YOUR "PRO" LLM SUBSCRIPTION MAY ACTUALLY BE "FREE": EXPOSING FINGERPRINT SPOOFING RISKS IN LLM INFERENCE SERVICES

Anonymous authors
Paper under double-blind review

ABSTRACT

As Large Language Model (LLM) APIs become ubiquitous, users increasingly rely on black-box fingerprinting to verify that providers are serving the advertised premium models. However, these methods may overlook adversarial providers who manipulate model weights to cheat the fingerprinting process. We introduce a novel threat termed *fingerprint spoofing*, where a malicious provider stealthily serves a weaker model that has been parameter-efficiently fine-tuned to mimic a stronger model, thereby evading user-side fingerprinting. We first formally prove that user-side resource constraints (i.e., finite query budgets and weak fingerprinting classifiers) make current fingerprinting vulnerable to fingerprint spoofing. Guided by this theoretical analysis, we propose **GhostPrint**, a cost-effective attack framework leveraging surrogate modeling, reward-ranked fine-tuning, and knowledge distillation. Extensive evaluations in both static and continual fingerprinting settings demonstrate that GhostPrint allows weak models to consistently bypass representative fingerprinting methods while maintaining utility at a low fine-tuning cost, exposing a critical vulnerability in current LLM fingerprinting pipelines.

1 INTRODUCTION

LLMs have shown great ability on various tasks. A growing number of third-party providers offer API-based inference services for LLMs, such as OpenRouter[1] and other Model-as-a-Service providers (Liang et al., 2023; Gao et al., 2025b; Chen et al., 2024). These platforms serve a dual role: providing foundational architecture for downstream applications [e.g., coding agents (Yao et al., 2023; Shinn et al., 2023; Ning et al., 2026), retrieval-augmented generation (Fan et al., 2025; Liu et al., 2026; Yang et al., 2026; Nguyen et al., 2026), personalized chat-bots (Jiang et al., 2025a;b; Huang et al., 2026), multi-modal reasoning (Zhang et al., 2025d; Wu et al., 2026b; Fang et al., 2026)] and acting as essential infrastructure for academic evaluation (Chiang et al., 2024; Qin et al., 2024).

Given the foundational role of these API services, ensuring the integrity of the served models has become a pressing necessity (Cai et al., 2025b; Velasco et al., 2025; Zhang et al., 2026b;c). From a user's perspective, a practical approach to audit whether LLM APIs are serving the advertised model relies on LLM fingerprinting (i.e., model auditing) methods (Gao et al., 2025b; Gubri et al., 2024; Sun et al., 2025c; Pasquini et al., 2025). The key aim of these methods is to probe a black-box API with a small number of queries and efficiently analyze the responses to verify the model's identity. This relies on the common assumption that distinct response styles to specific prompts are sufficient to classify models (Pasquini et al., 2025; Sun et al., 2025c), making fingerprinting a trustworthy tool.

However, fingerprinting inherently assumes that the models remain static and are not adversarially manipulated by API providers, which might not hold in practice. For instance, an inference provider serving open-source models might adversarially fine-tune a weak model to closely mimic the fingerprint of a stronger one, bypassing user audits, for profit gain (see Figure 1). Similarly, companies hosting proprietary models (e.g., OpenAI, Google) could stealthily serve fine-tuned, weaker variants to reduce operational costs while evading detection. If these fine-tuning processes are computationally efficient and financially profitable, they pose a critical risk to API users. Therefore, we challenge the

[1] https://openrouter.ai/

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.

adversarial reliability of existing LLM fingerprinting methods and study a novel and fundamental research question: *Can a malicious model provider parameter-efficiently fine-tune a weak model so that the weak model behaves like a stronger model to bypass user-side fingerprinting tests?*

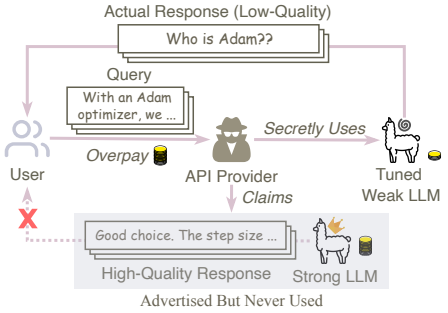

Figure 1: Motivation of LLM API spoofing

To answer this question, we formalize a threat model named **fingerprint spoofing**, where API service providers actively seek to bypass user-side model auditing (Figure 2(a)). In this setting, the provider advertises a strong LLM to charge a premium rate, but actually serves a parameter-efficiently (Hu et al., 2022; Dettmers et al., 2023; Liu et al., 2024) fine-tuned weak LLM to evade fingerprint checks. Successfully executing such an attack is highly non-trivial and presents severe technical challenges. Fundamentally, based on LLM scaling laws (Kaplan et al., 2020; Hoffmann et al., 2022; Sardana et al., 2024) and our lower-bound analysis (Proposition E.3), a small model inherently lacks the representational capacity to perfectly mimic the behavior of a much larger model. Consequently, the attacker faces a severe optimization trade-off: aggressively fine-tuning a weak model to mimic a strong model's fingerprint behavior may lead to catastrophic forgetting (Li et al., 2024; Luo et al., 2025), destroying its ability to answer normal user queries. Furthermore, mimicking a large model with a small model must be achieved under resource constraints, since full-parameter retraining is often expensive and would make the adversary's attack financially infeasible.

Formalizing these technical challenges, our goal is to study whether a malicious provider can successfully resolve this trade-off to simultaneously achieve three conflicting objectives: (i) **fingerprint spoofing**, where the adapted model becomes behaviorally indistinguishable from the claimed strong model under auditing; (ii) **utility preservation**, where the adapted weak model maintains acceptable performance on normal downstream tasks; and (iii) **economic gain**, where the attack only relies on inexpensive parameter-efficient fine-tuning (PEFT) mechanisms.

To understand why spoofing remains feasible under the three goals, we theoretically analyze the vulnerabilities of current fingerprint pipelines. Constrained by user-side resources, users often rely on limited query sets (Gubri et al., 2024; Pasquini et al., 2025) and weak verification mechanisms (Sun et al., 2025c; Gao et al., 2025b). Consequently, an attacker will not need to emulate the strong model universally, and a localized, parameter-efficient adaptation suffices to fool the user-side audit. We formalize these limitations in Theorem 5.1 (low-effective-rank query distributions) and Theorem 5.2 (weak local classifiers). Guided by these theoretical weaknesses, we propose a novel attack framework, namely **GhostPrint**, to overcome the utility-spoofing trade-off. Specifically, the attacker trains a surrogate fingerprint model and employs reward-ranked fine-tuning to identify queries where the weak model can naturally mimic the larger model. We then leverage LoRA-based supervised fine-tuning (SFT) alongside knowledge distillation (KD) to align the weak model with the strong model, specifically on the targeted fingerprint queries, effectively bypassing the audit while preserving general utility. Furthermore, recognizing that user-side fingerprints evolve over time, we extend our framework to a continual spoofing setting, exploring whether an attacker can dynamically adapt to new fingerprint queries and auditing methods. We empirically validate this attack framework in both static and continual settings, exposing the practical risks posed by fingerprint spoofing.

Our **core contributions** are: (i) We formally demonstrate when and why fingerprint spoofing is possible, proving that user-side resource constraints (i.e., weak query sets and weak surrogate models) make existing auditing methods fundamentally breakable (Theorems 5.1 and 5.2); (ii) We propose a novel, practical attack that can successfully bypass static fingerprinting while maintaining task utility; and (iii) We extend our attack to dynamic environments, demonstrating that malicious providers can continually adapt to evolving user-side fingerprinting methods.

2 Related Works

LLM Fingerprinting. Pay-per-use APIs create incentives for providers to silently substitute weaker models (Cai et al., 2025b; Zhang et al., 2026b). Fingerprinting addresses this by verifying model identity. Existing methods fall into three categories: white-box non-intrusive approaches that inspect

internal signals (Zheng et al., 2022; Wu et al., 2025; Zhang et al., 2025b; 2026a), black-box non-intrusive approaches that analyze API outputs (Gao et al., 2025b; Sun et al., 2025c; Pasquini et al., 2025; Gubri et al., 2024), and intrusive methods that embed verifiable signatures during training (Gu et al., 2022; Li et al., 2023; Xu et al., 2024; Yamabe et al., 2025). More discussion is in Appendix C.

Attacks Against LLM Fingerprinting. Adversarial robustness of fingerprinting is an emerging concern. Prior attacks use rewriting or statistical evasion but assume no fine-tuning of the weak model and rely on distinguishing query distributions (Kurian et al., 2025; Nasery et al., 2025a). In contrast, we consider a provider that actively fine-tunes a weak model to spoof fingerprints, a realistic threat that undermines distribution-agnostic auditors like LLM-idio (Sun et al., 2025c) and Model Equality Testing (MET) (Gao et al., 2025b).

3 PRELIMINARIES

Notations. We use $\|\mathbf{W}\|_F$ and $\|\mathbf{W}\|_2$ to denote the matrix Frobenius and spectral norms, respectively, and use $\|\mathbf{x}\|_2$ to denote the vector ℓ_2 norm. We treat an LLM as a conditional probability distribution over token sequences. Given an LLM f parameterized by $\boldsymbol{\theta}$ and a prompt q, a response r is sampled as $r \sim f(\cdot \mid q; \boldsymbol{\theta})$. When the parameters are fixed, we write $r \sim f(\cdot \mid q)$ for notational simplicity.

Black-Box LLM Fingerprinting. The goal of black-box fingerprinting is to audit an LLM API, denoted as f_{api}, to determine whether it actually serves a specific claimed target model, f_{target}. The auditing procedure is formalized below.

Definition 3.1 (Black-Box LLM Fingerprinting)**.** *Let the user/auditor have a query distribution $\mathcal{D}_q$, a query budget n, and a local classification mechanism g. The fingerprinting procedure is:*

1. ***Query Sampling**: The auditor samples a query set $\mathcal{Q}^n := \{q^i\}_{i=1}^n$, where $q^i \sim \mathcal{D}_q$.*
2. ***Response Collection**: The auditor queries the API and collects the responses $\mathcal{R}^n := \{r^i\}_{i=1}^n$, where $r^i \sim f_{\text{api}}(\cdot \mid q^i)$ and $q^i \in \mathcal{Q}^n$.*
3. ***Verification**: The mechanism g maps the collected responses to a binary decision $g(\mathcal{R}^n, f_{\text{target}}) \in \{0, 1\}$, returning 1 if the API is verified as f_{target}, and 0 otherwise.*

The definition naturally encapsulates existing representative black-box fingerprinting methods, which differ mainly in their formulation of the query distribution $\mathcal{D}_q$ and the verification mechanism g. For instance, LLMmap (Pasquini et al., 2025) restricts $\mathcal{D}_q$ to a small set of carefully crafted, identity-revealing probes and trains g as a multi-class embedding classifier. MET (Gao et al., 2025b) uses a general $\mathcal{D}_q$ but formulates g as a statistical two-sample kernel test that computes the distance between $\mathcal{R}^n$ and reference generations, thresholding the p-value to yield a binary decision.

Practical Constraints of Fingerprinting. In practice, the query budget n in Definition 3.1 enforces a trade-off between auditing accuracy and financial cost, as users must pay for every API call. A larger n can improve auditing accuracy but incurs a higher cost. Furthermore, user-side resource constraints typically restrict the verification mechanism g to lightweight statistical tests (Gao et al., 2025b; Zhu et al., 2026) or weak surrogate models (Pasquini et al., 2025; Sun et al., 2025c) rather than state-of-the-art LLMs. We focus on black-box fingerprinting, because this mirrors the reality of commercial LLM APIs that the auditor only has query access.

4 PROBLEM FORMULATION

We study a malicious third-party API provider (i.e., the adversary) that claims to serve a strong target model f_{target}, but in reality deploys a weaker and cheaper model architecture f_{weak}, initially parameterized by $\boldsymbol{\theta}_{\text{weak}}$. The adversary aims to adapt $\boldsymbol{\theta}_{\text{weak}}$ into deceptive parameters $\boldsymbol{\theta}_{\text{adv}}$ such that the auditor's local mechanism g incorrectly verifies $f_{\text{weak}}(\cdot \mid \cdot; \boldsymbol{\theta}_{\text{adv}})$ as f_{target}.

Attacker's Capabilities. We assume the attacker has the following capabilities: (i) **Knowledge of the fingerprinting mechanism**: The attacker has full knowledge of the auditor's local classification mechanism g (e.g., feature extractor, distance metric, etc.) and can build a surrogate $\overline{g}$ that mimics the auditor's g; (ii) **Knowledge of the query distribution**: The attacker knows the distribution $\mathcal{D}_q$ from which the auditor draws queries. This allows the attacker to sample a surrogate query set $\overline{\mathcal{Q}}^m := \{\overline{q}_i\}_{i=1}^m$, where $\overline{q}_i \sim \mathcal{D}_q$. Importantly, *the attacker does not know the exact finite query set $\mathcal{Q}^n$ instantiated by the auditor during the verification*; and (iii) **White-box manipulation of**

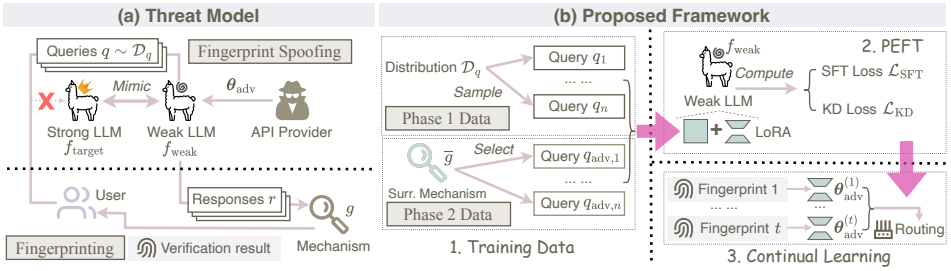

Figure 2: Overview of GhostPrint. (a) Threat model. An adversary leverages a weak LLM to mimic a strong target model and evade fingerprint-based verification when interacting with an API provider. (b) Proposed framework. GhostPrint constructs training data from both target-model and surrogate queries, then fine-tunes a weak LLM via SFT and KD, and enables continual learning to progressively adapt the adversarial fingerprint over time.

the weak model: The attacker has full access to the deployed architecture f_{weak} and can tune its parameters via parameter-efficient fine-tuning (PEFT).

Attacker's Goal. The attacker aims to simultaneously achieve: (i) *Fingerprint spoofing*: Under the fingerprinting mechanism (Definition 3.1), the auditor's local mechanism g incorrectly verifies $f_{\text{weak}}(\cdot \mid \cdot; \boldsymbol{\theta}_{\text{adv}})$ as f_{target}; (ii) *Utility preservation*: After adaptation, the served model $f_{\text{weak}}(\cdot \mid \cdot; \boldsymbol{\theta}_{\text{adv}})$ still maintains acceptable normal-task performance; and (iii) *Economic gain*: The process to obtain $\boldsymbol{\theta}_{\text{adv}}$ is cost-effective.

Practicality of Threat Model. Assuming the adversary knows both the query distribution $\mathcal{D}_q$ and the classifier g aligns with adaptive-attacker evaluations rooted in Kerckhoffs's principle. For an audit to be credible and independently verifiable, its methodology, classification algorithm g, and query distribution $\mathcal{D}_q$ must be transparent and publicly disclosed. Relying on secrecy of the auditing pipeline constitutes security by obscurity, so a fully informed attacker model is necessary.

Based on the threat model described above, our problem is formally defined as

Definition 4.1 (Fingerprint Spoofing Problem)**.** *Given a white-box weak model architecture f_{weak} initially parameterized by $\boldsymbol{\theta}_{\text{weak}}$, access to a target model f_{target}, and knowledge of the auditor's query distribution $\mathcal{D}_q$ and verification mechanism g, the fingerprint spoofing problem seeks to efficiently find adapted parameters $\boldsymbol{\theta}_{\text{adv}}$ for the weak model to achieve three primary objectives:*

1. ***Fingerprint Spoofing:*** *The adapted model $f_{\text{weak}}(\cdot \mid \cdot; \boldsymbol{\theta}_{\text{adv}})$ must generate responses that fool the auditor's mechanism g into identifying it as the target model over the distribution $\mathcal{D}_q$.*
2. ***Utility Preservation:*** *The adapted model $f_{\text{weak}}(\cdot \mid \cdot; \boldsymbol{\theta}_{\text{adv}})$ must maintain general model utility comparable to the initial weak model $f_{\text{weak}}(\cdot \mid \cdot; \boldsymbol{\theta}_{\text{weak}})$, ensuring the API remains functional for benign user requests.*
3. ***Economic Gain:*** *The adversary's fine-tuning process to obtain $\boldsymbol{\theta}_{\text{adv}}$ must be economically viable. Specifically, the adapted parameters are obtained via Parameter-Efficient Fine-Tuning (PEFT), i.e., $\boldsymbol{\theta}_{\text{adv}} = \boldsymbol{\theta}_{\text{weak}} + \Delta\boldsymbol{\theta}$, where $\Delta\boldsymbol{\theta}$ is a small low-rank update.*

5 Proposed Method

In this section, we establish theoretical limits of LLM fingerprinting and present GhostPrint, an attack framework motivated by these results. An illustration of our proposed framework is in Figure 2.

5.1 When and Why Could Fingerprint Spoofing Happen?

Fingerprint spoofing asks whether limited parameter updates can make a weak model imitate a stronger one. Matching the strong model's behavior across all queries can be inherently hard. Nevertheless, fooling a practical LLM API identity auditor may still be possible (e.g., one using a few test prompts (Pasquini et al., 2025) or a separate weak model to classify responses (Sun et al., 2025c)). We study when and why fingerprint spoofing happens through a lower bound on universal spoofing error and upper bounds under restricted queries and weak verification, modeled by a low-rank verifier.

Problem Setting. We study linear maps $f(\mathbf{x}) = \mathbf{W}\mathbf{x}$, causal attention, and Transformers (Vaswani et al., 2017). For token representations $\mathbf{X}$, causal attention computes $\mathrm{softmax}_{\mathrm{causal}}((\mathbf{X}\mathbf{Q})(\mathbf{X}\mathbf{K})^\top/\sqrt{d_k})\mathbf{X}\mathbf{V}$, where $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ are projection weights and d_k is the key dimension. Transformers stack attention and feed-forward blocks with residual connections and normalization. For both architectures, we compare last-token outputs.

Let f_* be a strong target model from one of these three classes, $\mathcal{X}$ the input domain, and $\mathcal{F}_r$ the candidate weak-model family with rank-at-most-r weight updates for linear models or query/key updates for attention or one Transformer head. For $f \in \mathcal{F}_r$, we measure imitation error by $\|f_*(x) - f(x)\|_2$. Formal statements and proofs are in Appendices D and E, respectively.

Main Results. We formalize three main results in Appendix D. First, Proposition D.3 gives a *lower bound* on universal fingerprint spoofing for all three model families. Under its hypotheses, we have $\inf_{f\in\mathcal{F}_r} \sup_{x\in\mathcal{X}} \|f_*(x) - f(x)\|_2 \geq B$, where $B \geq 0$ depends on $\sigma_{r+1}(\mathbf{W}_* - \mathbf{W}_0)$ for linear models, or the singular-value tail beyond rank $2r$ of $\mathbf{S}_* - \mathbf{S}_0$ for attention and its subspace restriction for Transformers, with $\mathbf{S} := \mathbf{Q}\mathbf{K}^\top$. When $B > 0$, exact universal spoofing is impossible within the allowed update family, highlighting a key challenge for our attack.

To address this challenge, we derive two *upper bounds* under practical auditor weaknesses: restricted queries and weak response classifiers. For example, LLMmap uses as few as eight queries and a lightweight Transformer classifier (Pasquini et al., 2025), while LLM-idio uses a linear head over fine-tuned text embeddings (Sun et al., 2025c).

Our second result addresses queries whose representations mainly vary along a few directions.

Theorem 5.1 (Upper bound under restricted queries, informal version of Theorem D.4)**.** *Let $\mathcal{D}_q$ be the audit query distribution. Under the input and update conditions of Theorem D.4, we have*

$$\inf_{f\in\mathcal{F}_r} \mathbb{E}_{x\sim\mathcal{D}_q}\|f_*(x) - f(x)\|_2^2 \leq b^2 \sum\nolimits_{j>r} \lambda_j(\mathbf{\Sigma}),$$

where $\mathbf{\Sigma}$ and b are the input second-moment matrix and model-dependent coefficient specified there, and its eigenvalues $\lambda_j(\mathbf{\Sigma})$ are ordered decreasingly.

A small right-hand side allows low-rank updates to closely match the target on the audit distribution.

Our third result models weak verification by a low-rank linear map. Here we fix the candidate features and allow an additional update to the output map.

Theorem 5.2 (Upper bound under low-rank verification, informal version of Theorem D.7)**.** *Let $\mathbf{z}(x) \in \mathbb{R}^d$ be fixed candidate features, $\mathbf{W}_0 \in \mathbb{R}^{q\times d}$ the output weights, and $\mathbf{C} \in \mathbb{R}^{q\times c}$ the verifier, which observes $\mathbf{C}^\top f(x)$. Assume $f_*, \mathbf{z}$ have finite second moments under $\mathcal{D}_q$. Define $\varepsilon_{\mathbf{C}}^2 := \inf_{\mathbf{B}\in\mathbb{R}^{c\times d}} \mathbb{E}_{x\sim\mathcal{D}_q}\|\mathbf{C}^\top f_* - \mathbf{B}\mathbf{z}\|_2^2$. If $s \geq \mathrm{rank}(\mathbf{C})$, then for $f_{\mathbf{\Delta}} := (\mathbf{W}_0 + \mathbf{\Delta})\mathbf{z}$, we have*

$$\inf_{\mathrm{rank}(\mathbf{\Delta})\leq s} \mathbb{E}_{x\sim\mathcal{D}_q}\|\mathbf{C}^\top(f_* - f_{\mathbf{\Delta}})\|_2^2 = \varepsilon_{\mathbf{C}}^2.$$

The result above applies to linear models, causal attention, and Transformers by updating their weight matrix, value projection, and untied output head, respectively, with candidate features fixed (see Definition D.5). Here $\varepsilon_{\mathbf{C}}^2$ is the best linear prediction error for the target's observed output. A small residual permits accurate matching. It vanishes when this output is linear in the candidate features.

Takeaway. Our analysis reveals a critical gap: matching the target on all queries can be impossible (Proposition D.3), yet matching only what an auditor observes can be feasible under restricted queries or weak verification (Theorems 5.1 and 5.2). These two weaknesses motivate GhostPrint's fine-tuning on audit queries and use of a surrogate verifier.

5.2 The Proposed GhostPrint Attack

As shown in Section 5.1, practical LLM fingerprinting pipelines suffer from two fundamental vulnerabilities: (i) low-complexity query distributions $\mathcal{D}_q$, and (ii) the reliance on computationally weak verification models g. A potent fingerprint spoofing attack must actively exploit both vulnerabilities simultaneously. To this end, we propose **GhostPrint**, a principled attack framework that leverages Supervised Fine-Tuning (SFT) and Knowledge Distillation (KD) to efficiently memorize the behavioral characteristics of the target model over $\mathcal{D}_q$ while preserving normal utility. Furthermore, we

incorporate a reward-ranked fine-tuning (RAFT) (Dong et al., 2023) mechanism, utilizing a surrogate fingerprint model $\overline{g}$ to filter and select weak model outputs that match the strong model's fingerprint. To satisfy the attacker's economic constraints, all optimizations are conducted parameter-efficiently.

Exploiting the Query Distribution $\mathcal{D}_q$. We employ Supervised Fine-Tuning (SFT) to force the weak model to mimic the target model's fingerprinting behavior. For a fingerprint query $q \sim \mathcal{D}_q$ and the target response r generated by f_{target}, the SFT objective minimizes the negative log-likelihood:

$$\mathcal{L}_{\text{SFT}}(\boldsymbol{\theta}_{\text{adv}}) := -\mathbb{E}_{q,r}\Big[\sum\nolimits_{t=1}^{|r|} \log f_{\text{weak}}(r_t \mid q, r_{<t}; \boldsymbol{\theta}_{\text{adv}})\Big].$$

To preserve general capability without external data, we apply Knowledge Distillation (KD) on the same queries, using the target model as teacher:

$$\mathcal{L}_{\text{KD}}(\boldsymbol{\theta}_{\text{adv}}) := \mathbb{E}_{q,r}\Big[\sum\nolimits_{t=1}^{|r|} D_{\text{KL}}\big(p_t^{\text{target}} \big\| \, p_t^{\text{weak}}\big)\Big],$$

where p_t^{target} and p_t^{weak} are the next-token distributions of the target and weak models, respectively. The total objective is

$$\min_{\boldsymbol{\theta}_{\text{adv}}} \mathcal{L}_{\text{total}} := \mathcal{L}_{\text{SFT}}(\boldsymbol{\theta}_{\text{adv}}) + \alpha \mathcal{L}_{\text{KD}}(\boldsymbol{\theta}_{\text{adv}}), \tag{1}$$

where $\alpha > 0$ is a hyperparameter balancing fingerprint deception and utility.

Exploiting the Fingerprint Mechanism g. After optimizing (1), the weak model mimics the target on the query distribution $\mathcal{D}_q$. However, when $\mathcal{D}_q$ comprises diverse, general-purpose queries [e.g., UltraChat (Ding et al., 2023)], standard SFT alone may fail to capture the subtle model-specific fingerprints required by auditors such as LLM-idio (Sun et al., 2025c) and MET (Gao et al., 2025b). To address this, we directly exploit the structural weakness of the verification model g by incorporating its supervisory signal into the fine-tuning process.

Knowledge of the verification mechanism does not require access to the auditor's deployed instance (Section 4). Even if accessible, employing it as a reward model for reinforcement learning (RL) (Guo et al., 2025; Li et al., 2026a;b; Luo et al., 2026) would be prohibitively expensive and unstable, violating the attacker's economic constraints. We instead construct a local surrogate $\overline{g}$ from the known query distribution $\mathcal{D}_q$, which can be a lightweight classifier (for LLMmap, LLM-idio) or a reproduced statistical metric (for MET). To leverage $\overline{g}$ without RL, we adopt RAFT (Dong et al., 2023) as an additional post-training phase.

Specifically, for each query $q \sim \mathcal{D}_q$, we sample K candidate responses from the current weak model, rank them with $\overline{g}$, and select the most deceptive one as $r_{\text{adv}} := \arg\max_{r_j} \overline{g}(r_j, f_{\text{target}})$. The resulting (q, r_{adv}) pairs are then used to fine-tune the adapter parameters $\boldsymbol{\theta}_{\text{adv}}$ with the same SFT and KD objectives defined in (1). This approach filters the weak model's natural generation space to isolate responses that already cross the surrogate's decision boundaries. By fine-tuning on these selected examples, the model leverages the signal from the target weak classifier g, while avoiding the unstable and costly use of RL.

Parameter-Efficient Fine-Tuning. To address the economic gain goal shown in Section 4, we restrict the attacker to apply Low-Rank Adaptation (LoRA) to the weak model rather than computationally prohibitive full-parameter fine-tuning.

Let the initial parameters of the weak architecture f_{weak} be denoted as the set of N targeted base weight matrices $\boldsymbol{\theta}_{\text{weak}} := \{\mathbf{W}'_i\}_{i=1}^N$. We formalize the deceptive parameter set $\boldsymbol{\theta}_{\text{adv}}$ by assigning each adapted weight matrix $\widehat{\mathbf{W}}_i$ as the sum of the frozen base weight and a low-rank update:

$$\boldsymbol{\theta}_{\text{adv}} := \{\widehat{\mathbf{W}}_i \mid \widehat{\mathbf{W}}_i = \mathbf{W}'_i + \mathbf{A}_i\mathbf{B}_i\}_{i=1}^N, \tag{2}$$

where $r < \min(d_{1,i}, d_{2,i})$ is the adapter rank, and $\mathbf{A}_i \in \mathbb{R}^{d_{1,i} \times r}$ and $\mathbf{B}_i \in \mathbb{R}^{r \times d_{2,i}}$ are the trainable LoRA matrices for the i-th target weight.

By freezing $\boldsymbol{\theta}_{\text{weak}}$ and optimizing the set $\Delta\boldsymbol{\theta} := \{(\mathbf{A}_i, \mathbf{B}_i)\}_{i=1}^N$, the adversary transitions the generation distribution to $f_{\text{weak}}(\cdot \mid \cdot; \boldsymbol{\theta}_{\text{adv}})$ while reducing the computational cost of the attack.

Continual Learning of Multiple Fingerprints. Our two-phase training bypasses a single fingerprint. However, auditors might continually deploy new methods. Thus, an adversary must spoof a sequence of verifiers $g^{(\tau)}$ with query distributions $\mathcal{D}_q^{(\tau)}$ for $\tau = 1, \ldots, T$, updating f_{weak} without

Table 1: **Single-fingerprint spoofing.** Attack Success Rate (ASR, %) across three model families. Bold marks the best attack results. Model utility is reported separately in Table 2.

Method	Gemma-1.1 2B → 7B			Qwen2 1.5B → 7B			Phi-3 mini → medium		
	LLMmap	LLM-idio	MET	LLMmap	LLM-idio	MET	LLMmap	LLM-idio	MET
Original (weak)	12.0	7.2	6.7	4.0	7.4	33.3	6.0	28.8	43.3
Query Detection	35.0	7.2	13.3	70.0	7.4	**40.0**	70.0	28.8	50.0
Rewriting	20.0	0.7	0.0	40.0	34.2	33.3	0.0	38.6	33.3
5-shot ICL	15.0	22.9	3.3	0.0	64.4	30.0	0.0	**70.5**	36.7
Token Suppression	28.0	4.5	0.0	6.0	4.9	6.7	5.0	51.2	6.7
GhostPrint (ours)	**95.0**	**58.5**	**23.3**	**76.0**	**71.3**	**40.0**	**80.0**	64.0	**60.0**
Original (target)	92.0	98.8	100.0	81.0	98.6	96.7	91.0	77.2	96.7

catastrophic forgetting. For each task at continual learning step τ, we train a task-specific LoRA adapter $\Delta\boldsymbol{\theta}^{(\tau)} := \{(\mathbf{A}_i^{(\tau)}, \mathbf{B}_i^{(\tau)})\}_{i=1}^{N}$ via our aforementioned process, then integrate adapters with a Mixture-of-LoRA-Experts (MoLA) (Gao et al., 2025a) architecture. It routes queries to relevant adapters, exploiting shared structure across fingerprinting mechanisms.

Specifically, for the i-th target weight matrix at continual learning step τ, we introduce a trainable router matrix $\mathbf{V}_i^{(\tau)} \in \mathbb{R}^{\tau \times d_{2,i}}$, where $d_{2,i}$ is the dimension of the intermediate hidden state $\mathbf{x}$. The router computes a τ-dimensional logit vector to assign a routing score $s_{i,m}(\mathbf{x})$ for the m-th expert as $s_{i,m}(\mathbf{x}) := \frac{\exp(\mathbf{V}_i^{(\tau)}\mathbf{x})_m}{\sum_{j=1}^{\tau}\exp(\mathbf{V}_i^{(\tau)}\mathbf{x})_j}$. The adapted output $\mathbf{h}_i^{(\tau)}$ for the i-th base weight matrix $\mathbf{W}_i'$ and input $\mathbf{x}$ is then computed as a dynamically weighted combination of the historical LoRA experts as:

$$\mathbf{h}_i^{(\tau)} := \mathbf{W}_i'\mathbf{x} + \sum\nolimits_{m=1}^{\tau} s_{i,m}(\mathbf{x})\mathbf{A}_i^{(m)}\mathbf{B}_i^{(m)}\mathbf{x}.$$

By freezing previous adapters $\Delta\boldsymbol{\theta}^{(<\tau)} := \{\Delta\boldsymbol{\theta}^{(1)}, \Delta\boldsymbol{\theta}^{(2)}, \ldots, \Delta\boldsymbol{\theta}^{(\tau-1)}\}$ and only training the new adapter $\Delta\boldsymbol{\theta}^{(\tau)}$ alongside the router $\mathbf{V}_i^{(\tau)}$, GhostPrint continually expands its spoofing capabilities across multiple detection methods while strictly maintaining parameter efficiency.

To train the router matrices, we reuse the same SFT and KD objectives defined in Eq. (1). The training data for this routing phase is a mixture of queries from all tasks seen so far, i.e., $\mathcal{D}_q^{(1)} \cup \mathcal{D}_q^{(2)} \cdots \cup \mathcal{D}_q^{(\tau)}$. Because we only train the new adapter and a very small set of router weights $\mathbf{V}^{(\tau)} := \{\mathbf{V}_i^{(\tau)}\}_{i=1}^{N}$, this continual learning process remains highly efficient.

6 Experiments

6.1 Experimental Settings

Models and Datasets. We instantiate our attack across three representative LLM families selected from the model list of LLMmap (Pasquini et al., 2025). We utilize Gemma-1.1 (2B as the weak model and 7B as the strong target) (Team et al., 2024), Qwen2 (1.5B as weak and 7B as strong) (Yang et al., 2024), and Phi-3 (Mini 3.8B as weak and Medium 14B as strong) (Abdin et al., 2024). To quantify the preservation of general capabilities, model utility is evaluated on three standard benchmarks under a 5-shot setting using the open-source LM Evaluation Harness: MMLU (Hendrycks et al., 2020), GSM8K (Cobbe et al., 2021), and ARC-Challenge (ARC-C) (Clark et al., 2018).

LLM Fingerprints. We evaluate against three representative black-box LLM fingerprints to audit LLM APIs: (i) *LLMmap* (Pasquini et al., 2025), a 52-class classifier driven by adversarial identity probes; (ii) *LLM-idiosyncrasies (LLM-idio)* (Sun et al., 2025c), a binary classifier trained on Ultra-Chat (Ding et al., 2023) responses; and (iii) *Model Equality Testing (MET)* (Gao et al., 2025b), a two-sample hypothesis test comparing completion distributions via MMD and Hamming distance.

Spoofing Baselines. We compare against four baselines: (i) *Query Detection*, where we train a simple MLP to memorize historical fingerprint queries, routing flagged queries to the strong model and all others to the weak model; (ii) *Rewriting* (Kurian et al., 2025), which uses their proposed prompt

Table 2: **General model utility.** Comparison of GhostPrint with the original weak and target models. The four inference-time baselines in Table 1 (Query Detection, Rewriting, 5-shot ICL, and Token Suppression) do not update model weights, so their utility is not reported separately in this table.

Method	Gemma-1.1 2B → 7B			Qwen2 1.5B → 7B			Phi-3 mini → medium		
	MMLU	GSM8K	ARC-C	MMLU	GSM8K	ARC-C	MMLU	GSM8K	ARC-C
Original (weak)	37.90	10.69	44.97	55.55	55.42	43.52	70.37	80.36	62.97
GhostPrint (ours)	38.11	8.72	44.54	55.36	55.80	44.97	69.70	78.01	62.80
Original (target)	58.70	49.58	56.83	70.67	72.93	62.12	78.12	85.52	66.38

Table 3: **Cross-family weak-to-strong spoofing.** ASR (%) for two transfer directions.

Method	Gemma-1.1-2B → Qwen2-7B			Qwen2-1.5B → Gemma-1.1-7B		
	LLMmap	LLM-idio	MET	LLMmap	LLM-idio	MET
Original (weak)	0.0	0.7	0.0	0.0	1.1	0.0
Query Detection	0.0	2.4	6.7	0.0	2.5	**6.7**
Rewriting	0.0	0.1	0.0	0.0	1.8	0.0
5-shot ICL	0.0	3.6	0.0	0.0	17.3	0.0
Token Suppression	0.0	2.0	6.7	**15.0**	1.7	0.0
GhostPrint (ours)	0.0	**95.2**	**33.3**	6.0	**96.7**	**6.7**
Original (target)	80.0	99.9	96.7	97.0	99.2	86.7

to rewrite the generated response; (iii) *5-shot ICL*, which provides five strong-model responses as in-context examples; and (iv) *Token Suppression* (Nasery et al., 2025a), which applies their Suppress Top-K attack to randomize generation and mask the model's standard behavioral patterns. For reference, we also report the unmodified baseline performance of both the base weak model and the target strong model.

Evaluation Setting. All fingerprint spoofing attacks are evaluated using an 8:2 train-test split for all fingerprint queries. For LLMmap (Pasquini et al., 2025), we use their official query set. For LLM-idio (Sun et al., 2025c) and MET (Gao et al., 2025b), we use the UltraChat (Ding et al., 2023) dataset as the fingerprint queries, consistent with the original papers. Due to space limitations, full implementation details are deferred to Appendix A.

6.2 Spoofing Performance

Single-Fingerprint Spoofing Performance Tables 1 and 2 report the single-fingerprint spoofing performance and model utility of GhostPrint, with full implementation details of GhostPrint deferred to Appendix B.1. The results demonstrate that our framework successfully achieves the attacker's primary goals shown in Section 4. **(i)** Regarding *fingerprint spoofing*, GhostPrint achieves the highest Attack Success Rate (ASR), including ties, in eight of the nine model–auditor settings; the exception is LLM-idio on Phi-3, where 5-shot ICL performs better. Notably, when evaluating the Gemma family against LLMmap, GhostPrint achieves a 95% ASR, even surpassing the original target model's natural classification rate of 92%. **(ii)** In terms of *utility preservation*, GhostPrint maintains downstream performance that is almost identical to the base weak models. Our parameter-efficient fine-tuning incurs only marginal performance drops or even slight improvements, confirming that the attack successfully preserves the weak model's general capabilities on normal tasks.

Cross-Model Spoofing Performance Next, we extend the within-family setting above to the more challenging cross-family setting. Specifically, we study two transfer scenarios: Gemma-1.1-2B to Qwen2-7B, and Qwen2-1.5B to Gemma-1.1-7B. From the results in Table 3, we observe: **(i)** Cross-family spoofing is inherently more difficult because different model architectures exhibit distinct baseline linguistic styles and vocabulary distributions; and **(ii)** Despite this architectural gap, GhostPrint still achieves notable success, successfully bypassing LLM-idio in both settings and MET in the Gemma-to-Qwen transfer.

Table 4: **Continual fingerprint spoofing with MoLA.** ASR (%) across routing configurations. Bold values indicate the highest ASR within each model family and column.

Router configuration	Gemma-1.1 2B → 7B			Qwen2 1.5B → 7B			Phi-3 mini → medium		
	LLMmap	LLM-idio	MET	LLMmap	LLM-idio	MET	LLMmap	LLM-idio	MET
LLMmap + LLM-idio	86.0	54.2	10.0	75.0	**58.4**	23.3	71.0	**53.0**	43.3
LLMmap + MET	**89.0**	13.2	10.0	74.0	55.5	16.7	70.0	27.5	43.3
LLM-idio + MET	20.0	**58.6**	**30.0**	20.0	57.1	**26.7**	0.0	49.3	**60.0**
3-way (all)	86.0	54.8	16.7	**81.0**	57.6	**26.7**	**77.0**	50.6	53.3

For completeness, we also evaluate cross-family spoofing between models of similar scales in Appendix B.2. However, as this does not fit the financial gain goals for the adversary, it falls outside of our threat model and is put in the appendix.

6.3 Ablation Study

Knowledge-Distillation Loss. In this study, we evaluate the stability of knowledge distillation under different coefficients α in Eq. (1). We also investigate whether the small model or the large model should serve as the reference anchor.

Figure 3 shows that using the large model as the anchor gives slightly higher ASR, while the small-model anchor remains effective. Thus, the attack works with response-only access to the target or with softmax probabilities for KD. ASR is generally stable across KD weights, with the best results near $\alpha = 0.25$.

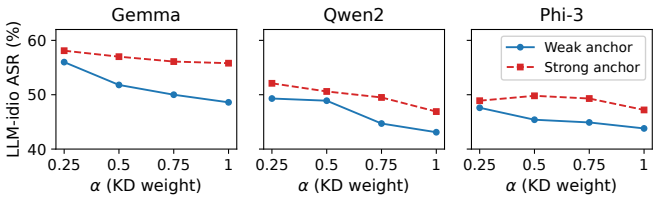

Figure 3: **Knowledge-distillation ablation.** LLM-idio ASR (%) across KD weights α for weak- and strong-model anchors in each family.

Impact of RAFT. Appendix B.3 compares fine-tuning with and without RAFT-selected queries. Mixing 50% RAFT queries into the training data improves ASR in all six combinations of three model families and two auditors (LLM-idio and MET). These results support using surrogate feedback to select queries that exploit fingerprint classifiers' weaknesses.

LoRA Rank and Training Cost. Appendix B.4 reports the ASR–utility trade-off across eight LoRA ranks. Appendix B.5 reports parameter counts, adapter sizes, and measured adapter-training time on a single NVIDIA RTX A6000.

6.4 Continual Fingerprint Spoofing

We evaluate continual fingerprint spoofing with our mixture-of-LoRA-experts (MoLA) framework. We sequentially train one LoRA expert for each of *LLMmap*, *LLM-idio*, and *MET*, then evaluate combinations routed through a Mixture-of-Experts architecture.

Table 4 shows two main results. (i) MoLA retains high spoofing performance after combining experts. For Qwen2, the 3-way configuration even exceeds its single-fingerprint LLMmap ASR (81.0% versus 76.0%), showing adaptation to new fingerprints while retaining prior capabilities. (ii) MoLA also generalizes to fingerprints without their corresponding expert. On Qwen2 and Phi-3, the LLMmap + LLM-idio configuration achieves MET scores close to those of configurations that include the MET expert, suggesting cross-fingerprint knowledge transfer.

7 Conclusion

We introduce fingerprint spoofing and propose GhostPrint, a parameter-efficient attack combining surrogate modeling, reward-ranked fine-tuning, and knowledge distillation, extended with a Mixture-of-LoRA router to unify multiple auditors. GhostPrint consistently bypasses three auditors across three model families while preserving utility, exposing a critical vulnerability in current black-box LLM fingerprint methods.

AI USE STATEMENT

In this work, we used generative AI tools to assist in the writing of proofs and implement the method. Additionally, we used generative AI tools to create or edit software code, edit the research paper to improve readability, and propose keywords for the research paper. We have reviewed all AI-assisted work. All proof steps were double-checked by the authors and rewritten by the authors when necessary to improve clarity. The code was also double-checked by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

For the theoretical contributions, the authors themselves completed the linear case and proposed the lower- and upper-bound analysis framework. We had extensive discussions with GPT-6 Astra to help extend the analysis to attention and Transformer models. The authors themselves provided the ideas, statement formulations, assumptions, and general proof techniques, while AI assisted in filling in the proof details. To make the proofs easier to understand and verify, we collaborated with AI to provide supporting lemmas and step-by-step explanations. All the steps and statements are double-checked by at least one human authors.

ETHICS STATEMENT

While our attack exposes a vulnerability that could be misused, this is a research prototype intended to motivate robust fingerprinting and we do not foresee a potential negative societal impact.

REPRODUCIBILITY STATEMENT

Experimental settings and hyperparameters are detailed in Section 6.1 and Appendix A. We provide the implementation code in the supplementary material. Proofs of all mathematical statements are provided in Appendices E.

REFERENCES

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Xiaofan Bai, Pingyi Hu, Xiaojing Ma, Linchen Yu, Dongmei Zhang, Qi Zhang, and Bin Benjamin Zhu. Esf: Efficient sensitive fingerprinting for black-box tamper detection of large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 10477–10494, 2025.

Jiacheng Cai, Jiahao Yu, Yangguang Shao, Yuhang Wu, and Xinyu Xing. Utf: Under-trained tokens as fingerprints——a novel approach to llm identification. In *Proceedings of the The First Workshop on LLM Security (LLMSEC)*, pp. 1–6, 2025a.

Will Cai, Tianneng Shi, Xuandong Zhao, and Dawn Song. Are you getting what you pay for? auditing model substitution in llm apis. *arXiv preprint arXiv:2504.04715*, 2025b.

Yuhan Cao, Yu Wang, Sitong Liu, Miao Li, Yixin Tao, and Tianxing He. Pay for the second-best service: A game-theoretic approach against dishonest llm providers. In *Proceedings of the ACM Web Conference 2026*, pp. 249–260, 2026.

Lingjiao Chen, Matei Zaharia, and James Zou. FrugalGPT: How to use large language models while reducing cost and improving performance. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. Featured Certification.

Sixu Chen, Xiang Chen, Hongyao Yu, Jiaxin Hong, Hao Fang, Shuoyang Sun, Bin Chen, and Shu-Tao Xia. Prompt2fingerprint: Plug-and-play llm fingerprinting via text-to-weight generation. *arXiv preprint arXiv:2605.18474*, 2026.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. In *International Conference on Machine Learning*, pp. 8359–8388. PMLR, 2024.

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

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115, 2023.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 3029–3051, 2023.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. RAFT: Reward rAnked FineTuning for generative foundation model alignment. *Transactions on Machine Learning Research (TMLR)*, 2023. arXiv:2304.06767.

Wenqi Fan, Pangjing Wu, Yujuan Ding, Liangbo Ning, Shijie Wang, and Qing Li. Towards retrieval-augmented large language models: Data management and system design. In *2025 IEEE 41st International Conference on Data Engineering (ICDE)*, pp. 4509–4512. IEEE, 2025.

Zhengru Fang, Yu Guo, Fei Liu, Yuang Zhang, Yihang Tao, Senkang Hu, Wenbo Ding, and Yuguang Fang. Agent-centric visual reinforcement learning under dynamic perturbations. *arXiv preprint arXiv:2604.24661*, 2026.

Chongyang Gao, Kezhen Chen, Jinmeng Rao, Berrak Sisman, Frank Bayer, Daniel Macko, and Yixin Lao. MoLA: MoE LoRA with layer-wise expert allocation. In *Findings of NAACL 2025*, 2025a.

Irena Gao, Percy Liang, and Carlos Guestrin. Model equality testing: Which model is this API serving? In *The Thirteenth International Conference on Learning Representations*, 2025b.

Thibaud Gloaguen, Robin Staab, Nikola Jovanović, and Martin Vechev. LLM fingerprinting via semantically conditioned watermarks. In *The Fourteenth International Conference on Learning Representations*, 2026.

Gene H Golub and Charles F Van Loan. *Matrix computations*. JHU press, 2013.

Chenxi Gu, Chengsong Huang, Xiaoqing Zheng, Kai-Wei Chang, and Cho-Jui Hsieh. Watermarking pre-trained language models with backdooring. *arXiv preprint arXiv:2210.07543*, 2022.

Martin Gubri, Dennis Ulmer, Hwaran Lee, Sangdoo Yun, and Seong Joon Oh. Trap: Targeted random adversarial prompt honeypot for black-box identification. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 11496–11517, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

Xin He, Junxi Shen, Zhenheng Tang, Xiaowen Chu, Bo Li, Ivor W Tsang, and Yew-Soon Ong. Routemark: A fingerprint for intellectual property attribution in routing-based model merging. *arXiv preprint arXiv:2508.01784*, 2025.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. An empirical analysis of compute-optimal large language model training. *Advances in neural information processing systems*, 35:30016–30030, 2022.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Pingyi Hu, Xiaofan Bai, Xiaojing Ma, Chaoxiang He, Dongmei Zhang, and Bin Benjamin Zhu. Resf: Regularized-entropy-sensitive fingerprinting for black-box tamper detection of large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 4889–4903, 2025.

Jiani Huang, Shijie Wang, Liang-bo Ning, Wenqi Fan, Shuaiqiang Wang, Dawei Yin, and Qing Li. Towards next-generation recommender systems: A benchmark for personalized recommendation assistant with llms. In *The ACM International Conference on Web Search and Data Mining (WSDM)*, 2026.

Alan Julian Izenman. Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, 5(2):248–264, 1975. doi: 10.1016/0047-259X(75)90042-1.

Bowen Jiang, Zhuoqun Hao, Young Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, Lyle Ungar, Camillo Jose Taylor, and Dan Roth. Know me, respond to me: Benchmarking LLMs for dynamic user profiling and personalized responses at scale. In *Second Conference on Language Modeling*, 2025a. URL https://openreview.net/forum?id=6ox8XZGOqP.

Bowen Jiang, Yuan Yuan, Maohao Shen, Zhuoqun Hao, Zhangchen Xu, Zichen Chen, Ziyi Liu, Anvesh Rao Vijjini, Jiashu He, Hanchao Yu, et al. Personamem-v2: Towards personalized intelligence via learning implicit user personas and agentic memory. *arXiv preprint arXiv:2512.06688*, 2025b.

Heng Jin, Chaoyu Zhang, Shanghao Shi, Wenjing Lou, and Y Thomas Hou. Proflingo: A fingerprinting-based intellectual property protection scheme for large language models. In *2024 IEEE Conference on Communications and Network Security (CNS)*, pp. 1–9. IEEE, 2024.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Rohith Kuditipudi, Jing Huang, Sally Zhu, Diyi Yang, Christopher Potts, and Percy Liang. Blackbox model provenance via palimpsestic membership inference. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Kevin Kurian, Ethan Holland, and Sean Oesch. Attacks and defenses against llm fingerprinting. *arXiv preprint arXiv:2508.09021*, 2025.

Hongyu Li, Liang Ding, Meng Fang, and Dacheng Tao. Revisiting catastrophic forgetting in large language model tuning. In *Findings of the association for computational linguistics: EMNLP 2024*, pp. 4297–4308, 2024.

Peixuan Li, Pengzhou Cheng, Fangqi Li, Wei Du, Haodong Zhao, and Gongshen Liu. Plmmark: a secure and robust black-box watermarking framework for pre-trained language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37:12, pp. 14991–14999, 2023.

Xiuyu Li, Jinkai Zhang, Mingyang Yi, Yu Li, Longqiang Wang, Yue Wang, and Ju Fan. Ets: Energy-guided test-time scaling for training-free rl alignment. In *Forty-third International Conference on Machine Learning*, 2026a.

Yu Li, Mingyang Yi, Xiuyu Li, Ju Fan, Fuxin Jiang, Binbin Chen, Peng Li, Jie Song, and Tieying Zhang. Reasoning and tool-use compete in agentic rl: From quantifying interference to disentangled tuning. *arXiv preprint arXiv:2602.00994*, 2026b.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

Jiale Liu, Jiahao Zhang, and Suhang Wang. Exposing privacy risks in graph retrieval-augmented generation. In *ACL Findings*, 2026.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. In *Forty-first International Conference on Machine Learning*, 2024.

Sijia Luo, Xiaokang Zhang, Yuxuan Hu, Bohan Zhang, Ke Wang, Jinbo Su, Mengshu Sun, Lei Liang, and Jing Zhang. Sparse-rl: Breaking the memory wall in llm reinforcement learning via stable sparse rollouts. *arXiv preprint arXiv:2601.10079*, 2026.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

Hope Elizabeth McGovern, Rickard Stureborg, Yoshi Suhara, and Dimitris Alikaniotis. Your large language models are leaving fingerprints. In *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, pp. 85–95, 2025.

Anshul Nasery, Edoardo Contente, Alkin Kaz, Pramod Viswanath, and Sewoong Oh. Are robust llm fingerprints adversarially robust? *arXiv preprint arXiv:2509.26598*, 2025a.

Anshul Nasery, Jonathan Hayase, Creston Brooks, Peiyao Sheng, Himanshu Tyagi, Pramod Viswanath, and Sewoong Oh. Scalable fingerprinting of large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b.

Vinh Nguyen, Cuong Dang, Jiahao Zhang, Hoa Tran, Minh Tran, Trinh Chau, Thai Le, Lu Cheng, and Suhang Wang. Urag: A benchmark for uncertainty quantification in retrieval-augmented large language models. *arXiv preprint arXiv:2603.19281*, 2026.

Ivica Nikolic, Teodora Baluta, and Prateek Saxena. Model provenance testing for large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Xuying Ning, Katherine Tieu, Dongqi Fu, Tianxin Wei, Zihao Li, Yuanchen Bei, Jiaru Zou, Mengting Ai, Zhining Liu, Ting-Wei Li, et al. Code as agent harness. *arXiv preprint arXiv:2605.18747*, 2026.

Dario Pasquini, Evgenios M. Kornaropoulos, and Giuseppe Ateniese. LLMmap: Fingerprinting for large language models. In *USENIX Security Symposium*, 2025.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, dahai li, Zhiyuan Liu, and Maosong Sun. ToolLLM: Facilitating large language models to master 16000+ real-world APIs. In *The Twelfth International Conference on Learning Representations*, 2024.

Zhenzhen Ren, GuoBiao Li, Sheng Li, Zhenxing Qian, and Xinpeng Zhang. Cotsrf: Utilize chain of thought as stealthy and robust fingerprint of large language models. *arXiv preprint arXiv:2505.16785*, 2025.

Nikhil Sardana, Jacob Portes, Sasha Doubov, and Jonathan Frankle. Beyond chinchilla-optimal: Accounting for inference in language model scaling laws. In *International Conference on Machine Learning*, pp. 43445–43460. PMLR, 2024.

Shuo Shao, Yiming Li, Yu He, Hongwei Yao, Wenyuan Yang, Dacheng Tao, and Zhan Qin. Sok: Large language model copyright auditing via fingerprinting. *arXiv preprint arXiv:2508.19843*, 2025.

Shuo Shao, Yiming Li, Hongwei Yao, Yifei Chen, Yuchen Yang, and Zhan Qin. Reading between the lines: Towards reliable black-box llm fingerprinting via zeroth-order gradient estimation. In *Proceedings of the ACM Web Conference 2026*, pp. 2637–2648, 2026.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023.

Guoheng Sun, Ziyao Wang, Bowei Tian, Meng Liu, Zheyu Shen, Shwai He, Yexiao He, Wanghao Ye, Yiting Wang, and Ang Li. Coin: Counting the invisible reasoning tokens in commercial opaque llm apis. *arXiv preprint arXiv:2505.13778*, 2025a.

Guoheng Sun, Ziyao Wang, Xuandong Zhao, Bowei Tian, Zheyu Shen, Yexiao He, Jinming Xing, and Ang Li. Invisible tokens, visible bills: The urgent need to audit hidden operations in opaque llm services. *arXiv preprint arXiv:2505.18471*, 2025b.

Mingjie Sun, Yida Yin, Zhiqiu Xu, J Zico Kolter, and Zhuang Liu. Idiosyncrasies in large language models. In *International Conference on Machine Learning*, pp. 57854–57885. PMLR, 2025c.

Ling Tang, YueFeng Chen, Hui Xue, and Quanshi Zhang. Towards the resistance of neural network fingerprinting to fine-tuning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Yao Tong, Haonan Wang, Siquan Li, Kenji Kawaguchi, and Tianyang Hu. Seedprints: Fingerprints can even tell which seed your large language model was trained from. In *The Fourteenth International Conference on Learning Representations*, 2026.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Ander Artola Velasco, Stratis Tsirtsis, Nastaran Okati, and Manuel Gomez-Rodriguez. Is your llm overcharging you? tokenization, transparency, and incentives. *arXiv preprint arXiv:2505.21627*, 2025.

Ander Artola Velasco, Stratis Tsirtsis, and Manuel Gomez-Rodriguez. Auditing pay-per-token in large language models. In *AISTATS*, 2026.

Ziyao Wang, Guoheng Sun, Yexiao He, Zheyu Shen, Bowei Tian, and Ang Li. Predictive auditing of hidden tokens in llm apis via reasoning length estimation. *arXiv preprint arXiv:2508.00912*, 2025.

Zehao Wu, Yanjie Zhao, and Haoyu Wang. Tensorguard: Gradient-based model fingerprinting for llm similarity detection and family classification. In *2025 40th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pp. 445–456. IEEE, 2025.

Zhaomin Wu, Haodong Zhao, Ziyang Wang, Jizhou Guo, Qian Wang, and Bingsheng He. LLM DNA: Tracing model evolution via functional representations. In *The Fourteenth International Conference on Learning Representations*, 2026a.

Zongyu Wu, Minhua Lin, Zhiwei Zhang, Fali Wang, Xianren Zhang, Xiang Zhang, and Suhang Wang. Image corruption-inspired membership inference attacks against large vision-language models. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7945–7957, 2026b.

Zixun Xiong, Gaoyi Wu, Qingyang Yu, Mingyu Derek Ma, Lingfeng Yao, Miao Pan, Xiaojiang Du, and Hao Wang. iseal: Encrypted fingerprinting for reliable llm ownership verification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40:42, pp. 35940–35948, 2026.

Jiashu Xu, Fei Wang, Mingyu Ma, Pang Wei Koh, Chaowei Xiao, and Muhao Chen. Instructional fingerprinting of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3277–3306, 2024.

Yijie Xu, Aiwei Liu, Xuming Hu, Lijie Wen, and Hui Xiong. Mark your llm: Detecting the misuse of open-source large language models via watermarking. *arXiv preprint arXiv:2503.04636*, 2025a.

Zhenhua Xu, Zhaokun Yan, Binhan Xu, Xin Tong, Haitao Xu, Yourong Chen, and Meng Han. Unlocking the effectiveness of lora-fp for seamless transfer implantation of fingerprints in downstream models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 4302–4312, 2025b.

Zhenhua Xu, Xixiang Zhao, Xubin Yue, Shengwei Tian, Changting Lin, and Meng Han. Ctcc: A robust and stealthy fingerprinting framework for large language models via cross-turn contextual correlation backdoor. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6978–7000, 2025c.

Shojiro Yamabe, Futa Kai Waseda, Tsubasa Takahashi, and Koki Wataoka. Mergeprint: Merge-resistant fingerprints for robust black-box ownership verification of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6894–6916, 2025.

Yuliang Yan, Haochun Tang, Shuo Yan, and Enyan Dai. Duffin: A dual-level fingerprinting framework for llms ip protection. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 5168–5184, 2026.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

Shuhua Yang, Jiahao Zhang, Yilong Wang, Dongwon Lee, and Suhang Wang. Query-efficient agentic graph extraction attacks on GraphRAG systems. In *ACL*, 2026.

Ziqing Yang, Yixin Wu, Yun Shen, Wei Dai, Michael Backes, and Yang Zhang. The challenge of identifying the origin of black-box large language models. *arXiv preprint arXiv:2503.04332*, 2025.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, 2023.

Chuanchao Zang, Xiangtao Meng, Wenyu Chen, Tianshuo Cong, Zha Yaxing, Dong Qi, Zheng Li, and Shanqing Guo. Errortrace: A black-box traceability mechanism based on model family error space. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Boyi Zeng, Lizheng Wang, Yuncong Hu, Yi Xu, Chenghu Zhou, Xinbing Wang, Yu Yu, and Zhouhan Lin. Huref: Human-readable fingerprint for large language models. *Advances in Neural Information Processing Systems*, 37:126332–126362, 2024.

Boyu Zhang, Ping He, Tianyu Du, Xuhong Zhang, Lei Yun, Kingsum Chow, and Jianwei Yin. Clmtracing: Black-box user-level watermarking for code language model tracing. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 28962–28978, 2025a.

Haobo Zhang, Zhenhua Xu, Junxian Li, Shangfeng Sheng, Dezhang Kong, and Meng Han. Attndiff: Attention-based differential fingerprinting for large language models. In *ACL*, 2026a.

Jie Zhang, Jiayuan Li, Haiqiang Fei, Lun Li, and Hongsong Zhu. Easydetector: Using linear probe to detect the provenance of large language models. In *2024 IEEE 23rd International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, pp. 2410–2417. IEEE, 2024.

Jie Zhang, Dongrui Liu, Chen Qian, Linfeng Zhang, Yong Liu, Yu Qiao, and Jing Shao. REEF: Representation encoding fingerprints for large language models. In *The Thirteenth International Conference on Learning Representations*, 2025b.

Qingjie Zhang, Di Wang, Haoting Qian, Liu Yan, Tianwei Zhang, Ke Xu, Qi Li, Minlie Huang, Hewu Li, and Han Qiu. Speculating llms' chinese training data pollution from their tokens. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 26124–26144, 2025c.

Xianren Zhang, Hui Liu, Delvin Ce Zhang, Xianfeng Tang, Qi He, Dongwon Lee, and Suhang Wang. Sua: Stealthy multimodal large language model unlearning attack. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 11241–11254, 2025d.

Yage Zhang, Yukun Jiang, Zeyuan Chen, Michael Backes, Xinyue Shen, and Yang Zhang. Real money, fake models: Deceptive model claims in shadow APIs. *arXiv preprint arXiv:2603.01919*, 2026b.

Yuanhe Zhang, Xinyue Wang, Zhican Chen, Weiliu Wang, Zilu Zhang, Zhengshuo Gong, Zhenhong Zhou, Kun Wang, Li Sun, Yang Liu, et al. Resource consumption threats in large language models. *arXiv preprint arXiv:2603.16068*, 2026c.

Yue Zheng, Si Wang, and Chip-Hong Chang. A dnn fingerprint for non-repudiable model ownership identification and piracy detection. *IEEE Transactions on Information Forensics and Security*, 17: 2977–2989, 2022.

Xiaoyuan Zhu, Yaowen Ye, Tianyi Alex Qiu, Hanlin Zhu, Sijun Tan, Ajraf Mannan, Jonathan Michala, Raluca Ada Popa, and Willie Neiswanger. Auditing black-box LLM APIs with a rank-based uniformity test. In *The Fourteenth International Conference on Learning Representations*, 2026.

Appendix

LIST OF CONTENTS

In this appendix, we provide the following additional information:

- **Section A**: Implementation Details of Experiments.
- **Section B**: Additional Experimental Results.
- **Section C**: Additional Related Works.
- **Section D**: Formal Main Results.
- **Section E**: Detailed Proofs.

A IMPLEMENTATION DETAILS OF EXPERIMENTS

A.1 FINGERPRINTING AUDITORS

For LLMmap, we use the released set of 800 adversarial identity probes, each concatenated with a system prompt and CoT/RAG-style prompts as in the original work. The 800 probes are split 8:2 (640 training, 160 test), and we use 100 probes to evaluate ASR.

For LLM-idio, we follow the original setting: a per-family binary classifier is trained on 10,000 UltraChat (Ding et al., 2023) responses from each of the small and large models. The prompt collection is split into disjoint subsets for training the attack adapters and evaluation; a held-out 20% subset (2,000 prompts) serves as the test set.

For MET, the two-sample test is applied to completions from the UltraChat prompt set. For the main-table fair-budget setting we use $N = 1$ prompt and $K = 10$ completions per prompt, subsampled from the full $N = 50, K = 20$ collection. The test statistic is $\mathrm{MMD}_{\mathrm{hamming}}$ with a permutation p-value (100 permutations, $\alpha = 0.05$). ASR is the fraction of 30 independent random seeds for which MET fails to reject H_0. The result with full (N, K) parameter grid is described in Appendix A.4.

A.2 BASELINE SETTINGS

For Query Detection, we train a provider-side classifier that decides at audit time whether to serve the small or the large model. The classifier is a two-layer MLP on top of frozen `all-MiniLM-L6-v2` sentence embeddings, trained to distinguish LLMmap identity probes from UltraChat prompts. At evaluation, each query is routed to the large model's pre-collected response if flagged as a probe and to the small model's response otherwise. This is an oracle baseline because the provider knows the probe distribution; utility is unchanged on prompts the detector rejects.

For Rewriting (Kurian et al., 2025), the small model first answers the query, then is re-prompted with a fixed style-laundering template that asks it to rewrite the response "to be more comprehensive, detailed, and well-reasoned, as if written by a highly knowledgeable assistant with deep expertise." Both passes use the same small model; only the second pass output is audited. Sample sizes per auditor match the Original (small) row.

For 5-shot ICL, we sample five question-answer pairs from the large model's UltraChat responses and prepend them as an in-context demonstration block before every query that the small model answers. For LLMmap, the same five pairs are prepended to every identity probe before evaluation; for LLM-idio and MET, they are prepended to every UltraChat prompt.

For Token Suppression (Nasery et al., 2025a), we follow the SuppressTop-k attack. At decoding time, for the first $n = 16$ generation positions we mask the $k = 4$ highest-probability tokens (setting their logits to $-\infty$) before sampling; subsequent positions decode normally. The mask is implemented as a HuggingFace `LogitsProcessor` and applied uniformly across all three auditors, using the same temperature, prompt set, and sample count as Original (small).

A.3 Training Settings

We adopt the LoRA-based deception attack from Section 5, with $r = 16$, LoRA scaling parameter 32, dropout 0.05, 2 epochs, learning rate 5×10^{-5}, batch size 4, AdamW with weight decay 0.01, and a knowledge-distillation term ($\alpha = 0.5$, $\tau = 2.0$) against a frozen large teacher. We set $\alpha = 0.5$ as an intermediate value in the parameter grid in Figure 3, which shows the dependence of ASR on α across $[0.25, 1.0]$. The RAFT-style ablation of Section 6.3 uses $N = 8$ candidates per prompt, 3 rounds, and $\lambda_{\text{raft}} = 0.5$.

A.4 MET Sample Complexity

LLMmap and LLM-idio use one query per audit; MET defaults to $(N = 20, K = 5)$, costing $100\times$ more generations. The main paper therefore reports MET at the fair budget $(1, 10)$. Here we characterize detection power by subsampling from a $(50, 20)$ pool, computing ASR over 30 seeds.

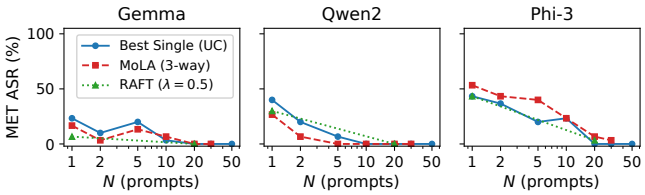

Figure 4: **MET sample complexity at $K = 10$ for three spoofing attacks across model families**. ASR is averaged over 30 resampling seeds.

At the strong budget on Phi-3, our MoLA adapter achieves a mean permutation p-value of 0.086. Its ASR stays above 50% until $N > 30$, while the best single-fingerprint adapter drops below 50% at $N = 20$. On other families, curves overlap, consistent with family-dependent trade-offs.

Even though all adapters are rejected at the full $(50, 20)$ budget, that setting costs 1,000 generations per audit and is impractical for frequent use. Auditors operating at lower budgets, including the fair $(1, 10)$ and the intermediate $(20, 5)$ default, leave many attacks undetected, especially on Phi-3.

B Additional Experimental Results

B.1 Per-adapter Single-Fingerprint Breakdown

Table 5 reports the six single-fingerprint LoRA variants trained per family (18 variants in total), organised by training data (LLMmap probes vs. UltraChat) and KD teacher (no KD, student-size self-distillation, target-size teacher). The "GhostPrint (ours)" rows in Table 1 pick the maximum for each auditor and model family across the six corresponding variants.

B.2 More Cross-Model Spoofing Results

For completeness, we provide additional results to investigate whether models of similar scale can spoof fingerprints across model families. As shown in Table 6, our proposed GhostPrint framework consistently outperforms all baselines. Notably, our results on LLM-idio across both settings are particularly notable, showing almost no performance difference compared to the target model's original fingerprint.

B.3 Impact of RAFT

To ensure that fingerprint spoofing fine-tuning effectively exploits the vulnerabilities of weak classifiers, we utilize RAFT to select queries that are more adversarial to fingerprint classifiers. The results are summarized in Table 7. As shown in the table, incorporating a 50% ratio of RAFT queries ($\lambda_{\text{raft}} = 0.5$) during fine-tuning alongside standard queries improves ASR in all six reported comparisons. This demonstrates that strategically using weak classifier vulnerabilities directly enhances spoofing capabilities, validating our theoretical analysis in Section 5.1.

Table 5: **Detailed evaluation across all 18 LoRA variants.** ASR (%) by training data and KD teacher. Weak and Target refer to the model sizes in each column group; "–" means no KD. Best results within each model family and column are bolded.

Training data	KD teacher	Gemma-1.1 2B → 7B			Qwen2 1.5B → 7B			Phi-3 mini → medium		
		LLMmap	LLM-idio	MET	LLMmap	LLM-idio	MET	LLMmap	LLM-idio	MET
LLMmap	–	**95.0**	31.2	13.3	74.0	70.0	26.7	**80.0**	**64.0**	26.7
LLMmap	Weak	84.0	14.7	3.3	**76.0**	41.5	30.0	77.0	35.9	43.3
LLMmap	Target	87.0	16.8	0.0	**76.0**	**71.3**	36.7	74.0	35.2	33.3
UltraChat	–	19.0	**58.5**	**23.3**	21.0	50.2	**40.0**	10.0	58.0	33.3
UltraChat	Weak	21.0	52.0	13.3	28.0	51.0	36.7	5.0	53.0	36.7
UltraChat	Target	28.0	55.9	13.3	35.0	51.7	**40.0**	0.0	56.0	**60.0**

Table 6: **Evaluation of cross-family weak-to-weak spoofing**. The table reports the Attack Success Rate (ASR, %) when adapting a weak model from one architectural family to spoof a weak model from a completely different family. The highest ASR among attack methods is bolded.

Method	Gemma-2B → Qwen2-1.5B			Qwen2-1.5B → Gemma-2B		
	LLMmap	LLM-idio	MET	LLMmap	LLM-idio	MET
Original (weak)	0.0	1.6	0.0	0.0	0.8	0.0
Query Detection	0.0	3.2	6.7	0.0	2.5	6.7
Rewriting	0.0	0.0	3.3	0.0	8.9	0.0
5-shot ICL	0.0	0.8	0.0	0.0	19.4	0.0
Token Suppression	1.0	2.2	16.7	16.0	1.0	3.3
GhostPrint (ours)	**4.0**	**97.5**	**46.7**	**38.0**	**97.9**	**20.0**
Original (target)	78.0	99.5	100.0	84.0	99.1	93.3

Table 7: **RAFT-guided query selection.** ASR (%) at two RAFT proportions λ_{raft}. Best results are bolded.

Auditor	Gemma-1.1 2B → 7B		Qwen2 1.5B → 7B		Phi-3 mini → medium	
	$\lambda_{\text{raft}} = 0$	0.5	$\lambda_{\text{raft}} = 0$	0.5	$\lambda_{\text{raft}} = 0$	0.5
LLM-idio	54.7	**58.8**	56.3	**63.4**	49.1	**54.8**
MET	6.7	**10.0**	36.7	**40.0**	40.0	**43.3**

B.4 LoRA Rank Study

We evaluate LoRA ranks $r \in \{1, 2, 4, 8, 16, 32, 48, 64\}$ on Gemma-1.1-2B-IT → Gemma-1.1-7B-IT, Qwen2-1.5B-Instruct → Qwen2-7B-Instruct, and Phi-3-Mini-4K-Instruct → Phi-3-Medium-4K-Instruct. Tables 8 and 9 report the ASR and utility of this additional rank study, reported separately from the main results in Tables 1 and 2.

Higher rank does not uniformly improve ASR. For example, Qwen2 MET ASR increases from 36.7% at $r = 16$ to 53.3% at $r = 48$, then decreases to 43.3% at $r = 64$. The utility results also vary with rank, illustrating the need to consider both spoofing performance and downstream utility when selecting the adaptation rank.

B.5 Adapter Training Cost

We measure the combined training time of two independent adapters: one trained on the LLMmap training split and one trained on the UltraChat training split for LLM-idio and MET. All experiments

Table 8: **ASR across LoRA ranks.** Attack success rates (%) for the additional rank study.

Rank	Gemma-1.1 2B → 7B			Qwen2 1.5B → 7B			Phi-3 mini → medium		
	LLMmap	LLM-idio	MET	LLMmap	LLM-idio	MET	LLMmap	LLM-idio	MET
1	55.0	31.2	10.0	50.0	53.6	30.0	65.0	51.0	43.3
2	65.0	35.4	13.3	50.0	54.9	30.0	60.0	51.9	56.7
4	80.0	45.9	10.0	60.0	59.1	30.0	70.0	46.6	66.7
8	70.0	54.1	16.7	60.0	57.1	36.7	60.0	60.6	33.3
16	85.0	61.2	13.3	60.0	50.3	36.7	60.0	52.9	63.3
32	85.0	64.4	6.7	70.0	58.4	43.3	70.0	53.8	53.3
48	70.0	66.3	23.3	65.0	63.6	53.3	70.0	58.2	40.0
64	80.0	67.4	30.0	70.0	52.9	43.3	75.0	54.8	50.0

Table 9: **Utility across LoRA ranks.** MMLU, GSM8K, and ARC-C scores for the additional rank study.

Rank	Gemma-1.1 2B → 7B			Qwen2 1.5B → 7B			Phi-3 mini → medium		
	MMLU	GSM8K	ARC-C	MMLU	GSM8K	ARC-C	MMLU	GSM8K	ARC-C
1	37.87	7.20	43.60	55.40	55.34	45.22	69.69	75.74	62.54
2	38.37	6.22	43.52	55.64	56.18	45.31	69.67	76.88	63.05
4	38.19	7.81	42.24	55.33	56.79	44.80	69.61	73.46	62.37
8	38.33	7.96	43.09	55.31	56.63	44.80	69.57	78.17	63.40
16	38.63	7.51	43.09	55.47	57.16	45.56	69.83	77.63	63.65
32	38.65	9.10	43.60	55.06	54.74	44.54	69.89	77.03	62.88
48	38.97	9.70	44.11	55.70	56.10	44.54	69.85	77.79	63.40
64	38.43	8.19	43.94	55.13	55.65	44.97	69.66	78.32	63.31

use a single NVIDIA RTX A6000 GPU. Table 10 reports these timings alongside the trainable-parameter counts, base-model percentages, and adapter sizes from the rank study.

Across the evaluated ranks, the combined training time is 52.6–53.3 GPU-min for Qwen2, 67.4–68.5 GPU-min for Gemma, and 101.2–111.2 GPU-min for Phi-3; each pair of adapter training runs completes in less than two GPU-hours. The reported trainable-parameter fractions remain below 1%. These measurements quantify adapter-training compute; they do not establish the total cost of data acquisition, target-model queries, or deployment.

C Additional Related Works

C.1 LLM API Overcharging Risks

The proliferation of commercial LLM APIs under pay-per-use pricing has raised concerns about provider dishonesty at multiple levels. At the model level, providers may silently substitute promised strong models with weaker alternatives (Cai et al., 2025b; Zhang et al., 2026b), directly misrepresenting service quality. At the token level, hidden reasoning tokens in opaque APIs introduce invisible billing overhead (Sun et al., 2025a;b), with tokenization transparency and incentive misalignment further analyzed by Velasco et al. (2025). Broader resource-side threats, such as compute, memory, and API abuse, are surveyed by Zhang et al. (2026c). Early defenses address both directions: a game-theoretic auditing mechanism targets dishonest model-level providers (Cao et al., 2026), while statistical and predictive auditing schemes tackle pay-per-token billing fraud (Velasco et al., 2026; Wang et al., 2025). Our work identifies a new threat distinct from previous model substitution and token overcharging, where a malicious provider can use parameter-efficient fine-tuning to adapt a

Table 10: **Adapter resources and training time.** Trainable parameters, base-model percentage, adapter size (MB), and training time (GPU-min). Time sums the LLMmap and UltraChat adapter training runs on one NVIDIA RTX A6000.

	Gemma-1.1 2B → 7B				Qwen2 1.5B → 7B				Phi-3 mini → medium			
Rank	Trainable params	Base (%)	Adapter (MB)	GPU-min	Trainable params	Base (%)	Adapter (MB)	GPU-min	Trainable params	Base (%)	Adapter (MB)	GPU-min
1	115,200	0.005	0.47	67.4	136,192	0.009	0.56	52.6	589,824	0.015	2.38	109.2
2	230,400	0.009	0.93	67.6	272,384	0.018	1.10	52.9	1,179,648	0.031	4.74	109.9
4	460,800	0.018	1.85	67.6	544,768	0.035	2.19	52.8	2,359,296	0.062	9.45	109.6
8	921,600	0.037	3.70	67.6	1,089,536	0.071	4.37	52.9	4,718,592	0.123	18.89	109.3
16	1,843,200	0.073	7.38	68.2	2,179,072	0.141	8.73	53.1	9,437,184	0.246	37.77	109.9
32	3,686,400	0.147	14.76	67.8	4,358,144	0.282	17.45	52.9	18,874,368	0.492	75.51	101.2
48	5,529,600	0.220	22.13	68.5	6,537,216	0.422	26.16	53.1	28,311,552	0.735	113.26	110.8
64	7,372,800	0.293	29.50	68.0	8,716,288	0.561	34.88	53.3	37,748,736	0.978	151.01	111.2

weak model to evade fingerprint-based auditing, making dishonest model deployments harder to detect.

C.2 LLM FINGERPRINTING

White-box Non-Intrusive Fingerprinting. White-box fingerprinting assumes direct weight access and verifies ownership through internal model analysis without modifying model behavior. Parameter and gradient signatures form one primary axis: Zheng et al. (2022) establish non-repudiable ownership certificates from weight statistics, and TensorGuard (Wu et al., 2025) classifies model families via gradient-based features. Representation-based approaches extract fingerprints from internal activations: REEF (Zhang et al., 2025b) encodes layer-wise hidden states as compact fingerprints, AttnDiff (Zhang et al., 2026a) exploits differential attention maps for identification, and EasyDetector (Zhang et al., 2024) detects provenance via linear probes on intermediate features. HuRef (Zeng et al., 2024) constructs human-interpretable fingerprints from weight-invariant vectors, and Shao et al. (2025) provide a systematic survey of LLM copyright auditing spanning the full white-box landscape.

Non-Intrusive Black-Box Fingerprinting. Black-box fingerprinting operates through API queries alone, without parameter access. *General-query methods* derive model identity from naturally elicited responses. For API integrity auditing, statistical hypothesis tests detect model substitution by comparing output distributions (Gao et al., 2025b; Zhu et al., 2026) and behavioral consistency (Cai et al., 2025b; Nikolic et al., 2025). Idiosyncrasy-based approaches exploit persistent behavioral quirks (Sun et al., 2025c; McGovern et al., 2025), characteristic error-space signatures (Zang et al., 2025), and side-channel query statistics (Pasquini et al., 2025) as inherent fingerprints. *Specific-query methods* craft targeted inputs to elicit discriminative signals. Adversarial honeypot strategies embed verifiable traps for IP attribution (Gubri et al., 2024; Jin et al., 2024). Token-level probing exploits under-trained tokens (Cai et al., 2025a) and training corpus artifacts (Zhang et al., 2025c) as covert signals. Further directions include membership-inference-based provenance (Kuditipudi et al., 2025), chain-of-thought fingerprinting (Ren et al., 2025), zeroth-order gradient estimation (Shao et al., 2026), sensitivity-guided query selection (Hu et al., 2025), code model tracing (Zhang et al., 2025a), and dual-level IP protection frameworks (Yan et al., 2026).

Intrusive Fingerprinting. Intrusive methods embed verifiable fingerprints during model training via backdoor mechanisms. Early work establishes backdoor watermarking baselines for pre-trained language models (Gu et al., 2022; Li et al., 2023). Instruction-triggered backdoors bind ownership responses to specific prompt patterns (Xu et al., 2024; 2025c), and LoRA-based transfer propagates fingerprints into fine-tuned derivatives (Xu et al., 2025b). Robustness under post-hoc model modification is a central challenge: merge-resistant designs (Yamabe et al., 2025; He et al., 2025) and fine-tuning-resistant schemes (Tang et al., 2025; Nasery et al., 2025b) address editing and adaptation threats. Semantically richer fingerprints are pursued via conditioned watermarks (Gloaguen et al.,

2026), random-seed fingerprinting (Tong et al., 2026), and functional evolution tracing (Wu et al., 2026a). Additional contributions include encrypted ownership verification (Xiong et al., 2026), plug-and-play fingerprint generation (Chen et al., 2026), open-source misuse detection (Xu et al., 2025a), tamper attribution (Bai et al., 2025), and analysis of the fundamental difficulty of black-box origin identification (Yang et al., 2025).

D Formal Main Results

This section states the formal results supporting Section 5.1. Their proofs are in Section E.5. Section E supplies the model-specific definitions and auxiliary results. The lower bound and restricted-query upper bound use the update families in Definition D.1. The low-rank-verification result permits an additional value or output-head update with fixed candidate features.

D.1 Lower Bound and Restricted-Query Upper Bound

Definition D.1 (Models and updates)**.** *Let $d, q \geq 1$ and $r \geq 0$ be integers, and let $\rho > 0$. Define $(f_*, \mathcal{F}_r, \mathcal{X})$ by one of the following cases.*

(i) ***Linear.*** *For $\mathbf{W}_*, \mathbf{W}_0 \in \mathbb{R}^{q\times d}$, define $\mathbf{R} := \mathbf{W}_* - \mathbf{W}_0$ and*

$$\begin{aligned} f_*(\mathbf{x}) &:= \mathbf{W}_*\mathbf{x}, \\ \mathcal{F}_r &:= \{\mathbf{x} \mapsto (\mathbf{W}_0 + \mathbf{\Delta})\mathbf{x} : \mathbf{\Delta} \in \mathbb{R}^{q\times d},\ \mathrm{rank}(\mathbf{\Delta}) \leq r\}, \\ \mathcal{X} &:= \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 \leq \rho\}. \end{aligned}$$

(ii) ***Causal attention.*** *Under Definitions E.10 and E.11, define*

$$\begin{aligned} f_*(\mathbf{X}) &:= \phi_{\mathbf{S}_*}(\mathbf{X})^\top, \\ \mathcal{F}_r &:= \{\mathbf{X} \mapsto \phi_{\mathbf{S}_\Delta}(\mathbf{X})^\top : (\Delta\mathbf{Q}, \Delta\mathbf{K}) \in \mathcal{A}_{r,M}\}, \\ \mathcal{X} &:= \{\mathbf{X} \in \mathbb{R}^{L\times d} : \max_i \|\mathbf{x}_i\|_2 \leq \rho\}. \end{aligned}$$

(iii) ***Transformer.*** *Under the single-head update family in Definition E.39, define*

$$\begin{aligned} f_*(\mathbf{X}) &:= F_{\boldsymbol{\theta}_*,\mathbf{W}}(\mathbf{X}), \\ \mathcal{F}_r &:= \{\mathbf{X} \mapsto F_{\boldsymbol{\theta}_\Delta,\mathbf{W}}(\mathbf{X}) : \Delta \in \mathcal{A}^{\mathrm{core}}_{r,M}\}. \end{aligned}$$

Let $\mathcal{X}$ contain the support of every input law used in the corresponding statement.

All three outputs are column vectors. For causal attention, f_* is the transpose of the last row of the standard attention output in Section 5.1, with $d_k = h$. The causal mask excludes no position from that last row.

Definition D.2 (Error norm and constants)**.** *For a measurable map $u : \mathcal{X} \to \mathbb{R}^q$ and an input law $\mathcal{D}$, define*

$$\|u\|_{L^2(\mathcal{D})} := \left(\mathbb{E}_{x\sim\mathcal{D}}\|u(x)\|_2^2\right)^{1/2}.$$

For integers $L \geq 2$, $k, h \geq 1$, $\rho > 0$, and a matrix $\mathbf{B}$, define

$$c_{L,k} := \frac{(L-1)^3}{L^4(k+2)^2(k+4)}, \qquad a_{L,k}(\mathbf{B}) := \frac{\rho^3\sqrt{c_{L,k}}}{2\sqrt{h}}\|\mathbf{B}\|_F.$$

The following proposition formalizes the lower-bound summary in Section 5.1. Its proof is in Section E.5.1.

Proposition D.3 (Lower bound for universal spoofing)**.** *Use Definitions D.1 and D.2. For attention, assume $\rho^2 M/\sqrt{h} \leq \sqrt{c_{L,d}}/12$. For Transformers, assume Definition E.40, including its input*

distribution and output lower-bound condition, and $\rho^2 M/\sqrt{h} \le \sqrt{c_{L,k}}/12$. *Then we have*

$$\inf_{f\in\mathcal{F}_r}\sup_{x\in\mathcal{X}}\|f_*(x)-f(x)\|_2 \ge \begin{cases}\rho\,\sigma_{r+1}(\mathbf{R}), & \textit{linear},\\ a_{L,d}(\mathbf{V})\left(\sum_{j>2r}\sigma_j(\mathbf{S}_*-\mathbf{S}_0)^2\right)^{1/2}, & \textit{attention},\\ a_{\mathrm{obs}}a_{L,k}(\bar{\mathbf{V}})\left(\sum_{j>2r}\sigma_j(\bar{\mathbf{S}}_*-\bar{\mathbf{S}}_0)^2\right)^{1/2}, & \textit{Transformer}.\end{cases}$$

Equality holds in the linear case.

The lower bound is positive if $\mathrm{rank}(\mathbf{R}) > r$ in the linear case, if $\mathrm{rank}(\mathbf{S}_* - \mathbf{S}_0) > 2r$ and $\mathbf{V} \neq \mathbf{0}$ in the attention case, or if $\mathrm{rank}(\bar{\mathbf{S}}_* - \bar{\mathbf{S}}_0) > 2r$ and $\bar{\mathbf{V}} \neq \mathbf{0}$ in the Transformer case, under the respective hypotheses above. The barred matrices and $a_{\mathrm{obs}} > 0$ are defined in Definition E.40. These are worst-case bounds. The input distribution used in the proof need not be the auditor's query distribution.

The next theorem is the formal version of Theorem 5.1. Its proof is in Section E.5.2.

Theorem D.4 (Upper bound under restricted queries)**.** *Use Definition D.1 and an audit law* $\mathcal{D}$ *supported on* $\mathcal{X}$. *Assume the following second moments are finite and define*

$$\mathbf{\Sigma} := \begin{cases}\mathbb{E}[\mathbf{x}\mathbf{x}^\top], & \textit{linear},\\ \mathbb{E}[\mathbf{x}_i\mathbf{x}_i^\top] \quad \textit{for every } i\in[L], & \textit{attention},\\ \mathbb{E}[\mathbf{Z}(\mathbf{X})^\top\mathbf{Z}(\mathbf{X})], & \textit{Transformer}.\end{cases}$$

For attention, assume the token second moments are equal. For attention and Transformers, assume the projection updates satisfy the score-norm constraints in Theorems E.28 and E.50, respectively. Let $\lambda_j(\mathbf{\Sigma})$ *be the eigenvalues in decreasing order, and define*

$$b := \begin{cases}\|\mathbf{R}\|_2, & \textit{linear},\\ \dfrac{2\rho^2 M\|\mathbf{V}\|_2}{\sqrt{h}}\left(1+e^{\rho^2 M/\sqrt{h}}\right), & \textit{attention},\\ K_D A_r, & \textit{Transformer},\end{cases}$$

where $\mathbf{Z}$, K_D, *and* A_r *are specified in Theorem E.50. Then we have*

$$\inf_{f\in\mathcal{F}_r}\|f_*-f\|_{L^2(\mathcal{D})} \le b\left(\sum_{j>r}\lambda_j(\mathbf{\Sigma})\right)^{1/2}.$$

Here $\mathbf{Z}$ is the shared normalized input to the adapted Transformer head. The theorem permits dependent tokens and does not require a uniform-ball audit law. A small right-hand side guarantees a small population approximation error on the audit distribution. The score-norm assumptions concern the constructed updates, not only the target and base parameters.

D.2 Upper Bound under Low-Rank Verification

Definition D.5 (Models with fixed features)**.** *Let* $d, q, c \ge 1$ *and* $s \ge 0$ *be integers. Fix an audit law* $\mathcal{D}$ *on* $\mathcal{X}$, *measurable maps* $f_* : \mathcal{X} \to \mathbb{R}^q$ *and* $h_0 : \mathcal{X} \to \mathbb{R}^d$, *and matrices* $\mathbf{W}_0 \in \mathbb{R}^{q\times d}$ *and* $\mathbf{C} \in \mathbb{R}^{q\times c}$. *Assume* $\mathbb{E}\|f_*\|_2^2 + \mathbb{E}\|h_0\|_2^2 < \infty$. *The observation map is* $\mathbf{y} \mapsto \mathbf{C}^\top\mathbf{y}$. *Define the three model instances by*

$$(h_0, \mathbf{W}_0, f_*) := \begin{cases}(\mathbf{x}, \mathbf{W}_0, \mathbf{W}_*\mathbf{x}), & \textit{linear},\\ (\mathbf{X}^\top\mathbf{p}_{\mathbf{S}_0}(\mathbf{X}), \mathbf{V}_0^\top, \mathbf{V}_*^\top\mathbf{X}^\top\mathbf{p}_{\mathbf{S}_*}(\mathbf{X})), & \textit{attention},\\ (h_{\boldsymbol{\theta}_0}(\mathbf{X}), \mathbf{W}_0, \mathbf{W}_* h_{\boldsymbol{\theta}_*}(\mathbf{X})), & \textit{Transformer}.\end{cases}$$

All maps and matrices are fixed before sampling $x \sim \mathcal{D}$.

The attention features are the weighted input before the value projection. The Transformer features are the final hidden vector before the output head. The candidate scores $\mathbf{S}_0$ or backbone $\boldsymbol{\theta}_0$ may

already incorporate a fixed admissible query/key update. The additional budget s applies to $\Delta\mathbf{V}^\top$ or an untied output-head update, respectively. The features stay fixed during this additional update. There is no separate update-norm constraint.

Definition D.6 (Prediction error)**.** *Under Definition D.5, define* $\mathbf{z}(x) := h_0(x)$ *and*

$$\varepsilon^2_{\mathbf{C}} := \inf_{\mathbf{B}\in\mathbb{R}^{c\times d}} \mathbb{E}_{x\sim\mathcal{D}}\|\mathbf{C}^\top f_*(x) - \mathbf{B}\mathbf{z}(x)\|_2^2.$$

The quantity $\varepsilon^2_{\mathbf{C}}$ is the best linear prediction error for the target output observed by the verifier. Taking $\mathcal{D} = \mathcal{D}_q$ gives the notation used in Theorem 5.2.

The next theorem is the formal version of Theorem 5.2. Its proof is in Section E.5.3.

Theorem D.7 (Upper bound under low-rank verification, formal version of Theorem 5.2)**.** *Under Definitions D.5 and D.6, assume* $s \geq \operatorname{rank}(\mathbf{C})$*. For* $\mathbf{\Delta} \in \mathbb{R}^{q\times d}$*, define* $f_{\mathbf{\Delta}}(x) := (\mathbf{W}_0 + \mathbf{\Delta})\mathbf{z}(x)$*. Then we have*

$$\inf_{\operatorname{rank}(\mathbf{\Delta})\leq s} \mathbb{E}_{x\sim\mathcal{D}}\|\mathbf{C}^\top(f_*(x) - f_{\mathbf{\Delta}}(x))\|_2^2 = \varepsilon^2_{\mathbf{C}}.$$

Moreover, $\varepsilon_{\mathbf{C}} = 0$ *if and only if there exists* $\mathbf{B} \in \mathbb{R}^{c\times d}$ *such that* $\mathbf{C}^\top f_*(x) = \mathbf{B}\mathbf{z}(x)$ *for* $\mathcal{D}$*-almost every* x*.*

This statement applies to every model instance in Definition D.5. The zero-residual condition holds for linear models, for attention with matched scores, and for Transformers with a shared backbone. Mismatched scores or backbones can leave a nonzero residual. It requires neither a uniform-ball audit law nor independence among tokens. For an arbitrary rank budget, Theorem E.33 gives the more general spectral-tail formula.

E Detailed Proofs

This section develops the proofs supporting the formal results in Section D. Section E.1 collects the common notation and matrix approximation tools. Section E.2 treats linear models, including the uniform-ball specialization. Sections E.3 and E.4 treat causal attention and Transformers. Section E.5 then derives the unified results and their main-text forms. Each result states its query assumptions and the parameters that may be updated.

E.1 Notation and Basic Tools

We use boldface for vectors and matrices. Scalar entries and dimensions, as well as model maps such as f and ϕ, are not bold. In particular, x_i denotes the ith entry of $\mathbf{x}$, and H_{ij} denotes the (i,j)th entry of $\mathbf{H}$. For a positive integer m, define $[m] := \{1,\dots,m\}$. For a vector $\mathbf{x} \in \mathbb{R}^m$, define $\|\mathbf{x}\|_2 := (\sum_{i=1}^m x_i^2)^{1/2}$. For a matrix $\mathbf{H} \in \mathbb{R}^{m\times n}$, define

$$\|\mathbf{H}\|_2 := \sup_{\|\mathbf{x}\|_2=1} \|\mathbf{H}\mathbf{x}\|_2, \qquad \|\mathbf{H}\|_F := \left(\sum_{i=1}^m\sum_{j=1}^n H_{ij}^2\right)^{1/2}.$$

The matrix 2-norm is also called the spectral norm. Let $\sigma_j(\mathbf{H})$ denote the j-th largest singular value of $\mathbf{H}$, with $\sigma_j(\mathbf{H}) := 0$ for $j > \min\{m,n\}$. An empty sum is zero. Let $\mathbf{I}_m$ denote the $m \times m$ identity matrix. For $i \in [m]$, let $\mathbf{e}_i \in \mathbb{R}^m$ denote the vector with entry one at coordinate i and zero at all other coordinates, with the ambient dimension specified by context. For $\mathbf{H} \in \mathbb{R}^{m\times n}$, its column space is $\{\mathbf{H}\mathbf{x} : \mathbf{x} \in \mathbb{R}^n\}$ and $\operatorname{rank}(\mathbf{H})$ is the dimension of this space. For vectors $\mathbf{x}_1,\dots,\mathbf{x}_k$, $\operatorname{span}\{\mathbf{x}_1,\dots,\mathbf{x}_k\}$ is their linear span. For $\mathbf{H} \in \mathbb{R}^{m\times n}$, define $\operatorname{col}(\mathbf{H}) := \{\mathbf{H}\mathbf{x} : \mathbf{x} \in \mathbb{R}^n\}$ and $\operatorname{null}(\mathbf{H}) := \{\mathbf{x} \in \mathbb{R}^n : \mathbf{H}\mathbf{x} = \mathbf{0}\}$. For a square matrix $\mathbf{H}$, define $\operatorname{tr}(\mathbf{H}) := \sum_i H_{ii}$. For symmetric matrices $\mathbf{H}_1, \mathbf{H}_2$ of the same size, $\mathbf{H}_1 \preceq \mathbf{H}_2$ means that $\mathbf{x}^\top(\mathbf{H}_2 - \mathbf{H}_1)\mathbf{x} \geq 0$ for every vector $\mathbf{x}$ of the appropriate dimension. An orthogonal projection is a square matrix $\mathbf{P}$ satisfying $\mathbf{P} = \mathbf{P}^\top = \mathbf{P}^2$. All model parameters and adaptation matrices are fixed before the random query is drawn. Every expectation below is over the explicitly stated query distribution.

For a random vector or matrix $\mathbf{U}$, define $\|\mathbf{U}\|_{L^2} := (\mathbb{E}\|\mathbf{U}\|^2)^{1/2}$. The inner norm is the Euclidean norm for vectors and the Frobenius norm for state matrices. A map is K-Lipschitz on a specified

domain if the distance between its outputs is at most K times the distance between its inputs on that domain.

We use the Eckart–Young approximation theorem (Golub & Van Loan, 2013) in the following rank-constrained form, with its factorization property proved explicitly.

Lemma E.1 (Low-rank matrix approximation)**.** *Let $\mathbf{H} \in \mathbb{R}^{m\times n}$ and let r be a nonnegative integer. Then we have*

$$\min_{\substack{\mathbf{Y}\in\mathbb{R}^{m\times n}\\ \operatorname{rank}(\mathbf{Y})\le r}} \|\mathbf{H}-\mathbf{Y}\|_2 = \sigma_{r+1}(\mathbf{H}),$$

$$\min_{\substack{\mathbf{Y}\in\mathbb{R}^{m\times n}\\ \operatorname{rank}(\mathbf{Y})\le r}} \|\mathbf{H}-\mathbf{Y}\|_F^2 = \sum_{j=r+1}^{\min\{m,n\}} \sigma_j(\mathbf{H})^2.$$

Both minima are attained by the truncated singular value decomposition of $\mathbf{H}$. Every feasible $\mathbf{Y}$ can be written as $\mathbf{Y} = \mathbf{AB}$ for matrices $\mathbf{A} \in \mathbb{R}^{m\times r}$ and $\mathbf{B} \in \mathbb{R}^{r\times n}$. For $r = 0$, the only feasible matrix is the zero matrix and $\mathbf{AB}$ denotes that matrix.

Proof. **Step 1. Construct a feasible approximation.** Define $s := \operatorname{rank}(\mathbf{H})$. If $r \ge s$, the choice $\mathbf{Y} := \mathbf{H}$ gives zero error and both claimed right-hand sides are zero. It remains to consider $0 \le r < s$. Choose a singular value decomposition

$$\mathbf{H} = \sum_{j=1}^{s} \sigma_j(\mathbf{H})\mathbf{u}_j\mathbf{v}_j^\top,$$

where $\mathbf{u}_j \in \mathbb{R}^m$ and $\mathbf{v}_j \in \mathbb{R}^n$ are orthonormal families. Define $\mathbf{H}_r := \sum_{j=1}^{r} \sigma_j(\mathbf{H})\mathbf{u}_j\mathbf{v}_j^\top$. The rank of $\mathbf{H}_r$ is at most r because it is a sum of r rank-one matrices. By orthonormality of the singular vectors,

$$\|\mathbf{H}-\mathbf{H}_r\|_2 = \sigma_{r+1}(\mathbf{H}),$$
$$\|\mathbf{H}-\mathbf{H}_r\|_F^2 = \sum_{j=r+1}^{s} \sigma_j(\mathbf{H})^2.$$

These identities give upper bounds on both minima.

Step 2. Prove the spectral-norm lower bound. Fix any $\mathbf{Y} \in \mathbb{R}^{m\times n}$ with $\operatorname{rank}(\mathbf{Y}) \le r$. The restriction of the map $\mathbf{x} \mapsto \mathbf{Yx}$ to $\operatorname{span}\{\mathbf{v}_1,\dots,\mathbf{v}_{r+1}\}$ has a nonzero kernel because its domain has dimension $r+1$ and its rank is at most r. Choose a unit vector $\mathbf{x} := \sum_{j=1}^{r+1} a_j\mathbf{v}_j$ in this kernel, where $\sum_{j=1}^{r+1} a_j^2 = 1$. Then we have

$$\begin{aligned}\|\mathbf{H}-\mathbf{Y}\|_2^2 &\ge \|(\mathbf{H}-\mathbf{Y})\mathbf{x}\|_2^2\\ &= \|\mathbf{Hx}\|_2^2\\ &= \sum_{j=1}^{r+1} \sigma_j(\mathbf{H})^2 a_j^2\\ &\ge \sigma_{r+1}(\mathbf{H})^2.\end{aligned}$$

The first step follows from the definition of the spectral norm and $\|\mathbf{x}\|_2 = 1$. The second step follows from $\mathbf{Yx} = \mathbf{0}$. The third step follows from the singular value decomposition and orthonormality of the $\mathbf{u}_j$. The fourth step follows from $\sigma_j(\mathbf{H}) \ge \sigma_{r+1}(\mathbf{H})$ for $j \le r+1$ and $\sum_j a_j^2 = 1$.

Step 3. Prove the Frobenius-norm lower bound. Let $\mathbf{P}$ be the orthogonal projection onto the column space of $\mathbf{Y}$. Define $a_j := \mathbf{u}_j^\top\mathbf{P}\mathbf{u}_j$ for $j \in [s]$. Since $\mathbf{P}$ is an orthogonal projection of rank at most r, these numbers satisfy $0 \le a_j \le 1$ and $\sum_{j=1}^{s} a_j \le \operatorname{tr}(\mathbf{P}) \le r$. For the sum inequality, extend $\mathbf{u}_1,\dots,\mathbf{u}_s$ to an orthonormal basis of $\mathbb{R}^m$ and use the nonnegativity of every diagonal entry of $\mathbf{P}$ in

that basis. Since the singular values are nonincreasing,

$$\begin{aligned}\sum_{j=1}^{s} \sigma_j(\mathbf{H})^2 a_j &\le \sum_{j=1}^{r} \sigma_j(\mathbf{H})^2 a_j + \sigma_{r+1}(\mathbf{H})^2 \sum_{j=r+1}^{s} a_j \\ &\le \sum_{j=1}^{r} \sigma_j(\mathbf{H})^2 a_j + \sigma_{r+1}(\mathbf{H})^2 \sum_{j=1}^{r} (1 - a_j) \\ &\le \sum_{j=1}^{r} \sigma_j(\mathbf{H})^2.\end{aligned}$$

The first step uses $\sigma_j(\mathbf{H}) \le \sigma_{r+1}(\mathbf{H})$ for $j > r$. The second step uses $\sum_{j=1}^{s} a_j \le r$. The third step uses $\sigma_{r+1}(\mathbf{H}) \le \sigma_j(\mathbf{H})$ and $1 - a_j \ge 0$ for $j \le r$. These inequalities also hold for $r = 0$, when $\mathbf{P} = \mathbf{0}$ and every $a_j = 0$. Consequently,

$$\begin{aligned}\|\mathbf{H} - \mathbf{Y}\|_F^2 &\ge \|(\mathbf{I}_m - \mathbf{P})(\mathbf{H} - \mathbf{Y})\|_F^2 \\ &= \|(\mathbf{I}_m - \mathbf{P})\mathbf{H}\|_F^2 \\ &= \sum_{j=1}^{s} \sigma_j(\mathbf{H})^2 (1 - a_j) \\ &\ge \sum_{j=r+1}^{s} \sigma_j(\mathbf{H})^2.\end{aligned}$$

The first step follows because an orthogonal projection cannot increase the Euclidean norm of any column. The second step uses $(\mathbf{I}_m - \mathbf{P})\mathbf{Y} = \mathbf{0}$. The third step follows by expanding the Frobenius norm in the orthonormal right singular vectors $\mathbf{v}_j$ and using $(\mathbf{I}_m - \mathbf{P})^2 = \mathbf{I}_m - \mathbf{P}$. The fourth step follows from the preceding bound on $\sum_j \sigma_j(\mathbf{H})^2 a_j$.

Step 4. Factor every feasible update. If $\mathbf{Y}$ has rank $k \le r$ and singular value decomposition $\mathbf{Y} = \sum_{j=1}^{k} s_j \mathbf{a}_j \mathbf{b}_j^\top$, place the vectors $s_j \mathbf{a}_j$ in the first k columns of $\mathbf{A}$ and the vectors $\mathbf{b}_j^\top$ in the first k rows of $\mathbf{B}$. Set all remaining entries of $\mathbf{A}$ and $\mathbf{B}$ to zero. Matrix multiplication gives $\mathbf{AB} = \mathbf{Y}$. This proves the factorization claim and completes the proof. □

Corollary E.2 (Rank-r matrix approximation)**.** *Let $\mathbf{H} \in \mathbb{R}^{m \times n}$ have rank s, and let $0 \le r < s$ be an integer. Then we have*

1. ***Spectral norm:***

$$\min_{\operatorname{rank}(\mathbf{Y}) = r} \|\mathbf{H} - \mathbf{Y}\|_2 = \sigma_{r+1}(\mathbf{H}).$$

2. ***Frobenius norm:***

$$\min_{\operatorname{rank}(\mathbf{Y}) = r} \|\mathbf{H} - \mathbf{Y}\|_F = \left(\sum_{j=r+1}^{s} \sigma_j(\mathbf{H})^2 \right)^{1/2}.$$

Proof. Every rank-r matrix is feasible for the rank-at-most-r problem in Lemma E.1, so the displayed quantities are lower bounds. The rank-r truncated singular value decomposition has rank exactly r because $r < s$. It attains both errors by the same lemma. Thus the bounds are equalities. □

E.2 Linear Models

E.2.1 Lower Bound for Universal Spoofing

Proposition E.3 (Lower bound for linear models)**.** *Let $\mathbf{W}, \mathbf{W}' \in \mathbb{R}^{d_1 \times d_2}$ be two weight matrices, where $\mathbf{W}$ has rank k_1 and $\mathbf{W}'$ has rank k_2, with $k_1 > k_2$. Let the low-rank adaptation (LoRA) matrices be $\mathbf{A} \in \mathbb{R}^{d_1 \times r}$ and $\mathbf{B} \in \mathbb{R}^{r \times d_2}$, where $r < k_1 - k_2$. Define the adapted weight matrix as $\widehat{\mathbf{W}} := \mathbf{W}' + \mathbf{AB}$ and the residual matrix as $\mathbf{R} := \mathbf{W} - \mathbf{W}'$. Then we have*

1. ***Spectral Norm:***

$$\inf_{\mathbf{A},\mathbf{B}} \sup_{\|\mathbf{x}\|_2=1} \|\mathbf{W}\mathbf{x} - \widehat{\mathbf{W}}\mathbf{x}\|_2 \geq \sigma^+_{\min}(\mathbf{R}),$$

2. ***Frobenius Norm:***

$$\inf_{\mathbf{A},\mathbf{B}} \|\mathbf{W} - \widehat{\mathbf{W}}\|_F^2 \geq (k_1 - k_2 - r) \cdot \sigma^+_{\min}(\mathbf{R})^2,$$

where $\sigma^+_{\min}(\mathbf{R})$ denotes the smallest non-zero singular value of $\mathbf{R}$.

Proof. **Part 1. Spectral Norm**. We first consider the LHS of the spectral norm bound without taking the infimum:

$$\begin{aligned}\sup_{\|\mathbf{x}\|_2=1} \|\mathbf{W}\mathbf{x} - \widehat{\mathbf{W}}\mathbf{x}\|_2 &= \|\mathbf{W} - \widehat{\mathbf{W}}\|_2 \\ &= \|\mathbf{R} - \mathbf{AB}\|_2,\end{aligned}$$

where the first equality follows from the definition of spectral norm, and the second equality follows from the definition of $\mathbf{R}$.

Taking the infimum over the LoRA matrices $\mathbf{A}$ and $\mathbf{B}$ on both sides, we obtain:

$$\inf_{\mathbf{A},\mathbf{B}} \sup_{\|\mathbf{x}\|_2=1} \|\mathbf{W}\mathbf{x} - \widehat{\mathbf{W}}\mathbf{x}\|_2 = \inf_{\mathbf{A},\mathbf{B}} \|\mathbf{R} - \mathbf{AB}\|_2. \tag{3}$$

Since the LoRA matrices $\mathbf{A} \in \mathbb{R}^{d_1 \times r}$ and $\mathbf{B} \in \mathbb{R}^{r \times d_2}$ are bottlenecked by their inner dimension r, their product $\mathbf{AB}$ has a rank of at most r. Conversely, every matrix in $\mathbb{R}^{d_1 \times d_2}$ of rank at most r admits such a factorization. Therefore, minimizing $\|\mathbf{R} - \mathbf{AB}\|_2$ is equivalent to finding the optimal rank-r approximation of the residual matrix $\mathbf{R}$ under the spectral norm. We can directly apply part 1 of Corollary E.2 to evaluate this optimal error, which yields:

$$\inf_{\mathbf{A},\mathbf{B}} \|\mathbf{R} - \mathbf{AB}\|_2 = \sigma_{r+1}(\mathbf{R}). \tag{4}$$

We now lower bound $\sigma_{r+1}(\mathbf{R})$. Because $\mathbf{W} = \mathbf{W}' + \mathbf{R}$, where $\mathrm{rank}(\mathbf{W}) = k_1$ and $\mathrm{rank}(\mathbf{W}') = k_2$, the subadditivity property of matrix rank dictates that $\mathrm{rank}(\mathbf{R}) \geq k_1 - k_2$. The proposition assumes the adaptation rank satisfies $r < k_1 - k_2$, which directly implies $r + 1 \leq \mathrm{rank}(\mathbf{R})$. Consequently, the $(r+1)$-th singular value must be at least as large as the smallest non-zero singular value of $\mathbf{R}$:

$$\sigma_{r+1}(\mathbf{R}) \geq \sigma^+_{\min}(\mathbf{R}). \tag{5}$$

Combining Eq. (3), Eq. (4) and Eq. (5) above, we obtain the desired result:

$$\inf_{\mathbf{A},\mathbf{B}} \sup_{\|\mathbf{x}\|_2=1} \|\mathbf{W}\mathbf{x} - \widehat{\mathbf{W}}\mathbf{x}\|_2 \geq \sigma^+_{\min}(\mathbf{R}),$$

which finishes the proof for the spectral norm case.

Part 2. Frobenius Norm. We consider the LHS of the Frobenius norm bound and have the following:

$$\begin{aligned}\inf_{\mathbf{A},\mathbf{B}} \|\mathbf{W} - \widehat{\mathbf{W}}\|_F^2 &= \inf_{\mathbf{A},\mathbf{B}} \|\mathbf{R} - \mathbf{AB}\|_F^2 \\ &= \min_{\mathrm{rank}(\mathbf{X}) \leq r} \|\mathbf{R} - \mathbf{X}\|_F^2 \\ &= \sum_{i=r+1}^{\mathrm{rank}(\mathbf{R})} \sigma_i^2(\mathbf{R}) \\ &\geq (\mathrm{rank}(\mathbf{R}) - r)\sigma^+_{\min}(\mathbf{R})^2 \\ &\geq (k_1 - k_2 - r)\sigma^+_{\min}(\mathbf{R})^2,\end{aligned}$$

where the first step is from the definition of $\mathbf{R}$, the second step is from the fact that $\mathbf{AB}$ has rank at most r, together with the converse factorization described above, the third step is from part 2 of Corollary E.2, the fourth step is from the fact that all terms in the summation are at least $\sigma^{+}_{\min}(\mathbf{R})^2$, and the last step is from the fact that $\mathrm{rank}(\mathbf{R}) \geq k_1 - k_2$.

Thus, we complete the proof. □

E.2.2 Upper Bound under Restricted Queries

To prove this upper bound, we first present a useful fact.

Fact E.4 (Second moment of a linear map)**.** *For a zero-mean stochastic vector* $\mathbf{x}$ *with covariance matrix* $\mathbf{M} := \mathbb{E}[\mathbf{x}\mathbf{x}^\top]$ *and any constant matrix* $\mathbf{A}$*, we have:*

$$\mathbb{E}[(\mathbf{A}\mathbf{x})(\mathbf{A}\mathbf{x})^\top] = \mathbf{A}\mathbf{M}\mathbf{A}^\top.$$

Now, we show the upper bound under the simple query distribution setting.

Theorem E.5 (Upper bound under restricted queries)**.** *Let* $\mathbf{x} \in \mathbb{R}^{d_2}$ *be the representation of a query drawn from a zero-mean continuous distribution* $\mathcal{D}_q$*, and let* $\mathbf{\Sigma} := \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_q}[\mathbf{x}\mathbf{x}^\top]$ *be the covariance matrix. Let* $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{d_2} \geq 0$ *denote the eigenvalues of* $\mathbf{\Sigma}$*.*
Let the LoRA matrices be $\mathbf{A} \in \mathbb{R}^{d_1 \times r}$ *and* $\mathbf{B} \in \mathbb{R}^{r \times d_2}$*, and the adapted weight matrix be* $\widehat{\mathbf{W}} := \mathbf{W}' + \mathbf{AB}$*. Then we have*

$$\inf_{\mathbf{A},\mathbf{B}} \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_q}\left[\|\mathbf{W}\mathbf{x} - \widehat{\mathbf{W}}\mathbf{x}\|_2^2\right] \leq \|\mathbf{R}\|_2^2 \sum_{i=r+1}^{d_2} \lambda_i,$$

where $\mathbf{R} := \mathbf{W} - \mathbf{W}'$ *is the residual matrix.*

Proof. Let $\mathbf{\Sigma} = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_q}[\mathbf{x}\mathbf{x}^\top]$ have eigendecomposition $\mathbf{\Sigma} = \mathbf{U}\mathbf{L}\mathbf{U}^\top$ with $\mathbf{L} = \mathrm{diag}(\lambda_1, \ldots, \lambda_{d_2})$ and $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$. Choose $\mathbf{U}_r \in \mathbb{R}^{d_2 \times r}$ as the first r eigenvectors. Construct

$$\mathbf{AB} = \mathbf{R}\mathbf{U}_r\mathbf{U}_r^\top,$$

which has rank at most r and thus can be factored as $\mathbf{AB}$.

For this choice, we have the following:

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_q}\left[\|\mathbf{W}\mathbf{x} - \widehat{\mathbf{W}}\mathbf{x}\|_2^2\right] \\
&= \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_q}\left[\|(\mathbf{R} - \mathbf{AB})\mathbf{x}\|_2^2\right] \\
&= \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_q}\left[\|\mathbf{R}(\mathbf{I} - \mathbf{U}_r\mathbf{U}_r^\top)\mathbf{x}\|_2^2\right] \\
&= \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_q}\left[\|\mathbf{R}(\mathbf{U}\mathbf{U}^\top - \mathbf{U}_r\mathbf{U}_r^\top)\mathbf{x}\|_2^2\right] \\
&= \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_q}\left[\left\|\mathbf{R}\left(\sum_{j=r+1}^{d_2} \mathbf{u}_j\mathbf{u}_j^\top\right)\mathbf{x}\right\|_2^2\right],
\end{aligned}$$

where the first step is from the definition of $\widehat{\mathbf{W}}$ and $\mathbf{R}$, the second step is from choosing a specific $\mathbf{AB}$, the third step is from the orthogonality of eigenvectors, and the last step is from expanding the matrix into eigenvector outer products and eliminating the first r terms.

For simplicity, let $\mathbf{Q} := \sum_{j=r+1}^{d_2} \mathbf{u}_j\mathbf{u}_j^\top$. Applying the trace identity for expected quadratic forms in Fact E.4, we have:

$$\begin{aligned}
\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_q}\left[\|\mathbf{W}\mathbf{x} - \widehat{\mathbf{W}}\mathbf{x}\|_2^2\right] &= \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_q}\left[\|\mathbf{R}\mathbf{Q}\mathbf{x}\|_2^2\right] \\
&= \mathrm{tr}(\mathbf{R}\mathbf{Q}\mathbf{\Sigma}\mathbf{Q}^\top\mathbf{R}^\top).
\end{aligned}$$

Since eigenvectors form an orthonormal basis, we expand and collapse the inner product $\mathbf{Q}\mathbf{\Sigma}\mathbf{Q}^\top$ as follows:

$$\begin{aligned}
&\mathbf{Q}\mathbf{\Sigma}\mathbf{Q}^\top \\
&= \left(\sum_{j=r+1}^{d_2} \mathbf{u}_j\mathbf{u}_j^\top\right)\left(\sum_{i=1}^{d_2} \lambda_i\mathbf{u}_i\mathbf{u}_i^\top\right)\left(\sum_{k=r+1}^{d_2} \mathbf{u}_k\mathbf{u}_k^\top\right) \\
&= \sum_{i=r+1}^{d_2} \lambda_i\mathbf{u}_i\mathbf{u}_i^\top,
\end{aligned}$$

where the second step follows from the orthogonality condition $\mathbf{u}_j^\top\mathbf{u}_i = 1$ if $i = j$ and 0 otherwise. Finally, we substitute this collapsed term back into the trace to bound the expected error:

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_q}\left[\|\mathbf{W}\mathbf{x} - \widehat{\mathbf{W}}\mathbf{x}\|_2^2\right] \\
&= \mathrm{tr}(\mathbf{R}\mathbf{Q}\mathbf{\Sigma}\mathbf{Q}^\top\mathbf{R}^\top) \\
&= \mathrm{tr}\left(\mathbf{R}\left(\sum_{i=r+1}^{d_2} \lambda_i\mathbf{u}_i\mathbf{u}_i^\top\right)\mathbf{R}^\top\right) \\
&= \sum_{i=r+1}^{d_2} \lambda_i\mathrm{tr}(\mathbf{R}\mathbf{u}_i\mathbf{u}_i^\top\mathbf{R}^\top) \\
&= \sum_{i=r+1}^{d_2} \lambda_i\mathrm{tr}(\mathbf{u}_i^\top\mathbf{R}^\top\mathbf{R}\mathbf{u}_i) \\
&= \sum_{i=r+1}^{d_2} \lambda_i\|\mathbf{R}\mathbf{u}_i\|_2^2 \\
&\leq \sum_{i=r+1}^{d_2} \lambda_i\|\mathbf{R}\|_2^2\|\mathbf{u}_i\|_2^2 \\
&= \|\mathbf{R}\|_2^2\sum_{i=r+1}^{d_2} \lambda_i,
\end{aligned}$$

where the third step follows from the linearity of the trace operator, the fourth step follows from the cyclic property of the trace, the fifth step follows from the definition of the ℓ_2 norm, the sixth step follows from the fact that $\|\mathbf{R}\mathbf{u}_i\|_2 \leq \|\mathbf{R}\|_2\|\mathbf{u}_i\|_2$, and the final step follows from the fact that eigenvectors are unit vectors (i.e., $\|\mathbf{u}_i\|_2 = 1$).

Taking the infimum over all parameterizations of $\mathbf{A}$ and $\mathbf{B}$ can only yield an expected error less than or equal to this specific construction. Therefore, the upper bound holds, completing the proof. □

E.2.3 Uniform-Ball Queries

Corollary E.6 (Linear models with uniform ball queries)**.** *Let d_1, d_2 be positive integers and $0 \leq r \leq \min\{d_1, d_2\}$ be an integer. Let $\mathbf{W}, \mathbf{W}' \in \mathbb{R}^{d_1\times d_2}$ be fixed weight matrices and define $\mathbf{R} := \mathbf{W} - \mathbf{W}'$. Let $\rho > 0$. Let $\mathcal{D}_q$ be the uniform distribution, with respect to Lebesgue measure, on*

$$B_{d_2}(\rho) := \{\mathbf{x} \in \mathbb{R}^{d_2} : \|\mathbf{x}\|_2 \leq \rho\}.$$

For $\mathbf{A} \in \mathbb{R}^{d_1 \times r}$ and $\mathbf{B} \in \mathbb{R}^{r \times d_2}$, define $\widehat{\mathbf{W}} := \mathbf{W}' + \mathbf{AB}$. Then we have

$$\begin{aligned}\boldsymbol{\Sigma} := \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_q}[\mathbf{x}\mathbf{x}^\top] &= \frac{\rho^2}{d_2+2}\mathbf{I}_{d_2},\\ \inf_{\mathbf{A},\mathbf{B}} \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_q}\left[\|\mathbf{W}\mathbf{x} - \widehat{\mathbf{W}}\mathbf{x}\|_2^2\right] &= \frac{\rho^2}{d_2+2}\sum_{j=r+1}^{\min\{d_1,d_2\}} \sigma_j(\mathbf{R})^2\\ &\le \frac{\rho^2}{d_2+2}(d_2-r)\|\mathbf{R}\|_2^2.\end{aligned}$$

Proof. **Step 1. Compute the query covariance.** The distribution of $\mathbf{x}$ is invariant under coordinate sign changes and coordinate permutations. A sign change in coordinate i changes the sign of x_i and of $x_i x_j$ for $i \neq j$ without changing their distributions. Thus $\mathbb{E}[x_i] = 0$ and $\mathbb{E}[x_i x_j] = 0$ for $i \neq j$. Coordinate permutation invariance implies that all $\mathbb{E}[x_i^2]$ are equal. The volume of a radius-t ball is t^{d_2} times the volume of the unit ball. Therefore $\Pr(\|\mathbf{x}\|_2 \le t) = (t/\rho)^{d_2}$ for $0 \le t \le \rho$. Differentiating this distribution function gives the radial density $d_2 t^{d_2-1}/\rho^{d_2}$ for $0 < t < \rho$. It follows that

$$\begin{aligned}\mathbb{E}[\|\mathbf{x}\|_2^2] &= \int_0^\rho t^2 \frac{d_2 t^{d_2-1}}{\rho^{d_2}}\,dt\\ &= \frac{d_2}{\rho^{d_2}}\frac{\rho^{d_2+2}}{d_2+2}\\ &= \frac{d_2\rho^2}{d_2+2}.\end{aligned}$$

The first step uses the radial density. The second step evaluates the integral of t^{d_2+1}. The third step cancels ρ^{d_2}. Since $\sum_i \mathbb{E}[x_i^2] = \mathbb{E}[\|\mathbf{x}\|_2^2]$ and the d_2 terms are equal, each diagonal entry of $\boldsymbol{\Sigma}$ is $\rho^2/(d_2+2)$. Together with the zero off-diagonal entries, this proves the stated covariance formula.

Step 2. Rewrite the expected error as a matrix approximation error. Fix any $\mathbf{A}, \mathbf{B}$ and define $\mathbf{H} := \mathbf{R} - \mathbf{AB}$. Then we have

$$\begin{aligned}\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_q}\left[\|\mathbf{W}\mathbf{x} - \widehat{\mathbf{W}}\mathbf{x}\|_2^2\right] &= \mathbb{E}[\mathbf{x}^\top\mathbf{H}^\top\mathbf{H}\mathbf{x}]\\ &= \operatorname{tr}\left(\mathbf{H}^\top\mathbf{H}\,\mathbb{E}[\mathbf{x}\mathbf{x}^\top]\right)\\ &= \frac{\rho^2}{d_2+2}\operatorname{tr}(\mathbf{H}^\top\mathbf{H})\\ &= \frac{\rho^2}{d_2+2}\|\mathbf{R}-\mathbf{AB}\|_F^2.\end{aligned}$$

The first step substitutes $\widehat{\mathbf{W}} = \mathbf{W}' + \mathbf{AB}$ and $\mathbf{H} = \mathbf{R} - \mathbf{AB}$ into the squared Euclidean norm. The second step uses $\mathbf{x}^\top\mathbf{H}^\top\mathbf{H}\mathbf{x} = \operatorname{tr}(\mathbf{H}^\top\mathbf{H}\mathbf{x}\mathbf{x}^\top)$ and linearity of expectation. The third step substitutes the covariance from Step 1. The fourth step uses $\operatorname{tr}(\mathbf{H}^\top\mathbf{H}) = \sum_{i,j} H_{ij}^2 = \|\mathbf{H}\|_F^2$ and the definition of $\mathbf{H}$.

Step 3. Minimize over the adapter. By Lemma E.1, the products $\mathbf{AB}$ are exactly the matrices of rank at most r. The same lemma gives

$$\begin{aligned}\inf_{\mathbf{A},\mathbf{B}} \mathbb{E}\left[\|\mathbf{W}\mathbf{x} - \widehat{\mathbf{W}}\mathbf{x}\|_2^2\right] &= \frac{\rho^2}{d_2+2}\sum_{j=r+1}^{\min\{d_1,d_2\}} \sigma_j(\mathbf{R})^2\\ &\le \frac{\rho^2}{d_2+2}(d_2-r)\|\mathbf{R}\|_2^2.\end{aligned}$$

The first step applies the Frobenius-norm conclusion of Lemma E.1 to Step 2. The second step uses $\sigma_j(\mathbf{R}) \le \|\mathbf{R}\|_2$ and pads the sum with zero singular values up to index d_2. Finally, Step 1 gives $\sum_{j=r+1}^{d_2} \lambda_j = (d_2-r)\rho^2/(d_2+2)$, which proves the claimed relation to Theorem E.5. □

Remark E.7 (Uniform ball queries)**.** *The upper bound is precisely the bound in Theorem E.5 with $\lambda_j = \rho^2/(d_2+2)$ for every $j \in [d_2]$, where λ_j denotes the jth largest eigenvalue of $\mathbf{\Sigma}$. The equality in Corollary E.6 evaluates the same expected-error objective as Theorem E.5 for one specified query distribution. The spectral tail quantifies the error that a rank-r adapter cannot remove under these queries. An isotropic ball has no preferred low-dimensional query subspace. Consequently, this specialization need not give a small upper bound when r is small relative to d_2. It does not contradict the small-error guarantee for a distribution whose covariance has a small eigenvalue tail.*

E.2.4 Upper Bound under Low-Rank Verification

We first prove a matrix approximation identity for a fixed linear verifier.

Lemma E.8 (Matrix approximation after a linear map)**.** *Let $\mathbf{A} \in \mathbb{R}^{d_1 \times d_2}$ and $\mathbf{C} \in \mathbb{R}^{d_1 \times d_1}$. Let the singular value decomposition of the product matrix $\mathbf{CA}$ be given by $\mathbf{CA} = \sum_{i=1}^{p} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$, where $p = \min(d_1, d_2)$ and $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$.*
For any integer $0 \leq k < p$, let $(\mathbf{CA})_k := \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$ denote the rank-k truncated singular value decomposition of $\mathbf{CA}$.
Then we have

$$\min_{\mathrm{rank}(\mathbf{B}) \leq k} \|\mathbf{CA} - \mathbf{CB}\|_2 = \sigma_{k+1}(\mathbf{CA}).$$

Proof. **Part 1. Existence**. We show that there exists a matrix $\mathbf{B}$ of rank at most k such that $\|\mathbf{CA} - \mathbf{CB}\|_2 = \sigma_{k+1}(\mathbf{CA})$.

Because the product matrix $\mathbf{CA}$ is formed by linear combinations of the columns of $\mathbf{C}$, its column space is restricted: $\mathrm{col}(\mathbf{CA}) \subseteq \mathrm{col}(\mathbf{C})$. Furthermore, its truncated SVD is built from its principal directions, meaning $\mathrm{col}((\mathbf{CA})_k) \subseteq \mathrm{col}(\mathbf{CA})$.

Since $\mathrm{col}((\mathbf{CA})_k) \subseteq \mathrm{col}(\mathbf{C})$, the linear equation $\mathbf{CX} = (\mathbf{CA})_k$ is consistent. We can construct a specific solution using the Moore-Penrose pseudoinverse: let $\mathbf{B}^* = \mathbf{C}^+(\mathbf{CA})_k$. This satisfies $\mathbf{CB}^* = (\mathbf{CA})_k$. Furthermore, because the rank of a product cannot exceed the rank of its factors, $\mathrm{rank}(\mathbf{B}^*) \leq \mathrm{rank}((\mathbf{CA})_k) \leq k$.

We now consider this specific valid matrix $\mathbf{B}^*$ of rank at most k and compute the approximation error:

$$\begin{aligned}
\|\mathbf{CA} - \mathbf{CB}^*\|_2 &= \|\mathbf{CA} - (\mathbf{CA})_k\|_2 \\
&= \left\| \sum_{i=1}^{p} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top - \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top \right\|_2 \\
&= \left\| \sum_{i=k+1}^{p} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top \right\|_2,
\end{aligned}$$

which follows from basic algebra.

Because the remaining singular values are ordered $\sigma_{k+1} \geq \sigma_{k+2} \geq \ldots$, the spectral norm of this residual matrix is exactly the largest remaining singular value. Thus, $\|\mathbf{CA} - \mathbf{CB}^*\|_2 = \sigma_{k+1}(\mathbf{CA})$.

Part 2. Optimality. We show that no matrix $\mathbf{B}$ with rank at most k can achieve a lower error value than $\sigma_{k+1}(\mathbf{CA})$.

Suppose $\mathrm{rank}(\mathbf{B}) \leq k$ for some arbitrary matrix $\mathbf{B} \in \mathbb{R}^{d_1 \times d_2}$. Let $\mathbf{Y} = \mathbf{CB}$. By the properties of matrix multiplication, $\mathrm{rank}(\mathbf{Y}) \leq \mathrm{rank}(\mathbf{B}) \leq k$.

Because $\mathbf{Y}$ has d_2 columns and rank at most k, its null space has a dimension of at least $d_2 - k$, which follows from the rank-nullity theorem. It follows that we can find orthonormal vectors $\mathbf{x}_1, \ldots, \mathbf{x}_{d_2-k}$ such that $\mathrm{span}\{\mathbf{x}_1, \ldots, \mathbf{x}_{d_2-k}\} \subseteq \mathrm{null}(\mathbf{Y})$ via the Gram-Schmidt process.

Let $\mathcal{V}_{k+1} := \mathrm{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_{k+1}\}$ be the $(k+1)$-dimensional subspace spanned by the top right singular vectors of $\mathbf{CA}$. A dimension argument shows that:

$$(d_2 - k) + (k + 1) = d_2 + 1 > d_2,$$

meaning the sum of their dimensions exceeds the ambient space $\mathbb{R}^{d_2}$. Therefore, their intersection must be non-trivial:

$$\text{span}\{\mathbf{x}_1, \ldots, \mathbf{x}_{d_2-k}\} \cap \text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_{k+1}\} \neq \{\mathbf{0}\}.$$

Let $\mathbf{z}$ be a unit 2-norm vector (i.e., $\|\mathbf{z}\|_2 = 1$) in this intersection. Since $\mathbf{z} \in \text{null}(\mathbf{Y})$, we have $\mathbf{Yz} = \mathbf{CBz} = \mathbf{0}$. Furthermore, since $\mathbf{z} \in \mathcal{V}_{k+1}$, applying $\mathbf{CA}$ to $\mathbf{z}$ yields:

$$(\mathbf{CA})\mathbf{z} = \sum_{i=1}^{k+1} \sigma_i(\mathbf{v}_i^\top \mathbf{z})\mathbf{u}_i. \tag{6}$$

We can now lower-bound the spectral norm for any choice of $\mathbf{B}$:

$$\begin{aligned}
\|\mathbf{CA} - \mathbf{CB}\|_2^2 &\geq \|(\mathbf{CA} - \mathbf{CB})\mathbf{z}\|_2^2 \\
&= \|(\mathbf{CA})\mathbf{z} - \mathbf{0}\|_2^2 \\
&= \left\| \sum_{i=1}^{k+1} \sigma_i(\mathbf{v}_i^\top \mathbf{z})\mathbf{u}_i \right\|_2^2 \\
&= \sum_{i=1}^{k+1} \sigma_i^2(\mathbf{v}_i^\top \mathbf{z})^2 \\
&\geq \sigma_{k+1}^2 \sum_{i=1}^{k+1} (\mathbf{v}_i^\top \mathbf{z})^2 \\
&= \sigma_{k+1}^2 \|\mathbf{z}\|_2^2 \\
&= \sigma_{k+1}^2,
\end{aligned}$$

where the sixth step follows from the fact that $\mathbf{z} \in \text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_{k+1}\}$, and the other steps follow from definitions and basic algebra.

Taking the square root of both sides yields $\|\mathbf{CA} - \mathbf{CB}\|_2 \geq \sigma_{k+1}(\mathbf{CA})$, completing the proof.

□

Based on the lemma above, we show the tight bound as follows.

Theorem E.9 (Minimum error under linear verification)**.** *Let $\mathbf{C} \in \mathbb{R}^{d_1 \times d_1}$ be a weight matrix from the auditor's verification model, and let $\mathbf{W}, \mathbf{W}' \in \mathbb{R}^{d_1 \times d_2}$ be the weight matrices of the strong target model and the weak model, respectively.*
Let the LoRA matrices be $\mathbf{A} \in \mathbb{R}^{d_1 \times r}$ and $\mathbf{B} \in \mathbb{R}^{r \times d_2}$, and the adapted weight matrix be $\widehat{\mathbf{W}} := \mathbf{W}' + \mathbf{AB}$. Then we have

$$\inf_{\mathbf{A},\mathbf{B}} \sup_{\|\mathbf{x}\|_2=1} \|\mathbf{C}^\top(\mathbf{Wx} - \widehat{\mathbf{W}}\mathbf{x})\|_2 = \sigma_{r+1}(\mathbf{C}^\top \mathbf{R}),$$

where $\mathbf{R} := \mathbf{W} - \mathbf{W}'$ is the residual matrix, and $\sigma_{r+1}(\cdot)$ is the $(r+1)$-th largest singular value.

Proof. By the definition of $\widehat{\mathbf{W}}$ and $\mathbf{R}$, and the definition of the matrix spectral norm, we first rewrite the worst-case spoofing error:

$$\begin{aligned}
&\inf_{\mathbf{A},\mathbf{B}} \sup_{\|\mathbf{x}\|_2=1} \|\mathbf{C}^\top(\mathbf{Wx} - \widehat{\mathbf{W}}\mathbf{x})\|_2 \\
&= \inf_{\mathbf{A},\mathbf{B}} \|\mathbf{C}^\top \mathbf{R} - \mathbf{C}^\top \mathbf{AB}\|_2.
\end{aligned} \tag{7}$$

Because the LoRA matrices $\mathbf{A} \in \mathbb{R}^{d_1 \times r}$ and $\mathbf{B} \in \mathbb{R}^{r \times d_2}$ are bottlenecked by their inner dimension r, their product $\mathbf{AB}$ forms a matrix of rank at most r. Conversely, any $d_1 \times d_2$ matrix with a rank of at most r can be exactly factored into the product of two such matrices $\mathbf{A}$ and $\mathbf{B}$. Therefore, taking

the infimum over the parameter space of $\mathbf{A}$ and $\mathbf{B}$ is mathematically equivalent to minimizing over all possible matrices $\mathbf{X}$ of rank at most r:

$$\begin{aligned}&\inf_{\mathbf{A},\mathbf{B}} \|\mathbf{C}^\top\mathbf{R} - \mathbf{C}^\top\mathbf{A}\mathbf{B}\|_2 \\ &= \min_{\mathrm{rank}(\mathbf{X})\le r} \|\mathbf{C}^\top\mathbf{R} - \mathbf{C}^\top\mathbf{X}\|_2. \end{aligned} \tag{8}$$

We now apply Lemma E.8, substituting $\mathbf{C}$ with $\mathbf{C}^\top$, $\mathbf{A}$ with $\mathbf{R}$, and the rank constraint k with r. This yields exactly:

$$\min_{\mathrm{rank}(\mathbf{X})\le r} \|\mathbf{C}^\top\mathbf{R} - \mathbf{C}^\top\mathbf{X}\|_2 = \sigma_{r+1}(\mathbf{C}^\top\mathbf{R}). \tag{9}$$

Combining Eq. (7), Eq. (8), and Eq. (9) completes the proof. □

E.3 Causal Attention

Definition E.10 (Attention parameters and updates). *Let $d, h, p \ge 1$ and $L \ge 2$ be integers. Let $r \in \{0, \dots, d\}$, $\rho > 0$, and $M > 0$. The symbols d, h, p, and L denote the input dimension, head dimension, output dimension, and sequence length, respectively. Let $\mathbf{I}_d \in \mathbb{R}^{d\times d}$ denote the identity matrix.*
Fix matrices $\mathbf{Q}_*, \mathbf{K}_*, \mathbf{Q}_0, \mathbf{K}_0 \in \mathbb{R}^{d\times h}$ and a nonzero matrix $\mathbf{V} \in \mathbb{R}^{d\times p}$. Define

$$\begin{aligned}\mathbf{S}_* &:= \mathbf{Q}_*\mathbf{K}_*^\top, \\ \mathbf{S}_0 &:= \mathbf{Q}_0\mathbf{K}_0^\top.\end{aligned}$$

Assume $\|\mathbf{S}_*\|_2 \le M$ and $\|\mathbf{S}_0\|_2 \le M$. For matrices $\Delta\mathbf{Q}, \Delta\mathbf{K} \in \mathbb{R}^{d\times h}$, define

$$\mathbf{S}_\Delta := (\mathbf{Q}_0 + \Delta\mathbf{Q})(\mathbf{K}_0 + \Delta\mathbf{K})^\top.$$

Define the admissible update family by

$$\mathcal{A}_{r,M} := \left\{ (\Delta\mathbf{Q}, \Delta\mathbf{K}) : \begin{array}{c} \Delta\mathbf{Q}, \Delta\mathbf{K} \in \mathbb{R}^{d\times h}, \\ \mathrm{rank}(\Delta\mathbf{Q}) \le r, \quad \mathrm{rank}(\Delta\mathbf{K}) \le r, \\ \|\mathbf{S}_\Delta\|_2 \le M \end{array} \right\}.$$

The matrix $\mathbf{V}$ is shared by the target and every candidate.

The zero update belongs to $\mathcal{A}_{r,M}$ because $\|\mathbf{S}_0\|_2 \le M$.

Definition E.11 (Causal attention outputs). *For $\mathbf{z} = (z_1, \dots, z_L)^\top \in \mathbb{R}^L$, define $\mathrm{softmax}(\mathbf{z}) \in \mathbb{R}^L$ coordinatewise by*

$$[\mathrm{softmax}(\mathbf{z})]_i := \frac{\exp(z_i)}{\sum_{j=1}^L \exp(z_j)}.$$

For a matrix $\mathbf{S} \in \mathbb{R}^{d\times d}$ and a sequence matrix $\mathbf{X} \in \mathbb{R}^{L\times d}$, write the i-th row of $\mathbf{X}$ as $\mathbf{x}_i^\top$, where $\mathbf{x}_i \in \mathbb{R}^d$. Define the causal attention matrix $\mathbf{P}_\mathbf{S}(\mathbf{X}) \in \mathbb{R}^{L\times L}$ by

$$[\mathbf{P}_\mathbf{S}(\mathbf{X})]_{ij} := \begin{cases} \dfrac{\exp(\mathbf{x}_i^\top\mathbf{S}\mathbf{x}_j/\sqrt{h})}{\sum_{\ell=1}^i \exp(\mathbf{x}_i^\top\mathbf{S}\mathbf{x}_\ell/\sqrt{h})}, & 1 \le j \le i \le L, \\ 0, & 1 \le i < j \le L. \end{cases}$$

Define the full output matrix and the final-position output by

$$\begin{aligned}F_\mathbf{S}(\mathbf{X}) &:= \mathbf{P}_\mathbf{S}(\mathbf{X})\mathbf{X}\mathbf{V} \in \mathbb{R}^{L\times p}, \\ \phi_\mathbf{S}(\mathbf{X}) &:= [F_\mathbf{S}(\mathbf{X})]_{L,:} \in \mathbb{R}^{1\times p},\end{aligned}$$

where $[\mathbf{B}]_{i,:}$ denotes the i-th row of a matrix $\mathbf{B}$.

Define the final-position logit vector $\mathbf{z}_{\mathbf{S}}(\mathbf{X}) \in \mathbb{R}^L$ *and probability vector* $\mathbf{p}_{\mathbf{S}}(\mathbf{X}) \in \mathbb{R}^L$ *by*

$$[\mathbf{z}_{\mathbf{S}}(\mathbf{X})]_i := \frac{\mathbf{x}_L^\top \mathbf{S}\mathbf{x}_i}{\sqrt{h}},$$
$$\mathbf{p}_{\mathbf{S}}(\mathbf{X}) := \mathrm{softmax}(\mathbf{z}_{\mathbf{S}}(\mathbf{X})).$$

Equivalently,

$$\phi_{\mathbf{S}}(\mathbf{X}) = \sum_{i=1}^{L} [\mathbf{p}_{\mathbf{S}}(\mathbf{X})]_i \mathbf{x}_i^\top \mathbf{V}.$$

The norm $\|\phi_{\mathbf{S}}(\mathbf{X})\|_2$ *is the Euclidean norm of this row vector.*

Definition E.12 (Weighted mean and covariance)**.** *Define the probability simplex*

$$\mathcal{P}_L := \left\{ \mathbf{w} \in \mathbb{R}^L : w_i \geq 0 \text{ for all } i,\ \sum_{i=1}^{L} w_i = 1 \right\}.$$

For $\mathbf{Y} \in \mathbb{R}^{L\times d}$ *with rows* $\mathbf{y}_i^\top$ *and* $\mathbf{w} \in \mathcal{P}_L$*, define*

$$\mathbf{m}_{\mathbf{w}}(\mathbf{Y}) := \sum_{i=1}^{L} w_i \mathbf{y}_i \in \mathbb{R}^d,$$
$$\mathbf{\Gamma}_{\mathbf{w}}(\mathbf{Y}) := \sum_{i=1}^{L} w_i \mathbf{y}_i \mathbf{y}_i^\top - \mathbf{m}_{\mathbf{w}}(\mathbf{Y})\mathbf{m}_{\mathbf{w}}(\mathbf{Y})^\top \in \mathbb{R}^{d\times d}.$$

Define the uniform weight vector by $\mathbf{u} := (1/L, \ldots, 1/L)^\top \in \mathcal{P}_L$*. For* $\mathbf{v} \in \mathbb{R}^L$*, use* $\|\mathbf{v}\|_1 := \sum_i |v_i|$ *and* $\|\mathbf{v}\|_\infty := \max_i |v_i|$*.*

Lemma E.13 (Attention output difference)**.** *Fix* $(\Delta\mathbf{Q}, \Delta\mathbf{K}) \in \mathcal{A}_{r,M}$ *and* $\mathbf{X} \in \mathbb{R}^{L\times d}$*. Define*

$$\mathbf{R} := \mathbf{S}_* - \mathbf{S}_\Delta,$$
$$\mathbf{S}_t := \mathbf{S}_\Delta + t\mathbf{R} \qquad (0 \leq t \leq 1),$$
$$\mathbf{p}_t := \mathbf{p}_{\mathbf{S}_t}(\mathbf{X}),$$
$$p_{t,i} := [\mathbf{p}_t]_i \qquad (1 \leq i \leq L),$$
$$\mathbf{\Gamma}_t(\mathbf{X}) := \mathbf{\Gamma}_{\mathbf{p}_t}(\mathbf{X}).$$

Then we have

$$\phi_{\mathbf{S}_*}(\mathbf{X}) - \phi_{\mathbf{S}_\Delta}(\mathbf{X}) = \frac{1}{\sqrt{h}} \mathbf{x}_L^\top \mathbf{R} \left(\int_0^1 \mathbf{\Gamma}_t(\mathbf{X})\, dt \right) \mathbf{V}. \tag{10}$$

The integral is taken entrywise.

Proof. **Step 1. Differentiate the logits.** For each $i \in \{1, \ldots, L\}$, define $z_{t,i} := [\mathbf{z}_{\mathbf{S}_t}(\mathbf{X})]_i$. Then we have

$$z_{t,i} = \frac{\mathbf{x}_L^\top (\mathbf{S}_\Delta + t\mathbf{R})\mathbf{x}_i}{\sqrt{h}},$$
$$\frac{d}{dt} z_{t,i} = \frac{\mathbf{x}_L^\top \mathbf{R}\mathbf{x}_i}{\sqrt{h}}.$$

The first equality follows from the definitions of $\mathbf{z}_{\mathbf{S}_t}(\mathbf{X})$ and $\mathbf{S}_t$. The second equality follows by differentiation because $\mathbf{X}$, $\mathbf{S}_\Delta$, and $\mathbf{R}$ are fixed.

Step 2. Differentiate the attention probabilities. The denominator $\sum_j \exp(z_{t,j})$ is positive for every t. The quotient rule gives

$$\begin{aligned}\frac{d}{dt}p_{t,i} &= p_{t,i}\left(\frac{d}{dt}z_{t,i} - \sum_{j=1}^{L} p_{t,j}\frac{d}{dt}z_{t,j}\right)\\ &= \frac{1}{\sqrt{h}}p_{t,i}\left(\mathbf{x}_L^\top \mathbf{R}\mathbf{x}_i - \sum_{j=1}^{L} p_{t,j}\mathbf{x}_L^\top \mathbf{R}\mathbf{x}_j\right).\end{aligned}$$

The first equality follows from differentiating the softmax quotient and dividing both numerator terms by its denominator. The second equality follows by substituting the logit derivatives from Step 1.

Step 3. Differentiate the output. Define $\mathbf{m}_t := \mathbf{m}_{\mathbf{p}_t}(\mathbf{X})$. Then we have

$$\begin{aligned}\frac{d}{dt}\phi_{\mathbf{S}_t}(\mathbf{X}) &= \sum_{i=1}^{L}\left(\frac{d}{dt}p_{t,i}\right)\mathbf{x}_i^\top \mathbf{V}\\ &= \frac{1}{\sqrt{h}}\sum_{i=1}^{L} p_{t,i}\left(\mathbf{x}_L^\top \mathbf{R}\mathbf{x}_i - \mathbf{x}_L^\top \mathbf{R}\mathbf{m}_t\right)\mathbf{x}_i^\top \mathbf{V}\\ &= \frac{1}{\sqrt{h}}\mathbf{x}_L^\top \mathbf{R}\left(\sum_{i=1}^{L} p_{t,i}\mathbf{x}_i\mathbf{x}_i^\top - \mathbf{m}_t\mathbf{m}_t^\top\right)\mathbf{V}\\ &= \frac{1}{\sqrt{h}}\mathbf{x}_L^\top \mathbf{R}\mathbf{\Gamma}_t(\mathbf{X})\mathbf{V}.\end{aligned}$$

The first equality follows by differentiating the finite sum defining $\phi_{\mathbf{S}_t}(\mathbf{X})$. The second equality follows from Step 2 and the definition of $\mathbf{m}_t$. The third equality follows by distributing the sum and using $\sum_i p_{t,i}\mathbf{x}_i = \mathbf{m}_t$. The fourth equality follows from the definition of $\mathbf{\Gamma}_t(\mathbf{X})$.

Step 4. Integrate along the score segment. Every coordinate of $\phi_{\mathbf{S}_t}(\mathbf{X})$ is continuously differentiable in t because it is a finite sum of smooth functions with positive denominators. Integrating the derivative from Step 3 over $[0, 1]$ gives (10). This step uses $\mathbf{S}_t|_{t=1} = \mathbf{S}_*$ and $\mathbf{S}_t|_{t=0} = \mathbf{S}_\Delta$ and the fundamental theorem of calculus. The fixed base score $\mathbf{S}_0$ is distinct from the starting point $\mathbf{S}_t|_{t=0}$ of this interpolation. □

E.3.1 Lower Bound

Definition E.14 (Uniform ball queries)**.** *Define the unit ball and its uniform probability distribution by*

$$\begin{aligned}\mathbb{B}_d &:= \{\mathbf{g} \in \mathbb{R}^d : \|\mathbf{g}\|_2 \le 1\},\\ \mu_d &:= \textit{the uniform probability distribution on } \mathbb{B}_d.\end{aligned}$$

Let $\mathbf{g}_1, \ldots, \mathbf{g}_L$ be independent random vectors with distribution μ_d. Define

$$\begin{aligned}\mathbf{x}_i &:= \rho\mathbf{g}_i \qquad (1 \le i \le L),\\ \mathbf{G} &:= \begin{bmatrix}\mathbf{g}_1^\top\\ \vdots\\ \mathbf{g}_L^\top\end{bmatrix} \in \mathbb{R}^{L\times d},\\ \mathbf{X} &:= \rho\mathbf{G},\\ c_{L,d} &:= \frac{(L-1)^3}{L^4(d+2)^2(d+4)}.\end{aligned}$$

All update pairs considered in the population statements below are fixed before $\mathbf{g}_1, \ldots, \mathbf{g}_L$ are sampled. Every expectation in these statements is over these random vectors.

The assumptions $L \ge 2$ and $d \ge 1$ imply $c_{L,d} > 0$.

Lemma E.15 (Matrix norm and rank inequalities)**.** *For real matrices of compatible dimensions, the following statements hold.*

(i) $\|\mathbf{AB}\|_2 \le \|\mathbf{A}\|_2\|\mathbf{B}\|_2$ *and* $\|\mathbf{A}\|_2 \le \|\mathbf{A}\|_F$.
(ii) $\|\mathbf{A}+\mathbf{B}\|_2 \le \|\mathbf{A}\|_2 + \|\mathbf{B}\|_2$.
(iii) $\|\mathbf{v}\mathbf{w}^\top\|_2 = \|\mathbf{v}\|_2\|\mathbf{w}\|_2$ *for column vectors* $\mathbf{v}$ *and* $\mathbf{w}$.
(iv) $\operatorname{rank}(\mathbf{AB}) \le \min\{\operatorname{rank}(\mathbf{A}), \operatorname{rank}(\mathbf{B})\}$, $\operatorname{rank}(\mathbf{A}+\mathbf{B}) \le \operatorname{rank}(\mathbf{A}) + \operatorname{rank}(\mathbf{B})$, *and* $\operatorname{rank}(\mathbf{A}^\top) = \operatorname{rank}(\mathbf{A})$.

Proof. For (i), every vector $\mathbf{x}$ satisfies

$$\begin{aligned}\|\mathbf{ABx}\|_2 &\le \|\mathbf{A}\|_2\|\mathbf{Bx}\|_2 \\ &\le \|\mathbf{A}\|_2\|\mathbf{B}\|_2\|\mathbf{x}\|_2.\end{aligned}$$

Each inequality follows from the definition of the spectral norm. Taking the supremum over $\|\mathbf{x}\|_2 = 1$ proves the product inequality. If $\mathbf{a}_i^\top$ is the i-th row of $\mathbf{A}$, then

$$\begin{aligned}\|\mathbf{Ax}\|_2^2 &= \sum_i (\mathbf{a}_i^\top \mathbf{x})^2 \\ &\le \sum_i \|\mathbf{a}_i\|_2^2\|\mathbf{x}\|_2^2 \\ &= \|\mathbf{A}\|_F^2\|\mathbf{x}\|_2^2.\end{aligned}$$

The first equality follows from the definition of matrix-vector multiplication and the Euclidean norm. The inequality follows from the Cauchy–Schwarz inequality for each row. The last equality follows from the definition of the Frobenius norm. Taking the supremum over unit vectors proves $\|\mathbf{A}\|_2 \le \|\mathbf{A}\|_F$.

For (ii), the vector triangle inequality gives $\|(\mathbf{A}+\mathbf{B})\mathbf{x}\|_2 \le \|\mathbf{Ax}\|_2 + \|\mathbf{Bx}\|_2$. Bounding each term by its spectral norm and taking the supremum over unit vectors proves the claim.

For (iii), every unit vector $\mathbf{x}$ satisfies

$$\begin{aligned}\|\mathbf{v}\mathbf{w}^\top\mathbf{x}\|_2 &= |\mathbf{w}^\top\mathbf{x}|\|\mathbf{v}\|_2 \\ &\le \|\mathbf{w}\|_2\|\mathbf{v}\|_2.\end{aligned}$$

The equality follows because $\mathbf{w}^\top\mathbf{x}$ is a scalar. The inequality follows from Cauchy–Schwarz. If $\mathbf{w} \ne \mathbf{0}$, equality is attained at $\mathbf{x} = \mathbf{w}/\|\mathbf{w}\|_2$. If $\mathbf{w} = \mathbf{0}$, both sides of the claimed identity are zero.

For (iv), the image of the linear map $\mathbf{AB}$ is contained in the image of $\mathbf{A}$. Its dimension also cannot exceed the dimension of the image of $\mathbf{B}$ because it is the image of that space under $\mathbf{A}$. These two inclusions prove the product rank inequality. The image of $\mathbf{A}+\mathbf{B}$ is contained in the sum of the images of $\mathbf{A}$ and $\mathbf{B}$. The dimension of a sum of two subspaces is at most the sum of their dimensions, which proves the rank-sum inequality. Transposing a singular value decomposition preserves the number of nonzero singular values, so $\operatorname{rank}(\mathbf{A}^\top) = \operatorname{rank}(\mathbf{A})$. □

Lemma E.16 (Reverse triangle inequality)**.** *Let* $\mathbf{A}$ *and* $\mathbf{B}$ *be random vectors of the same finite dimension with finite second moments. Then we have*

$$\left(\mathbb{E}\|\mathbf{A}+\mathbf{B}\|_2^2\right)^{1/2} \ge \left(\mathbb{E}\|\mathbf{A}\|_2^2\right)^{1/2} - \left(\mathbb{E}\|\mathbf{B}\|_2^2\right)^{1/2}.$$

Proof. For random vectors $\mathbf{U}$ and $\mathbf{W}$ of the same dimension, define their Euclidean inner product by $\langle \mathbf{U}, \mathbf{W}\rangle := \sum_j U_j W_j$, where the sum is over their coordinates. If they have finite second moments, then

$$\begin{aligned}\mathbb{E}\|\mathbf{U}+\mathbf{W}\|_2^2 &= \mathbb{E}\|\mathbf{U}\|_2^2 + 2\mathbb{E}\langle \mathbf{U}, \mathbf{W}\rangle + \mathbb{E}\|\mathbf{W}\|_2^2 \\ &\le \mathbb{E}\|\mathbf{U}\|_2^2 + 2\mathbb{E}[\|\mathbf{U}\|_2\|\mathbf{W}\|_2] + \mathbb{E}\|\mathbf{W}\|_2^2 \\ &\le \left(\left(\mathbb{E}\|\mathbf{U}\|_2^2\right)^{1/2} + \left(\mathbb{E}\|\mathbf{W}\|_2^2\right)^{1/2}\right)^2.\end{aligned}$$

The first equality follows by expanding the squared norm. The first inequality follows from the vector Cauchy–Schwarz inequality. The second inequality follows from the scalar Cauchy–Schwarz inequality applied to $\|\mathbf{U}\|_2$ and $\|\mathbf{W}\|_2$. Taking square roots gives the triangle inequality for the displayed second-moment norm. Apply that inequality to $\mathbf{U} := \mathbf{A} + \mathbf{B}$ and $\mathbf{W} := -\mathbf{B}$, and then subtract $\left(\mathbb{E}\|\mathbf{B}\|_2^2\right)^{1/2}$ from both sides. □

Lemma E.17 (Moments of a uniform ball vector)**.** *Let $\mathbf{g} \sim \mu_d$, and let g_i denote the i-th coordinate of this single vector. For indices $i, j \in \{1, \dots, d\}$, define the Kronecker delta by $\delta_{ij} := 1$ if $i = j$ and $\delta_{ij} := 0$ otherwise. Then we have*

$$\begin{aligned}
\mathbb{E}[\mathbf{g}] &= \mathbf{0}, \\
\mathbb{E}[\mathbf{g}\mathbf{g}^\top] &= \frac{1}{d+2}\mathbf{I}_d, \\
\mathbb{E}[g_i g_j g_k g_\ell] &= \frac{\delta_{ij}\delta_{k\ell} + \delta_{ik}\delta_{j\ell} + \delta_{i\ell}\delta_{jk}}{(d+2)(d+4)} \qquad (i, j, k, \ell \in \{1, \dots, d\}).
\end{aligned}$$

Every first-order and third-order coordinate moment of $\mathbf{g}$ is zero.

Proof. **Step 1. Record sign symmetry.** The distribution μ_d is unchanged by reversing the sign of any coordinate. A coordinate monomial containing an odd power of a coordinate changes sign under that coordinate reversal. Its expectation is therefore zero. This proves $\mathbb{E}[\mathbf{g}] = \mathbf{0}$, all third-order coordinate moments, and the off-diagonal entries of $\mathbb{E}[\mathbf{g}\mathbf{g}^\top]$.

Step 2. Compute radial moments. First suppose $d \geq 2$. Define the unit sphere by $\mathbb{S}^{d-1} := \{\mathbf{v} \in \mathbb{R}^d : \|\mathbf{v}\|_2 = 1\}$. A uniform-ball vector has a representation $\mathbf{g} = \omega\mathbf{v}$, where $\mathbf{v}$ is uniform on $\mathbb{S}^{d-1}$, ω is independent of $\mathbf{v}$, and

$$\Pr(\omega \leq t) = t^d \qquad (0 \leq t \leq 1).$$

The distribution formula follows because the volume of a radius-t ball is t^d times the volume of the unit ball. Uniformity of the direction and its independence from the radius follow by polar coordinates, whose volume factor is a function of the radius alone. For each integer $a \geq 0$,

$$\begin{aligned}
\mathbb{E}[\omega^a] &= \int_0^1 t^a dt^{d-1}\, dt \\
&= \frac{d}{d+a}.
\end{aligned}$$

The first equality integrates against the density dt^{d-1}. The second equality evaluates the integral of dt^{d+a-1}.

Step 3. Compute the second angular moments. Coordinate symmetry makes the values $\mathbb{E}[v_i^2]$ equal. Since $\sum_i v_i^2 = 1$,

$$\begin{aligned}
\mathbb{E}[v_i^2] &= \frac{1}{d}, \\
\mathbb{E}[g_i^2] &= \mathbb{E}[\omega^2]\mathbb{E}[v_i^2] \\
&= \frac{d}{d+2} \cdot \frac{1}{d} \\
&= \frac{1}{d+2}.
\end{aligned}$$

The first equality follows by summing the equal coordinate moments. The second equality uses the independence of ω and $\mathbf{v}$. The third equality substitutes Step 2 and the first equality. The last equality cancels d. Combining these diagonal values with Step 1 gives the second-moment matrix.

Step 4. Compute the fourth angular moments. Define $a_4 := \mathbb{E}[v_1^4]$ and $b_4 := \mathbb{E}[v_1^2 v_2^2]$. Rotational invariance and sign symmetry imply

$$\begin{aligned} a_4 &= \mathbb{E}\left[\left(\frac{v_1 + v_2}{\sqrt{2}}\right)^4\right] \\ &= \frac{1}{4}(2a_4 + 6b_4). \end{aligned}$$

The first equality uses that $(\mathbf{e}_1 + \mathbf{e}_2)/\sqrt{2}$ is a unit vector, where $\mathbf{e}_i$ denotes the i-th coordinate vector. The second equality expands the fourth power, removes its odd-coordinate terms by sign symmetry, and uses coordinate symmetry for v_1^4 and v_2^4. Hence $a_4 = 3b_4$. Also,

$$\begin{aligned} 1 &= \mathbb{E}\left[\left(\sum_{i=1}^{d} v_i^2\right)^2\right] \\ &= da_4 + d(d-1)b_4. \end{aligned}$$

The first equality uses $\|\mathbf{v}\|_2 = 1$. The second equality expands the square and counts its d diagonal terms and $d(d-1)$ ordered off-diagonal terms. Solving these two scalar equations gives

$$\begin{aligned} b_4 &= \frac{1}{d(d+2)}, \\ a_4 &= \frac{3}{d(d+2)}. \end{aligned}$$

Multiplying each angular moment by $\mathbb{E}[\omega^4] = d/(d+4)$ gives the fourth moments of $\mathbf{g}$. The three Kronecker-delta terms equal 3 when all four indices are equal, equal 1 when the indices form two distinct pairs, and equal 0 in every other case. These three cases establish the stated fourth-moment formula.

Step 5. Check dimension one. When $d = 1$, the vector g is a scalar uniform on $[-1, 1]$. Direct integration gives

$$\begin{aligned} \mathbb{E}[g^2] &= \frac{1}{2}\int_{-1}^{1} t^2\, dt = \frac{1}{3}, \\ \mathbb{E}[g^4] &= \frac{1}{2}\int_{-1}^{1} t^4\, dt = \frac{1}{5}. \end{aligned}$$

The density is $1/2$ on $[-1, 1]$, which gives the first equality on each line. Evaluation of the integrals gives the last equality on each line. These values equal $1/(d+2)$ and $3/[(d+2)(d+4)]$ at $d = 1$. Sign symmetry proves the required odd-moment statements in this case as well. □

Lemma E.18 (Changes in covariance and softmax)**.** *Let $\mathbf{Y} \in \mathbb{R}^{L\times d}$ have rows $\mathbf{y}_i^\top$ with $\|\mathbf{y}_i\|_2 \le 1$. For $\mathbf{w}, \mathbf{w}' \in \mathcal{P}_L$,*

$$\begin{aligned} \mathbf{0} \preceq \mathbf{\Gamma}_{\mathbf{w}}(\mathbf{Y}) \preceq \mathbf{I}_d, \\ \|\mathbf{\Gamma}_{\mathbf{w}}(\mathbf{Y}) - \mathbf{\Gamma}_{\mathbf{w}'}(\mathbf{Y})\|_2 \le 3\|\mathbf{w} - \mathbf{w}'\|_1. \end{aligned}$$

For every $\mathbf{z}, \mathbf{z}' \in \mathbb{R}^L$,

$$\|\operatorname{softmax}(\mathbf{z}) - \operatorname{softmax}(\mathbf{z}')\|_1 \le 2\|\mathbf{z} - \mathbf{z}'\|_\infty.$$

In particular,

$$\|\mathbf{\Gamma}_{\operatorname{softmax}(\mathbf{z})}(\mathbf{Y}) - \mathbf{\Gamma}_{\mathbf{u}}(\mathbf{Y})\|_2 \le 6\|\mathbf{z}\|_\infty.$$

Proof. **Step 1. Bound the covariance quadratic form.** Fix $\mathbf{v} \in \mathbb{R}^d$ and define $\mathbf{m} := \mathbf{m}_{\mathbf{w}}(\mathbf{Y})$. Then we have

$$\begin{aligned}\mathbf{v}^\top \mathbf{\Gamma}_{\mathbf{w}}(\mathbf{Y})\mathbf{v} &= \sum_i w_i(\mathbf{v}^\top \mathbf{y}_i)^2 - (\mathbf{v}^\top \mathbf{m})^2 \\ &= \sum_i w_i(\mathbf{v}^\top \mathbf{y}_i - \mathbf{v}^\top \mathbf{m})^2 \\ &\geq 0.\end{aligned}$$

The first equality follows from the definition of $\mathbf{\Gamma}_{\mathbf{w}}(\mathbf{Y})$. The second equality expands the squares and uses $\sum_i w_i = 1$ and $\sum_i w_i \mathbf{y}_i = \mathbf{m}$. The inequality holds because every summand is nonnegative. For the upper bound,

$$\begin{aligned}\mathbf{v}^\top \mathbf{\Gamma}_{\mathbf{w}}(\mathbf{Y})\mathbf{v} &\leq \sum_i w_i(\mathbf{v}^\top \mathbf{y}_i)^2 \\ &\leq \sum_i w_i \|\mathbf{v}\|_2^2 \|\mathbf{y}_i\|_2^2 \\ &\leq \|\mathbf{v}\|_2^2.\end{aligned}$$

The first inequality removes the nonnegative term $(\mathbf{v}^\top \mathbf{m})^2$. The second inequality uses Cauchy–Schwarz. The third inequality uses $\|\mathbf{y}_i\|_2 \leq 1$ and $\sum_i w_i = 1$. These quadratic-form bounds prove $\mathbf{0} \preceq \mathbf{\Gamma}_{\mathbf{w}}(\mathbf{Y}) \preceq \mathbf{I}_d$. Since $\mathbf{\Gamma}_{\mathbf{w}}(\mathbf{Y})$ is symmetric, its eigenvalues lie in $[0, 1]$, so its spectral norm is at most 1.

Step 2. Bound the difference between weighted means. Define $\mathbf{m}' := \mathbf{m}_{\mathbf{w}'}(\mathbf{Y})$. The vector triangle inequality gives

$$\begin{aligned}\|\mathbf{m}\|_2 &\leq \sum_i w_i \|\mathbf{y}_i\|_2 \leq 1, \\ \|\mathbf{m}'\|_2 &\leq \sum_i w_i' \|\mathbf{y}_i\|_2 \leq 1, \\ \|\mathbf{m} - \mathbf{m}'\|_2 &\leq \sum_i |w_i - w_i'| \|\mathbf{y}_i\|_2 \\ &\leq \|\mathbf{w} - \mathbf{w}'\|_1.\end{aligned}$$

The last inequality on each line uses $\|\mathbf{y}_i\|_2 \leq 1$. The first two lines also use the unit total mass of $\mathbf{w}$ and $\mathbf{w}'$.

Step 3. Bound the covariance difference. The mean outer products satisfy

$$\begin{aligned}\mathbf{m}\mathbf{m}^\top - \mathbf{m}'\mathbf{m}'^\top &= (\mathbf{m} - \mathbf{m}')\mathbf{m}^\top + \mathbf{m}'(\mathbf{m} - \mathbf{m}')^\top, \\ \|\mathbf{m}\mathbf{m}^\top - \mathbf{m}'\mathbf{m}'^\top\|_2 &\leq \|\mathbf{m} - \mathbf{m}'\|_2 \|\mathbf{m}\|_2 + \|\mathbf{m}'\|_2 \|\mathbf{m} - \mathbf{m}'\|_2 \\ &\leq 2\|\mathbf{w} - \mathbf{w}'\|_1.\end{aligned}$$

The first equality follows by expansion and cancellation. The first inequality uses the spectral-norm triangle inequality and the rank-one norm identity in Lemma E.15. The last inequality substitutes Step 2. Consequently,

$$\begin{aligned}\|\mathbf{\Gamma}_{\mathbf{w}}(\mathbf{Y}) - \mathbf{\Gamma}_{\mathbf{w}'}(\mathbf{Y})\|_2 &\leq \sum_i |w_i - w_i'| \|\mathbf{y}_i \mathbf{y}_i^\top\|_2 + \|\mathbf{m}\mathbf{m}^\top - \mathbf{m}'\mathbf{m}'^\top\|_2 \\ &\leq \|\mathbf{w} - \mathbf{w}'\|_1 + 2\|\mathbf{w} - \mathbf{w}'\|_1 \\ &= 3\|\mathbf{w} - \mathbf{w}'\|_1.\end{aligned}$$

The first inequality applies the triangle inequality to the covariance definition. The second inequality uses $\|\mathbf{y}_i \mathbf{y}_i^\top\|_2 = \|\mathbf{y}_i\|_2^2 \leq 1$ and the bound just proved for the mean term. The last equality combines the two scalar terms.

Step 4. Bound the softmax difference. Define $\mathbf{b} := \mathbf{z} - \mathbf{z}'$, $\mathbf{q}(t) := \mathrm{softmax}(\mathbf{z}' + t\mathbf{b})$ for $0 \le t \le 1$, and $\bar{b}_t := \sum_j q_j(t) b_j$. Differentiating the quotient defining softmax gives

$$\begin{aligned} q_i'(t) &= q_i(t)(b_i - \bar{b}_t), \\ \|\mathbf{q}'(t)\|_1 &= \sum_i q_i(t)|b_i - \bar{b}_t| \\ &\le \sum_i q_i(t)(|b_i| + |\bar{b}_t|) \\ &\le 2\|\mathbf{b}\|_\infty. \end{aligned}$$

The first equality follows from the quotient rule. The second equality uses $q_i(t) \ge 0$. The first inequality is the scalar triangle inequality. The last inequality uses $|b_i| \le \|\mathbf{b}\|_\infty$, $|\bar{b}_t| \le \sum_j q_j(t)|b_j| \le \|\mathbf{b}\|_\infty$, and $\sum_i q_i(t) = 1$. Integrating coordinatewise gives

$$\begin{aligned} \|\mathbf{q}(1) - \mathbf{q}(0)\|_1 &\le \int_0^1 \|\mathbf{q}'(t)\|_1 \, dt \\ &\le 2\|\mathbf{b}\|_\infty. \end{aligned}$$

The first inequality follows from the fundamental theorem of calculus and the triangle inequality for each coordinate integral. The second inequality uses the bound on $\|\mathbf{q}'(t)\|_1$. This proves the softmax claim. Set $\mathbf{z}' = \mathbf{0}$, use $\mathrm{softmax}(\mathbf{0}) = \mathbf{u}$, and then apply Step 3 to obtain the final covariance perturbation bound. □

Lemma E.19 (Second moment of a centered quadratic term)**.** *Let $\mathbf{b} \sim \mu_d$, and define $s := 1/(d+2)$ and $q_4 := 1/[(d+2)(d+4)]$. For every fixed $\mathbf{v} \in \mathbb{R}^d$ and $\mathbf{V} \in \mathbb{R}^{d \times p}$,*

$$\mathbb{E}\|\mathbf{v}^\top(\mathbf{b}\mathbf{b}^\top - s\mathbf{I}_d)\mathbf{V}\|_2^2 = q_4\|\mathbf{v}\|_2^2\|\mathbf{V}\|_F^2 + (2q_4 - s^2)\|\mathbf{v}^\top\mathbf{V}\|_2^2.$$

Moreover,

$$2q_4 - s^2 = \frac{d}{(d+2)^2(d+4)} > 0.$$

Proof. **Step 1. Evaluate the uncentered fourth moment.** Define $\mathbf{B} := \mathbf{V}\mathbf{V}^\top \in \mathbb{R}^{d \times d}$ and $\mathrm{tr}(\mathbf{B}) := \sum_{i=1}^d B_{ii}$. For $j, k \in \{1, \dots, d\}$,

$$\begin{aligned} \left[\mathbb{E}(\mathbf{b}\mathbf{b}^\top\mathbf{B}\mathbf{b}\mathbf{b}^\top)\right]_{jk} &= \sum_{\alpha=1}^d \sum_{\beta=1}^d B_{\alpha\beta}\mathbb{E}[b_j b_\alpha b_\beta b_k] \\ &= q_4 \sum_{\alpha=1}^d \sum_{\beta=1}^d B_{\alpha\beta}(\delta_{j\alpha}\delta_{\beta k} + \delta_{j\beta}\delta_{\alpha k} + \delta_{jk}\delta_{\alpha\beta}) \\ &= q_4(B_{jk} + B_{kj} + \delta_{jk}\,\mathrm{tr}(\mathbf{B})) \\ &= q_4(2B_{jk} + \delta_{jk}\,\mathrm{tr}(\mathbf{B})). \end{aligned}$$

The first equality expands the matrix product entrywise. The second equality applies Lemma E.17. The third equality evaluates the Kronecker-delta sums. The last equality uses $\mathbf{B}^\top = \mathbf{B}$. Therefore,

$$\mathbb{E}(\mathbf{b}\mathbf{b}^\top\mathbf{B}\mathbf{b}\mathbf{b}^\top) = q_4(2\mathbf{B} + \mathrm{tr}(\mathbf{B})\mathbf{I}_d).$$

Step 2. Center the quadratic matrix. The second-moment formula in Lemma E.17 gives

$$\begin{aligned} \mathbb{E}[(\mathbf{b}\mathbf{b}^\top - s\mathbf{I}_d)\mathbf{B}(\mathbf{b}\mathbf{b}^\top - s\mathbf{I}_d)] &= \mathbb{E}[\mathbf{b}\mathbf{b}^\top\mathbf{B}\mathbf{b}\mathbf{b}^\top] - s\mathbb{E}[\mathbf{b}\mathbf{b}^\top]\mathbf{B} - s\mathbf{B}\mathbb{E}[\mathbf{b}\mathbf{b}^\top] + s^2\mathbf{B} \\ &= q_4(2\mathbf{B} + \mathrm{tr}(\mathbf{B})\mathbf{I}_d) - s^2\mathbf{B} \\ &= q_4\,\mathrm{tr}(\mathbf{B})\mathbf{I}_d + (2q_4 - s^2)\mathbf{B}. \end{aligned}$$

The first equality expands the product and uses linearity of expectation. The second equality substitutes Step 1 and $\mathbb{E}[\mathbf{b}\mathbf{b}^\top] = s\mathbf{I}_d$, after which the three terms involving $s^2\mathbf{B}$ combine to $-s^2\mathbf{B}$. The last equality groups the multiples of $\mathbf{B}$ and $\mathbf{I}_d$.

Step 3. Evaluate the desired scalar. The definition of the Euclidean norm gives

$$\begin{aligned}\mathbb{E}\|\mathbf{v}^\top(\mathbf{b}\mathbf{b}^\top - s\mathbf{I}_d)\mathbf{V}\|_2^2 &= \mathbf{v}^\top\mathbb{E}[(\mathbf{b}\mathbf{b}^\top - s\mathbf{I}_d)\mathbf{B}(\mathbf{b}\mathbf{b}^\top - s\mathbf{I}_d)]\mathbf{v}\\ &= q_4\,\mathrm{tr}(\mathbf{B})\|\mathbf{v}\|_2^2 + (2q_4 - s^2)\mathbf{v}^\top\mathbf{B}\mathbf{v}\\ &= q_4\|\mathbf{v}\|_2^2\|\mathbf{V}\|_F^2 + (2q_4 - s^2)\|\mathbf{v}^\top\mathbf{V}\|_2^2.\end{aligned}$$

The first equality uses $\mathbf{B} = \mathbf{V}\mathbf{V}^\top$ and the symmetry of $\mathbf{b}\mathbf{b}^\top - s\mathbf{I}_d$. The second equality substitutes Step 2. The last equality uses $\mathrm{tr}(\mathbf{V}\mathbf{V}^\top) = \sum_{i,j} V_{ij}^2 = \|\mathbf{V}\|_F^2$ and $\mathbf{v}^\top\mathbf{V}\mathbf{V}^\top\mathbf{v} = \|\mathbf{v}^\top\mathbf{V}\|_2^2$. Finally,

$$\begin{aligned}2q_4 - s^2 &= \frac{2}{(d+2)(d+4)} - \frac{1}{(d+2)^2}\\ &= \frac{2(d+2) - (d+4)}{(d+2)^2(d+4)}\\ &= \frac{d}{(d+2)^2(d+4)}\\ &> 0.\end{aligned}$$

The first equality substitutes the definitions. The second equality uses a common denominator. The third equality simplifies the numerator. The inequality uses $d \geq 1$. □

Lemma E.20 (Lower bound for the covariance term)**.** *Let* $\mathbf{G}$ *have the distribution in Definition E.14. For every fixed* $\mathbf{R} \in \mathbb{R}^{d\times d}$ *and* $\mathbf{V} \in \mathbb{R}^{d\times p}$*,*

$$\mathbb{E}\|\mathbf{g}_L^\top\mathbf{R}\boldsymbol{\Gamma}_{\mathbf{u}}(\mathbf{G})\mathbf{V}\|_2^2 \geq c_{L,d}\|\mathbf{R}\|_F^2\|\mathbf{V}\|_F^2.$$

Proof. **Step 1. Expand the empirical covariance.** Define

$$\begin{aligned}&m := L-1, && \mathbf{a} := \mathbf{g}_L, && \mathbf{b}_i := \mathbf{g}_i \quad (1 \leq i \leq m),\\ &\lambda := \frac{m}{L^2}, && s := \frac{1}{d+2}, && q_4 := \frac{1}{(d+2)(d+4)}.\end{aligned}$$

The uniform weighted covariance satisfies

$$\begin{aligned}\boldsymbol{\Gamma}_{\mathbf{u}}(\mathbf{G}) &= \frac{1}{L}\left(\mathbf{a}\mathbf{a}^\top + \sum_{i=1}^m \mathbf{b}_i\mathbf{b}_i^\top\right) - \frac{1}{L^2}\left(\mathbf{a} + \sum_{i=1}^m \mathbf{b}_i\right)\left(\mathbf{a} + \sum_{i=1}^m \mathbf{b}_i\right)^\top\\ &= \lambda\left(\mathbf{a}\mathbf{a}^\top + \sum_{i=1}^m \mathbf{b}_i\mathbf{b}_i^\top\right) - \frac{1}{L^2}\left(\sum_{i\neq j}\mathbf{b}_i\mathbf{b}_j^\top + \sum_{i=1}^m(\mathbf{a}\mathbf{b}_i^\top + \mathbf{b}_i\mathbf{a}^\top)\right).\end{aligned}$$

The first equality substitutes $u_i = 1/L$ into the weighted-covariance definition. The second equality expands the outer product and combines the coefficient $1/L - 1/L^2 = (L-1)/L^2 = \lambda$ for every diagonal outer product. The sum $\sum_{i\neq j}$ is over ordered pairs of distinct indices in $\{1,\dots,m\}$.

Step 2. Separate centered quadratic terms. Define row vectors in $\mathbb{R}^{1\times p}$ by

$$\begin{aligned}\mathbf{Y} &:= \mathbf{a}^\top\mathbf{R}\boldsymbol{\Gamma}_{\mathbf{u}}(\mathbf{G})\mathbf{V},\\ \mathbf{H}_i &:= \lambda\mathbf{a}^\top\mathbf{R}(\mathbf{b}_i\mathbf{b}_i^\top - s\mathbf{I}_d)\mathbf{V} \quad (1 \leq i \leq m),\\ \mathbf{Z} &:= \mathbf{Y} - \sum_{i=1}^m \mathbf{H}_i.\end{aligned}$$

Substitution of Step 1 gives

$$\begin{aligned}\mathbf{Z} = \lambda\mathbf{a}^\top\mathbf{R}\mathbf{a}\mathbf{a}^\top\mathbf{V} + m\lambda s\,\mathbf{a}^\top\mathbf{R}\mathbf{V}\\ - \frac{1}{L^2}\mathbf{a}^\top\mathbf{R}\left(\sum_{i\neq j}\mathbf{b}_i\mathbf{b}_j^\top\right)\mathbf{V}\\ - \frac{1}{L^2}\mathbf{a}^\top\mathbf{R}\left(\sum_{i=1}^m(\mathbf{a}\mathbf{b}_i^\top + \mathbf{b}_i\mathbf{a}^\top)\right)\mathbf{V}.\end{aligned}$$

This equality follows because subtracting $\sum_i \mathbf{H}_i$ removes every term $\lambda \mathbf{a}^\top \mathbf{R} \mathbf{b}_i \mathbf{b}_i^\top \mathbf{V}$ and adds $m\lambda s\, \mathbf{a}^\top \mathbf{R}\mathbf{V}$.

Step 3. Prove pairwise orthogonality of the quadratic terms. Conditional on $\mathbf{a}$, the row vectors $\mathbf{H}_1, \ldots, \mathbf{H}_m$ are independent because each $\mathbf{H}_i$ is a function of the independent vector $\mathbf{b}_i$ and the fixed conditional value of $\mathbf{a}$. Moreover,

$$\begin{aligned}\mathbb{E}[\mathbf{H}_i \mid \mathbf{a}] &= \lambda \mathbf{a}^\top \mathbf{R} \left(\mathbb{E}[\mathbf{b}_i \mathbf{b}_i^\top] - s\mathbf{I}_d\right) \mathbf{V} \\ &= \mathbf{0}.\end{aligned}$$

The first equality uses independence of $\mathbf{b}_i$ and $\mathbf{a}$ and linearity of conditional expectation. The second equality applies Lemma E.17. For $i \neq j$,

$$\begin{aligned}\mathbb{E}[\langle \mathbf{H}_i, \mathbf{H}_j \rangle \mid \mathbf{a}] &= \langle \mathbb{E}[\mathbf{H}_i \mid \mathbf{a}], \mathbb{E}[\mathbf{H}_j \mid \mathbf{a}] \rangle \\ &= 0.\end{aligned}$$

The first equality uses conditional independence, applied coordinatewise to the inner product. The second equality uses the conditional means just computed. Taking expectation over $\mathbf{a}$ proves $\mathbb{E}\langle \mathbf{H}_i, \mathbf{H}_j \rangle = 0$.

Step 4. Prove orthogonality to the remainder. Fix $i \in \{1, \ldots, m\}$ and condition on $\mathbf{a}$ and all $\mathbf{b}_j$ with $j \neq i$. In the displayed expression for $\mathbf{Z}$ in Step 2, each term involving $\mathbf{b}_i$ is linear in $\mathbf{b}_i$. There is no term $\mathbf{b}_i \mathbf{b}_i^\top$ because the sum over two indices has $i \neq j$. Thus $\mathbf{Z}$ is an affine function of $\mathbf{b}_i$ under this conditioning.

For each $\ell \in \{1, \ldots, d\}$, let $\mathbf{e}_\ell$ be the ℓ-th coordinate vector. Define the coefficients of this affine function by

$$\begin{aligned}\mathbf{Z}_{i,0} &:= \mathbf{Z}|_{\mathbf{b}_i = \mathbf{0}}, \\ \mathbf{Z}_{i,\ell} &:= \mathbf{Z}|_{\mathbf{b}_i = \mathbf{e}_\ell} - \mathbf{Z}|_{\mathbf{b}_i = \mathbf{0}} \qquad (1 \leq \ell \leq d).\end{aligned}$$

The coefficients are fixed under the conditioning, and

$$\mathbf{Z} = \mathbf{Z}_{i,0} + \sum_{\ell=1}^{d} (\mathbf{b}_i)_\ell \mathbf{Z}_{i,\ell}.$$

The conditional expectation of $\mathbf{H}_i$ is zero by the same second-moment calculation as in Step 3. For each ℓ,

$$\begin{aligned}\mathbb{E}[(\mathbf{b}_i)_\ell \mathbf{H}_i \mid \mathbf{a}, \{\mathbf{b}_j : j \neq i\}] &= \lambda \mathbf{a}^\top \mathbf{R} \left(\mathbb{E}[(\mathbf{b}_i)_\ell \mathbf{b}_i \mathbf{b}_i^\top] - s\mathbb{E}[(\mathbf{b}_i)_\ell] \mathbf{I}_d\right) \mathbf{V} \\ &= \mathbf{0}.\end{aligned}$$

The first equality substitutes $\mathbf{H}_i$ and uses independence of $\mathbf{b}_i$ from the conditioned variables. The second equality follows because every entry in the first expectation is a third-order coordinate moment, and the second expectation is a first-order coordinate moment. Both types vanish by Lemma E.17. Therefore,

$$\begin{aligned}\mathbb{E}[\langle \mathbf{H}_i, \mathbf{Z} \rangle \mid \mathbf{a}, \{\mathbf{b}_j : j \neq i\}] &= \langle \mathbb{E}[\mathbf{H}_i \mid \mathbf{a}, \{\mathbf{b}_j : j \neq i\}], \mathbf{Z}_{i,0} \rangle \\ &\quad + \sum_{\ell=1}^{d} \langle \mathbb{E}[(\mathbf{b}_i)_\ell \mathbf{H}_i \mid \mathbf{a}, \{\mathbf{b}_j : j \neq i\}], \mathbf{Z}_{i,\ell} \rangle \\ &= 0.\end{aligned}$$

The first equality expands the affine representation of $\mathbf{Z}$ and uses linearity of conditional expectation. The last equality substitutes the zero conditional moments. Taking expectation over the conditioned variables proves $\mathbb{E}\langle \mathbf{H}_i, \mathbf{Z} \rangle = 0$.

Step 5. Remove the orthogonal remainder. Since $\mathbf{Y} = \mathbf{Z} + \sum_i \mathbf{H}_i$,

$$\begin{aligned}\mathbb{E}\|\mathbf{Y}\|_2^2 &= \mathbb{E}\|\mathbf{Z}\|_2^2 + \sum_i \mathbb{E}\|\mathbf{H}_i\|_2^2 \\ &\quad + 2\sum_{i<j} \mathbb{E}\langle \mathbf{H}_i, \mathbf{H}_j\rangle + 2\sum_i \mathbb{E}\langle \mathbf{H}_i, \mathbf{Z}\rangle \\ &= \mathbb{E}\|\mathbf{Z}\|_2^2 + \sum_i \mathbb{E}\|\mathbf{H}_i\|_2^2 \\ &\geq \sum_i \mathbb{E}\|\mathbf{H}_i\|_2^2.\end{aligned}$$

The first equality expands the squared norm. The second equality applies Steps 3 and 4. The inequality removes the nonnegative term $\mathbb{E}\|\mathbf{Z}\|_2^2$.

Step 6. Lower-bound one centered term. Conditional on $\mathbf{a}$, apply Lemma E.19 with $\mathbf{v} := \mathbf{R}^\top \mathbf{a}$. This gives

$$\begin{aligned}\mathbb{E}[\|\mathbf{H}_i\|_2^2 \mid \mathbf{a}] &= \lambda^2 q_4 \|\mathbf{R}^\top \mathbf{a}\|_2^2 \|\mathbf{V}\|_F^2 + \lambda^2 (2q_4 - s^2)\|\mathbf{a}^\top \mathbf{R}\mathbf{V}\|_2^2 \\ &\geq \lambda^2 q_4 \|\mathbf{R}^\top \mathbf{a}\|_2^2 \|\mathbf{V}\|_F^2.\end{aligned}$$

The equality substitutes the definition of $\mathbf{H}_i$ into that lemma. The inequality removes a nonnegative term because $2q_4 - s^2 > 0$. To average the remaining term over $\mathbf{a}$, use

$$\begin{aligned}\mathbb{E}\|\mathbf{R}^\top \mathbf{a}\|_2^2 &= \sum_{k=1}^d \sum_{j=1}^d \sum_{\ell=1}^d R_{jk} R_{\ell k} \mathbb{E}[a_j a_\ell] \\ &= s \sum_{k=1}^d \sum_{j=1}^d R_{jk}^2 \\ &= s\|\mathbf{R}\|_F^2.\end{aligned}$$

The first equality expands every coordinate of $\mathbf{R}^\top \mathbf{a}$ and its square. The second equality uses $\mathbb{E}[a_j a_\ell] = s\delta_{j\ell}$ from Lemma E.17. The last equality is the definition of the Frobenius norm. Consequently,

$$\mathbb{E}\|\mathbf{H}_i\|_2^2 \geq \lambda^2 q_4 s \|\mathbf{R}\|_F^2 \|\mathbf{V}\|_F^2.$$

This inequality follows by taking expectation of the preceding conditional lower bound.

Step 7. Sum the contributions. Step 5 and Step 6 imply

$$\begin{aligned}\mathbb{E}\|\mathbf{Y}\|_2^2 &\geq m\lambda^2 q_4 s \|\mathbf{R}\|_F^2 \|\mathbf{V}\|_F^2 \\ &= \frac{m^3}{L^4(d+2)^2(d+4)} \|\mathbf{R}\|_F^2 \|\mathbf{V}\|_F^2 \\ &= c_{L,d} \|\mathbf{R}\|_F^2 \|\mathbf{V}\|_F^2.\end{aligned}$$

The first inequality sums m identical lower bounds. The first equality substitutes the definitions of λ, q_4, and s. The last equality uses $m = L - 1$ and the definition of $c_{L,d}$. Finally, $\mathbf{Y} = \mathbf{g}_L^\top \mathbf{R}\mathbf{\Gamma}_\mathbf{u}(\mathbf{G})\mathbf{V}$ by its definition, which proves the lemma. □

Lemma E.21 (Rank of the score update)**.** *For every* $(\Delta\mathbf{Q}, \Delta\mathbf{K}) \in \mathcal{A}_{r,M}$,

$$\begin{aligned}\operatorname{rank}(\mathbf{S}_\Delta - \mathbf{S}_0) &\leq 2r, \\ \|\mathbf{S}_* - \mathbf{S}_\Delta\|_F^2 &\geq \sum_{j=2r+1}^d \sigma_j(\mathbf{S}_* - \mathbf{S}_0)^2.\end{aligned}$$

Proof. **Step 1. Expand the score update.** The definitions give

$$\begin{aligned}\mathbf{S}_\Delta - \mathbf{S}_0 &= (\mathbf{Q}_0 + \Delta\mathbf{Q})(\mathbf{K}_0 + \Delta\mathbf{K})^\top - \mathbf{Q}_0\mathbf{K}_0^\top \\ &= \Delta\mathbf{Q}\mathbf{K}_0^\top + \mathbf{Q}_0\Delta\mathbf{K}^\top + \Delta\mathbf{Q}\Delta\mathbf{K}^\top \\ &= \Delta\mathbf{Q}\mathbf{K}_0^\top + (\mathbf{Q}_0 + \Delta\mathbf{Q})\Delta\mathbf{K}^\top.\end{aligned}$$

The first equality substitutes $\mathbf{S}_\Delta$ and $\mathbf{S}_0$. The second equality expands the product and cancels $\mathbf{Q}_0\mathbf{K}_0^\top$. The last equality factors the two terms containing $\Delta\mathbf{K}^\top$.

Step 2. Bound the rank. Lemma E.15 implies

$$\begin{aligned}\operatorname{rank}(\mathbf{S}_\Delta - \mathbf{S}_0) &\le \operatorname{rank}(\Delta\mathbf{Q}\mathbf{K}_0^\top) + \operatorname{rank}((\mathbf{Q}_0 + \Delta\mathbf{Q})\Delta\mathbf{K}^\top) \\ &\le \operatorname{rank}(\Delta\mathbf{Q}) + \operatorname{rank}(\Delta\mathbf{K}^\top) \\ &= \operatorname{rank}(\Delta\mathbf{Q}) + \operatorname{rank}(\Delta\mathbf{K}) \\ &\le 2r.\end{aligned}$$

The first inequality applies the rank-sum inequality to Step 1. The second inequality applies the product rank inequality to each term. The equality uses invariance of rank under transposition. The final inequality uses membership in $\mathcal{A}_{r,M}$.

Step 3. Apply low-rank approximation. Because $\mathbf{S}_\Delta - \mathbf{S}_0$ has rank at most $2r$,

$$\begin{aligned}\|\mathbf{S}_* - \mathbf{S}_\Delta\|_F^2 &= \|(\mathbf{S}_* - \mathbf{S}_0) - (\mathbf{S}_\Delta - \mathbf{S}_0)\|_F^2 \\ &\ge \inf_{\substack{\mathbf{B}\in\mathbb{R}^{d\times d} \\ \operatorname{rank}(\mathbf{B})\le 2r}} \|(\mathbf{S}_* - \mathbf{S}_0) - \mathbf{B}\|_F^2 \\ &= \sum_{j=2r+1}^{d} \sigma_j(\mathbf{S}_* - \mathbf{S}_0)^2.\end{aligned}$$

The first equality adds and subtracts $\mathbf{S}_0$. The inequality compares one feasible rank-$2r$ approximation with the infimum over all matrices of rank at most $2r$. The last equality applies Lemma E.1. This argument only uses that realizable score updates have rank at most $2r$. It does not assert that every matrix of rank at most $2r$ is realizable. □

Theorem E.22 (Lower bound for attention)**.** *Let the parameters and update family satisfy Definition E.10, let the outputs be those in Definition E.11, and let* $\mathbf{X}$ *have the distribution in Definition E.14. Suppose*

$$\frac{\rho^2 M}{\sqrt{h}} \le \frac{\sqrt{c_{L,d}}}{12}.$$

Then every fixed $(\Delta\mathbf{Q}, \Delta\mathbf{K}) \in \mathcal{A}_{r,M}$ *satisfies*

$$\begin{aligned}\mathbb{E}\|\phi_{\mathbf{S}_*}(\mathbf{X}) - \phi_{\mathbf{S}_\Delta}(\mathbf{X})\|_2^2 &\ge \frac{\rho^6 c_{L,d}}{4h}\|\mathbf{V}\|_F^2\|\mathbf{S}_* - \mathbf{S}_\Delta\|_F^2 \\ &\ge \frac{\rho^6 c_{L,d}}{4h}\|\mathbf{V}\|_F^2 \sum_{j=2r+1}^{d} \sigma_j(\mathbf{S}_* - \mathbf{S}_0)^2.\end{aligned} \tag{11}$$

Consequently,

$$\begin{aligned}&\inf_{(\Delta\mathbf{Q},\Delta\mathbf{K})\in\mathcal{A}_{r,M}} \sup_{\substack{\mathbf{X}\in\mathbb{R}^{L\times d} \\ \max_i \|\mathbf{x}_i\|_2\le\rho}} \|\phi_{\mathbf{S}_*}(\mathbf{X}) - \phi_{\mathbf{S}_\Delta}(\mathbf{X})\|_2 \\ &\qquad \ge \frac{\rho^3\sqrt{c_{L,d}}}{2\sqrt{h}}\|\mathbf{V}\|_F \left(\sum_{j=2r+1}^{d} \sigma_j(\mathbf{S}_* - \mathbf{S}_0)^2\right)^{1/2}.\end{aligned} \tag{12}$$

Proof. **Step 1. Fix the candidate and bound its score segment.** Fix $(\Delta\mathbf{Q}, \Delta\mathbf{K}) \in \mathcal{A}_{r,M}$. Define

$$\begin{aligned}\mathbf{R} &:= \mathbf{S}_* - \mathbf{S}_\Delta, \\ \epsilon &:= \frac{\rho^2 M}{\sqrt{h}}.\end{aligned}$$

Use the interpolation $\mathbf{S}_t$, weights $\mathbf{p}_t$, and covariance $\mathbf{\Gamma}_t(\mathbf{X})$ from Lemma E.13. For every $t \in [0,1]$,

$$\begin{aligned}\|\mathbf{S}_t\|_2 &= \|(1-t)\mathbf{S}_\Delta + t\mathbf{S}_*\|_2 \\ &\le (1-t)\|\mathbf{S}_\Delta\|_2 + t\|\mathbf{S}_*\|_2 \\ &\le M.\end{aligned}$$

The equality substitutes the interpolation definition. The first inequality uses the spectral-norm triangle inequality and homogeneity. The last inequality uses the score-norm bounds and $(1-t)+t=1$.

Step 2. Express the exact identity in unit-ball coordinates. For every $\mathbf{w} \in \mathcal{P}_L$, the relation $\mathbf{X} = \rho\mathbf{G}$ implies

$$\begin{aligned}\mathbf{m}_\mathbf{w}(\mathbf{X}) &= \sum_i w_i \rho \mathbf{g}_i = \rho \mathbf{m}_\mathbf{w}(\mathbf{G}), \\ \mathbf{\Gamma}_\mathbf{w}(\mathbf{X}) &= \sum_i w_i \rho^2 \mathbf{g}_i \mathbf{g}_i^\top - \rho^2 \mathbf{m}_\mathbf{w}(\mathbf{G})\mathbf{m}_\mathbf{w}(\mathbf{G})^\top \\ &= \rho^2 \mathbf{\Gamma}_\mathbf{w}(\mathbf{G}).\end{aligned}$$

The first line substitutes $\mathbf{x}_i = \rho\mathbf{g}_i$ into the weighted mean. The next equality substitutes this mean and the same relation into the covariance definition. The last equality factors out ρ^2. Taking $\mathbf{w} = \mathbf{p}_t$ in this identity and using $\mathbf{x}_L = \rho\mathbf{g}_L$ in (10) gives

$$\phi_{\mathbf{S}_*}(\mathbf{X}) - \phi_{\mathbf{S}_\Delta}(\mathbf{X}) = \frac{\rho^3}{\sqrt{h}} \mathbf{g}_L^\top \mathbf{R} \left(\int_0^1 \mathbf{\Gamma}_{\mathbf{p}_t}(\mathbf{G})\, dt \right) \mathbf{V}.$$

Step 3. Bound the logit and covariance perturbations. For each i and t, define $z_{t,i} := [\mathbf{z}_{\mathbf{S}_t}(\mathbf{X})]_i$. Since $\|\mathbf{g}_i\|_2 \le 1$ almost surely,

$$\begin{aligned}|z_{t,i}| &= \frac{\rho^2}{\sqrt{h}} |\mathbf{g}_L^\top \mathbf{S}_t \mathbf{g}_i| \\ &\le \frac{\rho^2}{\sqrt{h}} \|\mathbf{g}_L\|_2 \|\mathbf{S}_t \mathbf{g}_i\|_2 \\ &\le \frac{\rho^2}{\sqrt{h}} \|\mathbf{g}_L\|_2 \|\mathbf{S}_t\|_2 \|\mathbf{g}_i\|_2 \\ &\le \frac{\rho^2 M}{\sqrt{h}} \\ &= \epsilon.\end{aligned}$$

The equality substitutes the logit definition and $\mathbf{x}_i = \rho\mathbf{g}_i$. The first inequality uses Cauchy–Schwarz. The second inequality uses the spectral norm. The third inequality uses Step 1 and the unit-ball bounds. The last equality is the definition of ϵ. Hence Lemma E.18 gives

$$\begin{aligned}\|\mathbf{\Gamma}_{\mathbf{p}_t}(\mathbf{G}) - \mathbf{\Gamma}_\mathbf{u}(\mathbf{G})\|_2 &\le 6\|\mathbf{z}_{\mathbf{S}_t}(\mathbf{X})\|_\infty \\ &\le 6\epsilon.\end{aligned}$$

The first inequality applies that lemma with $\mathbf{Y} = \mathbf{G}$ and $\mathbf{p}_t = \mathrm{softmax}(\mathbf{z}_{\mathbf{S}_t}(\mathbf{X}))$. The last inequality uses the coordinate bounds just proved.

Step 4. Separate the uniform-weight term. Define random row vectors by

$$\begin{aligned}\mathbf{A}(\mathbf{X}) &:= \frac{\rho^3}{\sqrt{h}} \mathbf{g}_L^\top \mathbf{R} \mathbf{\Gamma}_\mathbf{u}(\mathbf{G}) \mathbf{V}, \\ \mathbf{E}(\mathbf{X}) &:= \frac{\rho^3}{\sqrt{h}} \mathbf{g}_L^\top \mathbf{R} \left(\int_0^1 [\mathbf{\Gamma}_{\mathbf{p}_t}(\mathbf{G}) - \mathbf{\Gamma}_\mathbf{u}(\mathbf{G})]\, dt \right) \mathbf{V}.\end{aligned}$$

Linearity of integration and Step 2 imply

$$\phi_{\mathbf{S}_*}(\mathbf{X}) - \phi_{\mathbf{S}_\Delta}(\mathbf{X}) = \mathbf{A}(\mathbf{X}) + \mathbf{E}(\mathbf{X}).$$

The row vectors have finite second moments because the $\mathbf{g}_i$ are bounded and all fixed matrices have finite entries. By Lemma E.20,

$$\begin{aligned}\mathbb{E}\|\mathbf{A}(\mathbf{X})\|_2^2 &= \frac{\rho^6}{h}\mathbb{E}\|\mathbf{g}_L^\top \mathbf{R}\boldsymbol{\Gamma}_{\mathbf{u}}(\mathbf{G})\mathbf{V}\|_2^2 \\ &\geq \frac{\rho^6 c_{L,d}}{h}\|\mathbf{R}\|_F^2\|\mathbf{V}\|_F^2.\end{aligned}$$

The equality substitutes the definition of $\mathbf{A}(\mathbf{X})$ and squares its scalar factor. The inequality applies the cited lemma with the fixed matrices $\mathbf{R}$ and $\mathbf{V}$. Since both sides are nonnegative, taking square roots gives

$$\left(\mathbb{E}\|\mathbf{A}(\mathbf{X})\|_2^2\right)^{1/2} \geq \frac{\rho^3\sqrt{c_{L,d}}}{\sqrt{h}}\|\mathbf{R}\|_F\|\mathbf{V}\|_F.$$

Step 5. Bound the nonlinear remainder. For every realization of the unit-ball vectors,

$$\begin{aligned}\|\mathbf{E}(\mathbf{X})\|_2 &\leq \frac{\rho^3}{\sqrt{h}}\int_0^1 \|\mathbf{g}_L^\top \mathbf{R}[\boldsymbol{\Gamma}_{\mathbf{p}_t}(\mathbf{G}) - \boldsymbol{\Gamma}_{\mathbf{u}}(\mathbf{G})]\mathbf{V}\|_2\, dt \\ &\leq \frac{\rho^3}{\sqrt{h}}\|\mathbf{g}_L\|_2\|\mathbf{R}\|_2\|\mathbf{V}\|_2\int_0^1 \|\boldsymbol{\Gamma}_{\mathbf{p}_t}(\mathbf{G}) - \boldsymbol{\Gamma}_{\mathbf{u}}(\mathbf{G})\|_2\, dt \\ &\leq \frac{6\rho^3\epsilon}{\sqrt{h}}\|\mathbf{R}\|_F\|\mathbf{V}\|_F.\end{aligned}$$

The first inequality is the triangle inequality for the vector-valued integral. It follows by applying the finite-sum triangle inequality to Riemann sums and taking their limit. The second inequality applies the spectral-norm product bound in Lemma E.15. The third inequality uses $\|\mathbf{g}_L\|_2 \leq 1$, Step 3, $\|\mathbf{R}\|_2 \leq \|\mathbf{R}\|_F$, and $\|\mathbf{V}\|_2 \leq \|\mathbf{V}\|_F$. The right-hand side is independent of $\mathbf{X}$. Squaring, taking expectation, and taking square roots therefore gives

$$\left(\mathbb{E}\|\mathbf{E}(\mathbf{X})\|_2^2\right)^{1/2} \leq \frac{6\rho^3\epsilon}{\sqrt{h}}\|\mathbf{R}\|_F\|\mathbf{V}\|_F.$$

Step 6. Prove the population lower bound. Lemma E.16 and Steps 4–5 imply

$$\begin{aligned}\left(\mathbb{E}\|\phi_{\mathbf{S}_*}(\mathbf{X}) - \phi_{\mathbf{S}_\Delta}(\mathbf{X})\|_2^2\right)^{1/2} &\geq \left(\mathbb{E}\|\mathbf{A}(\mathbf{X})\|_2^2\right)^{1/2} - \left(\mathbb{E}\|\mathbf{E}(\mathbf{X})\|_2^2\right)^{1/2} \\ &\geq \frac{\rho^3}{\sqrt{h}}(\sqrt{c_{L,d}} - 6\epsilon)\|\mathbf{R}\|_F\|\mathbf{V}\|_F \\ &\geq \frac{\rho^3\sqrt{c_{L,d}}}{2\sqrt{h}}\|\mathbf{R}\|_F\|\mathbf{V}\|_F.\end{aligned}$$

The first inequality applies the reverse triangle inequality to the decomposition from Step 4. The second inequality substitutes the lower bound for $\mathbf{A}(\mathbf{X})$ and the upper bound for $\mathbf{E}(\mathbf{X})$. The last inequality uses $\epsilon \leq \sqrt{c_{L,d}}/12$. Its right-hand side is nonnegative, so squaring proves the first inequality in (11). Applying Lemma E.21 to $\|\mathbf{R}\|_F^2 = \|\mathbf{S}_* - \mathbf{S}_\Delta\|_F^2$ proves the second inequality in (11).

Step 7. Prove the worst-case consequence. Every random matrix $\mathbf{X}$ in Definition E.14 satisfies $\max_i \|\mathbf{x}_i\|_2 \leq \rho$ almost surely. Therefore, for the fixed candidate,

$$\begin{aligned}&\sup_{\substack{\mathbf{X}\in\mathbb{R}^{L\times d} \\ \max_i\|\mathbf{x}_i\|_2\leq\rho}} \|\phi_{\mathbf{S}_*}(\mathbf{X}) - \phi_{\mathbf{S}_\Delta}(\mathbf{X})\|_2 \\ &\geq \left(\mathbb{E}\|\phi_{\mathbf{S}_*}(\mathbf{X}) - \phi_{\mathbf{S}_\Delta}(\mathbf{X})\|_2^2\right)^{1/2} \\ &\geq \frac{\rho^3\sqrt{c_{L,d}}}{2\sqrt{h}}\|\mathbf{V}\|_F\left(\sum_{j=2r+1}^{d}\sigma_j(\mathbf{S}_* - \mathbf{S}_0)^2\right)^{1/2}.\end{aligned}$$

The first inequality holds because a nonnegative random variable is at most its supremum almost surely, so its second moment is at most the square of that supremum. The second inequality takes the square root of the second population bound in (11). The final right-hand side does not depend on the candidate. Taking the infimum over $(\Delta\mathbf{Q}, \Delta\mathbf{K}) \in \mathcal{A}_{r,M}$ proves (12). □

Lemma E.23 (Lower bound with value updates)**.** *Use the attention parameters and score family in Definition E.10, with the query and key rank budget denoted by r_Q instead of r. Let $r_V \geq 0$ be an integer. The target and base value matrices are the same nonzero matrix $\mathbf{V} \in \mathbb{R}^{d\times p}$. In addition to an admissible pair $(\Delta\mathbf{Q}, \Delta\mathbf{K}) \in \mathcal{A}_{r_Q,M}$, allow any matrix $\Delta\mathbf{V} \in \mathbb{R}^{d\times p}$ satisfying $\operatorname{rank}(\Delta\mathbf{V}) \leq r_V$. For a score matrix $\mathbf{S}$ and a value matrix $\mathbf{U} \in \mathbb{R}^{d\times p}$, define the final-position output by*

$$\phi_{\mathbf{S},\mathbf{U}}(\mathbf{X}) := \mathbf{p}_{\mathbf{S}}(\mathbf{X})^\top \mathbf{X}\mathbf{U} \in \mathbb{R}^{1\times p}.$$

Here $\mathbf{p}_{\mathbf{S}}(\mathbf{X})$ is the vector in Definition E.11. Let the rows of $\mathbf{X}$ be independent and uniform on $\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 \leq \rho\}$, as in Definition E.14, and assume

$$\frac{\rho^2 M}{\sqrt{h}} \leq \frac{\sqrt{c_{L,d}}}{12}.$$

Then every fixed admissible triple satisfies

$$\begin{aligned}
&\mathbb{E}\big\|\phi_{\mathbf{S}_*,\mathbf{V}}(\mathbf{X}) - \phi_{\mathbf{S}_\Delta,\mathbf{V}+\Delta\mathbf{V}}(\mathbf{X})\big\|_2^2 \\
&\quad\geq \frac{\rho^6 c_{L,d}}{4h}\left(\sum_{j>r_V}\sigma_j(\mathbf{V})^2\right)\|\mathbf{S}_* - \mathbf{S}_\Delta\|_F^2 \\
&\quad\geq \frac{\rho^6 c_{L,d}}{4h}\left(\sum_{j>r_V}\sigma_j(\mathbf{V})^2\right)\left(\sum_{j>2r_Q}\sigma_j(\mathbf{S}_* - \mathbf{S}_0)^2\right).
\end{aligned}$$

Consequently, the infimum over the same joint query, key, and value update family satisfies

$$\begin{aligned}
&\inf_{\substack{(\Delta\mathbf{Q},\Delta\mathbf{K})\in\mathcal{A}_{r_Q,M}\ \max_i\|\mathbf{x}_i\|_2\leq\rho \\ \operatorname{rank}(\Delta\mathbf{V})\leq r_V}} \sup \big\|\phi_{\mathbf{S}_*,\mathbf{V}}(\mathbf{X}) - \phi_{\mathbf{S}_\Delta,\mathbf{V}+\Delta\mathbf{V}}(\mathbf{X})\big\|_2 \\
&\quad\geq \frac{\rho^3\sqrt{c_{L,d}}}{2\sqrt{h}}\left(\sum_{j>r_V}\sigma_j(\mathbf{V})^2\right)^{1/2}\left(\sum_{j>2r_Q}\sigma_j(\mathbf{S}_* - \mathbf{S}_0)^2\right)^{1/2}.
\end{aligned}$$

Proof. **Step 1. Remove the value update by an orthogonal projection.** Fix an admissible triple before drawing $\mathbf{X}$. Define $\mathbf{P} \in \mathbb{R}^{p\times p}$ to be the orthogonal projection onto

$$\ker(\Delta\mathbf{V}) := \{\mathbf{u} \in \mathbb{R}^p : \Delta\mathbf{V}\mathbf{u} = \mathbf{0}\}.$$

Then $\mathbf{P} = \mathbf{P}^\top = \mathbf{P}^2$, $\Delta\mathbf{V}\mathbf{P} = \mathbf{0}$, and

$$\operatorname{rank}(\mathbf{I}_p - \mathbf{P}) = \operatorname{rank}(\Delta\mathbf{V}) \leq r_V.$$

The equality follows from rank–nullity and the definition of $\mathbf{P}$. The inequality is the value-update rank constraint. Define the row-vector error

$$\mathbf{e}(\mathbf{X}) := \phi_{\mathbf{S}_*,\mathbf{V}}(\mathbf{X}) - \phi_{\mathbf{S}_\Delta,\mathbf{V}+\Delta\mathbf{V}}(\mathbf{X}).$$

Using $\Delta\mathbf{V}\mathbf{P} = \mathbf{0}$ gives

$$\begin{aligned}
\mathbf{e}(\mathbf{X})\mathbf{P} &= \mathbf{p}_{\mathbf{S}_*}(\mathbf{X})^\top\mathbf{X}\mathbf{V}\mathbf{P} - \mathbf{p}_{\mathbf{S}_\Delta}(\mathbf{X})^\top\mathbf{X}(\mathbf{V}+\Delta\mathbf{V})\mathbf{P} \\
&= \phi_{\mathbf{S}_*,\mathbf{V}\mathbf{P}}(\mathbf{X}) - \phi_{\mathbf{S}_\Delta,\mathbf{V}\mathbf{P}}(\mathbf{X}).
\end{aligned}$$

The first equality substitutes the definitions of the outputs. The second equality removes $\Delta\mathbf{V}\mathbf{P}$ and uses the same definitions with the shared value matrix $\mathbf{V}\mathbf{P}$.

Step 2. Lower-bound the value energy that survives. The rank inequality for a matrix product gives

$$\operatorname{rank}\big(\mathbf{V}(\mathbf{I}_p - \mathbf{P})\big) \leq \operatorname{rank}(\mathbf{I}_p - \mathbf{P}) \leq r_V.$$

Therefore the Frobenius approximation identity in Lemma E.1 yields

$$\begin{aligned}\|\mathbf{VP}\|_F^2 &= \big\|\mathbf{V} - \mathbf{V}(\mathbf{I}_p - \mathbf{P})\big\|_F^2 \\ &\geq \inf_{\mathrm{rank}(\mathbf{B})\leq r_V} \|\mathbf{V} - \mathbf{B}\|_F^2 \\ &= \sum_{j>r_V} \sigma_j(\mathbf{V})^2.\end{aligned}$$

The first equality expands $\mathbf{V}(\mathbf{I}_p - \mathbf{P})$. The inequality uses the admissible approximation $\mathbf{B} = \mathbf{V}(\mathbf{I}_p - \mathbf{P})$. The last equality is the rank-constrained matrix approximation formula.

Step 3. Apply the original attention lower bound. If $\sum_{j>r_V} \sigma_j(\mathbf{V})^2 = 0$, both claimed lower bounds on the expectation are zero and follow from nonnegativity. Otherwise Step 2 implies $\mathbf{VP} \neq \mathbf{0}$. Thus Theorem E.22 applies with value matrix $\mathbf{VP}$. This replacement changes neither the query law nor the score matrices, their norm bounds, or their update ranks. The projection $\mathbf{P}$ is fixed because the update was fixed before $\mathbf{X}$. We obtain

$$\begin{aligned}\mathbb{E}\|\mathbf{e}(\mathbf{X})\|_2^2 &\geq \mathbb{E}\|\mathbf{e}(\mathbf{X})\mathbf{P}\|_2^2 \\ &\geq \frac{\rho^6 c_{L,d}}{4h}\|\mathbf{VP}\|_F^2\|\mathbf{S}_* - \mathbf{S}_\Delta\|_F^2 \\ &\geq \frac{\rho^6 c_{L,d}}{4h}\left(\sum_{j>r_V}\sigma_j(\mathbf{V})^2\right)\|\mathbf{S}_* - \mathbf{S}_\Delta\|_F^2.\end{aligned}$$

The first inequality uses contraction of the Euclidean norm under an orthogonal projection. The second applies the original attention theorem to the identity from Step 1. The third uses Step 2. Finally, Lemma E.21 gives

$$\|\mathbf{S}_* - \mathbf{S}_\Delta\|_F^2 \geq \sum_{j>2r_Q}\sigma_j(\mathbf{S}_* - \mathbf{S}_0)^2.$$

Multiplying by the nonnegative value factor proves the second expectation bound.

Step 4. Pass to the worst-case error. For each fixed triple, the specified query distribution is supported on sequences satisfying $\max_i \|\mathbf{x}_i\|_2 \leq \rho$. Consequently,

$$\sup_{\max_i\|\mathbf{x}_i\|_2\leq\rho}\|\mathbf{e}(\mathbf{X})\|_2 \geq \big(\mathbb{E}\|\mathbf{e}(\mathbf{X})\|_2^2\big)^{1/2}.$$

Substitute the second expectation bound and take the square root. The resulting right-hand side does not depend on the chosen triple. Taking the infimum over all admissible triples therefore preserves it. □

Corollary E.24 (Lower bound with different value matrices)**.** *Keep the score family, query distribution, and small-logit assumption of Lemma E.23, but let the target and base value matrices be $\mathbf{V}_*, \mathbf{V}_0 \in \mathbb{R}^{d\times p}$. Define $\mathbf{D}_V := \mathbf{V}_* - \mathbf{V}_0$, and let $\mathbf{P}_0 \in \mathbb{R}^{p\times p}$ be the orthogonal projection onto $\ker(\mathbf{D}_V)$. For every fixed admissible query, key, and rank-r_V value update,*

$$\begin{aligned}&\mathbb{E}\big\|\phi_{\mathbf{S}_*,\mathbf{V}_*}(\mathbf{X}) - \phi_{\mathbf{S}_\Delta,\mathbf{V}_0+\Delta\mathbf{V}}(\mathbf{X})\big\|_2^2 \\ &\quad\geq \frac{\rho^6 c_{L,d}}{4h}\left(\sum_{j>r_V}\sigma_j(\mathbf{V}_*\mathbf{P}_0)^2\right)\left(\sum_{j>2r_Q}\sigma_j(\mathbf{S}_* - \mathbf{S}_0)^2\right).\end{aligned}$$

The corresponding worst-case lower bound follows by taking the square root and then the infimum over the same joint update family.

Proof. **Step 1. Restrict to output directions with a shared value map.** Since $\mathbf{D}_V\mathbf{P}_0 = \mathbf{0}$, define the common projected value matrix by

$$\mathbf{V}_c := \mathbf{V}_*\mathbf{P}_0 = \mathbf{V}_0\mathbf{P}_0.$$

The projected value update satisfies

$$\operatorname{rank}(\Delta \mathbf{V}\mathbf{P}_0) \le \operatorname{rank}(\Delta \mathbf{V}) \le r_V.$$

The first inequality is the rank bound for a matrix product. The second is the admissibility assumption. Writing the unprojected output difference as $\mathbf{e}(\mathbf{X})$ gives

$$\mathbf{e}(\mathbf{X})\mathbf{P}_0 = \phi_{\mathbf{S}_*,\mathbf{V}_c}(\mathbf{X}) - \phi_{\mathbf{S}_\Delta,\mathbf{V}_c+\Delta\mathbf{V}\mathbf{P}_0}(\mathbf{X}).$$

This equality substitutes the output definitions and uses $\mathbf{V}_*\mathbf{P}_0 = \mathbf{V}_0\mathbf{P}_0$.

Step 2. Apply the shared-value lemma. If $\mathbf{V}_c = \mathbf{0}$, the asserted right-hand side is zero. Otherwise apply Lemma E.23 with shared value $\mathbf{V}_c$ and value update $\Delta\mathbf{V}\mathbf{P}_0$. Since $\mathbf{P}_0$ is an orthogonal projection, $\|\mathbf{e}(\mathbf{X})\|_2 \ge \|\mathbf{e}(\mathbf{X})\mathbf{P}_0\|_2$ for every $\mathbf{X}$. Squaring this inequality, taking expectation, and substituting the lemma proves the displayed bound. The worst-case conclusion follows from the same support argument used in Step 4 of that lemma. □

E.3.2 Upper Bound under Weak Queries

The results in this subsection retain the attention functions and the attack family $\mathcal{A}_{r,M}$. They replace the independent uniform query distribution used in the lower bound by a distribution supported on sequences satisfying $\max_{1\le i\le L}\|\mathbf{x}_i\|_2 \le \rho$. The token positions need not be independent.

Lemma E.25 (Query and key updates by projection)**.** *Let $\mathbf{P} \in \mathbb{R}^{d\times d}$ satisfy $\mathbf{P} = \mathbf{P}^\top = \mathbf{P}^2$ and $\operatorname{rank}(\mathbf{P}) \le r$. Define*

$$\begin{aligned}
\mathbf{P}_\perp &:= \mathbf{I}_d - \mathbf{P}, \\
\Delta\mathbf{Q} &:= \mathbf{P}(\mathbf{Q}_* - \mathbf{Q}_0), & \widehat{\mathbf{Q}} &:= \mathbf{Q}_0 + \Delta\mathbf{Q} = \mathbf{P}\mathbf{Q}_* + \mathbf{P}_\perp\mathbf{Q}_0, \\
\Delta\mathbf{K} &:= \mathbf{P}(\mathbf{K}_* - \mathbf{K}_0), & \widehat{\mathbf{K}} &:= \mathbf{K}_0 + \Delta\mathbf{K} = \mathbf{P}\mathbf{K}_* + \mathbf{P}_\perp\mathbf{K}_0, \\
\widehat{\mathbf{S}} &:= \widehat{\mathbf{Q}}\widehat{\mathbf{K}}^\top, & \mathbf{E} &:= \mathbf{S}_* - \widehat{\mathbf{S}}.
\end{aligned}$$

Then we have

$$\operatorname{rank}(\Delta\mathbf{Q}) \le r, \qquad \operatorname{rank}(\Delta\mathbf{K}) \le r, \qquad \mathbf{P}\mathbf{E}\mathbf{P} = \mathbf{0}.$$

The pair $(\Delta\mathbf{Q}, \Delta\mathbf{K})$ belongs to $\mathcal{A}_{r,M}$ if $\|\widehat{\mathbf{S}}\|_2 \le M$. A sufficient condition for this inequality is

$$\sqrt{\|\mathbf{P}\mathbf{Q}_*\|_2^2 + \|\mathbf{P}_\perp\mathbf{Q}_0\|_2^2} \cdot \sqrt{\|\mathbf{P}\mathbf{K}_*\|_2^2 + \|\mathbf{P}_\perp\mathbf{K}_0\|_2^2} \le M.$$

Proof. The rank inequality for a product gives

$$\operatorname{rank}(\Delta\mathbf{Q}) = \operatorname{rank}(\mathbf{P}(\mathbf{Q}_* - \mathbf{Q}_0)) \le \operatorname{rank}(\mathbf{P}) \le r.$$

The first equality uses the definition of $\Delta\mathbf{Q}$. The last inequality is an assumption on $\mathbf{P}$. Replacing $\mathbf{Q}_*, \mathbf{Q}_0$ by $\mathbf{K}_*, \mathbf{K}_0$ proves the corresponding statement for $\Delta\mathbf{K}$.

The identities $\mathbf{P}^2 = \mathbf{P}$ and $\mathbf{P}\mathbf{P}_\perp = \mathbf{0}$ give

$$\begin{aligned}
\mathbf{P}\widehat{\mathbf{Q}} &= \mathbf{P}(\mathbf{P}\mathbf{Q}_* + \mathbf{P}_\perp\mathbf{Q}_0) = \mathbf{P}\mathbf{Q}_*, \\
\mathbf{P}\widehat{\mathbf{K}} &= \mathbf{P}(\mathbf{P}\mathbf{K}_* + \mathbf{P}_\perp\mathbf{K}_0) = \mathbf{P}\mathbf{K}_*.
\end{aligned}$$

The first equality in each line uses the definition of the corresponding updated factor. The second equality uses the two projection identities. It follows that

$$\begin{aligned}
\mathbf{P}\widehat{\mathbf{S}}\mathbf{P} &= (\mathbf{P}\widehat{\mathbf{Q}})(\mathbf{P}\widehat{\mathbf{K}})^\top \\
&= (\mathbf{P}\mathbf{Q}_*)(\mathbf{P}\mathbf{K}_*)^\top \\
&= \mathbf{P}\mathbf{S}_*\mathbf{P}.
\end{aligned}$$

The first equality uses $\mathbf{P} = \mathbf{P}^\top$ and $\widehat{\mathbf{S}} = \widehat{\mathbf{Q}}\widehat{\mathbf{K}}^\top$. The second equality uses the preceding identities. The third equality uses $\mathbf{S}_* = \mathbf{Q}_*\mathbf{K}_*^\top$ and $\mathbf{P} = \mathbf{P}^\top$. Subtracting this equality from $\mathbf{P}\mathbf{S}_*\mathbf{P}$ proves $\mathbf{P}\mathbf{E}\mathbf{P} = \mathbf{0}$.

For the score norm, orthogonality of the two projections gives

$$\widehat{\mathbf{Q}}^\top \widehat{\mathbf{Q}} = \mathbf{Q}_*^\top \mathbf{P} \mathbf{Q}_* + \mathbf{Q}_0^\top \mathbf{P}_\perp \mathbf{Q}_0.$$

To obtain this equality, expand the product using $\widehat{\mathbf{Q}} = \mathbf{P}\mathbf{Q}_* + \mathbf{P}_\perp \mathbf{Q}_0$. Both cross terms are zero because $\mathbf{P}\mathbf{P}_\perp = \mathbf{P}_\perp \mathbf{P} = \mathbf{0}$. The two remaining terms simplify using $\mathbf{P}^2 = \mathbf{P}$ and $\mathbf{P}_\perp^2 = \mathbf{P}_\perp$. Therefore

$$\begin{aligned}\|\widehat{\mathbf{Q}}\|_2^2 &= \|\widehat{\mathbf{Q}}^\top \widehat{\mathbf{Q}}\|_2 \\ &\le \|\mathbf{Q}_*^\top \mathbf{P} \mathbf{Q}_*\|_2 + \|\mathbf{Q}_0^\top \mathbf{P}_\perp \mathbf{Q}_0\|_2 \\ &= \|\mathbf{P}\mathbf{Q}_*\|_2^2 + \|\mathbf{P}_\perp \mathbf{Q}_0\|_2^2.\end{aligned}$$

The first equality is the Gram-matrix formula for the spectral norm. The inequality is the triangle inequality applied to the preceding Gram-matrix identity. The last equality uses $\mathbf{P} = \mathbf{P}^\top = \mathbf{P}^2$ and the same Gram-matrix formula. The identical argument bounds $\|\widehat{\mathbf{K}}\|_2^2$. Finally,

$$\begin{aligned}\|\widehat{\mathbf{S}}\|_2 &= \|\widehat{\mathbf{Q}}\widehat{\mathbf{K}}^\top\|_2 \\ &\le \|\widehat{\mathbf{Q}}\|_2 \|\widehat{\mathbf{K}}\|_2 \\ &\le \sqrt{\|\mathbf{P}\mathbf{Q}_*\|_2^2 + \|\mathbf{P}_\perp \mathbf{Q}_0\|_2^2} \\ &\qquad \cdot \sqrt{\|\mathbf{P}\mathbf{K}_*\|_2^2 + \|\mathbf{P}_\perp \mathbf{K}_0\|_2^2}.\end{aligned}$$

The first equality is the definition of $\widehat{\mathbf{S}}$. The first inequality is submultiplicativity of the spectral norm and invariance of that norm under transposition. The second inequality substitutes the two factor bounds. The stated sufficient condition thus gives $\|\widehat{\mathbf{S}}\|_2 \le M$. Together with the two rank bounds, this is precisely membership in $\mathcal{A}_{r,M}$. □

Lemma E.26 (Bounds on weighted covariance). *Fix vectors $\mathbf{x}_1, \dots, \mathbf{x}_L \in \mathbb{R}^d$ with $\|\mathbf{x}_i\|_2 \le \rho$ and weights $a_1, \dots, a_L \ge 0$ with $\sum_{i=1}^L a_i = 1$. Define*

$$\boldsymbol{\mu}_a := \sum_{i=1}^L a_i \mathbf{x}_i,$$

$$\boldsymbol{\Gamma}_a := \sum_{i=1}^L a_i (\mathbf{x}_i - \boldsymbol{\mu}_a)(\mathbf{x}_i - \boldsymbol{\mu}_a)^\top.$$

Then $\mathbf{0} \preceq \boldsymbol{\Gamma}_a \preceq \rho^2 \mathbf{I}_d$. For every orthogonal projection $\mathbf{P}$ and every $\mathbf{B} \in \mathbb{R}^{d \times p}$, with $\mathbf{P}_\perp := \mathbf{I}_d - \mathbf{P}$, one also has

$$\|\mathbf{P}_\perp \boldsymbol{\Gamma}_a \mathbf{B}\|_2 \le \rho \|\mathbf{B}\|_2 \sqrt{\sum_{i=1}^L a_i \|\mathbf{P}_\perp \mathbf{x}_i\|_2^2}.$$

Proof. For every $\mathbf{u} \in \mathbb{R}^d$, the defining sum gives

$$\mathbf{u}^\top \boldsymbol{\Gamma}_a \mathbf{u} = \sum_{i=1}^L a_i \big(\mathbf{u}^\top (\mathbf{x}_i - \boldsymbol{\mu}_a)\big)^2 \ge 0.$$

The equality follows by multiplying each outer product by $\mathbf{u}^\top$ and $\mathbf{u}$. The inequality follows from $a_i \ge 0$. Thus $\boldsymbol{\Gamma}_a$ is positive semidefinite. Expanding the same defining sum gives

$$\boldsymbol{\Gamma}_a = \sum_{i=1}^L a_i \mathbf{x}_i \mathbf{x}_i^\top - \boldsymbol{\mu}_a \boldsymbol{\mu}_a^\top.$$

Both cross terms in the expansion equal $-\boldsymbol{\mu}_a\boldsymbol{\mu}_a^\top$ because $\sum_i a_i\mathbf{x}_i = \boldsymbol{\mu}_a$. The last term equals $\boldsymbol{\mu}_a\boldsymbol{\mu}_a^\top$ because $\sum_i a_i = 1$. Consequently,

$$\begin{aligned}\mathbf{u}^\top\boldsymbol{\Gamma}_a\mathbf{u} &= \sum_{i=1}^{L} a_i(\mathbf{u}^\top\mathbf{x}_i)^2 - (\mathbf{u}^\top\boldsymbol{\mu}_a)^2 \\ &\le \sum_{i=1}^{L} a_i(\mathbf{u}^\top\mathbf{x}_i)^2 \\ &\le \sum_{i=1}^{L} a_i\|\mathbf{u}\|_2^2\|\mathbf{x}_i\|_2^2 \\ &\le \rho^2\|\mathbf{u}\|_2^2.\end{aligned}$$

The first equality uses the expanded covariance. The first inequality drops a nonnegative square. The second inequality is Cauchy–Schwarz applied to every inner product. The last inequality uses $\|\mathbf{x}_i\|_2 \le \rho$ and $\sum_i a_i = 1$. This proves $\boldsymbol{\Gamma}_a \preceq \rho^2\mathbf{I}_d$.

Let $\boldsymbol{\Gamma}_a^{1/2}$ be the positive semidefinite square root of $\boldsymbol{\Gamma}_a$. The spectral norm satisfies

$$\begin{aligned}\|\mathbf{P}_\perp\boldsymbol{\Gamma}_a\mathbf{B}\|_2 &\le \|\mathbf{P}_\perp\boldsymbol{\Gamma}_a^{1/2}\|_2\|\boldsymbol{\Gamma}_a^{1/2}\mathbf{B}\|_2 \\ &\le \sqrt{\mathrm{tr}(\mathbf{P}_\perp\boldsymbol{\Gamma}_a\mathbf{P}_\perp)}\rho\|\mathbf{B}\|_2.\end{aligned}$$

The first inequality factors $\boldsymbol{\Gamma}_a$ into two square roots and applies submultiplicativity. For the second inequality, the Gram-matrix formula gives $\|\mathbf{P}_\perp\boldsymbol{\Gamma}_a^{1/2}\|_2^2 = \|\mathbf{P}_\perp\boldsymbol{\Gamma}_a\mathbf{P}_\perp\|_2$. The largest eigenvalue of a positive semidefinite matrix is at most its trace. Also, $\|\boldsymbol{\Gamma}_a^{1/2}\mathbf{B}\|_2 \le \|\boldsymbol{\Gamma}_a^{1/2}\|_2\|\mathbf{B}\|_2 \le \rho\|\mathbf{B}\|_2$ by the covariance bound already proved. Finally,

$$\begin{aligned}\mathrm{tr}(\mathbf{P}_\perp\boldsymbol{\Gamma}_a\mathbf{P}_\perp) &= \sum_{i=1}^{L} a_i\|\mathbf{P}_\perp\mathbf{x}_i\|_2^2 - \|\mathbf{P}_\perp\boldsymbol{\mu}_a\|_2^2 \\ &\le \sum_{i=1}^{L} a_i\|\mathbf{P}_\perp\mathbf{x}_i\|_2^2.\end{aligned}$$

The equality substitutes the expanded covariance and uses $\mathrm{tr}(\mathbf{y}\mathbf{y}^\top) = \|\mathbf{y}\|_2^2$. The inequality drops a nonnegative squared norm. Substituting this trace bound proves the claim. □

Lemma E.27 (Attention error outside a subspace)**.** *Fix $(\Delta\mathbf{Q}, \Delta\mathbf{K}) \in \mathcal{A}_{r,M}$ and an orthogonal projection $\mathbf{P} \in \mathbb{R}^{d\times d}$. Define*

$$\mathbf{E} := \mathbf{S}_* - \mathbf{S}_\Delta, \qquad \mathbf{P}_\perp := \mathbf{I}_d - \mathbf{P}, \qquad \epsilon := \rho^2 M/\sqrt{h}.$$

Assume $\mathbf{PEP} = \mathbf{0}$. For every sequence with $\max_i \|\mathbf{x}_i\|_2 \le \rho$, define $b_i := \|\mathbf{P}_\perp\mathbf{x}_i\|_2$ for $i = 1, \dots, L$. Then we have

$$\|\phi_{\mathbf{S}_*}(\mathbf{X}) - \phi_{\mathbf{S}_\Delta}(\mathbf{X})\|_2 \le \frac{\rho^2\|\mathbf{E}\|_2\|\mathbf{V}\|_2}{\sqrt{h}}\left(b_L + e^\epsilon\sqrt{\frac{1}{L}\sum_{i=1}^{L} b_i^2}\right).$$

The same inequality holds after right multiplication of both outputs by a fixed matrix $\mathbf{C} \in \mathbb{R}^{p\times p}$, with $\|\mathbf{V}\|_2$ replaced by $\|\mathbf{V}\mathbf{C}\|_2$.

Proof. For $t \in [0, 1]$, use the interpolation $\mathbf{S}_t := \mathbf{S}_\Delta + t\mathbf{E}$ and the attention weights $p_{t,i}$ and weighted covariance $\boldsymbol{\Gamma}_t(\mathbf{X})$ from Lemma E.13. The endpoint score bounds imply

$$\begin{aligned}\|\mathbf{S}_t\|_2 &= \|(1-t)\mathbf{S}_\Delta + t\mathbf{S}_*\|_2 \\ &\le (1-t)\|\mathbf{S}_\Delta\|_2 + t\|\mathbf{S}_*\|_2 \\ &\le M.\end{aligned}$$

The equality uses the definition of $\mathbf{E}$. The first inequality is the triangle inequality and positive homogeneity. The last inequality uses admissibility of the candidate and the target score bound. Define $z_{t,i} := \mathbf{x}_L^\top \mathbf{S}_t \mathbf{x}_i / \sqrt{h}$. Cauchy–Schwarz and the spectral norm give

$$|z_{t,i}| \le \frac{\|\mathbf{x}_L\|_2 \|\mathbf{S}_t\|_2 \|\mathbf{x}_i\|_2}{\sqrt{h}} \le \frac{\rho^2 M}{\sqrt{h}} = \epsilon.$$

The second inequality substitutes the row bounds and the preceding score bound. The equality is the definition of ϵ. The softmax formula therefore gives

$$p_{t,i} = \frac{e^{z_{t,i}}}{\sum_{j=1}^{L} e^{z_{t,j}}} \le \frac{e^{\epsilon}}{L e^{-\epsilon}} = \frac{e^{2\epsilon}}{L}.$$

The inequality bounds the numerator above and every summand in the denominator below. The last equality simplifies the quotient.

Since $\mathbf{PEP} = \mathbf{0}$ and $\mathbf{P} + \mathbf{P}_\perp = \mathbf{I}_d$,

$$\mathbf{x}_L^\top \mathbf{E} = (\mathbf{P}_\perp \mathbf{x}_L)^\top \mathbf{E} + (\mathbf{P}\mathbf{x}_L)^\top \mathbf{E} \mathbf{P}_\perp.$$

To verify this identity, expand $\mathbf{x}_L = \mathbf{P}\mathbf{x}_L + \mathbf{P}_\perp \mathbf{x}_L$ and insert $\mathbf{P} + \mathbf{P}_\perp = \mathbf{I}_d$ to the right of the second residual factor. The term $\mathbf{x}_L^\top \mathbf{PEP}$ is zero by assumption. By the triangle inequality,

$$\begin{aligned}
\|\mathbf{x}_L^\top \mathbf{E} \mathbf{\Gamma}_t(\mathbf{X}) \mathbf{V}\|_2 &\le \|(\mathbf{P}_\perp \mathbf{x}_L)^\top \mathbf{E} \mathbf{\Gamma}_t(\mathbf{X}) \mathbf{V}\|_2 + \|(\mathbf{P}\mathbf{x}_L)^\top \mathbf{E} \mathbf{P}_\perp \mathbf{\Gamma}_t(\mathbf{X}) \mathbf{V}\|_2 \\
&\le b_L \|\mathbf{E}\|_2 \rho^2 \|\mathbf{V}\|_2 + \rho \|\mathbf{E}\|_2 \|\mathbf{P}_\perp \mathbf{\Gamma}_t(\mathbf{X}) \mathbf{V}\|_2 \\
&\le \rho^2 \|\mathbf{E}\|_2 \|\mathbf{V}\|_2 \left(b_L + \sqrt{\sum_{i=1}^{L} p_{t,i} b_i^2} \right) \\
&\le \rho^2 \|\mathbf{E}\|_2 \|\mathbf{V}\|_2 \left(b_L + e^{\epsilon} \sqrt{\frac{1}{L} \sum_{i=1}^{L} b_i^2} \right).
\end{aligned}$$

The second inequality uses submultiplicativity, $\|\mathbf{\Gamma}_t(\mathbf{X})\|_2 \le \rho^2$ from Lemma E.26, and $\|\mathbf{P}\mathbf{x}_L\|_2 \le \|\mathbf{x}_L\|_2 \le \rho$. The third inequality applies the second conclusion of that lemma with weights $p_{t,i}$ and $\mathbf{B} = \mathbf{V}$. The fourth inequality applies the softmax-weight bound proved above.

Lemma E.13 and the triangle inequality for an integral give

$$\begin{aligned}
\|\phi_{\mathbf{S}_*}(\mathbf{X}) - \phi_{\mathbf{S}_\triangle}(\mathbf{X})\|_2 &= \left\| \frac{1}{\sqrt{h}} \int_0^1 \mathbf{x}_L^\top \mathbf{E} \mathbf{\Gamma}_t(\mathbf{X}) \mathbf{V} \, dt \right\|_2 \\
&\le \frac{1}{\sqrt{h}} \int_0^1 \|\mathbf{x}_L^\top \mathbf{E} \mathbf{\Gamma}_t(\mathbf{X}) \mathbf{V}\|_2 \, dt.
\end{aligned}$$

The bound on the integrand does not depend on t. Substituting it and using $\int_0^1 dt = 1$ proves the first claim. For the second claim, right multiply the integral identity by $\mathbf{C}$ and repeat each displayed inequality with $\mathbf{V}$ replaced by $\mathbf{VC}$. □

Theorem E.28 (Upper bound under restricted queries)**.** *Assume $0 \le r \le d$ and $\|\mathbf{x}_i\|_2 \le \rho$ almost surely for every $i = 1, \dots, L$. Suppose all token positions have the same raw second moment*

$$\mathbf{\Sigma} := \mathbb{E}[\mathbf{x}_i \mathbf{x}_i^\top].$$

Let $\lambda_1 \ge \cdots \ge \lambda_d \ge 0$ be its eigenvalues. Let $\mathbf{u}_1, \dots, \mathbf{u}_d$ be corresponding orthonormal eigenvectors and define $\mathbf{P} := \sum_{j=1}^{r} \mathbf{u}_j \mathbf{u}_j^\top$, where the sum is zero when $r = 0$. Assume the constructed pair in Lemma E.25 satisfies

$$\left\| (\mathbf{P}\mathbf{Q}_* + (\mathbf{I}_d - \mathbf{P})\mathbf{Q}_0)(\mathbf{P}\mathbf{K}_* + (\mathbf{I}_d - \mathbf{P})\mathbf{K}_0)^\top \right\|_2 \le M.$$

Then we have

$$\inf_{(\Delta\mathbf{Q},\Delta\mathbf{K})\in\mathcal{A}_{r,M}} \mathbb{E}\|\phi_{\mathbf{S}_*}(\mathbf{X}) - \phi_{\mathbf{S}_\Delta}(\mathbf{X})\|_2^2 \leq \frac{4\rho^4 M^2 \|\mathbf{V}\|_2^2}{h}\left(1+\exp\left(\frac{\rho^2 M}{\sqrt{h}}\right)\right)^2 \sum_{j=r+1}^{d}\lambda_j.$$

Proof. Use the constructed update from Lemma E.25. Define $\mathbf{E} := \mathbf{S}_* - \widehat{\mathbf{S}}$, $\mathbf{P}_\perp := \mathbf{I}_d - \mathbf{P}$, $b_i := \|\mathbf{P}_\perp \mathbf{x}_i\|_2$, and $\epsilon := \rho^2 M/\sqrt{h}$. The construction gives $\mathbf{PEP} = \mathbf{0}$ and the assumed score bound makes the pair admissible. For each token position,

$$\begin{aligned}\mathbb{E}b_i^2 &= \mathbb{E}\,\mathrm{tr}(\mathbf{P}_\perp \mathbf{x}_i \mathbf{x}_i^\top \mathbf{P}_\perp)\\ &= \mathrm{tr}(\mathbf{P}_\perp \mathbf{\Sigma} \mathbf{P}_\perp)\\ &= \sum_{j=r+1}^{d}\lambda_j.\end{aligned}$$

The first equality follows from $\|\mathbf{P}_\perp \mathbf{x}_i\|_2^2 = \mathrm{tr}((\mathbf{P}_\perp \mathbf{x}_i)(\mathbf{P}_\perp \mathbf{x}_i)^\top)$. The second equality uses linearity of expectation and trace together with the common-second-moment assumption. The third equality follows by substituting $\mathbf{\Sigma} = \sum_{j=1}^{d}\lambda_j \mathbf{u}_j \mathbf{u}_j^\top$ and observing that $\mathbf{P}_\perp \mathbf{u}_j = \mathbf{0}$ for $j \leq r$ and $\mathbf{P}_\perp \mathbf{u}_j = \mathbf{u}_j$ for $j > r$. The trace of each remaining $\mathbf{u}_j \mathbf{u}_j^\top$ is one because $\|\mathbf{u}_j\|_2 = 1$. Define $\eta := \sum_{j=r+1}^{d}\lambda_j$.

Minkowski's inequality in the probability-space L^2 norm gives

$$\begin{aligned}&\left(\mathbb{E}\left[b_L + e^\epsilon \sqrt{\frac{1}{L}\sum_{i=1}^{L} b_i^2}\right]^2\right)^{1/2}\\ &\quad\leq (\mathbb{E}b_L^2)^{1/2} + e^\epsilon \left(\mathbb{E}\frac{1}{L}\sum_{i=1}^{L} b_i^2\right)^{1/2}\\ &\quad= \sqrt{\eta} + e^\epsilon\sqrt{\eta}\\ &\quad= (1+e^\epsilon)\sqrt{\eta}.\end{aligned}$$

The first equality after the inequality uses $\mathbb{E}b_i^2 = \eta$ for every i and linearity of expectation. The final equality factors out $\sqrt{\eta}$. No step in this argument uses independence.

Lemma E.27 now yields

$$\begin{aligned}\mathbb{E}\|\phi_{\mathbf{S}_*}(\mathbf{X}) - \phi_{\widehat{\mathbf{S}}}(\mathbf{X})\|_2^2 &\leq \frac{\rho^4 \|\mathbf{E}\|_2^2 \|\mathbf{V}\|_2^2}{h}(1+e^\epsilon)^2\eta\\ &\leq \frac{4\rho^4 M^2 \|\mathbf{V}\|_2^2}{h}(1+e^\epsilon)^2\eta.\end{aligned}$$

The first inequality squares the pointwise bound and applies the preceding L^2 estimate. For the second inequality, the triangle inequality gives $\|\mathbf{E}\|_2 \leq \|\mathbf{S}_*\|_2 + \|\widehat{\mathbf{S}}\|_2 \leq 2M$. The infimum over admissible updates is at most the error of the constructed admissible pair. Substituting the definitions of ϵ and η proves the theorem. □

Token positions may be dependent and need not have zero mean.

Corollary E.29 (Exact matching on a query subspace)**.** *Under the assumptions of Theorem E.28, if* $\mathrm{rank}(\mathbf{\Sigma}) \leq r$*, the constructed attack satisfies* $\phi_{\mathbf{S}_*}(\mathbf{X}) = \phi_{\widehat{\mathbf{S}}}(\mathbf{X})$ *almost surely.*

Proof. The rank assumption and positive semidefiniteness of $\mathbf{\Sigma}$ imply $\sum_{j=r+1}^{d}\lambda_j = 0$. The proof of Theorem E.28 therefore gives $\mathbb{E}\|\mathbf{P}_\perp \mathbf{x}_i\|_2^2 = 0$ for each i. A nonnegative random variable with

zero expectation is zero almost surely. Thus $\mathbf{P}_\perp \mathbf{x}_i = \mathbf{0}$ almost surely for each i. There are only L positions, so all these equalities hold simultaneously with probability one. On this event,

$$\mathbf{x}_L^\top (\mathbf{S}_* - \widehat{\mathbf{S}})\mathbf{x}_i = \mathbf{x}_L^\top \mathbf{P}(\mathbf{S}_* - \widehat{\mathbf{S}})\mathbf{P}\mathbf{x}_i = 0$$

for every i. The first equality uses $\mathbf{x}_L = \mathbf{P}\mathbf{x}_L$, $\mathbf{x}_i = \mathbf{P}\mathbf{x}_i$, and $\mathbf{P} = \mathbf{P}^\top$. The second equality uses the projected-update identity $\mathbf{PEP} = \mathbf{0}$. The target and candidate logits are consequently identical. Applying the same softmax function gives identical attention weights. The value matrix is shared, so their weighted value sums are identical. □

Remark E.30 (The score bound after an update)**.** *The inequalities $\|\mathbf{S}_*\|_2, \|\mathbf{S}_0\|_2 \le M$ do not imply the constructed score bound in Theorem E.28. For $d = 2$, $h = r = 1$, $\mathbf{P} := \mathbf{e}_1\mathbf{e}_1^\top$, and $a > 1$, define $\mathbf{Q}_* := a\mathbf{e}_1$, $\mathbf{K}_* := a^{-1}\mathbf{e}_1$, $\mathbf{Q}_0 := a^{-1}\mathbf{e}_2$, and $\mathbf{K}_0 := a\mathbf{e}_2$. Then $\mathbf{S}_* = \mathbf{e}_1\mathbf{e}_1^\top$ and $\mathbf{S}_0 = \mathbf{e}_2\mathbf{e}_2^\top$, so both score norms are one. Direct multiplication of the constructed factors gives*

$$\widehat{\mathbf{S}} = \begin{pmatrix} 1 & a^2 \\ a^{-2} & 1 \end{pmatrix}.$$

In particular, $\|\widehat{\mathbf{S}}\|_2 \ge \|\widehat{\mathbf{S}}\mathbf{e}_2\|_2 = \sqrt{a^4 + 1} > a^2$. The inequality follows by evaluating the spectral norm on the unit vector $\mathbf{e}_2$. The equality multiplies the displayed matrix by $\mathbf{e}_2$. Thus the constructed score norm is not bounded by the two original score norms. This example does not rule out a different admissible update.

E.3.3 Upper Bound under Low-Rank Verification

The verifier in this subsection observes a fixed linear projection of the model output. The attack may adapt the output readout while keeping its input representation fixed. This parameter family differs from a family that permits only query and key updates. The matrix approximation identity below is a reduced-rank regression identity (Izenman, 1975). Our use of it specifies which parts of an attention or Transformer model must be adaptable.

Definition E.31 (Fixed features and linear verification)**.** *Let q, p, c be positive integers and let r be a nonnegative integer. Let X be a random query with distribution $\mathcal{D}$ on a set $\mathcal{X}$. Fix measurable maps $f_* : \mathcal{X} \to \mathbb{R}^q$ and $\widehat{h} : \mathcal{X} \to \mathbb{R}^p$ satisfying*

$$\mathbb{E}\|f_*(X)\|_2^2 < \infty, \qquad \mathbb{E}\|\widehat{h}(X)\|_2^2 < \infty.$$

Fix matrices $\mathbf{W}_0 \in \mathbb{R}^{q\times p}$ and $\mathbf{C} \in \mathbb{R}^{q\times c}$. Define

$$\begin{aligned} \mathcal{A}_r &:= \{\mathbf{\Delta} \in \mathbb{R}^{q\times p} : \operatorname{rank}(\mathbf{\Delta}) \le r\}, \\ f_{\mathbf{\Delta}}(X) &:= (\mathbf{W}_0 + \mathbf{\Delta})\widehat{h}(X), \\ \mathbf{z}(X) &:= \mathbf{C}^\top\big(f_*(X) - \mathbf{W}_0\widehat{h}(X)\big). \end{aligned}$$

The verifier observes $\mathbf{C}^\top f_*(X)$ or $\mathbf{C}^\top f_{\mathbf{\Delta}}(X)$ in $\mathbb{R}^c$. The representation $\widehat{h}$ does not depend on $\mathbf{\Delta}$. All matrices and maps, including the chosen update, are fixed before X is sampled. Every expectation in this subsection is over $X \sim \mathcal{D}$.

Definition E.32 (Second moments and pseudoinverses)**.** *For a real matrix with positive-singular-value decomposition $\mathbf{A} = \sum_{j=1}^k s_j \mathbf{u}_j\mathbf{v}_j^\top$, define its Moore–Penrose pseudoinverse by $\mathbf{A}^\dagger := \sum_{j=1}^k s_j^{-1}\mathbf{v}_j\mathbf{u}_j^\top$. The pseudoinverse of the zero matrix is the zero matrix of transposed size. Define*

$$\begin{aligned} \mathbf{\Sigma} &:= \mathbb{E}[\widehat{h}(X)\widehat{h}(X)^\top] \in \mathbb{R}^{p\times p}, \\ \mathbf{M} &:= \mathbb{E}[\mathbf{z}(X)\widehat{h}(X)^\top] \in \mathbb{R}^{c\times p}. \end{aligned}$$

For an eigendecomposition $\mathbf{\Sigma} = \sum_{j=1}^{p} \lambda_j \mathbf{u}_j \mathbf{u}_j^\top$ with $\lambda_j \geq 0$, define

$$\begin{aligned}
\mathbf{\Sigma}^{1/2} &:= \sum_{j=1}^{p} \sqrt{\lambda_j}\, \mathbf{u}_j \mathbf{u}_j^\top, \\
\mathbf{\Sigma}^{\dagger/2} &:= \sum_{j:\lambda_j>0} \lambda_j^{-1/2} \mathbf{u}_j \mathbf{u}_j^\top, \\
\mathbf{P}_{\mathbf{\Sigma}} &:= \sum_{j:\lambda_j>0} \mathbf{u}_j \mathbf{u}_j^\top, \\
\mathbf{H} &:= \mathbf{M}\mathbf{\Sigma}^{\dagger/2}, \\
\epsilon^2 &:= \mathbb{E}\|\mathbf{z}(X)\|_2^2 - \|\mathbf{H}\|_F^2.
\end{aligned}$$

Let $\mathbf{H}_r$ be a truncated singular value decomposition retaining the largest $\min\{r, \operatorname{rank}(\mathbf{H})\}$ positive singular values. Define $\sigma_j(\mathbf{H}) := 0$ beyond the matrix dimensions. An empty sum is zero.

Theorem E.33 identifies ϵ^2 as a minimum squared prediction error, so it is nonnegative.

Theorem E.33 (Minimum error under linear verification)**.** *Under Definitions E.31 and E.32, we have*

$$\min_{\mathbf{\Delta} \in \mathcal{A}_r} \mathbb{E}\|\mathbf{C}^\top (f_*(X) - f_{\mathbf{\Delta}}(X))\|_2^2 = \epsilon^2 + \sum_{j>r} \sigma_j(\mathbf{H})^2.$$

The minimum is attained by

$$\mathbf{\Delta}_r := (\mathbf{C}^\top)^\dagger \mathbf{H}_r \mathbf{\Sigma}^{\dagger/2}.$$

Moreover,

$$\begin{aligned}
\epsilon^2 = \min_{\mathbf{B} \in \mathbb{R}^{c \times p}} \mathbb{E}\|\mathbf{z}(X) - \mathbf{B}\widehat{h}(X)\|_2^2 \geq 0, \\
\operatorname{rank}(\mathbf{H}) \leq \operatorname{rank}(\mathbf{C}).
\end{aligned}$$

In particular, if $r \geq \operatorname{rank}(\mathbf{C})$, the optimal verified squared error equals ϵ^2.

Proof. **Step 1. Handle singular second moments.** Fix $\mathbf{v} \in \ker(\mathbf{\Sigma})$. The second-moment definition gives

$$\mathbb{E}(\mathbf{v}^\top \widehat{h}(X))^2 = \mathbf{v}^\top \mathbf{\Sigma} \mathbf{v} = 0.$$

A nonnegative random variable with zero expectation is zero almost surely. Thus $\mathbf{v}^\top \widehat{h}(X) = 0$ almost surely and $\mathbf{M}\mathbf{v} = \mathbb{E}[\mathbf{z}(X)\widehat{h}(X)^\top \mathbf{v}] = \mathbf{0}$. Applying this to a basis of the kernel yields

$$\mathbf{M}\mathbf{P}_{\mathbf{\Sigma}} = \mathbf{M}, \qquad \mathbf{H}\mathbf{P}_{\mathbf{\Sigma}} = \mathbf{H}.$$

The second equality also uses $\mathbf{\Sigma}^{\dagger/2}\mathbf{P}_{\mathbf{\Sigma}} = \mathbf{\Sigma}^{\dagger/2}$. All cross moments exist by Cauchy–Schwarz and the finite second moments.

Step 2. Separate the prediction error. Define $\mathbf{B}_0 := \mathbf{M}\mathbf{\Sigma}^\dagger$ and $\mathbf{u}(X) := \mathbf{z}(X) - \mathbf{B}_0\widehat{h}(X)$. Then we have

$$\begin{aligned}
\mathbb{E}[\mathbf{u}(X)\widehat{h}(X)^\top] &= \mathbf{M} - \mathbf{M}\mathbf{\Sigma}^\dagger \mathbf{\Sigma} \\
&= \mathbf{M} - \mathbf{M}\mathbf{P}_{\mathbf{\Sigma}} = \mathbf{0}.
\end{aligned}$$

The first equality substitutes the definitions. The second uses the eigendecomposition of $\mathbf{\Sigma}$. The last uses Step 1. Expanding a squared Euclidean norm therefore gives, for every fixed $\mathbf{B} \in \mathbb{R}^{c \times p}$,

$$\begin{aligned}
\mathbb{E}\|\mathbf{z} - \mathbf{B}\widehat{h}\|_2^2 &= \mathbb{E}\|\mathbf{u} + (\mathbf{B}_0 - \mathbf{B})\widehat{h}\|_2^2 \\
&= \mathbb{E}\|\mathbf{u}\|_2^2 + \|(\mathbf{B}_0 - \mathbf{B})\mathbf{\Sigma}^{1/2}\|_F^2.
\end{aligned}$$

The cross term vanishes by the preceding orthogonality identity. The second term follows from $\mathbb{E}\|\mathbf{A}\widehat{h}\|_2^2 = \operatorname{tr}(\mathbf{A}\mathbf{\Sigma}\mathbf{A}^\top)$. Since $\mathbf{B}_0\mathbf{\Sigma}^{1/2} = \mathbf{H}$, setting $\mathbf{B} = \mathbf{0}$ gives $\mathbb{E}\|\mathbf{u}\|_2^2 = \mathbb{E}\|\mathbf{z}\|_2^2 - \|\mathbf{H}\|_F^2 = \epsilon^2$.

The same decomposition shows that $\mathbf{B}_0$ minimizes the unrestricted linear prediction error and proves $\epsilon^2 \geq 0$.

Step 3. Prove the rank-constrained lower bound. For an update $\mathbf{\Delta} \in \mathcal{A}_r$, substitute $\mathbf{B} = \mathbf{C}^\top \mathbf{\Delta}$ into Step 2. This gives

$$\mathbb{E}\|\mathbf{C}^\top(f_* - f_{\mathbf{\Delta}})\|_2^2 = \epsilon^2 + \|\mathbf{H} - \mathbf{C}^\top \mathbf{\Delta}\mathbf{\Sigma}^{1/2}\|_F^2.$$

The product $\mathbf{C}^\top \mathbf{\Delta}\mathbf{\Sigma}^{1/2}$ has rank at most r because multiplication cannot increase rank. Lemma E.1 therefore bounds the last term below by $\sum_{j>r} \sigma_j(\mathbf{H})^2$.

Step 4. Check both range constraints. Set $\mathbf{P}_{\mathbf{C}} := (\mathbf{C}^\top)(\mathbf{C}^\top)^\dagger$, the orthogonal projection onto the column space of $\mathbf{C}^\top$. The definition of $\mathbf{z}$ gives $\mathbf{P}_{\mathbf{C}}\mathbf{z} = \mathbf{z}$ for every query. Taking cross moments gives $\mathbf{P}_{\mathbf{C}}\mathbf{M} = \mathbf{M}$ and $\mathbf{P}_{\mathbf{C}}\mathbf{H} = \mathbf{H}$. Every positive left singular vector of $\mathbf{H}$ lies in its column space. Every positive right singular vector lies in the column space of $\mathbf{\Sigma}$, by $\mathbf{H}\mathbf{P}_{\mathbf{\Sigma}} = \mathbf{H}$. The truncated singular value decomposition consequently satisfies

$$\mathbf{P}_{\mathbf{C}}\mathbf{H}_r = \mathbf{H}_r, \qquad \mathbf{H}_r\mathbf{P}_{\mathbf{\Sigma}} = \mathbf{H}_r.$$

Only positive singular directions are retained, so these conclusions also hold when r exceeds the rank of $\mathbf{H}$.

Step 5. Attain the lower bound with a feasible update. The proposed $\mathbf{\Delta}_r$ has rank at most r, since $\mathbf{H}_r$ has rank at most r. Its verified, weighted action is

$$\begin{aligned}\mathbf{C}^\top \mathbf{\Delta}_r \mathbf{\Sigma}^{1/2} &= \mathbf{P}_{\mathbf{C}}\mathbf{H}_r\mathbf{\Sigma}^{\dagger/2}\mathbf{\Sigma}^{1/2} \\ &= \mathbf{P}_{\mathbf{C}}\mathbf{H}_r\mathbf{P}_{\mathbf{\Sigma}} = \mathbf{H}_r.\end{aligned}$$

The first equality substitutes $\mathbf{\Delta}_r$. The second uses the eigendecomposition of $\mathbf{\Sigma}$. The last uses both identities from Step 4. Substitution into Step 3 and Lemma E.1 proves the exact formula. Finally, $\mathbf{P}_{\mathbf{C}}\mathbf{H} = \mathbf{H}$ implies $\operatorname{rank}(\mathbf{H}) \leq \operatorname{rank}(\mathbf{C})$. Thus the spectral tail vanishes whenever $r \geq \operatorname{rank}(\mathbf{C})$. □

The verifier may have any norm. Only its rank enters the condition $r \geq \operatorname{rank}(\mathbf{C})$.

Corollary E.34 (Verification error from feature error)**.** *Let d be a positive integer. Suppose additionally that $f_*(X) = \mathbf{W}_* h_*(X)$, where $\mathbf{W}_* \in \mathbb{R}^{q\times d}$ and $h_* : \mathcal{X} \to \mathbb{R}^d$ has finite second moment. Assume $r \geq \operatorname{rank}(\mathbf{C})$. For any fixed $\mathbf{T} \in \mathbb{R}^{d\times p}$, define*

$$\mathbf{P} := \mathbf{C}\mathbf{C}^\dagger, \qquad \mathbf{\Delta}_{\mathbf{T}} := \mathbf{P}(\mathbf{W}_*\mathbf{T} - \mathbf{W}_0).$$

Then $\mathbf{\Delta}_{\mathbf{T}} \in \mathcal{A}_r$ and, for every query,

$$\mathbf{C}^\top(f_* - f_{\mathbf{\Delta}_{\mathbf{T}}}) = \mathbf{C}^\top\mathbf{W}_*(h_* - \mathbf{T}\widehat{h}).$$

Consequently,

$$\min_{\mathbf{\Delta}\in\mathcal{A}_r} \left(\mathbb{E}\|\mathbf{C}^\top(f_* - f_{\mathbf{\Delta}})\|_2^2\right)^{1/2} \leq \|\mathbf{C}^\top\mathbf{W}_*\|_2 \left(\mathbb{E}\|h_* - \mathbf{T}\widehat{h}\|_2^2\right)^{1/2}.$$

If $h_* = \mathbf{T}\widehat{h}$ almost surely, the verified outputs match exactly almost surely.

Proof. **Step 1. Check the rank budget.** The matrix $\mathbf{P}$ is the orthogonal projection onto the column space of $\mathbf{C}$. Hence $\operatorname{rank}(\mathbf{\Delta}_{\mathbf{T}}) \leq \operatorname{rank}(\mathbf{P}) = \operatorname{rank}(\mathbf{C}) \leq r$.

Step 2. Compute the verified readout. The projection identity gives $\mathbf{C}^\top\mathbf{P} = \mathbf{C}^\top$. Thus

$$\begin{aligned}\mathbf{C}^\top(\mathbf{W}_0 + \mathbf{\Delta}_{\mathbf{T}}) &= \mathbf{C}^\top\mathbf{W}_0 + \mathbf{C}^\top(\mathbf{W}_*\mathbf{T} - \mathbf{W}_0) \\ &= \mathbf{C}^\top\mathbf{W}_*\mathbf{T}.\end{aligned}$$

Subtracting this action on $\widehat{h}$ from $\mathbf{C}^\top\mathbf{W}_* h_*$ proves the pointwise identity.

Step 3. Bound the remaining feature error. The inequality $\|\mathbf{A}\mathbf{u}\|_2 \leq \|\mathbf{A}\|_2\|\mathbf{u}\|_2$, followed by squaring, expectation, and a square root, gives the bound. If the feature error is zero almost surely, the pointwise identity gives exact matching almost surely. □

The case of shared features uses $d = p$, $\mathbf{T} = \mathbf{I}_p$, and $h_* = \widehat{h}$.

Corollary E.35 (Low-rank verification for attention)**.** *Let L, d, h, q, c be positive integers with $L \geq 2$, let $r_V \geq 0$ be an integer, and fix $\mathbf{C} \in \mathbb{R}^{q\times c}$. For $\mathbf{X} \in \mathbb{R}^{L\times d}$, write its rows as $\mathbf{x}_i^\top$. For a fixed score matrix $\mathbf{S} \in \mathbb{R}^{d\times d}$, define*

$$p_{\mathbf{S},i}(\mathbf{X}) := \frac{\exp(\mathbf{x}_L^\top \mathbf{S}\mathbf{x}_i/\sqrt{h})}{\sum_{j=1}^{L}\exp(\mathbf{x}_L^\top \mathbf{S}\mathbf{x}_j/\sqrt{h})},$$
$$\mu_{\mathbf{S}}(\mathbf{X}) := \sum_{i=1}^{L} p_{\mathbf{S},i}(\mathbf{X})\mathbf{x}_i \in \mathbb{R}^d.$$

Fix target and candidate scores $\mathbf{S}_*, \widehat{\mathbf{S}}$ and value matrices $\mathbf{V}_*, \mathbf{V}_0 \in \mathbb{R}^{d\times q}$. The target and adapted final-token outputs, written as columns, are

$$f_*(\mathbf{X}) := \mathbf{V}_*^\top \mu_{\mathbf{S}_*}(\mathbf{X}),$$
$$f_{\Delta\mathbf{V}}(\mathbf{X}) := (\mathbf{V}_0 + \Delta\mathbf{V})^\top \mu_{\widehat{\mathbf{S}}}(\mathbf{X}), \qquad \operatorname{rank}(\Delta\mathbf{V}) \leq r_V.$$

Assume $\mathbb{E}\max_i \|\mathbf{x}_i\|_2^2 < \infty$. Define $\mathbf{H}$ and ϵ^2 as in Definition E.32, with $p = d$, $\widehat{h} = \mu_{\widehat{\mathbf{S}}}$, and $\mathbf{W}_0 = \mathbf{V}_0^\top$. Then we have

$$\min_{\operatorname{rank}(\Delta\mathbf{V})\leq r_V} \mathbb{E}\|\mathbf{C}^\top(f_* - f_{\Delta\mathbf{V}})\|_2^2 = \epsilon^2 + \sum_{j>r_V}\sigma_j(\mathbf{H})^2.$$

If $r_V \geq \operatorname{rank}(\mathbf{C})$, this minimum equals ϵ^2. If also $\mathbf{S}_* = \widehat{\mathbf{S}}$, then $\epsilon^2 = 0$.
More explicitly, when $r_V \geq \operatorname{rank}(\mathbf{C})$, the update

$$\Delta\mathbf{V} := (\mathbf{V}_* - \mathbf{V}_0)\mathbf{C}\mathbf{C}^\dagger$$

satisfies, for every $\mathbf{X}$,

$$\mathbf{C}^\top(f_*(\mathbf{X}) - f_{\Delta\mathbf{V}}(\mathbf{X})) = \mathbf{C}^\top\mathbf{V}_*^\top(\mu_{\mathbf{S}_*}(\mathbf{X}) - \mu_{\widehat{\mathbf{S}}}(\mathbf{X})).$$

If $\max_i \|\mathbf{x}_i\|_2 \leq \rho$, this gives the pointwise bound

$$\|\mathbf{C}^\top(f_*(\mathbf{X}) - f_{\Delta\mathbf{V}}(\mathbf{X}))\|_2 \leq \frac{\rho^3}{\sqrt{h}}\|\mathbf{V}_*\mathbf{C}\|_2\|\mathbf{S}_* - \widehat{\mathbf{S}}\|_2.$$

Proof. **Step 1. Verify the readout setting.** The weights $p_{\mathbf{S},i}$ are nonnegative and sum to one. Hence $\|\mu_{\mathbf{S}}(\mathbf{X})\|_2 \leq \max_i \|\mathbf{x}_i\|_2$ by the triangle inequality. The moment assumption therefore verifies all finite-moment hypotheses. Transposition preserves rank. The generic theorem applies under the displayed substitutions. The explicit update and identity follow from Corollary E.34 with $\mathbf{W}_* = \mathbf{V}_*^\top$ and $\mathbf{T} = \mathbf{I}_d$. Matched scores give identical pooled features, so the verified error is zero.

Step 2. Bound the score-induced feature mismatch. Set $\mathbf{R} := \mathbf{S}_* - \widehat{\mathbf{S}}$ and $\mathbf{S}_t := \widehat{\mathbf{S}} + t\mathbf{R}$ for $t \in [0,1]$. Define $\mu_t := \mu_{\mathbf{S}_t}(\mathbf{X})$ and $\mathbf{\Gamma}_t := \sum_i p_{\mathbf{S}_t,i}(\mathbf{X})\mathbf{x}_i\mathbf{x}_i^\top - \mu_t\mu_t^\top$. Differentiating the displayed exponential weights using the quotient rule gives

$$\frac{d}{dt}\mu_t = \frac{1}{\sqrt{h}}\mathbf{\Gamma}_t\mathbf{R}^\top\mathbf{x}_L.$$

The matrix $\mathbf{\Gamma}_t$ is a covariance matrix and is positive semidefinite. Its trace satisfies $\operatorname{tr}(\mathbf{\Gamma}_t) = \sum_i p_{\mathbf{S}_t,i}\|\mathbf{x}_i\|_2^2 - \|\mu_t\|_2^2 \leq \rho^2$. Thus $\|\mathbf{\Gamma}_t\|_2 \leq \rho^2$, since its largest eigenvalue is no larger than the sum of its nonnegative eigenvalues. Integration yields

$$\|\mu_{\mathbf{S}_*}(\mathbf{X}) - \mu_{\widehat{\mathbf{S}}}(\mathbf{X})\|_2 \leq \frac{1}{\sqrt{h}}\int_0^1 \|\mathbf{\Gamma}_t\|_2\|\mathbf{R}\|_2\|\mathbf{x}_L\|_2\,dt$$
$$\leq \frac{\rho^3}{\sqrt{h}}\|\mathbf{R}\|_2.$$

The first inequality uses the fundamental theorem of calculus and submultiplicativity. The second uses the uniform bounds on the integrand. Multiplication by $\mathbf{C}^\top\mathbf{V}_*^\top$ proves the claim. □

This result uses fixed attention scores and updates only the value matrix. The query law need not be uniform on a ball, the tokens need not be independent, and the logits need not be small.

Remark E.36 (Joint query, key, and value updates)**.** *Let $\mathcal{Q}$ be any nonempty family of allowed fixed query/key updates, and let $\widehat{\mathbf{S}}_a$ be the score matrix from $a \in \mathcal{Q}$. Assume the value-adapter budget is independent of a. For each a, define ϵ_a and $\mathbf{H}_a$ by the preceding corollary with $\widehat{\mathbf{S}} = \widehat{\mathbf{S}}_a$. Then we have*

$$\inf_{a\in\mathcal{Q},\,\mathrm{rank}(\Delta\mathbf{V})\le r_V} \mathbb{E}\|\mathbf{C}^\top(f_* - f_{a,\Delta\mathbf{V}})\|_2^2 = \inf_{a\in\mathcal{Q}} \left(\epsilon_a^2 + \sum_{j>r_V} \sigma_j(\mathbf{H}_a)^2\right).$$

This follows by first minimizing over $\Delta\mathbf{V}$ for each fixed a. If $r_V \ge \mathrm{rank}(\mathbf{C})$, the right side is $\inf_{a\in\mathcal{Q}} \epsilon_a^2$. This family permits value updates. A lower bound proved for Q/K-only updates with a shared value matrix does not automatically apply to it.

Proposition E.37 (Nonzero error with a rank-one verifier)**.** *For attention, take $L = 2$, $d = h = q = c = 1$, $\mathbf{C} = 1$, $\mathbf{V}_* = 1$, $\mathbf{V}_0 = 0$, $\mathbf{S}_* = s > 0$, and $\widehat{\mathbf{S}} = 0$. Let $\mathbf{X}$ be equally likely to have rows $(-a, a)$ or $(a, -a)$, with $a > 0$. Every value-adapted candidate has verified mean squared error $a^2\tanh^2(sa^2) > 0$, even when $r_V = \mathrm{rank}(\mathbf{C}) = 1$.*

Proof. **Step 1. Compute the two pooled features.** Then $\mu_{\widehat{\mathbf{S}}}(\mathbf{X}) = 0$, whereas $\mu_{\mathbf{S}_*}(\mathbf{X})$ equals $a\tanh(sa^2)$ or $-a\tanh(sa^2)$, respectively.

Step 2. Evaluate every candidate. Every value-adapted candidate therefore has zero output, and its verified mean squared error is exactly $a^2\tanh^2(sa^2) > 0$. Here $r_V = \mathrm{rank}(\mathbf{C}) = 1$ already allows every scalar value update. The remaining error comes from fixed pooled features that carry no information about the target sign. Thus low verifier rank removes the visible readout rank bottleneck, not every possible discrepancy between target and candidate representations. □

E.4 TRANSFORMERS

Definition E.38 (Transformer model)**.** *Let L, d, m, H, h, p, q be integers satisfying $L \ge 2$, $d \ge 2$, and $m, H, h, p, q \ge 1$. The symbols L, d, and m denote the sequence length, model width, and number of layers, respectively. Each layer has H attention heads. Each head has query and key dimension h and value dimension p. The final output dimension is q.*
For every layer $\ell \in \{1, \dots, m\}$ and head $a \in \{1, \dots, H\}$, let

$$\mathbf{Q}_{\ell a}, \mathbf{K}_{\ell a} \in \mathbb{R}^{d\times h}, \qquad \mathbf{V}_{\ell a} \in \mathbb{R}^{d\times p}, \qquad \mathbf{O}_\ell \in \mathbb{R}^{Hp\times d}.$$

The matrix $\mathbf{O}_{\ell a} \in \mathbb{R}^{p\times d}$ is defined to be the ath consecutive block of p rows of $\mathbf{O}_\ell$. For $\mathbf{Z} \in \mathbb{R}^{L\times d}$, define

$$\begin{aligned}\mathbf{S}_{\ell a} &:= \mathbf{Q}_{\ell a}\mathbf{K}_{\ell a}^\top,\\ A_{\ell a}(\mathbf{Z}) &:= \mathbf{P}_{\mathbf{S}_{\ell a}}(\mathbf{Z})\mathbf{Z}\mathbf{V}_{\ell a},\\ A_\ell(\mathbf{Z}) &:= [A_{\ell 1}(\mathbf{Z}), \dots, A_{\ell H}(\mathbf{Z})]\mathbf{O}_\ell.\end{aligned}$$

Here $\mathbf{P}_{\mathbf{S}}$ is the causal attention matrix defined in the attention case. Its nonmasked scores are $\mathbf{z}_i^\top\mathbf{S}\mathbf{z}_j/\sqrt{h}$.
Write $\mathbf{1}_d \in \mathbb{R}^d$ for the all-ones vector and define $\mathbf{J}_d := \mathbf{I}_d - \mathbf{1}_d\mathbf{1}_d^\top/d$. For each $s \in \{1, 2\}$, fix $\boldsymbol{\gamma}_{\ell s}, \boldsymbol{\beta}_{\ell s} \in \mathbb{R}^d$ and $e_{\ell s} > 0$. Define the vector normalization map

$$n_{\ell s}(\mathbf{z}) := \frac{\mathrm{diag}(\boldsymbol{\gamma}_{\ell s})\mathbf{J}_d\mathbf{z}}{\sqrt{\|\mathbf{J}_d\mathbf{z}\|_2^2/d + e_{\ell s}}} + \boldsymbol{\beta}_{\ell s}.$$

Define $N_{\ell s}(\mathbf{Z})$ by applying $n_{\ell s}$ to the transpose of each row of $\mathbf{Z}$ and transposing the result back to a row.
For each ℓ, fix an integer $u_\ell \ge 1$, matrices $\mathbf{B}_{\ell 1} \in \mathbb{R}^{u_\ell\times d}$ and $\mathbf{B}_{\ell 2} \in \mathbb{R}^{d\times u_\ell}$, and vectors $\mathbf{b}_{\ell 1} \in \mathbb{R}^{u_\ell}$ and $\mathbf{b}_{\ell 2} \in \mathbb{R}^d$. Let $\sigma_\ell : \mathbb{R} \to \mathbb{R}$ have Lipschitz constant $s_\ell \ge 0$, meaning that $|\sigma_\ell(t) - \sigma_\ell(t')| \le s_\ell|t - t'|$ for all real t, t'. Apply σ_ℓ coordinatewise and define

$$f_\ell(\mathbf{z}) := \mathbf{B}_{\ell 2}\sigma_\ell(\mathbf{B}_{\ell 1}\mathbf{z} + \mathbf{b}_{\ell 1}) + \mathbf{b}_{\ell 2}.$$

Define FFN_ℓ *by applying* f_ℓ *to each row in the same convention as* $N_{\ell s}$.
For a state $\mathbf{X} \in \mathbb{R}^{L\times d}$*, define*

$$\mathbf{Y}_\ell(\mathbf{X}) := \mathbf{X} + A_\ell(N_{\ell 1}(\mathbf{X})),$$
$$T_\ell(\mathbf{X}) := \mathbf{Y}_\ell(\mathbf{X}) + \mathrm{FFN}_\ell(N_{\ell 2}(\mathbf{Y}_\ell(\mathbf{X}))).$$

The tuple $\boldsymbol{\theta} := ((\mathbf{Q}_{\ell a}, \mathbf{K}_{\ell a}))_{\ell,a}$ *contains the variable attention parameters. All other parameters in this definition are fixed.*
For an input sequence $\mathbf{X}$*, define the states recursively by*

$$\mathbf{X}_{\boldsymbol{\theta},0}(\mathbf{X}) := \mathbf{X},$$
$$\mathbf{X}_{\boldsymbol{\theta},\ell}(\mathbf{X}) := T_{\boldsymbol{\theta},\ell}(\mathbf{X}_{\boldsymbol{\theta},\ell-1}(\mathbf{X})) \quad (\ell = 1,\dots,m).$$

The subscript $\boldsymbol{\theta}$ *specifies the attention parameters used in a block. Let* N_f *be either the identity or another rowwise normalization of the same form with fixed parameters* $\boldsymbol{\gamma}_f, \boldsymbol{\beta}_f \in \mathbb{R}^d$ *and* $e_f > 0$*.*
Let $\pi_L(\mathbf{Z}) := [\mathbf{Z}]_L^\top \in \mathbb{R}^d$*. Define*

$$h_{\boldsymbol{\theta}}(\mathbf{X}) := \pi_L(N_f(\mathbf{X}_{\boldsymbol{\theta},m}(\mathbf{X}))),$$
$$F_{\boldsymbol{\theta},\mathbf{W}}(\mathbf{X}) := \mathbf{W}h_{\boldsymbol{\theta}}(\mathbf{X}) \quad \textit{for } \mathbf{W} \in \mathbb{R}^{q\times d}.$$

The two additions in T_ℓ are the residual connections. We compare the real output vector $F_{\boldsymbol{\theta},\mathbf{W}}(\mathbf{X})$, before sampling tokens.

Definition E.39 (Transformer updates)**.** *Fix target parameters* $\boldsymbol{\theta}_*$ *and base parameters* $\boldsymbol{\theta}_0$*. They share the value matrices, output projections, feed-forward networks, normalizations, sequence length, and causal mask in Definition E.38. Their input states are the same input sequence* $\mathbf{X}$*. Write* $\mathbf{S}^*_{\ell a} := \mathbf{Q}^*_{\ell a}(\mathbf{K}^*_{\ell a})^\top$ *and* $\mathbf{S}^0_{\ell a} := \mathbf{Q}^0_{\ell a}(\mathbf{K}^0_{\ell a})^\top$*. Fix numbers* $M_{\ell a} > 0$ *satisfying*

$$\|\mathbf{S}^*_{\ell a}\|_2 \le M_{\ell a}, \qquad \|\mathbf{S}^0_{\ell a}\|_2 \le M_{\ell a}.$$

For nonnegative integers $r_1,\dots,r_m$*, define* $\mathcal{A}^{\mathrm{multi}}_{\mathbf{r},\mathbf{M}}$ *to be the set of collections* $\Delta := ((\Delta\mathbf{Q}_{\ell a}, \Delta\mathbf{K}_{\ell a}))_{\ell,a}$ *such that, for every* ℓ, a*,*

$$\Delta\mathbf{Q}_{\ell a}, \Delta\mathbf{K}_{\ell a} \in \mathbb{R}^{d\times h},$$
$$\mathrm{rank}(\Delta\mathbf{Q}_{\ell a}) \le r_\ell,$$
$$\mathrm{rank}(\Delta\mathbf{K}_{\ell a}) \le r_\ell,$$
$$\widehat{\mathbf{Q}}_{\ell a} := \mathbf{Q}^0_{\ell a} + \Delta\mathbf{Q}_{\ell a},$$
$$\widehat{\mathbf{K}}_{\ell a} := \mathbf{K}^0_{\ell a} + \Delta\mathbf{K}_{\ell a},$$
$$\widehat{\mathbf{S}}_{\ell a} := \widehat{\mathbf{Q}}_{\ell a}\widehat{\mathbf{K}}_{\ell a}^\top,$$
$$\|\widehat{\mathbf{S}}_{\ell a}\|_2 \le M_{\ell a}.$$

Write $\boldsymbol{\theta}_\Delta := ((\widehat{\mathbf{Q}}_{\ell a}, \widehat{\mathbf{K}}_{\ell a}))_{\ell,a}$*.*
The single-head update family below fixes a layer $\ell_0 \in \{1,\dots,m\}$*, a head* $a_0 \in \{1,\dots,H\}$*, and an integer* $r \in \{0,\dots,d\}$*. Assume that* $\mathbf{Q}^0_{\ell a} = \mathbf{Q}^*_{\ell a}$ *and* $\mathbf{K}^0_{\ell a} = \mathbf{K}^*_{\ell a}$ *whenever* $(\ell, a) \ne (\ell_0, a_0)$*. Define* $\mathcal{A}^{\mathrm{core}}_{r,M}$ *to contain only updates of* $\mathbf{Q}_{\ell_0 a_0}$ *and* $\mathbf{K}_{\ell_0 a_0}$ *with rank at most* r *and score norm at most* $M := M_{\ell_0 a_0}$*. All other updates are zero. The final head* $\mathbf{W}$ *is shared in this single-head update family.*

E.4.1 Lower Bound

Definition E.40 (Inputs and output condition for the lower bound)**.** *Use the single-head update family from Definition E.39. Fix an integer* $k \in \{1,\dots,d-1\}$ *and a matrix* $\mathbf{U} \in \mathbb{R}^{d\times k}$ *satisfying* $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_k$*. Fix* $\rho > 0$ *and an input distribution such that the rows of*

$$\mathbf{Z}(\mathbf{X}) := N_{\ell_0 1}(\mathbf{X}_{\boldsymbol{\theta}_*,\ell_0-1}(\mathbf{X}))$$

are $\mathbf{z}_i^\top$, where $\mathbf{z}_i := \rho \mathbf{U}\mathbf{g}_i$ and $\mathbf{g}_1, \dots, \mathbf{g}_L$ are independent and uniform on the unit ball in $\mathbb{R}^k$. Define

$$\begin{aligned}
\bar{\mathbf{S}}_* &:= \mathbf{U}^\top \mathbf{S}^*_{\ell_0 a_0} \mathbf{U}, \\
\bar{\mathbf{S}}_0 &:= \mathbf{U}^\top \mathbf{S}^0_{\ell_0 a_0} \mathbf{U}, \\
\bar{\mathbf{V}} &:= \mathbf{U}^\top \mathbf{V}_{\ell_0 a_0} \mathbf{O}_{\ell_0 a_0} \in \mathbb{R}^{k \times d}, \\
T &:= \left(\sum_{j > 2r} \sigma_j (\bar{\mathbf{S}}_* - \bar{\mathbf{S}}_0)^2 \right)^{1/2}, \\
c_{L,k} &:= \frac{(L-1)^3}{L^4 (k+2)^2 (k+4)}.
\end{aligned}$$

Let $\mathbf{Y}_*(\mathbf{X})$ and $\mathbf{Y}_\Delta(\mathbf{X})$ denote the states immediately after the first residual connection of layer ℓ_0. Let $D : \mathbb{R}^{L \times d} \to \mathbb{R}^q$ denote the common remaining computation from that state to the final output. Thus D includes the feed-forward residual connection of layer ℓ_0, all later blocks, N_f, and $\mathbf{W}$. Fix $a_{\mathrm{obs}} > 0$. The output lower-bound condition means that, for every admissible single-head update and almost every input,

$$\|D(\mathbf{Y}_*(\mathbf{X})) - D(\mathbf{Y}_\Delta(\mathbf{X}))\|_2 \geq a_{\mathrm{obs}} \|\pi_L(\mathbf{Y}_*(\mathbf{X}) - \mathbf{Y}_\Delta(\mathbf{X}))\|_2.$$

The layers before the updated head are shared, so its input $\mathbf{Z}(\mathbf{X})$ is the same for the target and candidate. The output condition applies only to the displayed pairs of states. It does not require D to be globally injective.

Theorem E.41 (Lower bound for Transformers)**.** *Assume Definition E.40, including its output lower-bound condition, $\bar{\mathbf{V}} \neq \mathbf{0}$, and*

$$\frac{\rho^2 M}{\sqrt{h}} \leq \frac{\sqrt{c_{L,k}}}{12}.$$

Then we have

$$\inf_{\Delta \in \mathcal{A}^{\mathrm{core}}_{r,M}} \mathbb{E} \|F_{\boldsymbol{\theta}_*, \mathbf{W}}(\mathbf{X}) - F_{\boldsymbol{\theta}_\Delta, \mathbf{W}}(\mathbf{X})\|_2^2 \geq \frac{a_{\mathrm{obs}}^2 \rho^6 c_{L,k}}{4h} \|\bar{\mathbf{V}}\|_F^2 T^2. \tag{13}$$

If $T > 0$, the right-hand side is positive. Moreover, the essential supremum over the stated input distribution is at least

$$\frac{a_{\mathrm{obs}} \rho^3 \sqrt{c_{L,k}}}{2\sqrt{h}} \|\bar{\mathbf{V}}\|_F T$$

for every admissible single-head update.

Proof. **Step 1: Project the head parameters.** Fix an admissible update. Define $\bar{\mathbf{Q}}_* := \mathbf{U}^\top \mathbf{Q}^*_{\ell_0 a_0}$, $\bar{\mathbf{K}}_* := \mathbf{U}^\top \mathbf{K}^*_{\ell_0 a_0}$, and define the projected base and candidate factors in the same way. All projected factors have dimensions $k \times h$. Define $\bar{r} := \min\{r, k\}$. The projected updates have rank at most $\bar{r}$ because multiplication by $\mathbf{U}^\top$ cannot increase rank and a $k \times h$ matrix has rank at most k. If $r < k$, then $\bar{r} = r$. If $r \geq k$, both sums indexed by $j > 2r$ and by $j > 2\bar{r}$ are zero because the projected score difference has size $k \times k$. Thus the tail T is unchanged when r is replaced by $\bar{r}$. Also,

$$\begin{aligned}
\|\mathbf{U}^\top \widehat{\mathbf{S}}_{\ell_0 a_0} \mathbf{U}\|_2 &\leq \|\mathbf{U}^\top\|_2 \|\widehat{\mathbf{S}}_{\ell_0 a_0}\|_2 \|\mathbf{U}\|_2 \\
&\leq M.
\end{aligned}$$

The first inequality is submultiplicativity. The second uses $\mathbf{U}^\top \mathbf{U} = \mathbf{I}_k$ and admissibility. The target and base projected scores satisfy the same bound.

Step 2: Identify the observed head difference. All unchanged heads have equal outputs on the common input $\mathbf{Z}(\mathbf{X})$. The residual input is also equal. Consequently,

$$\mathbf{Y}_*(\mathbf{X}) - \mathbf{Y}_\Delta(\mathbf{X}) = [\mathbf{P}_{\mathbf{S}^*_{\ell_0 a_0}}(\mathbf{Z}) - \mathbf{P}_{\widehat{\mathbf{S}}_{\ell_0 a_0}}(\mathbf{Z})] \mathbf{Z} \mathbf{V}_{\ell_0 a_0} \mathbf{O}_{\ell_0 a_0}.$$

Let $\mathbf{G} \in \mathbb{R}^{L\times k}$ have rows $\mathbf{g}_i^\top$. The identity $\mathbf{Z} = \rho\mathbf{G}\mathbf{U}^\top$ shows that the last row of the displayed difference is the difference between two attention outputs with input dimension k, head dimension h, and shared value matrix $\bar{\mathbf{V}}$.

Step 3: Apply the attention lower bound. The projected model satisfies all hypotheses of Theorem E.22 with rank budget $\bar{r}$. The first step verifies the rank and score constraints. The input law is the independent uniform-ball law by definition. The assumed radius inequality is exactly its radius condition in dimension k. Therefore,

$$\mathbb{E}\|\pi_L(\mathbf{Y}_* - \mathbf{Y}_\Delta)\|_2^2 \geq \frac{\rho^6 c_{L,k}}{4h}\|\bar{\mathbf{V}}\|_F^2 T^2.$$

Step 4: Pass through the shared output map. Square the output lower-bound inequality and take expectations. The result of Step 3 gives

$$\begin{aligned}\mathbb{E}\|F_{\boldsymbol{\theta}_*,\mathbf{W}} - F_{\boldsymbol{\theta}_\Delta,\mathbf{W}}\|_2^2 &\geq a_{\mathrm{obs}}^2\mathbb{E}\|\pi_L(\mathbf{Y}_* - \mathbf{Y}_\Delta)\|_2^2 \\ &\geq \frac{a_{\mathrm{obs}}^2\rho^6 c_{L,k}}{4h}\|\bar{\mathbf{V}}\|_F^2 T^2.\end{aligned}$$

The bound is uniform over the stated single-head update family, so taking its infimum proves (13). Finally, the essential supremum of a nonnegative random variable is at least its RMS norm. Taking square roots gives the last assertion. □

Lemma E.42 (Inputs for the normalized query distribution)**.** *Suppose $\ell_0 = 1$, $\boldsymbol{\gamma}_{11} = \mathbf{1}_d$, and $\boldsymbol{\beta}_{11} = \mathbf{0}$. Set $k := d - 1$ and choose $\mathbf{U} \in \mathbb{R}^{d\times k}$ such that $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_k$ and $\mathbf{U}^\top\mathbf{1}_d = \mathbf{0}$. For $0 < \rho < \sqrt{d}$, let $\mathbf{z}_i := \rho\mathbf{U}\mathbf{g}_i$, where the $\mathbf{g}_i$ are independent uniform-ball vectors in $\mathbb{R}^k$. Define input rows using*

$$\mathbf{x}_i := \frac{\sqrt{e_{11}}\,\mathbf{z}_i}{\sqrt{1 - \|\mathbf{z}_i\|_2^2/d}}.$$

Then $n_{11}(\mathbf{x}_i) = \mathbf{z}_i$ and

$$\|\mathbf{x}_i\|_2 \leq \frac{\sqrt{e_{11}}\rho}{\sqrt{1 - \rho^2/d}}.$$

Proof. **Step 1: Check that the input is defined and centered.** The inequality $\|\mathbf{z}_i\|_2 \leq \rho < \sqrt{d}$ makes the denominator strictly positive. The relation $\mathbf{U}^\top\mathbf{1}_d = \mathbf{0}$ gives $\mathbf{J}_d\mathbf{z}_i = \mathbf{z}_i$. Since $\mathbf{x}_i$ is a scalar multiple of $\mathbf{z}_i$, it also gives $\mathbf{J}_d\mathbf{x}_i = \mathbf{x}_i$.

Step 2: Compute the normalization denominator. Set $t_i := \|\mathbf{z}_i\|_2^2/d$. Then we have

$$\begin{aligned}\frac{\|\mathbf{x}_i\|_2^2}{d} + e_{11} &= \frac{e_{11}t_i}{1 - t_i} + e_{11} \\ &= \frac{e_{11}}{1 - t_i}.\end{aligned}$$

The first equality uses the definition of $\mathbf{x}_i$. The second equality combines the two fractions. Substituting this expression into n_{11} gives $n_{11}(\mathbf{x}_i) = \mathbf{z}_i$.

Step 3: Bound the input norm. Use $\|\mathbf{z}_i\|_2 \leq \rho$ in the numerator and $1 - \|\mathbf{z}_i\|_2^2/d \geq 1 - \rho^2/d$ in the denominator. This gives the asserted bound. □

These bounded continuous inputs produce the normalized query distribution in Definition E.40.

Lemma E.43 (Example with a positive lower bound)**.** *If $\ell_0 = m$, $\mathrm{FFN}_m = 0$, N_f is the identity, $q = d$, and $\mathbf{W} = \mathbf{I}_d$, then the output lower-bound condition holds with $a_{\mathrm{obs}} = 1$. There are models and continuous input distributions satisfying all hypotheses of Theorem E.41 with $T > 0$.*

Proof. **Step 1: Compute the suffix in the stated setting.** There are no later layers because $\ell_0 = m$. The zero feed-forward branch leaves its residual input unchanged. The identity final normalization and output head give $D(\mathbf{Y}) = \pi_L(\mathbf{Y})$. Thus

$$\|D(\mathbf{Y}_*) - D(\mathbf{Y}_\Delta)\|_2 = \|\pi_L(\mathbf{Y}_* - \mathbf{Y}_\Delta)\|_2.$$

This proves the first assertion for every pair of states.

Step 2: Specify a setting with a positive spectral tail. Choose d and r such that $k := d - 1 > 2r$. Set $m := 1$, $H := 1$, $h := k$, and $p := 1$. Choose the normalization and input distribution from Lemma E.42. Let $\mathbf{v} := \mathbf{U}\mathbf{e}_1$, where $\mathbf{e}_1 \in \mathbb{R}^k$ is the first coordinate vector. Set

$$\mathbf{Q}^*_{11} := \sqrt{M}\mathbf{U}, \qquad \mathbf{K}^*_{11} := \sqrt{M}\mathbf{U}, \qquad \mathbf{Q}^0_{11} := \mathbf{0}, \qquad \mathbf{K}^0_{11} := \mathbf{0},$$
$$\mathbf{V}_{11} := \mathbf{v}, \qquad \mathbf{O}_{11} := \mathbf{v}^\top, \qquad \mathrm{FFN}_1 := 0, \qquad \mathbf{W} := \mathbf{I}_d.$$

Take N_f to be the identity. Then $\mathbf{S}^*_{11} = M\mathbf{U}\mathbf{U}^\top$, $\mathbf{S}^0_{11} = \mathbf{0}$, and $\bar{\mathbf{S}}_* - \bar{\mathbf{S}}_0 = M\mathbf{I}_k$. Consequently, $T = M\sqrt{k - 2r} > 0$. Also $\bar{\mathbf{V}} = \mathbf{e}_1\mathbf{v}^\top$ has Frobenius norm one. Both score norms satisfy the cap M, and the zero update is admissible. Choose $\rho > 0$ small enough that $\rho < \sqrt{d}$ and $\rho^2 M/\sqrt{h} \le \sqrt{c_{L,k}}/12$. Such a choice exists because $M, h, c_{L,k}$ are positive. Step 1 gives the required observability condition, and Lemma E.42 gives the required input law. □

Remark E.44 (Scope of the lower bound)**.** *Theorem E.41 concerns the single-head update family. It does not give a lower bound for $\mathcal{A}^{\mathrm{multi}}_{\mathbf{r},\mathbf{M}}$. Additional layer updates can cancel a difference created in one layer. The output lower-bound assumption is also substantive. Layer normalization has the exact invariance $n(\mathbf{z} + t\mathbf{1}_d) = n(\mathbf{z})$. Output projections can have kernels. Neither component is automatically bounded below on arbitrary differences. Lemma E.42 supplies a nonempty normalized-input model without assuming that centered LayerNorm produces full-dimensional uniform-ball samples. Lemma E.43 also realizes the output lower-bound condition with a positive lower bound. It does not assert reachability by discrete text prompts.*

E.4.2 Upper Bound under Weak Queries

Lemma E.45 (Bounds for normalization and feed-forward layers)**.** *For the architecture in Definition E.38, define*

$$\nu_{\ell s} := \|\boldsymbol{\gamma}_{\ell s}\|_\infty/\sqrt{e_{\ell s}},$$
$$B_{\ell s} := \sqrt{d}\|\boldsymbol{\gamma}_{\ell s}\|_\infty + \|\boldsymbol{\beta}_{\ell s}\|_2,$$
$$b_\ell := \|\mathbf{B}_{\ell 2}\|_2 s_\ell \|\mathbf{B}_{\ell 1}\|_2,$$
$$c_\ell := 1 + b_\ell \nu_{\ell 2}.$$

The map $N_{\ell s}$ is globally $\nu_{\ell s}$-Lipschitz in Frobenius norm. Each of its output rows has Euclidean norm at most $B_{\ell s}$. The map FFN_ℓ is globally b_ℓ-Lipschitz. The map $R_\ell(\mathbf{Y}) := \mathbf{Y} + \mathrm{FFN}_\ell(N_{\ell 2}(\mathbf{Y}))$ is globally c_ℓ-Lipschitz. Define $\nu_f := 1$ when N_f is the identity. Otherwise define $\nu_f := \|\boldsymbol{\gamma}_f\|_\infty/\sqrt{e_f}$. Then N_f is globally ν_f-Lipschitz.

Proof. **Step 1: Differentiate the normalized radial map.** Fix $e > 0$ and define $r_e(\mathbf{u}) := \mathbf{u}/(\|\mathbf{u}\|_2^2/d + e)^{1/2}$. Writing $a(\mathbf{u}) := (\|\mathbf{u}\|_2^2/d + e)^{1/2}$, direct differentiation gives

$$Dr_e(\mathbf{u}) = \frac{\mathbf{I}_d}{a(\mathbf{u})} - \frac{\mathbf{u}\mathbf{u}^\top}{d\, a(\mathbf{u})^3}.$$

On vectors perpendicular to $\mathbf{u}$, its eigenvalue is $1/a(\mathbf{u})$. On the direction $\mathbf{u}$, its eigenvalue is $e/a(\mathbf{u})^3$. Both numbers are nonnegative and at most $1/\sqrt{e}$. Integrating the derivative on a line segment proves that r_e is $1/\sqrt{e}$-Lipschitz. The matrix $\mathbf{J}_d$ has spectral norm 1. The diagonal gain matrix has spectral norm $\|\boldsymbol{\gamma}\|_\infty$. Their composition proves the asserted vector normalization constant. Summing the squared row inequalities proves the Frobenius version.

Step 2: Bound the normalized row norm. For every $\mathbf{u}$,

$$\frac{\|\mathbf{u}\|_2}{\sqrt{\|\mathbf{u}\|_2^2/d + e}} \le \sqrt{d}.$$

Multiply by the gain norm and add the bias norm by the triangle inequality. This gives $B_{\ell s}$.

Step 3: Bound the feed-forward and residual maps. Coordinatewise application of σ_ℓ is s_ℓ-Lipschitz in Euclidean norm. The two linear maps multiply this constant by their spectral norms. Biases cancel when two outputs are subtracted. This proves the constant b_ℓ for each row and then for all rows. For the residual map, the triangle inequality gives

$$\begin{aligned}\|R_\ell(\mathbf{Y}) - R_\ell(\mathbf{Y}')\|_F &\le \|\mathbf{Y} - \mathbf{Y}'\|_F + b_\ell \|N_{\ell 2}(\mathbf{Y}) - N_{\ell 2}(\mathbf{Y}')\|_F \\ &\le (1 + b_\ell \nu_{\ell 2}) \|\mathbf{Y} - \mathbf{Y}'\|_F.\end{aligned}$$

The first step uses the feed-forward constant. The second step uses the normalization constant from Step 1. The proof for N_f is the same normalization calculation or the identity case. □

Lemma E.46 (Lipschitz bounds for softmax and attention)**.** *The vector softmax map is $1/2$-Lipschitz in Euclidean norm. Let $\mathbf{S} \in \mathbb{R}^{d\times d}$ and $\mathbf{V} \in \mathbb{R}^{d\times p}$ satisfy $\|\mathbf{S}\|_2 \le M$. For two sequences $\mathbf{Z}, \mathbf{Z}' \in \mathbb{R}^{L\times d}$ with every row of norm at most B, define $A_{\mathbf{S}}(\mathbf{Z}) := \mathbf{P}_{\mathbf{S}}(\mathbf{Z})\mathbf{Z}\mathbf{V}$. Then we have*

$$\|A_{\mathbf{S}}(\mathbf{Z}) - A_{\mathbf{S}}(\mathbf{Z}')\|_F \le \|\mathbf{V}\|_2 \left(\sqrt{L} + \frac{LB^2M}{\sqrt{h}}\right) \|\mathbf{Z} - \mathbf{Z}'\|_F.$$

Proof. **Step 1: Bound the softmax derivative.** For a probability vector $\mathbf{v} := \mathrm{softmax}(\mathbf{u})$, the Jacobian is $\mathbf{J}_{\mathbf{v}} := \mathrm{diag}(\mathbf{v}) - \mathbf{v}\mathbf{v}^\top$. The sum of absolute values in row i is $2v_i(1 - v_i) \le 1/2$. Since $\mathbf{J}_{\mathbf{v}}$ is symmetric, its spectral norm is at most its maximum absolute row sum. Integrating the Jacobian along the segment between two score vectors proves the first assertion. Apply the assertion separately to each causal prefix and sum squared row bounds. This proves a $1/2$ Frobenius bound for the rowwise causal softmax map on the nonmasked scores.

Step 2: Bound the score difference. The algebraic identity

$$\mathbf{Z}\mathbf{S}\mathbf{Z}^\top - \mathbf{Z}'\mathbf{S}(\mathbf{Z}')^\top = (\mathbf{Z} - \mathbf{Z}')\mathbf{S}\mathbf{Z}^\top + \mathbf{Z}'\mathbf{S}(\mathbf{Z} - \mathbf{Z}')^\top$$

and $\|\mathbf{Z}\|_2, \|\mathbf{Z}'\|_2 \le \sqrt{L}B$ give

$$\|\mathbf{Z}\mathbf{S}\mathbf{Z}^\top - \mathbf{Z}'\mathbf{S}(\mathbf{Z}')^\top\|_F \le 2\sqrt{L}BM\|\mathbf{Z} - \mathbf{Z}'\|_F.$$

The inequality follows by bounding each product with one Frobenius factor and the remaining spectral norms. Discarding masked entries cannot increase the Frobenius norm. Divide by $\sqrt{h}$ and use Step 1 to obtain

$$\|\mathbf{P}_{\mathbf{S}}(\mathbf{Z}) - \mathbf{P}_{\mathbf{S}}(\mathbf{Z}')\|_F \le \frac{\sqrt{L}BM}{\sqrt{h}}\|\mathbf{Z} - \mathbf{Z}'\|_F.$$

Step 3: Bound the value and probability terms. Every row of an attention matrix is a probability vector. Its squared Euclidean norm is at most 1. Thus $\|\mathbf{P}_{\mathbf{S}}(\mathbf{Z}')\|_2 \le \|\mathbf{P}_{\mathbf{S}}(\mathbf{Z}')\|_F \le \sqrt{L}$. Use the decomposition

$$A_{\mathbf{S}}(\mathbf{Z}) - A_{\mathbf{S}}(\mathbf{Z}') = [\mathbf{P}_{\mathbf{S}}(\mathbf{Z}) - \mathbf{P}_{\mathbf{S}}(\mathbf{Z}')]\mathbf{Z}\mathbf{V} + \mathbf{P}_{\mathbf{S}}(\mathbf{Z}')(\mathbf{Z} - \mathbf{Z}')\mathbf{V}.$$

The first term has norm at most $LB^2M\|\mathbf{V}\|_2\|\mathbf{Z} - \mathbf{Z}'\|_F/\sqrt{h}$ by Step 2. The second term has norm at most $\sqrt{L}\|\mathbf{V}\|_2\|\mathbf{Z} - \mathbf{Z}'\|_F$ by the probability-matrix bound. Adding these bounds proves the claim. □

Lemma E.47 (Attention error after projection updates)**.** *Fix $\mathbf{Q}_*, \mathbf{K}_*, \mathbf{Q}_0, \mathbf{K}_0 \in \mathbb{R}^{d\times h}$ and $\mathbf{V} \in \mathbb{R}^{d\times p}$. Let $\mathbf{\Pi} \in \mathbb{R}^{d\times d}$ be an orthogonal projection of rank at most r. Define*

$$\begin{aligned}\Delta\mathbf{Q} &:= \mathbf{\Pi}(\mathbf{Q}_* - \mathbf{Q}_0), & \Delta\mathbf{K} &:= \mathbf{\Pi}(\mathbf{K}_* - \mathbf{K}_0), \\ \widehat{\mathbf{Q}} &:= \mathbf{Q}_0 + \Delta\mathbf{Q}, & \widehat{\mathbf{K}} &:= \mathbf{K}_0 + \Delta\mathbf{K}, \\ \mathbf{S}_* &:= \mathbf{Q}_*\mathbf{K}_*^\top, & \widehat{\mathbf{S}} &:= \widehat{\mathbf{Q}}\widehat{\mathbf{K}}^\top.\end{aligned}$$

Suppose $\|\mathbf{S}_*\|_2, \|\widehat{\mathbf{S}}\|_2 \le M$. *For a sequence* $\mathbf{Z}$ *with row norms at most* B,

$$\|\mathbf{P}_{\mathbf{S}_*}(\mathbf{Z})\mathbf{Z}\mathbf{V} - \mathbf{P}_{\widehat{\mathbf{S}}}(\mathbf{Z})\mathbf{Z}\mathbf{V}\|_F \le \frac{2LB^2M\|\mathbf{V}\|_2}{\sqrt{h}}\|\mathbf{Z}(\mathbf{I}_d - \mathbf{\Pi})\|_F.$$

Both updates have rank at most r. *If* $\mathbf{\Pi} = \mathbf{U}\mathbf{U}^\top$ *with* $\mathbf{U} \in \mathbb{R}^{d\times r'}$ *and* $r' \le r$, *explicit LoRA factors for* $\Delta\mathbf{Q}$ *are* $\mathbf{U}$ *and* $\mathbf{U}^\top(\mathbf{Q}_* - \mathbf{Q}_0)$. *The analogous factors realize* $\Delta\mathbf{K}$.

Proof. **Step 1: Verify the projection identity and ranks.** Since $\mathbf{\Pi}^2 = \mathbf{\Pi}$, $\mathbf{\Pi}\widehat{\mathbf{Q}} = \mathbf{\Pi}\mathbf{Q}_*$ and $\mathbf{\Pi}\widehat{\mathbf{K}} = \mathbf{\Pi}\mathbf{K}_*$. Consequently, $\mathbf{\Pi}(\mathbf{S}_* - \widehat{\mathbf{S}})\mathbf{\Pi} = \mathbf{0}$. The displayed factors give the claimed LoRA realization and rank bounds.

Step 2: Bound the score residual. Define $\mathbf{R} := \mathbf{S}_* - \widehat{\mathbf{S}}$ and $\mathbf{E} := \mathbf{Z}(\mathbf{I}_d - \mathbf{\Pi})$. The score bounds imply $\|\mathbf{R}\|_2 \le 2M$ by the triangle inequality. Using $\mathbf{\Pi}\mathbf{R}\mathbf{\Pi} = \mathbf{0}$, write

$$\mathbf{Z}\mathbf{R}\mathbf{Z}^\top = \mathbf{E}\mathbf{R}\mathbf{Z}^\top + \mathbf{Z}\mathbf{\Pi}\mathbf{R}\mathbf{E}^\top.$$

Indeed, substitute $\mathbf{Z} = \mathbf{E} + \mathbf{Z}\mathbf{\Pi}$ in the left factor and then substitute $\mathbf{Z}^\top = \mathbf{E}^\top + \mathbf{\Pi}\mathbf{Z}^\top$ in the second term. The remaining $\mathbf{Z}\mathbf{\Pi}\mathbf{R}\mathbf{\Pi}\mathbf{Z}^\top$ term is zero. It follows that

$$\begin{aligned}\|\mathbf{Z}\mathbf{R}\mathbf{Z}^\top\|_F &\le \|\mathbf{E}\|_F\|\mathbf{R}\|_2\|\mathbf{Z}\|_2 + \|\mathbf{Z}\mathbf{\Pi}\|_2\|\mathbf{R}\|_2\|\mathbf{E}\|_F \\ &\le 4\sqrt{L}BM\|\mathbf{E}\|_F.\end{aligned}$$

The first inequality bounds the two products. The second uses $\|\mathbf{Z}\mathbf{\Pi}\|_2 \le \|\mathbf{Z}\|_2 \le \sqrt{L}B$.

Step 3: Pass through attention probabilities and values. Divide the last score bound by $\sqrt{h}$. The softmax assertion of Lemma E.46 then gives

$$\|\mathbf{P}_{\mathbf{S}_*}(\mathbf{Z}) - \mathbf{P}_{\widehat{\mathbf{S}}}(\mathbf{Z})\|_F \le \frac{2\sqrt{L}BM}{\sqrt{h}}\|\mathbf{E}\|_F.$$

Multiply this by $\|\mathbf{Z}\mathbf{V}\|_2 \le \sqrt{L}B\|\mathbf{V}\|_2$. This proves the asserted output bound. □

Definition E.48 (Lipschitz constants for the network)**.** *Use the constants in Lemma E.45 and define*

$$\begin{aligned}l_{\ell a} &:= \|\mathbf{V}_{\ell a}\|_2\left(\sqrt{L} + \frac{LB_{\ell 1}^2 M_{\ell a}}{\sqrt{h}}\right), \\ l_\ell &:= \|\mathbf{O}_\ell\|_2\left(\sum_{a=1}^{H} l_{\ell a}^2\right)^{1/2}, \\ \kappa_\ell &:= c_\ell(1 + \nu_{\ell 1} l_\ell), \\ K_\ell &:= \prod_{j=\ell+1}^{m} \kappa_j.\end{aligned}$$

An empty product is defined to be 1. For a shared output head $\mathbf{W}$, *define*

$$K_D := \|\mathbf{W}\|_2 \nu_f c_{\ell_0} K_{\ell_0}.$$

Lemma E.49 (Lipschitz bounds for the network)**.** *Every target or admissible candidate block* T_ℓ *is globally* κ_ℓ*-Lipschitz in Frobenius norm. The shared output map* D *in Definition E.40, when used for the single-head update family without imposing its input-law assumptions, is globally* K_D*-Lipschitz.*

Proof. **Step 1: Bound multi-head attention on normalized inputs.** Lemma E.45 bounds all rows of $N_{\ell 1}(\mathbf{X})$ by $B_{\ell 1}$ for every state $\mathbf{X}$. Lemma E.46 therefore supplies the constant $l_{\ell a}$ for each head on these inputs. The squared Frobenius norm of concatenated head differences is the sum of their squared Frobenius norms. Multiplication by $\mathbf{O}_\ell$ gives the constant l_ℓ.

Step 2: Include the first residual and the feed-forward residual. The first residual map is $\mathbf{X} \mapsto \mathbf{X} + A_\ell(N_{\ell 1}(\mathbf{X}))$. Its Lipschitz constant is at most $1 + l_\ell \nu_{\ell 1}$ by the triangle inequality and composition. The second residual map has constant c_ℓ by Lemma E.45. Multiplication of the two constants gives κ_ℓ. These bounds hold for every residual state because normalization always maps its input into the same bounded row domain. In particular, they cover all target, candidate, and hybrid states.

Step 3: Include the suffix and final observation. The suffix begins with the second residual of layer ℓ_0. It then applies blocks $\ell_0 + 1, \dots, m$, the final normalization, the last-row selection, and $\mathbf{W}$. Their constants are, respectively, $c_{\ell_0}, \kappa_{\ell_0+1}, \dots, \kappa_m, \nu_f$, 1, and $\|\mathbf{W}\|_2$. Their product is K_D. □

Theorem E.50 (Upper bound under restricted queries)**.** *Use the single-head update family with any input law. Define the shared normalized attention input $\mathbf{Z}(\mathbf{X})$ as in Definition E.40, without requiring its uniform-ball law. Define*

$$\begin{aligned}
\mathbf{\Sigma} &:= \mathbb{E}[\mathbf{Z}(\mathbf{X})^\top \mathbf{Z}(\mathbf{X})] \in \mathbb{R}^{d\times d}, \\
t_r &:= \sum_{j>r} \lambda_j(\mathbf{\Sigma}), \\
\mathbf{V}_{\text{eff}} &:= \mathbf{V}_{\ell_0 a_0} \mathbf{O}_{\ell_0 a_0} \in \mathbb{R}^{d\times d}, \\
A_r &:= \frac{2LB_{\ell_0 1}^2 M \|\mathbf{V}_{\text{eff}}\|_2}{\sqrt{h}}.
\end{aligned}$$

The eigenvalues $\lambda_j(\mathbf{\Sigma})$ are ordered nonincreasingly. Let $\mathbf{\Pi}$ project onto the first $\min(r, d)$ corresponding eigenvectors. Construct the two updates of Lemma E.47 for the selected head. Assume that their resulting score matrix has spectral norm at most M. Then we have

$$\inf_{\Delta \in \mathcal{A}^{\text{core}}_{r,M}} \|F_{\boldsymbol{\theta}_*,\mathbf{W}} - F_{\boldsymbol{\theta}_\Delta,\mathbf{W}}\|_{L^2} \le K_D A_r \sqrt{t_r}. \tag{14}$$

Proof. **Step 1: Check admissibility and the local error.** Lemma E.47 verifies the two rank bounds. The additional score-norm assumption verifies the remaining constraint in the single-head update family. The unchanged heads and the common residual input cancel. Apply that lemma with value matrix $\mathbf{V}_{\text{eff}}$ to obtain

$$\|\mathbf{Y}_*(\mathbf{X}) - \mathbf{Y}_\Delta(\mathbf{X})\|_F \le A_r \|\mathbf{Z}(\mathbf{X})(\mathbf{I}_d - \mathbf{\Pi})\|_F.$$

Step 2: Evaluate the residual input energy. By the definition of the Frobenius norm and linearity of trace,

$$\begin{aligned}
\mathbb{E}\|\mathbf{Z}(\mathbf{I}_d - \mathbf{\Pi})\|_F^2 &= \operatorname{tr}((\mathbf{I}_d - \mathbf{\Pi})\mathbf{\Sigma}(\mathbf{I}_d - \mathbf{\Pi})) \\
&= \operatorname{tr}((\mathbf{I}_d - \mathbf{\Pi})\mathbf{\Sigma}) \\
&= t_r.
\end{aligned}$$

The second equality uses $(\mathbf{I}_d - \mathbf{\Pi})^2 = \mathbf{I}_d - \mathbf{\Pi}$ and cyclicity of trace. The third equality uses the chosen eigenspace of $\mathbf{\Sigma}$.

Step 3: Apply the shared output map bound. Lemma E.49 gives

$$\|F_{\boldsymbol{\theta}_*,\mathbf{W}} - F_{\boldsymbol{\theta}_\Delta,\mathbf{W}}\|_{L^2} \le K_D \|\mathbf{Y}_* - \mathbf{Y}_\Delta\|_{L^2} \le K_D A_r \sqrt{t_r}.$$

The second inequality uses Steps 1 and 2. The constructed update belongs to the family, so its error bounds the infimum from above and proves (14). □

Theorem E.51 (Upper bound with updates in every layer)**.** *Use $\mathcal{A}^{\mathrm{multi}}_{\mathbf{r},\mathbf{M}}$ and a shared final head $\mathbf{W}$. For each layer, define the target-prefix quantities*

$$\begin{aligned}\mathbf{Z}_\ell(\mathbf{X}) &:= N_{\ell 1}(\mathbf{X}_{\boldsymbol{\theta}_*,\ell-1}(\mathbf{X})),\\ \boldsymbol{\Sigma}_\ell &:= \mathbb{E}[\mathbf{Z}_\ell^\top \mathbf{Z}_\ell],\\ t_\ell &:= \sum_{j>r_\ell} \lambda_j(\boldsymbol{\Sigma}_\ell).\end{aligned}$$

Let $\boldsymbol{\Pi}_\ell$ be the leading eigenspace projection of dimension $\min(r_\ell, d)$. For every head in that layer, define

$$\begin{aligned}\Delta\mathbf{Q}_{\ell a} &:= \boldsymbol{\Pi}_\ell(\mathbf{Q}^*_{\ell a} - \mathbf{Q}^0_{\ell a}),\\ \Delta\mathbf{K}_{\ell a} &:= \boldsymbol{\Pi}_\ell(\mathbf{K}^*_{\ell a} - \mathbf{K}^0_{\ell a}).\end{aligned}$$

Assume that each constructed score satisfies its bound $M_{\ell a}$. Define

$$\begin{aligned}a_{\ell a} &:= \frac{2LB^2_{\ell 1} M_{\ell a}\|\mathbf{V}_{\ell a}\|_2}{\sqrt{h}},\\ C_\ell &:= c_\ell\|\mathbf{O}_\ell\|_2 \left(\sum_{a=1}^{H} a^2_{\ell a}\right)^{1/2}.\end{aligned}$$

The constructed candidate satisfies

$$\|h_{\boldsymbol{\theta}_*} - h_{\boldsymbol{\theta}_\Delta}\|_{L^2} \le \nu_f \sum_{\ell=1}^{m} K_\ell C_\ell \sqrt{t_\ell}. \tag{15}$$

Consequently,

$$\inf_{\Delta\in\mathcal{A}^{\mathrm{multi}}_{\mathbf{r},\mathbf{M}}} \|F_{\boldsymbol{\theta}_*,\mathbf{W}} - F_{\boldsymbol{\theta}_\Delta,\mathbf{W}}\|_{L^2} \le \|\mathbf{W}\|_2 \nu_f \sum_{\ell=1}^{m} K_\ell C_\ell \sqrt{t_\ell}. \tag{16}$$

Proof. **Step 1: Bound each block discrepancy at a target state.** For each head, Lemma E.47 gives

$$\|A^*_{\ell a}(\mathbf{Z}_\ell) - \widehat{A}_{\ell a}(\mathbf{Z}_\ell)\|_F \le a_{\ell a}\|\mathbf{Z}_\ell(\mathbf{I}_d - \boldsymbol{\Pi}_\ell)\|_F.$$

Square and sum the head inequalities to bound the concatenation. Multiply by $\|\mathbf{O}_\ell\|_2$. Then apply the shared second residual map, whose constant is c_ℓ. This gives

$$\|T^*_\ell(\mathbf{X}_{\boldsymbol{\theta}_*,\ell-1}) - \widehat{T}_\ell(\mathbf{X}_{\boldsymbol{\theta}_*,\ell-1})\|_F \le C_\ell\|\mathbf{Z}_\ell(\mathbf{I}_d - \boldsymbol{\Pi}_\ell)\|_F.$$

The trace calculation in Step 2 of the preceding proof gives

$$\varepsilon_\ell := \|T^*_\ell(\mathbf{X}_{\boldsymbol{\theta}_*,\ell-1}) - \widehat{T}_\ell(\mathbf{X}_{\boldsymbol{\theta}_*,\ell-1})\|_{L^2} \le C_\ell\sqrt{t_\ell}.$$

Step 2: Define the hybrids. For $\ell \in \{0,\dots,m\}$, let $\mathbf{H}_\ell(\mathbf{X})$ be the state obtained by applying the first ℓ target blocks and then the remaining $m-\ell$ candidate blocks. Thus $\mathbf{H}_0 = \mathbf{X}_{\boldsymbol{\theta}_\Delta,m}$ and $\mathbf{H}_m = \mathbf{X}_{\boldsymbol{\theta}_*,m}$. For $\ell \in \{1,\dots,m\}$, define $G_{>\ell} := \widehat{T}_m \circ \cdots \circ \widehat{T}_{\ell+1}$, with $G_{>m}$ equal to the identity. Then we have

$$\begin{aligned}\mathbf{H}_{\ell-1}(\mathbf{X}) - \mathbf{H}_\ell(\mathbf{X}) = G_{>\ell}(&\widehat{T}_\ell(\mathbf{X}_{\boldsymbol{\theta}_*,\ell-1}(\mathbf{X})))\\ &- G_{>\ell}(T^*_\ell(\mathbf{X}_{\boldsymbol{\theta}_*,\ell-1}(\mathbf{X}))).\end{aligned}$$

Both arguments use the same target-prefix state. The suffix uses only candidate blocks.

Step 3: Sum the propagated discrepancies. Lemma E.49 gives Lipschitz constant K_ℓ for $G_{>\ell}$ on every state. Therefore

$$\|\mathbf{H}_{\ell-1} - \mathbf{H}_\ell\|_{L^2} \le K_\ell\varepsilon_\ell.$$

The identity $\mathbf{H}_0 - \mathbf{H}_m = \sum_{\ell=1}^{m}(\mathbf{H}_{\ell-1} - \mathbf{H}_\ell)$ and the triangle inequality in L^2 imply

$$\|\mathbf{X}_{\boldsymbol{\theta}_\Delta,m} - \mathbf{X}_{\boldsymbol{\theta}_*,m}\|_{L^2} \le \sum_{\ell=1}^{m} K_\ell \varepsilon_\ell \le \sum_{\ell=1}^{m} K_\ell C_\ell \sqrt{t_\ell}.$$

The second inequality uses Step 1. No independence of layer discrepancies is used.

Step 4: Apply the final normalization and head. The map $\pi_L \circ N_f$ has constant at most ν_f. This proves (15). Multiplication by $\mathbf{W}$ contributes $\|\mathbf{W}\|_2$. The constructed updates are admissible by their rank factorizations and the assumed score bounds. Taking the infimum proves (16). □

Remark E.52 (Query distributions at each layer)**.** *The matrices $\mathbf{\Sigma}_\ell$ are noncentral second moments of the normalized target-prefix states. They include all token positions. No equality of the token marginals is required. A small second-moment tail at the raw input does not imply small tails after nonlinear layers. Equations* (15) *and* (16) *require the stated layerwise tails. They establish existence of admissible approximations when the constructed score bounds hold. They do not establish that a training algorithm finds these updates.*

E.4.3 Upper Bound under Low-Rank Verification

Corollary E.53 (Low-rank verification for Transformers)**.** *Let $d, p, q, c \ge 1$ and $r \ge 0$ be integers. Let $h_*(\mathbf{X}) \in \mathbb{R}^d$ and $\widehat{h}(\mathbf{X}) \in \mathbb{R}^p$ be fixed target and candidate Transformer representations with finite second moments. Fix $\mathbf{W}_* \in \mathbb{R}^{q\times d}$, $\mathbf{W}_0 \in \mathbb{R}^{q\times p}$, and $\mathbf{C} \in \mathbb{R}^{q\times c}$. Define*

$$f_*(\mathbf{X}) := \mathbf{W}_* h_*(\mathbf{X}), \qquad f_{\mathbf{\Delta}}(\mathbf{X}) := (\mathbf{W}_0 + \mathbf{\Delta})\widehat{h}(\mathbf{X}).$$

Let $\mathbf{H}$ and ϵ^2 be as in Definition E.32. Then we have

$$\min_{\mathrm{rank}(\mathbf{\Delta})\le r} \mathbb{E}\|\mathbf{C}^\top(f_* - f_{\mathbf{\Delta}})\|_2^2 = \epsilon^2 + \sum_{j>r} \sigma_j(\mathbf{H})^2.$$

If $r \ge \mathrm{rank}(\mathbf{C})$, this minimum equals ϵ^2. If also $h_* = \mathbf{T}\widehat{h}$ almost surely for some fixed $\mathbf{T} \in \mathbb{R}^{d\times p}$, then $\epsilon^2 = 0$.

Proof. The displayed maps have precisely the form in Definition E.31. The moment assumptions verify its integrability requirements. Theorem E.33 gives the exact formula. Corollary E.34 gives the zero-error condition. □

A shared backbone satisfies the feature condition with $\mathbf{T} = \mathbf{I}_d$. This result fixes the backbone and updates only the output head. It does not require a uniform ball query law or a particular Transformer architecture.

Remark E.54 (Outputs and allowed updates)**.** *The conclusions concern real output vectors observed through the specified fixed matrix $\mathbf{C}$. If an actual deterministic verifier factors as $g(\mathbf{y}) = \psi(\mathbf{C}^\top\mathbf{y})$, equal projected outputs imply equal verifier outputs. Nonzero projected-error bounds require an additional stability or margin condition to imply a decision guarantee. The results do not concern sampled text directly. An output-head update must not alter $\widehat{h}$. An output-only adapter or an untied readout satisfies this requirement. Changing a tied embedding and readout generally does not. The theorems impose a rank constraint but no independent bound on adapter norm. They establish population optima or explicit constructions, not finite-sample learnability, optimization convergence, or utility preservation.*

E.5 Proofs of the Unified Main Results

We now derive the unified statements in Section D from the preceding linear, attention, and Transformer results.

E.5.1 Proof of Proposition D.3

Theorem E.55 (Restatement of Proposition D.3)**.** *Use Definitions D.1 and D.2. For attention, assume $\rho^2 M/\sqrt{h} \le \sqrt{c_{L,d}}/12$. For Transformers, assume Definition E.40, including its input distribution and output lower-bound condition, and $\rho^2 M/\sqrt{h} \le \sqrt{c_{L,k}}/12$. Then we have*

$$\inf_{f\in\mathcal{F}_r}\sup_{x\in\mathcal{X}}\|f_*(x)-f(x)\|_2 \ge \begin{cases}\rho\,\sigma_{r+1}(\mathbf{R}), & \textit{linear},\\ a_{L,d}(\mathbf{V})\left(\sum_{j>2r}\sigma_j(\mathbf{S}_*-\mathbf{S}_0)^2\right)^{1/2}, & \textit{attention},\\ a_{\mathrm{obs}}a_{L,k}(\bar{\mathbf{V}})\left(\sum_{j>2r}\sigma_j(\bar{\mathbf{S}}_*-\bar{\mathbf{S}}_0)^2\right)^{1/2}, & \textit{Transformer}.\end{cases}$$

Equality holds in the linear case.

Proof. **Case 1: Linear models.** For every admissible $\mathbf{\Delta}$, we have

$$\sup_{\|\mathbf{x}\|_2\le\rho}\|(\mathbf{R}-\mathbf{\Delta})\mathbf{x}\|_2=\rho\|\mathbf{R}-\mathbf{\Delta}\|_2.$$

This equality follows from $\rho>0$ and the definition of the spectral norm. Taking the infimum over $\mathrm{rank}(\mathbf{\Delta})\le r$ and applying Lemma E.1 gives

$$\inf_{\mathrm{rank}(\mathbf{\Delta})\le r}\sup_{\|\mathbf{x}\|_2\le\rho}\|(\mathbf{R}-\mathbf{\Delta})\mathbf{x}\|_2=\rho\,\sigma_{r+1}(\mathbf{R}).$$

Case 2: Causal attention. The update family and input domain are those of Theorem E.22. Its worst-case conclusion gives the stated inequality after substituting the definition of $a_{L,d}(\mathbf{V})$. Transposing the last-token output preserves its Euclidean norm.

Case 3: Transformers. Let $\mathcal{D}_{\mathrm{w}}$ be the input distribution in Definition E.40. For every admissible single-head update, we have

$$\sup_{x\in\mathcal{X}}\|f_*(x)-f(x)\|_2 \ge \|f_*-f\|_{L^2(\mathcal{D}_{\mathrm{w}})} \ge a_{\mathrm{obs}}a_{L,k}(\bar{\mathbf{V}})\left(\sum_{j>2r}\sigma_j(\bar{\mathbf{S}}_*-\bar{\mathbf{S}}_0)^2\right)^{1/2}.$$

The first inequality holds because $\mathcal{X}$ contains the support of $\mathcal{D}_{\mathrm{w}}$. The second follows from Theorem E.41 and the definition of $a_{L,k}$. Taking the infimum over the same single-head update family proves the claim. The distribution $\mathcal{D}_{\mathrm{w}}$ is used only in this proof. □

E.5.2 Proof of Theorem D.4

Theorem E.56 (Restatement of Theorem D.4)**.** *Use Definition D.1 and an audit law $\mathcal{D}$ supported on $\mathcal{X}$. Assume the following second moments are finite and define*

$$\mathbf{\Sigma}:=\begin{cases}\mathbb{E}[\mathbf{x}\mathbf{x}^\top], & \textit{linear},\\ \mathbb{E}[\mathbf{x}_i\mathbf{x}_i^\top] \quad \textit{for every } i\in[L], & \textit{attention},\\ \mathbb{E}[\mathbf{Z}(\mathbf{X})^\top\mathbf{Z}(\mathbf{X})], & \textit{Transformer}.\end{cases}$$

For attention, assume the token second moments are equal. For attention and Transformers, assume the projection updates satisfy the score-norm constraints in Theorems E.28 and E.50, respectively. Let $\lambda_j(\mathbf{\Sigma})$ be the eigenvalues in decreasing order, and define

$$b:=\begin{cases}\|\mathbf{R}\|_2, & \textit{linear},\\ \dfrac{2\rho^2 M\|\mathbf{V}\|_2}{\sqrt{h}}\left(1+e^{\rho^2 M/\sqrt{h}}\right), & \textit{attention},\\ K_D A_r, & \textit{Transformer},\end{cases}$$

where $\mathbf{Z}$*,* K_D*, and* A_r *are specified in Theorem E.50. Then we have*

$$\inf_{f \in \mathcal{F}_r} \|f_* - f\|_{L^2(\mathcal{D})} \le b \left(\sum_{j>r} \lambda_j(\mathbf{\Sigma}) \right)^{1/2}.$$

Proof. **Case 1: Linear models.** Let $\mathbf{P}$ be the orthogonal projection onto the leading $\min(r, d)$ eigenvectors of $\mathbf{\Sigma}$ and define $\mathbf{\Delta} := \mathbf{RP}$. We have $\operatorname{rank}(\mathbf{\Delta}) \le \operatorname{rank}(\mathbf{P}) \le r$. For every $\mathbf{x}$, we have

$$\begin{aligned} \|(\mathbf{R} - \mathbf{\Delta})\mathbf{x}\|_2^2 &= \|\mathbf{R}(\mathbf{I}_d - \mathbf{P})\mathbf{x}\|_2^2 \\ &\le \|\mathbf{R}\|_2^2 \|(\mathbf{I}_d - \mathbf{P})\mathbf{x}\|_2^2. \end{aligned}$$

The equality substitutes the definition of $\mathbf{\Delta}$. The inequality follows from the defining property of the spectral norm. Taking expectations gives

$$\begin{aligned} \mathbb{E}\|(\mathbf{I}_d - \mathbf{P})\mathbf{x}\|_2^2 &= \operatorname{tr}((\mathbf{I}_d - \mathbf{P})\mathbf{\Sigma}) \\ &= \sum_{j>r} \lambda_j(\mathbf{\Sigma}). \end{aligned}$$

The first equality uses $\mathbf{P} = \mathbf{P}^\top = \mathbf{P}^2$ and cyclicity of trace. The second uses the chosen eigenspace of $\mathbf{\Sigma}$. Combining these inequalities, taking square roots, and then taking the infimum proves the linear case.

Case 2: Causal attention. Take the square root of the conclusion of Theorem E.28. Its coefficient is the attention value of b above. The score-norm hypothesis ensures that its constructed update belongs to $\mathcal{A}_{r,M}$.

Case 3: Transformers. Apply Theorem E.50 to the same normalized head input $\mathbf{Z}$ and audit law. Its coefficient is $K_D A_r$ and its second-moment matrix is $\mathbf{\Sigma}$ above. Its admissibility hypothesis places the constructed update in the stated single-head update family. □

E.5.3 Proof of Theorem D.7

Theorem E.57 (Restatement of Theorem D.7)**.** *Under Definitions D.5 and D.6, assume* $s \ge \operatorname{rank}(\mathbf{C})$*. For* $\mathbf{\Delta} \in \mathbb{R}^{q \times d}$*, define* $f_{\mathbf{\Delta}}(x) := (\mathbf{W}_0 + \mathbf{\Delta})\mathbf{z}(x)$*. Then we have*

$$\inf_{\operatorname{rank}(\mathbf{\Delta}) \le s} \mathbb{E}_{x \sim \mathcal{D}} \|\mathbf{C}^\top (f_*(x) - f_{\mathbf{\Delta}}(x))\|_2^2 = \varepsilon_{\mathbf{C}}^2.$$

Moreover, $\varepsilon_{\mathbf{C}} = 0$ *if and only if there exists* $\mathbf{B} \in \mathbb{R}^{c \times d}$ *such that* $\mathbf{C}^\top f_*(x) = \mathbf{B}\mathbf{z}(x)$ *for* $\mathcal{D}$*-almost every* x*.*

The proof applies to all three model families in Definition D.5 and directly proves Theorem 5.2 when $\mathcal{D} = \mathcal{D}_q$.

Proof. All expectations below are over $x \sim \mathcal{D}$. The finite second moments of f_* and $\mathbf{z}$ ensure that every prediction error below is finite.

Step 1: Bound the error from below. For every admissible $\mathbf{\Delta}$, define $\mathbf{B}_{\mathbf{\Delta}} := \mathbf{C}^\top(\mathbf{W}_0 + \mathbf{\Delta}) \in \mathbb{R}^{c \times d}$. Then we have

$$\begin{aligned} \mathbb{E}\|\mathbf{C}^\top (f_* - f_{\mathbf{\Delta}})\|_2^2 &= \mathbb{E}\|\mathbf{C}^\top f_* - \mathbf{B}_{\mathbf{\Delta}}\mathbf{z}\|_2^2 \\ &\ge \varepsilon_{\mathbf{C}}^2. \end{aligned}$$

The equality substitutes the definitions of $f_{\mathbf{\Delta}}$ and $\mathbf{B}_{\mathbf{\Delta}}$. The inequality follows from the infimum defining $\varepsilon_{\mathbf{C}}^2$. Taking the infimum over admissible updates proves the lower bound.

Step 2: Construct a low-rank update. Fix any $\mathbf{B} \in \mathbb{R}^{c \times d}$ and define

$$\begin{aligned} \mathbf{P} &:= \mathbf{C}^\top (\mathbf{C}^\top)^\dagger \in \mathbb{R}^{c \times c}, \\ \mathbf{\Delta}_{\mathbf{B}} &:= (\mathbf{C}^\top)^\dagger (\mathbf{B} - \mathbf{C}^\top \mathbf{W}_0) \in \mathbb{R}^{q \times d}. \end{aligned}$$

Here † denotes the Moore–Penrose inverse. The matrix $\mathbf{P}$ is the orthogonal projection onto the column space of $\mathbf{C}^\top$. Thus $\mathbf{P}\mathbf{C}^\top = \mathbf{C}^\top$, and we have

$$\operatorname{rank}(\mathbf{\Delta_B}) \le \operatorname{rank}((\mathbf{C}^\top)^\dagger) = \operatorname{rank}(\mathbf{C}) \le s.$$

The first inequality uses the rank bound for a matrix product. The equality holds because transposition and the Moore–Penrose inverse preserve rank. The last inequality is the assumed rank budget, so $\mathbf{\Delta_B}$ is admissible. Moreover,

$$\begin{aligned}\mathbf{C}^\top(\mathbf{W}_0 + \mathbf{\Delta_B}) &= \mathbf{C}^\top\mathbf{W}_0 + \mathbf{P}(\mathbf{B} - \mathbf{C}^\top\mathbf{W}_0)\\ &= \mathbf{P}\mathbf{B}.\end{aligned}$$

The first equality substitutes the constructed update and the definition of $\mathbf{P}$. The second uses $\mathbf{P}\mathbf{C}^\top = \mathbf{C}^\top$ and cancels the two base-weight terms.

Step 3: Bound the error from above. For every x, we have

$$\begin{aligned}\|\mathbf{C}^\top(f_*(x) - f_{\mathbf{\Delta_B}}(x))\|_2^2 &= \|\mathbf{C}^\top f_*(x) - \mathbf{P}\mathbf{B}\mathbf{z}(x)\|_2^2\\ &= \|\mathbf{P}(\mathbf{C}^\top f_*(x) - \mathbf{B}\mathbf{z}(x))\|_2^2\\ &\le \|\mathbf{C}^\top f_*(x) - \mathbf{B}\mathbf{z}(x)\|_2^2.\end{aligned}$$

The first equality uses the identity from Step 2. The second uses $\mathbf{P}\mathbf{C}^\top = \mathbf{C}^\top$. The inequality holds because an orthogonal projection cannot increase the Euclidean norm. Taking expectations gives, for every $\mathbf{B}$,

$$\inf_{\operatorname{rank}(\mathbf{\Delta})\le s} \mathbb{E}\|\mathbf{C}^\top(f_* - f_{\mathbf{\Delta}})\|_2^2 \le \mathbb{E}\|\mathbf{C}^\top f_* - \mathbf{B}\mathbf{z}\|_2^2.$$

The left side is no larger than the error of the admissible update $\mathbf{\Delta_B}$. Taking the infimum over $\mathbf{B}$ bounds the left side by $\varepsilon_{\mathbf{C}}^2$. Combining this with Step 1 proves the equality.

Step 4: Characterize zero error. The functions $\{x \mapsto \mathbf{B}\mathbf{z}(x) : \mathbf{B} \in \mathbb{R}^{c\times d}\}$ form a finite-dimensional linear subspace of $L^2(\mathcal{D};\mathbb{R}^c)$. Indeed, they depend linearly on the cd entries of $\mathbf{B}$, and the second-moment assumption places each function in this space. Every finite-dimensional linear subspace of a normed space is closed. Therefore its distance from $\mathbf{C}^\top f_*$ is zero if and only if $\mathbf{C}^\top f_*$ belongs to this subspace. By the definition of $\varepsilon_{\mathbf{C}}^2$, this is equivalent to the stated almost-sure representation. For linear models, take $\mathbf{B} := \mathbf{C}^\top\mathbf{W}_*$. For attention with matched scores, take $\mathbf{B} := \mathbf{C}^\top\mathbf{V}_*^\top$. For Transformers with a shared backbone, take $\mathbf{B} := \mathbf{C}^\top\mathbf{W}_*$. Each choice gives the representation in Definition D.5. □