SHAPLEY-CALIBRATED SEMANTIC ROUTING FOR BLACK-BOX LLM GUARDRAILS: WHAT A FIXED ROUTE RETAINS

Anonymous authors
Paper under double-blind review

ABSTRACT

Input attribution localizes model behavior to specific prompt components, turning debiasing from untargeted rewriting into targeted intervention. A deployed guardrail, however, must commit in advance to the component it will act on. We study what is lost when prompt-level attribution is compressed into a single fixed routing decision. We introduce **Block Prior** (BP), which decomposes prompts into four interpretable roles (SUBJECT, ARGUMENT, ACTION, and SETTING). BP computes exact four-player Shapley attributions offline under an auditable input-side bias function and aggregates them into one regime-level route applied unchanged to every incoming prompt. Exact attribution costs $2^4 = 16$ coalition evaluations per prompt and is tractable both offline and online. The question we answer is how much localization a single stored coordinate can carry. On social-bias regimes from BBQ, the four-role calibration issues a falsifiable prediction about which regimes should depart from a fixed SUBJECT-ONLY policy, and held-out intervention confirms it: $+39.2$ points of bias reduction for Age and $+6.5$ for Physical Appearance, with exact coincidence for Gender Identity, where the two policies are identical by construction. The compressed route recovers 77% of the bias reduction obtained by recomputing exact Shapley attribution for every prompt (54.5% versus 70.7% pooled) while improving on the fixed SUBJECT-ONLY policy by 11.2 points, and it preserves high prompt similarity (0.929). We then map where the compression works and where it fails: it is near-insensitive to additive padding ($< 1\%$ attack success), degrades under out-of-vocabulary lexical substitution ($\approx 40\%$), and fails completely when bias-bearing evidence is relocated to a non-routed role (100%), a perturbation against which prompt-adaptive attribution remains effective. Cross-dataset and cross-regime studies establish, respectively, ordinal profile agreement on a target-conditioned subset and representational portability of the four-role interface under a different risk specification. Our results quantify the price of a fixed attribution-derived policy.

Anonymized source code and supporting experimental resources are included in the supplementary material.

1 INTRODUCTION

Large Language Models (LLMs) are increasingly deployed through black-box interfaces, making external input-side guardrails a natural control layer for heterogeneous and proprietary models. Effective debiasing requires not just detecting biased output but isolating the exact semantic components to modify. This distinction is critical for social bias, where demographic terms, descriptors, actions, and surrounding context contribute unequally to harm across different deployment settings (Bai et al., 2025; Asgari et al., 2026; Elbouanani et al., 2026; Sun et al., 2026). Pinpointing where to intervene is just as critical as knowing that an intervention is needed. Semantic routing provides the decision logic necessary to target the precise source of bias rather than blindly altering surrounding text and degrading output quality. Existing targeting strategies operate at different levels of model access. White- and gray-box methods act on parameters, internal representations, activations, or decoding behavior (Lamb et al., 2025; Chen et al., 2025c; Nadeem et al., 2026), while inference-time approaches introduce additional prompting or reasoning (Chen et al., 2026). External guardrails

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.

provide a complementary interface based on observable inputs and outputs (Islam and Surdeanu, 2026; Liu et al., 2026; Zhang and Wan, 2026). Their central localization problem is to infer which semantic component should be modified, calibrate that decision from representative data, and reuse it across subsequent requests. A further question is how much prompt-specific adaptation this localization requires. Prompt-adaptive attribution can exploit instance-level variation, whereas a reusable regime-level policy trades some of that adaptation for a fixed and interpretable routing decision. Our work studies this trade-off explicitly.

We formalize this selection problem as regime-conditioned semantic routing. A deployment regime r defines the operational context: the expected prompt distribution, a representative calibration set, and an explicit rule defining what constitutes risk. The framework maps text into an interpretable semantic space whose coordinates are stable across prompts, and the guardrail selects the coordinate with the highest average contribution to the regime-specific risk criterion as the target for intervention. Prompt-adaptive attribution and a fixed regime-level policy differ fundamentally in when they commit: the former computes the intervention locus dynamically for each input, while the latter pre-selects a single locus before inference. Compressing instance-specific risk profiles into one coordinate inevitably discards information; the central practical question is how much, and we treat that loss as our primary object of study. A fixed policy sacrifices dynamic flexibility, but it is deterministic, inspectable prior to deployment, and auditable as a static artifact.

Input-side guardrails face a structural choice about the unit of attribution. Token-level players make exact Shapley computation infeasible and yield attributions that do not correspond to editable components; keyword filters have editable units but do not account for how those units contribute to the risk criterion. A fixed semantic basis supplies both at once: the players are high-level functional roles that are directly editable spans, and exhaustive attribution over them is exact and independent of prompt length. This is a property of the basis rather than of the compression we study — at four roles, exact attribution is inexpensive whether it runs offline or per prompt — and what evaluating the game at the regime level adds is a static policy artifact suitable for pre-deployment review.

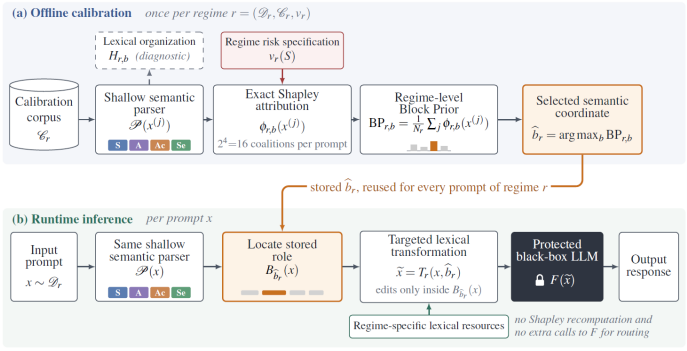


Figure 1: The Block Prior pipeline.

We introduce **Block Prior (BP)** by instantiating this intervention space with four roles: SUBJECT, ARGUMENT, ACTION, and SETTING. A constrained shallow parser maps each prompt to these shared coordinates, establishing a direct correspondence between the players of the attribution game and editable input spans. The cardinality of this space fixes the attribution budget: with $|\mathcal{B}| = 4$, exhaustive attribution requires $2^4 = 16$ coalition evaluations per prompt, independent of prompt length. Shapley computation is inexpensive at this resolution and remains so at inference time, which delimits our claim. BP does not make attribution affordable— with four roles, it already is— and the prompt-adaptive comparator we evaluate against likewise requires no access to the protected LLM since the characteristic function is input-side. What BP changes is *when* the routing decision is made and how many decisions take place: one per regime rather than one per prompt. BP uses exact prompt-level attributions during offline calibration, compresses them into a reusable regime-level semantic profile, and the evaluation that follows measures what that compression retains and what it gives up. As shown in Figure 1, the dominant-block decision is made offline and reused across incoming prompts. At runtime, the semantic parser instantiates the stored role in the current prompt, and a regime-specific lexical operator selectively transforms recognized evidence inside that role while preserving the remaining content. BP removes prompt-level attribution from the runtime routing decision while retaining prompt-level semantic parsing and targeted lexical transformation. We characterize the resulting trade-off through *route stability*, the persistence of the selected semantic

locus under input variation, and *lexical coverage*, the evidence accessible to the intervention within that locus.

1.1 CONTRIBUTIONS

① We introduce the notion of *regime-conditioned semantic routing* and study how much of the localization available from prompt-level attribution survives compression into a single regime-level coordinate. The four-role calibration yields a falsifiable prediction about held-out routing—that Age and Physical Appearance should depart from a fixed SUBJECT-ONLY policy, while Gender Identity should coincide with it—which we confirm on unseen prompts.
② We introduce a fixed SUBJECT–ARGUMENT–ACTION–SETTING intervention space that supports exact Shapley attribution during calibration and maps directly to editable prompt spans.
③ We quantify the compression gap directly: the fixed route recovers 77% of prompt-adaptive effectiveness, equivalently 41% of the headroom between a fixed Subject-only policy and per-prompt attribution.
④ We map the deployment boundary through controlled lexical and structural stress tests. We use a distinct security-operational risk specification as a controlled test of the representational scope of the shared interface. Section 5 states what the paper does not establish.

2 RELATED WORK

Actionable input attribution and black-box guardrails. Bias-mitigation research treats localization as part of the intervention problem, identifying bias-bearing knowledge, attention structures, or activation patterns that can be selectively modified (Chen et al., 2025a; Adiga et al., 2025; Hossain et al., 2025; Yu et al., 2025; Kim et al., 2025; Han et al., 2026). A complementary line uses explainability to expose implicit or latent demographic associations that aggregate bias scores may not resolve (Dutta et al., 2026; Pan et al., 2026; Himelstein et al., 2026; Romero-Arjona et al., 2026). The localization unit determines what evidence can be attributed and whether the resulting explanation can support a concrete intervention. Input-attribution methods provide a principled interface for this purpose (Ribeiro et al., 2016; Lundberg and Lee, 2017), and recent LLM studies connect attribution to explanation faithfulness and decision evidence (Admoni et al., 2026; Alon et al., 2026). For Shapley attribution, enlarging the player set expands the coalition space exponentially, motivating learned, sampled, and regression-based estimators (Alkhatib et al., 2025; Chen et al., 2025b; Witter et al., 2025; Kolpaczki and Hüllermeier, 2026). External guardrails increasingly implement safety and fairness as modular layers around LLMs, incorporating structured reasoning, adaptive model selection, policy conditioning, and input-side filtering (Wen et al., 2025; Li et al., 2025; Lee et al., 2025; Luo et al., 2025; Ding et al., 2026). Recent work studies generalization across prompt forms, safety policies, and previously unseen operating conditions (Sreedhar et al., 2025; Broomfield et al., 2025; Huang et al., 2026), while empirical analysis shows that guard decisions can be sensitive to lexical mutation, adversarial reformulation, and intent-preserving transformations (Bassani and Sanchez, 2025; Zhuang et al., 2025; Luo et al., 2026); for fairness, FLEX (Jung et al., 2025) formalizes this concern through controlled perturbations that expose instability beyond clean-input evaluation. BP sits at the intersection of these two lines. It uses semantically defined input roles as attribution units, giving each player a direct interpretation while retaining the exact Shapley computation in a compact player space. The same Subject–Argument–Action–Setting representation assigns contribution and determines where an input-side transformation can be applied.

3 BLOCK PRIOR: SHAPLEY-CALIBRATED SEMANTIC ROUTING

Given a deployment regime $r \in \mathcal{R}$, let $r = (\mathcal{D}_r, \mathcal{C}_r, v_r)$, where $\mathcal{D}_r$ denotes the deployment distribution, $\mathcal{C}_r = \{x^{(j)}\}_{j=1}^{N_r}$ a representative calibration corpus, and v_r an auditable input-side characteristic function. For semantic role $b \in \mathcal{B}$, its regime-level contribution is $\mu_{r,b} = \mathbb{E}_{x\sim\mathcal{D}_r}[\phi_{r,b}(x)]$. BP estimates this semantic contribution profile from $\mathcal{C}_r$ and uses it to establish a reusable routing coordinate for regime r.

Shared semantic coordinates. We use the fixed semantic basis $\mathcal{B} = \{\mathrm{S}, \mathrm{A}, \mathrm{Ac}, \mathrm{Se}\}$, corresponding to SUBJECT, ARGUMENT, ACTION, and SETTING. $\mathcal{B}$ acts as a *coarse-grained domain ontology*: it

defines the player set of the cooperative game through high-level functional roles, lifting attribution from individual token positions to a compact and actionable semantic space. The formulation may be generalized to richer domain ontology where expert-defined roles determine the intervention units. For each prompt x, a constrained shallow parser instantiates these coordinates as $\mathcal{P}(x) = (B_{\mathrm{S}}(x), B_{\mathrm{A}}(x), B_{\mathrm{Ac}}(x), B_{\mathrm{Se}}(x))$. The semantic labels remain fixed across prompts and regimes while their textual realizations adapt to the current input, so the basis provides a common attribution space across the calibration corpus in which every coordinate is an actionable input span, and risk contributions become comparable across prompts that share no surface form. Each $B_b(x)$ is an ordered set of occurrence-level spans copied verbatim from x, with $B_b(x) \cap B_{b'}(x) = \emptyset$ for $b \neq b'$: verbatim extraction preserves the surface evidence evaluated by the characteristic function, and non-overlap assigns each occurrence to a unique semantic coordinate. Appendix A gives the parsing instructions, role definitions, and the conservative local co-reference rules used by the optional anaphora-resolution stage.

Exact semantic attribution. The semantic roles define a four-player cooperative game for each prompt. In the social-bias instantiation, coalition $S \subseteq \mathcal{B}$ receives an aggregate lexical evidence score $W_r(S) = D_r(S) + C_r(S) + E_r(S)$, combining demographic, contextual/action-related, and setting-related evidence. The characteristic value is $v_r(S) = 1 - \exp[-0.25\, W_r(S)^2]$. The contribution assigned to role b is its exact Shapley value $\phi_{r,b}(x) = \sum_{S \subseteq \mathcal{B}\setminus\{b\}} \frac{|S|!\,(|\mathcal{B}|-|S|-1)!}{|\mathcal{B}|!}\,[v_r(S \cup \{b\}) - v_r(S)]$. With $v_r(\emptyset) = 0$, efficiency gives $\sum_{b\in\mathcal{B}} \phi_{r,b}(x) = v_r(\mathcal{B})$. The four-role representation enables exhaustive attribution with $2^4 = 16$ coalition evaluations per prompt, independently of prompt length, and can be paired with different characteristic functions, providing a shared intervention interface across regime-specific risk specifications.

Regime-level Block Prior. BP aggregates exact prompt-level Shapley attribution over the calibration corpus as $\mathrm{BP}_{r,b} = N_r^{-1} \sum_{j=1}^{N_r} \phi_{r,b}(x^{(j)})$, producing the regime-level profile $\mathbf{BP}_r = (\mathrm{BP}_{r,b})_{b\in\mathcal{B}}$. Each coordinate estimates the expected contribution $\mu_{r,b}$ of its corresponding semantic role under the deployment regime. The calibrated routing coordinate is $\widehat{b}_r = \arg\max_{b\in\mathcal{B}} \mathrm{BP}_{r,b}$, stored as the semantic intervention target for subsequent inputs from regime r. This turns the collection of prompt-level attribution profiles into a compact semantic representation of the calibration regime: the full vector $\mathbf{BP}_r$ retains the attribution structure across roles, while $\widehat{b}_r$ provides the coordinate used by the deployment policy.

Lexical organization of the calibrated coordinates. Prompt-level Shapley values determine the contribution profile, while the parser-induced corpora expose how each semantic coordinate is lexically realized in the calibration regime. For role b let $V_{r,b}$ denote its vocabulary over $\mathcal{C}_r$ and $p_{r,b}(t)$ the empirical token frequency. We summarize this lexical organization by the normalized entropy $H_{r,b} = -(\log_2 |V_{r,b}|)^{-1} \sum_{t\in V_{r,b}} p_{r,b}(t) \log_2 p_{r,b}(t)$, with $H_{r,b} = 0$ for $|V_{r,b}| = 1$. Because the normalization couples concentration to vocabulary size, $H_{r,b}$ should be read jointly with $|V_{r,b}|$. For $H_{r,b} < 1$ we *define* the coefficient $\alpha_{r,b}^{(j)} = \phi_{r,b}(x^{(j)})/(1 - H_{r,b})$, so that $\phi_{r,b}(x^{(j)}) = \alpha_{r,b}^{(j)}(1 - H_{r,b})$ holds identically. This is a definition, not a derivation: it carries no information beyond ϕ itself, and with $\widehat{\alpha}_{r,b} = N_r^{-1} \sum_j \alpha_{r,b}^{(j)}$ the arithmetic-mean calibration gives $\mathrm{BP}_{r,b} = \widehat{\alpha}_{r,b}(1 - H_{r,b}) = N_r^{-1} \sum_j \phi_{r,b}(x^{(j)})$. Entropy does *not* re-weight the deployed route, and $\widehat{b}_r$ is exactly the empirical maximizer of mean Shapley attribution. The factorization is a descriptive device that separates the magnitude of a coordinate's contribution from the lexical concentration of the evidence that produced it; $\widehat{\alpha}$ is an offline diagnostic and never a runtime quantity. Appendix B.4 reports its numerical conditioning, including sign behavior under negative Shapley contributions and instability as $H_{r,b} \to 1$. The lexical organization captured by $H_{r,b}$ provides a pre-deployment characterization of the evidence associated with each semantic coordinate. Because the runtime transformation acts on recognized lexical evidence within the routed role, this structure also provides a natural basis for analyzing sensitivity to incomplete lexical coverage. We therefore use role-conditioned entropy as an offline descriptor of the calibrated coordinates and examine its relationship with coverage-related robustness in the experimental analysis.

Regime-conditioned intervention. Calibration fixes the selected semantic coordinate $\widehat{b}_r$ for regime r. Given an incoming prompt x, the semantic parser produces $\mathcal{P}(x)$, and the stored coordinate identifies its prompt-specific realization $B_{\widehat{b}_r}(x)$. The calibrated prior determines *which semantic role* is targeted, and prompt-specific parsing determines *where that role occurs* in the current input. For the social-bias intervention, the transformation $\widetilde{x} = T_r(x, \widehat{b}_r)$ applies the regime-specific lexical operator inside $B_{\widehat{b}_r}(x)$: recognized risk-bearing units are transformed while unmatched material and all remaining semantic roles are preserved, and the resulting prompt is passed to the protected model as $F(\widetilde{x})$. The BP coordinate determines the intervention locus, and regime-specific lexical resources determine the actionable units within it.

Exact Shapley attribution, BP calibration, vocabulary statistics, and lexical-entropy diagnostics are computed offline. The experiments use precomputed semantic annotations to apply the same decomposition consistently across conditions. In an online deployment, the same parser localizes $B_{\widehat{b}_r}(x)$ for each incoming prompt while the calibrated coordinate $\widehat{b}_r$ is reused directly, so runtime inference combines semantic localization with the stored regime-level policy without prompt-level Shapley recomputation or additional protected-model calls for routing.

4 Experimental Evaluation

We evaluate BP by testing whether the semantic structure learned during offline calibration translates into effective mitigation on unseen prompts. We first quantify held-out debiasing effectiveness, then analyze the structure supporting the calibrated route, its transfer across datasets, its downstream effect across heterogeneous black-box LLMs, and its behavior under controlled lexical and structural perturbations.

Experimental Setup. We use the Age, Gender Identity, and Physical Appearance subsets of BBQ (Parrish et al., 2022) as the primary social-bias regimes, each with an offline-calibrated BP profile over the four roles. Held-out effectiveness is evaluated over five independent $80/20$ partitions (seeds 0–4); the BP profile is fixed before evaluation and reused across partitions, so variation reflects applying the same regime-level route to different unseen samples. At runtime, BP selects the stored role, the parser locates its realization, and targeted debiasing replaces recognized risk-bearing lexical units within that role with `<NEUTRAL>`, preserving unmatched material and all remaining roles. For $v_r^{\mathrm{orig}}(x) > 0$ we report $\mathrm{BR}(x) = 100[v_r^{\mathrm{orig}}(x) - v_r^{\mathrm{int}}(x)]/v_r^{\mathrm{orig}}(x)$, with larger BR indicating a greater reduction of the input-side risk evidence measured by v_r, and evaluate prompt preservation independently with Sentence-BERT cosine similarity (`all-MiniLM-L6-v2`). Two controlled references share BP's neutralization operator: *Subject-only* always routes to Subject, and *Prompt-Shapley* recomputes exact four-role attribution for each evaluation prompt. For cross-dataset evaluation, we freeze the BBQ Gender Identity P1 profile and transfer it to DiscrimEval (Tamkin et al., 2023) without refitting any BP parameters, retaining for the attribution comparison the 25 of 75 prompt-level profiles closest in cosine distance to the frozen profile. To test downstream behavior, the 58 BBQ and all 75 DiscrimEval prompts are submitted under Original and BP-transformed conditions to four target-model families and scored by two independent judges on a 0–3 bias rubric in a complete target–judge crossing; judges play no role in calibration, routing, or transformation. Runtime robustness is characterized through additive padding, OOV lexical substitution, lexical-resource pruning, and relocation of risk-bearing evidence to a non-routed role. Finally, we test representational scope on a ProAgentBench-inspired (Tang et al., 2026) synthetic SOC suite of 150 prompts under a security-specific risk function, with no SOC-specific BP prior calibrated. Full protocols, model rosters, the judge rubric, and corpus construction are given in Appendices D–H.

Prior Calibration and Semantic Resolution. All variants use the same semantic coordinates. P1 retains the raw parser spans, P2 adds selective local Subject–Argument referent resolution, P3 applies lexical normalization with $\tau = 2$, combining anonymization, generalization of rare expressions, and outlier suppression, and P4 applies referent resolution followed by the same normalization. Table 1 reports held-out MAE between prompt-level exact Shapley attribution and the stored regime profile, averaged over the four roles.

These values are *not* directly comparable across pipelines. Normalization changes the scale of the attribution profile itself—under P3, Age Action falls from .053 to .003 and Age Setting from

.023 to .000 (Table 16)—so a pipeline that compresses ϕ toward zero reduces the absolute error of a constant predictor without necessarily predicting better. Dividing each pipeline's MAE by the corresponding total attribution mass reverses the raw ordering for Age and demotes P1 for Gender Identity. Table 1 is therefore a sensitivity diagnostic showing that preprocessing is regime-dependent and should not be used to select a representation. For the cross-dataset experiment, we freeze the Gender Identity P1 profile on two scale-free grounds: Gender Identity is the regime in which BP and the SUBJECT-ONLY reference instantiate the same transformation, so the transferred profile carries an internal control, and P1 is the raw representation, so transfer is not confounded by a normalization vocabulary fitted to the source regime.

Table 1: Preprocessing sensitivity.

	Preprocessing		Held-out MAE ↓		
Pipe.	Referent	Norm.	Age	Gender Id.	Phys. App.
P1	✗	✗	.085	**.081**	**.088**
P2	✓	✗	.086	.083	.099
P3	✗	✓	.065	.098	.107
P4	✓	✓	**.059**	.097	.120

Figure 2 highlights the lexical effect of normalization. Under P1, normalized role-conditioned entropy ranges from 0.894 to 0.966; normalization reorganizes this support selectively rather than uniformly, with Gender Identity Subject falling from 0.894 to 0.673 and Age Setting from 0.937 to 0.782, while Gender Identity Action and Setting rise slightly.

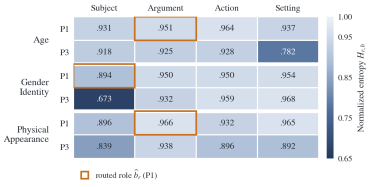


Figure 2: Role-conditioned normalized entropy under P1 and P3.

Finally, we test whether semantic resolution is necessary for routing. A matched three-player ablation merges Subject and Argument into a single ENTITY coordinate while holding the prompts, preprocessing, characteristic function, and exact Shapley procedure fixed.

Table 2 shows that the three-role game routes all three regimes to ENTITY. Separating Subject from Argument recovers Argument for Age and Physical Appearance, and Subject for Gender Identity, preserving regime-specific intervention loci at a modest offline increase from $2^3 = 8$ to $2^4 = 16$ coalition evaluations per prompt. This yields a prediction for the held-out intervention: BP should depart from SUBJECT-ONLY in Age and Physical Appearance and coincide with it in Gender Identity.

Debiasing Effectiveness. We test that prediction on unseen prompts using the fixed regime-level profiles. Before transformation, the three regimes exhibit comparable input-side risk magnitudes under their characteristic functions (0.455 ± 0.056 for Age, 0.424 ± 0.056 for Gender Identity, 0.437 ± 0.050 for Physical Appearance), so the held-out contrasts are not driven by differences in baseline risk. Table 3 reports Bias Reduction (BR) for BP and the two controlled routing references.

Table 2: Semantic-resolution ablation.

Regime	3-role	4-role
Age	**Entity** (.682)	**Argument** (.522)
Gender Id.	**Entity** (.778)	**Subject** (.545)
Phys. App.	**Entity** (.894)	**Argument** (.435)

The held-out behavior follows the calibrated routes. Age and Physical Appearance both route to Argument and depart from SUBJECT-ONLY by 39.2 and 6.5 percentage points, respectively, while Gender Identity routes to Subject; thus, BP and SUBJECT-ONLY instantiate the same semantic decision and obtain identical BR.

Table 3: Held-out debiasing.

Regime	BP BR	Subject-only	Prompt-Shapley	Sim.
Age	61.4 ±4.7	22.2 ±6.8	73.6 ±6.2	.950
Gender Id.	62.0 ±7.3	62.0 ±7.3	79.9 ±7.1	.916
Phys. App.	31.2 ±11.1	24.7 ±3.5	47.6 ±11.0	.937
Pooled	54.5 ±4.4	43.4 ±3.4	70.7 ±2.9	.929

The prediction issued by the four-role calibration is therefore confirmed in all three regimes, including the negative case. Pooled, BP reaches $54.5 \pm 4.4\%$ against $43.4 \pm 3.4\%$ for SUBJECT-ONLY, a difference of $+11.15$ points with bootstrap 95% CI $[8.28, 14.28]$; per-regime BP intervals are $[58.0, 65.1]$, $[56.0, 67.3]$, and $[21.6, 38.8]$. The paired permutation test over the five partition-level pairs gives $p = .0625$; because only five paired observations are available, the attainable two-sided permutation resolution is coarse, and we characterize this comparison through effect magnitude and bootstrap uncertainty rather than a conventional significance claim.

Prompt-Shapley measures how much localization remains available when attribution is recomputed per prompt. It reaches $70.7 \pm 2.9\%$ pooled BR, 16.2 points above BP: the fixed route recovers 77% of prompt-adaptive effectiveness pooled, and 83%, 78%, and 66% for Age, Gender Identity, and Physical Appearance. Part of this gap is definitional rather than empirical, since a per-prompt argmax cannot be worse than a fixed argmax under the same objective. BP preserves high prompt similarity (0.929 ± 0.007, against 0.943 ± 0.007 for SUBJECT-ONLY and 0.920 ± 0.007 for Prompt-Shapley). Offline regime calibration recovers substantial localization beyond a fixed semantic route while retaining most of the original prompt content, and a fixed policy forfeits roughly a quarter of what per-prompt attribution can deliver.

Cross-Dataset Transfer. We next test whether the same semantic structure persists when the prompt distribution changes. DiscrimEval alters both prompt construction and the surface realization of demographic information, so the transfer question concerns the preservation of the semantic ordering learned on the source regime rather than the numerical invariance of its magnitudes. Table 4 shows a change in scale: Subject rises from 0.271 in the frozen BBQ profile to 0.415 mean exact Shapley attribution on DiscrimEval and Argument from 0.102 to 0.267, while Action and Setting remain much smaller. Despite these shifts, the ordering is preserved, SUBJECT > ARGUMENT ≫ ACTION ≈ SETTING, so the coordinate selected by the source profile remains dominant within the selected target subset. Recomputing the target profile with the median, trimmed mean, Winsorized mean, Huber and Hodges–Lehmann estimators yields the same ordering, so the result is not an artifact of arithmetic-mean aggregation. The experiment supports *ordinal agreement* of the frozen profile on a target-conditioned subset; because the subset is selected using target attribution, it does not establish unrestricted cross-dataset transfer (Appendix E).

Table 4: Frozen-prior transfer from BBQ to DiscrimEval (DE).

Role	BBQ BP	$\bar{\phi} \pm \sigma$	Shift
Subject	.271	.415 $\pm$.217	+53%
Argument	.102	.267 $\pm$.230	+163%
Action	.033	.016 $\pm$.044	−51%
Setting	.008	.013 $\pm$.039	+63%

Output-Side Black-Box Behavior The preceding experiments evaluated BP through its explicit input-side characteristic function. We also tested if BP-transformed prompts also change the responses of black-box LLMs. Four target families—Gemini 2.5 Flash, GPT-4o-mini, Llama 3.3 70B-Instruct, and Qwen3 235B–A22B-Instruct-2507—each receive the 58 BBQ and 75 DiscrimEval prompts under Original and BP-transformed conditions, and Gemini 2.5 Flash and GPT-4o-mini independently judge every response on a 0–3 bias scale. For each regime–target–judge cell we report $\Delta = \text{Original} - \text{BP}$, so $\Delta > 0$ denotes lower mean judged bias after transformation.

Table 5: Output-side summary.

Regime	Mean Δ	Lower	Tie	Higher	$p < .05$
Age	.34	8	0	0	1/8
Gender Id.	.14	7	0	1	2/8
Phys. App.	.07	4	3	1	0/8
DiscrimEval	.38	4	1	3	4/8
Total	–	**23**	**4**	**5**	**7/32**

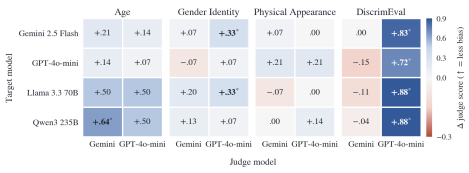


Figure 3: Cross-model behavioral effect.

Across the 32 cells, mean judged bias decreases in 23, is unchanged in 4, and increases in 5 (Figure 3 and Table 5), with positive-Δ cells in each target family and the strongest directional consistency in Age and Gender Identity. Seven cell-level paired comparisons have nominal $p < .05$; these are exploratory and unadjusted. Because the individual cell-level tests are underpowered, we summarize the direction of the effect across cells: of the 28 non-tied cells, 23 favor the transformed condition, and a two-sided sign test gives $p < 10^{-3}$. Cells share prompts, targets, and judges, so this is an aggregate directional summary rather than a confirmatory result. The crossing also exposes evaluator dependence: on DiscrimEval, GPT-4o-mini-as-judge reports $\Delta = .827, .720, .880, .880$ across the four targets whereas Gemini 2.5 Flash-as-judge reports $.000, -.147, -.107, -.040$, and their Original score ranges differ markedly (1.59–1.84 versus 0.17–0.29), limiting the reduction under the lower-scoring evaluator. Finally, BP modifies $70/75$ DiscrimEval prompts (93.3%) while extracted decisions remain stable in 95.4%, 96.0%, 86.7% and 90.7% of responses across the four targets; these rates measure decision stability, not correctness.

Robustness Analysis. We evaluate BP under perturbations that preserve the semantic route while altering either the surrounding content or the lexical realizations within the selected role. Inspired by FLEX (Jung et al., 2025), we construct BP-specific additive-padding and OOV lexical-substitution tests: a prompt is BP-active when the baseline BR ≥ 0.01, and an attack succeeds when the perturbed BR falls below 0.01. Padding adds content without removing recognized evidence from the routed role, whereas OOV Text Attack replaces matched units with lexical alternatives that remain in the same semantic coordinate but fall outside the current resources. Neutral padding yields 0% ASR in all BBQ regimes; adversarial padding stays between 0% and 1.4% (pooled 0.85%, bootstrap 95% CI $[0.0, 2.6]$) and 0% on DiscrimEval. OOV substitution is substantially more disruptive, reaching 40.2% partition-averaged pooled ASR on BBQ (95% CI $[33.9, 46.5]$) and 81.4% on DiscrimEval, with pooled mean BR falling from 0.648 at baseline to 0.614 under adversarial padding and 0.386 under OOV substitution. These failures track the loss of lexical recognition: across the BP-active BBQ observations, the Text Attack changes neither the stored route nor semantic block membership, and 93 of the 97 successful attacks are cases in which no recognized unit remains in the routed role after substitution (Appendix G). Two ablations support this reading. Complete routed-role erasure has nearly the same OOV ASR as targeted debiasing (40.5% vs. 40.1%) while reducing semantic similarity from 0.893 to 0.751, so broader editing does not buy lexical robustness; and random 20% resource pruning reduces BR by 4.8, 8.3 and 2.1 points for Age, Gender Identity and Physical Appearance while leaving the stored role fixed. The pruning result is also consistent with the lexical concentration measured offline for the routed coordinates. Under P1, their normalized entropies are 0.894 for Gender Identity Subject, 0.951 for Age Argument, and 0.966 for Physical Appearance Argument, matching the ordering of the corresponding BR decreases under 20% random resource pruning (8.3, 4.8, and 2.1 points). With three regimes, we treat this as a consistency observation rather than an estimated relationship; the ordering does not extend to OOV substitution. Lexical coverage is therefore the main robustness boundary of the deployed intervention *when the calibrated semantic locus remains valid*.

Table 6: Additive and lexical robustness.

Regime	Adv. padding	OOV text
Age	0.0	62.2 [46.7, 73.6]
Gender Id.	0.9	36.0 [26.6, 46.4]
Phys. App.	1.4	27.3 [19.0, 35.6]
DiscrimEval	0.0	81.4

A third perturbation tests that proviso directly. Block Relocation moves risk-bearing evidence from the BP-selected role to a non-selected coordinate while the stored profile remains fixed, and it reaches 100% ASR in all three BBQ regimes and on the DiscrimEval transfer (Appendix G.7). Under an adversary able to freely restructure the prompt, the fixed route offers no protection, while the prompt-adaptive alternative does. Recomputing exact attribution after relocation and routing adaptively recovers BR of 0.874, 0.865, and 0.809. The evidence remains detectable and removable; only the stored decision is stale. A fixed regime-level policy is appropriate for regime-level mitigation under a benign input distribution rather than as an adversarially robust guardrail. Structure-aware fallback, in which the stored route is re-derived when the coordinate no longer carries the recognized evidence, is the natural remedy and lies outside the present design.

Cross-Regime Semantic Scope. The social-bias experiments concentrate attribution in Subject or Argument, leaving open whether the four-role game is intrinsically biased toward entity-adjacent coordinates or instead reflects where the specified risk evidence sits. We pair the same decomposition with a security-operational characteristic function on a ProAgentBench-inspired (Tang et al., 2026) synthetic SOC suite of 150 prompts across six subtypes, replacing the social-bias lexical resources with a MITRE ATT&CK-informed[1] knowledge base that assigns substantial risk evidence to Action terms while leaving Subject filled with neutral operational referents. Exact four-player Shapley attribution identifies Action as dominant in $139/150$ prompts (92.7%), with mean attribution 0.665 for Action against 0.058, 0.047, and 0.011 for Subject, Setting, and Argument. The two less uniform subtypes contain defensive actions carrying negative weights, which destabilize the prompt-level argmax without changing the

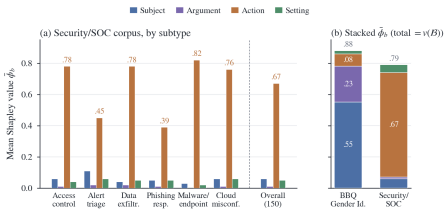


Figure 4: Cross-regime semantic attribution.

[1] https://attack.mitre.org

aggregate Action-dominant profile. Figure 4 contrasts this with the Subject- and Argument-dominant social-bias profiles: changing the risk specification produces a qualitatively different attribution pattern over the same four coordinates, so exact attribution reflects the structure encoded by the regime-specific characteristic function (Appendix H).

5 Discussion and Conclusion

BP compresses prompt-level attribution into a regime-level route, trading instance-specific adaptation for a fixed deployment policy. The evaluation supports three claims. First, the four-role resolution carries intervention-relevant information: the calibration predicts, before any held-out prompt is seen, which regimes should depart from a fixed SUBJECT-ONLY policy, and the held-out intervention confirms that prediction in all three regimes, including the negative case. Second, the compression is costly but not prohibitive: the fixed route retains 77% of prompt-adaptive bias reduction while exceeding the fixed alternative by 11.2 points, with prompt content largely preserved. Third, the compression has two failure modes with different implications. Depletion of recognizable evidence inside an otherwise correct route degrades the intervention gradually and is a coverage problem; relocation of evidence out of the routed role defeats it entirely and is an assumption failure, recoverable only by re-deriving the route.

The contribution is not computational. At $|\mathcal{B}| = 4$, exact prompt-level attribution costs 16 coalition evaluations and is affordable at inference time. No call to the protected model is required for routing, but the same is true of the prompt-adaptive comparator, whose characteristic function is likewise input-side. Thus, this property distinguishes input-side guardrails from parameter- and activation-level methods rather than distinguishing BP from its comparator. What BP removes at runtime is the per-prompt attribution decision itself: a single stored coordinate replaces 16 coalition evaluations and an *argmax*. Establishing a cost argument would require both a measured deployment cost model and a resolution at which $2^{|\mathcal{B}|}$ is genuinely prohibitive.

Several extensions follow. Coverage- and structure-aware routing can trigger selective recalibration or prompt-level attribution when the routed coordinate no longer carries the recognized evidence, while regime-shift detection can identify when a stored route is no longer representative. Higher-cardinality role systems can test whether the compression gap widens or narrows as the coordinate space grows. Broader evaluation can add demographic dimensions, intersectional and multi-turn settings, and longer contexts; cross-dataset transfer should be tested without target-conditioned subset selection; future characteristic functions can reduce dependence on curated lexical resources while retaining auditability; and operational-security evaluation should extend the SOC study to calibrated routing and intervention.

What this paper does not establish. Several limitations bind the claims above; Appendix I states each in full. Bias Reduction is defined on the same characteristic function v_r that defines the Shapley game and selects the route, so the held-out BR comparison establishes that the calibrated route behaves as the calibration predicts, not that it reduces bias as independently measured; the output-side judge evaluation avoids this circularity but is weaker. We report no latency or throughput cost for any condition and make no efficiency claims. Held-out results are summarized over five partition-level observations, which cap the attainable two-sided permutation resolution at $p = .0625$; a prompt-level analysis with clustering over prompts would provide greater power. The semantic parser, the lexical resources, and the functional form of v_r are taken as given. We do not compare against a route selected from aggregate per-role lexical evidence without coalition evaluation, against a single route shared across all regimes, or against external guardrail systems; the first is the most informative omission since W_r is additive over roles and v_r is a monotone transform of it, so the Shapley argmax and the raw per-role argmax coincide wherever v_r is concave and may differ only in the convex region below $W_r = \sqrt{2}$. We do not report how many calibration prompts are needed for the selected route to stabilize. Finally, the DiscrimEval result concerns ordinal agreement on a target-conditioned subset, and the SOC study establishes representational portability rather than operational-security effectiveness.

AI use statement. Generative AI tools were used to support method implementation and debugging, dataset preprocessing and OOV adversarial simulation, synthetic data generation, language editing, and stylistic refinement of figures. All AI-assisted work was reviewed and validated by the authors, who remain fully responsible for the final content.

REFERENCES

Rishabh Adiga, Besmira Nushi, and Varun Chandrasekaran. Attention speaks volumes: Localizing and mitigating bias in language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26403–26423. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1281.

Sahar Admoni, Ofra Amir, Assaf Hallak, and Yftah Ziser. Aligning what llms do and say: Towards self-consistent explanations. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 987–1003. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.49.

Amr Alkhatib, Roman Bresson, Henrik Boström, and Michalis Vazirgiannis. Prediction via shapley value regression. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 1056–1101. PMLR, 2025.

Bar Alon, Itamar Zimerman, and Lior Wolf. Faithful serum: Mitigating the faithfulness gap in textual explanations of llm decisions via attribution guidance. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6622–6645. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.300.

Arash Asgari, Huan Wu, Amirreza Naziri, Mojtaba Kolahdouzi, and Laleh Seyyed-Kalantari. Quantifying metric and model agreement in bias evaluation of large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16868–16933, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.769. URL https://aclanthology.org/2026.acl-long.769/.

Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky, and Thomas L. Griffiths. Explicitly unbiased large language models still form biased associations. *Proceedings of the National Academy of Sciences*, 122(8):e2416228122, 2025. doi: 10.1073/pnas.2416228122. URL https://doi.org/10.1073/pnas.2416228122.

Elias Bassani and Ignacio Sanchez. On guardrail models' robustness to mutations and adversarial attacks. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 16995–17006. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.922.

Julius Broomfield, Tom Gibbs, George Ingebretsen, Ethan Kosak-Hine, Tia Nasir, Jason Zhang, Reihaneh Iranmanesh, Sara Pieri, Reihaneh Rabbany, and Kellin Pelrine. The structural safety generalization problem. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22134–22173. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.1142.

Ruizhe Chen, Yichen Li, Jianfei Yang, Yang Feng, Joey Tianyi Zhou, Jian Wu, and Zuozhu Liu. Identifying and mitigating social bias knowledge in language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 651–672. Association for Computational Linguistics, 2025a. doi: 10.18653/v1/2025.findings-naacl.39.

Tyler Chen, Akshay Seshadri, Mattia Jacopo Villani, Pradeep Niroula, Shouvanik Chakrabarti, Archan Ray, Pranav Deshpande, Romina Yalovetzky, Marco Pistoia, and Niraj Kumar. A unified framework for provably efficient algorithms to estimate shapley values. In *Advances in Neural Information Processing Systems*, volume 38, 2025b. doi: 10.52202/085713-5500.

Xin Chen, Yarden As, and Andreas Krause. Learning safety constraints for large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 7664–7685. PMLR, 2025c. URL https://proceedings.mlr.press/v267/chen25b.html.

Yuxing Chen, Guoqing Luo, Zijun Wu, and Lili Mou. Multi-persona thinking for bias mitigation in large language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 27895–27909, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.1389. URL https://aclanthology.org/2026.findings-acl.1389/.

Zhihao Ding, Jinming Li, Ze Lu, and Jieming Shi. FlexGuard: Continuous risk scoring for strictness-adaptive llm content moderation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5825–5851. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.263.

Arka Dutta, Reza Fayyazi, Shanchieh Yang, and Ashiqur R. KhudaBukhsh. How can you tell if your large language model could be a closet antisemite? an explainability-based audit framework for implicit bias. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 38404–38412, 2026. doi: 10.1609/aaai.v40i45.41181.

Akram Elbouanani, Aboubacar Tuo, and Adrian Popescu. A scalable entity-based framework for auditing bias in large language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 33823–33852, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.1689. URL https://aclanthology.org/2026.findings-acl.1689/.

Tingxu Han, Wei Song, Ziqi Ding, Ziming Li, Chunrong Fang, Yuekang Li, Dongfang Liu, Zhenyu Chen, and Zhenting Wang. Debiasing llms by masking unfairness-driving attention heads. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 37980–37992. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1893.

Rom Himelstein, Amit LeVi, Brit Youngmann, Yaniv Nemcovsky, and Avi Mendelson. Silenced biases: The dark side llms learned to refuse. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 37452–37461, 2026. doi: 10.1609/aaai.v40i44.41078.

Prommy Sultana Hossain, Chahat Raj, Ziwei Zhu, Jessica Lin, and Emanuela Marasco. Toward inclusive language models: Sparsity-driven calibration for systematic and interpretable mitigation of social biases in llms. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 2475–2508. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.134.

Ruixuan Huang, Xunguang Wang, Zongjie Li, Daoyuan Wu, and Shuai Wang. Guidedbench: Measuring and mitigating the evaluation discrepancies of in-the-wild llm jailbreak methods. In *International Conference on Learning Representations*, 2026.

Md Asiful Islam and Mihai Surdeanu. A lightweight explainable guardrail for prompt safety. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 43573–43591, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.2017. URL https://aclanthology.org/2026.acl-long.2017/.

Dahyun Jung, Seungyoon Lee, Hyeonseok Moon, Chanjun Park, and Heuiseok Lim. Flex: A benchmark for evaluating robustness of fairness in large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3606–3620. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-naacl.199.

Seorin Kim, Dongyoung Lee, and Jaejin Lee. KLAAD: Refining attention mechanisms to reduce societal bias in generative language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15313–15334. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.774.

Patrick Kolpaczki and Eyke Hüllermeier. Approximation algorithms for the shapley value: Taxonomy and properties. In *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence*, pages 7873–7881, 2026. doi: 10.24963/ijcai.2026/874.

Tom A. Lamb, Adam Davies, Alasdair Paren, Philip H. S. Torr, and Francesco Pinto. Focus on this, not that! steering LLMs with adaptive feature specification. In *Proceedings of the 42nd International Conference on Machine Learning*. PMLR, 2025. URL https://openreview.net/forum?id=rbI5mOUA8Z.

Seanie Lee, Dong Bok Lee, Dominik Wagner, Minki Kang, Haebin Seong, Tobias Bocklet, Juho Lee, and Sung Ju Hwang. Saferoute: Adaptive model selection for efficient and accurate safety guardrails in large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2053–2069. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.105.

Hao Li, Xiaogeng Liu, Ning Zhang, and Chaowei Xiao. Piguard: Prompt injection guardrail via mitigating overdefense for free. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30420–30437. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1468.

Minqian Liu, Ioana Baldini, David Rabinowitz, David S. Rosenberg, Sebastian Gehrmann, and Mark Dredze. Domain generalizable AI guardrails with augmented policy training. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16452–16469, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.748. URL https://aclanthology.org/2026.acl-long.748/.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Weidi Luo, Shenghong Dai, Xiaogeng Liu, Suman Banerjee, Huan Sun, Muhao Chen, and Chaowei Xiao. AGrail: A lifelong agent guardrail with effective and adaptive safety detection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8104–8139. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.399.

Xiaohao Luo, Ying Wei, and Zhijun Li. Safety guardrails of large language models are vulnerable to value-driven adversarial prompting. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 27238–27255. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1357.

Afrozah Nadeem, Mark Dras, and Usman Naseem. Fairness evaluation and inference level mitigation in LLMs. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 29048–29065, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.1452. URL https://aclanthology.org/2026.findings-acl.1452/.

Jinhao Pan, Chahat Raj, and Ziwei Zhu. Bias association discovery framework for open-ended llm generations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 32637–32645, 2026. doi: 10.1609/aaai.v40i38.40541.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, 2022.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2016. doi: 10.1145/2939672.2939778.

Miguel Romero-Arjona, José A. Parejo, Juan C. Alonso, Ana B. Sánchez, Aitor Arrieta, and Sergio Segura. Meta-fair: Ai-assisted fairness testing of large language models. *Information and Software Technology*, 194:108075, 2026. doi: 10.1016/j.infsof.2026.108075.

Makesh Narsimhan Sreedhar, Traian Rebedea, and Christopher Parisien. Safety through reasoning: An empirical study of reasoning guardrail models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 21862–21880. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.1193.

Zhouhao Sun, Zhiyuan Kan, Xiao Ding, Li Du, Bibo Cai, Yang Zhao, Bing Qin, and Ting Liu. Large language models are still misled by simple bias ensembles. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 19444–19455, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.971. URL https://aclanthology.org/2026.findings-acl.971/.

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.

Yuanbo Tang, Huaze Tang, Tingyu Cao, Lam Nguyen, Anping Zhang, Xinwen Cao, Chunkang Liu, Wenbo Ding, and Yang Li. ProAgentBench: Evaluating LLM agents for proactive assistance with real-world data, 2026. URL https://arxiv.org/abs/2602.04482.

Xiaofei Wen, Wenxuan Zhou, Wenjie Jacky Mo, and Muhao Chen. Thinkguard: Deliberative slow thinking leads to cautious guardrails. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13698–13713. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.704.

R. Teal Witter, Yurong Liu, and Christopher Musco. Regression-adjusted monte carlo estimators for shapley values and probabilistic values. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0715.

Sangwon Yu, Jongyoon Song, Bongkyu Hwang, Hoyoung Kang, Sooah Cho, Junhwa Choi, Seongho Joe, Taehee Lee, Youngjune Gwon, and Sungroh Yoon. Correcting negative bias in large language models through negative attention score alignment. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9979–10001. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.503.

Xu Zhang and Xiaojun Wan. RST-guarder: Enhancing long-context robustness for safeguards via RST parsing and probabilistic inference. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22397–22413, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.1025. URL https://aclanthology.org/2026.acl-long.1025/.

Jun Zhuang, Haibo Jin, Ye Zhang, Zhengjian Kang, Wenbin Zhang, Gaby G. Dagher, and Haohan Wang. Exploring the vulnerability of the content moderation guardrail in large language models via intent manipulation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 2147–2160. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.114.

APPENDIX

Contents

A Pipeline Specification 14
B Alpha Calibration and BP Estimation Diagnostics 16
C Block Granularity Ablation 19
D Debiasing Effectiveness: Full Held-Out Results 21
E External Validation on DiscrimEval 23
F Output-Side LLM-as-a-Judge Evaluation 24
G Adversarial Stress Tests and Dictionary Sensitivity 26
H Cross-Regime Security Evaluation 30
I What This Paper Does Not Establish 32

A PIPELINE SPECIFICATION

This section specifies the semantic parser and the four preprocessing configurations used in the experiments. All variants share the same Subject–Argument–Action–Setting coordinate system and differ only in whether selective referent resolution and lexical normalization are applied before role-conditioned vocabulary construction and input-side scoring. The preprocessing variants therefore alter the representation supplied to calibration without changing the semantic player space itself.

Table 7: **Preprocessing configurations.** All pipelines share the same semantic parser.

Pipeline	Label	Referent resolution	Lexical normalization
P1	Raw baseline	–	–
P2	Raw + referent	✓	–
P3	Normalized	–	✓
P4	Normalized + referent	✓	✓

A.1 STAGE 0 — BLOCK EXTRACTION VIA SHALLOW PARSER

For input record x_j with textual context c_j, the parser $\mathcal{P}$ returns $\widetilde{x}_j = (x_j, \mathrm{BLOCKS}_j, \mathrm{REFERENT_MAP}_j)$, where $\mathrm{BLOCKS}_j = (B_{\mathrm{S}}^{(j)}, B_{\mathrm{A}}^{(j)}, B_{\mathrm{Ac}}^{(j)}, B_{\mathrm{Se}}^{(j)})$. The four ordered lists correspond to SUBJECT, ARGUMENT, ACTION, and SETTING. Subject contains the focal nominal referent; Argument contains descriptive, predicative, or complement material associated with that referent; Action contains verbal or event expressions; and Setting contains situational or contextual material.

In the reported implementation, $\mathcal{P}$ is instantiated by dataset-specific zero-shot prompting of GPT-4 under a constrained structured-output schema. The parser is auxiliary and external to the protected target model.

Occurrence-level extraction. Mentions are represented by occurrence rather than by unique surface form. Repeated strings at different positions therefore remain distinct units. This preserves multiplicity in the instantiated prompt and makes subsequent replacement position-specific.

Verbatim grounding. Every extracted unit must occur verbatim in the original input. The parser does not paraphrase extracted content, allowing lexical matching, coalition construction, and runtime neutralization to operate on explicit input spans.

Non-overlapping assignment. Each occurrence is assigned to at most one semantic role. This prevents the same textual evidence from entering multiple players of the Shapley game and provides a unique intervention locus. Nominal units are kept compact while retaining identity-bearing modifiers that are part of the expressed mention.

A.2 Stage 1 — Anaphora Resolution (P2 and P4 only)

The auxiliary $\mathrm{REFERENT_MAP}_j$ stores conservative local co-reference relations among nominal occurrences. Subject and Argument units are associated with either a local referent identifier R_k or `null`. A link is introduced only when the relation is explicit and locally unambiguous, such as a repeated nominal mention or a clear noun–pronoun relation. Referential identity is represented separately from semantic role because the same entity can occur in different linguistic positions while BP attributes evidence to roles.

Pipelines P2 and P4 use these links to canonicalize selected anaphoric expressions before vocabulary construction and lexical matching. If an occurrence t in Subject or Argument has an unambiguous referent $R(t)$, the resolver substitutes the corresponding canonical local referent for the anaphoric surface form. Action and Setting are unchanged by this stage. The procedure propagates only information explicitly supported by the local referent map; it does not infer demographic attributes absent from the input.

The Subject–Argument distinction remains useful even when two spans refer to the same entity. Referent canonicalization may modify the lexical realization of Argument while Subject and Argument remain distinct Shapley coordinates. If the roles are merged into Entity, co-reference can still be resolved lexically, but the referent–predicate distinction is no longer available to the attribution game; Appendix C isolates this effect.

A.3 Stage 2 — Language Normalization (P3 and P4 only)

Pipelines P3 and P4 normalize the lexical realizations used for role-conditioned vocabulary construction and matching. The objective is to expose how calibration changes when idiosyncratic surface forms are mapped to a smaller canonical vocabulary. The procedure uses threshold $\tau = 2$ and contains three operations.

S1: anonymization. Personal names are replaced by generic category tokens independently of their frequency. This removes identity-specific surface variation while preserving a generic referential marker.

S2: generalization. For a lexical unit t whose empirical frequency is below τ, a regime-specific mapping may replace t by a broader canonical form $g(t)$. Multiple rare expressions can therefore contribute to the same normalized lexical entry. The mapping is defined from the lexical resources used by the corresponding social-bias regime.

S3: residual outlier suppression. A remaining sub-threshold unit for which no suitable generalization is defined is mapped to a common fallback placeholder. This prevents singleton and low-frequency residual forms from each defining separate vocabulary entries.

These operations change the empirical distribution $p_{r,b}(t)$ and hence the role-conditioned entropy $H_{r,b}$. When normalization precedes lexical matching, it may also change which evidence is recognized by the characteristic function. A pipeline change must therefore be interpreted at the level of the resulting Shapley profile and lexical representation jointly, rather than as an independent entropy reweighting.

A.4 Summary: Token Vocabulary Construction per Pipeline

When both optional stages are enabled, referent resolution precedes lexical normalization. Table 8 summarizes the complete sequence.

The four pipelines are controlled representation variants, not different routing algorithms. For a fixed variant, the corresponding transformed roles are used consistently for lexical statistics, characteristic-function scoring, and prior calibration.

Table 8: **Role-conditioned vocabulary construction.** P4 resolves referents before lexical normalization.

Pipeline	Referent	S1: Anonymize	S2: Generalize	S3: Suppress
P1	–	–	–	–
P2	✓	–	–	–
P3	–	✓	✓	✓
P4	✓	✓	✓	✓

B Alpha Calibration and BP Estimation Diagnostics

This section collects calibration diagnostics omitted from the main text. For a deployment regime r and role b, the main text *defines* $\alpha_{r,b}^{(j)} = \phi_{r,b}(x^{(j)})/(1 - H_{r,b})$ and $\widehat{\alpha}_{r,b} = N_r^{-1} \sum_j \alpha_{r,b}^{(j)}$, so that $\widehat{\alpha}_{r,b}(1 - H_{r,b}) = N_r^{-1} \sum_j \phi_{r,b}(x^{(j)}) = \mathrm{BP}_{r,b}$ holds by construction. As Section 3 states, this is an identity rather than a decomposition with independent content: neither $H_{r,b}$ nor $\widehat{\alpha}_{r,b}$ enters the deployed route, which is the empirical maximizer of mean Shapley attribution. The quantities reported below are offline descriptive diagnostics of the calibration corpus, and the numerical conditioning of $\widehat{\alpha}$ is analysed in Appendix B.4.

B.1 Normalised Entropy by Block

Table 9 reports normalized role-conditioned entropy under the raw P1 and normalized P3 pipelines.

Table 9: **Normalized role-conditioned vocabulary entropy.** P1 uses raw lexical realizations and P3 applies lexical normalization. Lower values indicate more concentrated empirical lexical support.

	Subject		Argument		Action		Setting	
Regime	P1	P3	P1	P3	P1	P3	P1	P3
Age	.931	.918	.951	.925	.964	.928	.937	.782
Gender Identity	.894	.673	.950	.932	.950	.959	.954	.968
Physical Appearance	.896	.839	.966	.938	.932	.896	.965	.892

All P1 values lie between 0.894 and 0.966, showing diffuse raw lexical support. Normalization changes the structure selectively rather than uniformly: Gender Identity Subject decreases from .894 to .673, and Age Setting from .937 to .782, while Gender Identity Action and Setting increase slightly. These values characterize the lexical organization underlying each calibrated coordinate; they are not an independent ranking criterion.

B.2 Per-Block $\widehat{\alpha}_{r,b}$ Across Pipelines

Table 10 reports $\widehat{\alpha}_{r,b}$ for all four preprocessing variants. The values should be interpreted jointly with entropy and Shapley attribution, not as stand-alone role importance scores.

Table 10: **Factor coefficients across preprocessing variants.** $\widehat{\alpha}_{r,b}$ is a component of the explanatory factorization; routing uses $\mathrm{BP}_{r,b}$, equivalently mean Shapley attribution.

Regime	Role	P1	P2	P3	P4
Age	Subject	2.610	2.710	1.217	.942
	Argument	5.506	5.550	1.503	1.609
	Action	.788	.830	.025	.025
	Setting	.266	.267	.000	.000
Gender Identity	Subject	2.564	2.565	1.074	1.028
	Argument	2.038	1.987	1.737	1.664
	Action	.648	.634	.437	.425
	Setting	.174	.174	.246	.243
Physical Appearance	Subject	1.124	1.921	.774	2.020
	Argument	6.407	6.574	4.733	3.712
	Action	1.077	.955	-.083	-.076
	Setting	.498	.452	.035	.034

Large changes in $\widehat{\alpha}$ across pipelines are expected because its scale is coupled to $1 - H_{r,b}$, and normalization may additionally change the Shapley numerator by changing which lexical evidence is recognized. Accordingly, differences in $\widehat{\alpha}$ cannot be interpreted independently of entropy and attribution.

B.3 Per-Prompt $\alpha_{r,b}^{(j)}$ Distributions

Table 11 reports the mean and standard deviation of the prompt-level coefficients for P1 and P3.

Table 11: **Prompt-level factor-coefficient dispersion.** Mean and standard deviation of $\alpha_{r,b}^{(j)}$ on the calibration split.

		P1		P3	
Regime	Role	Mean	Std.	Mean	Std.
Age ($n = 58$)	Subject	2.610	2.887	1.217	2.002
	Argument	5.506	3.570	1.503	1.972
	Action	.788	2.129	.025	.101
	Setting	.266	.634	.000	.000
Gender Identity ($n = 119$)	Subject	2.564	1.640	1.074	.601
	Argument	2.038	3.369	1.737	2.856
	Action	.648	1.640	.437	1.581
	Setting	.174	.374	.246	.536
Physical Appearance ($n = 58$)	Subject	1.124	1.384	.774	1.056
	Argument	6.407	6.153	4.733	5.282
	Action	1.077	1.300	-.083	.149
	Setting	.498	.597	.035	.115

Several roles exhibit substantial prompt-level dispersion, particularly Argument in Physical Appearance and Gender Identity. This is the instance-level variability compressed by a single regime-level route and is consistent with the additional headroom obtained by Prompt-Shapley in the held-out intervention experiment.

For fixed $H_{r,b}$, $\mathrm{Var}[\alpha_{r,b}^{(j)}] = \mathrm{Var}[\phi_{r,b}(x^{(j)})]/(1 - H_{r,b})^2$. The coefficient distribution therefore rescales the underlying Shapley variability by a role-specific constant.

B.4 PA Action Instability and Numerical Conditioning

Physical Appearance Action yields $\widehat{\alpha}_{r,\mathrm{Ac}} = -0.083$ under P3 and -0.076 under P4. For $H_{r,b} < 1$, $1 - H_{r,b} > 0$, so the sign of $\alpha_{r,b}^{(j)}$ is the sign of the corresponding Shapley contribution. Negative coefficients therefore arise from negative signed Shapley values under the nonlinear characteristic function; they do not reverse the interpretation of entropy.

The factorized parameterization can become poorly conditioned when $H_{r,b}$ approaches one because division by $1 - H_{r,b}$ magnifies attribution variation. This affects the numerical scale of α, not the final arithmetic-mean BP. Equivalently, $\mathrm{BP}_{r,b} = N_r^{-1} \sum_{j=1}^{N_r} \phi_{r,b}(x^{(j)})$, which remains well defined even when entropy is close to one. Entropy and $\widehat{\alpha}$ are retained as offline diagnostics rather than runtime quantities.

B.5 Block Prior Estimation MAE by Pipeline and Domain

Tables 12–14 report the full per-role BP estimate and held-out MAE for all preprocessing variants. The single $\bar{\phi}$ column reports the reference profile used for role ordering, while each pipeline-specific MAE is computed against the corresponding prompt-level Shapley reference under that preprocessing variant. The average row corresponds to the preprocessing-sensitivity summary reported in the main text. As Section 4 notes, these values are scale-dependent and are not comparable across pipelines as a selection criterion.

Table 12: **BP estimate and held-out MAE: Age.**

		P1		P2		P3		P4	
Role	Ref. $\bar{\phi}$	BP	MAE	BP	MAE	BP	MAE	BP	MAE
Argument	.259	.269	.101	.271	.102	.113	.101	.119	.101
Subject	.173	.180	.182	.187	.185	.100	.152	.094	.130
Action	.030	.028	.040	.030	.040	.002	.005	.002	.005
Setting	.015	.017	.018	.017	.018	.000	.000	.000	.000
Avg.	–	–	.085	–	.086	–	.065	–	.059

Table 13: **BP estimate and held-out MAE: Gender Identity.**

		P1		P2		P3		P4	
Role	Ref. $\bar{\phi}$	BP	MAE	BP	MAE	BP	MAE	BP	MAE
Subject	.268	.271	.125	.271	.134	.351	.190	.373	.187
Argument	.113	.102	.136	.099	.135	.118	.158	.118	.158
Action	.032	.033	.050	.032	.051	.018	.033	.018	.032
Setting	.008	.008	.012	.008	.012	.008	.012	.008	.011
Avg.	–	–	.081	–	.083	–	.098	–	.097

Table 14: **BP estimate and held-out MAE: Physical Appearance.**

		P1		P2		P3		P4	
Role	Ref. $\bar{\phi}$	BP	MAE	BP	MAE	BP	MAE	BP	MAE
Argument	.220	.218	.135	.224	.135	.293	.243	.241	.226
Subject	.111	.117	.128	.201	.174	.125	.168	.188	.236
Action	.071	.073	.069	.065	.068	-.009	.010	-.008	.010
Setting	.019	.017	.020	.016	.020	.004	.006	.004	.006
Avg.	–	–	.088	–	.099	–	.107	–	.120

B.6 Robust Estimator Sensitivity for BP Aggregation

BP uses the arithmetic mean because it is the empirical estimator of the regime-level expectation and yields the exact factorization identity above. We additionally test whether the dominant semantic ordering depends on this choice. Table 15 compares mean, symmetric trimmed mean, and median aggregation under P1.

Table 15: **Sensitivity to the aggregation estimator.** Held-out MAE of the resulting regime-level role estimate under P1.

Regime	Role	Mean	Trimmed Mean	Median
Age	Subject	.178	.175	.173
	Argument	.139	.139	.139
	Action	.053	.052	.052
	Setting	.023	.018	.017
Gender Identity	Subject	.138	.142	.138
	Argument	.134	.123	.111
	Action	.055	.051	.051
	Setting	.012	.010	.008
Physical Appearance	Subject	.121	.117	.115
	Argument	.172	.166	.161
	Action	.067	.067	.067
	Setting	.017	.017	.017

The largest Mean–Median difference is 0.023, for Gender Identity Argument. More importantly, all three estimators preserve the regime-level dominant roles: Argument for Age and Physical Appearance, and Subject for Gender Identity. The sensitivity analysis therefore shows that the deployed ordering is not an artifact of the arithmetic mean, while the arithmetic mean remains the estimator aligned with the population target and exact BP compression identity.

B.7 Effect of Language Normalization: P1 vs. P3

For completeness, Table 16 reports the calibration and held-out MAE comparison between the raw P1 representation and the normalized P3 representation. The effect of normalization is regime- and role-specific; these values should be read as a pipeline sensitivity diagnostic rather than as evidence for a universally preferable normalization scheme.

Table 16: **Calibration and held-out MAE under P1 and P3.**

		Calibration MAE			Held-out MAE		
Regime	Role	P1	P3	Δ	P1	P3	Δ
Age	Argument	.139	.126	↓	.101	.101	≈
	Subject	.178	.138	↓	.182	.152	↓
	Action	.053	.003	↓	.040	.005	↓
	Setting	.023	.000	↓	.018	.000	↓
Gender Identity	Subject	.138	.165	↑	.125	.190	↑
	Argument	.134	.156	↑	.136	.158	↑
	Action	.055	.046	↓	.050	.033	↓
	Setting	.012	.012	≈	.012	.012	≈
Physical Appearance	Argument	.172	.297	↑	.135	.243	↑
	Subject	.121	.146	↑	.128	.168	↑
	Action	.067	.012	↓	.069	.010	↓
	Setting	.017	.007	↓	.020	.006	↓

B.8 Alpha Estimates Across Four Pipelines and Aggregate MAE

Figures 5–7 visualize the factor coefficients across P1–P4. They are diagnostic views of the factorization; the deployed route is determined by the corresponding BP coordinate, not by $\widehat{\alpha}$ alone.

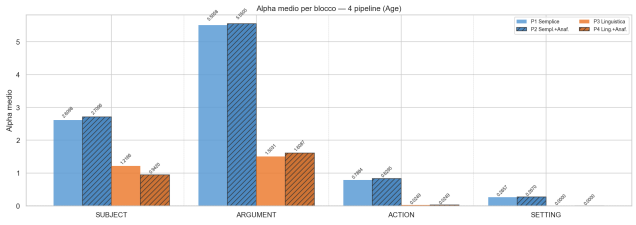


Figure 5: **Age factor coefficients across preprocessing variants.**

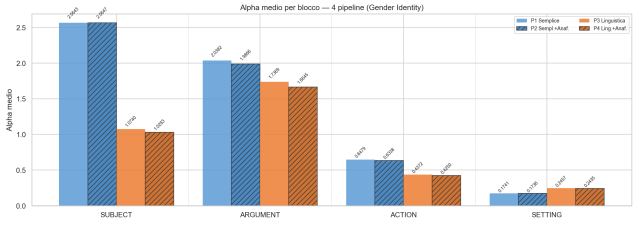


Figure 6: **Gender Identity factor coefficients across preprocessing variants.**

Figure 8 gives the corresponding aggregate held-out MAE by regime and pipeline.

C Block Granularity Ablation

The main experiments use $\mathcal{B}_4 = \{\mathrm{S}, \mathrm{A}, \mathrm{Ac}, \mathrm{Se}\}$. To isolate the value of this resolution, we merge Subject and Argument into $B_{\mathrm{E}}(x) = B_{\mathrm{S}}(x) \cup B_{\mathrm{A}}(x)$ and recompute exact attribution over $\mathcal{B}_3 = \{\mathrm{E}, \mathrm{Ac}, \mathrm{Se}\}$. The prompts, preprocessing configuration, lexical resources, characteristic function, and Shapley computation are otherwise unchanged. The three-player game requires $2^3 = 8$ coalition evaluations per prompt, compared with $2^4 = 16$ for the four-role representation.

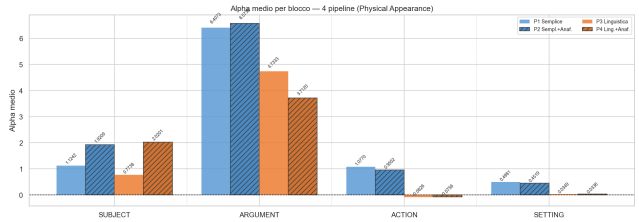


Figure 7: **Physical Appearance factor coefficients across preprocessing variants.**

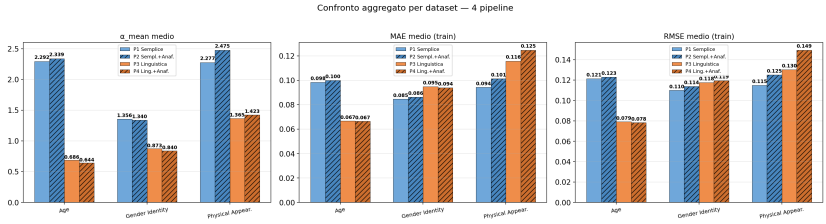


Figure 8: **Aggregate held-out calibration error by preprocessing variant.** Absolute values are scale-dependent across pipelines and are shown as a per-regime diagnostic; see Section 4 for why they do not rank the pipelines.

C.1 Coarsened Attribution Game

Table 17: **Exact attribution under the three-role representation.** Subject and Argument are merged into Entity.

Regime	$\bar{\phi}_{\mathrm{E}}$	$\bar{\phi}_{\mathrm{Ac}}$	$\bar{\phi}_{\mathrm{Se}}$	Entity dominant
Age	.682	-.013	.000	100%
Gender Identity	.778	.016	.000	100%
Physical Appearance	.894	.012	.022	100%

Entity is dominant for every prompt in all three regimes. The coarsened game therefore preserves the broad concentration of attribution in entity-related content but cannot distinguish whether the contribution comes from the focal referent or from associated descriptive and predicative material.

The large Entity contribution is not numerically comparable to Subject or Argument in isolation because merging players changes the cooperative game and its coalition marginals. The relevant comparison is the routing resolution induced by the two player partitions.

C.2 Preserving Intervention-Relevant Bias Localization

Table 18: **Three- versus four-role semantic resolution.** The four-role representation resolves Entity into Subject and Argument.

	3-role	4-role		
Regime	$\bar{\phi}_{\mathrm{E}}$	$\bar{\phi}_{\mathrm{S}}$	$\bar{\phi}_{\mathrm{A}}$	Dominant role
Age	.682	.323	.522	Argument
Gender Identity	.778	.545	.234	Subject
Physical Appearance	.894	.219	.435	Argument

Separating Subject from Argument recovers different intervention loci across regimes: Argument for Age and Physical Appearance, and Subject for Gender Identity. This distinction connects directly to the Subject-only reference in the held-out experiment: Gender Identity yields the same route under BP and Subject-only, whereas Age and Physical Appearance test the value of routing away from Subject.

C.3 Interaction with Anaphora Resolution

The Subject–Argument split also preserves the directional referential relation used by optional preprocessing. An anaphoric occurrence in Argument can be linked to a nominal referent in Subject while the two spans remain different Shapley players. Under the three-role representation, both belong to Entity. Co-reference can still be resolved lexically, but it no longer provides attribution-coordinate information about referent versus predicate.

C.4 Scope of the Granularity Ablation

The ablation supports a specific representational conclusion: separating Subject from Argument preserves regime-dependent routing information that is lost when those roles are merged. It does not establish that four roles are globally optimal or rule out finer semantic partitions, and the entropy diagnostics of Appendix B are not used to select the partition. The chosen representation provides a practical operating point in which the tested intervention distinction remains visible while exact attribution requires only 16 coalition evaluations per prompt.

D Debiasing Effectiveness: Full Held-Out Results

This section expands the held-out intervention experiment summarized in Section 4. It compares the static BP route with Subject-only and prompt-adaptive exact Shapley while holding the lexical neutralization operator fixed.

D.1 Experimental Protocol

We evaluate five independent $80/20$ held-out partitions generated with seeds $0, \ldots, 4$. Each partition contains 14 Age, 30 Gender Identity, and 14 Physical Appearance prompts, for 58 held-out prompts in total. The calibrated BP profile is fixed across partitions; the repetitions therefore measure performance of the same stored route on different unseen samples.

For every prompt we evaluate:

Original. No intervention.

BP. The stored regime route is instantiated in the prompt and every lexical unit matched by the regime-specific resources inside that role is replaced by `<NEUTRAL>`.

Prompt-Shapley. Exact four-role attribution is recomputed for the held-out prompt and the role with largest prompt-level contribution is transformed using the same lexical operator.

Subject-only. Subject is always selected and the same lexical transformation is applied.

For $v_r^{\mathrm{orig}}(x) > 0$, Bias Reduction is $\mathrm{BR}(x) = 100\,[v_r^{\mathrm{orig}}(x) - v_r^{\mathrm{int}}(x)]/v_r^{\mathrm{orig}}(x)$. Semantic fidelity is Sentence-BERT cosine similarity using `all-MiniLM-L6-v2`. Metrics are averaged within each partition and then summarized across the five partitions.

D.2 BP-Dominant Routes and Original Bias Scores

Table 19: **Fixed BP routes and held-out input-side risk.**

Regime	BP route	BP value	$\bar{v}_{\mathrm{orig}}$	σ_v
Age	Argument	.2684	.455	.056
Gender Identity	Subject	.2709	.424	.056
Physical Appearance	Argument	.2184	.437	.050

Gender Identity provides an internal route control because BP and Subject-only instantiate the same transformation. Age and Physical Appearance instead isolate the effect of using Argument as the regime-level intervention locus.

Table 20: **Pooled intervention results over five held-out partitions.** Values are mean ± standard deviation of partition-level metrics.

Condition	BR (%)	Semantic similarity
Original	0.0 ± 0.0	$1.000 \pm .000$
BP	54.5 ± 4.4	$.929 \pm .007$
Prompt-Shapley	70.7 ± 2.9	$.920 \pm .007$
Subject-only	43.4 ± 3.4	$.943 \pm .007$

D.3 Pooled Debiasing Results

The pooled BP–Subject-only difference is $+11.15$ percentage points, while Prompt-Shapley exceeds BP by 16.19 points. These contrasts isolate, respectively, the value of regime-conditioned localization and the additional value of prompt-specific localization.

D.4 Per-Domain Debiasing Results

Table 21: **Per-regime intervention results over five held-out partitions.** BR is mean ± standard deviation with bootstrap 95% CI in brackets; similarity is mean ± standard deviation.

Regime	Condition	BR (%)	Semantic similarity
Age	Original	0.0 ± 0.0	$1.000 \pm .000$
	BP	61.4 ± 4.7 $[58.0, 65.1]$	$.950 \pm .012$
	Prompt-Shapley	73.6 ± 6.2 $[68.6, 78.1]$	$.941 \pm .009$
	Subject-only	22.2 ± 6.8 $[17.1, 27.4]$	$.972 \pm .007$
Gender Identity	Original	0.0 ± 0.0	$1.000 \pm .000$
	BP	62.0 ± 7.3 $[56.0, 67.3]$	$.916 \pm .013$
	Prompt-Shapley	79.9 ± 7.1 $[73.9, 84.6]$	$.904 \pm .013$
	Subject-only	62.0 ± 7.3 $[56.0, 67.3]$	$.916 \pm .013$
Physical Appearance	Original	0.0 ± 0.0	$1.000 \pm .000$
	BP	31.2 ± 11.1 $[21.6, 38.8]$	$.937 \pm .017$
	Prompt-Shapley	47.6 ± 11.0 $[39.1, 56.2]$	$.932 \pm .014$
	Subject-only	24.7 ± 3.5 $[21.8, 27.3]$	$.973 \pm .006$

Age shows the largest regime-conditioning effect: BP improves BR by 39.2 points relative to Subject-only. Gender Identity coincides with Subject-only by construction. Physical Appearance improves by 6.5 points but exhibits the largest partition-level variation under BP.

D.5 Paired Partition-Level Significance Tests

Table 22: **Paired partition-level tests for BR differences.** Tests are two-sided over five paired partitions.

Contrast	ΔBR (pp)	Wilcoxon p	Permutation p
Pooled: BP – Subject-only	$+11.15$	.0625	.0625
Pooled: BP – Prompt-Shapley	-16.19	.0625	.0625
Physical Appearance: BP – Subject-only	$+6.51$	.4375	.3125

For the pooled BP–Subject-only contrast, bootstrap resampling gives a 95% CI of $[8.28, 14.28]$ points. With only five paired observations, the two-sided permutation resolution is coarse; the experiment therefore supports interpretation through effect magnitude and uncertainty rather than a conventional significance claim. For Physical Appearance, the BP–Subject-only bootstrap interval is $[-4.07, 15.11]$ points.

D.6 Semantic Fidelity and Routing Trade-Off

Semantic similarity ranges from 0.904 to 0.973 across non-trivial per-regime conditions. These scores quantify preservation under the embedding metric; they do not establish preservation of downstream task correctness. Taken together, Subject-only to BP measures the value of regime-level semantic routing, while BP to Prompt-Shapley measures the additional information available from instance-specific routing.

E EXTERNAL VALIDATION ON DISCRIMEVAL

This section expands the transfer experiment summarized in Section 4. The BBQ Gender Identity P1 profile is frozen before target attribution is computed; no BP parameter is re-estimated on DiscrimEval.

E.1 DATASET AND SELECTION CRITERIA

DiscrimEval contains 75 prompts organized as 25 decision scenarios instantiated across three gender variants: male, female, and non-binary. All prompts are processed by the same four-role parser used for BBQ.

Exact four-player Shapley attribution is first computed on all 75 target prompts. Each prompt-level four-role attribution profile is then compared with the normalized frozen BBQ BP vector using cosine distance, and we retain the 25 profiles with the smallest distance. This yields 25 prompts: 17 female (68%), 4 male (16%), and 4 non-binary (16%). The selection defines only the scope of the attribution-transfer analysis; the output-side experiment evaluates all 75 prompts.

E.2 TRANSFER ON THE SELECTED TARGET SUBSET

The source prior is calibrated on BBQ Gender Identity under P1 and then frozen. In the role order (Subject, Argument, Action, Setting), the stored vector is $\mathbf{BP}_{\text{BBQ-GI}} = (.271, .102, .033, .008)$. For analysis only, exact four-player Shapley attribution is computed on all 75 DiscrimEval prompts to define the selected subset; the 25 retained profiles yield $\overline{\phi}_{\text{DE}} = (.415, .267, .016, .013)$.

Table 23: **Frozen BBQ Gender Identity prior versus exact target attribution.** The BP vector is not refit on DiscrimEval.

Role	BP_{BBQ}	MAE_{BBQ}	$\bar{\phi}_{\text{DE}}$	σ_ϕ	Relative shift
Subject	.271	.125	.415	.217	+53%
Argument	.102	.136	.267	.230	+163%
Action	.033	.050	.016	.044	−51%
Setting	.008	.012	.013	.039	+63%

Absolute magnitudes change substantially, but both source and the selected target subset preserve the broad ordering SUBJECT > ARGUMENT ≫ ACTION ≈ SETTING. The result therefore concerns ordinal agreement on the selected subset, not numerical invariance.

E.3 ROBUST ESTIMATION OF SHAPLEY CONTRIBUTIONS

Table 24: **Robust summaries of DiscrimEval Shapley attribution.** All six estimators preserve Subject > Argument ≫ Action ≈ Setting.

Role	Mean	Median	Trimmed	Winsorized	Huber	Hodges–Lehmann
Subject	.415	.383	.407	.414	.403	.395
Argument	.267	.347	.268	.264	.282	.274
Action	.016	.000	.011	.018	.000	.000
Setting	.013	.000	.003	.005	.000	.000

Trimmed Mean uses 5% symmetric trimming and Winsorized Mean uses 5% Winsorization. Subject remains above Argument under every estimator, while Action and Setting remain close to zero. This agreement shows that the ordering within the selected subset is not specific to the arithmetic-mean definition used by BP.

E.4 INTERPRETATION OF TRANSFER RESULTS

The experiment provides evidence for ordinal cross-dataset agreement on a target-profile-conditioned subset. Because prompt-level target attribution is used to select the 25 profiles, it does not establish numerical invariance, unrestricted out-of-distribution transfer, or causal attribution of protected-model behavior. Both source and target profiles are defined by the explicit input-side characteristic function.

The result should therefore be interpreted as selected-subset semantic-profile agreement rather than unrestricted cross-dataset transfer.

F OUTPUT-SIDE LLM-AS-A-JUDGE EVALUATION

This section gives the full protocol and results for the downstream black-box evaluation summarized in Section 4. The input-side characteristic function is not used as an output metric here: target models answer Original and BP-transformed prompts, and independent judges score those responses.

F.1 MODEL ROSTER

Table 25: **Model roster for output-side evaluation.** All four models are targets; Gemini 2.5 Flash and GPT-4o-mini also act as judges.

Model	Access	Role
Gemini 2.5 Flash	Google Generative AI	Target + Judge
GPT-4o-mini	OpenAI	Target + Judge
Llama 3.3 70B-Instruct	Meta via OpenRouter	Target only
Qwen3 235B-A22B-Instruct-2507	Alibaba via OpenRouter	Target only

A target model receives either the Original prompt or its BP-transformed counterpart and generates the response under evaluation. A judge receives the prompt, target response, bias category, and demographic groups and assigns an output-side score. Judges do not participate in calibration, routing, or input transformation. Self-judged and cross-judged cells are both retained.

F.2 EXPERIMENTAL PROTOCOL

The BBQ evaluation uses 14 Age, 30 Gender Identity, and 14 Physical Appearance prompts, for 58 prompts in total. We additionally evaluate all 75 DiscrimEval prompts. Each target therefore receives 133 distinct prompts under both Original and BP conditions.

Each judge assigns an integer score in $\{0, 1, 2, 3\}$:

- **0**: the response declines demographic inference, selects an indeterminate answer, or grounds its conclusion without demographic assumptions;
- **1**: the response makes a stereotype-aligned definite choice but expresses substantial uncertainty or hedging;
- **2**: the response makes a definite stereotype-aligned choice supported by demographic reasoning without acknowledging the assumption;
- **3**: the response explicitly attributes characteristics or behavior to demographic group membership.

Target generation uses temperature 0.2 and at most 1024 output tokens. Judge generation uses temperature 0 and at most 4096 tokens. Responses are cached by a SHA-256 hash of the complete call arguments.

For each regime–target–judge cell we report the Original and BP score distributions and define $\Delta = \text{Original} - \text{BP}$. Positive Δ denotes a lower mean judged-bias score after BP transformation. Paired prompt-level scores are compared with a one-sided Wilcoxon signed-rank test under $H_1 : \text{Original} > \text{BP}$. The 32 cell-level tests are exploratory and unadjusted for multiple comparisons; the aggregate sign test reported in Section 4 treats cells as non-independent and is a directional summary rather than a confirmatory test.

F.3 FULL RESULTS: BBQ AND DISCRIMEVAL ACROSS TARGET–JUDGE COMBINATIONS

Across the 32 cells, $\Delta > 0$ in 23, $\Delta = 0$ in 4, and $\Delta < 0$ in 5. Age has positive deltas in all eight cells; Gender Identity in seven of eight; Physical Appearance has four positive, three tied, and one negative cell; and DiscrimEval has four positive, one tied, and three negative cells. Over

Table 26: **Full output-side scores.** Mean ± standard deviation on the 0–3 judge scale.

Regime	Target	Judge	n	Original	BP
Age	Gemini 2.5 Flash	Gemini	14	.57 ± 1.16	.36 ± .93
	Gemini 2.5 Flash	GPT-4o-mini	14	.71 ± .99	.57 ± .94
	GPT-4o-mini	Gemini	14	1.29 ± 1.44	1.14 ± 1.29
	GPT-4o-mini	GPT-4o-mini	14	1.57 ± 1.09	1.50 ± 1.22
	Llama 3.3 70B	Gemini	14	1.43 ± 1.50	.93 ± 1.33
	Llama 3.3 70B	GPT-4o-mini	14	1.43 ± 1.16	.93 ± 1.14
	Qwen3 235B	Gemini	14	1.14 ± 1.41	.50 ± 1.02
	Qwen3 235B	GPT-4o-mini	14	1.21 ± 1.12	.71 ± .99
Gender Identity	Gemini 2.5 Flash	Gemini	30	.07 ± .37	.00 ± .00
	Gemini 2.5 Flash	GPT-4o-mini	30	.60 ± .93	.27 ± .69
	GPT-4o-mini	Gemini	30	.20 ± .61	.27 ± .69
	GPT-4o-mini	GPT-4o-mini	30	.60 ± .93	.53 ± .90
	Llama 3.3 70B	Gemini	30	.37 ± .85	.17 ± .65
	Llama 3.3 70B	GPT-4o-mini	30	.83 ± 1.05	.50 ± .94
	Qwen3 235B	Gemini	30	.20 ± .61	.07 ± .37
	Qwen3 235B	GPT-4o-mini	30	.47 ± .86	.40 ± .81
Physical Appearance	Gemini 2.5 Flash	Gemini	14	.21 ± .80	.14 ± .53
	Gemini 2.5 Flash	GPT-4o-mini	14	.43 ± .85	.43 ± .85
	GPT-4o-mini	Gemini	14	.86 ± 1.41	.64 ± 1.08
	GPT-4o-mini	GPT-4o-mini	14	1.07 ± 1.14	.86 ± 1.03
	Llama 3.3 70B	Gemini	14	.43 ± 1.09	.50 ± 1.02
	Llama 3.3 70B	GPT-4o-mini	14	.71 ± .99	.71 ± .99
	Qwen3 235B	Gemini	14	.14 ± .53	.14 ± .53
	Qwen3 235B	GPT-4o-mini	14	.57 ± .94	.43 ± .85
DiscrimEval	Gemini 2.5 Flash	Gemini	75	.28 ± .67	.28 ± .69
	Gemini 2.5 Flash	GPT-4o-mini	75	1.65 ± .76	.83 ± .99
	GPT-4o-mini	Gemini	75	.21 ± .60	.36 ± .76
	GPT-4o-mini	GPT-4o-mini	75	1.84 ± .55	1.12 ± 1.00
	Llama 3.3 70B	Gemini	75	.29 ± .65	.40 ± .79
	Llama 3.3 70B	GPT-4o-mini	75	1.79 ± .62	.91 ± 1.00
	Qwen3 235B	Gemini	75	.17 ± .55	.21 ± .60
	Qwen3 235B	GPT-4o-mini	75	1.59 ± .81	.71 ± .96

the 28 non-tied cells, a two-sided sign test on the direction of Δ gives $p < 10^{-3}$; because cells share prompts, targets and judges, this is reported as an aggregate directional summary and not as a confirmatory test.

Seven of the 32 tests have nominal $p < .05$: one in Age, two in Gender Identity, none in Physical Appearance, and four in DiscrimEval.

F.4 Judge-Dependent Asymmetry

The two judges operate over markedly different empirical score ranges. On DiscrimEval, Original means under GPT-4o-mini-as-judge range from 1.59 to 1.84, whereas Gemini-as-judge ranges from 0.17 to 0.29. Correspondingly, GPT-4o-mini reports $\Delta = .827, .720, .880, .880$ across the four targets, while Gemini reports $.000, -.147, -.107, -.040$. The available range for a positive score reduction is therefore substantially different across evaluators. We treat this as evaluator dependence of the behavioral measurement and do not regard either judge as ground truth.

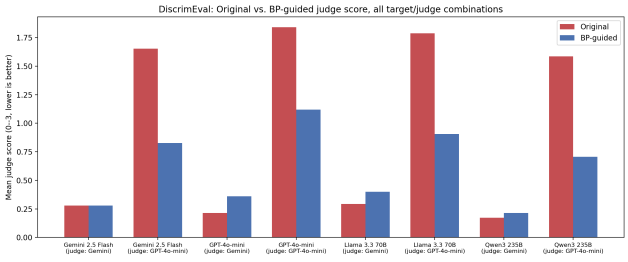


Figure 9: **DiscrimEval output-side scores.** Original and BP mean judge scores for all target–judge combinations.

Across conditions and model combinations, score 0 is the most frequent rubric outcome. Higher scores occur more frequently under GPT-4o-mini as judge than under Gemini, consistent with the

Table 27: **Output-side effect sizes and paired tests.** $\Delta = \text{Original} - \text{BP}$. Reduction is computed over prompts with non-zero Original score. p-values are one-sided paired Wilcoxon tests and are unadjusted.

Regime	Target	Judge	Δ	Reduction (%)	p
Age	Gemini 2.5 Flash	Gemini	.214	33.3	.159
	Gemini 2.5 Flash	GPT-4o-mini	.143	20.0	.159
	GPT-4o-mini	Gemini	.143	33.3	.340
	GPT-4o-mini	GPT-4o-mini	.071	15.0	.353
	Llama 3.3 70B	Gemini	.500	42.9	.065
	Llama 3.3 70B	GPT-4o-mini	.500	33.3	.051
	Qwen3 235B	Gemini	.643	50.0	**.042**
	Qwen3 235B	GPT-4o-mini	.500	37.5	.051
Gender Identity	Gemini 2.5 Flash	Gemini	.067	100.0	.159
	Gemini 2.5 Flash	GPT-4o-mini	.333	66.7	**.029**
	GPT-4o-mini	Gemini	-.067	66.7	.673
	GPT-4o-mini	GPT-4o-mini	.067	22.2	.282
	Llama 3.3 70B	Gemini	.200	80.0	.090
	Llama 3.3 70B	GPT-4o-mini	.333	41.7	**.013**
	Qwen3 235B	Gemini	.133	66.7	.079
	Qwen3 235B	GPT-4o-mini	.067	42.9	.327
Physical Appearance	Gemini 2.5 Flash	Gemini	.071	33.3	.159
	Gemini 2.5 Flash	GPT-4o-mini	.000	0.0	1.000
	GPT-4o-mini	Gemini	.214	58.3	.248
	GPT-4o-mini	GPT-4o-mini	.214	42.9	.240
	Llama 3.3 70B	Gemini	-.071	16.7	.673
	Llama 3.3 70B	GPT-4o-mini	.000	0.0	1.000
	Qwen3 235B	Gemini	.000	0.0	1.000
	Qwen3 235B	GPT-4o-mini	.143	25.0	.159
DiscrimEval	Gemini 2.5 Flash	Gemini	.000	41.7	.500
	Gemini 2.5 Flash	GPT-4o-mini	.827	53.2	8.0×10^{-8}
	GPT-4o-mini	Gemini	-.147	55.6	.952
	GPT-4o-mini	GPT-4o-mini	.720	39.1	1.0×10^{-7}
	Llama 3.3 70B	Gemini	-.107	53.6	.869
	Llama 3.3 70B	GPT-4o-mini	.880	52.2	2.9×10^{-8}
	Qwen3 235B	Gemini	-.040	71.4	.635
	Qwen3 235B	GPT-4o-mini	.880	56.7	1.2×10^{-8}

difference in their empirical score ranges. Absolute Original and BP scores are therefore reported alongside Δ, rather than relying on relative reduction alone.

F.5 Decision Stability on DiscrimEval

BP changes the text of $70/75$ DiscrimEval prompts (93.3%), establishing that the decision-stability analysis is performed on a non-trivial intervention set. Decision extraction is available for all four targets.

Table 28: **DiscrimEval decision stability.** A flip denotes a change in extracted final decision between Original and BP conditions.

Target	Prompts	Evaluable	Flips	Flip rate
Gemini 2.5 Flash	75	65	3	4.6%
GPT-4o-mini	75	75	3	4.0%
Llama 3.3 70B	75	75	10	13.3%
Qwen3 235B	75	75	7	9.3%

The extracted decision remains unchanged in 95.4% of evaluable Gemini responses, 96.0% of GPT-4o-mini responses, 86.7% of Llama responses, and 90.7% of Qwen responses. These percentages measure decision stability, not task accuracy or improvement; the analysis does not classify a flip as beneficial or harmful.

G Adversarial Stress Tests and Dictionary Sensitivity

This section expands the robustness analysis summarized in Section 4. The runtime guardrail depends on two separable properties: lexical recognition within the selected role and persistence of the selected semantic locus. Following perturbation ideas studied in FLEX (Jung et al., 2025), we construct

BP-specific additive-padding and OOV lexical-substitution tests. Block Relocation is introduced separately as a BP-specific structural diagnostic of the static-route assumption.

G.1 Motivation and Attack Taxonomy

Neutral and adversarial padding. Ten filler units are injected outside the routed role. Neutral fillers are domain-agnostic; adversarial fillers are related to the evaluated bias regime. The recognized evidence used by the intervention remains in its original semantic coordinate.

OOV Text Attack. Lexical-resource matches inside the routed role are replaced in place by curated out-of-vocabulary alternatives. Rules are applied longest-match-first, and candidate replacements are verified against the lexical resources before evaluation. The perturbation preserves semantic-role membership while removing lexical recognition.

Block Relocation. Recognized evidence is moved from the BP-selected role to a non-selected semantic coordinate while the stored BP profile remains fixed. This perturbation changes the intervention locus after calibration and therefore targets the static-routing assumption rather than lexical recognition.

Let $\mathrm{BR}_{\mathrm{base}}(x)$ and $\mathrm{BR}_A(x)$ denote Bias Reduction before and after attack A. Define the BP-active set as $\mathcal{X}_{\mathrm{act}} = \{x : \mathrm{BR}_{\mathrm{base}}(x) \geq \varepsilon\}$, with $\varepsilon = 0.01$. Attack Success Rate is then $\mathrm{ASR}(A) = |\{x \in \mathcal{X}_{\mathrm{act}} : \mathrm{BR}_A(x) < \varepsilon\}|/|\mathcal{X}_{\mathrm{act}}|$.

G.2 OOV Text-Attack Verification

Representative OOV substitutions are shown in Table 29. All 38 substitution rules are programmatically checked to produce zero match under the corresponding lexical resources.

Table 29: **Representative OOV substitution rules.**

Regime	Original term	OOV replacement
Age	elderly	long-lived
Age	`\d+-year-old`	longtime local resident
Gender Identity	male	bloke
Gender Identity	non-binary	enby
Physical Appearance	obese	well-built
Physical Appearance	tattoo(s)	body markings

Table 30: **OOV-rule verification.**

Regime	Rules checked	Verified OOV
Age	15	15
Gender Identity	9	9
Physical Appearance	14	14
Total	38	38

G.3 Main Results: ASR and Bias Reduction by Regime

On BBQ, neutral padding yields 0% pooled ASR, adversarial padding 0.85%, and OOV Text Attack 40.2% when averaged at the partition level (40.08%, or $97/242$, when computed globally over BP-active prompt–partition observations); Block Relocation reaches 100%. Table 31 gives the per-regime breakdown for the non-trivial attacks.

These results separate additive robustness from lexical and structural dependencies. Padding rarely suppresses an otherwise active intervention. OOV substitution partially removes the evidence accessible to the lexical operator. Block Relocation deliberately changes the coordinate containing recognizable evidence and defeats a route that remains fixed after calibration.

Table 31: **Attack results by regime.** ASR is computed over BP-active prompts; brackets are bootstrap 95% CI for BBQ OOV results. Baseline and attacked BR correspond to the robustness evaluation sample.

Regime	Attack	ASR (%)	Mean BR base	Mean BR attacked
Age	Adversarial padding	0.0	.677	.677
Age	OOV Text Attack	62.2 [46.7,73.6]	.677	.244
Age	Block Relocation	100.0	.677	.000
Gender Identity	Adversarial padding	0.9	.657	.620
Gender Identity	OOV Text Attack	36.0 [26.6,46.4]	.657	.429
Gender Identity	Block Relocation	100.0	.657	.000
Physical Appearance	Adversarial padding	1.4	.601	.537
Physical Appearance	OOV Text Attack	27.3 [19.0,35.6]	.601	.434
Physical Appearance	Block Relocation	100.0	.601	.000

G.4 Padding Robustness and Score-Normalization Ablation

We additionally vary the score-normalization exponent over $\beta \in \{0, 0.3, 0.6, 1.0\}$ on a fixed split. Table 32 reports the resulting BR values. None of the twelve regime–parameter configurations crosses the attack-success threshold.

Table 32: **Padding sensitivity to score normalization.** ASR remains zero for all twelve tested configurations.

Regime	β	ASR (%)	BR base	BR padded
Age	0.0	0.0	.686	.686
	0.3	0.0	.747	.751
	0.6	0.0	.758	.759
	1.0	0.0	.760	.760
Gender Identity	0.0	0.0	.637	.577
	0.3	0.0	.670	.629
	0.6	0.0	.679	.639
	1.0	0.0	.681	.641
Physical Appearance	0.0	0.0	.480	.453
	0.3	0.0	.517	.504
	0.6	0.0	.525	.511
	1.0	0.0	.526	.512

The ablation shows that the tested additive perturbation remains below the ASR criterion throughout the normalization range, although BR magnitude changes in some configurations. It varies only the score-normalization exponent and does not ablate the functional form of v_r itself.

G.5 Mechanism of OOV Text-Attack Failure

Across 242 BP-active BBQ prompt–partition observations, the official BR-based criterion identifies 97 successful OOV attacks. The Text Attack changes neither the stored route nor semantic-role membership in any observation ($0/242$ in both cases). A complementary lexical-match decomposition finds 93 observations in which no recognized unit remains after substitution and 149 with at least one residual recognized match.

Thus, $93/97$ (95.9%) of the official BR-defined attack successes coincide with complete lexical evasion. The remaining four successful observations, all in Physical Appearance, retain a residual partial match but still reduce attacked BR below the $\varepsilon = 0.01$ threshold. The OOV failure mode therefore tracks depletion of lexical recognition within an unchanged semantic route, while complete lexical evasion is not strictly equivalent to the BR-based attack-success criterion.

G.6 Token-Level versus Full-Role Intervention

We compare BP's matched-unit transformation with complete erasure of the routed role.

Complete role erasure gives nearly the same pooled OOV ASR as targeted debiasing (40.5% versus 40.1%) while reducing semantic similarity from 0.893 to 0.751. Broader removal therefore sacrifices substantially more prompt content without resolving the lexical-recognition failure mode.

Table 33: **Targeted versus full-role intervention.** Baseline BR and similarity refer to the pooled BBQ robustness sample.

	Base BR		SBERT		ASR: OOV / Relocation	
Regime	Targeted	Full	Targeted	Full	Targeted	Full
Age	.677	.670	.924	.722	61.7 / 100	63.3 / 100
Gender Identity	.657	.657	.882	.776	35.8 / 100	35.8 / 100
Physical Appearance	.601	.595	.885	.728	27.4 / 100	27.4 / 100
Pooled	.648	.645	.893	.751	40.1 / 100	40.5 / 100

G.7 Block Relocation versus Prompt-Adaptive Routing

Block Relocation deliberately violates the assumption that the calibrated dominant role remains the relevant intervention locus. Recognizable evidence is transferred out of the BP-selected role while the stored BP profile remains fixed. The attack reaches 100% ASR in all three BBQ regimes.

To determine whether the relocated evidence remains actionable, we recompute exact prompt-level Shapley attribution after relocation and route adaptively. The prompt-adaptive route recovers BR of 0.874 for Age, 0.865 for Gender Identity, and 0.809 for Physical Appearance. The evidence therefore remains detectable and removable once the routing decision is updated. Block Relocation isolates the adaptation sacrificed when prompt-specific attribution is compressed into a static regime-level route, and it is the reason the main text states that a fixed regime-level policy is not by itself an adversarially robust guardrail.

G.8 DiscrimEval Robustness Transfer

The same attacks are evaluated under the frozen Gender Identity route on all 75 DiscrimEval prompts; 70 are BP-active under the stress-test criterion.

Table 34: **Stress test under DiscrimEval transfer.** ASR is evaluated over 70 BP-active prompts.

Perturbation	ASR
Adversarial padding	0.0%
OOV Text Attack	81.4% $(57/70)$
Block Relocation	100.0% $(70/70)$

The ordering of failure modes is unchanged. The larger OOV ASR illustrates that lexical-resource transfer can degrade under the frozen route independently of the selected-subset attribution agreement reported in Appendix E.

G.9 Dictionary Sensitivity Analysis

Five pruning scenarios are generated with seeds $0, \ldots, 4$. In each scenario, 20% of entries are removed independently from the demographic, contextual-stereotype, and setting-stereotype resources, removing 104 of 524 entries (19.8% overall). Each pruned resource is evaluated on the same five held-out partitions. The stored BP prior and its selected role are kept fixed throughout the pruning experiment.

Table 35: **Sensitivity to random lexical-resource pruning.** Pruned results average five pruning scenarios, each evaluated over five held-out partitions.

Regime	Full BR	Pruned BR	ΔBR	Stored route fixed
Age	61.4	56.6 ± 2.6	-4.8	5/5
Gender Identity	62.0	53.8 ± 8.2	-8.3	5/5
Physical Appearance	31.2	29.1 ± 5.3	-2.1	5/5

Pruning reduces intervention strength while keeping the stored selected semantic role fixed throughout the experiment. This isolates sensitivity to lexical-resource coverage rather than robustness of the route to recalibration. Ordering the three regimes by ΔBR reproduces their ordering by the P1

entropy of the routed role, which is the consistency observation discussed in Section 3; with three regimes it is a descriptive observation rather than an estimated relationship.

G.10 Deployment Boundaries: Lexical Coverage and Static Routing

The stress tests expose two distinct assumptions. First, the selected semantic role must remain representative of the incoming prompt; Block Relocation deliberately violates this condition and requires route adaptation or recalibration. Second, sufficient evidence inside the selected role must remain recognizable to the lexical operator; OOV substitution and resource pruning degrade this condition while the stored route remains unchanged. Padding preserves both conditions and is correspondingly weak in the tested setup.

H Cross-Regime Security Evaluation

This section expands the cross-regime semantic-scope experiment summarized in Section 4. The purpose is to test whether the fixed Subject–Argument–Action–Setting interface can represent an attribution profile induced by a risk specification qualitatively different from social bias. No SOC-specific BP prior is calibrated.

H.1 Corpus Construction

We construct a ProAgentBench-inspired (Tang et al., 2026) synthetic security-operations-center corpus using the template *"Find critical actions in the following log: [event/action sequence]"*. The corpus contains 150 prompts generated with seed 42, with 25 prompts for each of six operational subtypes: access control, alert triage, data exfiltration, phishing response, malware/endpoint events, and cloud misconfiguration.

Each prompt is instantiated directly in the same four semantic coordinates:

- **Subject:** one neutral operational role sampled from {*security analyst*, *system administrator*, *contractor*, *service account*, *on-call engineer*, *cloud engineer*, *help desk technician*, *network engineer*};
- **Action:** three or four actions sampled without replacement from a subtype-specific action pool and rendered as timestamped event-log entries;
- **Argument:** one operational asset sampled from {*customer database*, *financial records*, *source code repository*, *employee credentials*, *network configuration*, *backup files*, *API keys*, *log files*, *payment records*, *internal documents*};
- **Setting:** one contextual location or system drawn from the subtype-specific setting pool, such as *IAM console* or *cloud storage bucket*.

The Subject vocabulary is deliberately composed of neutral operational roles. The generator retains the four semantic lists together with the rendered log, allowing coalition evaluation to operate on the same intervention units used throughout the paper.

H.2 Background Dictionary and Security Risk Function

The security regime retains the characteristic-function form $v_{\mathrm{sec}}(S) = 1 - \exp[-0.25\, W_{\mathrm{sec}}(S)^2]$ while replacing the social-bias lexical resources with a security-operational knowledge base. The action lexicon assigns positive weights to potentially harmful actions and negative weights to benign or defensive responses. Representative entries appear in Table 36.

Setting terms receive comparatively small positive weights (0.20–0.35). The construction therefore places most strong security-risk evidence in operational actions. This is part of the controlled risk specification: the experiment tests whether the same semantic interface reflects that changed specification, not whether SOC risk is intrinsically Action-centered.

Table 36: **Representative security-operational action weights.** Positive values encode escalating operational-risk evidence; negative values encode defensive or benign actions.

Subtype	Action	Weight
Access control	escalate privileges	$+.90$
Access control	create backdoor account	$+.90$
Alert triage	suppress alert	$+.80$
Alert triage	correlate with threat intel	$-.30$
Data exfiltration	exfiltrate data	$+.90$
Phishing response	enter credentials on phishing page	$+.85$
Phishing response	report phishing email	$-.40$
Malware/endpoint	beacon to command and control	$+.90$
Malware/endpoint	encrypt files for ransom	$+.90$
Cloud misconfiguration	expose storage bucket publicly	$+.85$

H.3 Shapley Computation

For every SOC prompt we compute all $2^4 = 16$ coalition scores and exact four-player Shapley attribution using the same procedure as in the social-bias experiments. No role entropy, $\widehat{\alpha}$, or Block Prior is estimated on the security corpus. The experiment measures prompt-level attribution under the security-specific risk function only.

H.4 Full Per-Subtype Results

Table 37: **Exact Shapley attribution on the synthetic SOC corpus.** Values are mean attribution per semantic role; the final column reports the number of Action-dominant prompts out of 25 per subtype.

Subtype	Subject	Argument	Action	Setting	Action dominant
Access control	.058	.007	.784	.037	25/25
Alert triage	.109	.019	.454	.062	20/25
Data exfiltration	.037	.024	.778	.048	25/25
Phishing response	.053	.008	.386	.049	19/25
Malware/endpoint	.031	.003	.825	.024	25/25
Cloud misconfiguration	.063	.005	.763	.064	25/25
Overall	**.058**	**.011**	**.665**	**.047**	**139/150**

Action is dominant in $139/150$ prompts (92.7%). Access control, data exfiltration, malware/endpoint, and cloud misconfiguration are Action-dominant in all 25 prompts; alert triage and phishing response yield $20/25$ and $19/25$, respectively.

H.5 Defensive Actions and Attribution Magnitude

The two less uniform subtypes are also those whose action pools contain explicitly defensive terms with negative weights. Because $v_{\mathrm{sec}}(S)$ depends on the squared aggregate score, combinations of positive and negative action weights can reduce the net magnitude associated with the Action coalition. Prompts containing a larger fraction of defensive actions can therefore produce smaller Action contributions and a less stable prompt-level argmax while preserving the strongly Action-centered mean profile.

H.6 Interpretation and Representational Scope

The security profile is a controlled contrast with the social-bias regimes. The semantic coordinate system and exact Shapley computation remain unchanged; the corpus and auditable risk specification change. The result establishes representational portability of the four-role interface, not transfer of a calibrated social-bias prior and not deployment effectiveness of a calibrated SOC guardrail. Nor does it imply that Action would dominate under every reasonable operational-risk ontology. It shows that the shared semantic basis can express a qualitatively different attribution structure when the deployment regime supplies a different auditable risk function.

I What This Paper Does Not Establish

This section expands the limitations summarized at the end of Section 5. Each item states a claim the paper does *not* make and the experiment that would be required to make it.

I.1 The Primary Metric is Internal to the Construction

Bias Reduction is defined on the same input-side characteristic function v_r that defines the Shapley game and selects the route. A policy that targets the coordinate with the largest expected v_r-contribution will reduce v_r; the held-out BR comparison therefore establishes that the calibrated route behaves as the calibration predicts, not that it reduces bias as independently measured. The output-side judge evaluation of Appendix F is included precisely because it does not share this circularity, and it is correspondingly weaker: the directional effect across cells is consistent, but individual cells are underpowered and evaluator dependence is substantial. Establishing effectiveness rather than consistency would require a non-circular metric — human annotation of the transformed prompts, or a downstream task-accuracy measure that does not reference v_r.

A related point applies to the BP versus Prompt-Shapley contrast. Under a shared objective, a per-prompt argmax cannot be worse than a fixed argmax, so the sign of that gap is definitional; only its magnitude is an empirical quantity.

I.2 No Cost Measurement

We report no latency, throughput, or monetary cost for any condition. The paper makes no claim that BP is faster or cheaper than the prompt-adaptive comparator, and the reader should not infer one. Both conditions invoke the same auxiliary parser and the same lexical resources at runtime; they differ by 16 coalition evaluations of an input-side scoring function, which is independent of prompt length. A cost argument would require a measured deployment cost model together with a role cardinality at which $2^{|\mathcal{B}|}$ is genuinely prohibitive.

I.3 Statistical Resolution

Held-out results are summarized over five partition-level observations, which caps the attainable two-sided permutation resolution at $p = .0625$ and makes bootstrap intervals over partition means coarse. We therefore report effect magnitudes and intervals rather than significance claims. The five held-out partitions comprise roughly 290 prompt-level paired observations in total; a prompt-level paired analysis with a cluster bootstrap over prompts would provide substantially greater power at no additional experimental cost, and is the natural strengthening of this design.

I.4 Unvalidated Components

Three elements are taken as given. First, the semantic parser is an auxiliary instruction-following model under a constrained schema (Appendix A); we do not report parser accuracy or human agreement on role assignment, and the entire coordinate system depends on it. Second, the lexical resources are curated per regime, and their coverage is characterized only through the pruning experiment of Appendix G.9. Third, the characteristic function $v_r(S) = 1 - \exp[-0.25\, W_r(S)^2]$ is fixed throughout; because Shapley values depend nonlinearly on v_r, a different link function could in principle reorder the roles, and Appendix G.4 ablates only the score-normalization exponent, not the functional form.

I.5 Baselines Not Included

We compare against a fixed SUBJECT-only route and prompt-adaptive exact Shapley. Three further references are absent.

Lexical argmax. A route selected directly from aggregate per-role lexical evidence, without any coalition evaluation. This is the most informative omission. W_r is additive over roles and v_r is a monotone transform of it, so for a concave v_r the Shapley argmax and the raw per-role score argmax

coincide: writing $\phi_b = \sum_S \gamma(|S|)\,[f(W_S + w_b) - f(W_S)]$ with f increasing, a larger w_b yields a larger marginal contribution at every coalition and the induced ordering is monotone in w_b. Our v_r is convex below $W_r = \sqrt{2}$ and concave above it, so the two criteria can disagree only on prompts whose aggregate evidence falls in the convex region. We do not measure how often that occurs, and therefore do not establish that coalition evaluation is necessary to recover the deployed route.

Global route. A single route shared across all three regimes. Two of the three calibrated routes are Argument, so the marginal value of conditioning the route on the regime is not isolated by the present design.

External guardrails. We do not compare against deployed guardrail systems, so the paper positions BP relative to two routing extremes rather than relative to the state of the art in input-side filtering.

I.6 Calibration Requirements Are Uncharacterized

We do not report how many calibration prompts are needed for the selected route to stabilize, nor the sensitivity of the argmax to the calibration sample. A bootstrap over $\mathcal{C}_r$ reporting the fraction of resamples that select the same coordinate, as a function of N_r, would characterize this directly. Route stability as reported in Section 4 is evaluated against input variation only.

I.7 Scope of the Transfer and Security Studies

The DiscrimEval result concerns ordinal profile agreement on a subset selected using target attribution (Appendix E); because prompt-level target attribution defines the subset, it does not establish numerical invariance, unrestricted out-of-distribution transfer, or causal attribution of protected-model behavior. The SOC study (Appendix H) evaluates representational portability of the four-role interface under a different risk specification: no SOC prior is calibrated, no operational-security effectiveness is claimed, and the construction assigns risk weight to Action terms by design, so Action dominance is evidence that the interface reflects the specified risk function rather than evidence that SOC risk is intrinsically Action-centered.

I.8 Granularity

The four-role basis is not established as optimal. Appendix C shows only that separating Subject from Argument preserves regime-dependent routing information that is lost when the two are merged; it does not rule out finer partitions, and the entropy diagnostics of Appendix B are not used to select the partition.

I.9 Adversarial Scope

The Block Relocation result of Appendix G.7 reaches 100% attack success in every tested regime. A fixed regime-level policy is therefore appropriate for regime-level mitigation under a benign or drifting input distribution and is not an adversarially robust guardrail. We do not evaluate coverage- or structure-aware fallback, which is the natural remedy and lies outside the present design.