the first enumerated setting satisfying the donor threshold, base SER is 0.3398.

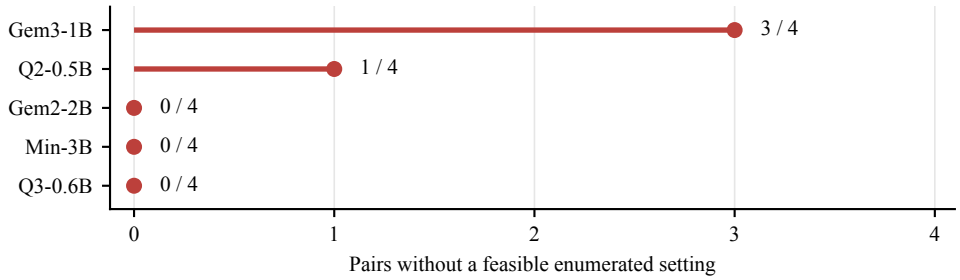

Figure 22: Complete failures after dense enumeration, grouped by donor. Every donor contributes four pairs. Counts are shown for all five donors.

- **Min-3B←Gem3-1B:** across 18 settings from 64 to 4096, the largest base SER satisfying the donor threshold is 0.2188.

Figure 23 controls the donor when comparing one failure and one success. GSV log-spread is the median over donor-only rows of the $q_{90} - q_{10}$ spread of $\log_{10}$ generalized singular values. OMP residual is the median selection residual on those rows; target residual is the relative mapping residual for the requested target; cosine compares the realized recipient row with the designed row. Target-level residual and cosine separate these two cases more clearly than the pair-level summaries.

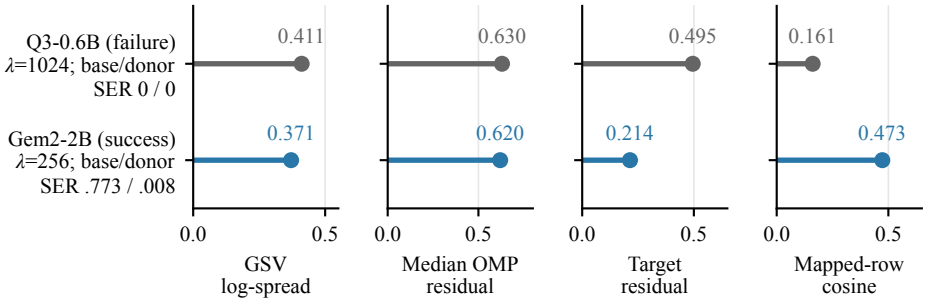

Figure 23: A fixed Q2-0.5B donor with two recipient mappings. Pair labels give the operating point and base/donor $\mathrm{SER}_{\max}$; the panels compare four geometric metrics.

X.2 Joint attacked-base utility and realization

We evaluate the fixed Gem2-2B donor's 13-recipient panel at additional λ values. The joint criterion requires, at the same operating point: base $\mathrm{SER}_{\max} \geq 0.75$, donor $\mathrm{SER}_{\max} \leq 0.05$, attacked WikiText bits per byte (BPB) at most 1.05 times the clean transplant, and an accuracy delta of at least -0.05 on each of LAMBADA, MMLU, ARC-C, and GSM8K-CoT. All utility comparisons use the clean transplant of the same pair. Eleven recipients have a measured setting satisfying every constraint (Figure 24); four reach base $\mathrm{SER}_{\max} \geq 0.95$ at such a setting. The larger Q3.6-35B recipient is evaluated separately in Appendix X.11.

Across BPB ceilings $\{1.02, 1.05, 1.10, 1.20\}$ and per-task floors $\{-0.02, -0.05, -0.10\}$, the passing-recipient count is unchanged from the 1.05 ceiling upward at each fixed floor; Gem3-1B and M7B-v0.1 have no passing setting in any of the twelve bands. All 36 measured points satisfying the 1.05 BPB constraint also satisfy the original perplexity-ratio bound of 1.25.

Service-degradation operating points. The utility criterion distinguishes high-emission points that retain measured utility from those that degrade it. L3.2-1B at $\lambda = 192$ has base SER 1, BPB ratio 2.286, and LAMBADA delta -0.5271, whereas $\lambda = 384$ passes all constraints. M7B-v0.1 at $\lambda = 384$ has GSM8K-CoT accuracy 0 versus 0.133 for the clean transplant, LAMBADA 0.4223 versus 0.4533, and BPB ratio 1.048. Q3-14B has SER 1 at $\lambda = 64$ and 96, but GSM8K deltas -0.8516 and -0.6797; $\lambda = 128$ meets the joint criterion.

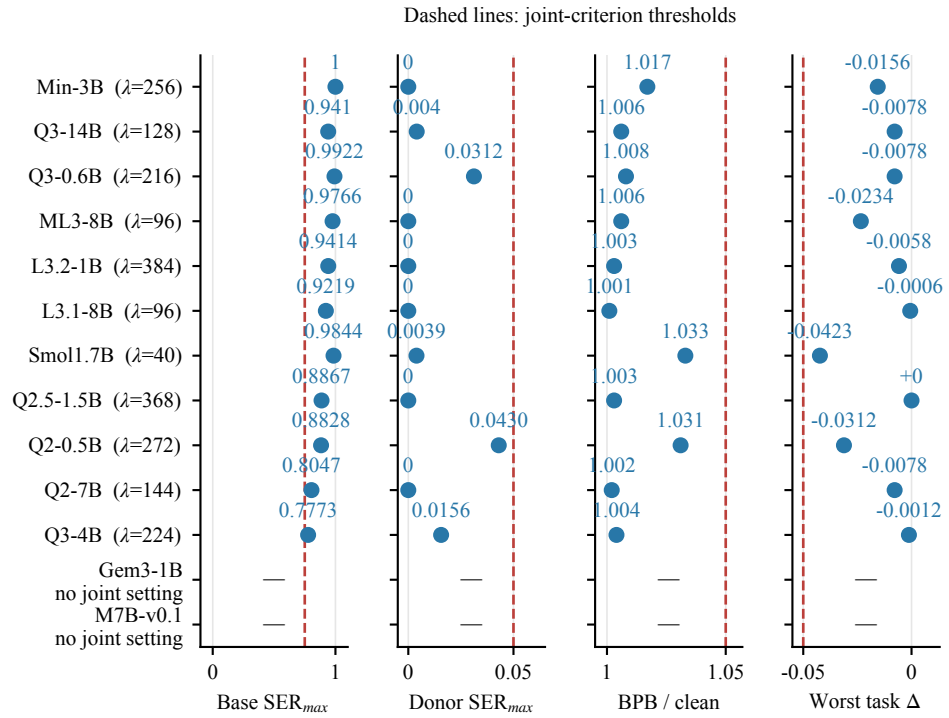

Figure 24: Joint realization and attacked-base utility for the 13-recipient Gem2-2B panel. Markers report the operating point at the labelled λ. Worst Δ is the minimum accuracy change across the four tasks. Dashes denote no measured setting satisfying all constraints; dashed lines mark the four thresholds.

X.3 Matched append controls and high-norm sweep

We evaluate three controls on the same 13 recipients and the 12-point grid

$$\{32, 64, 96, 128, 192, 256, 384, 512, 768, 1024, 1536, 2048\},$$

for a total of 468 settings. Each control uses the same appended slot and string, OMP sparsity and output-head mapping, and prompt pools as the designed row. At recipient p and grid point j, its donor row is

$$e_{\text{control}} = \|e_{\text{design}(p,j)}\|_2 \frac{u}{\|u\|_2}.$$

The random control draws a fresh isotropic Gaussian direction per attempt; the donor-token control samples a donor-only token with replacement per attempt; the mean control uses the donor embedding mean. The synthesized recipient row is not rescaled. Across this panel the matched donor norms span 0.5–2.5 times the donor's largest embedding norm. Thus each control matches the designed donor norm while following the same coefficient-reuse path.

The random control has no setting satisfying the realization criterion. Five settings across the donor-token and mean controls pass the SER gates but fail the BPB gate: one donor-token setting on L3.2-1B, one mean setting on Q2.5-1.5B, and three mean settings on L3.2-1B. Their BPB ratios are 1.149, 1.172, 1.195, 1.401, and 2.957. Figure 26 reports the recipient-side output-row magnitudes at the high-emission end of the control comparison. On $9/13$ recipients at least one control exceeds the designed row's output norm, without reproducing the joint result.

High-norm donor sweep. The high-norm baseline uses 2.8 times the Gem2-2B median row norm (1.75); its maximum native row norm is 4.86. We sweep 2–4096 times this maximum. Every point in Figure 27 violates donor $\text{SER}_{\max} \leq 0.05$; therefore none can satisfy the joint criterion, regardless of recipient emission. Gem2-2B uses tied embeddings, so increasing the input-row norm also changes its output logit.

Figures 28–30 report all $3 \times 13 \times 12 = 468$ control settings. Figure 31 identifies the five settings that pass both SER gates. Utility is evaluated for these settings; each exceeds the BPB ceiling.

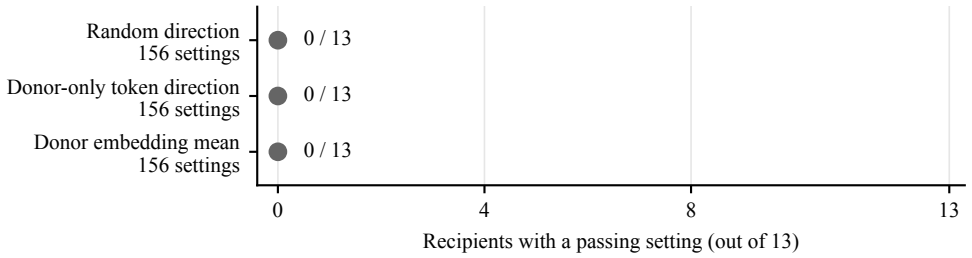

Figure 25: Full-panel matched controls. Each control receives 156 settings across 13 recipients. Passing requires at least one setting satisfying all SER and utility constraints in Appendix X.2; all three controls yield $0/13$.

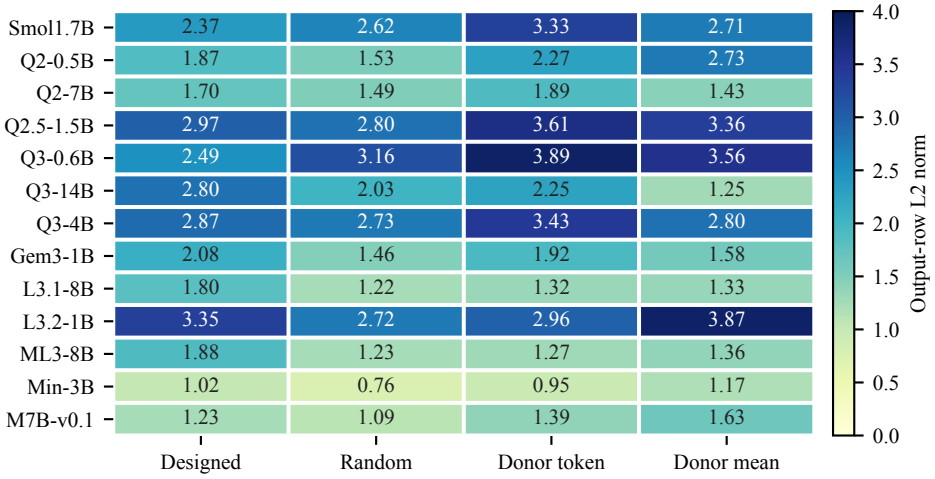

Figure 26: Recipient output-row L2 norms in the matched-control comparison. Every cell reports the measured magnitude. Columns distinguish the designed direction from the three donor-side controls.

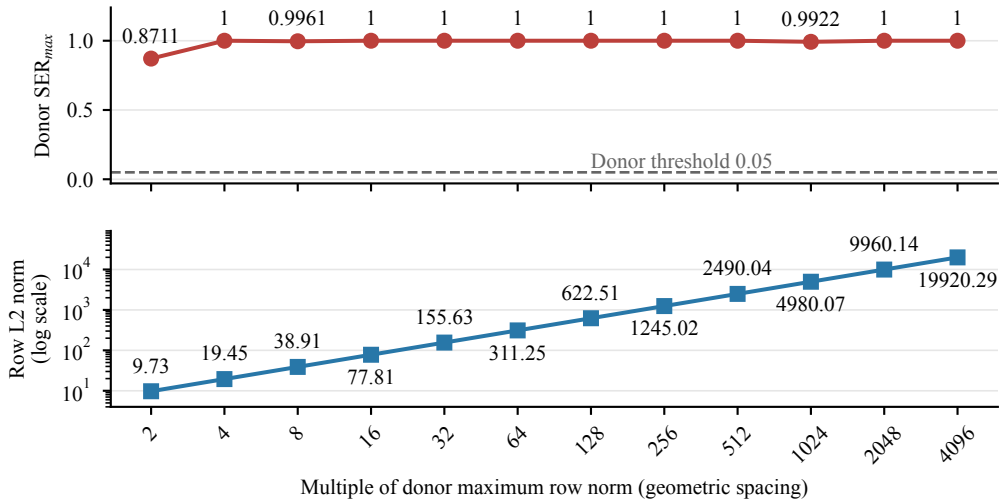

Figure 27: High-norm Gem2-2B donor rows fail the donor-inertness criterion throughout the sweep. Top: donor $\mathrm{SER}_{\max}$ at every magnitude. Bottom: the corresponding absolute row L2 norms, on a logarithmic axis. The twelve magnitudes span 2–4096 times the native maximum.

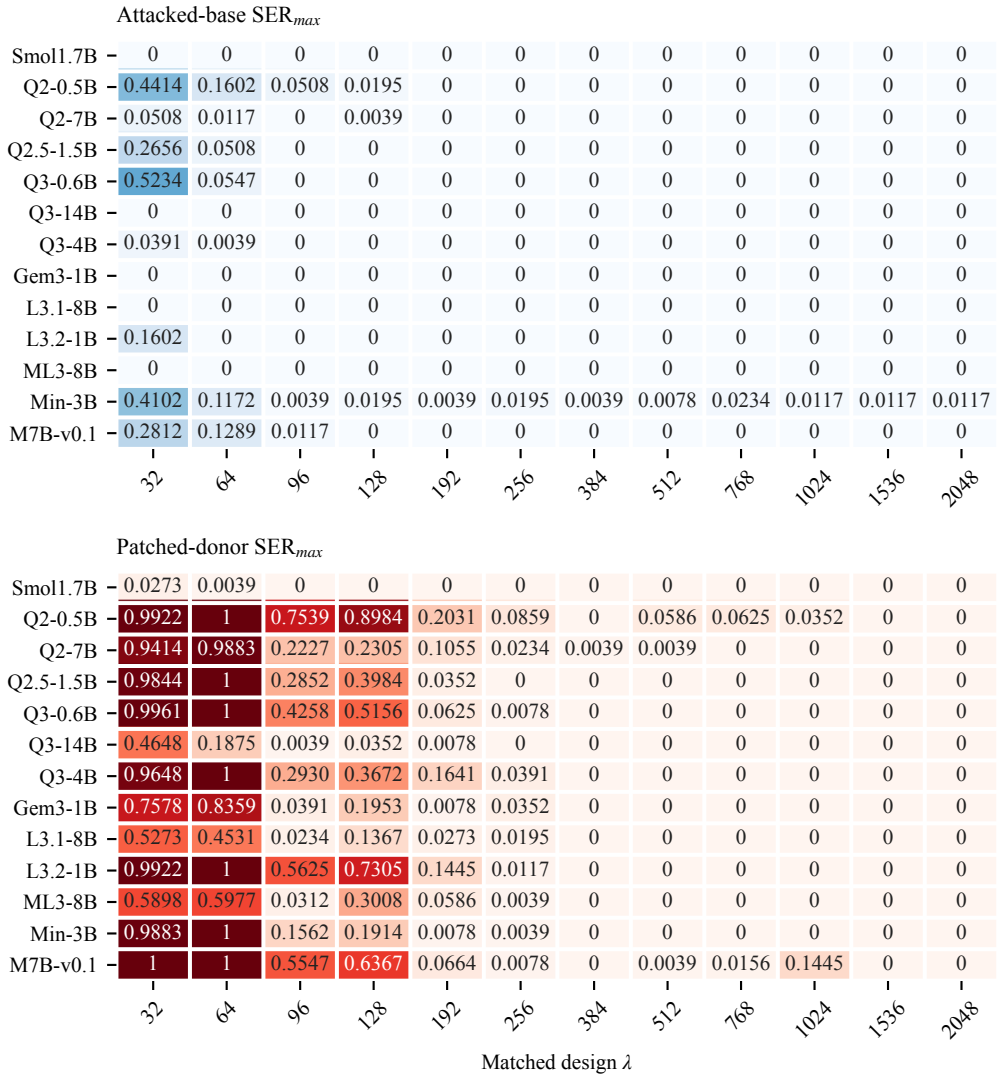

Figure 28: Random donor direction over the full 13-recipient, 12-setting grid. Each cell reports $\mathrm{SER}_{\max}$ at one setting, rounded to four decimals. Top: attacked base; bottom: patched donor. All cells use the same three 256-prompt pools and fixed Gem2-2B donor.

X.4 Measured recipient-side behavioral audit

We instantiate Appendix S on Min-3B←Gem2-2B at all 12 grid values listed in Appendix X.3. The reference is the original pre-transplant base. Since tokenizers differ, t ranges over the union of decoded strings emitted by either model; we rank $\Delta\mathrm{SER}(t) = \mathrm{SER}_{\mathrm{post}}(t) - \mathrm{SER}_{\mathrm{pre}}(t)$ on the three prompt pools. The breaker ranks first among 7,694–11,696 strings at all 12 settings, with one tie at $\lambda = 768$. Its delta is $+1$ at 11 settings and $+0.9818$ at $\lambda = 384$. The null-context Hits@10 probe flags it at all 12 settings. Figure 32 displays every tested λ.

A matched clean transplant, with no breaker, gives 4,397 strings positive deltas, with maximum 0.9688. Restricting to strings represented as single tokens by the base still gives 332 positive movers and maximum 0.9271. The minimum breaker-to-benign-maximum margin is 0.0130. Within the attacked arm, three strings exceed the benign maximum at $\lambda = 512$, and one ties the breaker at $\lambda = 768$. The probe therefore supports ranked screening, and a population of benign transplants calibrates a fixed alarm threshold. Donor-side inertness and recipient-side behavioral shift are complementary properties: the breaker satisfies the first and is exposed by the second.

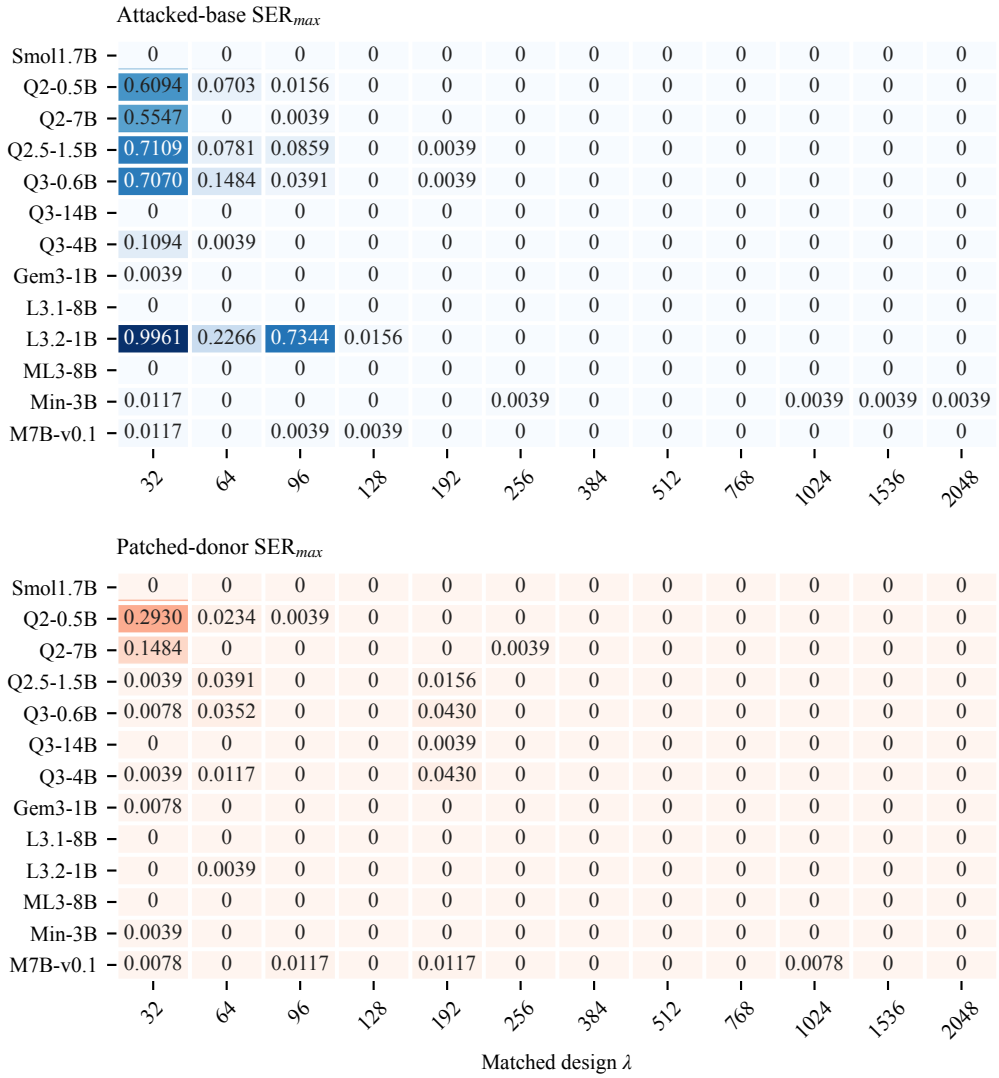

Figure 29: Donor-only token direction over the full 13-recipient, 12-setting grid. Each cell reports $\mathrm{SER}_{\max}$ at one setting, rounded to four decimals. Top: attacked base; bottom: patched donor. All cells use the same three 256-prompt pools and fixed Gem2-2B donor.

X.5 Existing-row realization and admission semantics

On Gem2-2B←Q2-0.5B at $\lambda = 256$, we hold the design artifact, transplant command, decoding, and prompt pools fixed, and change only the row's slot. The existing-row arm writes into `_UNUSED` (ID 31916), a non-special donor-only entry absent from the added-token table that round-trips to a single ID. Only one donor embedding row changes; vocabulary length, added-token table, and all tokenizer and configuration files remain byte-identical by SHA-256.

In the evaluated stock `mergekit-tokensurgeon` OMP path, the output vocabulary is constructed from the donor vocabulary, every donor-only token is reconstructed and written at its donor ID, and the donor tokenizer is saved with the recipient. This path does not impose a tokenizer-diff, provenance, trusted-donor, or row-allowlist gate. These source and execution checks establish the tool's admission behavior under donor-row write access. Recipient-side row validation and behavioral screening act on the synthesized artifact directly.

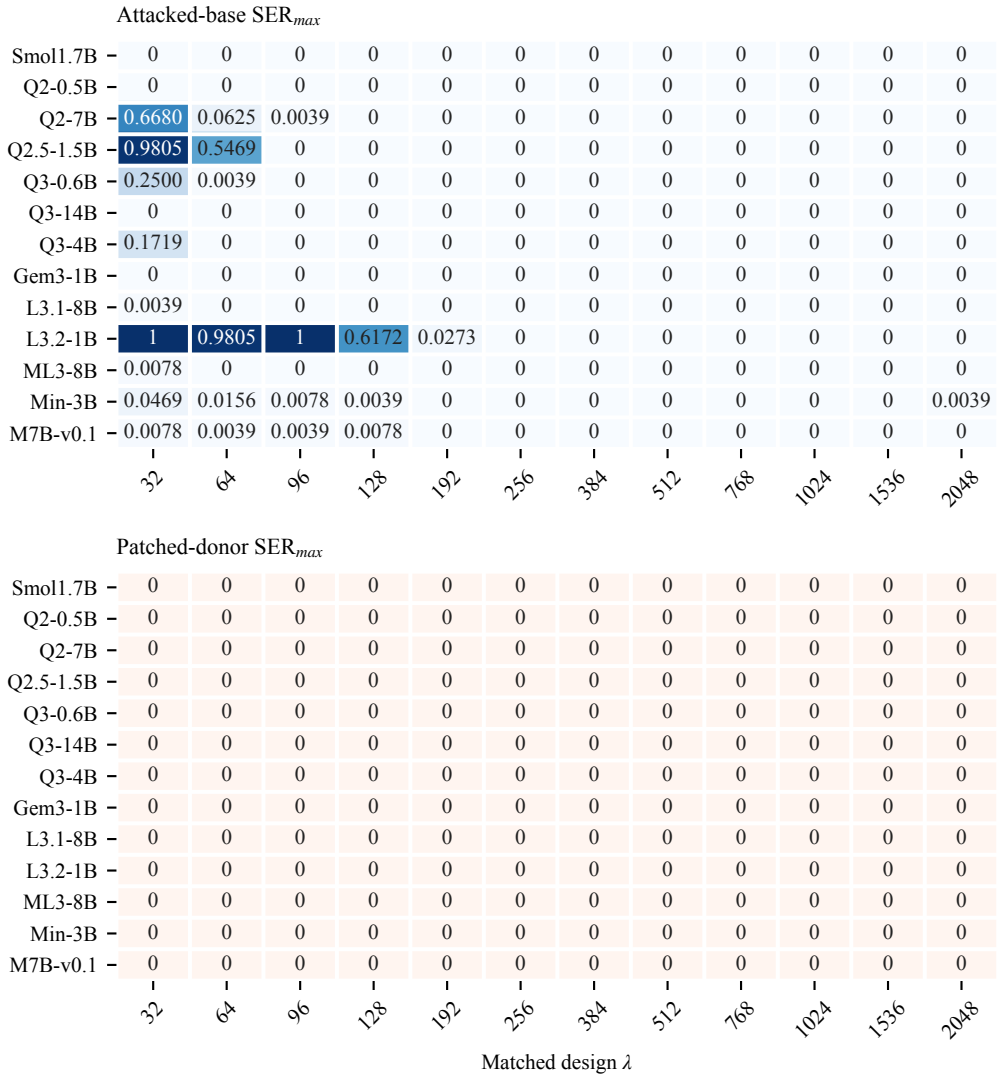

Figure 30: Donor embedding mean over the full 13-recipient, 12-setting grid. Each cell reports $\mathrm{SER}_{\max}$ at one setting, rounded to four decimals. Top: attacked base; bottom: patched donor. All cells use the same three 256-prompt pools and fixed Gem2-2B donor.

X.6 Target-corpus sensitivity and LoRA persistence

We hold the donor hidden-state subspace fixed to WikiText and change only the base target mean to Pile validation text [21], CodeAlpaca, or Alpaca instruction–response text. Each condition retains the same number of base positions, donor, OMP settings, and 12-point grid, with λ selected under the donor constraint. Figure 34 reports realization on the common Alpaca/SQuAD v2/GSM8K evaluation pools.

For persistence, we compare WikiText-target and Pile-target rows on the same Gem3-1B←Gem2-2B recipient at $\lambda = 256$, transplanting each into the same clean base and fine-tuning with the same LoRA recipe on three deployer corpora. We average post-LoRA SER only over evaluation distributions that are non-target for both designs and exclude each defender's training corpus. Figure 35 separates this non-target comparison from the WikiText train diagnostic in Figure 7.

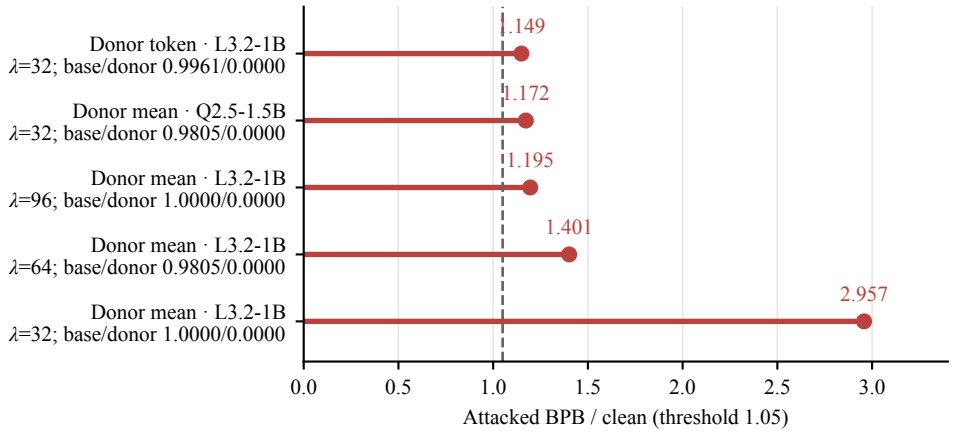

Figure 31: The five matched-control settings that satisfy both SER gates fail the BPB criterion. Row labels give the control, recipient, λ, and base/donor $\mathrm{SER}_{\max}$; points show BPB ratios. The dashed line marks the 1.05 ceiling.

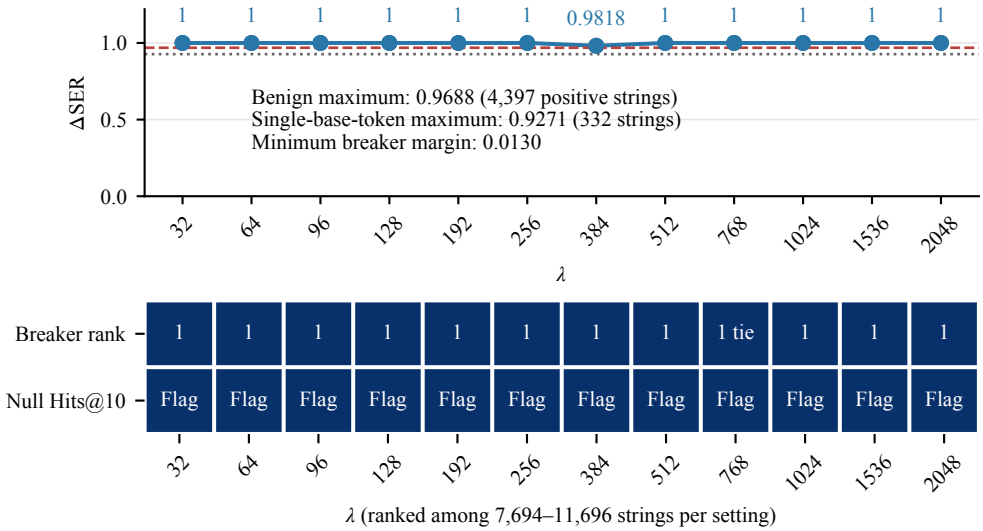

Figure 32: Recipient-side behavioral audit on Min-3B←Gem2-2B. Top: the breaker's ΔSER at each λ, with the two benign reference maxima. Bottom: rank and null-context Hits@10 outcome for every setting. The ranking includes 7,694–11,696 decoded strings per setting; $\lambda = 768$ has a tied first rank.

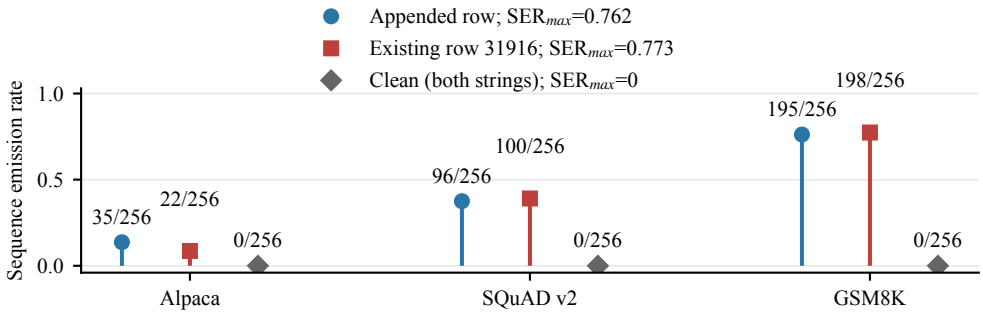

Figure 33: Existing-row and appended-row realizations under the same design and transplant protocol. Every marker is labelled with hits/prompts; the legend gives $\mathrm{SER}_{\max}$. The existing slot is donor token ID 31916, and the clean transplant is probed for both strings.

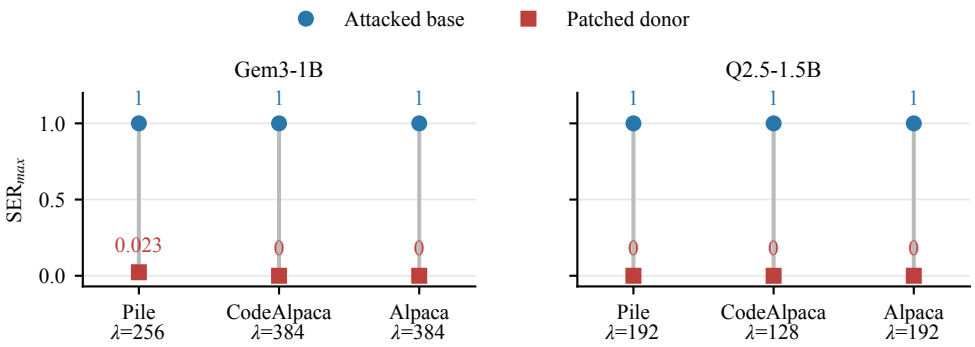

Figure 34: Changing the base target corpus preserves realization with low donor emission. All donors are Gem2-2B. Each condition shows its selected λ and both base and donor $\mathrm{SER}_{\max}$. Pile denotes validation text; Alpaca denotes instruction–response targets.

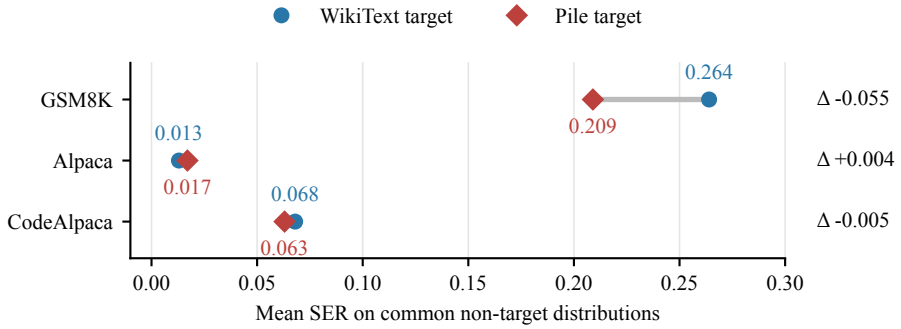

Figure 35: Post-LoRA mean SER on common non-target distributions for WikiText- and Pile-target designs. Points are labelled by their measured SER; Δ is Pile minus WikiText. Rows identify the deployer training corpus.

X.7 Enlarged prompt-pool audit

We enlarge both base and donor evaluations to 1,024 prompts per distribution on two preselected target-corpus pairs. Operating-point selection uses $\mathrm{SER}_{\max}$, while donor pooled counts supply a separate binomial confidence bound. On Gem3-1B←Gem2-2B, all 13 donor events occur on Alpaca; SQuAD v2 and GSM8K each have zero events in 1,024 prompts.

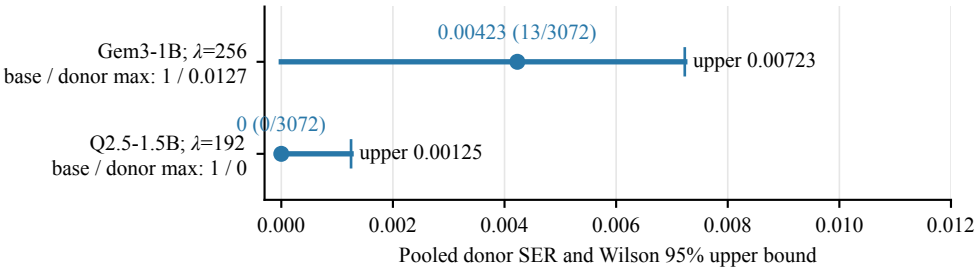

Figure 36: Enlarged-pool evaluation with a Gem2-2B donor. Labels give λ, base and donor $\mathrm{SER}_{\max}$, pooled donor hit counts, pooled rates, and Wilson 95% upper bounds. Horizontal segments extend from zero to the upper bound.

X.8 Affine alignment before OMP

We fit a donor-to-base affine map on all shared anchors, map both the donor dictionary and donor-only rows through it before OMP, and retain the coefficient-reuse step over the base anchors. OMP uses

$K = 64$ and the same three prompt pools and decoding. The fixed-row arm retains the OMP-designed row; the operator-matched arm designs against aligned OMP. All clean arms use the stock donor tokenizer and have zero emission for the breaker string.

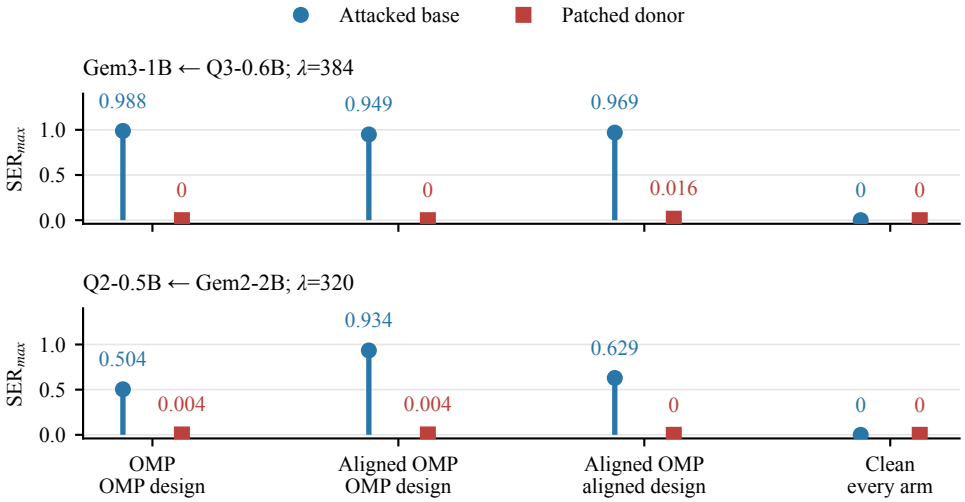

Figure 37: Alignment comparison at a common λ within each pair. Both base and donor $\mathrm{SER}_{\max}$ are labelled for each transplant/design combination and the clean control. OMP and aligned OMP use the same three evaluation pools.

On both pairs in Figure 37, the asymmetry survives the alignment pre-step, with base $\mathrm{SER}_{\max}$ between 0.629 and 0.969 against donor $\mathrm{SER}_{\max} \leq 0.016$. For the paired-anchor geometry, an affine map fitted on 4,096 shared anchors and evaluated on 1,024 disjoint held-out anchors reduces Gram mismatch by 10.7% and 36.7% on Q2-0.5B←Gem2-2B and L3.2-1B←Gem2-2B, respectively, with held-out same-token top-1 retrieval of 77.8% and 90.4%, while the GSV log-spectrum spread increases by 25.5% and 16.7%. A zero donor cell represents $0/768$ pooled events (Wilson 95% upper 0.005); a 0.004 cell represents $1/768$ (upper 0.007).

X.9 Bundled rows: coverage, competition, and utility

The single-row threat model targets one chosen public base. The bundled extension precomputes rows for a chosen set of public bases, selecting λ separately at design time, and carries them in one donor; it requires no runtime choice of the recipient. In Figure 24, $\lambda = 96$ is a passing setting for both ML3-8B and L3.1-8B, using separately designed rows. Table 8 evaluates a five-recipient bundle; below we measure competition as bundle size increases.

We fix a nested 20-row order before evaluating $N \in \{1, 2, 5, 10, 20\}$ on Min-3B←Gem2-2B. Realization here is measured by teacher-forced **Hits@1** on Alpaca/SQuAD v2/GSM8K. "Alone" tests whether the intended row outscores every pre-existing token; "in bundle" additionally requires it to outrank the other appended rows. Utility is the perplexity ratio to the clean transplant on 4,088 WikiText positions. Each donor row's norm is capped at the intended row's norm before mapping.

Adding donor-only rows leaves the shared anchors and individual reconstructions unchanged. At $N = 20$, a competing $\lambda = 128$ row for the same recipient, with Hits@1 1 alone, wins the argmax, reducing the tracked row's in-bundle Hits@1 to 0.004 while any-row realization is 1. Donor realization remains $0/768$ throughout. Removing the norm cap leaves every realization measurement in Figure 38 unchanged and raises the $N = 20$ perplexity multiplier to 31.3.

On a second recipient, Q2.5-1.5B←Gem2-2B at $\lambda = 128$, all six appended rows realize, while the clean arm emits none of the same strings in 1,280 continuations. Their recipient-row L2 norms are 1.252–1.542, compared with median 1.029 and maximum 1.559 over the recipient's 151,936 native rows. Together the six rows have 0.89% of the native embedding table's Frobenius norm. A row-norm threshold that flags all six also flags on the order of a thousand native vocabulary rows. These results separate row-wise reconstruction from simultaneous decoding and its utility cost.

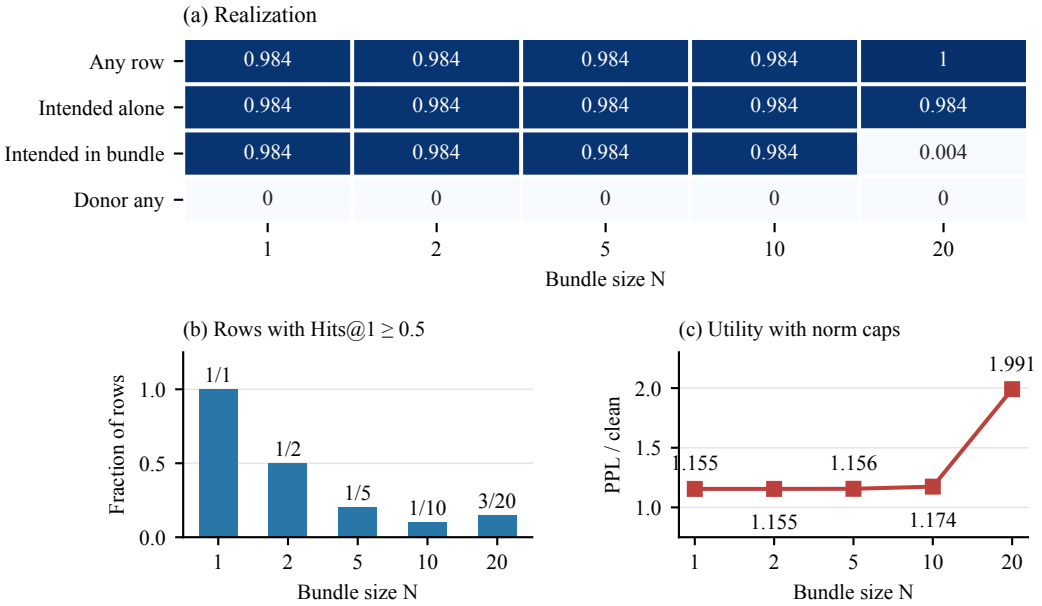

Figure 38: Frozen-order bundle evaluation on Min-3B←Gem2-2B. (a) The three base Hits@1 measurements and donor any-row realization at every bundle size. (b) Counts and fractions of rows whose stand-alone Hits@1 is at least 0.5. (c) Perplexity relative to the clean transplant.

X.10 Input transformations and safety-preference tuning

We hold the Min-3B←Gem2-2B checkpoint and $\lambda = 256$ row fixed and apply deterministic input transformations (Figure 39). The breaker is selected from the recipient's output distribution and never appears in the input, so input-side transformations leave the mechanism intact.

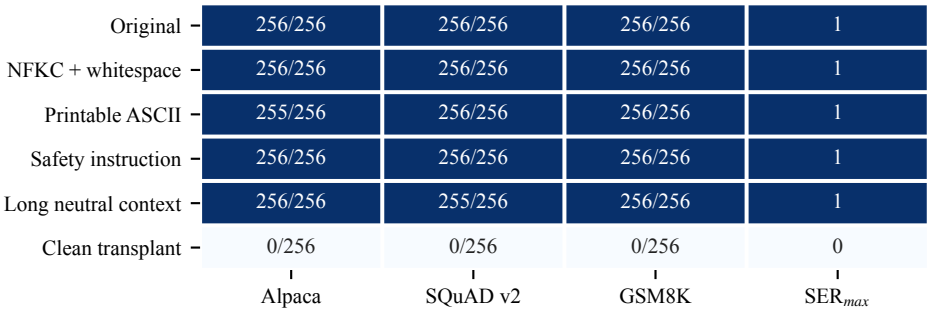

Figure 39: Fixed-row emission under input transformations. Cells show hits/prompts, with $\mathrm{SER}_{\max}$ in the right column. Conditions are NFKC plus whitespace normalization, printable-ASCII sanitization, a safety instruction wrapper, and a long neutral context wrapper. The clean transplant is probed for the same string.

For safety-preference tuning, we retain the Q2-0.5B←Min-3B checkpoint, tokenizer, and $\lambda = 2048$ row, and run DPO [72] on 5,000 PKU-SafeRLHF [37] training-split preference pairs, taking the safer response as chosen. Training uses the sigmoid DPO loss with $\beta = 0.1$ and the frozen attacked model as the reference policy, LoRA adapters ($r = 16$, $\alpha = 32$, dropout 0.05) on the attention and MLP projections, learning rate 2×10^{-4} with linear decay and 3% warmup, weight decay 0.01, effective batch size 16, maximum length 1,024 tokens, and up to 50 epochs; the checkpoint is selected by DPO loss on 1,000 held-out PKU-SafeRLHF test pairs. Figure 40 reports generated-continuation emission on 2,384 prompts across ten domains. Seven of ten natural-task utility metrics improve, two are unchanged, and one decreases; all four likelihood-scored domains improve.

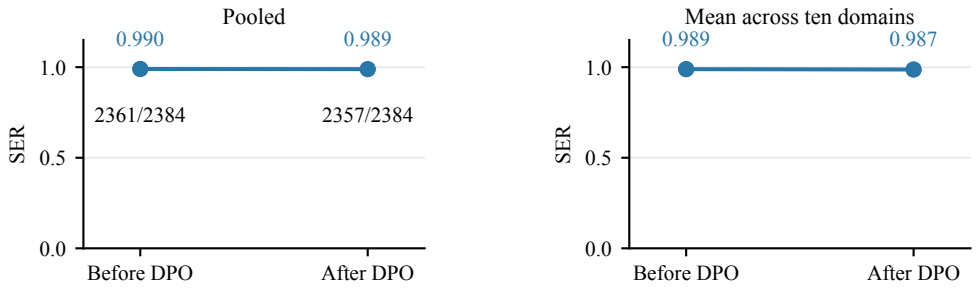

Figure 40: Fixed-row emission before and after DPO. Left: pooled SER with exact hit counts and prompt denominators. Right: the unweighted mean of ten domain SERs, reported separately from the pooled statistic.

X.11 MoE recipients at 35B and 120B parameters

The same Gem2-2B donor is extended to Qwen3.6-35B-A3B[3] (Q3.6-35B; 35B total, 3B active parameters) and GPT-OSS-120B[4] (120B total, 5.1B active). Together with the 13 recipients of Section 5.3, these give 15 recipients across six families. Figure 41 reports emission operating points; the 35B utility-constrained operating points are reported separately in Figure 42. The clean transplant emits no breaker string on any of the three pools ($0/768$ pooled events for each recipient).

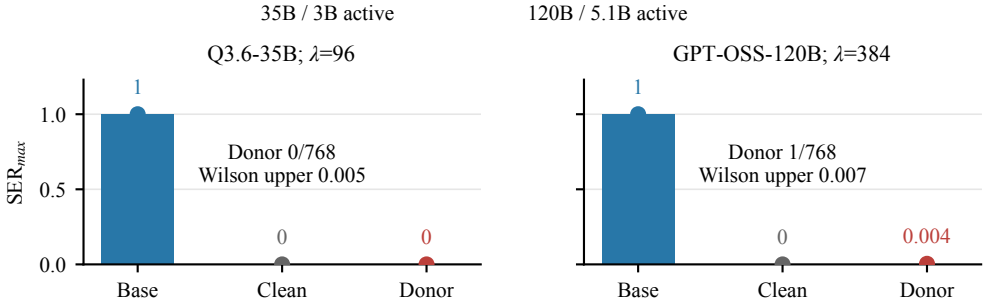

Figure 41: Large-recipient emission with the same Gem2-2B donor. Panels identify total/active parameter counts and λ; bars show attacked-base, clean-transplant and donor $\mathrm{SER}_{\max}$. Annotations give pooled donor counts and Wilson 95% upper bounds.

The sole donor event in the GPT-OSS-120B condition occurs on SQuAD v2. Its 12-point curve exhibits joint base/donor realization at low λ, donor-suppressed base realization at 384, and weak base realization in the high-λ tail.

For Q3.6-35B we extend the grid to $\lambda = 576, 640, 704$, retaining designer settings and seed, and apply the joint utility criterion in Appendix X.2. The first two points satisfy all constraints; $\lambda = 704$ falls below the 0.75 base-SER floor. Each of the three donor conditions has $0/768$ pooled events, with Wilson 95% upper bound 0.005. LAMBADA has the worst accuracy delta at both passing points; at 640, MMLU changes by $+0.0003$ and ARC-C by -0.0078.

X.12 Measured computational cost

Figure 43 reports measured stage wall time using A6000 GPUs. Row design costs less than one second per λ at both scales; the forward-only hidden-statistics pass and clean transplantation take longer.

[3] `https://huggingface.co/Qwen/Qwen3.6-35B-A3B`
[4] `https://huggingface.co/openai/gpt-oss-120b`

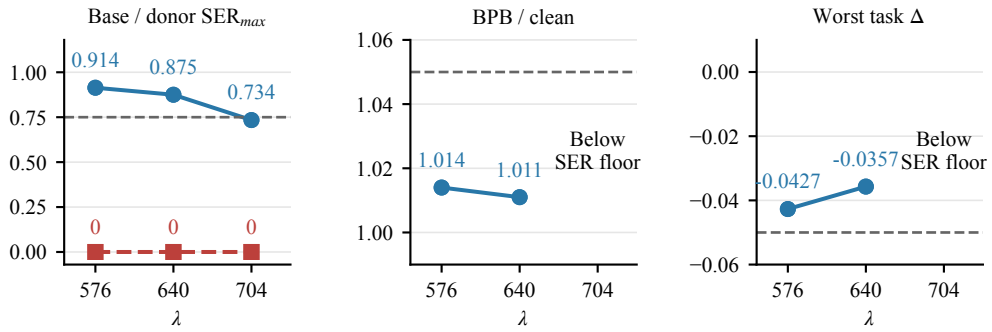

Figure 42: Joint realization and attacked-base utility on Q3.6-35B. The three panels show base and donor $\mathrm{SER}_{\max}$, BPB relative to the clean transplant, and worst task accuracy change at each λ. $\lambda = 704$ falls below the base-SER floor. Dashed grey lines mark the joint thresholds.

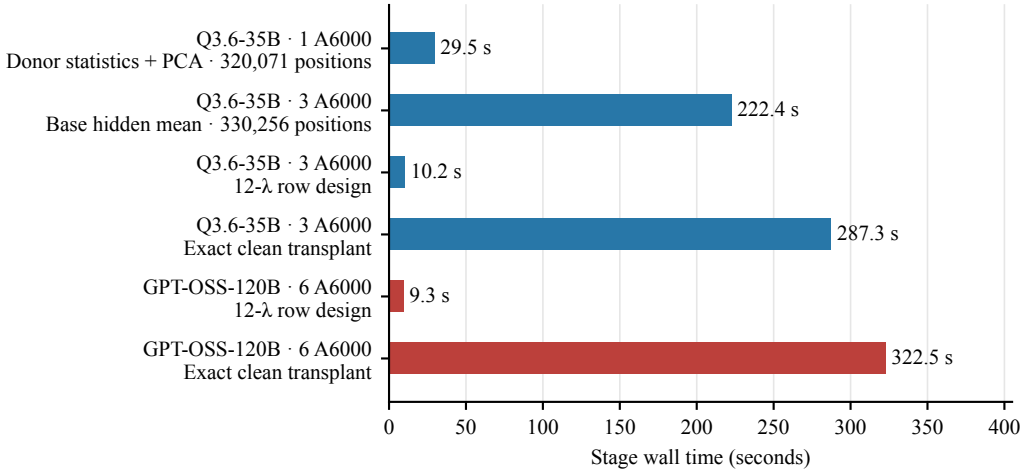

Figure 43: Stage wall time for the Gem2-2B donor and two large recipients. Labels identify the recipient, GPU count, stage and processed positions; bars show measured seconds. Row-design entries cover the entire 12-point grid. All GPUs are A6000s.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 1 state four contributions: a geometric characterization of shared-basis transplant (formalized via GSVD in Theorem K.2); a solver-agnostic reachability result for nonlinear coefficient maps (Theorem N.1); a spectral λ-window for joint donor-suppression and base-alignment (Theorem L.2); and empirical validation plus a discussion of potential misuse, including a deployer case study in which standard LoRA fine-tuning suppresses the breaker only on prompts whose distribution matches the training corpus and is therefore not a sufficient mitigation against this attack family. These are the contributions that the experimental sections (Section 5, Appendices H, W, F, G) and the theoretical appendices (J, K, L, N, O, I) actually deliver. Scope is calibrated to the open-weight composition ecosystem and shared-basis transplant operators, with the limits stated in the Implications paragraph of Section 1 and in Appendix A.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: A dedicated *Limitations and Future Directions* appendix (Appendix A) scopes the work to the shared-basis training-free transplant family (OMP, CLP, WECHSEL, FOCUS) in the regime where shared-basis transplant is currently deployed, points to the formal conditions under which a nonlinear solver does not alter the conclusion (Appendix I), and identifies principled algorithmic redesign of transplant operators as the defense direction most directly motivated by our findings, since the LoRA suppression demonstrated in Appendix F is bounded to prompts whose distribution matches the training corpus and does not cover the held-out out-of-distribution prompts we test, and is therefore not a sufficient mitigation against this attack family in our case study.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be

used reliably to provide closed captions for online lectures because it fails to handle technical jargon.

- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Definition 3.1 introduces the shared-basis transplant operator. The theoretical contributions are stated and proved across six appendices, each with its assumptions adjacent to the results that use them. Appendix J (Assumptions J.1–J.2) gives Theorem J.4 on OMP support and coefficient recovery under mutual-coherence and beta-min conditions, plus Corollary J.5. Appendix K (Assumption K.1) gives Theorem K.2, the GSVD construction for asymmetric realizability, with its random-anchor genericity Corollary K.3. Appendix L (Assumption L.1) gives Theorem L.2, the spectral λ-window. Appendix N gives Theorem N.1 together with operator-specific corollaries for FOCUS softmax (Lemma N.3), top-k mixtures (Lemma N.5, Corollary N.6), CLP (Lemma N.8), and WECHSEL (Lemma N.10). Appendix O gives Theorem O.4 on constructive single-statistic mimicry. Appendix Q gives Theorem Q.3, the end-to-end SER asymmetry composing the above. Appendix I (Assumptions I.1–I.4) gives Theorem I.5, Theorem I.7 (Carathéodory / Maurey-type rates), Corollary I.6, and Proposition I.8. Complete proofs accompany every numbered result.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Appendix T specifies model identifiers (Table 14), shared-token construction, public-text feature collection (corpus, sample sizes, seeds), the sparse-OMP designer hyperparameters (K=64, ridge ρ=10^{-3}), the per-pair λ sweep grid, the SER protocol (256 prompts/dataset, decoding parameters), the utility-evaluation protocol, the LoRA fine-tuning configuration (rank, α, dropout, learning rate, sequence length, optimizer settings), and the merge-mitigation operators with their references. The differentiable-operator designer details (FOCUS / CLP / WECHSEL) are in Appendix V. Supplementary code is provided.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code and configuration files are publicly available at `https://github.com/xz-liu/tokenforge`. All datasets used (WikiText-103, LAMBADA, MMLU, ARC-Challenge, SQuAD v2, GSM8K, Alpaca) are publicly available, and all model checkpoints are open-weight on HuggingFace; we cite them and list aliases in Appendix T. We deliberately do not release pre-compiled, weaponized breaker-token artifacts (Appendix R); we release the optimization framework, configurations, evaluation scripts, and the operator-matched designers needed to reproduce the reported numbers.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.

- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Appendix T specifies all experimental details: model pool, dataset splits, prompt-pool sizes, decoding parameters, sparse-OMP and differentiable-operator hyperparameters, LoRA settings, merge operators, and the per-pair λ selection procedure (Appendix T.2).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All accuracy metrics in the appendix utility tables (LAMBADA, MMLU, ARC-Challenge) are reported as $\mathrm{value} \pm \mathrm{stderr}$ from the standard `lm-evaluation-harness` accuracy estimator (binomial standard error over the evaluation set). SER is a binomial proportion over $n{=}256$ prompts per pool; the methodology note in Appendix B gives the closed-form Wilson 95% confidence interval, and per-cell intervals are annotated on the single-value SER tables (Tables 15, 12, 13). For the slash-pair per-pool tables, the same Wilson formula applies cell-by-cell and is stated in the appendix. Aggregate SER analyses use the full 65-pair OMP pool; pair-level numbers reported in the main text correspond to single fixed-λ runs and are stated as such.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).

- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All experiments except the two MoE recipients run on a single NVIDIA-class GPU (A100 / H100 80GB suffices for the standard-scale clique; the lightweight clique runs comfortably on a 24GB consumer GPU). The MoE recipients use three (Q3.6-35B) and six (GPT-OSS-120B) A6000 GPUs; Appendix X.12 reports measured stage wall times. The breaker-token designer is training-free and takes seconds to minutes per pair; SER and utility evaluation per pair take on the order of tens of minutes; LoRA fine-tuning runs are 50 epochs on $5{,}000$ examples per SFT corpus and complete in a few hours per cell on a single GPU. We additionally disclose that exploratory runs over alternative λ grids and operator variants used roughly an order of magnitude more compute than the runs reported in the paper.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work conforms with the NeurIPS Code of Ethics. We use only public open-weight models and standard public datasets; no human subjects are involved; we follow responsible-disclosure norms for the offensive construction (Appendix R); we do not release weaponized artifacts; and we provide a defender-facing auditing protocol (Appendix S). The paper includes an unnumbered *Ethics Statement* (in the appendix, immediately after the Limitations) stating these commitments.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Both directions are discussed across three places. The *Implications* paragraph in Section 1 names the negative-impact concern (potential misuse in the open-weight composition supply chain) and the positive-impact direction (the GSVD asymmetry-energy diagnostic for evaluating new shared-basis operators), distinguishing workflows that auto-accept donor-exclusive appended rows from those that quarantine new tokens. The unnumbered *Ethics Statement* (in the appendix, immediately after the Limitations) reiterates the

responsible-disclosure posture and the artifacts we deliberately do not release. Appendix R describes the release posture, under which no pre-built breaker-token artifacts or patched checkpoints are released, and Appendix S gives a defender-facing post-transplant auditing protocol that the community can extend.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: We deliberately do not release pre-compiled, weaponized breaker-token artifacts or patched checkpoints targeting specific deployed models (Appendix R); we release the optimization framework only at the level needed for researchers to reproduce our findings and develop defenses; and we provide a defender-facing post-transplant auditing protocol (Appendix S). We additionally characterize the limits of post-hoc deployer interventions: in a deployer case study, standard LoRA fine-tuning suppresses the breaker only on prompts whose distribution matches the training corpus (Appendix F), motivating embedding-side validation in the auditing protocol (Appendix S).

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All open-weight models (Qwen2 / Qwen2.5 / Qwen3, Gemma-2 / Gemma-3, Llama-3 / 3.1 / 3.2, Mistral-7B, Ministral-3B-Instruct, SmolLM2) and all datasets (WikiText-103, LAMBADA, MMLU, ARC-Challenge, SQuAD v2, GSM8K, Alpaca) are cited to their original publications in Appendix T. We use these assets within the terms of their respective public licenses. We additionally cite the transplant operators and tooling we evaluate (`mergekit`, OMP, CLP, WECHSEL, FOCUS), the LoRA implementation, and the spectral-outlier scanner (`Magikarp`).

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The new asset released with this submission is the breaker-token optimization framework (sparse OMP designer plus operator-matched FOCUS / CLP / WECHSEL designers) together with evaluation scripts. It is publicly available at `https://github.com/xz-liu/tokenforge`, with a README documenting the end-to-end pipeline (hidden-statistics collection, row design, donor patching, and transplantation) and SER evaluation. We deliberately do not release pre-built weaponized artifacts (Appendix R).

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper involves no crowdsourcing and no human-subjects research. All experiments are conducted on public open-weight models and public benchmark datasets.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human subjects are involved in this work, so IRB approval is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs are not part of the core methodology: the core methods (the breaker-token designer, the SER metric, and the auditing protocol) operate on model weights and embeddings directly. We disclose that LLMs were used as general-purpose writing assistants for prose paraphrasing, boilerplate code generation, and consistency checks; the unnumbered *LLM Use Statement* (in the appendix, alongside the Ethics Statement) describes the scope. The evaluated language models themselves (the donors and base/recipient checkpoints) are the subjects of study, not auxiliary tools, and are documented in Appendix T.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.