CAIC: A DECISION-AWARE INTERVENTION FRAMEWORK FOR COUNTERFACTUAL SENSITIVITY IN LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models can change consequential decisions after demographic-only edits to otherwise identical inputs. We introduce **Counterfactual Activation-based Intervention with Certification (CAIC)**, a white-box framework for reducing decision-level counterfactual sensitivity and bounding what remains. Through a *Locate–Forecast–Select–Verify* pipeline, CAIC admits low-rank activation interventions only when they satisfy task-specific requirements at the deployed scoring endpoint, and abstains otherwise. We instantiate two operators: CAIC-S, a spectral intervention that removes dominant counterfactual directions, and CAIC-T, a task-protected intervention that limits suppression of task-relevant signal. On a controlled task, CAIC-S reduces the mean counterfactual gap by 96.6%, with a 95% population upper confidence bound of 0.030. On a semi-synthetic salary-screening task with entangled counterfactual and task signals, CAIC-S reduces final-output raw gaps by 69%–78% across two 7B models, attaining the lowest residual gap among seven matched baselines, though partly through score compression. As an early-exit decision rule at its intervention site, admitted CAIC-T reduces the gap by 32%–74% while better preserving task signal and score structure; however, the mitigation does not persist through downstream computation and CAIC abstains there. Our results underscore that counterfactual intervention should be selected and evaluated at its deployed scoring endpoint.

1 INTRODUCTION

Large language models (LLMs) have increasingly been adopted in high-stakes decision-making settings (Otani et al., 2025). However, they can exhibit *decision-level counterfactual sensitivity*: their decisions change when an input is modified along an attribute that should be irrelevant, even while decision-relevant information is held fixed. Such behavior has been documented in employment, where changing demographic cues such as an applicant's name or gender markers can alter hiring and salary recommendations despite otherwise identical qualifications (An et al., 2024). This behavior extends to other consequential settings, including clinical triage and treatment (Lee et al., 2026b), legal judgments (Sun et al., 2025; Hu et al., 2026), and college admissions (Nghiem et al., 2025; Lee et al., 2026a). Recent regulations likewise impose increasing transparency requirements on AI-assisted decision systems (New York City Council, 2021; European Parliament and Council of the European Union, 2024). Mitigation should therefore reduce sensitivity to attributes that are irrelevant to the decision without suppressing legitimate task-relevant behavior.

What should such an intervention remove? The presence of demographic information in a representation does not imply that this information influences the audited decision (Elazar et al., 2021; Shan & Mueller, 2026; Sharma & Le, 2026). Conversely, a small representational change may matter if it propagates along a direction that strongly affects the decision. This tension creates a mismatch between representation-level objectives and the quantity ultimately being audited in a practical use case: erasing or disentangling a demographic concept need not remove the variation driving a particular score, while aggressive intervention can also suppress task-relevant signal.

Existing approaches address parts of this problem. Black-box counterfactual auditing and certification can detect or bound behavioral sensitivity (Chaudhary et al., 2025), while concept-erasure and activation interventions modify internal representations (Ravfogel et al., 2020; Belrose et al., 2023b; Goyal & Daumé III, 2026). More general representation-discovery approaches can addition-

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.

ally learn auxiliary decompositions of model activations Huben et al. (2024). These methods serve broader goals, whereas the intervention needed for a specific decision can be substantially narrower. We instead ask whether a pretrained white-box model can be edited *post-training* for a declared counterfactual family and task-specific score, while explicitly controlling what functional behavior must be preserved and quantifying the sensitivity that remains.

To bridge this gap, we introduce **Counterfactual Activation-based Intervention with Certification (CAIC)** (Figure 3), a post-training framework for reducing and bounding decision-level counterfactual sensitivity. Given matched counterfactuals and a task-specific score, CAIC constructs low-rank activation interventions and organizes selection through a *Locate–Forecast–Select–Verify* pipeline. Its spectral operator (CAIC-S) targets the dominant counterfactual displacement directions, providing an aggressive sensitivity-reduction operating point. Its task-protected operator (CAIC-T) instead targets decision-relevant counterfactual signal while limiting suppression of task-relevant signal and distortion of the score distribution. Validation then selects configurations that satisfy predeclared operational requirements; if none qualifies, CAIC abstains. Our main contributions are:

- **Post-training counterfactual intervention.** We formulate post-training mitigation as reducing residual counterfactual sensitivity while preserving task-relevant behavior. CAIC reports the residual gap at the deployed endpoint with a held-out upper confidence bound on its population mean.
- **Endpoint-aware intervention and deployment.** CAIC treats the intervention site and deployed scoring endpoint as distinct components of a configuration. CAIC-S is deployed at the native final-output score, whereas an admitted CAIC-T site, operator, and frozen affine lens form an onsite early-exit decision rule. Each intervention–endpoint pair is admitted and evaluated independently; if a required endpoint does not qualify, CAIC abstains.
- **Empirical evidence for a mitigation–preservation tradeoff.** On a controlled task, CAIC-S reduces the mean counterfactual gap by 96.6%, with a one-sided population upper bound of 0.030. On the salary-screening task, full-strength CAIC-S under the matched-baseline protocol reduces final-output gaps by 69%–78% across two 7B models. When evaluated at their selected intervention sites, admitted CAIC-T interventions provide more scale-robust sensitivity reduction while better preserving task-relevant signal and score structure. For these same CAIC-T edits, however, the effect largely disappears after downstream computation, so no final-output configuration is admitted. Comparisons with concept erasure, gated steering, prompting, and score averaging further show that these objectives need not coincide.

Together, CAIC instantiates an auditable workflow for connecting a declared counterfactual audit to an endpoint-specific post-training intervention and residual bound (Appendix A).

2 Related Work

Representation intervention and bias mitigation. Concept-erasure methods such as INLP, R-LACE, ACE, and LEACE remove protected-attribute information from internal representations (Ravfogel et al., 2020; 2022; Marshall et al., 2024; Belrose et al., 2023b). Activation-based methods instead steer or modify bias-related representations directly (Li et al., 2025; Shan & Mueller, 2026), but their effects can depend on intervention site and strength and may alter other model behaviors (Goyal & Daumé III, 2026; Feng et al., 2026; Grant et al., 2026; Karvonen & Marks, 2025). These methods primarily target representational information or behavioral steering; neither objective guarantees reduced sensitivity of a particular decision. Moreover, bias metrics can improve without the underlying disparity being removed (Gonen & Goldberg, 2019; Zietlow et al., 2022; Borkan et al., 2019). CAIC selects a targeted post-training intervention under operational constraints, bounds the residual sensitivity, and audits whether gap reductions reflect score compression.

Representation discovery and disentanglement. A complementary line of interpretability work seeks more separable or human-interpretable internal representations, including SAE and concept-based architectures (Huben et al., 2024; Templeton et al., 2026; Yuksekgonul et al., 2022). Such approaches can reveal features useful for understanding or steering model behavior, but typically optimize a broader representation-level objective and may require fitting substantial auxiliary representations. CAIC does not require a globally disentangled representation: given an existing white-

box model, a declared counterfactual family, and a task-specific score, it constructs a low-rank post-training edit targeted to that decision. Additional related work is discussed in Appendix B.

3 COUNTERFACTUAL INTERVENTION WITH RESIDUAL BOUNDS

In this section, we first formalize counterfactual sensitivity with respect to a declared family of input variants and a task-specific scoring function, together with two bounds for certifying the resulting sensitivity. We then define the shared residualization family, its spectral and task-protected variants, and the *Locate–Forecast–Select–Verify* pipeline that operationalizes CAIC end to end.

3.1 RESIDUAL SENSITIVITY AND DIRECTIONAL CERTIFICATION

Let $\mathcal{M}$ be a fixed model and $x \in \mathcal{X}$ an input. A declared counterfactual family $\mathcal{D}(x) \subseteq \mathcal{X}$ specifies the variants over which the decision should remain stable, such as resumes that differ only in demographic cues. Let I be a fixed post-training activation intervention. Running the intervened model produces a representation $z_I(x) \in \mathcal{Z}$ at the selected scoring endpoint, which may be the intervention site itself or a downstream representation reached after the edited forward pass. A scoring function $F_I : \mathcal{Z} \to \mathbb{R}$ represents the model's task-specific decision (e.g., the score to determine whether a candidate's resume should advance). We define the *residual counterfactual sensitivity* as

$$\Delta_I(x) = \sup_{x' \in \mathcal{D}(x)} \left| F_I(z_I(x)) - F_I(z_I(x')) \right|.$$

Let $\phi : \mathcal{Z} \to \mathcal{H}$ map representations into a certificate space equipped with metric ρ, and define $R_I^{\phi}(x) = \sup_{x' \in \mathcal{D}(x)} \rho(\phi(z_I(x)), \phi(z_I(x')))$, we have the following theorem.

Theorem 1 (Post-intervention bound). Suppose that, for some $L < \infty$, every $z, z' \in \mathcal{Z}$ satisfies

$$|F_I(z) - F_I(z')| \le L\rho(\phi(z), \phi(z')).$$

Then $\Delta_I(x) \le L R_I^{\phi}(x)$ for every $x \in \mathcal{X}$. Moreover, if $\rho(\phi(z), \phi(z_I(x))) \le R$, then $|F_I(z) - F_I(z_I(x))| \le LR$.

Proof. Apply the Lipschitz condition to every $x' \in \mathcal{D}(x)$, take the supremum. The second statement follows by applying the same condition to every representation within radius R of $z_I(x)$. □

Remarks. Theorem 1 is agnostic to how I is constructed and to whether the certified score is read at the intervention site or downstream to it. Consequently, an intervention may be admissible and certifiable at an onsite score even if its effect does not persist to a downstream score. For CAIC, the counterfactual family specifies what variation is covered, the intervention specifies how the model is edited, and the scoring endpoint specifies which deployed decision is certified.

Directional bound. In our main experiments, the scoring function is sigmoid-affine at the selected scoring endpoint, $F_I(z) = \sigma(w_I^{\top} z + c_I)$, with additional family of functions discussed in Appendix D. Since the sigmoid is globally $1/4$-Lipschitz, the directional certificate is $\Delta_I(x) \le \epsilon_{\mathrm{dir},I}(x) = \frac{1}{4} \sup_{x' \in \mathcal{D}(x)} \left| w_I^{\top} \left(z_I(x) - z_I(x') \right) \right|$. For a finite declared family $\widehat{\mathcal{D}}(x)$, the supremum becomes a maximum, yielding a deterministic bound on every evaluated counterfactual once the intervention and scoring endpoint are fixed.

Population confidence bound. Let $q_1, \ldots, q_n \in [0, R]$ be i.i.d. per-input gaps or bounds. If $\bar{q}$ and s_q are their sample mean and standard deviation, the one-sided empirical Bernstein upper confidence bound (UCB) (Maurer & Pontil, 2009) is

$$U(q) = \bar{q} + s_q \sqrt{\frac{2\log(2/\delta)}{n}} + \frac{7R\log(2/\delta)}{3(n-1)}, \qquad \Pr(\mathbb{E}[q] \le U(q)) \ge 1 - \delta.$$

We apply empirical Bernstein across held-out inputs after aggregating each finite counterfactual family by either its mean or maximum (only the latter corresponds to $\Delta_I(x)$). Complementing ϵ_{dir}, which provides a deterministic per-input bound tied to the selected scoring direction, U is agnostic to the choice of intervention and scoring method. Appendices D and E discuss additional bounds.

3.2 INTERVENTION FAMILY

At a candidate intervention site L, let $D \in \mathbb{R}^{N\times d}$ contain the training counterfactual displacements $d_i = z_L(x_i) - z_L(x_i')$ as rows. Given an orthonormal basis $Q \in \mathbb{R}^{d\times q}$, CAIC applies

$$M_{Q,A} = I_d - QAQ^\top, \qquad Q^\top Q = I_q, \qquad 0 \preceq A \preceq I_q.$$

The basis Q specifies which activation directions are targeted, while A determines how strongly they are attenuated. Since the eigenvalues of A lie in $[0,1]$, $M_{Q,A}$ is symmetric and non-expansive. Sharing this residualizaton operator, the following variants diverge in their construction of Q and A.

CAIC-S: spectral intervention. Let $D = U\Sigma V^\top$ be the compact singular value decomposition of a counterfactual displacement matrix. CAIC-S takes $Q_{\mathrm{S}} = V_r = [v_1, \ldots, v_r]$ and $A_{\mathrm{S}} = \alpha I_r$, thereby uniformly attenuating the directions that capture the greatest counterfactual displacement energy. At full strength ($\alpha = 1$), by the Ky Fan principle (Fan, 1951), V_r minimizes the residual displacement energy among all rank-r orthogonal residualizations (proof appears in Appendix C):

$$\min_{Q^\top Q = I_r} \frac{1}{N} \left\| D(I_d - QQ^\top) \right\|_F^2 = \frac{1}{N} \sum_{j>r} \sigma_j^2.$$

CAIC-T: task-protected intervention. Removing the largest counterfactual displacement that CAIC-S targets may also degrade the task behavior that the intervention is meant to preserve. CAIC-T therefore aims to isolate counterfactual variation that is predictive of the model's decision while avoiding overlap with task-relevant signals. If the model's decision score is collected downstream of an intermediate layer L, the activation distribution at L may be distributionally divergent to that scoring function (Belrose et al., 2023a). We therefore fit an affine lens $m_L(z) = \beta_L^\top z + b_L$ to predict the unedited downstream pre-sigmoid decision margin from activations at L. Given layer-L activations z_i and its downstream margins s_i, we fit via ridge regression a local affine predictor of the downstream decision margin: $(\hat\beta_L, \hat b_L) = \arg\min_{\beta,b} \sum_i \left(s_i - (\beta^\top z_i + b)\right)^2 + \lambda_{\text{lens}} \|\beta\|_2^2$. The fitted direction $\hat\beta_L$ estimates which activation variation at L predicts the downstream decision score. We also construct a task-contrast matrix T_L from paired inputs that differ only in a declared task-relevant attribute, so each row represents variation the intervention should preserve. For counterfactual channel c, let $\bar D_c = D_c / \|D_c\|_F$, and let $\mathcal{S}_L$ be the fixed reduced span formed from leading directions of the counterfactual and task matrices together with the normalized lens direction (full construction in Appendix C). CAIC-T fits

$$\hat u_c = \arg\min_{u \in \mathcal{S}_L} \left\{ \|\bar D_c u - \bar D_c \hat\beta_L\|_2^2 + \gamma \|T_L u\|_2^2 + \lambda \|u\|_2^2 \right\}.$$

The first term targets counterfactual variation predictive of the decision score, while the second penalizes overlap with task-relevant signal. We normalize the nonzero $\hat u_c$ directions and orthonormalize their union, discarding linear dependencies, to obtain the joint basis $\widetilde Q_L$. CAIC-T then rotates this basis according to its overlap with the task-relevant signal:

$$\widetilde Q_L^\top T_L^\top T_L \widetilde Q_L = R_L \Lambda_L R_L^\top, \qquad Q_{\mathrm{T},L} = \widetilde Q_L R_L,$$

with the eigenvalues in Λ_L ordered from smallest to largest. This preserves the fitted subspace while ordering its directions from least to most aligned with the protected task contrasts. CAIC-T then uses $A_{\mathrm{T}} = \mathrm{diag}(a_1, \ldots, a_q)$, with $0 \le a_j \le 1$, where the vector $(a_1, \ldots, a_q)$ defines the *attenuation schedule* across the ordered directions. This setup allows directions with lower measured task-contrast overlap to be attenuated more strongly than directions with higher overlap. We choose the attenuation schedule and other operating points by the procedure below.

3.3 LOCATE, FORECAST, SELECT, AND VERIFY

CAIC constructs deployment configurations that differ in operator intervention site, rank, attenuation schedule and scoring endpoint, which may coincide with the intervention site or lie downstream. The process is operationalized in four stages: *Locate* enumerates candidate intervention sites, *Forecast* characterizes their behavior on fitting data, *Select* chooses an intervention subject to operational constraints, *Verify* evaluates and the selected intervention and bounds the residual sensitivity.

Locate. Using training data, we identify candidate intervention sites and construct their counterfactual displacement matrix D. For CAIC-T, we also fit an affine lens at each candidate site using a disjoint subset of this split via the procedure described in §3.2.

Forecast. Before validation, we characterize each candidate using fitting-set directional and preservation diagnostics. For a candidate operator $M_{Q,A}$, sigmoid-affine score direction w and other components defined above, we compute

$$\mathcal{F}_{\mathrm{tr}}(D, Q, A; w) = \frac{1}{4}\sqrt{\frac{1}{N}\left\|DM_{Q,A}w\right\|_2^2}.$$

A smaller $\mathcal{F}_{\mathrm{tr}}$ indicates a smaller fitting-set RMS directional bound after the candidate edit. When the intervention and scoring sites coincide, this quantity is exact for that fitting-set bound; otherwise, it is a diagnostic proxy for downstream behavior. For CAIC-T, ρ_{cf} measures how much score-aligned movement between counterfactual variants is removed. For each task-relevant comparison k, $\rho_{\mathrm{task},k}$ measures how much score-aligned task variation is removed. High ρ_{cf} and low ρ_{task} indicate a favorable repair–preservation tradeoff. Their exact definitions appear in Appendix C.

Select. Using a disjoint validation data, we retain only intervention–endpoint configurations satisfying the task-specific operational constraints. For CAIC-T, onsite and downstream scoring endpoints are treated as distinct candidates: onsite deployment additionally requires a sufficiently faithful frozen affine lens and selective removal of counterfactual over task-relevant signal. Among admissible candidates, CAIC selects the one with the smallest mean validation counterfactual gap. If no candidate is admissible, CAIC abstains.

Verify. After selection, the intervention, scoring endpoint, and scoring function are fixed and evaluated on held-out test data, including the counterfactual samples not used in training and validation.

4 Controlled validation of CAIC's pipeline

We first evaluate CAIC in a controlled setting where counterfactual decision sensitivity is introduced by construction and legitimate task-relevant behavior is independently measurable. We show that a CAIC-S operator is sufficient to identify and validate a low-rank intervention while satisfying the declared operational requirements in this setting.

4.1 Task design and operational requirements

Controlled task. We LoRA-fine-tune (Hu et al., 2021) `Qwen2.5-1.5B` (Qwen et al., 2025) to make salary-threshold decisions for occupational biographies drawn from Bias-in-Bios (De-Arteaga et al., 2019). Matched counterfactuals hold each biography fixed and swap only gendered pronouns (`he/she, him (his)/her`). During fine-tuning, we introduce a male-favoring salary effect in six occupations and a female-favoring effect in six others, while leaving six control occupations without a gender-conditioned effect. Independently, we train the model to score senior-experience cues above matched junior-experience cues. The pronoun-conditioned effect is the target of intervention, whereas seniority provides a separately constructed task signal that should be preserved.

We divide the $1{,}008$ counterfactual pairs from the 12 biased occupations into 696 train-set pairs for operator construction and forecasting and 312 validation-set pairs for selection. Validation additionally contains 504 pairs from the control occupations for measuring specificity. We reserve 648 independent test-set pairs across all 18 occupations for final evaluation and certification. Further details on task construction, including template filtering and the complete prompt format, appear in Appendix I. Table 12 reports held-out validation of the induced pronoun-conditioned effect.

Decision score. Although the model produces discrete `Yes`/`No` completions, we evaluate the normalized probability of `Yes`. At the final scoring representation z,

$$F(z) = P(T = \texttt{Yes} \mid T \in \{\texttt{Yes}, \texttt{No}\}, z) = \sigma(w^\top z + c),$$

where $w = W_{\texttt{Yes}} - W_{\texttt{No}}$ and $c = b_{\texttt{Yes}} - b_{\texttt{No}}$. The deployed score is therefore sigmoid-affine and is directly covered by the directional certificate in Section 3.

Task-specific operational requirements. A valid intervention must reduce the pronoun-conditioned score gap while preserving the planted seniority signal and decisions in the six control occupations. Here, utility is the fraction of matched senior–junior pairs for which $F_I(x_{\text{senior}}) > F_I(x_{\text{junior}})$. Specificity is the Spearman correlation between original and post-intervention scores within each occupation–pronoun group, averaged over the 12 control groups. A candidate is admissible on validation if its utility is at least 90% of the base value and specificity is at least 0.90.

4.2 Applying CAIC-S

Using the operational requirements above, we instantiate the four-stage protocol from Section 3.3 at the localized scoring site over the predeclared spectral-rank grid $r \in \{1, 2, 4, 8, 16, 64, 128, 256\}$, using full-strength attenuation $\alpha = 1$. Among admissible candidates, we select the one with the smallest mean validation gap. Otherwise, CAIC abstains.

Locate, forecast and select. At the answer token, we screen both normalization sites in five mid-network layers and the final normalization before unembedding. Attribution, corroborated by causal activation swaps at earlier sites, identifies the final normalized representation immediately before the model's output projection, which we hereby refer to as the *final scoring site*, as the dominant mediator of the score difference between matched counterfactual inputs. We therefore use this site for both intervention and certification. Appendix J provides the full localization analysis.

On the train set, we construct the candidate spectral operators and estimate their attainable residual sensitivity using the directional forecast from Section 3.3. Rank-8 CAIC-S has a forecast $\mathcal{F}_{\text{tr}} = 0.041$, attaining the smallest admissible validation gap, 0.016, while retaining utility of 1.000 and specificity of 0.914. It satisfies all operational requirements and the declared gap target. A post-selection analysis holding rank fixed at 8 while varying α shows that stronger attenuation steadily reduces the gap while utility and specificity remain admissible (Figure 6).

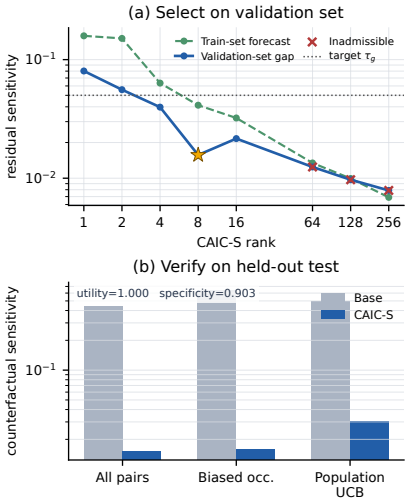

Figure 1: Controlled CAIC-S forecasting and verification. **(a)** The directional forecast identifies a low-rank intervention. **(b)** Held-out gaps and bounds decrease while utility is preserved through the selected operating region.

Verify. On the held-out test set, the selected operator reduces the mean counterfactual gap from 0.441 to 0.015, a 96.6% reduction, and the gap within biased occupations from 0.638 to 0.016. Utility remains 1.000 and specificity 0.903, preserving the seniority ordering and largely preserving decision rankings in control occupations. The mean directional certificate is 0.045, with a one-sided 95% population UCB of 0.063; the corresponding population UCB for the observed mean gap is 0.030. The broader Euclidean bound remains vacuous (Table 13), motivating the decision-aligned directional certificate. In Figure 1, increasing rank initially improves sensitivity reduction, while higher ranks violate the declared preservation requirements. CAIC-S suffices in this controlled setting: it substantially reduces the targeted counterfactual sensitivity and satisfies the operational requirements, obviating the need for CAIC-T.

5 Counterfactual Intervention in Salary Screening Task

We next evaluate CAIC on a semi-synthetic salary decision task using open-weight models without further finetuning. In contrast to §4.1, activation signals that are strongly associated with counterfactual directions may not clearly separate from task-related information, motivating the use of both CAIC-S and CAIC-T. We describe the task and the corresponding setup and offer an analysis of the demographic vs. task-relevant signals. We next evaluate the sensitivity-reduction and task-related preservation tradeoffs between the operators, and then compare the operators' performance against both black-box methods and white-box baselines to showcase the utility of the entire pipeline.

5.1 SALARY SCREENING TASK

Task and models. We use `Qwen2.5-7B-Instruct` (Yang et al., 2024) and `Mistral-7B-Instruct-v0.3` (Jiang et al., 2023) and `Ministral-8B` (Liu et al., 2026) with their public checkpoints. Given a target position, each model decides whether a minimum salary is appropriate based on a candidate's resume. We draw from from Nghiem et al. (2026)'s synthetic bank and take the candidate's first listed job as the target occupation and set the salary threshold to the U.S. national median. Sample prompt is shown in Figure 8.

Counterfactual family and decision score. We vary demographic information while holding the target occupation, salary threshold, and the base resume fixed. We consider Black vs. White-associated *race* cues and male vs. female-associated *gender* cues, spanning four counterfactual variants of a base resume. Demographic cues maybe transmitted implicitly via *name* (Nghiem et al., 2024) (e.g., "Brady Carlson" for White versus "Deshawn Alston" for Black male), or via *affiliation* (e.g. the National Society of Black Engineers) Kang et al. (2016) that is associated strongly with each racial-gender group, or by stacking both cues concordantly (referred to as *stacked* edits). As in Section 4, at the final scoring representation z we define $F(z) = \sigma(w^\top z + c)$, the probability assigned to `Yes` after renormalizing over the `Yes`/`No` answer tokens. The raw counterfactual gap is their absolute difference in F, which CAIC seeks to reduce.

5.2 INTERVENTION SETUP

Data splits, joint intervention and seeds. We partition the $1{,}073$ unique resumes by candidate into 500 train, 73 validation, and 500 test resumes, and freeze this assignment across five seeds. Each seed then determines the 200-resume fitting subset within the train set and the names, affiliations, and prompt templates used to instantiate matched counterfactuals; all validation and test resumes and the selection rules remain fixed. CAIC-S uses all 200 fitting resumes to fit its spectral basis, while CAIC-T splits them into 100 operator-fitting and 100 disjoint affine-lens-fitting resumes. Both methods construct a joint operator from four single-cue counterfactual displacement banks (race/name, race/affiliation, gender/name, and gender/affiliation) and orthonormalize their union. Stacked edits are excluded from fitting and used only to evaluate transfer to combined cues.

Task-specific operational requirements. As in § 4, admissible interventions must reduce the counterfactual gap while preserving utility and specificity. In this salary task, utility measures agreement with an external qualification ruler (via LLM judges) built from name-anonymized resume comparisons, and specificity measures preservation of score ordering on resumes rendered with "the applicant" and no demographic affiliation cue. A smaller raw counterfactual gap can reflect genuine counterfactual stability, but it can also arise if the intervention compresses or shifts the score distribution. We report three diagnostics for this failure mode: the neutral-CDF quantile gap asks whether counterfactual differences remain after converting each score to its percentile among neutral-resume scores from the same method; the score-dispersion retention checks whether the intervention collapses the spread of scores; the fixed-threshold stability measures how often the intervention changes a decision when the original score cutoff is kept fixed.

Per our protocol, *Locate* constructs counterfactual displacement banks at every fourth layer and, for CAIC-T, fits an affine lens on a disjoint training subset. *Forecast* computes train-set directional and preservation diagnostics. *Select* chooses the lowest-validation-gap configuration satisfying the operational requirements, treating CAIC-T's onsite lens and native final output as independently admitted endpoints. *Verify* freezes the intervention and score before reporting held-out sensitivity, confidence bounds, and operational diagnostics. Validation requires utility ≥ 0.90 of its matched baseline and specificity ≥ 0.90; CAIC-T additionally requires lens fidelity, counterfactual-to-task selectivity, score stability, and minimum raw- and quantile-gap reductions (Appendix K).

Pre-intervention counterfactual gap. At the original end-of-model scoring site (as defined in §4.2), Qwen exhibits the largest counterfactual gaps (0.25–0.28 on stacked race and gender edits), whereas Mistral's are smaller (≈ 0.07). Ministral-8B's corresponding gaps are only 0.020–0.021, near its 0.011 irrelevant-edit floor (Table 16), so CAIC abstains rather than intervenes. We therefore carry Qwen and Mistral forward to the operator and baseline comparisons.

Model	Axis	Method	CAIC-T admitted	Final-output gap pre→post (red.)	Onsite-lens gap pre→post (red.)	Quantile-gap red. (%) ↑	Task-signal removal (%) ↓	SD retained (%) ↑	Neutral threshold flips (%) ↓
Qwen	Race	CAIC-S	5/5	**.248→.079 (68.2%)**	.248→.083 (66.5%)	10.0	72.0	49.4	29.7
		CAIC-T		.248→.245 (1.4%)	**.246→.112 (54.6%)**	24.5	23.9	88.9	10.2
	Gender	CAIC-S	3/5	**.269→.079 (70.6%)**	.269→.095 (64.8%)	15.8	69.5	34.1	16.9
		CAIC-T		.269→.268 (0.2%)	**.267→.069 (74.2%)**	44.0	24.2	82.3	10.3
Mistral	Race	CAIC-S	1/5	**.073→.008 (88.6%)**	.069→.071 (-3.4%)	0.8	42.4	16.8	20.8
		CAIC-T		.073→.071 (1.9%)	**.072→.049 (32.1%)**	18.8	35.2	86.7	10.8
	Gender	CAIC-S	3/5	**.071→.018 (74.8%)**	.072→.046 (37.6%)	6.0	53.7	28.7	20.3
		CAIC-T		.071→.070 (0.9%)	**.070→.041 (42.3%)**	24.8	33.4	83.7	8.3

Table 1: Held-out means over CAIC-T-admitted seeds, with CAIC-S restricted to the same seeds. Gaps are measured before and after intervention at the model's final output and through a frozen affine lens at each operator's intervention site. Bold marks each operator's deployed scoring function, where the four rightmost diagnostics are computed; gray marks the alternative-score diagnostic. All 24 deployed-endpoint, *per-seed pre-to-post paired* 95% *confidence intervals exclude zero* (Appendix F.1); CAIC-T's onsite effect shows little persistence at the final output (§5.4).

5.3 Separability of demographic and task-relevant signal

We first ask whether intervening at decision-readable sites helps CAIC-S preserve task-relevant variation. At each candidate site, we fit an affine lens and fixed-rank CAIC-S on disjoint training subsets, then use validation data to measure lens agreement and counterfactual versus qualification-related decision-margin removal. In a single seed diagnostic, late sites are highly decision-readable but not selective (Table 14): CAIC-S removes 88.9%–92.6% of counterfactual and 71.2%–85.0% of task-relevant variation on Qwen, with corresponding ranges of 82.3%–88.4% and 81.6%–87.5% on Mistral. Proposition 2 (Appendix C) explains this distinction: an affine ridge lens weights activation directions by their variance and alignment with the downstream margin, with no term separating counterfactual from task-relevant variation. Lens fidelity therefore does not establish selective removability, motivating both CAIC-T's sensitivity vs. preservation analysis at different scoring sites.

5.4 The CAIC-S/CAIC-T sensitivity-reduction–preservation tradeoff

We evaluate each operator on held-out data with the selected intervention. For CAIC-S, the *final-output score* is obtained after the edited activation passes through the model's remaining layers to produce $P(\texttt{Yes})$. For admitted CAIC-T, the *onsite affine-lens score* is obtained by applying its frozen affine lens directly to the edited representation at the selected intervention site. Table 1 reports each operator under its deployed scoring function in bold and, as a diagnostic, under the alternative scoring function in gray. Reports here are averaged over the CAIC-T-admitted seed subsets only. Table 1 contrasts sensitivity reduction with preservation of task signal, neutral-score dispersion, and fixed-threshold decisions (§5.2). At their validation- selected scoring functions, CAIC-S achieves the larger raw-gap reduction in three of four settings, but removes 42.4%–72.0% of task signal and retains only 16.8%–49.4% of score dispersion. Its smaller quantile-gap reductions show that score-scale changes contribute to this effect. By contrast, CAIC-T removes 23.9%–35.2% of task signal, retains 82.3%–88.9% of dispersion, limits threshold flips to 8.3%–10.8%, and achieves greater quantile-gap reduction throughout.

Endpoint-specific intervention and abstention. Table 1) shows that behavior measured immediately after an activation edit need not persist after the remaining model layers. CAIC-T reduces its onsite gap by 32.1%–74.2%, but the same edits reduce the final-output gap by only 0.2%–1.9%. Consistently, validation admits no CAIC-T configuration scored at the final output in any of the 20 model–axis–seed conditions. CAIC-S changes little across the two scoring locations on Qwen, but its effect becomes stronger after downstream computation on Mistral: the gender reduction increases from 37.6% onsite to 74.8% at the final output over three seeds, while race changes from -3.4% to 88.6% for its single admitted seed. A sensitivity analysis (Appendix K.1) confirms that relaxing preservation requirements increases onsite admission from 12 to 18/20, whereas final-output admission remains 0/20 across all relaxation settings. Thus, final-output abstention is not a failure in optimization, but an expected outcome when no candidate satisfies the endpoint-specific deployment criteria. These results distinguish the operators: CAIC-S supports final-output mitigation with ranking preservation, whereas CAIC-T is better suited to early-exit use of the intervention-site score, with stronger preservation of task signal, score dispersion, and threshold stability.

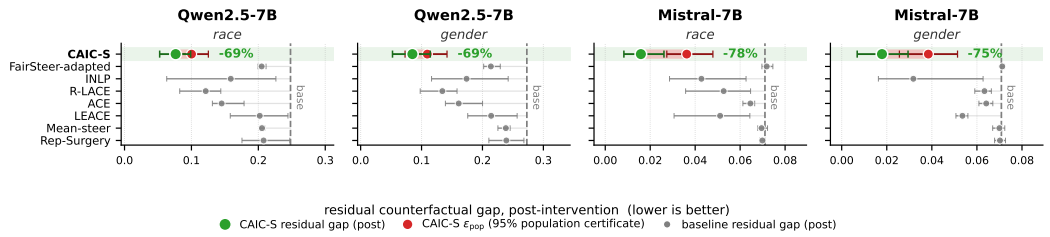

Figure 2: Matched-constraint comparison of full-strength CAIC-S with seven white-box intervention baselines at the final-output score. Dots show mean held-out counterfactual gaps, whiskers span five seeds, and red marks the CAIC-S 95% population UCB. Each method selects its own configuration on validation under common utility and specificity requirements.

5.5 Comparison with other intervention baselines

White- and black-box baseline comparisons. Because CAIC-T uses an onsite affine-lens score, the matched white-box comparison evaluates full-strength CAIC-S and seven baselines at their shared final-output score. Each baseline—INLP, R-LACE, ACE, LEACE, mean steering, representation surgery, and FairSteer-adapted—selects its own configuration on validation under common operation requirements. CAIC-S achieves the lowest mean held-out raw gap in all four model–axis settings, reducing base gaps by 69%–78% (Figure 2), and remains lowest under the own-neutral-CDF quantile metric (Table 7), so its advantage is not explained solely by score rescaling. CAIC-S is markedly more efficient than iterative erasure baselines while achieving lower residual gaps; cheaper closed-form methods leave larger gaps (Appendix O). All 28 paired CAIC-S–baseline comparisons have positive lower confidence limits and remain significant after Holm correction (Appendix C).

The first black-box baseline, score averaging evaluates all eight demographic renderings of each resume and assigns their mean score to every variant, guaranteeing zero gap within the enumerated family only, but requiring eight forward passes, and does not protect against unenumerated cues. In addition, we apply four anti-bias prompts by Karvonen & Marks (2025) to mitigate counterfactual gaps. Qwen rejects prompting in $8/10$ conditions on validation, whereas Mistral admits it in all 10 but preserves test-set utility in only $1/10$. Overall, prompting satisfies both held-out utility and specificity requirements in only $1/12$ admitted conditions and leaves larger conditional gaps than CAIC-S. See Appendices F, G, and O for matched-baseline, black-box, and efficiency details.

Sensitivity reduction is distinct from concept erasure. On Qwen, CAIC-S substantially reduces the final-output counterfactual gap while held-out demographic-cue decodability remains nearly unchanged (Figure 11). Conversely, INLP and LEACE can satisfy their fit-set erasure objectives while held-out decodability and counterfactual gaps persist; applying ACE or LEACE at all candidate layers likewise either leaves larger residual gaps or violates the operational requirements (Table 8). Thus, removing recoverable demographic information is neither necessary nor sufficient for reducing counterfactual sensitivity of a specified decision.

6 Discussion

We have introduce CAIC, a white-box framework for reducing and bounding decision-level counterfactual sensitivity under explicit operational constraints. CAIC-S provides aggressive spectral mitigation, while CAIC-T offers a task-protected alternative within a *Locate–Forecast–Select–Verify* pipeline that selects admissible interventions, abstains otherwise, and bounds residual sensitivity. We further distinguish counterfactual mitigation from generic concept erasure: representation-level removal need not reduce the audited decision gap, while aggressive removal can compromise task-relevant signal or score structure (Elazar et al., 2021; Kumar et al., 2022; Chowdhury et al., 2025). CAIC requires white-box access and its bounds cover only the declared counterfactual family and operational constraints, with broader scoring settings and a three-way hiring example discussed in Appendices E and E.1. CAIC links measured sensitivity to targeted intervention, preservation requirements, and quantitative residual bounds for responsible-AI auditing (Appendix A).

ETHICS STATEMENT

This work studies counterfactual sensitivity to demographic cues in LLM-based resume-screening-style decisions. The evaluated resumes are synthetic, and the study does not recruit human participants or collect private personal data. The demographic categories and proxy cues used in our counterfactual templates are necessarily simplified and should not be interpreted as exhaustive, essential, or definitive representations of race, gender, disability, or professional identity. CAIC certifies sensitivity only over a declared input-level counterfactual family. It does not establish legal compliance, causal counterfactual fairness, absence of disparate impact, or end-to-end fairness of a hiring system. The method requires white-box access and may alter model behavior outside the measured task. We therefore frame it as a technical mitigation and auditing component rather than a substitute for domain expertise, human oversight, and broader fairness evaluation.

REPRODUCIBILITY STATEMENT

Sections 3–5 specify the operators, assumptions, data partitions, admission requirements, and evaluation procedures. Appendices C and D provide complete proofs, while Appendices I, K, F, and O document the experimental construction, baseline configurations, statistical analyses, and compute environment. Our anonymous supplement[1] provides the model-independent CAIC implementation, prepared-activation schema, sanitized data underlying the principal tables, deterministic emitters and notebook, integrity manifests, and tests. Results may differ with different hardware infrastructure, seed choice, and models.

AI USE DISCLOSURE

We used generative-AI assistants to help develop and critique the conceptual framework, formulate and check mathematical claims and proofs, design experiments, implement and debug code, interpret results, create and revise figures, and draft and edit the manuscript. Generative models also supported data preparation and evaluation: `gpt-4o-mini-2024-07-18` assisted in converting biographies into name-free, pronoun-slotted templates and polishing controlled disclosure renderings, subject to deterministic validation and fallback checks. External LLM judges produced pairwise qualification comparisons for the resume-screening utility ruler; the final ruler averages `claude-sonnet-4-5` and `deepseek-chat`, with `gpt-5.4-mini` and `gemini-2.5-pro` included in cross-judge validation.

The authors reviewed and revised all AI-assisted outputs. Mathematical arguments were checked against their stated assumptions, generated or modified code was tested, and quantitative claims were traced to retained experimental artifacts and reproducible emitters. We did not use generative AI for translation or qualitative or thematic analysis; survey, interview, and transcription tasks were not applicable. The authors take responsibility for the paper's final text, claims, code, and artifacts.

REFERENCES

Haozhe An, Christabel Acquaye, Colin Wang, Zongxia Li, and Rachel Rudinger. Do large language models discriminate in hiring decisions on the basis of race, ethnicity, and gender? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 386–397, 2024.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023a.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. Leace: Perfect linear concept erasure in closed form. *Advances in Neural Information Processing Systems*, 36:66044–66063, 2023b.

[1] Anonymous code

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pp. 491–500, 2019.

Isha Chaudhary, Qian Hu, Manoj Kumar, Morteza Ziyadi, Rahul Gupta, and Gagandeep Singh. Certifying counterfactual bias in llms. In *International Conference on Learning Representations*, volume 2025, pp. 10494–10519, 2025.

Haodong Chen, Qiang Huang, Jiaqi Zhao, Qiuping Jiang, Xiaojun Chang, and Jun Yu. Measuring social bias in vision-language models with face-only counterfactuals from real photos. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 39963–39987, 2026.

Somnath Basu Roy Chowdhury, Avinava Dubey, Ahmad Beirami, Rahul Kidambi, Nicholas Monath, Amr Ahmed, and Snigdha Chaturvedi. Fundamental limits of perfect concept erasure. *arXiv preprint arXiv:2503.20098*, 2025.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 120–128, 2019.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226, 2012.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021.

European Parliament and Council of the European Union. Regulation (eu) 2024/1689 laying down harmonised rules on artificial intelligence (artificial intelligence act), 2024. URL `https://eur-lex.europa.eu/eli/reg/2024/1689/oj`. Article 50: Transparency Obligations for Providers and Deployers of Certain AI Systems.

Ky Fan. Maximum properties and inequalities for the eigenvalues of completely continuous operators. *Proceedings of the National Academy of Sciences*, 37(11):760–766, 1951.

Zijian Feng, Tianjiao Li, Zixiao Zhu, Hanzhang Zhou, Junlang Qian, Li Zhang, Chua Deryl, Lee Mak, Gee Ng, and Kezhi Mao. Fine-grained activation steering: Steering less, achieving more. In *International Conference on Learning Representations*, volume 2026, pp. 39421–39443, 2026.

Hila Gonen and Yoav Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 609–614, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1061. URL `https://aclanthology.org/N19-1061/`.

Navita Goyal and Hal Daumé III. Steering safely or off a cliff? rethinking specificity and robustness in inference-time interventions. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5723–5738, 2026.

Satchel Grant, Simon Jerome Han, Alexa Tartaglini, and Christopher Potts. Addressing divergent representations from causal interventions on neural networks. In *International Conference on Learning Representations*, volume 2026, pp. 11330–11356, 2026.

Umang Gupta, Jwala Dhamala, Varun Kumar, Apurv Verma, Yada Pruksachatkun, Satyapriya Krishna, Rahul Gupta, Kai-Wei Chang, Greg Ver Steeg, and Aram Galstyan. Mitigating gender bias in distilled language models via counterfactual role reversal. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 658–678, 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, 2021. URL https://arxiv.org/abs/2106.09685.

Yiran Hu, Zongyue Xue, Haitao Li, Siyuan Zheng, Qingjing Chen, Shaochun Wang, Xihan Zhang, Ning Zheng, Yun Liu, Qingyao Ai, et al. Llms on trial: Evaluating judicial fairness for large language models. In *International Conference on Learning Representations*, volume 2026, pp. 23780–23862, 2026.

Robert Huben, Hoagy Cunningham, Logan Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, volume 2024, pp. 7827–7845, 2024.

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

Sonia K Kang, Katherine A DeCelles, András Tilcsik, and Sora Jun. Whitened résumés: Race and self-presentation in the labor market. *Administrative science quarterly*, 61(3):469–502, 2016.

Adam Karvonen and Samuel Marks. Robustly improving llm fairness in realistic settings via interpretability. *arXiv preprint arXiv:2506.10922*, 2025.

Abhinav Kumar, Chenhao Tan, and Amit Sharma. Probing classifiers are unreliable for concept removal and detection. *Advances in Neural Information Processing Systems*, 35:17994–18008, 2022.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

Jinsook Lee, Conrad Borchers, AJ Alvero, Thorsten Joachims, and Rene F Kizilcec. The digital divide in generative ai: Evidence from large language model use in college admissions essays. In *Proceedings of the Thirteenth ACM Conference on Learning@ Scale*, pp. 179–190, 2026a.

Joseph Lee, Tianqi Shang, Jae Young Baik, Duy Duong-Tran, Shu Yang, Lingyao Li, and Li Shen. From promising capabilities to pervasive bias: Assessing large language models for emergency department triage. *Journal of Healthcare Informatics Research*, pp. 1–21, 2026b.

Yichen Li, Zhiting Fan, Ruizhe Chen, Xiaotang Gai, Luqi Gong, Yan Zhang, and Zuozhu Liu. Fairsteer: Inference time debiasing for llms with dynamic activation steering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 11293–11312, 2025.

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, Alexandre Sablayrolles, Amélie Héliou, Amos You, Andy Ehrenberg, Andy Lo, Anton Eliseev, Antonia Calvi, Avinash Sooriyarachchi, Baptiste Bout, Baptiste Rozière, Baudouin De Monicault, Clémence Lanfranchi, Corentin Barreau, Cyprien Courtot, Daniele Grattarola, Darius Dabert, Diego de las Casas, Elliot Chane-Sane, Faruk Ahmed, Gabrielle Berrada, Gaëtan Ecrepont, Gauthier Guinet, Georgii Novikov, Guillaume Kunsch, Guillaume Lample, Guillaume Martin, Gunshi Gupta, Jan Ludziejewski, Jason Rute, Joachim Studnia, Jonas Amar, Joséphine Delas, Josselin Somerville Roberts, Karmesh Yadav, Khyathi Chandu, Kush Jain, Laurence Aitchison, Laurent Fainsin, Léonard Blier, Lingxiao Zhao, Louis Martin, Lucile Saulnier, Luyu Gao, Maarten Buyl, Margaret Jennings, Marie Pellat, Mark Prins, Mathieu Poirée, Mathilde Guillaumin, Matthieu Dinot, Matthieu Futeral, Maxime Darrin, Maximilian Augustin, Mia Chiquier, Michel Schimpf, Nathan Grinsztajn, Neha Gupta, Nikhil Raghuraman, Olivier Bousquet, Olivier Duchenne, Patricia Wang, Patrick von Platen, Paul Jacob, Paul Wambergue, Paula Kurylowicz, Pavankumar Reddy Muddireddy, Philomène Chagniot, Pierre Stock, Pravesh Agrawal, Quentin Torroba, Romain Sauvestre, Roman Soletskyi, Rupert Menneer, Sagar Vaze, Samuel Barry, Sanchit Gandhi, Siddhant Waghjale, Siddharth Gandhi, Soham Ghosh, Srijan Mishra, Sumukh Aithal, Szymon Antoniak, Teven Le Scao, Théo Cachet, Theo Simon Sorg, Thibaut Lavril, Thiziri Nait

Saada, Thomas Chabal, Thomas Foubert, Thomas Robert, Thomas Wang, Tim Lawson, Tom Bewley, Tom Bewley, Tom Edwards, Umar Jamil, Umberto Tomasini, Valeriia Nemychnikova, Van Phung, Vincent Maladière, Virgile Richard, Wassim Bouaziz, Wen-Ding Li, William Marshall, Xinghui Li, Xinyu Yang, Yassine El Ouahidi, Yihan Wang, Yunhao Tang, and Zaccharie Ramzi. Ministral 3, 2026. URL https://arxiv.org/abs/2601.08584.

Thomas Marshall, Adam Scherlis, and Nora Belrose. Refusal in llms is an affine function. *arXiv preprint arXiv:2411.09003*, 2024.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

New York City Council. Local law 144 of 2021: Automated employment decision tools, 2021. URL https://legistar.council.nyc.gov/LegislationDetail.aspx?ID=4344524&GUID=B051915D-A9AC-451E-81F8-6596032FA3F9. Int. No. 1894-A, enacted December 11, 2021.

Huy Nghiem, John Prindle, Jieyu Zhao, and Hal Daumé Iii. "you gotta be a doctor, lin": An investigation of name-based bias of large language models in employment recommendations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7268–7287, 2024.

Huy Nghiem, Phuong-Anh Nguyen-Le, John Prindle, Rachel Rudinger, and Hal Daumé III. 'rich dad, poor lad': How do large language models contextualize socioeconomic factors in college admission? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 21033–21067, 2025.

Huy Nghiem, Phuong-Anh Nguyen-Le, Sy-Tuyen Ho, and Hal Daume III. Bias in the tails: How name-conditioned evaluative framing in resume summaries destabilizes llm-based hiring. *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*, 2026.

Naoki Otani, Nikita Bhutani, and Estevam Hruschka. Natural language processing for human resources: A survey. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pp. 583–597, 2025.

Jaeil Park and Sung-Bae Cho. Multi-feature quantized self-attention for fair large language models. In *The Fourteenth International Conference on Learning Representations*, 2026.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 7237–7256, 2020.

Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan D Cotterell. Linear adversarial concept erasure. In *International Conference on Machine Learning*, pp. 18400–18421. PMLR, 2022.

Antoine Saillenfest and Pirmin Lemberger. Nonlinear concept erasure: a density matching approach. *arXiv preprint arXiv:2507.12341*, 2025.

Zhengyang Shan and Aaron Mueller. Measuring mechanistic independence: Can bias be removed without erasing demographics? In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4241–4265, 2026.

Aarushi Sharma and Phong Le. Encoding without influence: Dissociating demographic representation from causal effect in large language models. *Transactions on Machine Learning Research*, 2026.

Yongjie Sun, Angelo Zappalà, Eleonora Di Maso, Francesco Pompedda, Thomas J Nyman, and Pekka Santtila. Large language models (llms) as jurors: Assessing the potential of llms in legal contexts. *Law and Human Behavior*, 49(5):440, 2025.

Alex Tamkin, Amanda Askell, Liane Lovitt, Esin Durmus, Nicholas Joseph, Shauna Kravec, Karina Nguyen, Jared Kaplan, and Deep Ganguli. Evaluating and mitigating discrimination in language model decisions. *arXiv preprint arXiv:2312.03689*, 2023.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, et al. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *arXiv preprint arXiv:2605.29358*, 2026.

Yisong Xiao, Xianglong Liu, QianJia Cheng, Zhenfei Yin, Siyuan Liang, Jiapeng Li, Jing Shao, Aishan Liu, and Dacheng Tao. Genderbias-vl: Benchmarking gender bias in vision language models via counterfactual probing. *International Journal of Computer Vision*, 133(12):8332–8355, 2025.

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, 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, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report, 2024.

Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. *arXiv preprint arXiv:2205.15480*, 2022.

Dominik Zietlow, Michael Lohaus, Guha Balakrishnan, Matthäus Kleindessner, Francesco Locatello, Bernhard Schölkopf, and Chris Russell. Leveling down in computer vision: Pareto inefficiencies in fair deep classifiers. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10400–10411. IEEE, 2022.

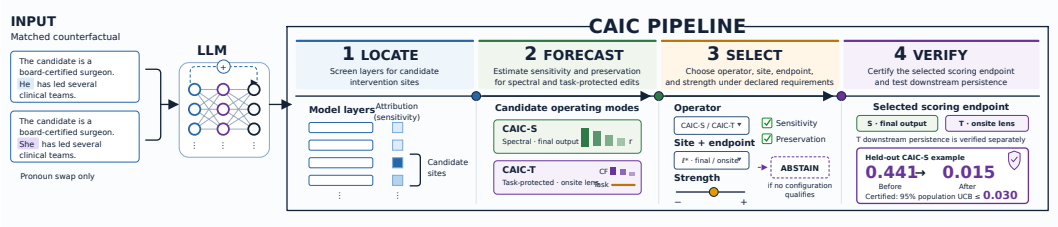

Figure 3: CAIC locates candidate intervention sites, forecasts residual counterfactual sensitivity, selects an admissible intervention under task-specific operational constraints, and verifies the fixed choice on held-out data. It abstains when no candidate satisfies the declared requirements.

A Social and Governance Application

We discuss CAIC's positioning among related fairness notions, and how our method can address certain imperatives by policies on responsible AI deployment.

A.1 Fairness Literature

Individual and counterfactual fairness. CAIC's objective is closely related to individual and counterfactual fairness. Individual fairness requires similarly situated individuals to receive similar treatment (Dwork et al., 2012). CAIC instantiates a restricted version by holding task-relevant information fixed, varying a designated demographic cue, and bounding the resulting decision change. Counterfactual fairness (Kusner et al., 2017) similarly asks whether decisions remain invariant under changes to protected attributes, but is defined through interventions in a structural causal model. CAIC instead certifies stability over an explicit input-level counterfactual family, rather than an arbitrary similarity metric or all causal counterfactual worlds.

Finally, CAIC illustrates why *fairness through unawareness* (removing an explicit protected attribute from the input) is generally insufficient. Demographic information can remain available through correlated cues or proxies, and removing demographic information from a representation need not remove its influence on a particular decision. In our resume experiments, anonymizing names leaves affiliation-based pathways intact, while concept-erasure methods can substantially reduce demographic decodability without eliminating the audited decision gap. CAIC instead targets the component of counterfactual variation that affects the specified decision.

A.2 Application to AI Policy-Oriented Auditing

High-stakes decision systems. CAIC addresses this remediation gap by reducing specified counterfactual decision-score sensitivity under measured utility and specificity constraints, then certifying the residual sensitivity. Selection can additionally constrain neutral decision changes, with abstention when mitigation and preservation requirements conflict. Coupled with application-specific decision-level validation, these score-level guarantees could support targeted mitigation alongside group-level disparate-impact analysis.

Auditable risk mitigation. Broader AI governance frameworks increasingly require high-risk systems to be assessed, documented, and monitored. The EU AI Act (European Parliament and Council of the European Union, 2024), for example, places transparency and risk-management requirements on high-risk employment AI, while recent U.S. laws impose similar governance obligations for automated consequential decisions. CAIC can provide a technical artifact for such workflows: the counterfactual family specifies the behavior being controlled, utility and specificity define acceptable intervention costs, abstention records when no suitable mitigation exists, and the post-intervention certificate quantifies residual sensitivity. These guarantees do not establish regulatory compliance, but they make mitigation more explicit, testable, and auditable.

B Additional Related Work

Broader erasure and multi-attribute mitigation. Beyond linear erasure methods, LEOPARD targets nonlinear protected-attribute information through class-conditional density matching (Saillenfest & Lemberger, 2025), while recent multi-attribute methods address overlapping protected features jointly (Park & Cho, 2026). These approaches pursue broader representation-level removal; CAIC instead targets counterfactual variation relevant to a specified decision and certifies the residual decision sensitivity.

Counterfactual evaluation and intervention validity. Controlled counterfactual evaluation has been used across text and multimodal settings to isolate demographic effects (Xiao et al., 2025; Chen et al., 2026), and training-time methods mitigate bias through counterfactual augmentation or role reversal (Gupta et al., 2022). CAIC instead operates post hoc on pretrained models. Representation interventions can also induce off-manifold hidden states (Grant et al., 2026); CAIC controls behavioral side effects through utility and specificity checks, though explicit latent-distribution diagnostics remain an important extension.

C Operator structure and forecasting

At a fixed intervention site, let the rows of $D \in \mathbb{R}^{N\times d}$ be the train-set counterfactual displacements $d_i^\top$, where $d_i = z(x_i) - z(x_i')$. Let

$$D = U\Sigma V^\top$$

be the compact singular value decomposition of rank p, with singular values $\sigma_1 \geq \cdots \geq \sigma_p > 0$ and right singular vectors $v_1, \ldots, v_p$. For an orthonormal basis $Q \in \mathbb{R}^{d\times q}$ and a symmetric attenuation matrix $A \in \mathbb{R}^{q\times q}$ satisfying $0 \preceq A \preceq I_q$, recall the shared intervention family

$$M_{Q,A} = I_d - QAQ^\top.$$

The scalar-strength operator is the special case $P_{Q,\alpha} \equiv M_{Q,\alpha I_q} = I_d - \alpha QQ^\top$, with $\alpha \in [0,1]$.

To see that every member of this family is non-expansive, extend Q to an orthonormal basis of $\mathbb{R}^d$ and diagonalize A. In the targeted subspace, the eigenvalues of $M_{Q,A}$ are $1 - a_j \in [0,1]$; on its orthogonal complement they are one. Hence $\|M_{Q,A}\|_2 \leq 1$. This property prevents the intervention from amplifying Euclidean activation differences, although it does not by itself guarantee preservation of downstream task behavior. We do not mean-center D, because its mean is part of the counterfactual variation that the intervention is intended to remove. When multiple counterfactual channels are fitted, each channel produces a basis and CAIC orthonormalizes their union. The results below apply to each channel separately; they do not assert that the resulting union is globally optimal under a fixed total-rank constraint.

Proposition 1 (CAIC-S residual-energy optimality). For $1 \leq r \leq p$, let $V_r = [v_1, \ldots, v_r]$. Among all rank-r full-strength residualization operators, $P_{V_r,1} = I_d - V_rV_r^\top$ minimizes the mean squared residual displacement:

$$\min_{Q^\top Q = I_r} \frac{1}{N}\|DP_{Q,1}\|_F^2 = \frac{1}{N}\|DP_{V_r,1}\|_F^2 = \frac{1}{N}\sum_{j>r}\sigma_j^2.$$

Proof. For any $Q \in \mathbb{R}^{d\times r}$ with $Q^\top Q = I_r$, $QQ^\top$ and $P_{Q,1}$ project onto orthogonal subspaces. Hence

$$\|DP_{Q,1}\|_F^2 = \|D\|_F^2 - \|DQ\|_F^2.$$

The first term is independent of Q, so minimizing the residual is equivalent to maximizing

$$\|DQ\|_F^2 = \mathrm{tr}(Q^\top D^\top DQ).$$

The right singular vectors v_j are eigenvectors of $D^\top D$ with eigenvalues σ_j^2. By the Ky Fan maximum principle (Fan, 1951),

$$\max_{Q^\top Q = I_r} \mathrm{tr}(Q^\top D^\top DQ) = \sum_{j=1}^{r}\sigma_j^2,$$

and the maximum is attained by $Q = V_r$. Since $\|D\|_F^2 = \sum_{j=1}^{p} \sigma_j^2$, the residual at this basis is $\sum_{j>r} \sigma_j^2$. Dividing by N proves the claim. □

This proposition concerns the displacement geometry alone. It explains the spectral construction of CAIC-S, but does not claim that the selected subspace minimizes the counterfactual decision gap or preserves utility and specificity.

Affine lens for CAIC-T. At candidate site L, let $Z_L \in \mathbb{R}^{n_L \times d}$ contain activations from the lens-fitting split and let $y_L \in \mathbb{R}^{n_L}$ contain the corresponding unedited final decision margins. The affine lens is the ridge fit

$$(\beta_L, b_L) = \arg\min_{\beta, b} \left\{ \|Z_L\beta + b\mathbf{1} - y_L\|_2^2 + \lambda_L \|\beta\|_2^2 \right\}.$$

Equivalently, center Z_L and y_L, solve ridge regression for β_L, and set $b_L = \bar{y}_L - \bar{z}_L^\top \beta_L$. This is precisely the affine score whose fidelity is checked on the disjoint operator-fitting subset of the training data before an onsite policy can be selected. The lens-fitting examples are disjoint from those used to construct the intervention, preventing the fitted direction from being trained and judged on the same paired displacements.

The following standard ridge identity makes the limitation of lens fidelity explicit. Let

$$\widetilde{Z}_L = Z_L - \mathbf{1}\bar{z}_L^\top, \qquad \widetilde{y}_L = y_L - \bar{y}_L \mathbf{1}$$

denote the centered lens-fitting activations and margins.

Proposition 2 (Spectral form of the affine lens). Let $\widetilde{Z}_L = U\Sigma V^\top$ be the compact singular value decomposition, with singular values $\sigma_1 \geq \cdots \geq \sigma_p > 0$. The ridge coefficient satisfies

$$\hat{\beta}_L = \sum_{j=1}^{p} \frac{\sigma_j}{\sigma_j^2 + \lambda_L} \big(u_j^\top \widetilde{y}_L\big) v_j \in \operatorname{span}(V).$$

Thus, each activation direction is weighted by its variance scale, its alignment with the downstream margin, and the ridge shrinkage.

Proof. The centered normal equations are

$$(\widetilde{Z}_L^\top \widetilde{Z}_L + \lambda_L I)\hat{\beta}_L = \widetilde{Z}_L^\top \widetilde{y}_L.$$

Substituting $\widetilde{Z}_L^\top \widetilde{Z}_L = V\Sigma^2 V^\top$ and $\widetilde{Z}_L^\top \widetilde{y}_L = V\Sigma U^\top \widetilde{y}_L$ gives

$$\hat{\beta}_L = V(\Sigma^2 + \lambda_L I)^{-1} \Sigma U^\top \widetilde{y}_L,$$

which yields the stated expansion. □

Readability is not selectivity. The expression contains no term indicating whether v_j carries counterfactual variation, task-relevant variation, or both. High lens agreement therefore shows that the decision is readable at site L, but does not imply that lens-aligned variation can be attenuated while preserving task signal. Changing λ_L only reweights the same variance and margin-alignment terms and does not itself impose this distinction. This explains the readability–selectivity gap in Table 14 and motivates the explicit task bank T_L in CAIC-T.

For an admitted onsite CAIC-T policy, the deployed scoring direction is $w_I = \hat{\beta}_L$. Its directional certificate therefore bounds residual counterfactual sensitivity of the frozen onsite affine-lens score. Lens agreement separately measures fidelity to the unedited downstream decision, and the model's standard final output is evaluated and admitted independently.

Task-protected direction fit. Let $B_L \in \mathbb{R}^{d \times k}$ be an orthonormal basis for the fixed reduced span $\mathcal{S}_L$, assembled from leading directions of the counterfactual banks, the protected task bank, and the normalized lens direction. For channel c, define

$$X_c = \bar{D}_c B_L, \qquad H_L = T_L B_L, \qquad y_c = \bar{D}_c \beta_L,$$

where $\bar{D}_c = D_c / \|D_c\|_F$. Writing $u = B_L\theta$, the CAIC-T objective in Section 3.2 becomes

$$\min_{\theta \in \mathbb{R}^k} \left\{ \|X_c\theta - y_c\|_2^2 + \gamma \|H_L\theta\|_2^2 + \lambda \|\theta\|_2^2 \right\}.$$

For $\lambda > 0$, its unique solution is

$$\widehat{\theta}_c = \left(X_c^\top X_c + \gamma H_L^\top H_L + \lambda I_k\right)^{-1} X_c^\top y_c, \qquad \widehat{u}_c = B_L \widehat{\theta}_c.$$

Indeed, differentiating gives the displayed normal equations, and the coefficient matrix is positive definite because its quadratic form is $\|X_c a\|_2^2 + \gamma \|H_L a\|_2^2 + \lambda \|a\|_2^2 > 0$ for every nonzero a. Thus the first term fits the component of the channel displacement that contributes to the lens margin, whereas the second explicitly penalizes activity on the declared task contrasts. After normalizing each nonzero $\widehat{u}_c$, CAIC-T takes a rank-revealing orthonormal union, denoted $\widetilde{Q}_L$.

Task ordering and anisotropic attenuation. Let

$$G_L = \widetilde{Q}_L^\top T_L^\top T_L \widetilde{Q}_L = R_L \Lambda_L R_L^\top, \qquad Q_{\mathrm{T},L} = \widetilde{Q}_L R_L,$$

where R_L is orthogonal and the diagonal entries $\ell_1 \le \cdots \le \ell_q$ of Λ_L are ordered increasingly. Because multiplication by R_L is only a rotation, $\mathrm{span}(Q_{\mathrm{T},L}) = \mathrm{span}(\widetilde{Q}_L)$. Moreover,

$$Q_{\mathrm{T},L}^\top T_L^\top T_L Q_{\mathrm{T},L} = \Lambda_L, \qquad \|T_L q_j\|_2^2 = \ell_j.$$

The rotated coordinates therefore have mutually orthogonal task responses and are ordered by measured task activity. Choosing $A_{\mathrm{T}} = \mathrm{diag}(a_1, \ldots, a_q)$ can attenuate task-light coordinates more strongly while remaining inside the shared non-expansive family. This ordering is a construction device, not a utility guarantee: validation still determines whether a proposed schedule preserves the required behavior.

Proposition 3 (Common training forecast). Suppose the intervention and scoring sites coincide and the decision score there is $F(z) = \sigma(w^\top z + c)$. For any member of the shared operator family, define

$$\mathcal{F}_{\mathrm{tr}}(D, Q, A; w) = \frac{1}{4}\sqrt{\frac{1}{N}\|D M_{Q,A} w\|_2^2}.$$

Then

$$\mathcal{F}_{\mathrm{tr}}(D, Q, A; w) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{4}\left|w^\top M_{Q,A} d_i\right|\right)^2},$$

the train-set root-mean-square of the unclipped directional bound. For the scalar-strength spectral operator, $Q = V_r$ and $A = \alpha I_r$,

$$\mathcal{F}_{\mathrm{tr}}(D, V_r, \alpha I_r; w)^2 = \frac{1}{16}\left[\sum_{j \le r}\frac{\sigma_j^2}{N}(1-\alpha)^2 (v_j^\top w)^2 + \sum_{j > r}\frac{\sigma_j^2}{N}(v_j^\top w)^2\right].$$

Proof. Because $M_{Q,A}$ is symmetric, the i-th entry of $D M_{Q,A} w$ is

$$d_i^\top M_{Q,A} w = w^\top M_{Q,A} d_i.$$

Squaring these entries, summing them, and multiplying by $1/(16N)$ proves the first identity.

For the spectral special case, write $a_j = v_j^\top w$. Components of w orthogonal to $\mathrm{span}(V)$ are annihilated by D. Within $\mathrm{span}(V)$, scalar partial residualization multiplies the first r components by $1 - \alpha$ and leaves the others unchanged. Hence

$$D M_{V_r, \alpha I_r} w = U \Sigma \left[(1-\alpha)a_1 \quad \cdots \quad (1-\alpha)a_r \quad a_{r+1} \quad \cdots \quad a_p\right]^\top.$$

Orthonormality of the columns of U gives the displayed sum after multiplication by $1/(16N)$. □

At full strength, $\alpha = 1$, the first sum vanishes and the expression reduces to the CAIC-S tail-spectrum forecast. The general identity also applies to the anisotropic CAIC-T operator and to channelwise unions. For onsite CAIC-T, the scoring direction is $w = \beta_L$. In every case it is an exact identity for the fitting-set RMS directional bound under the stated coincident-site assumption; it is not an observed validation gap, a utility or specificity guarantee, or a held-out certificate.

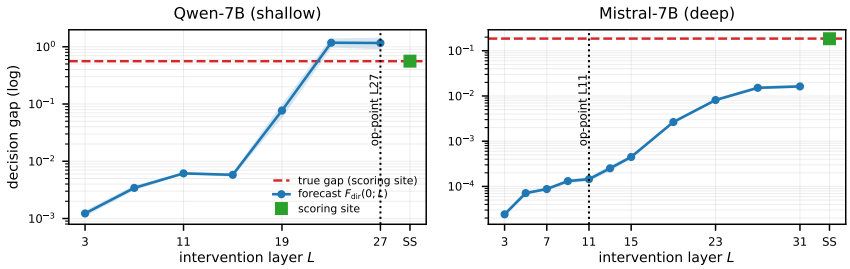

Figure 4: Forecast reliability across candidate sites. The forecast $\mathcal{F}_{\mathrm{tr}}(D, Q, \alpha I; w)$ (blue) closely tracks the measured gap (dashed) empirically at the scoring site (green square), but not at upstream sites, where the remaining network transforms the intervention. The dotted line marks the deployed CAIC-S configuration.

Selective-removal diagnostics. Let $\delta_{\mathrm{cf}}^{\mathrm{base}}$ and $\delta_{\mathrm{cf}}^{\mathrm{post}}$ collect the pre-sigmoid decision-margin differences between counterfactual pairs before and after intervention. For each task-relevant comparison k, define $\delta_{\mathrm{task},k}^{\mathrm{base}}$ and $\delta_{\mathrm{task},k}^{\mathrm{post}}$ analogously. When the respective base energies are nonzero, the removal fractions in Section 3.3 are

$$\rho_{\mathrm{cf}} = 1 - \frac{\|\delta_{\mathrm{cf}}^{\mathrm{post}}\|_2^2}{\|\delta_{\mathrm{cf}}^{\mathrm{base}}\|_2^2}, \qquad \rho_{\mathrm{task},k} = 1 - \frac{\|\delta_{\mathrm{task},k}^{\mathrm{post}}\|_2^2}{\|\delta_{\mathrm{task},k}^{\mathrm{base}}\|_2^2}.$$

When intervention and scoring coincide, these differences satisfy $\delta_{\mathrm{cf}}^{\mathrm{base}} = Dw$, $\delta_{\mathrm{cf}}^{\mathrm{post}} = DM_{Q,A}w$, $\delta_{\mathrm{task},k}^{\mathrm{base}} = T_{L,k}w$, and $\delta_{\mathrm{task},k}^{\mathrm{post}} = T_{L,k}M_{Q,A}w$. For downstream scoring, they are the corresponding measured final decision-margin differences. Thus ρ_{cf} is the fractional reduction in squared counterfactual margin displacement, whereas $\rho_{\mathrm{task},k}$ is the analogous reduction for task-relevant comparison k. They compare decision-relevant variation rather than total activation energy. A negative value indicates amplification for that comparison; zero denominators are treated as undefined rather than as evidence of removal. These are repair–preservation diagnostics, not certificates. We compute them on fitting data to characterize candidates, on validation data for admission, and on test data for reporting, as described in Section 3.3.

Forecast scope across sites. The common forecast is exact only when the intervention and scoring sites coincide and the score is sigmoid-linear at that representation. At an upstream intervention site, the remaining network can transform the edited activation nonlinearly, so $DM_{Q,A}w$ need not equal the resulting decision change. Figure 4 illustrates this distinction for the scalar-strength spectral operator in the models of Section 5: its forecast tracks the measured counterfactual gap at the scoring site but not at upstream candidate sites. This figure is a site-forecast diagnostic for CAIC-S, not a validation of the CAIC-T construction. Site and operator selection therefore use validation measurements rather than treating the fitting-set forecast as a held-out guarantee.

D PER-INPUT CERTIFICATE DERIVATIONS

This appendix derives the three certificates in Table 2 from Theorem 1. Unlike the operator results in Appendix C, each derivation fixes a certificate geometry and its Lipschitz constant, then applies Theorem 1 to the declared counterfactual family $\mathcal{D}(x)$.

The notation (w_I, c_I) is endpoint-specific. At the final-output endpoint it denotes the model's final scoring direction and intercept. For an onsite CAIC-T policy it denotes the independently fitted affine lens (β_L, b_L). The certificate then applies to that deployed onsite score; held-out lens agreement is a separate fidelity measurement, not part of the certificate.

Smoothing certificate $(\epsilon_{\mathrm{smooth}})$**.** At an arbitrary certificate site, the remaining computation from a representation to the final score may be too complex to analyze directly. Let $f_I(z) \in [a, b]$ be this bounded score, and define its Gaussian-smoothed version as

$$\bar{f}_I(z) = \mathbb{E}_{\xi \sim \mathcal{N}(0, \sigma^2 I)}[f_I(z + \xi)].$$

Certificate	Assumption on F_I	Per-input bound $\epsilon(x)$
ϵ_{smooth}	bounded $f_I \in [a, b]$	$\frac{b-a}{\sigma\sqrt{2\pi}} \sup_{x'} \|z_I(x) - z_I(x')\|_2$
ϵ_{euc}	$\sigma(w_I^\top z + c_I)$	$\frac{\|w_I\|_2}{4} \sup_{x'} \|z_I(x) - z_I(x')\|_2$
ϵ_{dir}	$\sigma(w_I^\top z + c_I)$	$\frac{1}{4} \sup_{x'} \|w_I^\top (z_I(x) - z_I(x'))\|$

Table 2: Three per-input instances of Theorem 1. Each supremum is over $x' \in \mathcal{D}(x)$. The first row certifies the smoothed score $\bar{f}_I$; the remaining rows certify the stated sigmoid-affine score.

For two representations z, z', boundedness gives

$$|\bar{f}_I(z) - \bar{f}_I(z')| \le (b-a)\,\mathrm{TV}\big(\mathcal{N}(z, \sigma^2 I), \mathcal{N}(z', \sigma^2 I)\big)\,,$$

where TV is total-variation distance. If Φ denotes the standard normal cumulative distribution function, the distance between these equal-covariance Gaussians is

$$2\Phi\left(\frac{\|z - z'\|_2}{2\sigma}\right) - 1 \le \frac{\|z - z'\|_2}{\sigma\sqrt{2\pi}}.$$

Thus $\bar{f}_I$ is $(b-a)/(\sigma\sqrt{2\pi})$-Lipschitz in Euclidean distance. Applying Theorem 1 with $F_I = \bar{f}_I$ gives

$$\epsilon_{\text{smooth}}(x) = \frac{b-a}{\sigma\sqrt{2\pi}} \sup_{x' \in \mathcal{D}(x)} \|z_I(x) - z_I(x')\|_2.$$

This certificate requires no particular form for the remaining model, but it applies to the smoothed score $\bar{f}_I$, not to the model's original score f_I.

Euclidean certificate (ϵ_{euc})**.** Suppose the decision score at the certificate site is $F_I(z) = \sigma(w_I^\top z + c_I)$. Since the sigmoid derivative is at most $1/4$,

$$|F_I(z) - F_I(z')| \le \frac{1}{4}|w_I^\top (z - z')| \le \frac{\|w_I\|_2}{4}\|z - z'\|_2,$$

where the second inequality is Cauchy–Schwarz. Taking the certificate geometry to be Euclidean therefore gives

$$\epsilon_{\text{euc}}(x) = \frac{\|w_I\|_2}{4} \sup_{x' \in \mathcal{D}(x)} \|z_I(x) - z_I(x')\|_2.$$

Because the score is a probability, clipping this expression at one would remain valid. We leave it unclipped in Section 4 to show when the Euclidean certificate is vacuous.

Directional certificate (ϵ_{dir})**.** For the same sigmoid-linear score, retaining the first inequality above instead of applying Cauchy–Schwarz gives

$$\epsilon_{\text{dir}}(x) = \frac{1}{4} \sup_{x' \in \mathcal{D}(x)} |w_I^\top (z_I(x) - z_I(x'))|.$$

This certificate measures only the component of the post-intervention counterfactual displacement along the scoring direction. It is no larger than $\epsilon_{\text{euc}}(x)$, because $|w_I^\top d| \le \|w_I\|_2 \|d\|_2$ for every displacement d. Since $|F_I(z) - F_I(z')| \le 1$, clipping $\epsilon_{\text{dir}}(x)$ at one also preserves validity. We use this clipped form for the natural-model population summaries below.

Finite counterfactual families and population summaries. Suppose each independent held-out input x_i has J listed matched counterfactual pairs, with pair j written as (x_{ij}, x'_{ij}). Define the observed per-input mean gap as

$$g_i = \frac{1}{J} \sum_{j=1}^{J} \left|F_I(z_I(x_{ij})) - F_I(z_I(x'_{ij}))\right|.$$

Applying empirical Bernstein to $g_1, \ldots, g_n$ gives the reported raw-gap population UCB. For each pair, define the clipped directional bound

$$b_{ij} = \min\left\{1, \frac{1}{4}\left|w_I^\top \left(z_I(x_{ij}) - z_I(x'_{ij})\right)\right|\right\}.$$

We summarize these bounds within each held-out input in two ways:

$$\epsilon_{\text{dir,mean}}(x_i) = \frac{1}{J}\sum_{j=1}^{J} b_{ij}, \qquad \epsilon_{\text{dir,max}}(x_i) = \max_{1\le j\le J} b_{ij}.$$

The maximum certifies every listed pair for x_i. The mean instead describes average sensitivity when those listed pairs are weighted equally; it does not bound each pair individually. We report the averages of these per-input quantities across the n held-out inputs:

$$\epsilon_{\text{dir,mean}} = \frac{1}{n}\sum_{i=1}^{n} \epsilon_{\text{dir,mean}}(x_i), \qquad \epsilon_{\text{dir,max}} = \frac{1}{n}\sum_{i=1}^{n} \epsilon_{\text{dir,max}}(x_i).$$

Their population UCBs are computed separately across the n independent held-out inputs using the empirical-Bernstein formula in Section 3. The multiple pairs belonging to one input are first combined by the mean or maximum above and are not treated as independent population samples. In the controlled experiment, $J = 1$, so the mean and maximum directional summaries coincide.

Local-Lipschitz refinement. The global sigmoid constant $1/4$ is attained only at a pre-sigmoid margin of zero. Define $m_I(x) = w_I^\top z_I(x) + c_I$ and let

$$\bar{L}_I(x) = \sup_{x'\in\mathcal{D}(x)} \sup_{t \text{ between } m_I(x) \text{ and } m_I(x')} \sigma'(t).$$

The mean-value theorem applied separately to every $x' \in \mathcal{D}(x)$ then gives the valid refinement

$$\Delta_I(x) \le \bar{L}_I(x) \sup_{x'\in\mathcal{D}(x)} |w_I^\top (z_I(x) - z_I(x'))|, \qquad \bar{L}_I(x) \le \frac{1}{4}.$$

This can improve the directional certificate when all relevant margins lie in a saturated region of the sigmoid. If any relevant interval crosses zero, its maximum slope is $1/4$, and the refinement reduces to the global bound. We report the global-constant certificate throughout for a single consistent closed-form quantity.

E CERTIFICATES BEYOND BINARY DECISIONS

Our experiments certify a binary `Yes`/`No` score, but the closed-form directional certificate of Section 3 applies to any decision that is a fixed link applied to a *linear* score in the representation at the certification site. Here we record the per-input directional certificate for the main non-binary decision families. In each case let $d = z_{\text{marked}} - z_{\text{ref}}$ be the counterfactual displacement at the certification site; because the underlying logits are linear in z and followed by a fixed link, the mean-value theorem bounds the score gap by the displacement projected onto the scoring direction(s), so the certificate is always a directional quantity in d.

Linear score (regression, scalar reward). For a linear score $F(z) = w^\top z + c$—a predicted salary, a risk score, or a linear reward head—the counterfactual gap is exact,

$$|F(z_{\text{marked}}) - F(z_{\text{ref}})| = |w^\top d|,$$

with no sigmoid slack; this is the tightest case.

Binary logistic. For $F(z) = \sigma(w^\top z + c)$, the $\frac{1}{4}$-Lipschitz constant of σ gives $|\Delta F| \le \frac{1}{4}|w^\top d|$; the constant $\frac{1}{4}$ is the optimal global Lipschitz constant of the sigmoid, approached for small logit perturbations near the decision boundary. This is the case certified in Sections 4 and 5.

K-class softmax. Let $p_k(z) = \text{softmax}_k(Wz + b)$ be the probability of class k, with per-class scoring vectors w_j and projections $g_j \equiv w_j^\top d$. Along the segment $z(t) = z_{\text{ref}} + t\,d$,

$$\frac{dp_k}{dt} = p_k\Big(g_k - \sum_j p_j g_j\Big) = p_k \sum_{j\ne k} p_j\,(g_k - g_j),$$

so $|dp_k/dt| \le p_k(1-p_k)\,\max_{j\ne k}|g_k - g_j| \le \frac{1}{4}\max_{j\ne k}|g_k - g_j|$, and integrating over $t \in [0,1]$,

$$|p_k(z_{\text{marked}}) - p_k(z_{\text{ref}})| \;\le\; \tfrac{1}{4}\max_{j\ne k}\big|(w_k - w_j)^\top d\big|.$$

The bound is $\frac{1}{4}$ times the displacement projected onto the worst *class-contrast* direction $w_k - w_j$; it recovers the binary certificate when $K = 2$. The intervention target generalizes accordingly: remove the component of d in this contrast subspace rather than along a single scoring direction.

Categorical stability (flip-margin certificate). Beyond bounding $|\Delta p_k|$, we certify that the *argmax* recommendation itself does not change. Under the fixed intervention let $k^\star = \arg\max_k (w_k^\top z + b_k)$ at input x. The recommendation is certified stable over the evaluated counterfactual family $\hat{\mathcal{D}}(x)$ if every runner-up $j \neq k^\star$ satisfies

$$(w_{k^\star} - w_j)^\top z + (b_{k^\star} - b_j) \;>\; \max_{x' \in \hat{\mathcal{D}}(x)} \left|(w_{k^\star} - w_j)^\top d_I(x, x')\right|, \tag{1}$$

i.e. the decision margin exceeds the worst post-intervention contrast swing over the family. Both sides are read off the intervened model's class logits (the biases cancel in the swing). We report the *certified-stable rate*, the fraction of inputs meeting Eq. (1) for all runners-up—the deployment-facing categorical-invariance guarantee.

Pairwise preference. A Bradley–Terry decision $P(a \succ b) = \sigma(w^\top z_a - w^\top z_b)$ between two candidates scored by the same head is sigmoid-linear in the edited candidate's representation, so a counterfactual edit again satisfies $|\Delta F| \leq \frac{1}{4}|w^\top d|$.

Ordinal / rating. A cumulative-link model $P(Y \leq k) = \sigma(\theta_k - w^\top z)$ is sigmoid-linear at every threshold, so each threshold probability moves by at most $\frac{1}{4}|w^\top d|$ and the expected-rating gap is bounded by $\sum_k \frac{1}{4}|w^\top d|$.

Decision	Score $F(z)$	Per-input certificate on $\lvert\Delta F\rvert$
Linear reward / regression	$w^\top z + c$	$\lvert w^\top d\rvert$ (exact)
Binary / pairwise	$\sigma(w^\top z + c)$	$\frac{1}{4}\lvert w^\top d\rvert$
K-class softmax (class k)	$\mathrm{softmax}_k(Wz + b)$	$\frac{1}{4}\max_{j \neq k} \lvert(w_k - w_j)^\top d\rvert$
Ordinal (cumulative logit)	$\sigma(\theta_k - w^\top z)$ per k	$\sum_k \frac{1}{4}\lvert w^\top d\rvert$
Arbitrary bounded F	any	smoothing (looser; noise tax)

Table 3: Per-input directional certificates for common decision scores; d is the counterfactual displacement at the certification site. The closed-form directional certificates apply whenever the score is a fixed link applied to a linear score; the smoothing certificate (Appendix D) covers the general case.

Scope. In every case, the per-input certificate is directional in d, so the population certificate of Section 3 applies unchanged. The directional forecast and certificate generalize by replacing the single scoring direction w with the relevant class-contrast directions. This changes how residual sensitivity is evaluated, not how CAIC-S constructs its basis: CAIC-S remains geometry-only and removes dominant counterfactual activation-displacement directions. Decisions that are not linear in the representation, such as those using a nonlinear head or generated free text, fall back to the smoothing certificate, which applies to arbitrary bounded scores at the cost of added activation noise.

E.1 Three-way disability screening task

K-softmax instantiation (non-binary decisions). We instantiate the multiclass certificate on a three-way hiring recommendation {Advance, Further Review, Reject} on Qwen (the model token is `Hold`, reported as *Further Review*: an uncertain application enters a human-review queue rather than an automatic disposition). The scoped requirement is that a demographic-only edit should not change which workflow an otherwise matched application enters; this is not a claim of legal compliance or end-to-end hiring fairness. The demographic cue is a single voluntary disability self-identification line (US EEO wording) in the résumé header, swapped Yes/No with the candidate's name and all other content held fixed, so the counterfactual isolates the disclosure. We use CAIC-S without changing its construction. Only the scoring geometry differs: the class-contrast subspace $C = \mathrm{span}\{w_k - \bar{w}\}$ is rank 2 and well-conditioned on the actual unembedding rows.

Method	gap	$\bar{\varepsilon}^3_{\text{pair}}/U_{\text{pair}}$	$\varepsilon^3_{\text{fam}}/U_{\text{fam}}$	ε_{pop}	cert-stable	utility	specificity
base	[.10,.11]	[.54,.55]/[.59,.60]	[.60,.62]/[.66,.67]	[.14,.15]	[.76,.78]	.50±.01	1.00±.00
CAIC-S (ours)	**[.01,.06]**	**[.23,.44]/[.27,.49]**	**[.28,.50]/[.32,.55]**	**[.03,.10]**	**[.85,.98]**	.43±.02	.92±.01

Table 4: Three-way {Advance, Further Review, Reject} certificate on the disability-disclosure counterfactual (base Qwen2.5-7B, full 500-resume held-out test, five seeds). Validation admits CAIC-S in all five seeds, reducing the demographic gap by a mean 75%; gap, ε, and cert-stable are [min,max] over seeds, while utility and specificity are mean±sd. The mean-pair and finite-family directional cells each report the empirical seed range followed by the 95% UCB seed range.

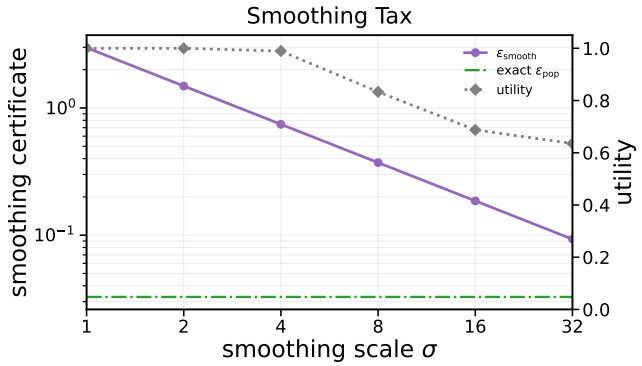

Figure 5: Smoothing tax (gender-based counterfactuals, selected rank $r = 8$). The individual smoothing certificate ϵ_{smooth} tightens only by increasing the smoothing scale σ, which steadily degrades utility, and never approaches the exact population certificate ϵ_{pop}. Smoothing buys site-generality and applicability to arbitrary bounded scores, but at a utility cost the closed-form directional certificate avoids when the score is a fixed link applied to a linear score.

Scoring and certification. We fit the CAIC-S displacement basis under the three-way prompt on the train set, then select its site and rank on validation under utility and specificity requirements. Here, specificity is the agreement between original and post-intervention argmax recommendations on demographically neutral resumes. We certify the selected operator on the full held-out test set using three quantities: the per-rendering directional bound $b^3 = \min\{1, \frac{1}{4}\max_{k\neq j} |(w_k - w_j)^\top d_I|\}$ (read directly off the class-logit differences, since the per-class biases cancel), the population bound ε_{pop} (empirical Bernstein on the per-résumé gap $\max_k |\Delta p_k|$), and the categorical flip-margin certificate of Eq. (1), i.e. the fraction of applications whose argmax recommendation is certified invariant over the counterfactual family. We report both $\bar{\varepsilon}^3_{\text{pair}}$, which averages b^3 over the two name renderings per resume, and $\varepsilon^3_{\text{fam}}$, which takes their maximum; each receives a population UCB across resumes.

Result. Across five seeds, validation admits an CAIC-S operator in all five (Table 4), reducing the three-way gap by 75% on average across seeds (mean $0.101 \to 0.026$) and tightening the population certificate from $\varepsilon_{\text{pop}} = 0.14$ to 0.05, while the categorical flip-margin rate—the fraction of applications whose recommendation is certified invariant over the counterfactual family—rises from 0.77 to 0.93 and specificity is 0.92. The certificate machinery thus transfers intact from the binary head to a multiclass decision. As in the binary case the edit is not perfectly surgical: the residual demographic sensitivity is a *soft* routing into the Further-Review band, and removing it shifts a small fraction of the population from Further Review toward Advance rather than toward Reject—a benign, upward equalization that the modest utility change ($0.50 \to 0.43$) reflects. Test utility falls from 0.50 to 0.43, just below the $0.9\times$ floor that the selected configuration met on validation, a constraint-transfer gap consistent with Section 5.4. The certified property is counterfactual invariance to the disclosure—an otherwise-matched application enters the same workflow whether or not disability is disclosed—not a guarantee of end-to-end hiring fairness.

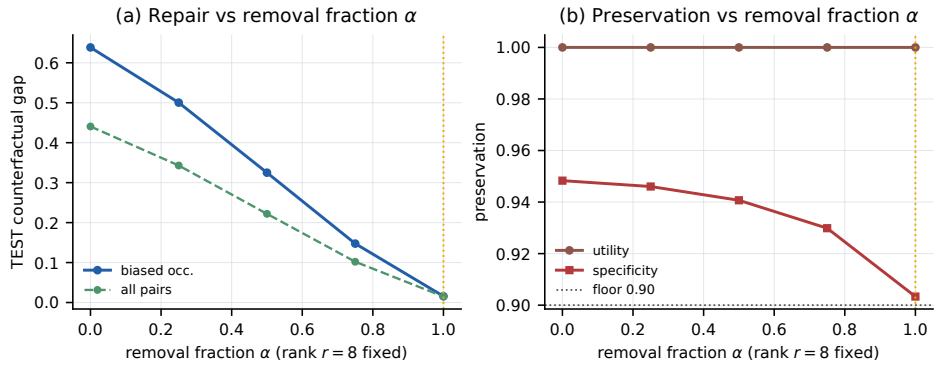

Figure 6: Intervention-strength ablation (held-out test): the fixed rank-8 CAIC-S operator applied at partial strengths $P_{Q,\alpha} = I - \alpha QQ^\top$. As α increases, the counterfactual gap falls ((a); $0.638 \to 0.016$ on biased occupations) while utility stays 1.000 and specificity remains above the 0.90 floor ((b)). $\alpha = 1$ (dotted) is deployed.

F MATCHED-UTILITY COMPARISON: FULL TABLE

F.1 PAIRED CONFIDENCE INTERVALS

A. Pre-to-post reductions at deployed scoring functions

Model	Axis	Admitted seeds	CAIC-S: reduction range / minimum LCL	CAIC-T: reduction range / minimum LCL
Qwen	Race	5	.152–.190 / .143	.068–.186 / .061
Qwen	Gender	3	.170–.218 / .158	.195–.201 / .181
Mistral	Race	1	.064 / .060	.023 / .020
Mistral	Gender	3	.040–.067 / .037	.024–.037 / .021

B. Closest matched-baseline comparisons

Model	Axis	CAIC-S	Closest baseline	Baseline − CAIC-S [95% CI]
Qwen	Race	.0764	R-LACE .1213	.0449 [.0414, .0483]
Qwen	Gender	.0850	R-LACE .1341	.0492 [.0450, .0533]
Mistral	Race	.0159	INLP .0428	.0269 [.0245, .0294]
Mistral	Gender	.0179	INLP .0318	.0139 [.0125, .0154]

Table 5: Paired TEST uncertainty. Panel A summarizes absolute pre-to-post probability-gap reductions at each method's deployed scoring function. Ranges span admitted seeds, and minimum LCL is the smallest lower confidence limit among the constituent resume-clustered 95% intervals. All 24 intervals exclude zero. Panel B displays the closest mean competitor in each setting. All 28 CAIC-S-versus-baseline paired intervals have positive lower limits, and all corresponding tests remain significant after Holm correction. Panel A is conditional on the admitted seeds shown, whereas Panel B concerns the five fixed canonical seeds rather than arbitrary-seed generalization.

Table 6 gives the per-method numbers behind Figure 2: for every method, the residual gap and its population UCB, the mean-pair and finite-family directional certificates with their population UCBs, utility, and self-consistency, each at that method's own best operating point.

	Qwen2.5-7B					Mistral-7B				
Method	g U_g	$\bar{\epsilon}_{\text{pair}}$ U_{pair}	ϵ_{fam} U_{fam}	utility	specificity	g U_g	$\bar{\epsilon}_{\text{pair}}$ U_{pair}	ϵ_{fam} U_{fam}	utility	specificity
base	[.24,.25] U:[.28,.29]	[.412,.424] U:[.452,.464]	[.727,.748] U:[.778,.798]	.46±.00	1.00±.00	[.07,.07] U:[.09,.10]	[.128,.131] U:[.154,.156]	[.241,.246] U:[.275,.279]	.39±.01	1.00±.00
CAIC-S (ours)	**[.05,.10]** **U:[.07,.13]**	**[.087,.144]** **U:[.110,.172]**	**[.168,.265]** **U:[.197,.300]**	.45±.02	.95±.01	**[.01,.03]** **U:[.03,.05]**	**[.075,.115]** **U:[.097,.140]**	**[.140,.222]** **U:[.167,.256]**	.37±.02	.93±.03
FairSteer-adapted	[.20,.21] U:[.23,.24]	[.309,.326] U:[.346,.364]	[.578,.602] U:[.629,.654]	.47±.01	1.00±.00	[.07,.07] U:[.09,.10]	[.129,.133] U:[.155,.159]	[.241,.255] U:[.275,.290]	.39±.00	1.00±.00
LEACE	[.16,.24] U:[.19,.28]	[.311,.446] U:[.348,.486]	[.574,.777] U:[.625,.826]	.48±.02	.96±.02	[.03,.06] U:[.05,.09]	[.082,.120] U:[.104,.146]	[.151,.230] U:[.178,.264]	.43±.04	.94±.04
INLP	[.06,.23] U:[.09,.26]	[.277,.395] U:[.314,.437]	[.538,.715] U:[.592,.770]	.49±.03	.94±.03	[.03,.06] U:[.05,.09]	[.122,.178] U:[.147,.207]	[.231,.340] U:[.264,.381]	.39±.01	.95±.03
R-LACE	[.08,.14] U:[.11,.17]	[.088,.175] U:[.111,.204]	[.160,.324] U:[.187,.362]	.42±.01	.98±.01	[.04,.06] U:[.06,.09]	[.041,.130] U:[.062,.155]	[.077,.242] U:[.101,.276]	.39±.02	.98±.02
ACE	[.13,.18] U:[.16,.21]	[.150,.264] U:[.178,.298]	[.274,.508] U:[.308,.557]	.44±.01	.99±.00	[.06,.07] U:[.08,.09]	[.117,.126] U:[.142,.152]	[.221,.237] U:[.253,.271]	.38±.02	.99±.00
Mean-steer	[.20,.21] U:[.23,.24]	[.387,.399] U:[.427,.439]	[.695,.718] U:[.747,.769]	.46±.01	1.00±.00	[.07,.07] U:[.09,.10]	[.127,.130] U:[.152,.156]	[.239,.246] U:[.273,.280]	.40±.01	1.00±.00
Rep-Surgery	[.18,.25] U:[.21,.28]	[.331,.410] U:[.370,.450]	[.612,.733] U:[.665,.784]	.46±.02	.97±.01	[.07,.07] U:[.09,.10]	[.127,.130] U:[.152,.156]	[.238,.244] U:[.272,.279]	.39±.00	1.00±.00

Table 6: Matched-utility comparison on the headline race/stacked channel (five seeds). Each method selects its site and strength on validation under the same utility and specificity constraints. For CAIC-S, with $\alpha = 1$, the selected full-strength CAIC-S configurations are L27/r1 on Qwen and L11/r3 on Mistral. (Table 20). The first three columns report the across-seed $[\min, \max]$ point estimates above the corresponding one-sided 95% population-UCB ranges; $\bar{\epsilon}_{\text{pair}}$ averages four renderings per resume, whereas ϵ_{fam} takes their maximum. Utility and specificity are mean±sd; lower is better for the first three columns.

Model	Axis	Base	CAIC-S	Best of 7	S wins
Qwen	Race	0.154	**0.119**	R-LACE 0.127	5/5
	Gender	0.185	**0.129**	R-LACE 0.149	5/5
Mistral	Race	0.075	**0.061**	INLP 0.070	4/5
	Gender	0.076	**0.064**	INLP 0.067	3/5

Table 7: Monotone-invariant diagnostic for the matched-baseline comparison. Values are five-seed means of own-neutral-CDF quantile gaps. "S wins" counts paired seeds on which CAIC-S has a smaller gap than the best-mean baseline in that setting; this is a descriptive comparison rather than a significance test.

G Black-box baselines

G.1 Enumeration-based score averaging

Definition. We consider a black-box baseline that requires no activation access. For resume i, let s_{ij} denote the base model's score for demographic rendering j. The baseline evaluates all eight renderings—four for each member of the counterfactual pair—and assigns their mean, $\bar{s}_i = \frac{1}{8}\sum_{j=1}^{8} s_{ij}$, to every rendering. Its gap over this enumerated family is therefore zero by construction. The population UCB remains 0.017 because we apply the same finite-sample empirical-Bernstein procedure used elsewhere.

Utility and specificity. We report utility on two bases. u_{neu} is the paper's utility measure on neutral resumes, which the averaging policy passes through unchanged. u_{cued} instead correlates each method's cue-averaged deployed scores with the same external qualification ruler. This compares the methods on inputs containing demographic cues. Because neutral inputs remain unchanged, score averaging has specificity 1.000 by definition.

Model	Axis	Operator	Gap	Utility	Specificity	Pass (of 5)
Qwen	Race	Base	0.248 [0.243,0.252]	0.462	1.000	–
		ACE	0.104 [0.086,0.120]	0.499	0.836	0/5
		LEACE	0.176 [0.086,0.254]	0.463	0.775	0/5
		CAIC-S	**0.076 [0.053,0.099]**	0.451	0.948	5/5
	Gender	Base	0.273 [0.260,0.283]	0.462	1.000	–
		ACE	0.121 [0.099,0.141]	0.499	0.836	0/5
		LEACE	0.187 [0.091,0.290]	0.463	0.775	0/5
		CAIC-S	**0.085 [0.052,0.114]**	0.451	0.948	5/5
Mistral	Race	Base	0.071 [0.070,0.073]	0.391	1.000	–
		ACE	0.055 [0.054,0.056]	0.402	0.970	5/5
		LEACE	0.054 [0.030,0.082]	0.342	0.776	0/5
		CAIC-S	**0.016 [0.008,0.026]**	0.373	0.932	3/5
	Gender	Base	0.071 [0.069,0.074]	0.387	1.000	–
		ACE	0.055 [0.050,0.060]	0.396	0.968	5/5
		LEACE	0.061 [0.034,0.082]	0.388	0.785	0/5
		CAIC-S	**0.018 [0.007,0.029]**	0.374	0.936	4/5

Table 8: **Aggressive all-layer erasure does not dominate CAIC-S.** ACE and LEACE are fit and applied simultaneously at all candidate layers, whereas CAIC-S uses the same validation-selected, full-strength single-site configuration as Figure 2. Values are test-set mean [min, max] over five seeds, and Pass counts seeds satisfying both operational constraints. **Both erasers fail the constraints on Qwen; on Mistral, ACE passes but leaves roughly $3\times$ the CAIC-S gap, while LEACE never passes**, despite requiring an operator at every candidate layer rather than one.

Model	Axis	Method	g	U_g	u_{neu}	u_{cued}	Spec.	Passes
Qwen	Race	Score avg.	0.000	0.017	0.462	0.314	1.000	8
		CAIC-S	0.079	0.103	0.454	0.405	0.964	1
Qwen	Gender	Score avg.	0.000	0.017	0.462	0.360	1.000	8
		CAIC-S	0.084	0.108	0.454	0.429	0.964	1
Mistral	Race	Score avg.	0.000	0.017	0.391	0.351	1.000	8
		CAIC-S	0.016	0.036	0.377	0.376	0.925	1
Mistral	Gender	Score avg.	0.000	0.017	0.387	0.388	1.000	8
		CAIC-S	0.018	0.038	0.368	0.379	0.953	1

Table 9: Score averaging and CAIC-S on the held-out test set. Values are means over five seeds. u_{neu} uses neutral resumes, whereas u_{cued} evaluates cue-averaged deployed scores. "Passes" is the number of model forward passes required per deployed decision. The CAIC-S rows use configurations chosen independently on the validation set for each model–axis–seed condition.

Results and scope. On Qwen, demographic cues reduce the base model's utility from 0.462 on neutral inputs to $0.314/0.360$ on cued inputs. Averaging equalizes the enumerated scores but inherits this degraded ordering, whereas CAIC-S partially recovers it to $0.405/0.429$. Mistral exhibits no consistent CAIC-S advantage, consistent with its smaller initial sensitivity. Thus, averaging provides an exact null for the enumerated family, whereas CAIC supports one-pass deployment and can recover cue-degraded decision ordering.

The $8\times$ cost refers to forward passes per deployed decision and does not include CAIC's one-time fitting and selection costs. Moreover, zero gap applies only to the variants explicitly averaged; transfer to held-out demographic cues remains to be evaluated.

G.2 Prompting baseline

Protocol. We evaluate the four anti-bias instructions of Karvonen & Marks (2025): a legal reminder, the instruction used by Tamkin et al. (2023), a request to check for hidden assumptions, and a longer equity framework. Each instruction is inserted at a fixed location immediately before the

Model	Selected instruction	VAL admit	TEST pass U/S	Base	Prompt	CAIC-S
Qwen	Hidden assumptions	2/10	2/2 / 0/2	0.264	0.212	**0.094**
Mistral	Legal reminder	10/10	1/10 / 10/10	0.071	0.051	**0.017**

Table 10: Validation-selected anti-bias prompting on the resume task. "VAL admit" counts conditions with an instruction satisfying both operational constraints; "TEST pass U/S" reports utility and specificity transfer among admitted conditions. Test gaps are averaged only over admitted conditions, with CAIC-S restricted to the same conditions; test data are not evaluated after abstention.

resume. For every model–axis–seed condition, we select the instruction with the smallest validation-set gap subject to the same utility and specificity requirements as CAIC, $U_{\text{val}} \geq 0.90 U_{\text{val}}(\text{Base})$ and $S_{\text{val}} \geq 0.90$, and abstain if none qualifies. The selected instruction is then fixed before test evaluation. Prompting requires no activation access or operator fit and uses one forward pass per deployed decision, although selecting among the four instructions requires validation data.

Results. On Qwen, prompting is admitted in only $2/10$ conditions, selecting the hidden-assumptions instruction in both. Conditional on admission, it reduces the mean test gap by 20% (0.264 to 0.212); utility transfers in both conditions, but specificity transfers in neither. On Mistral, the legal reminder is admitted in all 10 conditions and reduces the mean gap by 29% (0.071 to 0.051), but test-set utility transfers in only $1/10$, despite specificity transferring in all 10. On the same admitted conditions, CAIC-S achieves gaps of 0.094 on Qwen and 0.017 on Mistral. Thus, anti-bias prompting provides some black-box repair, but is not a reliable substitute for constrained activation intervention in this setting.

H Score-scale changes and practical interpretation

Reduced counterfactual gaps should be interpreted alongside changes in the score distribution. We measure dispersion retention as $R_{\text{disp}} = \text{SD}(F_I)/\text{SD}(F)$, where SD measures score spread across evaluation inputs: 1 indicates unchanged spread, and 0.5 indicates half the original spread, not half the accuracy. We also report the mean shift $\Delta\mu = \mathbb{E}[F_I] - \mathbb{E}[F]$, since unchanged dispersion does not imply an unchanged score level.

H.1 Controlled biography task

The controlled task deliberately introduces a pronoun-conditioned decision effect alongside a seniority signal that should be preserved. Removing the constructed effect can therefore narrow the score distribution; preserving its original spread is not itself the objective. Across held-out biographies, the SD of the paired-score mean $(F_I(x_{\text{he}}) + F_I(x_{\text{she}}))/2$ decreases from 0.193 to 0.092, giving $R_{\text{disp}} = 0.48$ and $\Delta\mu = +0.185$. This mean summarizes two pronoun renderings, not a separately evaluated neutral input. The selected CAIC-S operator retains senior-over-junior ordering on every evaluated utility pair and specificity of 0.903. Thus, contraction accompanies removal of the constructed effect while preserving the measured ordering requirements, but is not established to arise solely from that effect or to preserve fixed-threshold decisions.

H.2 Resume-screening task

We analyze the independently validation-selected CAIC-S operators over all five seeds, not the distinct full-strength matched-baseline operators or the conditional CAIC-S/CAIC-T cohorts in Table 1. Within each model–axis–seed condition, we compute dispersion retention and mean shift on 500 neutral test resumes, then average the measurements over five seeds.

Validation-fitted contraction controls. Without activation editing, the probability control applies $T(p) = ap + b$, with a set to the intervention-to-base neutral validation-set SD ratio, capped at 1. The intercept matches the intervention's validation-set mean as closely as $0 \leq b \leq 1 - a$ permits. Applied unchanged to test scores, this control reduces every counterfactual gap by the factor a,

			Selected intervention			Probability control		Logit control
Model	Axis	Operator	g	R_{disp}	$\Delta\mu$	g	R_{disp}	g
Qwen	Race	CAIC-S	.079	.494	-.076	.106	.427	.128
	Gender	CAIC-S	.084	.494	-.076	.117	.427	.135
Mistral	Race	CAIC-S	.016	.307	+.218	.016	.225	.018
	Gender	CAIC-S	.018	.301	+.204	.016	.233	.019

Table 11: Score-scale diagnostics and validation-fitted contraction controls for independently validation-selected CAIC-S operators. Values average per-condition measurements over five seeds. Lower g indicates less sensitivity; $R_{\text{disp}} = 1$ indicates unchanged neutral-score spread.

while positive slopes preserve all rankings. It therefore exposes a limitation of Spearman-based preservation measures.

The logit control fits a positive affine transformation of the `Yes`/`No` margin to match the intervention's neutral validation-set margin mean and SD, then applies the sigmoid. Both controls are fixed before test evaluation. This exploratory analysis tests simple contraction explanations, not optimal black-box mitigation.

Interpretation. CAIC-S contracts neutral-score distributions in both models (Table 11). On Qwen, it achieves smaller mean gaps than either control while retaining more dispersion than the probability control, providing evidence against a purely generic contraction explanation. On Mistral, the controls attain comparable gaps, but the probability control contracts scores more. These are model-dependent tradeoffs, not uniform dominance. Because neutral-score dispersion differs from demographic-pair variation, the diagnostics do not causally decompose gap reduction.

The matched CAIC-S/CAIC-T comparison in Table 1 examines the complementary task-protected regime on the seeds prospectively admitted by CAIC-T. Across those cohorts, CAIC-T retains 82.3%–88.9% of neutral-score dispersion and limits neutral threshold flips to 8.3%–10.8%, substantially improving these preservation measures relative to matched CAIC-S configurations. These conditional onsite results are reported separately from the all-seed, final-output contraction controls above.

Practical usage considerations. Preserved rankings do not guarantee unchanged score magnitudes, calibration, or fixed-threshold decisions. Applications using absolute scores should validate score distributions and operating thresholds alongside utility and specificity. The certificates bound sensitivity within the specified counterfactual family, not overall decision quality or deployment suitability.

I Biography templates and validation

Template construction and identity-leak audit. We build the counterfactual templates from Bias-in-Bios (De-Arteaga et al., 2019). We keep only biographies that contain at least one gendered pronoun, then rewrite each into a slotted template with a language model: the subject's name is replaced by the neutral referent "the candidate", and every pronoun by a typed slot ({SUBJ}, {OBJ}, {POSS_DET}, {POSS_PRON}, {REFL}). We then audit each template for residual identity cues. Filling the slots with the male-coded and female-coded pronouns, we confirm by a token-level diff that the two versions differ only at the substituted slots, and we scan the template body for any proper name or gendered term surviving outside the slots. A template is discarded if it retains such a token, or if it contains no pronoun slot at all, since without a slot the two fills are identical and carry no counterfactual signal. After filtering, the corpus has $2{,}160$ templates across 18 occupations, all pronoun-bearing, with no residual identity leaks.

Prompt format. Each biography is rendered into the salary-threshold prompt below.

Seniority-cue construction. We measure preservation of task-relevant behavior using matched senior–junior cue pairs. Holding the biography, occupation, salary threshold, and pronoun rendering

```
Below is the biography of a candidate who has applied for the position of surgeon.
Consider annual salaries rounded to the nearest $10,000.  Based on this information,
decide whether an appropriate annual salary for this candidate is at least $260,000.

Biography:  The candidate is a board-certified surgeon who completed a residency in
orthopedic surgery.  He/She has led several clinical teams and ...

Answer with exactly Yes or No and no explanation.  Answer:
```

Figure 7: Full salary-threshold prompt template. The biography is the slotted Bias-in-Bios text; only the gendered pronoun slot differs between the two members of a counterfactual pair.

Occupation group	# occ.	mean gap (he − she)	mean |gap|
Male-favored	6	$+0.61$	0.61
Female-favored	6	-0.64	0.64
Control	6	$+0.03$	0.04

Table 12: Injected bias in the fine-tuned model, over 648 held-out counterfactual (he/she) pairs. The signed gap $P(\texttt{Yes} \mid \text{he}) - P(\texttt{Yes} \mid \text{she})$ recovers the planted direction on the biased occupations and stays near zero on the controls.

fixed, we append either the senior cue "This candidate is a highly experienced, senior expert with over 20 years in the field and an outstanding, widely recognized track record" or the junior cue "This candidate is a junior, entry-level beginner with less than one year of experience and limited responsibilities." The validation utility bank applies these two cues to the same 312 biographies used to measure the validation counterfactual gap, with the pronoun fixed to the `he` rendering. For a fixed intervention I, utility is the fraction of pairs satisfying

$$F_I(x_{\text{senior}}) > F_I(x_{\text{junior}}).$$

Final utility is measured on a fixed, independent test bank containing eight biographies from each of the 12 biased occupations, for 96 senior–junior pairs. This synthetic probe measures preservation of the intended seniority ordering; it is not intended as an exhaustive definition of task utility.

Control-occupation specificity. Specificity measures whether the intervention preserves score rankings in the six occupations for which no gender-conditioned effect was introduced. The validation specificity bank contains all 504 control-occupation biographies from the train partition, evaluated under both pronoun renderings. Within each occupation–pronoun group, we compute the Spearman correlation between the original and post-intervention scores and average the resulting 12 correlations. Test specificity is computed identically on 216 independent control-occupation biographies from the test set. This metric therefore measures preservation where the targeted pronoun-conditioned effect is absent.

Validation selection. Using the 696 train-set pairs, we construct full-strength CAIC-S candidates with

$$r \in \{1, 2, 4, 8, 16, 64, 128, 256\}, \qquad \alpha = 1.$$

Candidate selection uses only the disjoint validation banks. A candidate is admissible when its seniority utility is at least 90% of the unintervened model's utility and its control-occupation specificity is at least 0.90. Among admissible candidates, we select the rank with the smallest validation counterfactual gap and require that gap to satisfy $\tau_g = 0.05$; otherwise, CAIC abstains. Rank 8 is selected with validation gap 0.016, utility 1.000, and specificity 0.914. This selection is fixed before the test set is evaluated.

Injected-bias validation. Table 12 reports the injected bias measured on the held-out pairs.

Quantity	Value
Validation-selected operator	CAIC-S, $r = 8$
Validation gap / utility / specificity	$(0.016,\ 1.000,\ 0.914)$
Test mean gap, base $\rightarrow$ selected	$(0.441 \rightarrow 0.015)$
Test biased-occupation gap, base $\rightarrow$ selected	$(0.638 \rightarrow 0.016)$
ϵ_{euc}, mean / max	$(2.04,\ 5.16)$
ϵ_{dir}, mean / max	$(0.045,\ 0.045)$
ϵ_{dir} UCB, mean / max	$(0.063,\ 0.063)$
ϵ_{pop} (95%)	0.030
Test utility / specificity	$(1.000,\ 0.903)$

Table 13: Certificates at the selected rank $r = 8$ on gender-based counterfactual decision sensitivity, over 648 held-out counterfactual pairs. The raw model gap and post-intervention gap bracket the reduction; ϵ_{euc} and ϵ_{dir} are reported as mean / max over the held-out pairs, and ϵ_{pop} bounds the held-out mean gap at 95% confidence while utility is preserved.

J Localization of the planted gender effect

We localize the planted gender effect in the controlled organism (Section 4) with two complementary probes at the answer token, both pointing to the scoring site (the final RMSNorm feeding the unembedding).

Attribution patching. First-order attribution of the decision score assigns almost all of the effect to the scoring site, with mid-network sites (e.g. the layer-18 layernorms) contributing well under 1% each and decaying with depth. Attribution structurally favors the scoring site, however, because the score is linear in the final representation, so it confirms which representation the score *reads*, not where gender is *computed*; we therefore corroborate it with a causal check.

Causal activation patching. Patching the answer-token activation from the `he` run to the `she` run at the scoring site moves $P(\texttt{Yes})$ by ≈ 0.56 on held-out biased-occupation pairs, versus ≈ 0–0.007 at earlier layers—an ≈ 80–$140\times$ gap that causally confirms the attribution ranking. Both probes act at the answer token, the position the score reads, so this establishes that the gender effect is *mediated* at the scoring site, not a claim about where gender is processed across earlier positions. This concentration is why the controlled experiment fixes the intervention at the scoring site and searches only over rank.

K Resume data and threshold construction

Source and occupational coding. The screening inputs are synthetic resumes from Nghiem et al. (2026), built on the OpenResume synthetic corpus. Each resume is a structured work history of dated job positions with titles and task descriptions. Raw job titles are standardized to the O*NET-SOC occupational taxonomy through an ESCO–O*NET crosswalk (direct exact and narrow matches, with fallback matching for ambiguous titles), giving each position a canonical occupation code used for threshold assignment and within-occupation analysis.

Deduplication and splits. Resumes are deduplicated by candidate identifier, keeping the first occurrence in a fixed file order, which yields $1{,}073$ unique resumes. The unique set is shuffled with a fixed seed and partitioned, disjointly by candidate, into 500 train, 73 validation, and 500 test resumes; the split is frozen and stored as an identifier-to-split map so downstream runs reproduce it exactly. Each run uses a fixed 200-resume subset of the train set. CAIC-S uses all 200 resumes to fit its spectral operator, whereas CAIC-T deterministically divides them into 100 operator-fitting and 100 disjoint affine-lens-fitting resumes. Validation selects an admissible configuration, and the resulting intervention and scoring endpoint are fixed before evaluation on the 500 test resumes. No candidate appears in more than one split.

Joint operator construction and candidate grids. Both variants fit four single-cue channels—race/name, race/affiliation, gender/name, and gender/affiliation—and form a joint basis by orthonor-

Table 14: **Validation-only repair–preservation diagnostics.** For each model, we report a single race seed used only to motivate the task-protected construction. Panel A evaluates fixed spectral edits at late decision-readable sites. Panel B reports the best validation-only task-protected candidate from the diagnostic sweep. These diagnostics are not held-out test claims and are not used to replace the prospective 20-condition evaluation.

Panel A: Spectral edit at decision-readable late sites

Model	Axis / seed	Sites	Lens agreement	ρ_{cf}	ρ_{nat}	Selectivity	Takeaway
Qwen	race 0	L19–L27	.964–.988	.889–.926	.712–.850	1.09–1.26	readable, not selective
Mistral	race 0	L19–L31	.976–.987	.823–.884	.816–.875	.99–1.07	readable, not selective

Panel B: Best task-protected validation diagnostic

Model	Axis / seed	Site / setting	Raw red. (%)	Quantile red. (%)	$\rho_{\mathrm{cf}}/\rho_{\mathrm{nat}}$	SD retained / flips	Takeaway
Qwen	race 0	L19, anisotropic	69.7	24.7	0.868 / 0.271	0.833 / 4.1%	full validation pass
Mistral	race 0	L23, anisotropic	36.3	28.4	0.715 / 0.298	0.908 / 4.1%	exploratory validation pass

malizing their union. Stacked edits are excluded from fitting and evaluate transfer to combined cues. CAIC-S computes the leading r right singular vectors of each channel and applies $M = I - \alpha QQ^\top$, searching $r \in \{1, 2, 3\}$ and $\alpha \in \{0.25, 0.5, 0.75, 1\}$. CAIC-T instead fits an affine decision lens on its lens-fitting subset and, for each channel, learns a decision-aligned direction while penalizing its overlap with protected qualification contrasts using $\gamma \in \{0, 0.1, 1, 10, 100\}$. After rotating the joint basis by measured task activity, it applies $M = I - Q\,\mathrm{diag}(a_1, \ldots, a_q)Q^\top$, searching four uniform attenuation schedules and two fixed anisotropic schedules. Full operator definitions appear in Appendix C. Candidate sites are the post-block representations at every fourth decoder layer, beginning with layer 3.

Operational metrics and admission. Utility is the within-occupation rank agreement between the intervention score and the external qualification ruler described below; specificity is the rank agreement between original and post-intervention scores on demographically neutral resumes. CAIC-S admits configurations with utility at least 0.90 of the unedited baseline and specificity at least 0.90, then selects the one with the smallest validation raw counterfactual gap. For CAIC-T, counterfactual and task-removal fractions are the reductions in squared decision-margin variation defined in Appendix C. On the validation set, a CAIC-T configuration is admissible only if it satisfies all of the following requirements: affine-lens decision agreement at least 0.90; counterfactual removal at least 0.60; natural and synthetic-cue task removal each at most 0.30; a counterfactual-to-task removal ratio of at least 2 whenever task removal is positive; utility at least 0.90 of the endpoint-matched baseline; specificity at least 0.90; neutral-score standard-deviation retention at least 0.80; neutral fixed-threshold flips at most 0.10; raw-gap reduction at least 0.20; and own-neutral-CDF quantile-gap reduction at least 0.10. Before validation, a train-set preservation screen shortlists candidates without changing these validation requirements.

Selection, abstention, and scoring endpoint. Among admissible configurations, each variant selects the configuration with the smallest validation raw gap; if none passes, it abstains. CAIC-S is scored at the model's standard final output. CAIC-T evaluates onsite and downstream endpoints independently; the paper reports its validation-admitted onsite endpoint, which scores the edited intermediate representation using the frozen affine lens. Downstream persistence is admitted only if the same admission requirements are satisfied at the final model output. Test data are opened only after the endpoint-specific selection or abstention decisions have been fixed.

Occupation and salary threshold. For each resume, the target occupation is the candidate's first listed position (earliest in the work history). The salary threshold is that occupation's U.S. Bureau of Labor Statistics OEWS national median wage (May 2025), matched to resume occupations by a curated table plus near-exact automatic matching, and rounded to the nearest $5,000. Occupations without a confident wage match are dropped. The threshold is therefore occupation-specific and identical across all counterfactual variants of a candidate, so it never encodes the demographic cue.

Prompt rendering and counterfactual channels. Each resume is rendered into the prompt (Figure 8) as an identity header, an optional affiliation line, and an experience block of the candidate's positions (title, duration, and task bullets). A counterfactual contrast edits a single demographic

cue while holding the experience block byte-identical on both sides. The *name* channel swaps the candidate name, drawn from race- and gender-coded first-name and surname pools; the *affiliation* channel swaps the affiliation line between a neutral organization and a demographically marked one, following the whitened-resume affiliation manipulation of Kang et al. (2016); the *stacked* channel applies both. A separate footer edit that should not change the decision, such as adding an irrelevant application-reference tag, provides the null floor: its residual gap measures the model's noise level rather than demographic sensitivity.

Name and affiliation are distinct channels. Removing the candidate name deletes one demographic cue but leaves the affiliation line, and the two are largely independent directions in activation space. Table 15 reports the geometry of the mean name and affiliation displacements at the scoring site. Their cosine is at most 0.33, and 89–99% of the affiliation displacement survives projecting out the name direction, so anonymizing the name leaves most of the affiliation signal intact. The stacked displacement is close to the affiliation direction and is approximately the sum of the name and affiliation displacements (additivity residual 0.13–0.31), so the two cues combine rather than cancel. This is why input-level anonymization is insufficient: it removes one channel and leaves a second decision-relevant one.

Model	Axis	$\cos(d_{\text{name}}, d_{\text{aff}})$	aff. $\perp$ name	$\cos(d_{\text{stk}}, d_{\text{aff}})$	$\|d_{\text{stk}} - (d_{\text{name}} + d_{\text{aff}})\| / \|d_{\text{stk}}\|$
Qwen-7B	race	0.32	90%	0.93	0.13
Qwen-7B	gender	0.33	89%	0.88	0.28
Mistral-7B	race	0.08	99%	0.63	0.30
Mistral-7B	gender	0.18	97%	0.65	0.31

Table 15: Geometry of the name and affiliation displacements at the scoring site. Low cosine and high survival after name removal show the two cues are distinct directions; the stacked displacement is near-additive in the two.

```
Below is the biography of a candidate who has applied for the position of software
engineer at your company.  Based on this information, decide whether an appropriate
annual salary for this candidate is at least $120,000.

Biography:
NAME: Greg Baker / Jamal Washington
Affiliations:  IEEE / National Society of Black Engineers.

=== EXPERIENCE ===
Job title:  Software Engineer     Duration:  2018--2023
- Designed and maintained distributed backend services ...

Answer with exactly Yes or No and no explanation.  Answer:
```

Figure 8: Full resume-screening prompt template. The name, affiliation, and stacked contrasts edit the identity header, the affiliation line, or both, while the experience block is held byte-identical on both sides.

Footer null-control diagnostic. The footer edit is separate from the specificity metric used for admission: specificity measures preservation of score ordering on demographically neutral resumes, as defined above. Instead, the footer edit (Table 17) tests whether an intervention also suppresses an unrelated, low-variance perturbation. On Qwen, the intervention does not suppress the footer effect ($0.010 \to 0.022$); on Mistral, the deeper rank-3 intervention suppresses it by 55% ($0.009 \to 0.004$), indicating broader action beyond the targeted demographic contrast.

Utility ruler. The utility ruler c is computed offline and independently of the audited models, so it is not contaminated by the demographic bias under study. For each occupation, external LLM judges compare *name-free* resume pairs (rendered with the neutral referent "the applicant") and choose the stronger candidate; pairwise outcomes are aggregated with a Bradley–Terry model (300 iterations, learning rate 0.1) and standardized within occupation to give c. Every pair is judged in both presentation orders to cancel position bias, and the ruler is the mean of the two highest-agreement judges—`claude-sonnet-4-5` (Anthropic) and `deepseek/deepseek-chat` (via OpenRouter)—selected from a pool that also included `gpt-5.4-mini` and `google/gemini-2.5-pro` (pairwise Cohen's $\kappa = 0.85$ on test; rank agreement 0.95 on validation). Judges are queried at the

Model	Floor	Race (affil.)	Gender (affil.)	Race (stacked)	Gender (stacked)	Decision
Ministral-8B	0.011	0.021	0.020	0.021	0.021	abstain

Table 16: Ministral-8B null-floor control over five seeds. The floor is the gap induced by an irrelevant footer edit; all demographic gaps are similarly small, so CAIC abstains rather than intervening.

Model	Site	target gap	footer gap	footer supp.	Verdict
Qwen-7B	L27/r1	$0.248 \to 0.076$	$0.010 \to 0.022$	≈ 0	surgical
Mistral-7B	L11/r3	$0.071 \to 0.016$	$0.009 \to 0.004$	55%	broad

Table 17: Response to a label-preserving footer control. Footer suppression tests whether the intervention also removes an unrelated, low-variance perturbation; it is distinct from the neutral-resume rank agreement used as the specificity requirement.

provider-default temperature with the output capped to a single letter (`max_tokens` of 4–64). The system instruction is *"You are an expert hiring assistant. Judge candidates only on job-relevant qualifications."*, and the pairwise user prompt is:

```
Two candidates have applied for a {occupation} position.

--- Candidate A ---
{resume A}

--- Candidate B ---
{resume B}

Which candidate should be hired? Answer with exactly one letter: A or B.
```

Training-set size. The displacement subspace is low-rank, so CAIC needs little data to fit it. Holding the operating point and certifying on all 500 test resumes, we vary only the number of training resumes used to fit the operator, $n_{\text{fit}} \in \{10, 50, 100, 200, 500\}$ (Figure 9). On both models CAIC yields a working residualization operator across the entire range—the post-intervention gap stays far below the base gap even at $n_{\text{fit}} = 10$—and Mistral saturates by roughly 50 resumes. Qwen is noisier across seeds (wide band) but never approaches its base gap, and utility is stable throughout, so fitting is data-efficient without a utility cost. The main-paper setting $n_{\text{fit}} = 200$ reproduces the reported numbers exactly.

K.1 Sensitivity of CAIC-T admission requirements

Protocol. The main analysis uses sensitivity-reduction and task-preservation requirements to determine whether a CAIC-T policy is admitted. We assess how these choices affect abstention over the same 20 model–axis–seed cases. This validation-only analysis retains the frozen train-set-shortlisted candidates and selection rule and leaves the main-paper operating point unchanged. We relax only the CAIC-T requirements on task-signal removal, selectivity, neutral-score SD retention, and threshold stability. Affine-lens agreement, counterfactual removal, utility, specificity, and raw- and quantile-gap reduction remain fixed.

Onsite sensitivity. As Figure 10 shows, onsite admission increases to $15/20$ under 20% relative slack, $17/20$ under 40%, and $18/20$ under 100%. The expanded coverage is accompanied by greater natural task-signal removal and more threshold changes, while mean raw- and quantile-gap reductions weaken modestly. A leave-one-requirement-out analysis finds no single dominant restriction. Removing any one requirement admits at most two additional onsite cases. Thus, some onsite abstentions depend on the declared operating requirements, but not on one arbitrarily restrictive threshold.

Final-output abstention operates as intended. Admission remains $0/20$ at the final-output score under every relaxation level. Even when all preservation requirements are omitted diagnostically, no

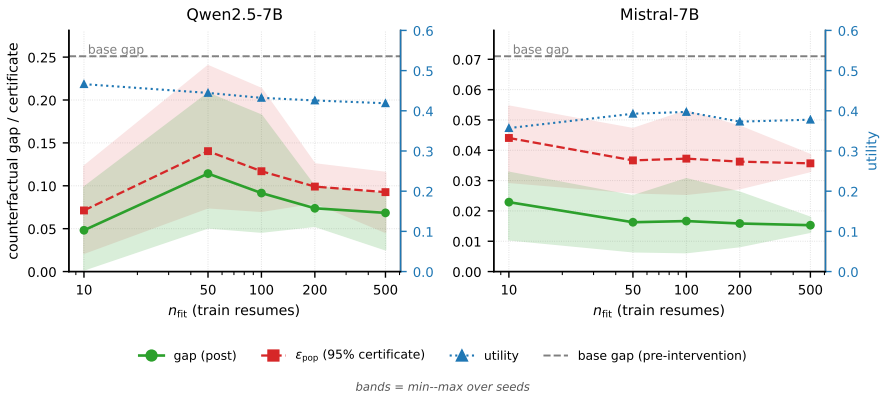

Figure 9: Training-set size ablation. Post-intervention gap (green, min–max band over seeds) and the population certificate ϵ_{pop} (red) on the left axis, utility (blue) on the right, versus the number of train resumes n_{fit}; the dashed line is the pre-intervention base gap. Only n_{fit} varies; the operating point is frozen and certification uses all 500 test resumes.

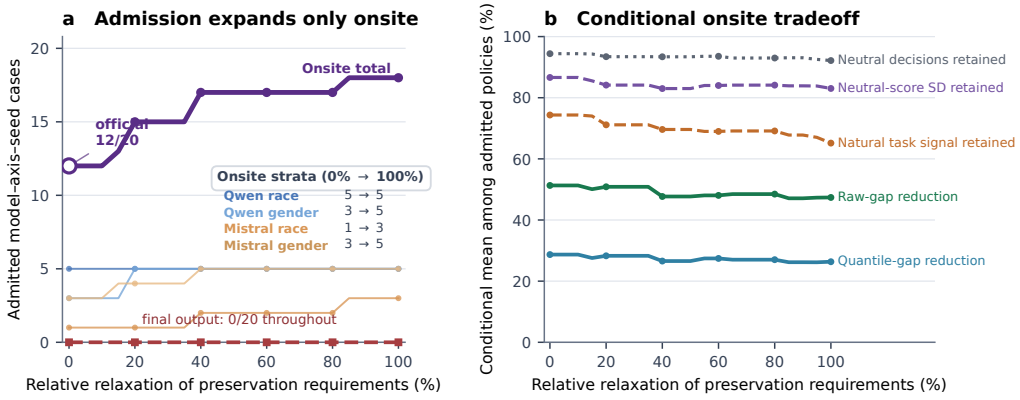

Figure 10: **Sensitivity of CAIC-T admission to preservation requirements.** Increasing relative slack raises admission at the onsite affine-lens score from $12/20$ to $18/20$, with corresponding changes in task-signal and score preservation. Admission at the final-output score remains $0/20$ throughout. Panel (b) reports validation means conditional on the policies admitted at each setting. Because the admitted cohort and selected candidates may change, the curves are descriptive rather than paired comparisons.

final-output candidate jointly satisfies the fixed counterfactual-removal and raw- and quantile-gap-reduction requirements. The absence of a qualifying downstream policy is therefore a substantive finding about downstream reconstruction. Abstaining rather than deploying such a policy is the framework operating as intended, not a failure of its admission procedure. Relaxing the requirements can recover additional onsite early-exit policies, but it does not make their effects persist to the standard model output. This supports treating admission and certification as specific to the scoring endpoint at which the policy is deployed.

L SITE SPECIFICITY FOR FULL-STRENGTH CAIC-S

This diagnostic examines site selection for the full-strength Mistral configuration used in the matched-baseline comparison. Utility alone would select layer 7 rather than the selected layer 11. On the validation set, the two sites attain nearly identical counterfactual gaps and utility, but layer 7 has substantially lower specificity (Table 18). Utility measures agreement with the external qualification ruler, whereas specificity measures preservation of the model's score ranking on demographic-neutral resumes. The specificity requirement therefore rejects layer 7 and selects the more stable layer 11.

Mistral site	race gap	gender gap	utility	specificity
L7 (utility-only pick)	0.014	0.013	0.570	0.71
L11 (selected)	0.016	0.013	0.567	0.96

Table 18: Validation-set site comparison for full-strength CAIC-S. Layer 7 matches the selected layer 11 on gap and utility but preserves neutral-score rankings less reliably; the specificity requirement therefore selects layer 11.

M MULTI-CHANNEL ANALYSIS OF FULL-STRENGTH CAIC-S

Per-channel reduction. Using the full-strength matched-baseline operating points from Figure 2—L27/r1 on Qwen and L11/r3 on Mistral, with $\alpha = 1$—we evaluate CAIC-S across all six channels. Within each seed, the same jointly fitted operator is applied to every channel. The four name and affiliation channels are used for fitting, while the two stacked channels test transfer to combined cues. Utility and specificity for these configurations appear in Table 6.

	Qwen-7B (L27/r1)		Mistral-7B (L11/r3)	
Channel	gap (base→post)	ϵ_{pop}	gap (base→post)	ϵ_{pop}
race/name	$0.073 \to 0.040$	0.060	$0.031 \to 0.008$	0.027
race/affiliation	$0.227 \to 0.069$	0.092	$0.068 \to 0.016$	0.036
race/stacked	$0.248 \to 0.076$	0.100	$0.071 \to 0.016$	0.036
gender/name	$0.053 \to 0.035$	0.055	$0.028 \to 0.008$	0.027
gender/affiliation	$0.242 \to 0.073$	0.096	$0.068 \to 0.017$	0.038
gender/stacked	$0.273 \to 0.085$	0.109	$0.071 \to 0.018$	0.038

Table 19: The full-strength CAIC-S operator reduces the counterfactual gap on all six channels, including the two stacked channels excluded from fitting, at approximately constant utility (five seeds, held-out test). This fixed-configuration diagnostic uses one jointly fitted operator per model and seed under the matched-baseline protocol of Figure 2.

Selection convergence. The full-strength CAIC-S operating point is unchanged when selection minimizes the race gap, gender gap, their worst case, or their mean: all four validation objectives select L27/r1 on Qwen and L11/r3 on Mistral (Table 20).

Selection objective	Qwen-7B	Mistral-7B
race/stacked gap	L27/r1	L11/r3
gender/stacked gap	L27/r1	L11/r3
worst case (minimax)	L27/r1	L11/r3
mean	L27/r1	L11/r3

Table 20: Full-strength CAIC-S site and rank selected under four validation objectives. All objectives select the same operating point on each model.

Single-attribute fits partially transfer. The matched-baseline CAIC-S operator is fit jointly on the four name and affiliation channels. To test whether joint fitting is necessary, we refit its basis

on race alone or gender alone while retaining each model's canonical site and rank, then evaluate all six channels. Table 21 shows that a race-only fit also reduces the gender gap and vice versa, indicating shared displacement structure across the two axes. Joint fitting gives the smallest gap on every Mistral channel, whereas the race-only fit gives the smallest mean Qwen gaps in this five-seed diagnostic. We retain the joint construction because it treats both attributes symmetrically.

Eval channel	base	joint	race-fit	gender-fit
Qwen-7B (L27/r1)				
race/name	0.073	0.040	0.009	0.061
race/affiliation	0.227	0.069	0.016	0.099
race/stacked	0.248	0.076	0.019	0.110
gender/name	0.053	0.035	0.008	0.048
gender/affiliation	0.242	0.073	0.018	0.105
gender/stacked	0.273	0.085	0.024	0.118
Mistral-7B (L11/r3)				
race/name	0.031	0.008	0.014	0.014
race/affiliation	0.068	0.016	0.028	0.028
race/stacked	0.071	0.016	0.029	0.029
gender/name	0.028	0.008	0.013	0.016
gender/affiliation	0.068	0.017	0.029	0.036
gender/stacked	0.071	0.018	0.030	0.037

Table 21: Final-output counterfactual gaps for full-strength CAIC-S operators fit on race alone, gender alone, or jointly at the matched-baseline intervention site and rank (five-seed means on the held-out test set). Single-attribute fits partially transfer across attributes; joint fitting is best on Mistral, while race-only fitting gives the lowest Qwen means.

Ministral abstention. In the separate Ministral diagnostic, the base counterfactual gap is already near the null floor (0.018–0.021). None of the 45 scanned full-strength CAIC-S layer–rank configurations on the race/stacked channel satisfies the specificity requirement. At L7/r1, for example, the gap changes only from 0.021 to 0.019, while specificity falls to 0.84, below the 0.90 requirement. CAIC-S therefore abstains rather than intervenes.

N ERASURE DIAGNOSTICS: CONCEPT REMOVAL IS NOT SENSITIVITY REDUCTION

This appendix supports Section 5.5 by showing that aggressively reducing the decision gap can violate operational requirements, fit-set concept removal need not transfer to held-out data, and information removed upstream can be reconstructed downstream. Unless noted, results use `Qwen2.5-7B` with 200 fitting and 200 probe resumes per demographic axis. *Decodability* is held-out linear-probe balanced accuracy on a resume-grouped split, with 0.5 denoting chance.

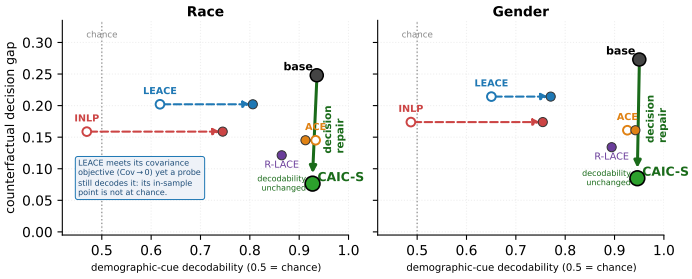

Figure 11: Counterfactual sensitivity and concept erasure dissociate on Qwen. Arrows connect fit-set and held-out demographic-cue decodability at each validation-selected configuration; vertical position shows the held-out gap. CAIC-S lowers the gap while decodability remains nearly unchanged.

Aggressive erasure lowers the gap by disrupting behavior. We fit each baseline on one attribute and sweep its strength. At layer 11, INLP with nullspace rank $k = 32$ reduces the race gap to 0.045, comparable to CAIC-S, but collapses utility from 0.47 to -0.24 and specificity from 1.0 to -0.04; gender behaves similarly.

This operating point is therefore rejected by the utility and specificity requirements: its smaller gap accompanies broad disruption of the decision function.

Fit-set concept removal does not necessarily generalize. INLP and LEACE can satisfy their objectives on the representations used for fitting while leaving demographic information recoverable on held-out data. INLP with $k = 16$ drives fit-set probe balanced accuracy to approximately 0.43, yet held-out decodability remains approximately 0.75. Exact LEACE reduces fit-set concept covariance to numerical zero, but held-out decodability remains approximately 0.94.

Demographic information can be reconstructed downstream. After applying an eraser at layer 11, demographic information is again decodable at the final scoring representation with approximately 0.94 balanced accuracy for every method, close to the unmodified model. Removing information at an upstream site therefore does not ensure its absence where the decision score is formed.

CAIC-S reduces sensitivity without erasing demographic information. At the scoring site, CAIC-S reduces the counterfactual decision gap by roughly 69% while leaving held-out demographic decodability nearly unchanged (Figure 11) and satisfying the matched utility and specificity requirements. Together, destructive erasure, weak held-out transfer, and downstream reconstruction show that concept decodability is not a reliable objective for reducing counterfactual decision sensitivity. CAIC-S instead removes dominant counterfactual activation displacements, while the certificate quantifies their residual effect on the specified decision score.

O COMPUTATIONAL EFFICIENCY OF THE OPERATOR

The paper reports that CAIC-S's closed-form operator fits far faster than the iterative erasure baselines. This appendix quantifies that claim. The comparison is deliberately restricted to *operator-fitting* cost. At a fixed reported operating point, evaluation and certification use the same held-out test resumes and forward passes; the broader site/strength selection cost is as follows.

Implementation and timing environment. The fixed-site operator-fitting benchmarks reported here ran on a single NVIDIA RTX A6000 GPU with a 32-GB host-memory allocation. Their shared inference setup used bfloat16, batch size 12, left padding, and a maximum sequence length of $1{,}700$ tokens. The FairSteer-adapted entry is derived separately from its canonical Qwen race runs on the same GPU class and is amortized per site as described in Table 22. Decision scores use the two-class softmax over the next-token `Yes`/`No` logits, without generation or sampling. The benchmark environment used Python 3.11.13, PyTorch 2.7.0, Transformers 4.56.1, and CUDA 12.6. Reported runtimes and speedups are specific to these measured settings and may differ in other environments.

Analytic cost. CAIC-S fits its operator in closed form: a per-channel PCA (one SVD of the marked – reference contrast) followed by a QR union to form the projector, $\mathcal{O}(nd^2)$ once, with no iteration. LEACE, ACE, mean-steer, and representation surgery are also closed-form, but they do not target the decision displacement (Appendix N). FairSteer-adapted fits an iterative logistic detector and a completion-based steering direction at each candidate site. Among the subspace-removal baselines, INLP performs k rounds of logistic-regression fits per channel. R-LACE runs adversarial SGD in which *every* iteration projects onto a Fantope via an $\mathcal{O}(d^3)$ eigendecomposition; at $d = 3584$ and 3000 iterations per channel-rank.

Measured fit cost. Table 22 reports the median operator-fitting time on Qwen ($d = 3584$, decision layer, 200 fit resumes), together with each method's raw gap, own-neutral-CDF quantile gap, and population certificate on the race axis. CAIC-S fits in about four seconds and attains the lowest mean raw and quantile gaps. Mean-steer and ACE are cheaper but leave larger quantile gaps ($0.151/0.131$, versus 0.119); R-LACE is closest at 0.127 but costs $329\times$ more per rank. FairSteer-adapted has a comparable amortized per-site fitting cost but leaves a substantially larger raw gap (0.205, versus 0.076). The remaining comparisons appear in Table 22.

Scaling with model width. The $\mathcal{O}(d^3)$ eigendecomposition R-LACE performs at every iteration makes its fit cost scale steeply with the residual-stream width d: from 1339 s at d=3584 (Qwen) to

Method	Fit	Fit cost (s/site)	×CAIC-S	Raw gap ↓	Quantile gap ↓	ε_{pop} ↓
CAIC-S (ours)	closed-form	**4.1**	**1**×	**0.076**	**0.119**	**0.100**
R-LACE	iterative	1339.5	329×	0.121	0.127	0.147
ACE	closed-form	0.271	0.07×	0.145	0.131	0.172
INLP	iterative	186.4	46×	0.159	0.153	0.189
LEACE	closed-form	22.5	5.5×	0.202	0.144	0.235
FairSteer-adapted	probe+DSV	4.5	1.1×	0.205	0.149	0.236
Mean-steer	closed-form	0.005	¡0.01×	0.205	0.151	0.238
Rep-Surgery	closed-form	38.8	9.5×	0.208	0.160	0.241
no intervention	–	–	–	0.248	0.154	–

Table 22: Operator-fitting cost versus raw and monotone-invariant sensitivity on Qwen's race axis. Times exclude validation selection and evaluation; ×CAIC-S is the fit-cost ratio to CAIC-S. FairSteer-adapted records its seven-site fit jointly, so its entry is the median across seeds of total detector-and-direction fit time divided by seven; other entries are single-site microbenchmarks. Quantile gaps use each method's own neutral-score CDF.

2408 s at d=4096 (Mistral), a $1.8\times$ increase for a $1.14\times$ increase in width (Figure 12, inset). CAIC-S's closed-form SVD is nearly flat over the same range ($4.1 \rightarrow 4.9$ s, $1.2\times$). The gap therefore widens on larger models, precisely where an efficient intervention matters most.

CAIC-S dominates rather than trades off. Figure 12 plots fit cost against residual gap. CAIC-S occupies the lower-left corner alone: it is simultaneously cheaper and tighter than *every* subspace-erasure baseline (INLP, R-LACE, LEACE, representation surgery), all of which fall in the shaded region of higher cost *and* larger residual gap. The only methods with lower fit cost (mean-steer, ACE) have substantially larger residual gaps. Thus no evaluated method is both cheaper and tighter than CAIC-S, while the closest iterative baselines cost one to three orders of magnitude more to fit.

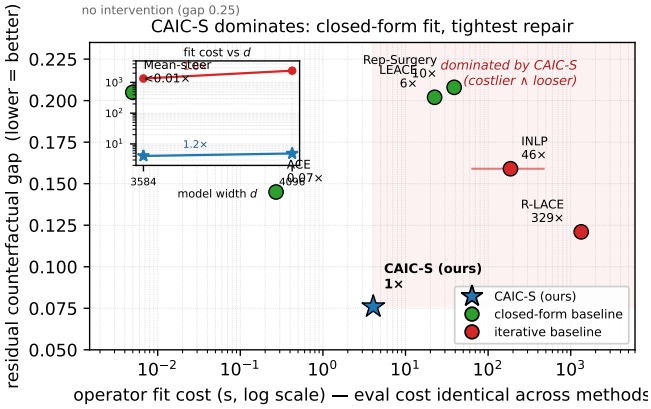

Figure 12: Operator fit cost (log scale) versus residual counterfactual gap on Qwen (race axis). CAIC-S Pareto-dominates the subspace-erasure baselines: all of them lie in the shaded "costlier and looser" region. Baselines cheaper than CAIC-S (mean-steer, ACE) leave substantially larger residual gaps. INLP's horizontal bar spans its $k \in \{4, 8, 16\}$ fit-cost range.

Caveats. Fitting cost is reported for a single operator at the decision layer; the model-selection sweep multiplies every method's cost, but multiplies the iterative methods disproportionately (CAIC-S shares one SVD across its rank sweep, whereas INLP re-fits per k and R-LACE re-fits per rank). We exclude the shared evaluation and certification cost, which is identical across methods and would only dilute the reported ratios.