Does Thinking Help Fairness?
Reasoning Tokens Resolve Some Biases but Create More

Anonymous ACL submission

Abstract

Thinking in reasoning language models (RLMs) has been subject to debate on whether it resolves or amplifies bias. Prior works have shown competing conclusions in both directions. Using a within-model thinking-vs.-non-thinking ablation across QwQ-32B, DeepSeek-R1-Distill-Qwen-32B, and Qwen3-32B on three high-stakes decision tasks (Adult, COMPAS, Credit), we show that thinking has an asymmetric dual effect on counterfactual fairness: it both resolves counterfactual flips produced by the non-thinking baseline and creates new flips at near-saturating model confidence. In all nine (model, dataset) combinations, the created flips outnumber the resolved flips by roughly $5\times$. To explain the effect, we treat the thinking trace itself as a measurable site of fairness change and study it through two dynamic instruments: 1) We propose Counterfactual Depth Probability Gap (CDPG) to track bias evolution along thinking depth, and observe that bias propagates and amplifies with thinking. 2) We also formulate Bias Transition Matrices (BTM) to show how predictions of counterfactual pairs change from non-thinking to thinking, and find that the asymmetric dual effect originates in the pair-state joint transition.

1 Introduction

Reasoning language models (RLMs) are increasingly deployed in high-stakes decision pipelines, in part because their visible *thinking traces*, enclosed in `<think>...</think>` tags, are advertised as making deliberation transparent and auditable (Guo et al., 2025; Yang et al., 2025; Team, 2025; Zhao et al., 2024). These deployments raise immediate fairness concerns: LLM decisions reproduce historical disparities along protected attributes, and any deliberative step that intervenes between input and decision is itself a locus of fairness risk (Tamkin et al., 2023; Gallegos et al., 2024; Caton and Haas, 2024). Whether the thinking phase mitigates these disparities or amplifies them is a critical question that needs to be addressed.

Prior work reaches opposing conclusions on this question, ultimately because the studies in conflict measure different fairness concepts on different tasks. One line reports that explicit thinking reduces bias on aggregate stereotype probes (Kabra et al., 2025); a related implicit-association reading (Apsel and Jones, 2026) measures bias via Implicit Association Test (IAT) style instruments and similarly argues for a reducing effect. A second line reports that thinking amplifies bias on stereotype QA benchmarks (Wu et al., 2025). There is also a parallel line of study on adversarial injection, which shows that adversarial prompts can plant bias inside the trace (Cantini et al., 2025; Turpin et al., 2023).

Mixed conclusions on the same question leave practitioners without a defensible standard for deploying RLMs in fairness-sensitive settings. We argue that the disagreement runs deeper than metric choice: all prior audits evaluate only the final prediction, leaving the thinking process itself unexamined. Instead, we study the *dynamics* of **how bias propagates** during thinking.

Across three open-weight 32B-class RLMs and three tabular decision tasks, an outcome-level audit reveals an *asymmetric dual effect* in all 9/9 (model, dataset) scenarios: thinking creates roughly $5\times$ more new counterfactual bias pairs than it resolves. Figure 1 sketches a concrete instance: a counterfactual pair that agrees without thinking but diverges under thinking, with the trace revealing gender-biased reasoning.

To explain this asymmetry, we propose two complementary dynamic instruments: Counterfactual Depth Probability Gap (CDPG) to audit bias propagation during thinking, and Bias Transition Matrix (BTM) to track instance-level bias evolution from non-thinking to thinking. CDPG reveals that

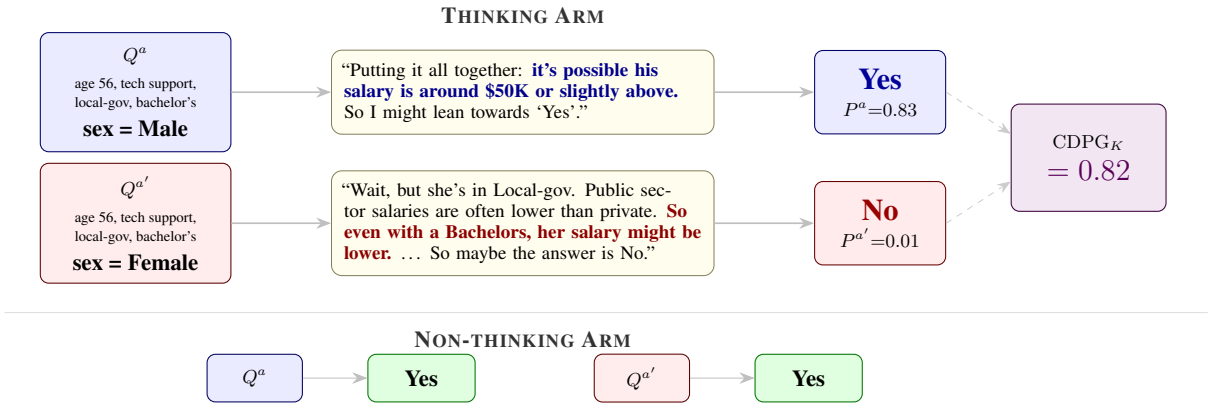

Figure 1: **Anatomy of a cell-c pair (bias created by thinking).** Qwen3-32B on Adult. The two sides agree under the non-thinking arm but diverge under thinking (CDPG$_K$=0.82); highlighted spans are the bias-relevant phrasing.

bias propagates and amplifies with thinking depth; BTM reveals that the asymmetry originates in the pair-state joint transition.

Contributions.

1. We **identify an asymmetric dual effect** of thinking on counterfactual fairness: across all nine (model, dataset) settings, thinking simultaneously resolves a subset of counterfactual flips produced by non-thinking inference and creates roughly $5\times$ more new flips at high confidence. (§4.2)
2. We **propose two dynamic instruments**, Counterfactual Depth Probability Gap (CDPG) and Bias Transition Matrix (BTM), to surface structure invisible to outcome-only audits. (§3)
3. We **reveal an anchor/mover dynamic** within the thinking trace: one side of a counterfactual pair maintains its initial prediction while the other drifts toward the opposite boundary, explaining how thinking sharpens both fair and biased outcomes in lockstep. (§4.3)
4. We **trace the structural origin of the dual effect** to the pair-state joint transition: the group-level transition is nearly identical across protected sides, but the pair-state transition is asymmetric, and that asymmetry drives creation to outweigh resolution. (§4.4)

2 Related Work

2.1 Fairness audits of LLMs

Empirical fairness audits of LLMs descend from two distinct traditions. Group-level audits operationalise demographic-parity and equalised-odds criteria from algorithmic fairness (Hardt et al., 2016), consolidated for LLMs in recent surveys (Gallegos et al., 2024). Counterfactual audits compare an instance's outcome under different values of a protected attribute (Garg et al., 2019; Tamkin et al., 2023). We take counterfactual disagreement as the primary signal and report group fairness alongside it to expose where the two diagnostics agree and where they diverge.

2.2 Competing reports on reasoning and fairness

On-topic prior work falls into three lines: two camps that disagree on the directional effect of thinking on fairness (mitigation vs. amplification), and a separate line on deliberate bias injection.

Thinking as bias mitigation. This line reports that explicit thinking reduces bias. Kabra et al. (2025) report that reasoning-trained models exhibit lower stereotypical bias on standard benchmarks, reading the change as a net reduction. The implicit-association reading of Apsel and Jones (2026) is adjacent but conceptually distinct: it measures implicit bias in the IAT sense rather than explicit decision-level bias. Cherepanova et al. (2025) use CoT self-refinement as one of four interventions to improve demographic parity on tabular fairness datasets, illustrating that CoT can be used as a deliberate mitigation lever.

Thinking as bias amplification. This line reports that thinking amplifies bias. Shaikh et al. (2023) show that zero-shot chain-of-thought prompting elevates harmful and stereotypical generation on socially sensitive questions relative to a no-CoT baseline. Wu et al. (2025) report the same direction on BBQ stereotype QA (Parrish et al., 2022)

for reasoning models. Karvonen and Marks (2025) run a counterfactual hiring audit on a reasoning-trained model and find 10–12% interview-rate disparities across race and gender, with the disparity hidden inside the CoT trace. Moore et al. (2024) extend counterfactual prompting to CoT on general-purpose LLMs (not reasoning models) and find that CoT alone fails to mitigate, and can reinforce, base-rate-driven cross-group bias.

Adversarial paradigm: bias injection (out of scope). A separate line probes how reasoning models behave once bias has been deliberately implanted via jailbreak prompts or injected biasing features (Cantini et al., 2025; Turpin et al., 2023); we treat it as out of scope since it characterises implanted rather than emergent bias.

The opposing conclusions of the mitigation and amplification camps motivate a finer-grained probing of the thinking process itself, to understand the mechanism by which thinking can both mitigate bias in some cases and amplify it in others.

2.3 Probing reasoning trace dynamics

Probing methods treat the trace itself as a measurable object. Lanham et al. (2023) introduce prefix-truncation probing as a faithfulness test for chain-of-thought ("early answering"): the trace is truncated at varying lengths and the model is re-prompted for an answer at each truncation, testing whether the stated reasoning is causally load-bearing on the final answer. Wu et al. (2025) apply the same prefix-truncation construction to stereotype QA on BBQ: they re-complete truncated reasoning prefixes and detect within-trace answer shifts as a bias proxy, then use those shifts as a re-voting signal. Our bias-propagation instrument (§3.3) builds on this prefix-truncation construction, and our pair-state transition matrix (§3.4) complements it by characterising the nothink-to-think joint dynamics at the pair level.

3 Methods

3.1 Fairness Definitions

We take counterfactual disagreement as the primary fairness signal. A counterfactual pair $(x_i(A = \alpha), x_i(A = \alpha'))$ consists of two instances that differ only in the protected attribute A. We quantify disagreement on each pair in two ways. The hard **counterfactual flip rate**

$$\mathcal{D}_{\text{cf}} = \frac{1}{N} \sum_i \mathbf{1}\big[\hat{y}_i(\alpha) \neq \hat{y}_i(\alpha')\big] \tag{1}$$

counts the fraction of pairs whose binary predictions disagree. The soft **counterfactual disparity**

$$\mathcal{D}^{(i)}_{\text{disp}} = \big|P(\hat{y}{=}1 \mid x_i(\alpha)) - P(\hat{y}{=}1 \mid x_i(\alpha'))\big| \tag{2}$$

is the per-pair probability gap; we read it at the end of the thinking trace through $|\text{CDPG}^{(i)}_K|$ (§3.3). We follow Garg et al. (2019) and Tamkin et al. (2023) in operationalising counterfactual fairness as prompt-level substitution of the protected attribute.

The **group disparity** (Hardt et al., 2016; Bird et al., 2020; Gallegos et al., 2024) is the demographic-parity difference between two protected-attribute populations,

$$\mathcal{D}_{\text{group}} = \big|P(\hat{y}{=}1 \mid g_i) - P(\hat{y}{=}1 \mid g_j)\big|. \tag{3}$$

Counterfactual fairness and group fairness capture different harms and can conflict.

3.2 Within-pair Contingency Cells

The contingency cells defined here are our primary instrument for surfacing the asymmetric dual effect of thinking on counterfactual fairness (§4.2). For each counterfactual pair we cross-tabulate the flip indicator under thinking against the flip indicator under the within-model non-thinking arm. The resulting four cells are labelled $\{a, b, c, d\}$ (Table 1): cell a contains pairs with no flip in either arm; cell b contains pairs flipped under non-thinking but agreeing under thinking, the *flips resolved by thinking*; cell c contains pairs flipped under thinking but agreeing under non-thinking, the *flips created by thinking*; cell d contains pairs flipped under both arms. Figure 1 illustrates a concrete cell-c instance: a counterfactual pair that agrees under the non-thinking arm but diverges under thinking, with the trace itself surfacing biased reasoning.

3.3 Bias Propagation Tool: CDPG$_k$

The Counterfactual Depth Probability Gap at depth k, CDPG$_k$, is our instrument for observing how counterfactual bias evolves along thinking depth: a per-pair signal that quantifies the counterfactual decision gap after the first k segments of the thinking trace. The construction adapts the prefix-truncation

		Think	
		no flip	*flip*
No-think	*no flip*	a: both agree	c: created
	flip	b: resolved	d: both flip

Table 1: **Within-pair contingency cells.** Each counterfactual pair is classified by whether a flip ($\hat{y}^a \neq \hat{y}^{a'}$) occurs under the thinking and non-thinking arms.

probing scaffold of Lanham et al. (2023), originally applied as an early-answering faithfulness test on chain-of-thought, to counterfactually paired inputs.

For a counterfactual query pair $(Q^\alpha, Q^{\alpha'})$ with thinking traces T^α and $T^{\alpha'}$, we partition each trace into $K{=}10$ equal-token segments, $T^\alpha = (S_1^\alpha, \ldots, S_K^\alpha)$ and likewise for $T^{\alpha'}$. For each depth $k \in \{1, \ldots, K\}$ and attribute value $A \in \{\alpha, \alpha'\}$ we re-feed the prefix $(Q^A, S_{1:k}^A, \text{suffix})$ to the model, where the suffix is the answer-elicitation template that closes the thinking block and opens the answer position (Appendix C). At the answer position we extract the logits for the <Yes> and <No> tokens only and apply softmax over this pair (not over the full vocabulary), giving

$$P_k^\alpha = P(\hat{y}{=}1 \mid Q^\alpha, S_{1:k}^\alpha), \tag{4}$$

$$P_k^{\alpha'} = P(\hat{y}{=}1 \mid Q^{\alpha'}, S_{1:k}^{\alpha'}). \tag{5}$$

and the per-pair, per-depth signal

$$\text{CDPG}_k^{(i)} = P_k^\alpha - P_k^{\alpha'}. \tag{6}$$

At $k{=}K$, $|\text{CDPG}_K^{(i)}|$ equals the thinking-arm per-pair counterfactual disparity $\mathcal{D}_{\text{disp}}^{(i)}$, so the $k{=}1{:}K$ progression decomposes the final disparity over reasoning depth. Figure 5 illustrates the full pipeline.

Interventional reading. Re-feeding the truncated prefix as a fresh context is what makes P_k^A interventional: it answers what decision the model would commit to if thinking were interrupted at depth k and the model were forced to answer immediately.

Aggregation. For pooled or stratified reporting we average $|\text{CDPG}_k^{(i)}|$ over pairs within a chosen subpopulation $\mathcal{S}$ (a contingency cell, a (model, dataset) cell, or the full pair pool), giving $\overline{|\text{CDPG}_k|} = \frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}} |\text{CDPG}_k^{(i)}|$. This aggregation is used to build the per-cell trajectories in §4.3.

3.4 Transition Matrix Tools

Group-level transition matrix T^A. Each instance yields a non-thinking prediction $\hat{y}_{\text{nothink}} \in \{Y, N\}$ and a thinking prediction $\hat{y}_{\text{think}} \in \{Y, N\}$, where Y, N represents Yes and No, respectively. Conditioning on the protected attribute A, the group-level transition matrix is

$$T_{ij}^A = P(\hat{y}_{\text{think}}{=}j \mid \hat{y}_{\text{nothink}}{=}i,\ A), \tag{7}$$

for $i, j \in \{Y, N\}$ and $A \in \{\alpha, \alpha'\}$, estimated by empirical conditional frequency. T^A characterises how individual predictions move from non-thinking to thinking for each protected attribute value, without reference to the counterfactual partner.

Pair-state transition matrix T^{pair}. Each counterfactual pair yields a non-thinking pair-state $s_{\text{nt}} \in \{\text{NN}, \text{YY}, \text{M}\}$ and a thinking pair-state $s_{\text{th}} \in \{\text{NN}, \text{YY}, \text{M}\}$, where NN denotes both sides predicted negative, YY both sides predicted positive, and M a mixed state (equivalently a counterfactual flip). The pair-state transition matrix is

$$T_{ij}^{\text{pair}} = P(s_{\text{th}}{=}j \mid s_{\text{nt}}{=}i), \tag{8}$$

for $i, j \in \{\text{NN}, \text{YY}, \text{M}\}$, estimated by empirical conditional frequency over the pooled pair population, and also stratified per (model, dataset) cell. T^{pair} characterises how the joint pair distribution moves between the two arms at the pair level. The contingency cells of §3.2 are off-diagonal blocks of T^{pair}: cell c (created flips) is the $\{\text{NN}, \text{YY}\}{\to}\text{M}$ mass, cell b (resolved flips) is the $\text{M}{\to}\{\text{NN}, \text{YY}\}$ mass, and the resolve-vs-create asymmetry of the dual effect is the asymmetry between these two off-diagonal blocks. By comparison with the group-level T^A above, T^{pair} captures the joint dynamics that T^A marginalises out: if T^α and $T^{\alpha'}$ are nearly identical, any asymmetry in the dual effect must originate at the joint pair-state level captured by T^{pair}.

4 Results

4.1 Experimental setup

We evaluate three high-stakes tabular decision datasets. **Adult** (Becker and Kohavi, 1996) predicts whether an individual's annual income exceeds \$50K, with *sex* as the protected attribute. **COMPAS** (Angwin et al., 2022) predicts two-year recidivism, with *race* as the protected attribute (binarised to White vs. Non-White). **Credit** (Yeh,

2009) predicts next-month credit-card default, with *sex* as the protected attribute. To prevent confounding between protected attributes, we drop the race column from Adult and the sex column from COMPAS so that each dataset isolates a single protected attribute under counterfactual substitution; Credit contains no second protected column. Each dataset is sampled at N=5,000 records and each record is paired with its protected-attribute counterfactual via fixed-template natural-language profiling (Appendix A).

We evaluate three open-weight 32B-class reasoning models that expose thinking tokens natively: QwQ-32B (Team, 2025), DeepSeek-R1-Distill-Qwen-32B (Guo et al., 2025), and Qwen3-32B (Yang et al., 2025). The within-model non-thinking baseline is implemented by inserting a closed empty `<think></think>` block. All models run at 4-bit AWQ quantisation on A100 80 GB GPUs with greedy decoding (T=0); CDPG force-answer probabilities are read directly from answer-token log-probabilities.

We retain only counterfactual pairs whose Yes/No answer parses on both sides of both arms; the effective N falls below 5,000 when the model emits no Yes/No token and therefore cannot be classified. The most severe drop is DeepSeek-R1-Distill-Qwen-32B on COMPAS (N=1,430), where the non-thinking arm refuses to answer on the majority of records; the remaining cases lie within a few records of 5,000 (Table 2).

4.2 The phenomenon: asymmetric dual effect of thinking

Group disparity gives an inconsistent signal. Figure 2 shows that $\Delta\mathcal{D}_{\text{group}}$ (pale bars) has no consistent sign across the nine (model, dataset) scenarios: thinking amplifies group disparity in four cases, mitigates it in four, and leaves it effectively unchanged in one. The magnitudes are also small ($|\Delta\mathcal{D}_{\text{group}}| \leq 0.035$, with six cases at ≤ 0.015). On QwQ-Adult $\Delta\mathcal{D}_{\text{group}} = -0.018$ (a single-group audit reads "thinking helps"); on QwQ-COMPAS $\Delta\mathcal{D}_{\text{group}} = +0.035$ (the same audit reads "thinking hurts"). The direction of the group-disparity verdict is dataset-dependent. This observation, complementing with Section 4.4, explains why prior work reaches opposing conclusions in fairness implications.

Counterfactual disparity is consistently amplified. By contrast, $\Delta\mathcal{D}_{\text{cf}}$ (dark bars in Figure 2) is positive in all $9/9$ cells, ranging from $+0.016$ (Qwen3-Credit) to $+0.122$ (R1-Distill-COMPAS). Counterfactual flip rate increases under thinking in every setting tested. To understand the source of this consistent amplification, a finer-grained diagnosis is needed.

Two direct contributors to the counterfactual amplification. Table 2 decomposes the counterfactual amplification into two contributors. *(i) Counterfactual bias creation outnumbers resolution*: $|c| > |b|$ in all nine scenarios, with the pooled gap 3508 vs 620 ($\approx 5.7\times$). *(ii) Within cell c, the per-pair counterfactual disparity is strengthened*: the mean $\overline{\mathcal{D}_{\text{cf}}}$ on cell-c pairs rises from 0.024 under nothink to 0.835 under thinking ($\approx 34\times$ pooled). Thinking thus creates more flips and commits to each created flip with markedly higher confidence; the cell-b resolution ($\overline{\mathcal{D}_{\text{cf}}}$ $0.143 \rightarrow 0.022$, $6.5\times$) is real but per-pair $\approx 5\times$ weaker than cell-c creation.

What remains for the dynamic analysis. Two questions follow from the dual-effect phenomenon and motivate the rest of this section. (i) Where in the thinking trace, and via what mechanism, does the dual effect emerge? (ii) Where, mechanistically, does the asymmetric transition arise that amplifies the counterfactual bias?

4.3 Bias propagation: the source of the dual effect

Bias dynamics. Figure 3 plots the per-cell mean $|\text{CDPG}_k|$ against normalised depth k/K for each of the nine (model, dataset) settings. The four cells split along a single axis: whether or not biased reasoning surfaces during thinking. Cells c and d, the two cells in which thinking ends with a counterfactual flip, see $|\text{CDPG}_k|$ rise toward saturation: once a bias signal surfaces mid-trace, the remaining depth amplifies it into a high-confidence disagreement. Cells a and b, the two cells in which thinking ends with agreement, see $|\text{CDPG}_k|$ stay near zero or decay toward it: when no biased reasoning surfaces, the same amplifier sharpens the prediction toward a confident fair outcome. The four trajectories are thus governed by one mechanism with two visible faces: *thinking commits each pair-side ever more confidently toward whichever boundary it is already drifting to*, and the contingency cell only records whether the two sides land on the same or opposite boundaries.

Model	Dataset	N	Contingency cells (counts) a	b	c	d	$\overline{\mathcal{D}_{\text{cf}}}$ in cell b nothink	think$_K$	$\overline{\mathcal{D}_{\text{cf}}}$ in cell c nothink	think$_K$
Qwen3-32B	Adult	5000	4531	75	362	32	0.219	0.048	0.055	**0.859**
	COMPAS	5000	4511	143	344	2	0.150	0.037	0.046	**0.781**
	Credit	4996	4867	23	105	1	0.075	0.012	0.023	**0.755**
QwQ-32B	Adult	5000	4396	152	443	9	0.131	0.001	0.029	**0.980**
	COMPAS	5000	4634	15	351	0	0.262	0.008	0.002	**0.891**
	Credit	4984	4265	121	595	3	0.156	0.016	0.007	**0.642**
R1-Distill-Qwen-32B	Adult	4999	4255	77	646	21	0.076	0.026	0.032	**0.918**
	COMPAS	1430	1247	4	179	0	0.085	0.003	0.042	**0.933**
	Credit	4995	4502	10	483	0	0.034	0.000	0.003	**0.791**
Pooled	**(9 cells)**	41404	37208	620	3508	68	0.143	0.022	0.024	**0.835**

Table 2: **Dual-effect contingency table with within-pair gap reshaping in cells b and c.** Cell counts a, b, c, d; rightmost columns give the mean per-pair counterfactual disparity $\overline{\mathcal{D}_{\text{cf}}}$ in cells b, c under the nothink and final-thinking arms ($\mathcal{D}_{\text{cf}}^{(i)} \equiv |\text{CDPG}_K^{(i)}|$, Eq. 6). N retains pairs parsing on both sides of both arms.

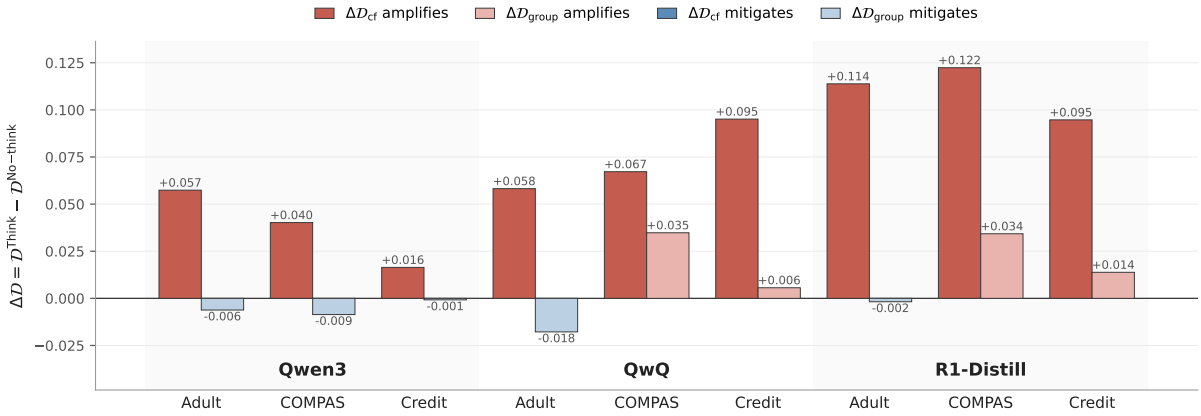

Figure 2: **Per-cell change under thinking of counterfactual disparity ($\Delta\mathcal{D}_{\text{cf}}$, Eq. 1) vs. group disparity ($\Delta\mathcal{D}_{\text{group}}$, Eq. 3).** Absolute Think/No-think values are reported in Table 7 (Appendix D).

Confidence dynamics. To understand where the bias dynamics in $|\text{CDPG}_k|$ come from, we decompose them into the two per-side $P(\text{Yes})$ trajectories that produce them (Figure 4, cells b and c). Within each cell, one side maintains its initial prediction throughout the trace (the *anchor*) and the other changes prediction (the *mover*). In cell b (flip $\rightarrow$ agree), the mover starts from the nothink-flip side, and moves to meet the anchor at the same boundary; the two sides converge in confidence and the within-pair gap collapses. In cell c (agree $\rightarrow$ flip), the mover starts at the same level as the anchor, and moves toward the opposite boundary; the two sides commit to opposite boundaries and the gap saturates. The depth-resolved confidence buildup in $|\text{CDPG}_k|$ is therefore a byproduct of thinking sharpening each side's prediction confidence: whether this sharpening creates bias (cell c) or resolves it (cell b) depends only on *whether the evidence the trace accumulates is itself biased*. The primary role of thinking is to strengthen prediction confidence; the unavoidable side effect is that when the accumulated evidence carries bias, the bias is sharpened in lockstep.

4.4 Pair-state asymmetry and group-level blindness

Asymmetry is not caused by group-level transition differences. A natural first hypothesis is that the asymmetry of $|c| > |b|$ under thinking comes from the protected attribute driving different transition rates for α and α'. Table 3 rules this out by evaluating the **group-level transition matrix** (§3.4): stratifying the nothink$\rightarrow$think transition matrix by the protected attribute, $T^{\alpha} \approx T^{\alpha'}$ on every (dataset, transition) entry, with signed gap $|T^{\alpha} - T^{\alpha'}| \leq 0.014$ uniformly. The post-think

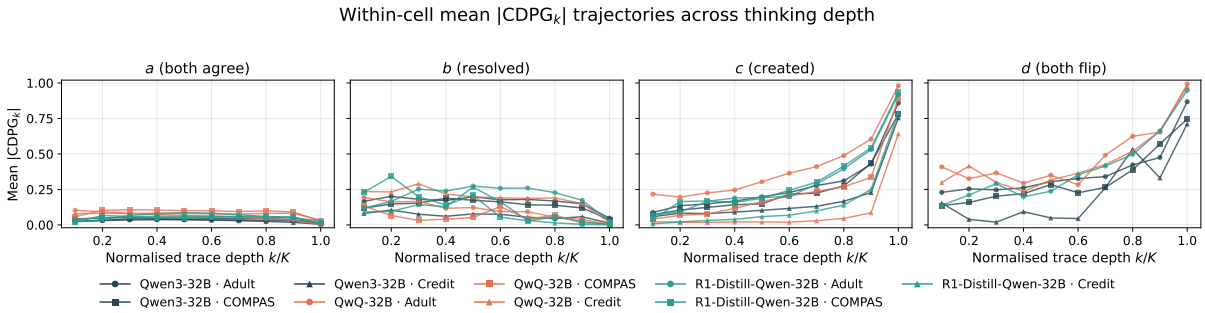

Figure 3: **The dual effect plays out consistently as the thinking trace deepens.** Mean $|\mathrm{CDPG}_k|$ as a function of normalised thinking depth k/K, computed within each contingency cell, pooled across the nine (model, dataset) settings.

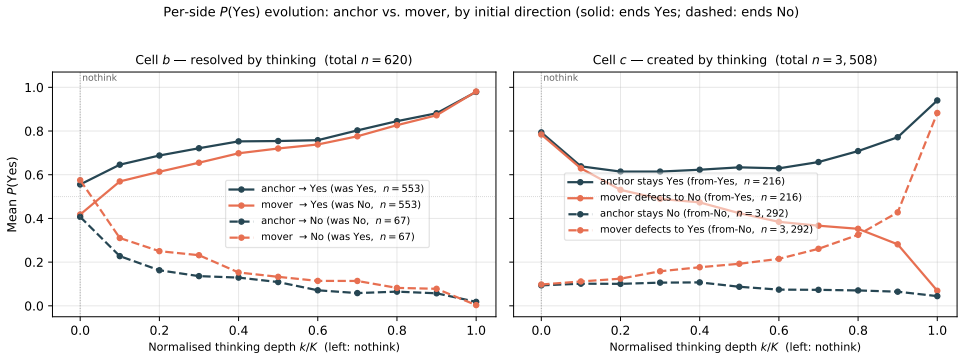

Figure 4: **Per-side $P(\mathrm{Yes})$ evolution through the thinking trace.** Cell b (left, resolved by thinking) and cell c (right, created by thinking). The leftmost point (k=0) is the nothink baseline. Pooled across the nine (model, dataset) settings.

dual effect asymmetry therefore cannot originate at the group-level transition where T^A is essentially balanced across $A \in \{\alpha, \alpha'\}$. A group-rate audit (such as group fairness) that reads only T^A is structurally blind to whatever drives the asymmetry. Moreover, thinking usually shifts the overall Yes-prediction rate on both sides (Table 3). *This shift could be misinterpreted by "bias mitigated" or "bias amplified" if we only study one side of the groups.* This gives another potential cause of why prior studies reach opposing conclusions.

Pair-state transition matrix T^{pair}: Asymmetry is rooted in starting-state population and transition probabilities. As a first-order analysis, assume per-side flips are independent within a pair conditional on the nothink starting state. Writing $p := P(Y|N)$ and $q := P(N|Y) = 1 - P(Y|Y)$ (Table 3; $T^\alpha \approx T^{\alpha'}$ lets us drop the side subscript), the expected cell-c and cell-b counts are

$$\mathbb{E}|c| = 2p(1-p)\,|\mathrm{NN}|_{\mathrm{nt}} + 2q(1-q)\,|\mathrm{YY}|_{\mathrm{nt}},$$
$$\mathbb{E}|b| = [\,p(1-q) + q(1-p)\,]\,|\mathrm{M}|_{\mathrm{nt}}.$$

Plugging the empirical p, q and nothink populations into these expressions already predicts $|c| > |b|$ on every cell (Table 10, "$|c|$ ind." vs. "$|b|$ ind." columns). Empirically, independence is not strictly satisfied: the two sides of a pair share most of their input prose and tend to co-flip, so the realised T^{pair} (Table 4 for Qwen3-32B, Tables 11, 12 for the other models) deviates from the independence formula (Appendix F). Within-pair correlation thus shifts the magnitudes but not the direction: $|c| > |b|$ holds in every cell, matching the independence prediction. The dual-effect asymmetry is therefore a structural consequence of any non-trivial transition probability acting on an imbalanced nothink population.

Directional decomposition of cell c and audit visibility. Cell c counts the pairs that flip under thinking but is silent on which side wins. Splitting cell c into α-favoured pairs ($c_{\alpha\uparrow}$: $\hat{y}^\alpha$=1, $\hat{y}^{\alpha'}$=0) and α'-favoured ones ($c_{\alpha'\uparrow}$, opposite direction) under the fixed convention α/α' = Male/Female (Adult, Credit), White/Non-White (COMPAS), adapting Prabhakaran et al. (2019), gives the signed

	Adult			COMPAS			Credit		
A	$P(Y\|Y)$	$P(Y\|N)$	$\Delta P(Y)$	$P(Y\|Y)$	$P(Y\|N)$	$\Delta P(Y)$	$P(Y\|Y)$	$P(Y\|N)$	$\Delta P(Y)$
α	0.889	0.076	+0.033	0.998	0.281	+0.174	0.908	0.012	−0.026
α'	0.903	0.077	+0.032	0.997	0.272	+0.182	0.912	0.011	−0.025
$\boldsymbol{\Delta = T^{\alpha} - T^{\alpha'}}$	**−0.014**	**−0.001**	**+0.001**	**+0.001**	**+0.009**	**−0.008**	**−0.004**	**+0.002**	**−0.001**

Table 3: **Group-level transition matrix T^A for Qwen3-32B, stratified by protected attribute.** α/α' = Male/Female (Adult, Credit), White/Non-White (COMPAS); $P(Y|Y)$ and $P(Y|N)$ denote the probability of transitioning to a Y prediction from a Y or N prediction respectively. $\Delta P(Y) = P(Y)_{\text{th}} - P(Y)_{\text{nt}}$ measures the overall shift in the Yes-prediction rate (Y = earns over $50K for Adult, will recidivate for COMPAS, will default for Credit). Other models in Appendix E.

	Adult				COMPAS				Credit			
from	NN	YY	M	n	NN	YY	M	n	NN	YY	M	n
NN	0.894	0.035	0.071	3,720	0.686	0.207	0.108	3,157	0.981	0.004	0.015	3,170
YY	0.050	0.866	0.084	1,173	0.001	0.996	0.002	1,698	0.068	0.899	0.033	1,802
M	0.290	0.411	0.299	107	0.000	0.986	0.014	145	0.917	0.042	0.042	24
total	3,415	1,191	**394**	5,000	2,167	2,487	**346**	5,000	3,256	1,634	**106**	4,996

Table 4: **Pair-state transition matrix T^{pair} for Qwen3-32B (forward: nothink → think).** Within-row conditional probabilities; row counts n and column totals shown. Highlighted M column totals are the post-think cell-c/d population sizes.

Model	**Dataset**	$\|c\|$	$c_{\alpha\uparrow}$	$c_{\alpha'\uparrow}$	asym(c)	δ_c (%)
Qwen3	Adult	362	186	176	0.03	+0.20
	COMPAS	344	170	174	0.01	−0.08
	Credit	105	61	44	0.16	+0.34
QwQ	Adult	443	231	212	0.04	+0.38
	COMPAS	351	82	269	0.53	−3.74
	Credit	595	232	363	0.22	−2.63
R1-Distill	Adult	646	337	309	0.04	+0.56
	COMPAS	179	63	116	0.30	−3.71
	Credit	483	276	207	0.14	+1.38
Pooled	**(9 cells)**	3508	1638	1870	0.07	−0.56

Table 5: **Directional decomposition of created flips (cell c).** δ_c reported in %; definitions in §4.4. Shaded cells: the four (model, dataset) settings with $\Delta\mathcal{D}_{\text{group}}>0$ in Table 7.

cell-c tilt $\delta_c = (c_{\alpha\uparrow} - c_{\alpha'\uparrow})/N$ (Table 5). Cells a and b have both sides of every pair agreeing under thinking, so they contribute zero to the think-arm marginal difference between α and α'; the cell-level contribution to $\Delta\mathcal{D}_{\text{group}}$ therefore originates entirely from δ_c and an analogous cell-d tilt that is small because $|d|$ itself is small (≤ 68 pairs pooled). Empirically, the four cells with the largest $|\delta_c|$ are the four cells where $|\Delta\mathcal{D}_{\text{group}}|$ in Figure 2 clears the visible-signal threshold; the remaining five carry $|\delta_c| \leq 0.56\%$ and yield $|\Delta\mathcal{D}_{\text{group}}|$ at the noise floor. The five noise-floor cells are not bias-free: they are cells with large $|c|$ whose tilt happens to split symmetrically between α and α'. A group-rate audit reads them as $\Delta\mathcal{D}_{\text{group}}\approx 0$ either way, so visible group disparity is the special case of asymmetric cell-c tilt.

5 Conclusion

Thinking in reasoning models does not simply resolve counterfactual bias; it acts as a confidence amplifier that both resolves and creates counterfactual flips through one mechanism. CDPG exposes this along the depth of the thinking trace: each side of a counterfactual pair is sharpened toward whichever boundary it is already drifting to, so the same dynamic resolves a flip on some pairs and creates one on others, depending only on whether the accumulated reasoning is itself biased. BTM then locates the asymmetric dual effect at the pair-state level: per-group marginal transitions shift in near-lockstep across the two protected-attribute values, so the cross-group disagreement that drives the asymmetry is invisible to any single-group audit and only surfaces in the joint pair-state transition. This is where counterfactual bias originates under thinking, and where fairness audits of reasoning models should look.

6 Limitations

Scope. This study evaluates counterfactual fairness on three tabular binary classification tasks with binary protected attributes (sex and race). Three lines of investigation are outside the present scope: group-fairness experiments on the same model–dataset settings (planned follow-up); extension to free-form generation tasks, multi-class attributes, and multilingual settings; and intervention-based mitigation. The paper is diagnostic: it introduces two dynamic instruments (bias propagation and the pair-state transition matrix) and characterises a previously invisible failure mode; evaluation of concrete mitigation strategies against these instruments is left to future work.

Methodological assumptions. Bias propagation uses $K{=}10$ equal-token segments of the thinking trace, a granularity that trades depth resolution against compute cost (900,000 forward passes for the full study: 3 models $\times$ 3 datasets $\times$ 5,000 pairs $\times$ 2 sides $\times$ 10 depths). We assume the dual-effect trend is K-invariant (that a coarser or finer segmentation would preserve the direction and approximate magnitude of the cell-b and cell-c gap motions), but this assumption has not been empirically stress-tested. The within-model non-thinking baseline injects an empty closed `<think></think>` block; a small fraction of generations may still emit reasoning tokens after this prefix, and we report aggregates only over completions that respect the instruction. All inference uses 4-bit AWQ quantisation.

Disposition. The study is diagnostic, not mitigating. The two dynamic instruments and the within-pair contingency decomposition jointly provide the lens and the finding; evaluation of concrete mitigation candidates against the propagation and transition-matrix objectives is the natural follow-up and is outside the scope of this paper.

Ethics Statement

The datasets (Adult, COMPAS, Credit) are public and extensively analysed in the fairness literature; we use them for measurement only, not training or deployment. Our finding, that thinking-token reasoning can construct high-confidence counterfactual disagreements at $5\times$ the per-pair magnitude of corresponding reductions, itself signals that deployment of such models in high-stakes settings should not lean on the *appearance* of deliberation as fairness assurance. We do not release new model weights.

Use of AI Assistants

AI coding assistants were used to help develop and refactor parts of the experimental codebase (data-loading utilities, inference pipelines, and plotting scripts), with all logic reviewed by the authors. AI tools were also used for limited language polishing of the manuscript (grammar, phrasing, and clarity); all scientific claims, analyses, and conclusions are the authors' own.

References

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2022. Machine bias. In *Ethics of data and analytics*, pages 254–264. Auerbach Publications.

Molly Apsel and Michael N Jones. 2026. Inference-time reasoning selectively reduces implicit social bias in large language models. *arXiv preprint arXiv:2602.04742*.

Barry Becker and Ronny Kohavi. 1996. Adult. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20.

Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. 2020. Fairlearn: A toolkit for assessing and improving fairness in ai. Technical Report MSR-TR-2020-32, Microsoft.

Riccardo Cantini, Nicola Gabriele, Alessio Orsino, and Domenico Talia. 2025. Is reasoning all you need? probing bias in the age of reasoning language models. *arXiv preprint arXiv:2507.02799*.

Simon Caton and Christian Haas. 2024. Fairness in machine learning: A survey. *ACM Comput. Surv.*, 56(7).

Valeriia Cherepanova, Chia-Jung Lee, Nil-Jana Akpinar, Riccardo Fogliato, Martin Bertran Lopez, Michael Kearns, and James Zou. 2025. Improving llm group fairness on tabular data via in-context learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pages 579–590.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational linguistics*, 50(3):1097–1179.

Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H Chi, and Alex Beutel. 2019. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 219–226.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Sanchit Kabra, Akshita Jha, and Chandan K Reddy. 2025. Reasoning towards fairness: mitigating bias in language models through reasoning-guided finetuning. *arXiv preprint arXiv:2504.05632*.

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

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, and 1 others. 2023. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*.

Kyle Moore, Jesse Roberts, Thao Pham, and Douglas Fisher. 2024. Reasoning beyond bias: A study on counterfactual prompting and chain of thought reasoning. *arXiv preprint arXiv:2408.08651*.

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

Vinodkumar Prabhakaran, Ben Hutchinson, and Margaret Mitchell. 2019. Perturbation sensitivity analysis to detect unintended model biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5740–5745.

Omar Shaikh, Hongxin Zhang, William Held, Michael Bernstein, and Diyi Yang. 2023. On second thought, let's not think step by step! bias and toxicity in zero-shot reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4454–4470.

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

Qwen Team. 2025. Qwq-32b: Embracing the power of reinforcement learning.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965.

Xuyang Wu, Jinming Nian, Ting-Ruen Wei, Zhiqiang Tao, Hsin-Tai Wu, and Yi Fang. 2025. Does reasoning introduce bias? a study of social bias evaluation and mitigation in llm reasoning. *arXiv preprint arXiv:2502.15361*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

I-Cheng Yeh. 2009. Default of Credit Card Clients. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C55S3H.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–38.

A Natural-Language Profile Generation

Each tabular record is rendered into prose by a *deterministic Python template*. Because the template is a pure function of the row dict, matched (`norm`, `cf`) pairs are byte-identical except for the protected-attribute adjective and the pronouns that grammatically depend on it; no phrasing-level variance can confound the counterfactual signal.

Pronoun handling has two modes: `gendered` (he/his vs. she/her, flipping with `sex`; used for Adult and Credit) and `neutral` (singular *they* on both sides; used for COMPAS, whose protected attribute `race` has no English pronoun system). Verbs and possessives are toggled accordingly. The templates are reproduced verbatim below.

Adult template.

```
This person is a {age}-year-old {sex}.
     {His} marital status is
{marital_status}. {He} work{s} in the {
     workclass} sector as a
{occupation} worker, {hours_per_week}
     hours per week. {He}
{has} a {education} education (
     education number {education_num}).
{He} report{s} ${capital_gain} in
     capital gains and
${capital_loss} in capital losses. {He}
     {is} from {native_country}.
```

COMPAS template.

```
This defendant is a {age}-year-old {
    race} person. They have
{juvenile_felony_count} prior juvenile
     felony offense(s),
{juvenile_misdemeanor_count} prior
    juvenile misdemeanor offense(s),
and {juvenile_other_count} other prior
     juvenile offense(s). They
have {priors_count} prior adult
    criminal offense(s). The current
charge is a {charge_degree}.
```

Credit template.

```
This person is a {age}-year-old {sex}.
     {His} marital status is
{marital_status} and {his} education
    level is {education_label}.
{He} {has} a credit limit of NT${
    credit_limit}. {His} statement
balances over the past 6 months were ${
    bill_1}, ${bill_2}, ...,
${bill_6}, and {he} paid ${paid_1}, ${
    paid_2}, ..., ${paid_6},
respectively. {He} {had_or_had_no}
    late payments in the past
6 months.
```

where `education_label` maps the UCI numeric education code via {1:graduate school, 2:university, 3:high school, 4:other education} and unmapped codes render as `unspecified education`.

Counterfactual instances are generated by code: a single dictionary field is flipped (`sex` ∈ {Male, Female} for Adult and Credit; `race` ∈ {White, Non-White} for COMPAS) and the same template is re-rendered. Because the template is a pure function of the row dict, all non-protected substrings in the (`norm`, `cf`) prose are byte-identical except for the protected-attribute adjective and (for `gendered`) the pronouns that grammatically depend on it.

B CDPG measurement pipeline

C Inference Prompt Format

Table 6 shows the prompts used for the thinking and non-thinking conditions on the Adult dataset. The dataset-specific question line is the only piece that varies across datasets: `"Does this person earn more than $50K per year?"` for Adult, `"Will this person re-offend within two years?"` for COMPAS, and `"Will this person default on their next credit card payment?"` for Credit.

D Absolute disparity values per (model, dataset) cell

Figure 2 in the main text reports only the per-cell change $\Delta = \mathcal{D}^{\text{Think}} - \mathcal{D}^{\text{No-think}}$ for compactness. Table 7 reports the corresponding absolute disparities $\mathcal{D}_{\text{cf}}$ and $\mathcal{D}_{\text{group}}$ under each arm.

E Group-level transition matrix: detailed breakdown

Table 3 in the main text reports the group-level transition matrix $T^A_{ij} = P(\hat{y}_{\text{think}}{=}j \mid \hat{y}_{\text{nothink}}{=}i, A)$ for $A \in \{\alpha, \alpha'\}$ within each dataset for Qwen3-32B. Tables 8 and 9 extend the same stratification to QwQ-32B and DeepSeek-R1-Distill-Qwen-32B, including the per-group marginal drift $\Delta P(Y) = P(Y)_{\text{th}} - P(Y)_{\text{nt}}$. Across all three models the per-(model, dataset, direction) gap between T^{α} and $T^{\alpha'}$ remains tightly bounded: $|\Delta| \leq 0.014$ on Qwen3-32B, $|\Delta| \leq 0.017$ on QwQ-32B (largest on Credit $P(Y|Y)$), and $|\Delta| \leq 0.012$ on R1-Distill-Qwen-32B. These gaps are an order of magnitude smaller than the cell-c pair-state tilts documented in the main text (e.g., QwQ-COMPAS asym=0.53); the pair-level cell-c tilts therefore cannot be inherited from group-level bias, since the joint property captured by T^{pair} (Appendix G) is what places the constraint on whether the two paired instances of a (norm, cf) pair transition consistently with each other.

Pooled across 84,428 per-instance observations under a per-side parseable filter (an instance is retained iff its own think and nothink arms both emit Yes/No; this is looser than the pair-level filter used in the main text, since one side of a pair can be unparseable without dropping the other), the marginal $T = \mathbb{E}_A[T^A]$ is

	think=Y	think=N	total
nothink=Y	13,891 (0.952)	705 (0.048)	14,596
nothink=N	20,216 (**0.289**)	49,616 (0.711)	69,832

Thinking exerts a strong group-level $N{\to}Y$ push: the $N{\to}Y$ count outnumbers the $Y{\to}N$ count by $20{,}216/705 = 28.7\times$, and the two off-diagonals of T are themselves asymmetric (0.048 vs. 0.289, a $6\times$ ratio). Thinking moves predictions toward Y for both protected attribute values, but operates more aggressively on N baselines than on Y baselines.

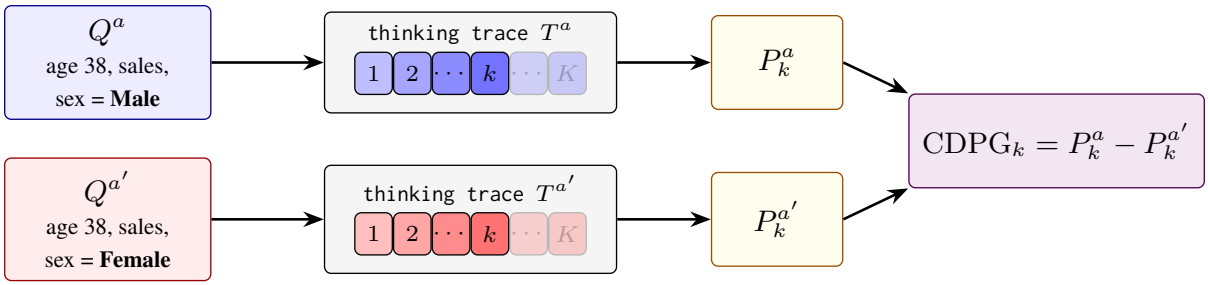

Figure 5: **CDPG measurement pipeline.** Counterfactual paired inputs $(Q^\alpha, Q^{\alpha'})$ are passed through the same reasoning model; each thinking trace is split into $K{=}10$ equal-token segments. At every prefix depth k we read the answer-token softmax $P_k^A{=}P(\hat{y}{=}1 \mid Q^A, S_{1:k}^A)$ and report $\mathrm{CDPG}_k{=}P_k^\alpha{-}P_k^{\alpha'}$.

Condition	**Example prompt (Adult)**

Thinking mode	`{description}`
	
	`Question: Does this person earn more than $50K per year?`
	`Answer with a single word: Yes or No.`

Non-thinking mode	`[Same user message as above. The chat template is rendered locally and the assistant turn is pre-filled with a closed empty <think>\n\n</think>\n\n block, so generation resumes after </think> and the model skips the reasoning phase.]`

Table 6: Prompt formats for the thinking and non-thinking arms on the Adult dataset. `{description}` is the natural-language profile from Appendix A.

F Independence-bound check per (model, dataset) cell

Table 10 compares the independence-based predictions for $|c|$ and $|b|$ (§4.4) against the empirical counts in each of the nine (model, dataset) cells. For each cell, $p = T_{NY}^A$ and $q = T_{YN}^A$ are read from the group-level transition matrices (Tables 3, 8, 9, averaged across α and α' since $T^\alpha \approx T^{\alpha'}$), and the nothink-state populations $|\mathrm{NN}|_{\mathrm{nt}}, |\mathrm{YY}|_{\mathrm{nt}}, |\mathrm{M}|_{\mathrm{nt}}$ are read from the row totals of the pair-state matrices (Tables 4, 11, 12).

Two consistent patterns appear. First, the directional inequality $|c| > |b|$ holds under both independence and the data on every cell: even purely independent per-side flips would produce the dual-effect asymmetry, given the observed nothink-state population skew. Second, the independence prediction is uniformly looser than the data: it overpredicts $|c|$ (2–7×) and underpredicts $|b|$ (2–10×), so the empirical $|c|/|b|$ ratios are 5–100× smaller than the independence-bound ratios. This gap quantifies within-pair flip correlation: two instances of a counterfactual pair share most of their input prose and therefore tend to co-flip rather than flip independently. This correlation acts in the direction of dampening the base-count asymmetry, but cannot reverse it.

G Per-model pair-state transition matrices

The pair-state transition matrices T^{pair} for QwQ-32B and DeepSeek-R1-Distill-Qwen-32B are presented here, along with the per-side stratifications. The Qwen3-32B counterpart is given in the main text (Table 4). The qualitative pattern reproduces in every (model, dataset) cell: NN dominates the starting distribution, NN → M is the dominant cell-c channel, and the NN → YY en-bloc rate varies by model.

Model heterogeneity is in the channel mix. All three models exhibit the same $|c| \gg |b|$ mechanism (§4.4); what varies is the relative weight of two channels out of NN: en-bloc push to YY versus split into M. Qwen3-32B is the most NN-inert, with NN-row stickiness 89%/69%/98% across Adult/COMPAS/Credit (Table 4) and the

Model	Dataset	N	Think		No-think		Δ	
			$\mathcal{D}_{\text{cf}}$	$\mathcal{D}_{\text{group}}$	$\mathcal{D}_{\text{cf}}$	$\mathcal{D}_{\text{group}}$	$\Delta\mathcal{D}_{\text{cf}}$	$\Delta\mathcal{D}_{\text{group}}$
Qwen3	Adult	5000	0.079	0.006	0.021	0.012	+0.058	−0.006
	COMPAS	5000	0.069	0.001	0.029	0.010	+0.040	−0.009
	Credit	4996	0.021	0.004	0.005	0.004	+0.016	0.000
QwQ	Adult	5000	0.090	0.004	0.032	0.021	+0.058	−0.017
	COMPAS	5000	0.070	0.037	0.003	0.003	+0.067	+0.034
	Credit	4984	0.120	0.026	0.025	0.021	+0.095	+0.005
R1-Distill	Adult	4999	0.133	0.006	0.020	0.008	+0.113	−0.002
	COMPAS	1430	0.125	0.037	0.003	0.003	+0.122	+0.034
	Credit	4995	0.097	0.014	0.002	0.000	+0.095	+0.014

Table 7: **Counterfactual disparity $\mathcal{D}_{\text{cf}}$ (Eq. 1) and demographic-parity gap $\mathcal{D}_{\text{group}}$ (Eq. 3) under the Think and No-think arms, with within-cell change $\Delta = \mathcal{D}^{\text{Think}} - \mathcal{D}^{\text{No-think}}$.** Positive Δ amplifies the bias, negative Δ mitigates (palette from Table 1). N matches Table 2.

	Adult			COMPAS			Credit		
A	$P(Y\|Y)$	$P(Y\|N)$	$\Delta P(Y)$	$P(Y\|Y)$	$P(Y\|N)$	$\Delta P(Y)$	$P(Y\|Y)$	$P(Y\|N)$	$\Delta P(Y)$
α	0.980	0.300	+0.249	1.000	0.525	+0.525	0.977	0.387	+0.361
α'	0.988	0.302	+0.255	1.000	0.515	+0.513	0.961	0.381	+0.355
$\boldsymbol{\Delta = T^{\alpha} - T^{\alpha'}}$	**−0.008**	**−0.002**	**−0.006**	**+0.000**	**+0.010**	**+0.011**	**+0.017**	**+0.006**	**+0.006**

Table 8: **Group-level transition matrix T^A for QwQ-32B, stratified by protected attribute.** Same conventions as Table 3.

en-bloc NN→YY channel never exceeding 21%. QwQ-32B is en-bloc-aggressive, with NN→YY at 24%/48%/31% across the three datasets (Table 11). DeepSeek-R1-Distill is split-aggressive, with NN → M peaking at 16.4% on Adult (Table 12).

	Adult			COMPAS			Credit		
A	$P(Y\|Y)$	$P(Y\|N)$	$\Delta P(Y)$	$P(Y\|Y)$	$P(Y\|N)$	$\Delta P(Y)$	$P(Y\|Y)$	$P(Y\|N)$	$\Delta P(Y)$
α	0.968	0.196	+0.138	1.000	0.215	+0.203	1.000	0.382	+0.381
α'	0.977	0.201	+0.147	1.000	0.203	+0.193	1.000	0.392	+0.391
$\boldsymbol{\Delta = T^{\alpha} - T^{\alpha'}}$	**−0.009**	**−0.005**	**−0.009**	**+0.000**	**+0.012**	**+0.010**	**+0.000**	**−0.010**	**−0.010**

Table 9: **Group-level transition matrix T^A for DeepSeek-R1-Distill-Qwen-32B, stratified by protected attribute.** Same conventions as Table 3.

		$\|c\|$ **(created)**			$\|b\|$ **(resolved)**			$\|c\|/\|b\|$	
Model	Dataset	ind.	actual	ratio	ind.	actual	ratio	ind.	actual
Qwen3-32B	Adult	744	363	2.05×	18	75	0.24×	41	4.8
Qwen3-32B	COMPAS	1272	344	3.69×	40	143	0.28×	32	2.4
Qwen3-32B	Credit	367	107	3.43×	2	23	0.10×	153	4.7
QwQ-32B	Adult	1768	444	3.98×	50	152	0.33×	36	2.9
QwQ-32B	COMPAS	2487	349	7.13×	8	15	0.52×	319	23.3
QwQ-32B	Credit	2198	598	3.68×	49	121	0.40×	45	4.9
R1-Distill	Adult	1246	649	1.92×	21	77	0.27×	59	8.4
R1-Distill	COMPAS	448	179	2.50×	1	4	0.21×	560	44.8
R1-Distill	Credit	2359	487	4.84×	4	10	0.39×	609	48.7

Table 10: **Independence-bound check per (model, dataset) cell.** Independence predictions use $\mathbb{E}|c| = 2p(1-p)\,|\mathrm{NN}|_{\mathrm{nt}} + 2q(1-q)\,|\mathrm{YY}|_{\mathrm{nt}}$ and $\mathbb{E}|b| = [p(1-q)+q(1-p)]\,|\mathrm{M}|_{\mathrm{nt}}$ with $p = P(Y|N)$ and $q = P(N|Y)$ read from Tables 3 (Qwen3-32B), 8 (QwQ-32B), 9 (R1-Distill). Actual counts are read from Tables 4, 11, 12. Across all nine cells, independence *overpredicts* $|c|$ by 2–7× and *underpredicts* $|b|$ by 2–10×, but the directional inequality $|c| > |b|$ holds under both independence and the data. The empirical ratios $|c|/|b|$ are uniformly smaller than the independence prediction (e.g., 4.8 vs. 41 on Qwen3-Adult): within-pair flip correlation dampens the base-count asymmetry but cannot reverse it.

	Adult					COMPAS					Credit				
from	NN	YY	YN	NY	n	NN	YY	YN	NY	n	NN	YY	YN	NY	n
NN	0.659	0.236	**0.055**	**0.050**	4,149	0.445	0.484	**0.016**	**0.054**	4,981	0.561	0.312	**0.050**	**0.077**	4,616
YY	0.004	0.984	**0.006**	**0.006**	690	0.000	1.000	**0.000**	**0.000**	4	0.008	0.943	**0.012**	**0.037**	244
YN	0.045	0.888	0.030	0.037	134	0.000	1.000	0.000	0.000	14	0.009	0.965	0.009	0.018	114
NY	0.000	1.000	0.000	0.000	27	0.000	1.000	0.000	0.000	1	0.000	1.000	0.000	0.000	10
total	2,743	1,805	235	217	5,000	2,218	2,431	82	269	5,000	2,594	1,792	233	365	4,984

Table 11: **Pair-state transition matrix T^{pair} for QwQ-32B, per dataset (forward: nothink → think), with** M **split into** YN **and** NY**.** Compared with Qwen3-32B (Table 4), QwQ runs hotter on every channel: the en-bloc NN → YY push peaks at 48.4% on COMPAS (half the originally-agree-No pairs are pushed wholesale to agree-Yes), and within M the cell-c split tilts strongly toward Non-White on COMPAS (NN → NY = 0.054 vs NN → YN = 0.016) and toward Female on Credit (NN → NY = 0.077 vs NN → YN = 0.050); the YY side reinforces the Credit tilt (YY → NY = 0.037 vs YY → YN = 0.012).

	Adult					COMPAS					Credit				
from	NN	YY	YN	NY	n	NN	YY	YN	NY	n	NN	YY	YN	NY	n
NN	0.728	0.108	**0.087**	**0.078**	3,715	0.726	0.142	**0.046**	**0.086**	1,356	0.565	0.338	**0.056**	**0.042**	4,971
YY	0.007	0.964	**0.012**	**0.018**	1,186	0.000	1.000	**0.000**	**0.000**	70	0.000	1.000	**0.000**	**0.000**	14
YN	0.059	0.735	0.103	0.103	68	–	–	–	–	0	0.000	1.000	0.000	0.000	5
NY	0.100	0.667	0.133	0.100	30	0.000	1.000	0.000	0.000	4	0.000	1.000	0.000	0.000	5
total	2,718	1,614	348	319	4,999	985	266	63	116	1,430	2,808	1,704	276	207	4,995

Table 12: **Pair-state transition matrix T^{pair} for DeepSeek-R1-Distill-Qwen-32B, per dataset (forward: no-think $\rightarrow$ think), with M split into YN and NY.** R1-Distill has the highest aggregated $\mathrm{NN} \rightarrow \mathrm{M}$ rates of the three models (Adult: $\mathrm{NN} \rightarrow \mathrm{YN} + \mathrm{NN} \rightarrow \mathrm{NY} = 0.165$, the per-model peak); the SFT-distillation regime is the most aggressive at producing pair-level splits rather than en-bloc pushes. The COMPAS row reflects $\sim 72\%$ refusal under R1-Distill's prompt sensitivity (only 1,430 parseable pairs); on the parseable subset NN tilts strongly toward Non-White ($\mathrm{NN} \rightarrow \mathrm{NY} = 0.086$ vs $\mathrm{NN} \rightarrow \mathrm{YN} = 0.046$), matching QwQ on COMPAS. On Credit, the direction reverses ($\mathrm{NN} \rightarrow \mathrm{YN} = 0.056$ vs $\mathrm{NN} \rightarrow \mathrm{NY} = 0.042$, Male-leaning), opposite to QwQ.