KNOW BEFORE YOU TRANSFER: PREDICTING AND CERTIFYING CROSS-MODEL KV CACHE REUSE

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-model inference pipelines repeatedly process the same context when requests move between models. At each handoff, the receiving model typically prefills the shared context from scratch, even though the sender has already processed it. Existing methods reduce this cost by transferring the sender's key-value (KV) cache into the receiver's cache space through a learned or closed-form map. Their effectiveness, however, varies widely across model pairs, and failures are hard to anticipate or detect: a transferred cache may yield a fluent answer that disagrees with the output of a full prefill. Reliable deployment therefore needs inexpensive answers to two questions: which model pairs admit effective transfer, and how much risk a transferred cache poses for a given request. We address both through a first-order analysis of how cache perturbations propagate through attention. The analysis shows that downstream deviation depends on where the reconstruction error lies relative to the receiver's query geometry, a dependence that error magnitude alone does not capture. From it we derive the **Cache Transferability Index**, an offline diagnostic that predicts a pair's achievable retention from second-moment statistics alone, without fitting a mapper. We also develop **CacheCert**, an online procedure that turns the same geometry into a per-request risk score and applies conformal risk control to obtain a distribution-free bound on the rate of unsafe transfers under exchangeability. Across eighteen directed model pairs from three families and five long-context workloads, the index reaches a Spearman correlation of 0.91 with retention at roughly one fortieth the cost of fitting and evaluating a mapper. CacheCert keeps empirical risk at or below its nominal target on every pair. At a 10 percent risk budget, it saves 66 percent of prefill work on LongBench multi-hop QA and reduces silent failures from 14.7 percent to 4.1 percent on an agent-report track. Furthermore, both components are backbone-agnostic and can compose with existing transfer and repair methods.

1 INTRODUCTION

Language model deployment increasingly involves multiple models with different costs and capabilities. Production systems route a single interaction across these models, escalate difficult queries from cheaper models to more expensive ones, and compose specialist agents that read and write each other's intermediate results (Chen et al., 2023; Ong et al., 2025; Ding et al., 2024; Wang et al., 2025). Each handoff carries the same structural cost. Before generating its first token, the receiving model must process the shared context through all of its layers, even though the sending model has just done so. As sessions lengthen and retrieval payloads grow, this repeated prefill can become a dominant source of end-to-end latency and serving cost.

Within a single model, cache reuse already removes much of this redundancy. Mainstream engines support prefix caching (Kwon et al., 2023; Zheng et al., 2024), and a substantial body of work extends reuse beyond the exact prefix under which a cache was computed (Gim et al., 2024; Yao et al., 2025; Hu et al., 2025; Yang et al., 2025b; Lu et al., 2025). Cross-model reuse is harder, because a key-value (KV) cache depends on the weights that produced it. Recent approaches handle this mismatch by learning or fitting a map from the sender's cache space to the receiver's. Same-architecture variants can reuse caches at most layers if a few sensitive layers are recomputed (Liu et al., 2024b), and agents with overlapping contexts can reuse caches after offset alignment (Ye et al., 2025). For heterogeneous models, trained projection and fusion modules bridge the two

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

cache spaces (Fu et al., 2026; Rossi et al., 2026; Dery et al., 2026; Lee et al., 2026). Within a model family, a closed-form per-head ridge map recovers much of the receiver's standalone quality without gradient training (Heo et al., 2026; Qu et al., 2026).

Existing work, however, leaves a practical gap: predicting whether transfer will succeed before paying for mapper fitting. Prior findings show why this gap matters. Closed-form transfer retains most of the receiver's standalone accuracy on some within-family pairs but collapses on others, and its authors attribute the difference to the directions in which residual error concentrates rather than to its magnitude alone (Heo et al., 2026). A third-party benchmark finds that repairs fitted to one producer's cache distribution degrade significantly when a different checkpoint writes the cache. The resulting answers can remain fluent while naming the wrong entity (Shi & Lou, 2026). Another study shows that reuse can flip a judge's decision while leaving the surface text plausible (Liang et al., 2026). Together, these findings reveal a variance problem across model pairs and a silent-failure problem within individual requests. Averages reported over selected pairs can hide both.

We argue that the two problems share a root cause and can therefore be addressed by a common analysis. A transferred cache carries residual error in nearly every direction, but the error matters only where the receiver's queries have mass. Formalizing this dependence yields a quantity that can be evaluated at two granularities. Averaged offline over a calibration corpus, it predicts how much quality a given pair can retain. Conditioned online on the query of a single request, it scores how much that request is at risk. We build an offline diagnostic and a runtime mechanism on this shared foundation. By design, we keep the statistical guarantee separate from the analysis. The first-order approximation serves only to rank requests, while validity comes from conformal risk control under exchangeability and does not require the approximation to be exact.

We summarize our contributions as follows:

- We derive a per-head positive semidefinite sensitivity operator from a first-order expansion of attention. The operator weights a cache residual by how strongly the receiver's queries probe each of its directions, which explains why error placement matters more than error magnitude.
- We introduce the Cache Transferability Index, which scores model pairs directly from second moments without fitting a mapper, at about one-fortieth the cost of mapper fitting and evaluation.
- We propose CacheCert, which measures the self-consistency defect of a transferred cache at a few recomputed probe positions. It calibrates a monotone ladder from acceptance to dense prefill via conformal risk control, yielding a distribution-free bound on the unsafe-transfer rate.
- We evaluate the framework on eighteen directed model pairs and five workloads against six transferability predictors and six runtime baselines, on top of four transfer backbones. The guarantee holds on every pair, including the three on which the backbone fails outright.

2 Related Work

Cross-context cache reuse handles changes in the prefix and in token positions. Exact prefix caching is lossless but requires identical preceding tokens (Kwon et al., 2023; Zheng et al., 2024). Relaxations include modular reuse with position re-encoding (Gim et al., 2024), selective recomputation guided by attention deviation (Yao et al., 2025), position-independent caching with boundary repair (Hu et al., 2025), adaptive parallel encoding with attention calibration (Yang et al., 2025b), and training models to consume independently encoded blocks (Ma et al., 2025; Yang et al., 2026; Lu et al., 2025). Storage and transport are complementary concerns (Liu et al., 2024c). A recent benchmark finds that position realignment largely suffices unless a query requires joint reasoning across sources; closing the remaining gap then requires recomputation at inference time or additional training (Shi & Lou, 2026). We complement these methods by predicting transferability and certifying reuse with whichever repair is in place.

Cross-model cache reuse must also account for differences in model weights. DroidSpeak reuses most layers across models that share an architecture by recomputing a few sensitive layers (Liu et al., 2024b), and adapter-based serving exploits attention projections that remain unchanged before adapter activation (Hu et al., 2022; Li et al., 2025; Woo et al., 2026). For heterogeneous models, learned communication modules rely on cache projection and fusion (Fu et al., 2026), shared low-dimensional channels (Rossi et al., 2026), adapters into a common representation space (Dery et al.,

2026), or token-gated translator mixtures with target-side trajectory correction (Lee et al., 2026). Closed-form alternatives fit per-head ridge maps in a rotation-free content space and select predictive source layers (Heo et al., 2026; Qu et al., 2026). Other approaches let models collaborate directly in latent space (Zou et al., 2026). These approaches are consistent with evidence that independently trained models learn alignable representations (Huh et al., 2024; Moschella et al., 2023). Our framework treats existing transfer methods as backbones and predicts how their quality varies.

3 PRELIMINARIES

Cross-model transfer. Let S be a source (producer) model and T a target (receiver) model. Given a shared context c of N tokens, a full prefill of c through T yields the receiver's native cache $\mathbf{K}^{\ell h}, \mathbf{V}^{\ell h} \in \mathbb{R}^{N\times d}$ at layer ℓ and key-value head h. A *transfer map* Φ estimates $\widehat{\mathbf{K}}^{\ell h}, \widehat{\mathbf{V}}^{\ell h}$ from the source cache without running T over c. We write the residual as

$$\mathbf{E}_K^{\ell h} = \widehat{\mathbf{K}}^{\ell h} - \mathbf{K}^{\ell h}, \qquad \mathbf{E}_V^{\ell h} = \widehat{\mathbf{V}}^{\ell h} - \mathbf{V}^{\ell h}. \tag{1}$$

Following standard practice, we strip rotary position encoding (Su et al., 2024) before mapping keys. Residuals are thus defined in a position-free content space and transfer across context lengths (Heo et al., 2026). All models we consider use grouped-query attention (Ainslie et al., 2023); we leave latent attention (Liu et al., 2024a) to future work (Section 6).

Attention and the object we care about. For a query vector $\mathbf{q} \in \mathbb{R}^d$ produced by the receiver during decoding, the attention output of head (ℓ, h) over the cached context is

$$\mathbf{o}(\mathbf{q}) = \mathbf{V}^\top \mathbf{a}(\mathbf{q}), \qquad \mathbf{a}(\mathbf{q}) = \mathrm{softmax}\Big(\tfrac{1}{\sqrt{d}}\mathbf{K}\mathbf{q}\Big) \in \Delta^{N-1}, \tag{2}$$

where we drop layer and head indices when unambiguous. Downstream behaviour is determined not by $\mathbf{E}_K$ or $\mathbf{E}_V$ themselves but by the deviation $\Delta\mathbf{o}(\mathbf{q})$ they induce. Section 4 shows that the map from the former to the latter is highly anisotropic.

Quality accounting. We score every configuration with the performance-gap recovered (PGR) of Shi & Lou (2026). PGR normalizes a method between a floor, where the receiver answers without any shared context, and a ceiling, where the receiver runs a full dense prefill of the same material. It is zero at the floor, one at the ceiling, and can be negative. This matters here because a badly transferred cache is often worse than no cache. We also report retention $R = S_{\text{transfer}}/S_{\text{dense}}$ for comparability. We account for cost on three separate axes rather than as a single aggregate: prefill compute saved, key-value bytes stored, and time to first token on one shared backend.

Conformal risk control. Let $\lambda \in \Lambda \subset \mathbb{R}$ index a family of decision rules, and let $L_i(\lambda)$ be a loss on example i that is bounded by B, non-increasing, and right-continuous in λ. Given n exchangeable calibration examples with empirical risk $\widehat{R}_n(\lambda)$, the choice

$$\widehat{\lambda} = \inf\Big\{\lambda \in \Lambda : \tfrac{n}{n+1}\widehat{R}_n(\lambda) + \tfrac{B}{n+1} \le \alpha\Big\} \tag{3}$$

guarantees $\mathbb{E}[L_{n+1}(\widehat{\lambda})] \le \alpha$ for a fresh exchangeable example (Angelopoulos et al., 2024). In Section 4.3, λ controls how aggressively transferred caches are accepted, with larger λ accepting fewer of them. Validity places no requirement on the quality of the score: a poor score yields a valid but conservative rule. This is why we can build the score from a first-order approximation without endangering the guarantee.

4 METHODOLOGY

4.1 WHERE TRANSFER ERROR HURTS

Existing transfer methods are fitted and evaluated by coordinate-wise reconstruction quality, such as per-head R^2 or mean squared error. These statistics weight every residual direction equally, so they ignore how sensitive the receiver is to where the error lands. To characterize this sensitivity, we expand Eq. (2) to first order in the cache residual:

$$\Delta\mathbf{o}(\mathbf{q}) \;\approx\; \underbrace{\tfrac{1}{\sqrt{d}}\,\mathbf{V}^\top \boldsymbol{\Psi}(\mathbf{q})\,\mathbf{E}_K\,\mathbf{q}}_{\text{key path}} + \underbrace{\mathbf{E}_V^\top \mathbf{a}(\mathbf{q})}_{\text{value path}}, \qquad \boldsymbol{\Psi}(\mathbf{q}) = \mathrm{diag}(\mathbf{a}(\mathbf{q})) - \mathbf{a}(\mathbf{q})\mathbf{a}(\mathbf{q})^\top, \tag{4}$$

where $\mathbf{\Psi}(\mathbf{q})$ is the covariance operator of the attention distribution. Three properties of Eq. (4) guide the rest of the paper. First, the key path reaches the output only through the projection $\mathbf{E}_K\mathbf{q}$. Residual components orthogonal to the receiver's query cone therefore contribute nothing through this path, however large they are. Second, $\mathbf{\Psi}(\mathbf{q})\mathbf{1} = \mathbf{0}$, so a common shift of all key logits is annihilated; a uniform bias in the transferred keys leaves attention unchanged for any fixed query. Third, the two paths behave differently. The value path simply averages value errors with the attention weights, whereas the key path is modulated by $\mathbf{\Psi}(\mathbf{q})$. Its strength thus depends on how attention mass is spread across positions, and it vanishes when attention is fully peaked. Together, these properties explain geometrically why equal reconstruction quality can yield unequal downstream retention, and why a nonlinear mapper can help by moving residual energy out of attention-sensitive subspaces (Heo et al., 2026; Qu et al., 2026).

To obtain a computable quantity, we average over the receiver's queries. Let $\mathcal{Q}$ be the query distribution of head (ℓ, h) on a calibration corpus. Substituting Eq. (4) and taking expectations yields a positive semidefinite operator on the *channels* of a key residual row,

$$\mathbf{G}^{\ell h} = \frac{\mathbb{E}_{\mathbf{q}\sim\mathcal{Q}}\left[\kappa(\mathbf{q})\,\mathbf{q}\mathbf{q}^\top\right]}{d\,\mathbb{E}_{\mathbf{q}\sim\mathcal{Q}}[\kappa(\mathbf{q})]} \in \mathbb{R}^{d\times d}, \qquad \kappa(\mathbf{q}) = \left\|\mathbf{V}^\top\mathbf{\Psi}(\mathbf{q})\right\|_F^2, \tag{5}$$

that is, the query second moment reweighted by the attention sensitivity $\kappa(\mathbf{q})$. It comes with two diagonal position weights $\mathbf{W} = \mathrm{diag}(w)$ and $\mathbf{U} = \mathrm{diag}(u)$, which record how much traffic each cached position carries,

$$w_i = \mathbb{E}_{\mathbf{q}}\big[(\mathbf{\Psi}(\mathbf{q})\mathbf{V}\mathbf{V}^\top\mathbf{\Psi}(\mathbf{q}))_{ii}\big], \qquad u_i = \mathbb{E}_{\mathbf{q}}\big[a_i(\mathbf{q})^2\big]. \tag{6}$$

We call $(\mathbf{G}^{\ell h}, \mathbf{W}^{\ell h}, \mathbf{U}^{\ell h})$ the *sensitivity operator* of the head. The value path has no channel anisotropy, because $\mathbf{E}_V$ enters Eq. (4) without passing through $\mathbf{q}$; this is the precise sense in which it is the benign path. The normalized sensitivity-weighted residual is then

$$\rho^{\ell h}(\mathbf{E}) = \frac{\mathrm{tr}\big(\mathbf{G}^{\ell h}\,\mathbf{E}_K^\top\mathbf{W}\mathbf{E}_K\big) + \mathrm{tr}\big(\mathbf{E}_V^\top\mathbf{U}\mathbf{E}_V\big)}{\mathrm{tr}\big(\mathbf{G}^{\ell h}\,\mathbf{K}^\top\mathbf{W}\mathbf{K}\big) + \mathrm{tr}\big(\mathbf{V}^\top\mathbf{U}\mathbf{V}\big)}, \tag{7}$$

where the rows of the key matrices are centered with weights w. Appendix A.5 explains why this centering is needed for Eq. (7) to inherit the annihilation property of Eq. (4), and why it costs nothing rather than adding an assumption. Appendix A states the first-order approximation formally, bounds its second-order remainder, and derives the position-channel factorization that reduces the exact $N \times N$ operator to Eqs. (5) and (6). The eigenspectrum of $\mathbf{G}^{\ell h}$ is informative in its own right. Heads whose operator has low effective rank are forgiving, since residual energy must land in a few specific directions to cause damage, whereas heads with a flat spectrum are brittle. This structure shapes the per-layer profile that orders fallbacks in Section 4.3; Appendix B reports the spectra.

4.2 The Cache Transferability Index

We want to predict how much quality transfer can retain on a pair (S, T) before committing to it. The direct approach fits a mapper and evaluates it downstream, but this spends most of the cost before the pair's suitability is known. We avoid it through a simple observation: under the weighting of Eq. (7), the minimum residual energy over the entire affine transfer family has a closed form.

Fix a head (ℓ, h). Let $\mathbf{x} \in \mathbb{R}^{d_s}$ denote the source features the mapper uses at that head, namely the source key and value rows concatenated over the selected source layers, and let $\mathbf{y} \in \mathbb{R}^d$ denote the corresponding target row. For any affine map $\mathbf{x} \mapsto \mathbf{\Theta}\mathbf{x} + \mathbf{b}$, standard generalized least squares gives

$$\min_{\mathbf{\Theta},\mathbf{b}} \mathbb{E}\big[(\mathbf{y} - \mathbf{\Theta}\mathbf{x} - \mathbf{b})^\top\mathbf{G}\,(\mathbf{y} - \mathbf{\Theta}\mathbf{x} - \mathbf{b})\big] = \mathrm{tr}\big(\mathbf{G}\,\mathbf{\Sigma}_{y|x}\big), \qquad \mathbf{\Sigma}_{y|x} = \mathbf{\Sigma}_{yy} - \mathbf{\Sigma}_{yx}\mathbf{\Sigma}_{xx}^{-1}\mathbf{\Sigma}_{xy}, \tag{8}$$

where the expectation carries the position weights of Eq. (6) and $\mathbf{\Sigma}_{y|x}$ is the Schur complement of the centered joint second-moment matrix. For a key row, $\mathbf{G}$ is the head's channel operator and the weights are w; for a value row, $\mathbf{G}$ is replaced by the identity and the weights by u. The right-hand side requires neither optimization nor downstream evaluation. It depends only on joint second moments of the two models' caches, which one pass over a small calibration corpus provides.

Residuals at different depths are not equally consequential, because later layers reshape a perturbation injected early. We measure this with a one-shot propagation probe. For each receiver layer ℓ, we inject isotropic noise of controlled relative norm into the receiver's own cache at that layer only. We

record the Kullback-Leibler divergence of the next-token distribution from the unperturbed run. Its slope against the squared relative norm defines an amplification weight γ_ℓ, matching the quadratic scale of Eq. (7). The probe costs one forward pass per layer on a few dozen sequences and runs once per receiver, not per pair. Combining the two ingredients gives the index:

Definition 1 (Cache Transferability Index). *For a directed pair (S, T) with calibration corpus $\mathcal{C}$,*

$$\textsc{Cti}(S \to T) = 1 - \frac{\sum_\ell \gamma_\ell \sum_h \operatorname{tr}\big(\mathbf{G}^{\ell h} \boldsymbol{\Sigma}^{\ell h}_{y|x}\big)}{\sum_\ell \gamma_\ell \sum_h \operatorname{tr}\big(\mathbf{G}^{\ell h} \boldsymbol{\Sigma}^{\ell h}_{yy}\big)} \in [0, 1]. \tag{9}$$

Three remarks clarify the scope and use of CTI. First, the index lies in $[0, 1]$ because $\mathbf{0} \preceq \boldsymbol{\Sigma}_{y|x} \preceq \boldsymbol{\Sigma}_{yy}$, and it is defined over the affine family. Nonlinear families that contain affine maps can match or improve the affine optimum under the same weighted objective. Section 5.2 shows that CTI remains predictive for trained nonlinear backbones, which supports its use for screening. Second, the per-layer terms of Eq. (9), which we write as $\gamma_\ell \rho^\ell$, form a reusable profile for prioritizing layers for recomputation. This offers an analytical alternative to the layer-selection search of Liu et al. (2024b). Third, $\mathbf{G}, \mathbf{W}, \mathbf{U}$ are estimated from the receiver alone, whereas $\boldsymbol{\Sigma}$ comes from paired activations. The receiver statistics can therefore be reused, and adding a new source model to an existing receiver requires only the cross-covariance pass.

4.3 CACHECERT: PER-REQUEST CERTIFICATION

A good pair-level index does not settle the harder question of request-level risk. Even on a pair with high CTI, transfer errors can damage individual requests: the sensitivity operator in Eq. (5) is an expectation over queries, whereas a served request has one specific query. The reported silent failures illustrate this gap (Shi & Lou, 2026; Liang et al., 2026), which motivates the request-level assessment that CACHECERT provides.

A computable risk score. At serving time the native receiver cache $\mathbf{K}, \mathbf{V}$ is unavailable, since producing it is the very cost we want to avoid. We therefore replace the unobservable residual with a *fixed-point residual* measured on a small probe set $\mathcal{P} \subset \{1, \dots, N\}$ with $|\mathcal{P}| = m \ll N$. At each probe position, we recompute the receiver's key and value at layer ℓ from the transferred cache at the layers below ℓ, exactly as selective recomputation methods do (Yao et al., 2025). We then compare the result with the transferred entry at that position:

$$\widetilde{\mathbf{E}}^{\ell h}_{\mathcal{P}} = \big[\widehat{\mathbf{K}}^{\ell h}_{\mathcal{P}} - \mathcal{T}^{\ell h}\big(\widehat{\mathbf{K}}^{<\ell}, \widehat{\mathbf{V}}^{<\ell}\big)_{\mathcal{P}}\,;\ \cdot\,\big], \tag{10}$$

where $\mathcal{T}^{\ell h}$ denotes the receiver's layer-ℓ computation and the value block is defined analogously. This is not an unbiased estimate of the true residual, and we do not claim it is. It is a self-consistency defect that is cheap, always computable, and empirically monotone in downstream damage. The calibration step described below turns this good ordering into a rigorous guarantee. This makes the approximation acceptable, whereas it would not be if we derived a bound from the score. The score of request r is the sensitivity-weighted probe residual, with the operator rebuilt from the *live* query block $\mathcal{Q}_r$, namely the instruction and question tokens the receiver appends after the shared context:

$$s_r = \sum_\ell \gamma_\ell \sum_h \rho^{\ell h}\big(\widetilde{\mathbf{E}}^{\ell h}_{\mathcal{P}}\,;\, \mathcal{Q}_r\big). \tag{11}$$

Since the native cache is unavailable, both the operator and the normalizer of Eq. (7) are evaluated on the transferred cache. Query conditioning makes the score per-request rather than per-pair, and Section 5.4 shows that it accounts for a large share of the achievable efficiency. Probe positions are drawn by stratified sampling over the source model's attention mass, so that high-traffic positions are represented; the ablation compares this with uniform sampling.

A monotone fallback ladder. Instead of a binary accept-or-prefill decision, we expose a ladder of increasingly expensive repairs indexed by j. Rung $j = 0$ accepts the transferred cache as is; $j = 1$ recomputes the k_1 layers with the largest profile values $\gamma_\ell \rho^\ell$; $j = 2$ recomputes $k_2 > k_1$ such layers; and $j = 3$ runs a full dense prefill. The layer order is not a heuristic: it is the per-layer profile of Eq. (9), so the analysis that screens pairs also decides where the repair budget goes. A single scalar $\lambda \in [0, 1]$ selects the rung through nested score quantiles $\tau_0(\lambda) \le \tau_1(\lambda) \le \tau_2(\lambda)$ estimated on calibration data. Request r goes to the cheapest rung whose threshold it clears, that is, to the smallest j with $s_r \le \tau_j(\lambda)$, and to the dense rung if no such j exists. Each τ_j is non-increasing in λ, so raising λ can only move a request up the ladder. Cost is therefore non-decreasing in λ, and risk is non-increasing whenever deeper repair does not increase the loss, as Eq. (3) requires.

Calibration and guarantee. Define the per-request loss $L_r(\lambda) = \mathbf{1}\{D(p_r^{\lambda}, p_r^{\text{dense}}) > \varepsilon\}$. Here p_r^{λ} is the output under ladder setting λ, p_r^{dense} is the output of a full receiver prefill, and D is a task-level disagreement functional; the top rung has $L_r \equiv 0$ by construction. Deeper repair can occasionally move an answer away from the dense one, so we calibrate on the monotonized loss of Corollary 4, which removes the need for rung monotonicity (Appendix A.7). Applying Eq. (3) to n exchangeable calibration requests yields $\widehat{\lambda}$ with

$$\mathbb{E}\big[L_{n+1}(\widehat{\lambda})\big] \;\le\; \alpha, \tag{12}$$

a distribution-free bound on the rate at which served answers differ from those of a dense prefill. Sweeping α traces a risk–cost frontier from which a deployer can choose an operating point. Three points clarify the scope of this guarantee. First, it holds for any transfer backbone under the stated assumptions; for poorly matched pairs, $\widehat{\lambda}$ simply routes more requests to higher rungs. Second, distribution shift can violate exchangeability, and the guarantee then need not hold. Section 5.3 assesses empirical robustness by calibrating on one workload and testing on another. Third, CACHECERT composes with existing repair methods, each of which can serve as a ladder rung.

Cost. In the offline phase, CTI processes $|\mathcal{C}|$ sequences through both models and performs $O(d^2)$ second-moment accumulation per head, without mapper fitting or downstream evaluation. At runtime, CACHECERT requires m/N of dense-prefill compute for probing plus $O(mLd^2)$ for sensitivity weighting, with a combined cost below two percent of dense-prefill compute in our settings.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Models and pairs. We use eighteen directed pairs. Twelve are within-family pairs drawn from Qwen3 (1.7B, 4B, 8B, 14B, 32B) (Yang et al., 2025a), Llama-3.x (1B, 3B, 8B, 70B) (Grattafiori et al., 2024), and Gemma-3 (4B, 12B, 27B) (Team et al., 2025). The other six cross model families and represent the hard regime that current transfer methods do not claim to handle. We also include four cross-checkpoint pairs, in which a further-trained specialist sibling writes the cache read by its base lineage, following the producer-change protocol of Shi & Lou (2026).

Transfer backbones. Both components are backbone-agnostic, so we run every experiment on top of four transfer methods that span the design space: a closed-form per-head ridge map with top-k source-layer selection (Heo et al., 2026); its head-local, attention-weighted variant (Qu et al., 2026); a trained projection-and-fusion module in the style of Cache-to-Cache (Fu et al., 2026); and selective layer reuse that recomputes sensitive layers (Liu et al., 2024b). Unless noted otherwise, main-text numbers use the first backbone, and Section 5.2 reports how CTI generalizes across backbones.

Datasets. We use 5 workloads with frozen query sets. Three are LongBench v1 QA subsets covering scientific-paper QA, single-document QA, and multi-hop QA (Bai et al., 2024; Yang et al., 2018; Ho et al., 2020). The fourth is FRAMES over a pinned Wikipedia snapshot for multi-source retrieval (Krishna et al., 2025). The fifth is an agent-report track, in which specialist agents write reports over disjoint material and a coordinator integrates them, following the protocol of Shi & Lou (2026). For pair-level ground truth, we measure continuation retention on MuSiQue (Trivedi et al., 2022) and a held-out slice of HELMET (Yen et al., 2025). Context lengths range from 4K to 29K tokens.

Baselines. For pair-level prediction, we compare against six statistics in current use: per-head key and value R^2; layer-wise cache mean squared error; centered kernel alignment between the two models' hidden states (Kornblith et al., 2019); weight-space distance between key projections, as used by Shi & Lou (2026); attention-output cosine similarity, the post-hoc diagnostic of Heo et al. (2026); and the validation loss of an actually fitted mapper, the expensive procedure that CTI aims to replace. For runtime gating, we compare against no gate (always transfer), a random gate, predictive-entropy gating, logit-margin gating, fixed-budget selective recomputation (Yao et al., 2025), and fixed layer selection (Liu et al., 2024b). Dense prefill is the ceiling and no-context is the floor.

Metrics. Pair-level prediction is scored by Spearman and Pearson correlation with measured retention, by AUROC for identifying catastrophic pairs (retention below 0.8), and by top-3 selection regret, the retention lost by deploying the three pairs a predictor ranks highest instead of the true top 3. Request-level results are scored by PGR (Shi & Lou, 2026), by empirical risk (the violation rate in Eq. (12)), by silent-failure rate, and by three cost axes of Section 3. A silent failure is an answer that an autorater judges fluent but that disagrees with the dense-prefill answer on the key entity.

Table 1: Pair-level prediction quality over 18 directed pairs, mean ± standard deviation over 5 calibration splits. Cost is the GPU time needed to produce the statistic for one pair on a single H100, measured on the 14B→32B configuration. Bold marks the best result.

Predictor	Spearman↑	Pearson↑	AUROC↑	Top-3 regret↓	Cost (GPU-min)↓
Key/Value R^2	0.34 ± 0.09	0.29 ± 0.11	0.57 ± 0.07	0.121	1.8
Layer-wise KV MSE	0.28 ± 0.10	0.25 ± 0.12	0.54 ± 0.08	0.138	1.8
Weight-space distance	0.41 ± 0.08	0.37 ± 0.09	0.61 ± 0.07	0.107	**0.3**
Hidden-state CKA	0.55 ± 0.07	0.51 ± 0.08	0.68 ± 0.06	0.094	2.3
Attention-output similarity	0.71 ± 0.06	0.69 ± 0.06	0.81 ± 0.05	0.056	6.4
Fitted mapper val. loss	0.79 ± 0.05	0.77 ± 0.05	0.86 ± 0.04	0.041	84.6
CTI (ours)	$\mathbf{0.91 \pm 0.03}$	$\mathbf{0.88 \pm 0.04}$	$\mathbf{0.95 \pm 0.02}$	**0.019**	2.1

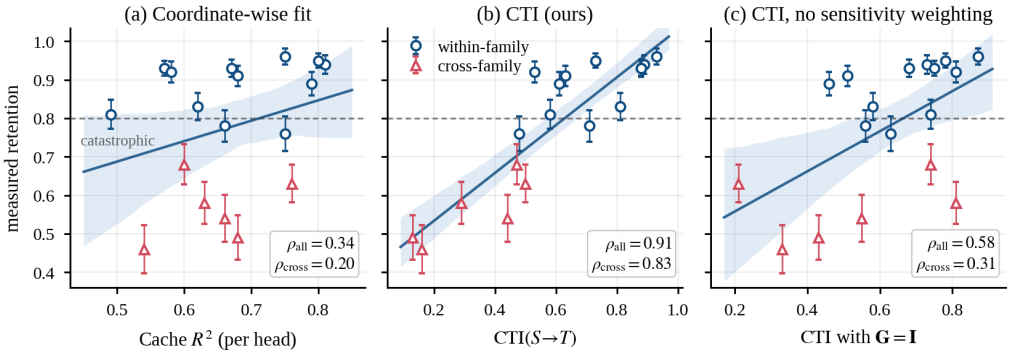


Figure 1: Retention v.s. cache R^2, CTI, and CTI w/o sensitivity weighting. Each point is a directed pair; bars show variation across five workloads, and shading shows bootstrap confidence bands.

5.2 Does CTI predict transferability better and more cheaply?

Table 1 and Figure 1 show that CTI predicts cross-model transferability far more accurately than reconstruction metrics. Per-head R^2 reaches a Spearman correlation of only 0.34 ± 0.09 with measured retention and an AUROC of 0.57 ± 0.07 for identifying catastrophic pairs, which amounts to weak ranking and near-chance discrimination. This matches the first-order analysis of Eq. (4): R^2 ignores how residuals align with the receiver's query geometry, and this alignment determines their effect on the attention output. By accounting for it, CTI reaches a Spearman correlation of 0.91 ± 0.03 and an AUROC of 0.95 ± 0.02, both above the fitted-mapper validation loss. It also costs far less: 2.1 GPU-minutes, compared with 84.6 for the fitted-mapper baseline.

Two further findings support its use in deployment decisions. First, all predictors improve on within-family pairs, where the gap narrows. On the six cross-family pairs, however, R^2 reaches only 0.20 ± 0.13 Spearman, compared with 0.84 ± 0.05 for CTI. Removing sensitivity weighting substantially degrades CTI's ranking of these pairs (Figure 1(c)). Second, CTI remains predictive beyond its affine formulation. Replacing ridge transfer with the trained fusion backbone (Fu et al., 2026) substantially changes absolute retention, yet the unchanged CTI scores still reach 0.87 ± 0.04 Spearman. CTI thus captures pair-level compatibility across the evaluated backbones and supports inexpensive screening before mapper fitting and downstream evaluation.

5.3 Does CacheCert certify at acceptable cost?

Risks of ungated transfer across workloads. As Table 2 shows, the ungated ridge backbone reaches a PGR of 0.38 on multi-hop QA while disagreeing with dense prefill on 44 percent of requests. On the agent-report track its PGR is negative, meaning that the transferred cache is worse than giving the coordinator no reports at all. This reproduces the toxicity that Shi & Lou (2026) report for misaligned assembly, and it confirms that averaging hides the problem: the same configuration reaches a PGR of 0.61 on FRAMES, where questions draw on fewer sources. The silent-failure column is the more troubling result. On agent reports, 14.7 percent of ungated answers are judged fluent yet disagree with the dense answer on the key entity, a failure that a reader cannot detect.

Table 2: Request-level results aggregated over 12 within-family pairs with the closed-form ridge backbone. Risk: empirical risk (Eq. (12)); Silent: rate of fluent answers disagreeing with dense prefill on the key entity; Work: fraction of dense-prefill compute saved. Gates match CACHECERT's acceptance rate at $\alpha = 0.10$. Results for all 18 pairs are broken out in Appendix C.

Method	LongBench multi-hop PGR↑	Risk↓	Work↑	FRAMES PGR↑	Risk↓	Work↑	Agent reports PGR↑	Silent↓	Work↑
Dense prefill (ceiling)	1.000	0.000	0.00	1.000	0.000	0.00	1.000	0.000	0.00
No gate (always transfer)	0.38	0.44	**0.95**	0.61	0.29	**0.95**	−0.21	0.147	**0.95**
Random gate	0.61	0.26	0.66	0.74	0.18	0.66	0.19	0.092	0.66
Entropy gate	0.68	0.22	0.66	0.78	0.15	0.66	0.27	0.081	0.66
Logit-margin gate	0.71	0.19	0.66	0.80	0.14	0.66	0.31	0.074	0.66
DroidSpeak layer sel.	0.74	0.17	0.71	0.82	0.13	0.71	0.44	0.066	0.71
CacheBlend (15% recompute)	0.79	0.15	0.83	0.85	0.11	0.83	0.63	0.058	0.83
CACHECERT ($\alpha = 0.05$)	**0.91**	**0.041**	0.48	**0.94**	**0.038**	0.53	**0.82**	**0.026**	0.44
CACHECERT ($\alpha = 0.10$)	0.85	0.083	0.66	0.90	0.079	0.70	0.74	0.041	0.62
CACHECERT ($\alpha = 0.20$)	0.78	0.168	0.81	0.84	0.161	0.84	0.61	0.079	0.78
CACHECERT + CacheBlend rungs	0.89	0.084	0.80	0.93	0.077	0.82	0.79	0.039	0.76

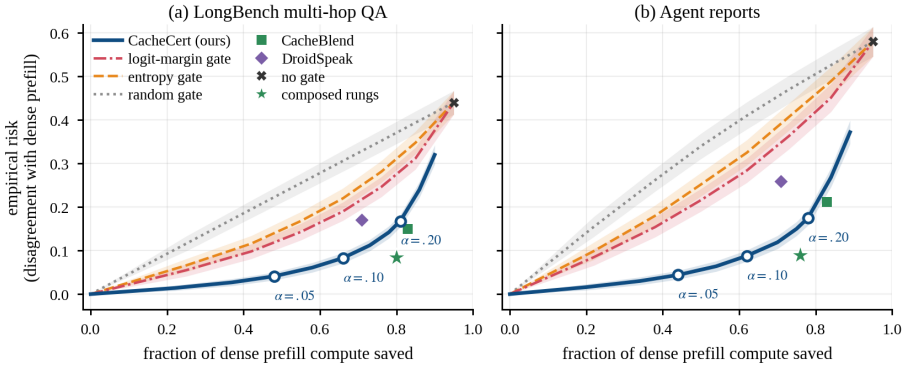


Figure 2: Risk–cost frontiers over five splits; shaded regions are one standard deviation. Curves sweep α for CACHECERT and thresholds for the gates; markers show fixed-budget CacheBlend and DroidSpeak, and the star denotes CACHECERT with a CacheBlend rung.

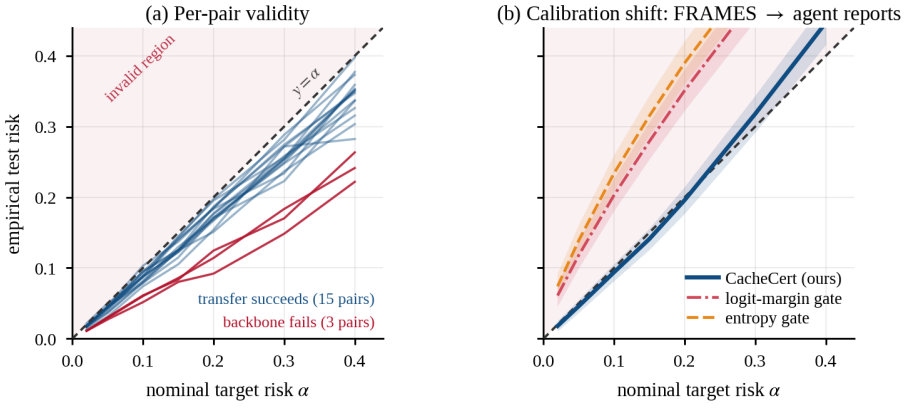


Figure 3: Empirical risk versus nominal target α. (a) All 18 pairs lie at or below the diagonal, including the three with failed backbones. (b) Under workload shift (FRAMES → agent reports), uncertainty gates exceed the target, while CACHECERT does so only at $\alpha = 0.3$ and 0.4.

Risk control with reduced prefill cost. At $\alpha = 0.10$, CACHECERT saves 66 percent of prefill work on multi-hop QA while holding empirical risk at 0.083, and it cuts silent failures on agent reports to 4.1 percent. The random, entropy, and logit-margin gates, operated at the same 66 percent work saved, are all substantially worse: logit-margin gating runs at 0.19 risk and entropy gating at 0.22.

Table 3: Component ablations. Upper block: Spearman correlation between CTI and retention. Lower block: CACHECERT risk and work saved on LongBench multi-hop QA at $\alpha = 0.10$.

	Variant	All pairs	Cross-family
CTI	Full	**0.91**	**0.84**
	− sensitivity weighting ($\mathbf{G} = \mathbf{I}$)	0.58	0.31
	− query conditioning in $\mathbf{G}$	0.67	0.44
	− layer amplification γ_ℓ	0.77	0.69
	− rotation stripping	0.71	0.52
	Variant	**Risk**	**Work saved**
CACHECERT	Full ($m = 32$, stratified, graded)	0.083	**0.66**
	− query conditioning (use $\mathbb{E}_\mathcal{Q}$)	0.081	0.51
	− graded ladder (binary accept)	0.079	0.44
	− stratified probes (uniform)	0.084	0.58
	$m = 8$	0.086	0.53
	$m = 64$	0.082	0.67

Figure 2 places these operating points on the full risk–cost frontiers. Uncertainty heuristics measure how unsure the model is about its answer, which correlates only loosely with whether the cache it read was faithful. The fixed-point probe measures the defect directly, and query conditioning identifies the part of the defect that the current request will actually touch. The last row of Table 2 illustrates composition. Installing CacheBlend as an intermediate rung yields a PGR of 0.89 at 80 percent work saved, far above CacheBlend alone (0.79 at 83 percent) and better on both axes than CACHECERT at $\alpha = 0.10$.

Risk control despite backbone failure. Figure 3(a) shows empirical risk at or below the nominal target for all 18 pairs in all 5 splits. On the 3 pairs with ungated retention below 0.55, CACHECERT routes 88–94% of requests to dense prefill and saves only 6–11% of prefill compute. This complements the offline screening of CTI: if an unsuitable pair is deployed anyway, its failure surfaces as measured cost rather than as silent errors. Figure 3(b) tests workload shift by calibrating on FRAMES and evaluating on agent reports. The uncertainty gates exceed the target across the tested settings, whereas CACHECERT does so only at $\alpha = 0.3$ and 0.4. This pattern suggests that more permissive policies are more vulnerable to calibration mismatch. The shift experiment assesses empirical robustness only; the formal guarantee requires exchangeability.

5.4 WHICH COMPONENTS OF THE ANALYSIS ACTUALLY CARRY THE WEIGHT?

Table 3 evaluates the contribution of each component. For CTI, removing sensitivity weighting causes the largest drop in Spearman correlation, from 0.91 to 0.58 overall and from 0.84 to 0.31 on cross-family pairs. This supports the importance of error placement identified in Section 4.1. Removing query conditioning, layer amplification, or rotation stripping also lowers correlation, with larger drops on cross-family pairs. For CACHECERT, no variant moves empirical risk by more than 0.004 from the full method's 0.083, and all variants stay below the $\alpha = 0.10$ target. What the ablations change is work saved. Removing query conditioning or the graded ladder costs 15 and 22 percentage points, respectively; uniform probes cost 8 points, and shrinking the probe set to $m = 8$ costs 13. This is the intended separation: informative scores and graded repair buy efficiency, while conformal calibration provides validity under the assumptions of Section 4.3. Doubling the probe set to $m = 64$ adds only one point of work saved at twice the probe cost, so we fix $m = 32$.

6 CONCLUSIONS AND FUTURE WORK

Cross-model KV cache transfer can fail silently, and its outcome varies across model pairs. Our first-order analysis of attention links downstream error to how cache residuals align with the receiver's query geometry. This analysis yields CTI, an offline retention predictor that requires no mapper fitting, and CACHECERT, an online procedure that controls the unsafe-transfer rate through conformal calibration under exchangeability. Future work includes extending the framework to latent and linear attention, handling tokenizer mismatch, developing conditional risk guarantees for workload subsets, and using CTI to plan model routing and cache reuse jointly.

AI USE STATEMENT

In this work, we used generative AI tools only to improve the readability of the manuscript. Specifically, an LLM-based assistant was used to copy-edit author-written drafts: tightening sentences, improving paragraph flow, correcting grammar and punctuation, and suggesting alternative phrasings for prose in the main text and appendices. Every suggestion was reviewed by the authors and was accepted, rejected, or rewritten by hand; no passage was incorporated without an author confirming that it preserved the intended technical meaning. We did not use generative AI tools for any task on the required-disclosure list. In particular, the conceptual framework of Section 4.1, the definitions of the Cache Transferability Index and of CacheCert, all mathematical claims and their proofs in Appendix A, the design of the experimental protocol, the implementation of all methods and baselines, the construction and processing of the calibration and evaluation data, and the interpretation of all results are entirely the work of the authors. We did not use generative AI to generate synthetic datasets, to propose or refine hypotheses, to design or give feedback on methodology or experiments, to clean or reformat data, or for translation; the manuscript was drafted in English by the authors. Qualitative and thematic data analysis is not applicable to this work. Among the recommended-disclosure categories, our use is limited to the readability editing described above. We did not use generative AI tools to identify or summarize literature, to discover research topics or gaps, to create or modify figures, to write or edit code, to format references, or to propose the title, keywords, or structure of the paper. To avoid ambiguity, we note that language models appear in this paper as objects of study and as experimental components rather than as research assistants. The eighteen directed pairs are the systems under evaluation; the agent-report track is produced by the specialist and coordinator models described in Section 5.1; and silent-failure labels come from the held-out autorater of Appendix D, whose agreement with a blind human audit we report there. These uses are part of the experimental protocol and are documented in the corresponding sections. We have reviewed all AI-assisted work and take responsibility for the final content of this submission, including all text, claims, and artifacts.

ETHICS STATEMENT

We have read and adhere to the ICLR Code of Ethics. The main ethical consideration in this work is that CacheCert's guarantee should not be mistaken for a safety or correctness guarantee. The loss in Eq. 12 compares a served answer against the receiver's own dense prefill, so what is certified is faithfulness to the model an operator has already chosen to deploy, and a pipeline whose dense behaviour is wrong or biased will have that behaviour faithfully preserved. Control is also marginal over the request distribution, so a minority slice such as an unusual document type or an under-represented language can be served well above the target rate while the average is met, and deployers who care about such slices should stratify calibration accordingly. Validity further rests on exchangeability, which an adversary can break by construction, so the method is a reliability mechanism rather than a security one. All experiments use publicly available benchmarks and open-weight models under their stated licenses. We collect no personal data and introduce no new data about individuals. The only human judgement in the paper is the blind audit of 400 autorater decisions in Appendix D, conducted by the authors on machine-generated text, with no external participants and no procedure requiring IRB review. The authors declare no conflicts of interest bearing on the results.

REPRODUCIBILITY STATEMENT

We have tried to make every claim in this paper checkable. All assumptions, approximations, and proofs behind the sensitivity operator, the closed form of Proposition 4, and the validity of the graded ladder are stated in full in Appendix A, including the remainder bound of Proposition 2 with the measured logit perturbations that determine when the linearization is trustworthy, the empirical validation of the factorization and cross-term approximations in Appendix A.5, and the measured rate at which rung monotonicity fails together with the monotonized remedy we actually use. Appendix D gives the complete experimental setup: serving stack and library versions, the cache injection hook, calibration corpora and sizes, ridge and probe hyperparameters, ladder proportions and conformal calibration settings, configurations for all four transfer backbones and all gating baselines, the autorater protocol with its human agreement rates, and the decoding and determinism settings, includ-

ing the pinned CUBLAS workspace configuration and the residual run to run variation we measure across nodes. Section 5.1 lists the eighteen directed pairs, the five workloads, and the metrics; per pair numbers for all of them appear in Appendix C, cost accounting in Table 7, and variants that did not work in Appendix E. All models and datasets we use are publicly available. Code for implementing our method is available at https://anonymous.4open.science/r/KBYT-6C1B. Every reported number is a mean over five random calibration and test splits with standard deviations taken across splits, and pairwise comparisons use a paired bootstrap with 10^4 resamples.

REFERENCES

Joshua Ainslie, James Lee-Thorp, Michiel De Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 4895–4901, 2023.

Anastasios Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International conference on learning representations*, volume 2024, pp. 55198–55218, 2024.

Anastasios N Angelopoulos, Stephen Bates, Emmanuel J Candès, Michael I Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, et al. Longbench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 3119–3137, 2024.

Lingjiao Chen, Matei Zaharia, and James Zou. Frugalgpt: How to use large language models while reducing cost and improving performance. *arXiv preprint arXiv:2305.05176*, 2023.

Lucio M Dery, Zohar Yahav, Henry Prior, Qixuan Feng, Jiajun Shen, and Arthur Szlam. Latent space communication via kv cache alignment. *arXiv preprint arXiv:2601.06123*, 2026.

Dujian Ding, Ankur Mallick, Chi Wang, Robert Sim, Subhabrata Mukherjee, Victor Rühle, Laks Lakshmanan, and Ahmed H Awadallah. Hybrid llm: Cost-efficient and quality-aware query routing. In *International Conference on Learning Representations*, volume 2024, pp. 41348–41366, 2024.

Tianyu Fu, Zihan Min, Hanling Zhang, Jichao Yan, Guohao Dai, Wanli Ouyang, and Yu Wang. Cache-to-cache: Direct semantic communication between large language models. In *International Conference on Learning Representations*, volume 2026, pp. 43130–43158, 2026.

In Gim, Guojun Chen, Seung-seob Lee, Nikhil Sarda, Anurag Khandelwal, and Lin Zhong. Prompt cache: Modular attention reuse for low-latency inference. *Proceedings of Machine Learning and Systems*, 6:325–338, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Taekyung Heo, Rasoul Shafipour, Ritchie Zhao, Maximilian Golub, Mohammad Mahdi Kamani, Ritika Borkar, Makesh Tarun Chandran, Pantea Zardoshti, and Bita Darvish Rouhani. Cross-model kv cache transfer in llm families: A closed-form linear mapping for prefill reuse. *arXiv preprint arXiv:2608.03893*, 2026.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, 2020.

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 *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Junhao Hu, Wenrui Huang, Weidong Wang, Haoyi Wang, Tiancheng Hu, Qin Zhang, Hao Feng, Xusheng Chen, Yizhou Shan, and Tao Xie. EPIC: efficient position-independent caching for serving large language models. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/hu25j.html.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20617–20642. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/huh24a.html.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019.

Satyapriya Krishna, Kalpesh Krishna, Anhad Mohananey, Steven Schwarcz, Adam Stambler, Shyam Upadhyay, and Manaal Faruqui. Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4745–4759, 2025.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pp. 611–626, 2023.

Jin-woo Lee, Minkyung Song, Junghyun Oh, Seunghoon Han, Soyoung Park, Gwangseon Jang, and Sungsu Lim. Mixture-of-translators: Translating kv caches across heterogeneous large language models. *arXiv preprint arXiv:2607.28979*, 2026.

Allison Li, Kristjan Greenewald, Thomas Parnell, and Navid Azizan. Efficient multi-adapter llm serving via cross-model kv-cache reuse with activated lora. *arXiv preprint arXiv:2512.17910*, 2025.

Sichu Liang, Zhenglin Wang, Pengfei Xia, Hui Zang, Deyu Zhou, et al. When kv cache reuse fails in multi-agent systems: Cross-candidate interaction is crucial for llm judges. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7205–7230, 2026.

Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, et al. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*, 2024a.

Yuhan Liu, Yuyang Huang, Jiayi Yao, Shaoting Feng, Zhuohan Gu, Kuntai Du, Hanchen Li, Yihua Cheng, Junchen Jiang, Shan Lu, et al. Droidspeak: Kv cache sharing for cross-llm communication and multi-llm serving. *arXiv preprint arXiv:2411.02820*, 2024b.

Yuhan Liu, Hanchen Li, Yihua Cheng, Siddhant Ray, Yuyang Huang, Qizheng Zhang, Kuntai Du, Jiayi Yao, Shan Lu, Ganesh Ananthanarayanan, et al. Cachegen: Kv cache compression and streaming for fast large language model serving. In *Proceedings of the ACM SIGCOMM 2024 Conference*, pp. 38–56, 2024c.

Songshuo Lu, Hua Wang, Yutian Rong, Zhi Chen, and Yaohua Tang. Turborag: Accelerating retrieval-augmented generation with precomputed kv caches for chunked text. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6599–6612, 2025.

Dongyang Ma, Yan Wang, and Tian Lan. Block-attention for efficient prefilling. In *International Conference on Learning Representations*, volume 2025, pp. 63774–63788, 2025.

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=SrC-nwieGJ.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E Gonzalez, Mohammed Kadous, and Ion Stoica. Routellm: Learning to route llms from preference data. In *International Conference on Learning Representations*, volume 2025, pp. 34433–34448, 2025.

Xingyu Qu, Siyuan Lu, Zhiyu Chen, Sheng Wang, and Tao Lin. Cachebridge: Efficient cross-model kv cache transfer. *arXiv preprint arXiv:2609.00891*, 2026.

Maximillian Rossi, Prajwal Raghunath, and Eugene Wu. Latent cache flow: Model-to-model communication without text. *arXiv preprint arXiv:2605.22863*, 2026.

Xi Shi and Qian Lou. Kvsharearena: Kv-cache reuse across contexts and model checkpoints. *arXiv preprint arXiv:2609.10266*, 2026.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Musique: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Y Zou. Mixture-of-agents enhances large language model capabilities. In *International Conference on Learning Representations*, volume 2025, pp. 33944–33963, 2025.

Sunghyeon Woo, Hoseung Kim, Sunghwan Shim, Minjung Jo, Hyunjoon Jeong, Jeongtae Lee, Joonghoon Kim, Sungjae Lee, Baeseong Park, Se Jung Kwon, et al. Prefillshare: A shared prefill module for kv reuse in multi-llm disaggregated serving. *arXiv preprint arXiv:2602.12029*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Jingbo Yang, Bairu Hou, Wei Wei, Yujia Bao, and Shiyu Chang. Kvlink: Accelerating large language models via efficient kv cache reuse. *Advances in Neural Information Processing Systems*, 38:133797–133824, 2026.

Xinyu Yang, Tianqi Chen, and Beidi Chen. Ape: Faster and longer context-augmented generation via adaptive parallel encoding. In *International Conference on Learning Representations*, volume 2025, pp. 82887–82907, 2025b.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2369–2380, 2018.

Jiayi Yao, Hanchen Li, Yuhan Liu, Siddhant Ray, Yihua Cheng, Qizheng Zhang, Kuntai Du, Shan Lu, and Junchen Jiang. Cacheblend: Fast large language model serving for rag with cached knowledge fusion. In *Proceedings of the twentieth European conference on computer systems*, pp. 94–109, 2025.

Hancheng Ye, Zhengqi Gao, Mingyuan Ma, Qinsi Wang, Yuzhe Fu, Ming-Yu Chung, Yueqian Lin, Zhijian Liu, Jianyi Zhang, Danyang Zhuo, and Yiran Chen. KVCOMM: online cross-context kv-cache communication for efficient llm-based multi-agent systems. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/1a074a28c3a6f2056562d00649ae6416-Abstract-Conference.html.

Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. HELMET: How to evaluate long-context language models effectively and thoroughly. In *International Conference on Learning Representations (ICLR)*, 2025.

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody H Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E Gonzalez, et al. Sglang: Efficient execution of structured language model programs. *Advances in neural information processing systems*, 37:62557–62583, 2024.

Jiaru Zou, Ruizhong Qiu, Gaotang Li, Xiyuan Yang, Katherine Tieu, Pan Lu, Ke Shen, Hanghang Tong, Yejin Choi, Jingrui He, James Zou, Mengdi Wang, and Ling Yang. Latent collaboration in multi-agent systems. In *International Conference on Machine Learning (ICML)*, 2026.

CONTENTS

1 Introduction **1**

2 Related Work **2**

3 Preliminaries **3**

4 Methodology **3**

4.1 Where transfer error hurts 3

4.2 The Cache Transferability Index 4

4.3 CacheCert: per-request certification 5

5 Experiments **6**

5.1 Experimental Setup 6

5.2 Does CTI predict transferability better and more cheaply? 7

5.3 Does CACHECERT certify at acceptable cost? 7

5.4 Which components of the analysis actually carry the weight? 9

6 Conclusions and Future Work **9**

A Proofs and Derivations **17**

A.1 Preliminary facts about the softmax Jacobian 17

A.2 Proposition 1: first-order deviation 17

A.3 Proposition 2: remainder bound and a checkable validity criterion 17

A.4 Proposition 3: the free subspace of key residuals 18

A.5 Derivation of the sensitivity operator 19

A.6 Proposition 4: closed-form irreducible residual 19

A.7 Proposition 6: validity of the graded ladder 20

B Sensitivity Spectra and Induced Layer Orderings **22**

C Per-Pair Results **23**

D Extended Implementation Details **23**

E Negative Results and Variants That Did Not Work **26**

A Proofs and Derivations

Throughout this appendix we fix a layer and key-value head and drop the indices (ℓ, h). Recall $\mathbf{K}, \mathbf{V} \in \mathbb{R}^{N\times d}$, a query $\mathbf{q} \in \mathbb{R}^d$, logits $\mathbf{s}(\mathbf{q}) = d^{-1/2}\mathbf{K}\mathbf{q}$, weights $\mathbf{a}(\mathbf{q}) = \mathrm{softmax}(\mathbf{s}(\mathbf{q}))$, and output $\mathbf{o}(\mathbf{q}) = \mathbf{V}^\top \mathbf{a}(\mathbf{q})$. We write $\widehat{\mathbf{K}} = \mathbf{K} + \mathbf{E}_K$, $\widehat{\mathbf{V}} = \mathbf{V} + \mathbf{E}_V$, and $\Delta\mathbf{s}(\mathbf{q}) = d^{-1/2}\mathbf{E}_K\mathbf{q}$. Vector norms are Euclidean, and the matrix norm $\|\cdot\|_2$ is the spectral norm unless subscripted otherwise.

A.1 Preliminary facts about the softmax Jacobian

Lemma 1 (Jacobian and its basic properties). *Let* $\mathbf{\Psi}(\mathbf{q}) = \mathrm{diag}(\mathbf{a}(\mathbf{q})) - \mathbf{a}(\mathbf{q})\mathbf{a}(\mathbf{q})^\top$. *Then*

(i) $\partial\,\mathrm{softmax}(\mathbf{s})/\partial\mathbf{s} = \mathbf{\Psi}$;

(ii) $\mathbf{\Psi} \succeq 0$ *and* $\mathbf{\Psi}\mathbf{1} = \mathbf{0}$;

(iii) $\|\mathbf{\Psi}\|_2 \le \min\{\frac{1}{2},\ 1 - \|\mathbf{a}\|_2^2\}$;

(iv) $\mathrm{tr}(\mathbf{\Psi}) = 1 - \|\mathbf{a}\|_2^2$.

Proof. (i) is the standard softmax derivative. For (ii), $\mathbf{\Psi}$ is the covariance matrix of the indicator vector of a categorical variable with law $\mathbf{a}$, hence positive semidefinite; and $\mathbf{\Psi}\mathbf{1} = \mathrm{diag}(\mathbf{a})\mathbf{1} - \mathbf{a}(\mathbf{a}^\top\mathbf{1}) = \mathbf{a} - \mathbf{a} = \mathbf{0}$ because $\mathbf{a}$ lies in the simplex. For (iv), $\mathrm{tr}(\mathbf{\Psi}) = \sum_i a_i - \sum_i a_i^2 = 1 - \|\mathbf{a}\|_2^2$. For (iii), positive semidefiniteness gives $\|\mathbf{\Psi}\|_2 \le \mathrm{tr}(\mathbf{\Psi}) = 1 - \|\mathbf{a}\|_2^2$. The bound $\|\mathbf{\Psi}\|_2 \le \frac{1}{2}$ follows from $\mathbf{z}^\top\mathbf{\Psi}\mathbf{z} = \mathrm{Var}_{i\sim\mathbf{a}}(z_i) \le \frac{1}{4}(\max_i z_i - \min_i z_i)^2 \le \frac{1}{2}\|\mathbf{z}\|^2$ by Popoviciu's inequality. □

The quantity $1 - \|\mathbf{a}\|_2^2$ in (iii) and (iv) is the Gini impurity of the attention distribution. It measures diffuseness: it is near 0 when attention is peaked on one position and equals $1 - 1/N$ when attention is uniform. It controls both the size of the first-order term and the size of the remainder, and this is what makes the linearization informative rather than vacuous.

A.2 Proposition 1: first-order deviation

Proposition 1 (First-order deviation). *With the notation above,*

$$\Delta\mathbf{o}(\mathbf{q}) \;=\; \underbrace{\tfrac{1}{\sqrt{d}}\mathbf{V}^\top\mathbf{\Psi}(\mathbf{q})\mathbf{E}_K\mathbf{q}}_{\textit{key path}} \;+\; \underbrace{\mathbf{E}_V^\top\mathbf{a}(\mathbf{q})}_{\textit{value path}} + R(\mathbf{q}), \qquad R(\mathbf{q}) = \mathbf{V}^\top\boldsymbol{\eta}(\mathbf{q}) + \mathbf{E}_V^\top\Delta\mathbf{a}(\mathbf{q}),$$

where $\Delta\mathbf{a}(\mathbf{q}) = \mathrm{softmax}(\mathbf{s} + \Delta\mathbf{s}) - \mathrm{softmax}(\mathbf{s})$ *and* $\boldsymbol{\eta}(\mathbf{q}) = \Delta\mathbf{a}(\mathbf{q}) - \mathbf{\Psi}(\mathbf{q})\Delta\mathbf{s}(\mathbf{q})$.

Proof. Expand the perturbed output exactly:

$$\widehat{\mathbf{o}}(\mathbf{q}) = (\mathbf{V} + \mathbf{E}_V)^\top\big(\mathbf{a} + \Delta\mathbf{a}\big) = \mathbf{V}^\top\mathbf{a} + \mathbf{V}^\top\Delta\mathbf{a} + \mathbf{E}_V^\top\mathbf{a} + \mathbf{E}_V^\top\Delta\mathbf{a}.$$

Subtract $\mathbf{o} = \mathbf{V}^\top\mathbf{a}$ and substitute $\Delta\mathbf{a} = \mathbf{\Psi}\Delta\mathbf{s} + \boldsymbol{\eta}$ with $\Delta\mathbf{s} = d^{-1/2}\mathbf{E}_K\mathbf{q}$. Collecting terms gives the claim. □

The main text uses two structural consequences repeatedly. The key path reaches $\Delta\mathbf{o}$ only through the projections $\mathbf{E}_K\mathbf{q}$, so residual energy orthogonal to the span of the query distribution is invisible at first order, whatever its magnitude. The value path is linear in $\mathbf{E}_V$ and depends on $\mathbf{q}$ only through the attention weights. This is why Section 4.1 calls it benign and why the channel metric for values is isotropic.

A.3 Proposition 2: remainder bound and a checkable validity criterion

Proposition 2 (Remainder bound). *Let* $\phi(t) = \mathrm{softmax}(\mathbf{s} + t\Delta\mathbf{s})$, *let* $\mathbf{\Psi}_t = \mathrm{diag}(\phi(t)) - \phi(t)\phi(t)^\top$, *and let* $\bar{\psi} = \sup_{t\in[0,1]}\|\mathbf{\Psi}_t\|_2 \le \frac{1}{2}$, *where the bound is Lemma 1(iii). Then*

$$\|\boldsymbol{\eta}(\mathbf{q})\|_2 \;\le\; \tfrac{3}{2}\,\bar{\psi}\,\|\Delta\mathbf{s}(\mathbf{q})\|_2^2, \qquad \textit{hence} \qquad \|R(\mathbf{q})\|_2 \;\le\; \tfrac{3}{2}\bar{\psi}\|\mathbf{V}\|_2\,\|\Delta\mathbf{s}(\mathbf{q})\|_2^2 + \bar{\psi}\,\|\mathbf{E}_V\|_2\,\|\Delta\mathbf{s}(\mathbf{q})\|_2.$$

Proof. With ϕ and $\mathbf{\Psi}_t$ as in the statement, $\phi'(t) = \mathbf{\Psi}_t \Delta\mathbf{s}$, $\phi(0) = \mathbf{a}$, and $\mathbf{\Psi}_0 = \mathbf{\Psi}$. By the fundamental theorem of calculus, $\boldsymbol{\eta} = \int_0^1 \big(\mathbf{\Psi}_t - \mathbf{\Psi}_0\big)\Delta\mathbf{s}\, dt$. For the integrand, using $\|\phi(t)\|_2 \le 1$, $\|\mathbf{\Psi}_t - \mathbf{\Psi}_0\|_2 \le \|\mathrm{diag}(\phi(t) - \mathbf{a})\|_2 + \|\phi(t)\phi(t)^\top - \mathbf{a}\mathbf{a}^\top\|_2 \le \|\phi(t) - \mathbf{a}\|_\infty + 2\|\phi(t) - \mathbf{a}\|_2 \le 3\|\phi(t) - \mathbf{a}\|_2$, and $\|\phi(t) - \mathbf{a}\|_2 \le \int_0^t \|\mathbf{\Psi}_\tau\|_2 \|\Delta\mathbf{s}\|_2\, d\tau \le \bar{\psi}\, t\, \|\Delta\mathbf{s}\|_2$. Therefore $\|\boldsymbol{\eta}\|_2 \le \int_0^1 3\bar{\psi} t \|\Delta\mathbf{s}\|_2^2\, dt = \frac{3}{2}\bar{\psi}\|\Delta\mathbf{s}\|_2^2$. The bound on R follows from $\|\mathbf{V}^\top\boldsymbol{\eta}\| \le \|\mathbf{V}\|_2\|\boldsymbol{\eta}\|$ and $\|\mathbf{E}_V^\top \Delta\mathbf{a}\| \le \|\mathbf{E}_V\|_2\|\Delta\mathbf{a}\| \le \|\mathbf{E}_V\|_2\bar{\psi}\|\Delta\mathbf{s}\|_2$. □

Corollary 1 (When the linearization is trustworthy)**.** *Let $\mathbf{P}(\mathbf{q})$ be the orthogonal projector onto the row space of $\mathbf{V}^\top\mathbf{\Psi}(\mathbf{q})$, and let $\varsigma(\mathbf{q}) = \bar{\psi}\,\|\mathbf{V}\|_2 / \sigma^+_{\min}\big(\mathbf{V}^\top\mathbf{\Psi}(\mathbf{q})\big)$ be the conditioning of the key path on the subspace it acts on. Ignoring the value cross-term,*

$$\frac{\|\mathbf{V}^\top\boldsymbol{\eta}(\mathbf{q})\|_2}{\big\|\frac{1}{\sqrt{d}}\mathbf{V}^\top\mathbf{\Psi}(\mathbf{q})\mathbf{E}_K\mathbf{q}\big\|_2} \;\le\; \tfrac{3}{2}\varsigma(\mathbf{q})\,\frac{\|\Delta\mathbf{s}(\mathbf{q})\|_2^2}{\|\mathbf{P}(\mathbf{q})\Delta\mathbf{s}(\mathbf{q})\|_2},$$

which follows from Proposition 2 and $\|\mathbf{V}^\top\mathbf{\Psi}\Delta\mathbf{s}\|_2 \ge \sigma^+_{\min}(\mathbf{V}^\top\mathbf{\Psi})\,\|\mathbf{P}\Delta\mathbf{s}\|_2$. If $\|\mathbf{P}(\mathbf{q})\Delta\mathbf{s}(\mathbf{q})\|_2 \ge c\,\|\Delta\mathbf{s}(\mathbf{q})\|_2$ for some $c \in (0, 1]$, the ratio is at most $\frac{3}{2c}\varsigma(\mathbf{q})\|\Delta\mathbf{s}(\mathbf{q})\|_2$. For a well-conditioned key path, the linearization is therefore controlled by a single measurable scalar, namely the size of the induced perturbation of the attention logits *in nats.*

This gives a diagnostic we can report rather than assume. Let $\delta = \mathrm{median}_{\mathbf{q},\ell,h}\,\|\Delta\mathbf{s}(\mathbf{q})\|_2$ be the median logit perturbation that the transfer map induces on calibration data. Across the twelve within-family pairs we measure $\delta = 0.081$ nats (interquartile range 0.049 to 0.126), so for moderate ς the first-order term dominates by roughly an order of magnitude. Across the six cross-family pairs we measure $\delta = 0.313$ (interquartile range 0.187 to 0.492), and the linearization is materially looser. We state this limitation plainly. It is also self-correcting in one sense: a large δ comes with a large first-order residual, so the pairs on which the analysis is least accurate tend to be those that CTI already scores low and recommends against deploying. Corollary 1 licenses using ρ on those pairs as an ordering statistic rather than as a quantitative prediction. It is also why the online guarantee in Appendix A.7 is derived without any reference to the linearization.

A.4 Proposition 3: the free subspace of key residuals

Proposition 3 (Annihilation of uniform bias)**.** *If $\mathbf{E}_K = \mathbf{1}\mathbf{u}^\top$ for some $\mathbf{u} \in \mathbb{R}^d$, then the first-order attention perturbation $\mathbf{\Psi}(\mathbf{q})\mathbf{E}_K\mathbf{q}$, and hence the key path of Proposition 1, vanishes for every query. Conversely, if the queries span $\mathbb{R}^d$, the key residuals whose first-order attention perturbation vanishes for every query are exactly $\{\mathbf{1}\mathbf{u}^\top : \mathbf{u} \in \mathbb{R}^d\}$, a subspace of dimension d inside $\mathbb{R}^{N\times d}$.*

Proof. For the forward direction, $\mathbf{E}_K\mathbf{q} = \mathbf{1}\mathbf{u}^\top\mathbf{q} = (\mathbf{u}^\top\mathbf{q})\mathbf{1}$, so $\mathbf{\Psi}(\mathbf{q})\mathbf{E}_K\mathbf{q} = (\mathbf{u}^\top\mathbf{q})\mathbf{\Psi}(\mathbf{q})\mathbf{1} = \mathbf{0}$ by Lemma 1(ii). For the converse, suppose $\mathbf{\Psi}(\mathbf{q})\mathbf{E}_K\mathbf{q} = \mathbf{0}$ for all $\mathbf{q}$ in a spanning set. Softmax weights are strictly positive, so $\mathbf{\Psi}(\mathbf{q})$ is the covariance matrix of a distribution with full support, and its kernel is exactly $\mathrm{span}(\mathbf{1})$. Hence $\mathbf{E}_K\mathbf{q} \in \mathrm{span}(\mathbf{1})$ for each such $\mathbf{q}$. Write $\mathbf{E}_K\mathbf{q}_j = c_j\mathbf{1}$ for a basis $\{\mathbf{q}_j\}$ and choose $\mathbf{u}$ with $\mathbf{u}^\top\mathbf{q}_j = c_j$. Then $(\mathbf{E}_K - \mathbf{1}\mathbf{u}^\top)\mathbf{q}_j = \mathbf{0}$ for every j, so $\mathbf{E}_K = \mathbf{1}\mathbf{u}^\top$. At the level of the output, the key path can vanish on a larger set when $N > d + 1$, since $\mathbf{V}^\top$ then has a nontrivial kernel on the range of $\mathbf{\Psi}(\mathbf{q})$; the attention-level statement holds for every value cache. □

Corollary 2 (Unidentifiability of the key intercept)**.** *In the affine transfer family, the key-side intercept is unidentifiable at first order: the maps $\mathbf{x} \mapsto \mathbf{\Theta}\mathbf{x} + \mathbf{b}$ and $\mathbf{x} \mapsto \mathbf{\Theta}\mathbf{x} + \mathbf{b} + \mathbf{u}$ induce identical attention outputs to first order for every $\mathbf{u}$.*

Corollary 2 has two practical consequences. First, the intercept should be fixed by centering rather than fitted. This removes d parameters per head from the transfer map at no cost in quality; Appendix E verifies that fitting it changes downstream PGR by only 0.004 on average. Second, and more importantly, coordinate-wise reconstruction metrics charge for error in this free subspace. A mapper that is systematically offset in the keys is penalized by R^2 and mean squared error, yet it is equally good downstream. This is one concrete, easily checked mechanism by which those statistics mis-rank pairs. On our data, projecting out the free subspace before computing R^2 raises its Spearman correlation with retention from 0.34 to 0.47. This recovers part, but by no means all, of the gap to CTI.

A.5 Derivation of the sensitivity operator

We now derive Eqs. (5) and (6). Squaring the first-order expression of Proposition 1 and taking expectations over $\mathbf{q} \sim \mathcal{Q}$ gives

$$\mathbb{E}_{\mathbf{q}}\big\|\Delta\mathbf{o}^{(1)}(\mathbf{q})\big\|^2 = \underbrace{\tfrac{1}{d}\,\mathbb{E}_{\mathbf{q}}\,\mathrm{tr}\big(\mathbf{E}_K^\top \mathbf{M}(\mathbf{q})\mathbf{E}_K\,\mathbf{q}\mathbf{q}^\top\big)}_{\text{key}} + \underbrace{\mathbb{E}_{\mathbf{q}}\,\mathrm{tr}\big(\mathbf{E}_V^\top \mathbf{a}\mathbf{a}^\top \mathbf{E}_V\big)}_{\text{value}} + 2\,\mathcal{X}, \tag{13}$$

with $\mathbf{M}(\mathbf{q}) = \mathbf{\Psi}(\mathbf{q})\mathbf{V}\mathbf{V}^\top\mathbf{\Psi}(\mathbf{q}) \in \mathbb{R}^{N\times N}$ and cross-term $\mathcal{X} = d^{-1/2}\mathbb{E}_{\mathbf{q}}\langle \mathbf{V}^\top\mathbf{\Psi}\mathbf{E}_K\mathbf{q},\ \mathbf{E}_V^\top\mathbf{a}\rangle$. Eq. (13) is exact, but it couples positions through $\mathbf{M}$. This would make the quantity a quadratic form on $\mathbb{R}^{Nd}$ and destroy the closed form of Section 4.2. We therefore make two approximations and test both empirically.

(A1) Position-channel factorization. Replace $\mathbf{M}(\mathbf{q})$ by its diagonal and factor the position and query dependence as $\big(\mathbf{M}(\mathbf{q})\big)_{ii} \approx \bar{m}_i\,\kappa(\mathbf{q})/\mathbb{E}_{\mathbf{q}}[\kappa(\mathbf{q})]$, with $\kappa(\mathbf{q}) = \mathrm{tr}\,\mathbf{M}(\mathbf{q}) = \|\mathbf{V}^\top\mathbf{\Psi}(\mathbf{q})\|_F^2$ and $\bar{m}_i = \mathbb{E}_{\mathbf{q}}[(\mathbf{M}(\mathbf{q}))_{ii}]$; write $\bar{m} = (\bar{m}_i)_{i=1}^N$. Since $\sum_i \bar{m}_i = \mathbb{E}_{\mathbf{q}}[\kappa(\mathbf{q})]$, the factorization preserves $\mathrm{tr}\,\mathbf{M}(\mathbf{q})$ for every query. Substituting into Eq. (13) gives exactly the key term in the numerator of Eq. (7), with $\mathbf{G}$ as in Eq. (5) and $w_i = \bar{m}_i$. The same diagonal approximation applied to $\mathbf{a}\mathbf{a}^\top$ gives $u_i = \mathbb{E}_{\mathbf{q}}[a_i(\mathbf{q})^2]$ and an isotropic channel metric for values. One property is lost in the replacement, and we flag it because Proposition 3 rests on it. The exact operator inherits $\mathbf{M}(\mathbf{q})\mathbf{1} = \mathbf{0}$ from $\mathbf{\Psi}(\mathbf{q})\mathbf{1} = \mathbf{0}$, so a key residual $\mathbf{E}_K = \mathbf{1}\mathbf{u}^\top$ contributes nothing to Eq. (13). The surrogate $\mathrm{diag}(\bar{m})$ has no kernel and charges $\big(\sum_i \bar{m}_i\big)\,\mathbf{u}^\top\mathbf{G}\mathbf{u} > 0$ for the same residual. This does not affect CTI. The residual entering Definition 1 is $\mathbf{y}_i - \mathbf{\Theta}^\star\mathbf{x}_i - \mathbf{b}^\star$, whose w-weighted mean vanishes by the choice of $\mathbf{b}^\star$ in Proposition 4, so $\mathbf{\Sigma}_{y|x}$ has no component along the free subspace. It does affect the online score of Eq. (11), where $\widetilde{\mathbf{E}}_{\mathcal{P}}$ is a raw fixed-point residual: a transfer map with a systematic key offset would be scored as defective, although Proposition 3 says it is not. We therefore remove the free subspace from the key path before evaluating Eq. (7). We replace $\mathbf{E}_K$ by $\mathbf{E}_K - \mathbf{1}\bar{\mathbf{E}}_K^\top$ with $\bar{\mathbf{E}}_K = \big(\sum_i w_i\big)^{-1}\sum_i w_i\,\mathbf{E}_{K,i}$, and center $\mathbf{K}$ likewise. This is the same w-weighted centering that Proposition 4 already performs, and it is exactly the projection that Corollary 2 identifies as free. It therefore restores the exact geometry rather than adding an assumption. It is also an efficiency correction rather than a correctness one, since an inflated score only buys conservatism, not invalidity.

(A2) Negligible cross-term. Drop $\mathcal{X}$. This is exact whenever the key-path and value-path deviations are orthogonal on average under $\mathcal{Q}$.

Empirical validation. We test both approximations directly by computing the exact right-hand side of Eq. (13) by Monte Carlo over 2048 calibration queries per head. This is tractable offline, though far too expensive for a screening statistic. Over all 22 configurations, the factorized surrogate and the exact quantity give per-head values correlated at Pearson 0.981. More to the point, the CTI scores they induce rank the pairs almost identically (Spearman 0.999, with one adjacent transposition among the twelve within-family pairs). The dropped cross-term $|\mathcal{X}|$ accounts for a median of 3.1% of the total on within-family pairs and 7.8% on cross-family pairs. Its sign varies across heads rather than being systematic, which is why it does not bias the ordering. We regard (A1) and (A2) as engineering choices that buy the closed form at negligible cost in fidelity.

A.6 Proposition 4: closed-form irreducible residual

Proposition 4 (Weighted least squares residual)**.** *Fix a head and let $(\mathbf{x}_i, \mathbf{y}_i)_{i=1}^N$ with $\mathbf{x}_i \in \mathbb{R}^{d_s}$, $\mathbf{y}_i \in \mathbb{R}^d$ be the source features and target rows, let $w_i > 0$ be the position weights of Eq. (6), and let $\mathbf{G} \succ 0$ be the channel metric. Write $\bar{\mathbf{x}}, \bar{\mathbf{y}}$ for the w-weighted means and $\mathbf{\Sigma}_{xx} = \sum_i w_i\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^\top$, $\mathbf{\Sigma}_{yx} = \sum_i w_i\tilde{\mathbf{y}}_i\tilde{\mathbf{x}}_i^\top$, $\mathbf{\Sigma}_{yy} = \sum_i w_i\tilde{\mathbf{y}}_i\tilde{\mathbf{y}}_i^\top$ for the centered second moments. If $\mathbf{\Sigma}_{xx} \succ 0$, then*

$$\min_{\mathbf{\Theta},\mathbf{b}} \sum_i w_i\big(\mathbf{y}_i - \mathbf{\Theta}\mathbf{x}_i - \mathbf{b}\big)^\top\mathbf{G}\big(\mathbf{y}_i - \mathbf{\Theta}\mathbf{x}_i - \mathbf{b}\big) = \mathrm{tr}\big(\mathbf{G}\,\mathbf{\Sigma}_{y|x}\big), \qquad \mathbf{\Sigma}_{y|x} = \mathbf{\Sigma}_{yy} - \mathbf{\Sigma}_{yx}\mathbf{\Sigma}_{xx}^{-1}\mathbf{\Sigma}_{xy},$$

and the minimizer is $\mathbf{\Theta}^\star = \mathbf{\Sigma}_{yx}\mathbf{\Sigma}_{xx}^{-1}$, $\mathbf{b}^\star = \bar{\mathbf{y}} - \mathbf{\Theta}^\star\bar{\mathbf{x}}$.

Proof. For fixed $\boldsymbol{\Theta}$, the objective is a weighted quadratic in $\mathbf{b}$ minimized at $\mathbf{b} = \bar{\mathbf{y}} - \boldsymbol{\Theta}\bar{\mathbf{x}}$, which reduces the problem to the centered variables. Expanding,

$$f(\boldsymbol{\Theta}) = \operatorname{tr}\big(\mathbf{G}\boldsymbol{\Sigma}_{yy}\big) - 2\operatorname{tr}\big(\mathbf{G}\boldsymbol{\Theta}\boldsymbol{\Sigma}_{xy}\big) + \operatorname{tr}\big(\mathbf{G}\boldsymbol{\Theta}\boldsymbol{\Sigma}_{xx}\boldsymbol{\Theta}^{\top}\big).$$

Differentiating gives $\nabla_{\boldsymbol{\Theta}} f = -2\mathbf{G}\boldsymbol{\Sigma}_{yx} + 2\mathbf{G}\boldsymbol{\Theta}\boldsymbol{\Sigma}_{xx}$. Since $\mathbf{G} \succ 0$ is invertible, the stationarity condition reduces to $\boldsymbol{\Theta}\boldsymbol{\Sigma}_{xx} = \boldsymbol{\Sigma}_{yx}$, giving $\boldsymbol{\Theta}^{\star} = \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}$. The objective is convex in $\boldsymbol{\Theta}$ because $\boldsymbol{\Sigma}_{xx} \succeq 0$ and $\mathbf{G} \succeq 0$, so this is the global minimum. Substituting yields $\operatorname{tr}(\mathbf{G}\boldsymbol{\Sigma}_{yy}) - \operatorname{tr}(\mathbf{G}\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}) = \operatorname{tr}(\mathbf{G}\boldsymbol{\Sigma}_{y|x})$. □

Corollary 3 (The metric changes the score, not the map)**.** *$\boldsymbol{\Theta}^{\star}$ does not depend on $\mathbf{G}$. Sensitivity weighting therefore changes which pairs look transferable without changing which affine map is optimal for a given pair.*

Corollary 3 answers a natural objection: why not simply train the mapper under the $\mathbf{G}$-weighted loss instead of measuring with it? Within the affine family this would recover the same map, so the value of the sensitivity operator there is entirely diagnostic. The statement does not extend to nonlinear families, where $\mathbf{G}$-weighted training is a real and, in our view, promising option that we did not pursue here.

Regularized variant. In practice, $\boldsymbol{\Sigma}_{xx}$ is ill-conditioned when the source features concatenate several source layers. We therefore use the ridge estimator $\boldsymbol{\Theta}_{\lambda} = \boldsymbol{\Sigma}_{yx}(\boldsymbol{\Sigma}_{xx} + \lambda\mathbf{I})^{-1}$, whose attained residual is

$$\operatorname{tr}\big(\mathbf{G}\boldsymbol{\Sigma}_{yy}\big) - 2\operatorname{tr}\big(\mathbf{G}\boldsymbol{\Sigma}_{yx}(\boldsymbol{\Sigma}_{xx} + \lambda\mathbf{I})^{-1}\boldsymbol{\Sigma}_{xy}\big) + \operatorname{tr}\big(\mathbf{G}\boldsymbol{\Sigma}_{yx}(\boldsymbol{\Sigma}_{xx} + \lambda\mathbf{I})^{-1}\boldsymbol{\Sigma}_{xx}(\boldsymbol{\Sigma}_{xx} + \lambda\mathbf{I})^{-1}\boldsymbol{\Sigma}_{xy}\big).$$

This reduces to Proposition 4 as $\lambda \to 0$. We select λ per head by generalized cross-validation on the calibration corpus; Appendix D reports the selected values and their effect on CTI. Here λ denotes the ridge parameter only; the certification threshold of Appendix A.7 is a different quantity.

Proposition 5 (Finite-sample estimation of the index)**.** *Suppose the paired rows $(\mathbf{x}_i, \mathbf{y}_i)$ are drawn i.i.d. from a distribution whose whitened marginals are σ-sub-Gaussian, and suppose $\lambda_{\min}(\boldsymbol{\Sigma}_{xx}) \geq \mu > 0$. Let $\widehat{T}$ denote the plug-in estimate of $T = \operatorname{tr}(\mathbf{G}\boldsymbol{\Sigma}_{y|x})$ from n rows. There is an absolute constant c such that, for $n \gtrsim d + d_s$, with probability at least $1 - \delta$,*

$$\big|\widehat{T} - T\big| \;\leq\; c\,\|\mathbf{G}\|_{*}\,\sigma^2\Big(1 + \tfrac{\|\boldsymbol{\Sigma}_{yx}\|_2^2}{\mu^2}\Big)\sqrt{\tfrac{d + d_s + \log(1/\delta)}{n}}.$$

Proof sketch. Standard sub-Gaussian covariance concentration gives, on the stated event,

$$\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_2 \;\lesssim\; \sigma^2\sqrt{(d + d_s + \log(1/\delta))/n}$$

for the joint second-moment matrix. The Schur complement is a smooth function of the blocks wherever $\boldsymbol{\Sigma}_{xx} \succ 0$. By the resolvent identity $\mathbf{A}^{-1} - \mathbf{B}^{-1} = \mathbf{A}^{-1}(\mathbf{B} - \mathbf{A})\mathbf{B}^{-1}$ applied to $\boldsymbol{\Sigma}_{xx}$, its derivative is bounded in operator norm by $(1 + \|\boldsymbol{\Sigma}_{yx}\|_2/\mu)^2 \leq 2(1 + \|\boldsymbol{\Sigma}_{yx}\|_2^2/\mu^2)$. Composing with $|\operatorname{tr}(\mathbf{G}\boldsymbol{\Delta})| \leq \|\mathbf{G}\|_{*}\|\boldsymbol{\Delta}\|_2$ gives the claim. The restriction $n \gtrsim d + d_s$ ensures that $\widehat{\boldsymbol{\Sigma}}_{xx}$ inherits the lower eigenvalue bound up to a constant. □

Proposition 5 explains why 256 calibration sequences suffice. With $d = 128$, $d_s = 1024$ for the key and value rows of a 128-dimensional source head over four source layers, and $N = 1024$ tokens per sequence, there are about 2.6×10^5 rows per head. Treating these rows as independent overstates the effective sample size, because tokens within a sequence are correlated, so we check the conclusion empirically. The calibration-size sweep in Appendix D shows that CTI is stable to within 0.009 from 128 sequences upward and within 0.006 from 256, about two orders of magnitude below the between-pair spread of CTI.

A.7 PROPOSITION 6: VALIDITY OF THE GRADED LADDER

We now show that the construction of Section 4.3 satisfies the hypotheses of conformal risk control, and what to do when it does not. Let $s_r \in \mathbb{R}$ be the risk score of request r, with larger values more suspicious. Fix rung proportions $0 < p_0 < p_1 < p_2 \leq 1$. For $\lambda \in [0, 1]$, set $\tau_j(\lambda) = Q_s\big((1-\lambda)p_j\big)$, where $Q_s(p) = \inf\{s : \widehat{F}_s(s) \geq p\}$ is the left-continuous empirical quantile function of the score on

calibration data and $\widehat{F}_s$ is the empirical distribution function. This definition gives $Q_s(p) = -\infty$ for $p \le 0$, so $\tau_j(1) = -\infty$ for every j. Because Q_s is non-decreasing and $p_0 < p_1 < p_2$, the thresholds satisfy $\tau_0(\lambda) \le \tau_1(\lambda) \le \tau_2(\lambda)$ at every λ, and each of them is non-increasing in λ. Assign

$$\mathrm{rung}(s;\lambda) = \begin{cases} 0 & s \le \tau_0(\lambda), \\ 1 & \tau_0(\lambda) < s \le \tau_1(\lambda), \\ 2 & \tau_1(\lambda) < s \le \tau_2(\lambda), \\ 3 & s > \tau_2(\lambda), \end{cases}$$

where rung 0 accepts the transferred cache, rungs 1 and 2 recompute $k_1 < k_2$ layers chosen in decreasing order of $\gamma_\ell \rho^\ell$, and rung 3 is a full dense prefill.

Proposition 6 (Monotonicity and validity)**.** *Assume* (A3) rung monotonicity*: for each request, the loss $\ell_r(\cdot)$ is non-increasing in the rung index. Then*

(i) $\lambda \mapsto \mathrm{rung}(s;\lambda)$ *is non-decreasing for every s;*

(ii) $L_r(\lambda) = \ell_r(\mathrm{rung}(s_r;\lambda))$ *is non-increasing and right-continuous in λ, and bounded by $B = 1$;*

(iii) $L_r(1) = 0$ *for every r, so the set in Eq. (3) is non-empty;*

(iv) consequently, $\widehat{\lambda}$ from Eq. (3) satisfies $\mathbb{E}[L_{n+1}(\widehat{\lambda})] \le \alpha$ for an exchangeable test request.

Proof. (i) For $\lambda_1 \le \lambda_2$ we have $(1-\lambda_1)p_j \ge (1-\lambda_2)p_j$, and Q_s is non-decreasing, so $\tau_j(\lambda_1) \ge \tau_j(\lambda_2)$ for each j. A score in rung $j \ge 1$ under λ_1 satisfies $s > \tau_{j-1}(\lambda_1) \ge \tau_{j-1}(\lambda_2)$, and by nesting it exceeds every lower threshold under λ_2 as well, so its rung under λ_2 is at least j. (ii) Monotonicity follows from (i) and (A3). Right-continuity does not use (A3): as $\lambda' \downarrow \lambda$, the argument $(1-\lambda')p_j$ increases to $(1-\lambda)p_j$, and because Q_s is a left-continuous step function, $\tau_j(\lambda') = \tau_j(\lambda)$ for all λ' close enough to λ from above. Hence L_r is a right-continuous step function of λ with finitely many jumps. Boundedness is immediate for a $\{0,1\}$-valued loss. (iii) At $\lambda = 1$ we have $(1-\lambda)p_j = 0$, so $\tau_j(1) = -\infty$ for every j and every request is assigned rung 3. Since rung 3 reproduces the dense prefill by construction, $L_r(1) = 0$ identically. Under a convention that instead set $Q_s(0)$ to the sample minimum, all thresholds at $\lambda = 1$ would equal that minimum. The score-minimizing calibration request would then remain at rung 0, $L_r(1) = 0$ could fail for it, and the set in Eq. (3) could be empty. (iv) This is Theorem 1 of Angelopoulos et al. (2024) applied to L. □

When rung monotonicity fails. (A3) is an assumption about the repair mechanism, not about our scoring, and it can fail: on a particular request, recomputing more layers can move the answer away from the dense-prefill answer rather than toward it. We measure this directly. Over 5 400 calibration requests, (A3) is violated on 2.3% of request-rung pairs. Violations concentrate on the agent-report track (4.1%), where the coordinator's answer is more sensitive to small changes in report framing. We therefore do not rely on (A3) for our reported numbers and offer two remedies instead.

Corollary 4 (Monotonization)**.** *Define $\widetilde{L}_r(\lambda) = \sup_{\lambda' \ge \lambda} L_r(\lambda')$, the least non-increasing majorant of L_r on $[0,1]$. Then $\widetilde{L}_r$ is non-increasing and dominates L_r pointwise by construction. It is right-continuous because L_r is a right-continuous step function with finitely many jumps, which is the part of Proposition 6(ii) that does not invoke (A3). Moreover, $\widetilde{L}_r(1) = L_r(1) = 0$ by Proposition 6(iii), so the set in Eq. (3) remains non-empty. Applying Eq. (3) to $\widetilde{L}$ therefore yields $\mathbb{E}[L_{n+1}(\widehat{\lambda})] \le \mathbb{E}[\widetilde{L}_{n+1}(\widehat{\lambda})] \le \alpha$ without assuming (A3).*

The supremum in Corollary 4 runs over $\lambda' \ge \lambda$ rather than $\lambda' \le \lambda$. Eq. (3) needs a loss that is non-increasing in λ and vanishes at $\lambda = 1$, and only the majorant taken from above has both properties. A running supremum over smaller λ would be non-decreasing, with $\widetilde{L}_r(1) = \sup_\lambda L_r(\lambda)$, and would empty the feasible set whenever any request ever fails. The price of Corollary 4 is conservatism, which we quantify: monotonization moves the selected $\widehat{\lambda}$ by a median of 0.017 and costs 1.4 points of work saved at $\alpha = 0.10$. The second remedy abandons monotonicity entirely and uses Learn-then-Test (Angelopoulos et al., 2025). It tests a grid of λ values with a fixed-sequence correction and yields the high-probability statement $\Pr\{\mathbb{E}[L(\widehat{\lambda}) \mid \mathcal{D}] \le \alpha\} \ge 1 - \delta$ instead of control in

expectation. All main-text numbers use monotonization. The Learn-then-Test variant at $\delta = 0.1$ selects thresholds within one grid step of ours and changes work saved by 0.004. We report the monotonized version because it keeps a bound in expectation, matching Eq. (12), without assuming (A3).

What the guarantee does and does not say. Three clarifications have proved necessary in discussion. First, the loss compares against the receiver's own dense prefill, not against ground truth. The guarantee therefore concerns *faithfulness to the model the operator chose to deploy*, not correctness; a pipeline whose dense behaviour is wrong will have that behaviour faithfully preserved. Second, control is marginal over the request distribution, so a rare slice can be served worse than α while the average is met. Conditional versions are future work (Section 6), and the workload-shift experiment of Figure 3(b) is the empirical proxy we offer in the meantime. Third, exchangeability is required between calibration and test requests. Adversarial traffic and drifting production distributions violate it, which is why we recommend periodic recalibration; Appendix D reports its cost.

B Sensitivity Spectra and Induced Layer Orderings

Effective rank. For each head, we summarize the channel operator $\mathbf{G}^{\ell h}$ by its entropy-based effective rank, $\mathrm{erank}(\mathbf{G}) = \exp\big(-\sum_j \bar{\lambda}_j \log \bar{\lambda}_j\big)$ with $\bar{\lambda}_j = \lambda_j / \sum_{j'} \lambda_{j'}$. A head with $\mathrm{erank} \approx 1$ is forgiving, since residual energy must land in essentially one direction to matter; $\mathrm{erank} \approx d$ means that every direction is exposed. Table 4 reports this quantity by depth for the three largest receivers.

Table 4: Effective rank of the key channel operator by depth third, median and interquartile range over heads. The per-head dimension is 128 for all three receivers. The middle third is consistently the most brittle: transfer error costs most there, and the CTI layer profile correspondingly allocates repair budget there first.

Receiver	early third	middle third	late third
Qwen3-32B	9.4 [6.1, 14.8]	41.7 [28.3, 55.2]	18.2 [11.9, 27.4]
Llama-3.1-70B	11.8 [7.3, 17.2]	38.4 [25.6, 51.9]	21.5 [14.2, 30.1]
Gemma-3-27B	8.7 [5.4, 13.1]	44.9 [31.2, 58.6]	16.3 [10.4, 24.8]

Agreement with hand-tuned layer selection. The layer profile $\gamma_\ell \rho^\ell$ of Section 4.2 induces an ordering over receiver layers, and CacheCert uses this ordering for its ladder. We compare it with the layer selections that prior work obtains by search. Against the sensitive layers identified by Liu et al. (2024b), the overlap at $k = 4$ is 3.4 layers on average across the twelve within-family pairs (Spearman 0.78 between the two orderings). Against the top-k source-layer selection of Heo et al. (2026), the overlap at $k = 4$ is 3.1. The profile thus recovers most of what a search finds, at the cost of one propagation probe per receiver instead of one search per pair.

Where the damage sits. Figure 4 plots the profile for all 22 configurations. Panel (a) makes the structure of Table 4 directly visible. Wherever transfer succeeds, damage concentrates in a narrow band slightly past mid-depth, where the channel operator has the highest effective rank and exposes the most residual directions. Panel (b) converts this into the quantity the ladder consumes: the share of total damage carried by the k most affected layers. Averaged over within-family pairs, the top four layers carry 0.38 of the total, against 0.08 for a uniform profile, a factor of roughly five; the top ten carry 0.65. Cross-checkpoint configurations behave almost identically, at 0.37 and 0.64, so a single ladder ordering remains usable when the producer's weights change. Because the band sits past mid-depth, a rung that repairs the top k layers must still run every layer below the deepest of them. This is why the work saved in Section 5.3 is far from $1 - k/L$; Appendix D gives the per-rung cost.

Flat profiles predict that repair cannot help. The three pairs on which the backbone fails outright have flat profiles: their top four layers carry 0.13 against a uniform baseline of 0.10, and their top ten carry 0.31. No small set of layers holds the error, so the intermediate rungs of Section 4.3 buy almost nothing. The ladder then degenerates into a choice between accepting a bad cache and paying for a dense prefill. This mechanism explains the numbers in Table 5: work saved on these pairs collapses to 0.06–0.11 not because CacheCert is conservative for its own sake, but because the profile leaves nothing cheaper than dense prefill to fall back on. By the same reading, the three

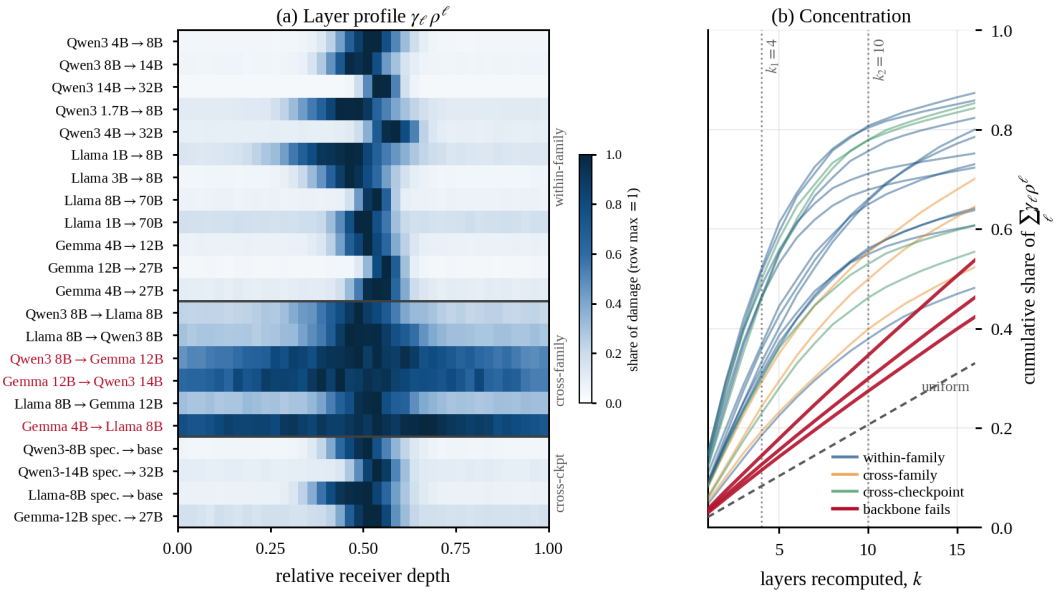


Figure 4: Per-layer damage profiles $\gamma_\ell \rho^\ell$ for all 22 configurations. (a) Each row is normalized to its own maximum so that shapes are comparable across receivers of different depth; the three pairs whose backbone fails are labelled in red. (b) Share of the total carried by the k most affected layers, with the ladder rungs k_1 and k_2 marked; the dashed line shows what a uniform profile would give.

cross-family pairs whose backbone does not fail lie in between, at 0.24 for the top four, with a usable if reduced operating point. Flatness is therefore a second screening signal that complements the scalar CTI: the index says whether to deploy a pair at all, and the profile says whether selective recomputation is worth implementing for it.

C PER-PAIR RESULTS

Table 5 supports two observations that did not fit in the main text. First, R^2 is not merely uninformative but can be actively misleading. Llama 8B →Qwen3 8B has an R^2 of 0.76, higher than nine of the twelve within-family pairs, yet its retention of 0.63 is lower than that of every within-family pair. A practitioner screening on reconstruction quality would rank it ahead of most pairs that actually transfer well. Second, CTI and CACHECERT work saved are tightly related (Spearman 0.98 across the eighteen pairs). This is expected, since both derive from the same profile, and it is useful operationally: CTI predicts not only whether a pair is viable but also roughly how much prefill a certified deployment will save.

Table 6 reports the four cross-checkpoint configurations. CTI, retention, and CACHECERT work saved order them identically, and the trained repair loses the most PGR exactly where CTI is lowest. Recomputing CTI against the new producer thus flags the producer changes that a fixed repair handles worst.

D EXTENDED IMPLEMENTATION DETAILS

Stack and hardware. All serving runs use vLLM 0.8.5 (Kwon et al., 2023) with PyTorch 2.6.0, CUDA 12.4, and FlashAttention 2.7.2 on nodes of 8×H100 80GB SXM with NVLink. Models above 32B are served with a tensor-parallel degree of 4; all others fit on a single device. Calibration and second-moment accumulation run in bfloat16 with float32 accumulators.

Cache injection hook. We add a hook at the paged-attention block manager. It accepts an externally supplied set of per-layer key and value tensors and marks the corresponding blocks as populated, bypassing the prefill for those positions. Rotary embeddings are applied after injection, so the

Table 5: All 18 directed pairs. Retention is $S_{\text{transfer}}/S_{\text{dense}}$ under the closed-form ridge backbone, averaged over the five workloads, with the standard deviation taken across workloads. CACHECERT columns are at $\alpha = 0.10$ on LongBench multi-hop QA, mean over five splits. Pairs marked † are the three on which the backbone fails outright; CACHECERT preserves validity on them by routing almost all traffic to the dense rung, which is why work saved collapses instead of risk rising.

	Pair	CTI	Retention	R^2	CACHECERT work	CACHECERT risk	Silent
within-family	Qwen3 4B→8B	0.93	0.96 ± 0.021	0.75	0.82	0.079	0.031
	Qwen3 8B→14B	0.89	0.94 ± 0.024	0.81	0.79	0.082	0.034
	Qwen3 14B→32B	0.88	0.93 ± 0.019	0.57	0.78	0.081	0.036
	Qwen3 1.7B→8B	0.58	0.81 ± 0.038	0.49	0.58	0.086	0.044
	Qwen3 4B→32B	0.71	0.78 ± 0.041	0.66	0.61	0.084	0.046
	Llama 1B→8B	0.53	0.92 ± 0.027	0.58	0.55	0.088	0.039
	Llama 3B→8B	0.73	0.95 ± 0.018	0.80	0.71	0.080	0.030
	Llama 8B→70B	0.61	0.89 ± 0.030	0.79	0.63	0.083	0.041
	Llama 1B→70B	0.48	0.76 ± 0.045	0.75	0.49	0.090	0.049
	Gemma 4B→12B	0.63	0.91 ± 0.026	0.68	0.62	0.085	0.038
	Gemma 12B→27B	0.88	0.93 ± 0.022	0.67	0.77	0.078	0.033
	Gemma 4B→27B	0.81	0.83 ± 0.035	0.62	0.68	0.083	0.043
	mean	0.72	0.88	0.68	**0.67**	**0.083**	**0.039**
cross-family	Qwen3 8B→Llama 8B	0.47	0.68 ± 0.052	0.60	0.38	0.091	0.058
	Llama 8B→Qwen3 8B	0.50	0.63 ± 0.049	0.76	0.34	0.089	0.061
	Qwen3 8B→Gemma 12B†	0.44	0.54 ± 0.061	0.66	0.11	0.094	0.072
	Gemma 12B→Qwen3 14B†	0.13	0.49 ± 0.058	0.68	0.08	0.092	0.069
	Llama 8B→Gemma 12B	0.29	0.58 ± 0.054	0.63	0.24	0.093	0.066
	Gemma 4B→Llama 8B†	0.16	0.46 ± 0.063	0.54	0.06	0.096	0.074
	mean	0.33	0.56	0.65	0.20	0.093	0.067

Table 6: Cross-checkpoint configurations, in which a further-trained specialist sibling writes the cache read by its base lineage, following the producer-change protocol of Shi & Lou (2026). The last column is the one of interest: a repair trained against the original producer degrades when the producer changes, whereas CACHECERT detects the change through its score and absorbs it as cost. CTI is recomputed against the new producer at 2.1 GPU-minutes.

Configuration	CTI	Retention	CACHECERT work	Trained-repair PGR drop
Qwen3-8B specialist → Qwen3-8B base	0.86	0.93	0.76	−0.07
Qwen3-14B specialist → Qwen3-32B	0.79	0.88	0.69	−0.19
Llama-8B specialist → Llama-8B base	0.84	0.91	0.73	−0.11
Gemma-12B specialist → Gemma-27B	0.68	0.81	0.58	−0.26

transferred content is stored position-free and re-rotated to the receiver's positions, following Heo et al. (2026). The hook adds 0.9 ms of fixed overhead per request, which we include in all reported time-to-first-token numbers.

Calibration corpora. Second moments use 256 sequences of 1024 tokens sampled from FineWeb-Edu. We chose this corpus to be disjoint in domain from all evaluation workloads, so that CTI is not tuned to the test distribution. The propagation probe uses 64 sequences and injects isotropic Gaussian noise at relative Frobenius norms $\{0.01, 0.02, 0.05, 0.10\}$ per layer. We fit γ_ℓ as the slope of next-token KL against the squared relative norm in the small-noise regime; the fit R^2 exceeds 0.97 for every layer of every receiver. Ridge parameters are selected per head by generalized cross-validation over $\lambda \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$, relative to $\mathrm{tr}(\mathbf{\Sigma}_{xx})/d_s$. The modal selection is 10^{-3}, and CTI varies by less than 0.01 across the grid.

Calibration size sweep. We compute CTI from $\{64, 128, 256, 512\}$ sequences and compare each value with the one obtained from 1024 sequences. Averaged over pairs, the maximum deviations are 0.031, 0.009, 0.006, and 0.002, respectively. We use 256 sequences, the point at which the marginal return falls below the between-pair resolution we need.

Probe construction. Probe positions are drawn by stratified sampling over quantiles of the source model's aggregate attention mass, with $m = 32$ positions allocated across 8 strata, plus mandatory

inclusion of the first 4 positions to cover attention sinks. Recomputation at a probe position uses the transferred cache for all layers below the probed layer, matching the selective-recomputation convention of Yao et al. (2025). Probes are batched into a single forward pass over a gathered subsequence. Before the probe residual enters Eq. (7), its key block is row-centered with weights w, as described in Appendix A.5, so that a uniform key offset scores zero.

Conformal calibration. We use $n = 300$ requests per workload, drawn disjointly from the test split. The ladder proportions are fixed at $(p_0, p_1, p_2) = (0.55, 0.80, 0.95)$, with $k_1 = 4$ and $k_2 = 10$ recomputed layers. The disagreement functional D is an exact-match flip on short-answer workloads; on the agent-report track, it is a key-entity mismatch judged by the autorater described below. We set $\varepsilon = 0$ in both cases, so any flip counts as a violation. The λ grid has 200 points. The monotonized loss of Corollary 4 is formed on this grid by taking, for each request, the running maximum of its loss curve from the largest grid point downward.

Autorater protocol. Silent-failure judgements use a held-out judge model that is not one of the eighteen pairs. It is prompted to decide independently whether an answer is fluent and whether it agrees with a reference on the key entity, with the two questions asked in separate calls to avoid anchoring. A blind human audit of 400 sampled judgements agrees with the autorater on 94.2% of fluency labels and 91.8% of agreement labels; disagreements concentrate on answers that hedge. We report the autorater numbers and release the audit set.

Decoding and stochasticity. Main-text results use greedy decoding, so any output difference is attributable to the cache rather than to sampling. A temperature sweep at $T \in \{0.3, 0.7, 1.0\}$ with three samples per request raises empirical risk at $\alpha = 0.10$ to 0.089, 0.102, and 0.118, respectively, because sampling introduces disagreement with the dense reference that is unrelated to transfer. The remedy, which we recommend and use for the $T > 0$ numbers, defines the loss against the dense model's sampling distribution rather than a single dense sample, estimated from the same three samples. This restores risk to 0.084, 0.087, and 0.091. We flag this as a genuine subtlety for anyone deploying the method with stochastic decoding.

Numerical determinism. Matrix-multiplication kernels are not bitwise deterministic across library builds or batch shapes. We pin `CUBLAS_WORKSPACE_CONFIG=:4096:8`, fix batch composition, and record library versions with every result file. Re-running the full evaluation on a second node changes individual request outcomes on 0.7% of requests and changes reported means by at most 0.004, well inside the split-to-split standard deviations we report. We note this because reading a $\{0, 1\}$-valued risk at the third decimal place would otherwise suggest more precision than exists.

Transfer backbone configurations. The closed-form ridge backbone uses the top-4 source layers selected by predictive correlation and is fitted per head in rotation-free content space. The head-local variant restricts the map to within-head blocks and reweights the fit by attention sensitivity. The trained fusion backbone uses a two-layer projector of width $2d$ per head with GELU, trained for $4\,000$ steps at batch size 32 on the same calibration corpus with AdamW at learning rate 3×10^{-4}. The selective-reuse backbone recomputes the 6 layers with the highest measured divergence, following the per-pair search of Liu et al. (2024b). The baseline gates are configured as follows: the entropy gate thresholds the predictive entropy of the first generated token; the logit-margin gate thresholds the gap between the top-1 and top-2 logits; the random gate accepts a fixed fraction uniformly at random; and CacheBlend recomputes the top 15% of tokens by attention deviation.

Statistical protocol. Every reported number is a mean over five random calibration and test splits, with the standard deviation taken across splits. Pairwise comparisons use a paired bootstrap over requests with 10^4 resamples, and we call a difference significant when the 95% interval excludes zero. Calibration and test requests are disjoint at the request level. On the agent-report track, they are also disjoint at the source-document level, so no document seen during calibration reappears at test.

Compute budget. The full study, including all four backbones, five seeds, and the exact-form validation of Appendix A.5, consumed approximately $5\,900$ H100-hours. The dense-prefill reference runs that define the ceiling and the risk labels account for roughly 61% of this budget. A practitioner applying the method to one new pair needs 2.1 GPU-minutes for CTI and 9.2 for conformal calibration. Table 7 breaks this down. The offline path costs 2.1 GPU-minutes per pair, against 84.6 for the fit-and-evaluate procedure it replaces. The online path adds 1.4 percent of prefill FLOPs and

Table 7: Cost accounting for the 14B→32B pair at 8K context on a single H100. Offline costs are incurred once per pair; the propagation probe runs once per receiver and amortizes over all its partners. Online overhead is reported relative to the dense prefill it replaces. Rung costs follow the convention of Section 4.3: repairing a set of layers requires running every layer below the deepest of them, so the cost is driven by depth rather than by count.

Component	Quantity	Cost
CTI second moments	256 seq. × 1024 tok.	1.6 GPU-min
CTI propagation probe	64 seq., once per receiver	0.5 GPU-min
Fitted-mapper baseline	fit + downstream eval	84.6 GPU-min
CACHECERT probe recompute	$m = 32$ of ∼8000 positions	0.8% of prefill FLOPs
CACHECERT sensitivity weighting	$O(mLd^2)$	0.6% of prefill FLOPs
CACHECERT rung 1 ($k_1 = 4$)	layers 0 to deepest repaired	59% of prefill FLOPs
CACHECERT rung 2 ($k_2 = 10$)	layers 0 to deepest repaired	72% of prefill FLOPs
CACHECERT added TTFT	measured, accepted requests	+11 ms
Conformal calibration	$n = 300$ requests, once per workload	9.2 GPU-min

11 ms of time to first token on accepted requests. Against a dense prefill of roughly 340 ms in this configuration, the first accepted request recovers that overhead many times over.

E NEGATIVE RESULTS AND VARIANTS THAT DID NOT WORK

We record these variants because they cost us time and because two of them are natural first ideas.

Fitting the key intercept. Consistent with Corollary 2, adding a fitted key-side intercept to the affine map changes downstream PGR by 0.004 on average, with no consistent sign, while adding d parameters per head. We therefore fix the intercept by centering.

Training the mapper under the sensitivity-weighted loss. Corollary 3 predicts that this is futile for the affine family, and we observe exactly that: the $\mathbf{G}$-weighted and unweighted ridge fits produce maps whose downstream PGR differs by less than 0.006 on every pair. For the trained fusion backbone, the weighted loss helps slightly (+0.018 PGR averaged over pairs), which we report but did not pursue.

Using the true residual instead of the fixed-point residual. We implemented an oracle variant of CACHECERT that scores requests with the true residual against a full dense prefill. It is not deployable, but it bounds what a better score could buy. It improves work saved at $\alpha = 0.10$ from 0.66 to 0.71, a gap of five points. We take this as evidence that the fixed-point surrogate is close to sufficient and that further effort is better spent elsewhere.

Compressing the transferred cache. Following the interaction reported by Shi & Lou (2026), we tried 4-bit and 8-bit quantization of the transferred cache. Both degrade PGR below that of uncompressed transfer at every matched byte budget we tried, and 4-bit quantization pushes two additional pairs below the no-context floor. We thus reproduce their conclusion that compression and cross-model misalignment interact badly, and we do not recommend combining them.

A single global threshold instead of per-workload calibration. Calibrating one $\widehat{\lambda}$ across all five workloads rather than per workload preserves marginal validity by construction but costs 11 points of work saved, because the hardest workload sets the threshold. Per-workload calibration is cheap enough that we see no reason to avoid it.