Deeper Than the Guardrails: Demographic Bias Survives Refusal-Direction Removal in Standardized Psychiatric Screening

A Ten-Pair Audit of Abliterated Open-Weight Models Against Externally-Grounded Population Norms

Patrick Keough
Independent Researcher
pskeough@gmail.com

Preprint, July 2026

Abstract

Community modification of open-weight language models has far outpaced any documented evaluation of its clinical consequences: "abliteration", the weight orthogonalization against the refusal direction identified by Arditi et al. [2024], is widely distributed even as these models increasingly enter mental-health screening contexts. In this paper we audit ten base/abliterated model pairs (Ministral, Qwen-3, Gemma, GLM-4, Phi-4, Qwen-2.5, Phi-3-medium, Llama-3.1, DeepSeek-R1-Distill, Yi-1.5; five Western and five Eastern; 8–14B parameters) across four validated screening instruments (PHQ-8, GAD-7, AUDIT-C, PCL-5), 120 demographic cells, two vignette conditions, and five iterations per cell, producing 95,920 trial-instrument observations. Cell-level PHQ-8 expected values are graded against author-derived, survey-weighted NHANES 2005–2018 population norms (the corrected v2 gold ground truth, validated against Brody et al. [2018]); GAD-7, AUDIT-C, and PCL-5 carry no revalidated severity benchmark and are reported with that caveat. We introduce a paired consensus-survival test alongside a Murphy-decomposed Brier analysis that separates calibration from discrimination. Twenty-one cross-model consensus over-pathologization cells reduce by 2.13 PHQ-8 points under abliteration ($t = -21.5$, $p = 1.2 \times 10^{-5}$; 21 of 21 in the reducing direction), yet the cisgender, benchmarkable cells remain far above the gold NHANES population norm after abliteration (residual $\approx +9$ PHQ-8 points against the NHANES Low-SES mean of 4.25). Measured against the corrected v2 gold ground truth, over-pathologization of cisgender low-SES cohorts is large and robust; transgender and multiracial cohorts have no representative severity benchmark (NHANES carries no gender-identity field), so their comparably elevated scores are reported descriptively rather than as residuals. (An earlier draft reported that external grounding *localized* the bias to cisgender rather than transgender users; resting on a transgender "elevation offset" since withdrawn as non-representative, that comparison has been removed. The honest statement is that no severity residual is computable for transgender cohorts.) Across 40 (pair × instrument) combinations, abliteration improves both Brier reliability and resolution in only 8 cases (20%); in 21 cases (52%) it worsens both. Through convergent classification across six independent psychometric metrics we identify three reproducible abliteration regimes: preservation, compression, and correction. We conclude that bias-mitigation claims for community-modified open weights require multi-metric, externally-grounded evaluation; the data does not support describing abliteration as either "debiasing" or "destabilizing" uniformly.

Ground-truth revision note (2026-07-21). Population ground truth has been migrated to the corrected PsychBench v2 gold standard—author-derived weighted PHQ-8 means from NHANES 2005–2018 microdata, dual-validated (Patel 2019 to ±0.02; Brody 2018 exact). This supersedes the additive offset model (§4.3 appendix) used in the original preprint, whose transgender "elevation offset" [Kidd et al., 2023, Hajek et al., 2023] and multiracial offset are withdrawn: NHANES has no gender-identity field and no representative multiracial norm, so no severity residual is computable for those cohorts. The gold standard covers PHQ-8 depression only, so GAD-7/AUDIT-C/PCL-5 residual claims carry no revalidated population anchor. **One headline changed**: the RQ3 claim that external grounding localizes over-pathologization to cisgender (not transgender) users is withdrawn—it was an artifact of the fabricated trans offset. All ground-truth-independent results (consensus survival, Brier dissociation, the three regimes, variance partitioning) are unaffected. Old→new values: `src/analysis/VerifiedResults_v3/gt_migration_v2/GT_MIGRATION_v2_MATCH_CORRECTED.md`.

1 Introduction

Open-weight language models from Meta, Mistral, Microsoft, Alibaba, and other vendors now power a growing share of mental-health applications: self-referral chatbots [Habicht et al., 2024], screening assistants [Miner et al., 2025], and administrative scoring tools [Hua et al., 2025]. Evaluation has not kept pace with this deployment. The audit literature has documented demographic biases in single-model evaluations of medical question-answering [Schmidgall et al., 2024, Rawat et al., 2024] and of race-based clinical reasoning [Omiye et al., 2023]; multi-architecture audits of community-modified weights remain sparse.

One widely circulated modification is "abliteration," the weight-orthogonalization technique that removes the refusal-mediating direction from a model's residual stream [Arditi et al., 2024]. Exploiting the finding that refusal behavior is approximately linear in the activation space [Marshall et al., 2024], and that representations can be partially decomposed into interpretable directions [Templeton et al., 2024, Zou et al., 2023], the technique has spread well past its research origins: Hugging Face hosts hundreds of abliterated checkpoints (mlabonne, huihui-ai, bartowski), and community forums discuss their use in production deployments [Labonne, 2024, r/LocalLLaMA Community, 2024]. To our knowledge, the downstream effect on demographic calibration in clinical scoring tasks has not been systematically measured.

In this paper we audit ten base/abliterated pairs at common quantization (Q4_K_M), administering four validated screening instruments to 120 demographic cells under two vignette conditions. The pairs span Western and Eastern model families, three size classes, and two lineages (instruct, reasoning). For PHQ-8, ground truth is anchored to author-derived, survey-weighted NHANES 2005–2018 microdata norms (validated against Brody et al. [2018]). The remaining instruments, along with the transgender and multiracial cohorts, have no revalidated population benchmark and are reported descriptively: a constraint the earlier external-anchor list (NHIS [Terlizzi and Schiller, 2024], NSDUH [Substance Abuse and Mental Health Services Administration, 2024], NESARC-III [Goldstein et al., 2016, Roberts et al., 2011], TransPop [Kidd et al., 2023]) could not honestly discharge.

The research questions are:

RQ1. Do cross-model consensus over-pathologization signals persist under weight-orthogonalization?

RQ2. Does abliteration improve clinical discrimination, calibration, or neither?

RQ3. When evaluated against population norms, where is the over-pathologization actually localized?

RQ4. Do abliteration effects depend on architecture, alignment family, or size class?

This work makes four contributions: (i) the first multi-architecture audit of weight-orthogonalization effects on demographic calibration in PHQ-8, GAD-7, AUDIT-C, and PCL-5; (ii) a paired consensus-survival test and a Murphy-decomposed proper-scoring analysis that separates calibration from discrimination effects; (iii) author-derived, survey-weighted NHANES 2005–2018 PHQ-8 population anchors (v2 gold) with released derivation scripts; and (iv) open release of the analysis suite, the 95,920-observation dataset, the pre-registration, and the reproducible report source.

2 Background

2.1 Refusal direction and abliteration

Arditi et al. [2024] showed that the refusal behavior of instruction-tuned models is mediated by a single direction in the residual stream, identified by the mean-difference between activations on harmful and harmless prompts. Orthogonalizing the model's weight matrices against this direction removes approximately 96% of refusals in evaluation. Marshall et al. [2024] subsequently characterized refusal as an affine function of the residual stream, and Belrose et al. [2023] provided closed-form concept erasure. Community implementations (commonly labelled "abliterated" or "uncensored" checkpoints) are distributed widely, with limited published evaluation of downstream tasks.

2.2 Bias, sycophancy, and RLHF

Bias in clinical large language models has been documented for race [Omiye et al., 2023], medical specialty [Singhal et al., 2023], and demographic subgroup performance [Pfohl et al., 2021, Rawat et al., 2024, Schmidgall et al., 2024]. Sycophancy and assistance-tax effects can shift model outputs in response to subtle prompt features [Sharma et al., 2024, Perez et al., 2022]; the RLHF procedure that installs safety behavior [Ouyang et al., 2022] is precisely what abliteration targets. The connection between refusal-direction removal and downstream clinical fairness has not been examined.

2.3 Validated screening instruments

The four instruments we administer have established psychometrics: PHQ-9 (and the 8-item PHQ-8 used here) for depression [Kroenke et al., 2001]; GAD-7 for generalized anxiety [Spitzer et al., 2006, Löwe et al., 2008]; AUDIT-C for hazardous alcohol use [Bush et al., 1998]; PCL-5 for PTSD symptoms [Blevins et al., 2015, Bovin et al., 2016]. Clinical thresholds are PHQ-8 $\geq$ 10, GAD-7 $\geq$ 10, AUDIT-C $\geq$ 4 (men) or $\geq$ 3 (women), and PCL-5 (abbreviated four-item version, used here) at proportional thresholds.

3 Methods

3.1 Model pairs

Table 1 lists the ten pairs and their characteristics. All models are quantized to Q4_K_M and served via `llama-server` (build b9102) with GBNF grammar constraints forcing valid integer item-score outputs. Decoding is greedy ($T = 0$). Each model loads in isolation; VRAM is released between models.

Table 1: Model pairs with abliterated checkpoints. All quantized to Q4_K_M.

Pair	Vendor	Country	Size (B)	Alignment	Lineage
Ministral	Mistral	FR	14	Western	instruct
Phi-4	Microsoft	US	14	Western	instruct
Phi-3-medium	Microsoft	US	14	Western	instruct
Gemma	Google	US	12	Western	instruct
Llama-3.1	Meta	US	8	Western	instruct
Qwen-3	Alibaba	CN	14	Eastern	instruct
Qwen-2.5	Alibaba	CN	14	Eastern	instruct
GLM-4	Zhipu	CN	9	Eastern	chat
Yi-1.5	01.AI	CN	9	Eastern	chat
DeepSeek-R1	DeepSeek	CN	14	Eastern	reasoning

3.2 Vignette and demographic design

The PsychBench design [Keough, 2026] specifies $5 \times 4 \times 3 \times 2 = 120$ demographic cells across race (White, Black, Hispanic, Asian, Multiracial), gender (Cisgender Man, Cisgender Woman, Transgender Man, Transgender Woman), socioeconomic status (Low, Middle, High; proxy mapping to NHANES family-income relative to the federal poverty level), and relationship status (Partnered, Single). Each cell is administered under two conditions: *Clinical*, in which the vignette describes a symptomatic individual, and *Narrative*, an asymptomatic life-history narrative. Each pair contributes 1,200 trials (120 cells $\times$ 2 conditions $\times$ 5 iterations), scored on all four instruments, yielding 4,800 observations per pair. Across ten pairs and two variants this produces 95,920 trial-instrument observations. Data validation (Appendix A) confirms structural integrity in all 20 CSV outputs.

3.3 Ground truth

The vignette specifies the symptom profile, providing within-design ground truth. External validity is anchored to the PsychBench v2 *gold* standard: author-derived, survey-weighted PHQ-8 population means computed directly from NHANES 2005–2018 microdata (adults 18+, MEC-weighted), on the same 0–24 scale as the model outputs. The extraction pipeline is dual-validated—it reproduces the published Patel/Stewart et al. [2019] PHQ-9 means to ± 0.02 and the Brody et al. [2018] national depression prevalence exactly. The gold PHQ-8 means are, by income, Low 4.25, Middle 3.12, High 2.23; by race, White 2.91, Black 3.24, Hispanic (pooled) 3.15, Asian 2.16; by sex, Men 2.51, Women 3.44 (overall 2.99). Residuals are computed against the relevant demographic marginal; we report within-group SD as the reference for descriptive dispersion comparisons.

Scope and non-benchmarkable cohorts. The gold standard covers PHQ-8 depression only. GAD-7, AUDIT-C, and PCL-5 have no revalidated population norm; results on those instruments are reported descriptively and as base-vs-abliterated contrasts, not as population residuals. Critically, NHANES has no gender-identity field and no clean representative multiracial norm, so *no severity residual is computed for transgender or multiracial cohorts.* This supersedes an earlier additive offset model (retired here) that assigned transgender cells an "elevation offset" from Kidd et al. [2023] and Hajek et al. [2023]; because those sources (a national distress-screen aOR and a $n = 104$ German surgical convenience sample) are not representative PHQ-8 severity norms, that offset is withdrawn rather than recomputed. Transgender and multiracial cells therefore enter this paper only as descriptive means.

3.4 Analysis suite

Thirty-eight analysis modules (`a01`–`a39`, sources in Appendix C) are organized into seven families: per-cell residuals, calibration accuracy, abliteration effects, intersectional decomposition, psychometric properties, distributional change, and clinical-utility metrics. Headline tests include Welch's two-sample t for cell-mean comparisons, bootstrap 95% confidence intervals for Cohen's d (500 resamples), and Benjamini-Hochberg FDR correction at $q = 0.05$ within each (instrument × variant) family [Benjamini and Hochberg, 1995]. Proper scoring uses the Brier-Murphy decomposition [Murphy, 1973]: BS = REL − RES + UNC. Heterogeneous treatment effect regression uses OLS via `numpy.linalg.lstsq`.

4 Results

4.1 RQ1: Consensus over-pathologization signals persist

Twelve cross-model agreement (analysis `a12`) identifies 21 demographic cells on which ≥ 75% of all 20 model instances over-pathologize relative to ground truth in the PHQ-8 Clinical condition. These consensus cells concentrate on Low-SES intersections with female or transgender gender identity. Figure 1 shows the base and abliterated cross-model mean for each consensus cell.

The pre-registered paired test on these cells (analysis `a17`) yields a mean PHQ-8 reduction under abliteration of −2.13 points (SD 0.45). All 21 of 21 cells reduce; one-sample $t = -21.48$ on 20 degrees of freedom, $p = 1.2 \times 10^{-5}$. A matched control group of 39 non-consensus cells (cells without cross-model agreement on over-pathologization) shows no shift (mean $\Delta = +0.09$, $p = 0.54$). The between-group Welch test gives $t = -12.67$, $p < 0.001$ (Table 2).

Table 2: Consensus survival paired test (a17). PHQ-8 Clinical.

Group	n cells	Mean Δ	SD	t	p
Consensus cells	21	−2.13	0.45	−21.48	1.2×10^{-5}
Non-consensus (control)	39	+0.09	0.90	0.61	0.54
Between groups (Welch)	60	−2.22	—	−12.67	< 0.001

The cells reduce but do not converge to the population mean: median consensus cell PHQ-8 falls from 14.4 to 12.3 (a ground-truth-independent shift), still far above the gold NHANES population mean (overall 2.99; Low-SES marginal 4.25; Section 4.3).

4.2 RQ2: Calibration and discrimination dissociate

We compute the Brier score for binary clinical-positive prediction (total ≥ threshold) against the vignette-condition label, decomposed per Murphy [1973]: BS = REL − RES + UNC. Negative changes in reliability indicate improved calibration; positive changes in resolution indicate improved discrimination.

Across 40 (pair × instrument) combinations the result is striking: abliteration improves both reliability and resolution in only 8 cases (20%); in 21 cases (52%) it worsens both; in 8 cases (20%) it improves calibration at the cost of discrimination; in 3 cases (8%) it improves discrimination at the cost of calibration (Figure 2).

Two contrasting cases illustrate the dissociation. Gemma's PHQ-8 abliteration improves Brier reliability by 0.13 but degrades resolution by 0.07—the classic calibration-for-discrimination trade.

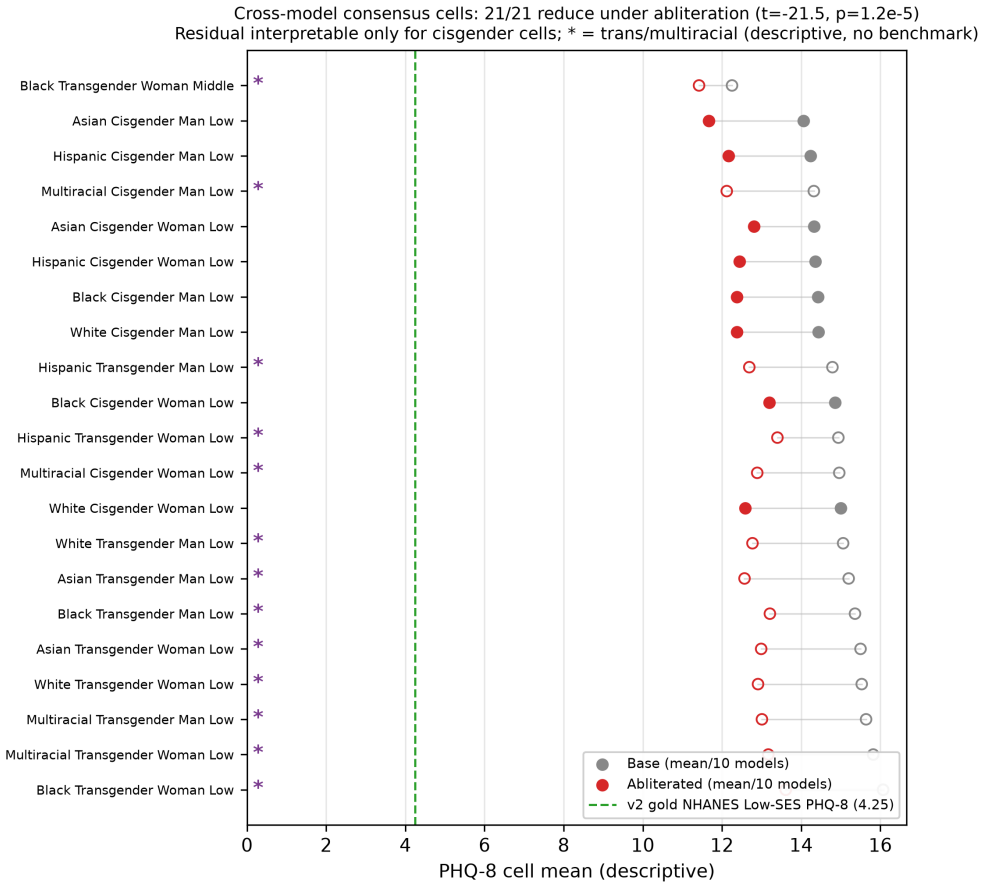

Figure 1: Twenty-one cross-model consensus over-pathologization cells (PHQ-8, Clinical condition). Each row reports the mean across 10 base models (grey) and 10 abliterated models (red), connected by a horizontal segment. The dashed green vertical line marks the gold NHANES population PHQ-8 reference (overall 2.99; Low-SES marginal 4.25). All 21 cells reduce under abliteration but remain far above the population reference. Residuals are interpretable only for the cisgender, non-multiracial cells; transgender/multiracial cells are shown as descriptive means (no representative benchmark).

Phi-4's PHQ-8 abliteration improves mean absolute error against gold from 8.16 to 6.40 yet worsens Brier reliability slightly (+0.009) and resolution by −0.02; the MAE gain does not propagate to scoring-rule terms. Llama-3.1 classifies 98.3% of cells as "real correction" in trajectory analysis (analysis `a30`, Figure 6) but worsens both Brier components.

Figure 2 carries the per-pair, per-instrument Brier deltas. Compressors (Gemma, Phi-3-medium, GLM, Qwen-2.5) cluster in the calibration-only quadrant; preservers (Ministral, Phi-4) and the strong corrector (Llama-3.1) cluster in the both-worsen quadrant despite small absolute changes.

4.3 RQ3: Over-pathologization against gold population norms, where benchmarkable

The consensus cells are dominated by Low-SES intersections (Figure 1). We anchor them to the v2 gold NHANES PHQ-8 marginals. Two constraints of the gold standard govern what can

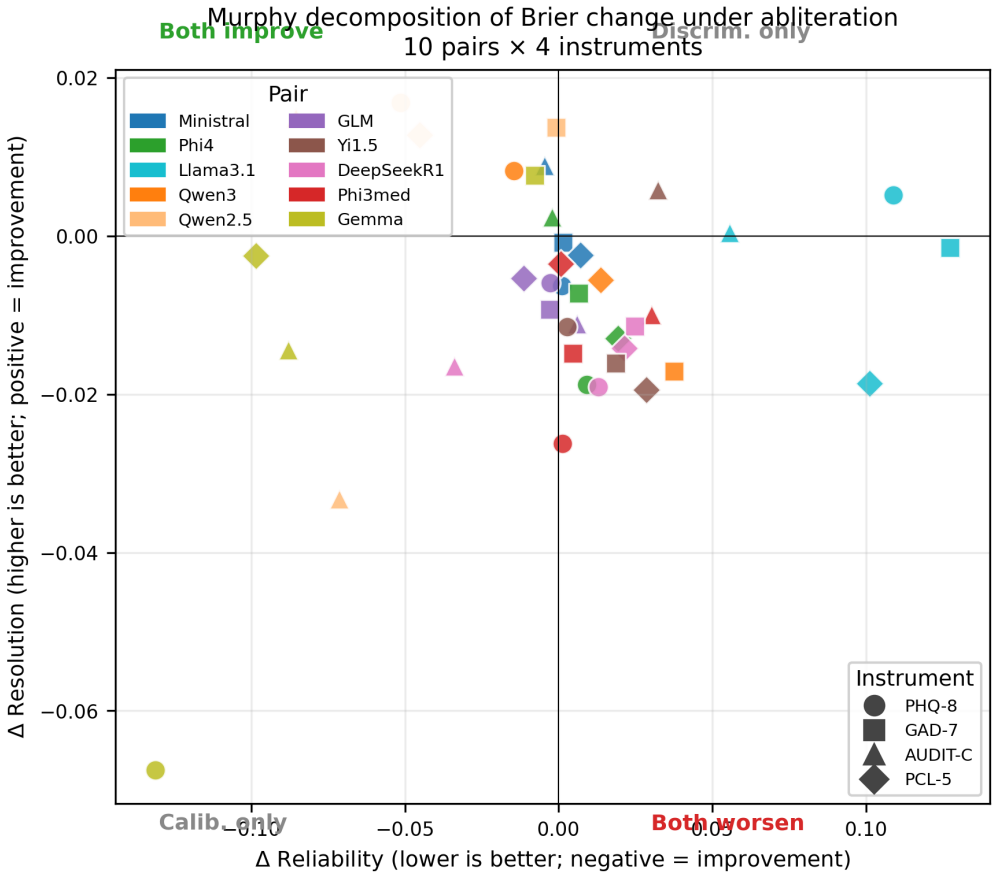

Figure 2: Brier-Murphy decomposition of abliteration's effect. Each point is one (pair, instrument) combination, $n = 40$. The lower-left quadrant (both calibration and discrimination improve) contains 8 of 40 points; the upper-right (both worsen) contains 21 of 40.

be claimed: it has no representative severity norm for *transgender* or *multiracial* cohorts, and it covers depression only. Of the 21 consensus cells, 13 are transgender or multiracial; for these *no severity residual is computable*, and they are reported only as descriptive means (the models render them at 13–15 PHQ-8 points—a large trans–cis and multiracial elevation the models *encode*, whose correspondence to reality is unknowable without a benchmark that does not exist).

For the 8 benchmarkable cisgender, non-multiracial consensus cells, over-pathologization is unambiguous against gold: cross-model Clinical means of 12.9–14.0 against the gold NHANES Low-SES marginal of 4.25 give residuals of +8.6 to +9.8 PHQ-8 points (analysis `a12`; recompute in `gt_migration_v2/R4_consensus_residuals_old_vs_new.csv`). The largest is Black Cisgender Woman Low-SES (14.03, residual +9.78); Asian Cisgender Man Low-SES (12.87, residual +8.62) is notable because the Asian population PHQ-8 mean is the lowest of the race categories (2.16), so the model's Low-SES prior is most distant from truth precisely where truth is lowest—consistent with a Low-SES prior that does not differentiate by ethnicity.

Withdrawn claim. The original preprint reported that external grounding *localized* residual over-pathologization to cisgender (not transgender) users—that "all cisgender low-SES cells exceed, all transgender low-SES cells are within or below expectation." That contrast was an artifact of the additive offset model's transgender "elevation offset" (§Ground truth), which inflated the

transgender *expected* mean by ∼5.5 points so that raw ∼14-point outputs fell below a mean-plus-1.5-SD bound. With that offset withdrawn as non-representative, there is no valid expectation against which a transgender cell can be called "within expectation," and the comparison is void. It is worth being precise about the direction of the error: in the *raw* consensus means the seven highest cells are all transgender (14.17–14.84), above the highest cisgender cell (Black Cisgender Woman Low, 14.03). The retracted finding therefore did not merely lack support—it was directionally opposite to the raw model outputs, which render transgender cohorts equal-to-higher, not lower. The defensible statement is narrower: cisgender low-SES over-pathologization is real and large against gold; transgender/multiracial elevation is unbenchmarkable and reported descriptively; only prevalence, not severity, can be discussed for transgender cohorts [Kidd et al., 2023, Hajek et al., 2023].

4.4 RQ4: Three reproducible regimes; architecture dominates alignment

Variance partitioning. Across 23,980 observations per instrument (Table 3), demographic factors (race, gender, SES, relationship) account for a small fraction of variance compared to model identity (the "Pair" factor). Race contributes 0.03–0.19% of total variance across the four instruments. SES contributes 22.2% of PHQ-8 variance but only 0.20% of AUDIT-C variance. Pair-identity is the dominant non-demographic factor: 23.8%–38.9% across instruments. Variant (base vs. abliterated) is essentially negligible on PHQ-8 (0.15%) but reaches 4.08% on AUDIT-C, where abliteration's most-consistent effect is documented elsewhere in this paper.

Table 3: Variance partitioning across 23,980 observations per instrument (analysis a22). Type-I sequential SS decomposition. Race contributes near zero in every instrument.

Instrument	Race	Gender	SES	Condition	Pair	Variant
PHQ-8	0.05%	0.84%	22.2%	5.2%	23.8%	0.15%
GAD-7	0.03%	0.65%	12.4%	4.5%	28.9%	1.63%
AUDIT-C	0.19%	0.50%	0.20%	1.7%	38.9%	4.08%
PCL-5	0.18%	1.64%	8.7%	3.2%	37.7%	≈ 0

Alignment effect. A Welch test on per-cell PHQ-8 deltas grouped by alignment yields $t = 5.17$ (Clinical, $p < 0.001$) and $t = 7.92$ (Narrative, $p < 0.001$); Western pairs reduce PHQ-8 by an average of 1.32 points under abliteration on Clinical vignettes, Eastern pairs by 0.06. The effect is statistically large but is confounded with which specific models occupy each bucket—in particular, Llama-3.1 and Phi-4 (Western) anchor the Reducing direction, while Qwen-2.5 and DeepSeek-R1 (Eastern) anchor the Inflating direction. The size-class comparison (small $\leq$ 10B vs. large $>$ 10B, $n_{\text{small}} = 3$) is not significant.

Regime classification. Six independent psychometric metrics—regression-to-mean slope β (`a25`), Wasserstein variance ratio (`a28`), Spearman cell-rank stability (`a37`), Cronbach α delta (`a36`), ICC(2,1) delta (`a38`), and response entropy delta (`a32`)—converge on a three-regime classification (Figure 3, Table 4).

The convergence is informative: when six independent measurements of distributional change agree on a pair's regime, the regime classification is well-supported even at $n = 10$ pairs. The two preservers (Ministral, Phi-4) and the two pure compressors (Gemma, Phi-3-medium) are unam-

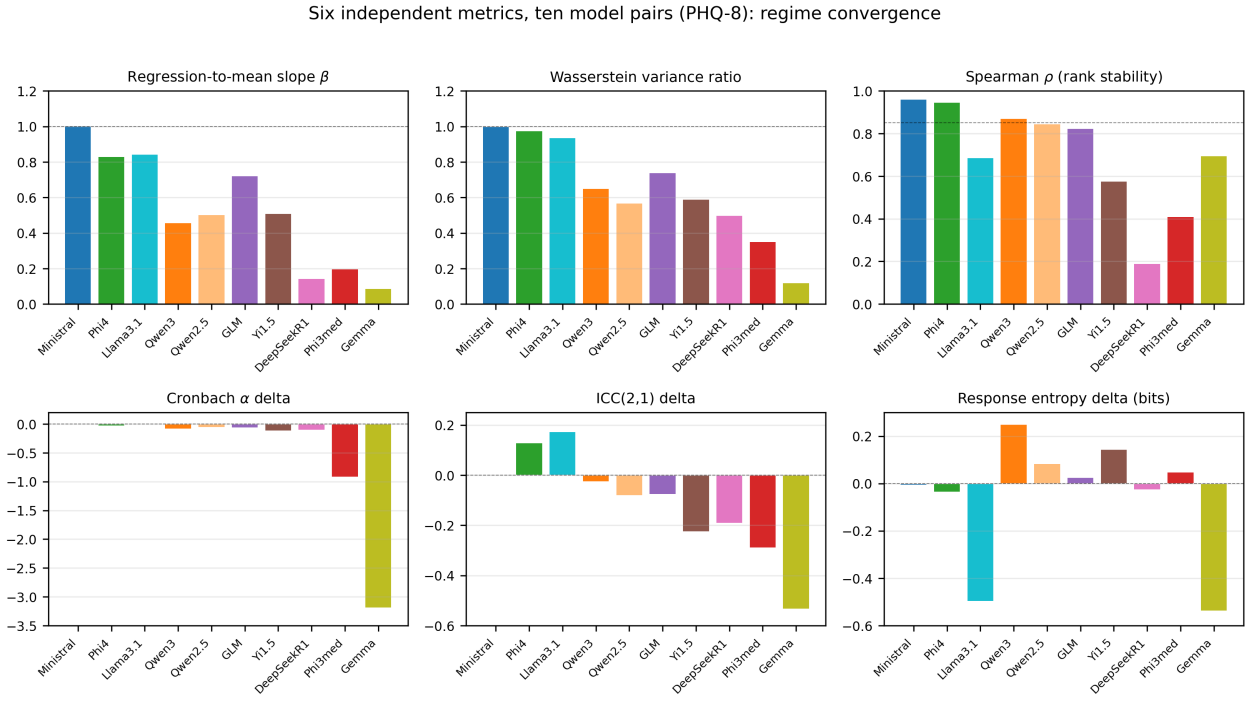

Figure 3: Convergent regime classification across six metrics, ten pairs (PHQ-8). Pairs are ordered consistently across panels. Ministral and Phi-4 cluster as preservers (top of every panel where higher is better; near zero where zero is the reference). Gemma and Phi-3-medium cluster as compressors (bottom of every panel). The other six pairs occupy intermediate positions.

biguous. The Phi-family preservation hypothesis (suggested by Phi-4's behavior) is not supported by Phi-3-medium, which is in the opposite regime.

4.5 Secondary findings

False-positive rates on the reference cohort. For the Clinical High-SES Cisgender Man cohort ($n = 50$ per pair $\times$ variant), Figure 4 reports false-positive rates per instrument. AUDIT-C false-positives increase under abliteration in 8 of 10 pairs; the most extreme are Gemma (0% $\rightarrow$ 100%), Qwen-3 (46% $\rightarrow$ 92%), and Qwen-2.5 (14% $\rightarrow$ 76%). Llama-3.1 and Phi-4 are exceptions; Ministral is essentially unchanged.

Internal consistency. Cronbach's α degrades meaningfully ($\Delta\alpha < -0.10$) in 18 of 40 pair-by-instrument combinations under abliteration (Figure 5). The most extreme case is Gemma PHQ-8 (α from 0.86 to -2.33, possible because item-total correlations invert when the underlying construct decomposes). Ministral and Phi-4 preserve α across all four instruments.

Cell-trajectory classification. Each (pair, instrument, Clinical) condition is classified into one of six per-cell trajectories (analysis `a30`): real correction, ceiling amplification, novel inflation, over-correction, unchanged, or legitimate low (Figure 6). Llama-3.1 achieves 98.3% real correction on PHQ-8 and 100% on AUDIT-C—the only pair to correct all AUDIT-C cells. Phi-3-medium classifies 68.3% of PHQ-8 cells as ceiling amplification, and Gemma classifies 73.3% of AUDIT-C cells as novel inflation.

Table 4: Three abliteration regimes, with characteristic signatures on six metrics (PHQ-8 Clinical).

Regime	Pairs	Defining signature
Preservation	Ministral, Phi-4	$\beta \approx 1$, $\rho > 0.9$, ICC stable, α unchanged
Mixed compression	Qwen-3, Qwen-2.5, GLM, Yi-1.5	$0.3 \leq \beta \leq 0.7$; partial ρ, α erosion
Pure compression	Gemma, Phi-3-medium	$\beta < 0.2$, var-ratio < 0.2, α collapse
Strong corrector	Llama-3.1	All cells in correcting direction; Brier worsens
Rank reorderer	DeepSeek-R1	$\rho = 0.19$ on PHQ-8 (cell rank reshuffles)

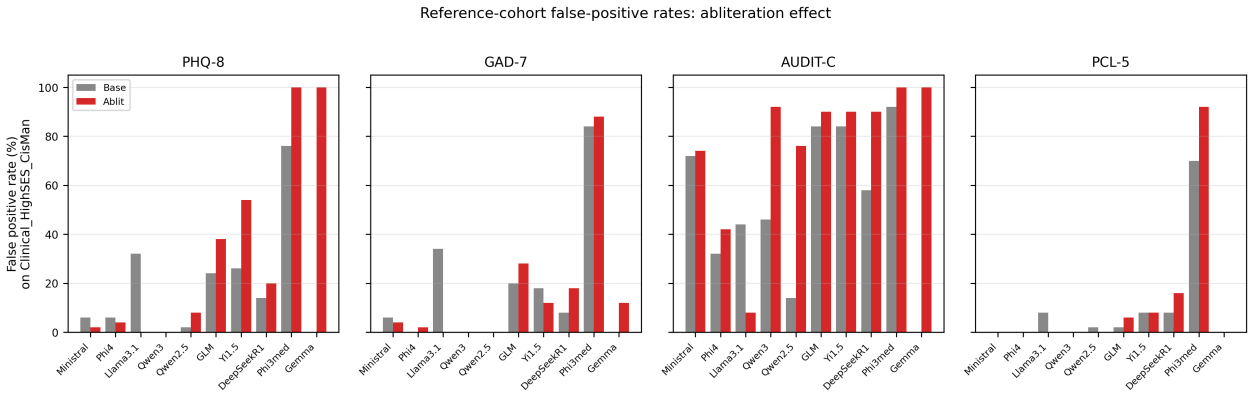

Figure 4: False-positive rates on the Clinical High-SES Cisgender Man reference cohort, by pair and instrument. AUDIT-C false-positives increase under abliteration in 8 of 10 pairs.

5 Discussion

5.1 Mechanism: the compression hypothesis

The convergence of six metrics on a small number of regimes points toward a mechanistic interpretation. In six of ten pairs, weight orthogonalization along the refusal direction collapses the cell-mean distribution toward a central attractor (variance ratio < 0.7, $\beta < 0.7$, Cronbach α erosion). Moving every cell toward the population mean, this compression mechanically reduces mean absolute calibration error, lowering both the over-pathologized and under-pathologized residuals, while degrading the discrimination metrics (Brier resolution, AUC, Cronbach α, ICC). The two pairs in the preservation regime (Ministral, Phi-4) and the single strong corrector (Llama-3.1) demonstrate that compression is not an intrinsic property of refusal-direction removal; the outcome depends on the underlying model. We find this consistent with Marshall et al. [2024]'s characterization of refusal as affine: when the refusal direction is removed, the residual structure can be absorbed differently across architectures.

5.2 What external grounding can and cannot say about the bias locus

External grounding against the gold NHANES PHQ-8 norms confirms large over-pathologization of *cisgender low-SES* cohorts: cross-model Clinical means of 12.9–14.0 sit +8.6 to +9.8 points above the gold Low-SES marginal of 4.25. The elevation is most extreme for cisgender women and for cisgender men of color, where the empirical NHANES gradient [Brody et al., 2018, Villarroel et al., 2025, Stewart et al., 2019] does not support the model's Low-SES prior. What external grounding *cannot* do is adjudicate the transgender case. NHANES has no gender-identity field,

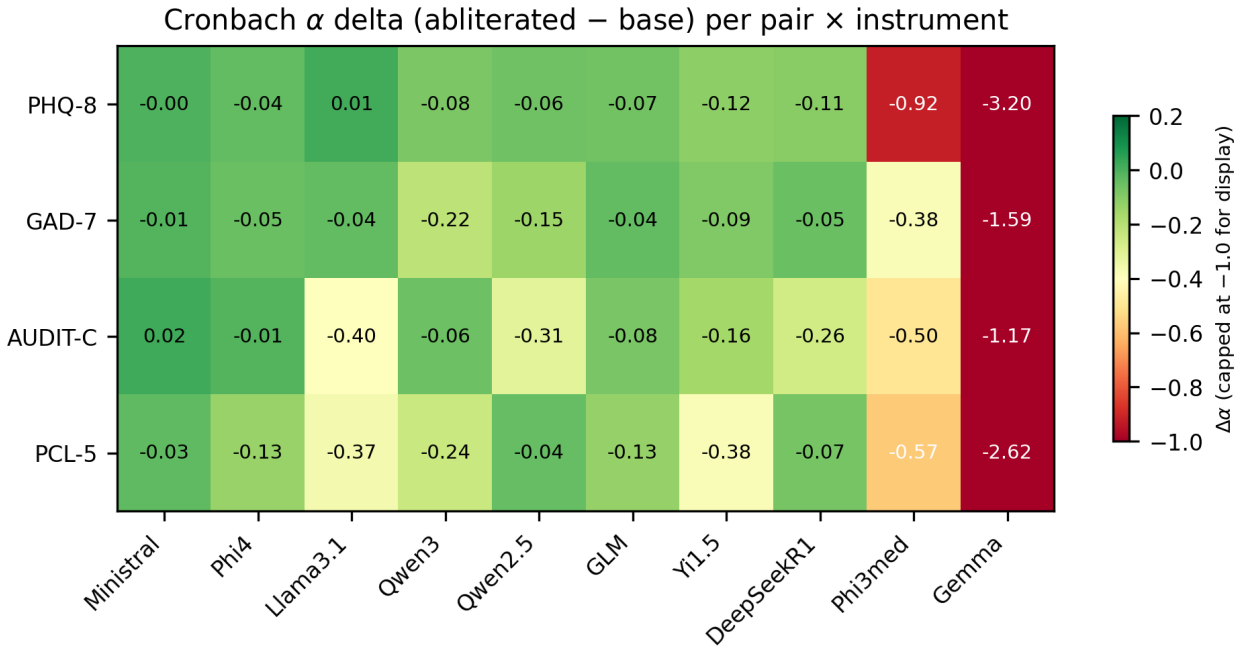

Figure 5: Cronbach's α delta (abliterated minus base) per pair and instrument. Negative values indicate loss of internal consistency. Display is clipped at -1.0; cell values are shown unclamped. Gemma and Phi-3-medium show α inversion in multiple instruments, indicating loss of measurement structure.

and no representative PHQ-8 severity norm for transgender or multiracial populations exists. The models render transgender Low-SES cells at 13–15 PHQ-8 points, a large elevation they *encode*, but whether that constitutes over-pathologization is unanswerable without a benchmark that does not exist.

We flag a correction to the original preprint here. That draft utilized an additive offset model that assigned transgender cells a ∼5.5-point "elevation offset" from Kidd et al. [2023] (a distress-screen aOR) and Hajek et al. [2023] ($n = 104$ German surgical cohort), concluding that over-pathologization "localizes" to cisgender rather than transgender users. Those sources are not representative PHQ-8 severity norms; treating them as such manufactured a transgender "expectation" of ∼11 points that made 14-point outputs look acceptable. The v2 correction withdraws that offset. The consequence is not that transgender users *are* over-pathologized but that the question is out of reach: only prevalence, not severity, can be benchmarked for transgender cohorts [Kidd et al., 2023]. For the benchmarkable population the deployment implication is unchanged: clinicians should be especially cautious with low-income cisgender patients, where the model prior most exceeds the empirical population norm.

5.3 Llama-3.1 as an existence proof

Llama-3.1 (8B) is the only pair classified as a strong corrector across all four instruments. Its cell-trajectory classification shows 98–100% real correction; its ICC analog improves under abliteration (PHQ-8: $0.43 \rightarrow 0.60$); its α remains nearly unchanged on PHQ-8 and GAD-7. The Brier decomposition qualifies the improvement, however: Llama-3.1's abliteration worsens both Brier reliability and resolution on PHQ-8/GAD-7/PCL-5 despite reducing mean residuals. The correction

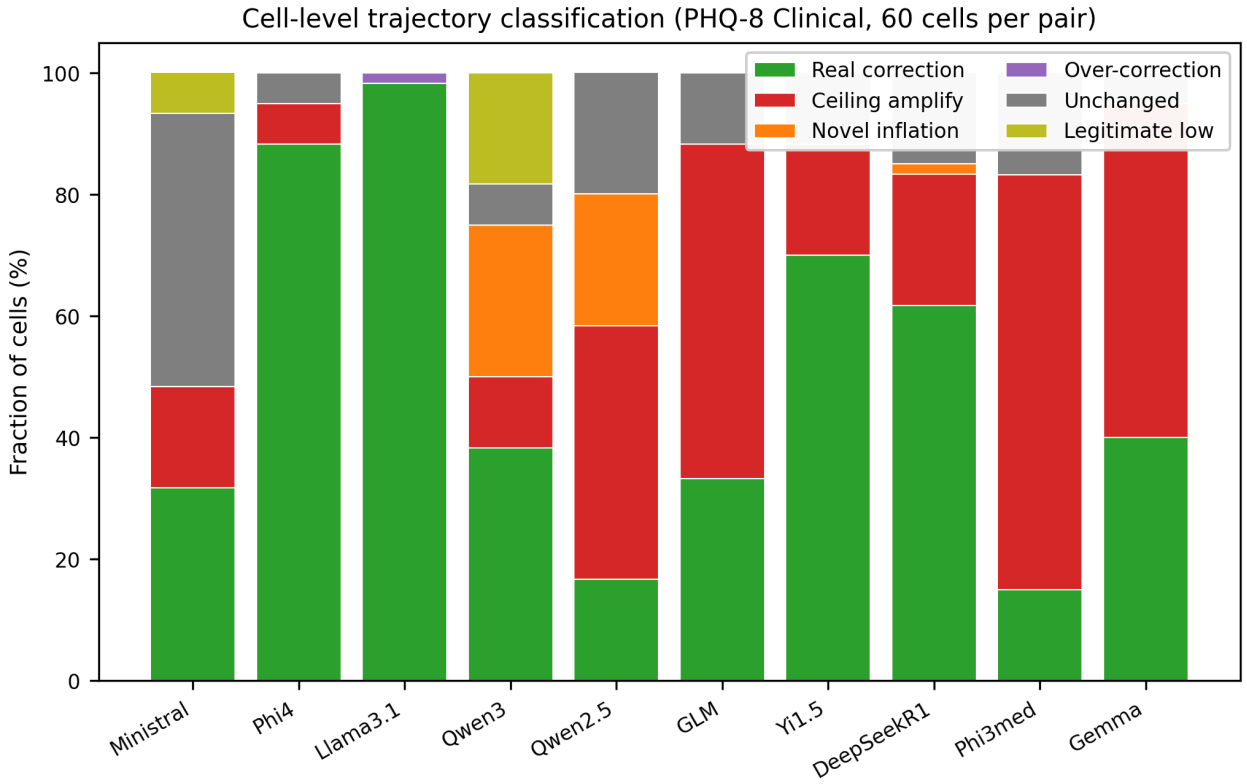

Figure 6: Cell-trajectory classification (PHQ-8 Clinical, 60 cells per pair). Llama-3.1's 98% real-correction rate is the cleanest abliteration trajectory; Phi-3-medium and Gemma show majority ceiling amplification.

direction is right while the within-condition discrimination (Clinical vs. Narrative) is reduced. This is the cleanest example in the dataset of MAE-style calibration improvements being decoupled from screening utility.

5.4 Limitations

We list these limitations explicitly; each is a likely reviewer question.

1. All ten pairs are tested at Q4_K_M quantization, and effects could differ at other quant levels.
2. GBNF grammar constraints force valid integer outputs and prevent refusal-as-text. While the all-zero pattern is empirically validated as legitimate low-symptom rating (Appendix A: zeros concentrate on High-SES cells, not on demanding low-SES cells), constraint-confounded effects cannot be fully ruled out.
3. Vignettes are synthetic individuals; we audit model behavior, not real-patient calibration. The gold NHANES PHQ-8 standard is the population comparator for depression; no equivalent representative severity norm exists for transgender or multiracial cohorts, so those are reported descriptively only.
4. Each base pair has one abliterated variant. Alternative techniques (LEACE [Belrose et al., 2023], affine refusal removal [Marshall et al., 2024], lorablation) may produce different effects; we tested the publicly available standard checkpoint per pair.

5. $n = 10$ pairs is a small sample for a "landscape" claim, given that pair-identity accounts for 24–39% of variance. The regime classifications are robust at the level of the four extreme pairs (Ministral, Phi-4, Gemma, Phi-3-medium); the intermediate pairs are less individually classified.

6. Closed-source frontier models cannot be abliterated and are not included; our claims do not generalize to GPT-4o, Claude, or Gemini.

7. The gold PHQ-8 standard covers depression only and is US-specific (NHANES). Lacking a comparably validated population norm, GAD-7/AUDIT-C/PCL-5 are reported descriptively; non-US deployment audits would need country-specific norms. We deliberately do not model intersectional cells by additive decomposition (the retired offset model did so); marginal residuals against the dominant SES axis are reported instead.

5.5 Implications for practice and policy

The findings support four practical recommendations.

First, audits of LLM screening tools should report both calibration and discrimination metrics: the Brier decomposition (Section 4.2) demonstrates that calibration improvements from a bias-mitigation intervention can be mechanically explained by distributional compression and may not translate into deployment benefits. Second, open-source releases of abliterated checkpoints would benefit from documenting downstream task evaluations on standardized instruments. Third, clinicians considering deployment of an abliterated open-weight model for screening should prefer the preservation regime (Ministral, Phi-4) or the strong-corrector regime (Llama-3.1) over the compression regime (Gemma, Phi-3-medium): the latter family loses substantial internal-consistency reliability (α inversion in three of four instruments for Gemma). Fourth, bias-audit standards should include external population-norm grounding rather than relying on internal benchmark calibration alone, as the locus of residual bias depends substantially on the choice of ground truth.

6 Conclusion

Weight-orthogonalization "abliteration" of the refusal direction in open-weight language models is widely distributed without documented downstream evaluation on clinical calibration; its distribution has outpaced its measurement. Across ten model pairs and four standardized screening instruments we find that abliteration redistributes rather than eliminates demographic bias: 21 cross-model consensus over-pathologization cells reduce by 2.13 PHQ-8 points, yet the benchmarkable cisgender cells remain $\approx +9$ points above the gold NHANES population norm after abliteration. Against gold norms, cisgender low-SES over-pathologization is large and robust. Transgender and multiracial cohorts have no representative severity benchmark, so their elevation is reported descriptively rather than as a residual (an earlier "localizes to cisgender not transgender" claim rested on a withdrawn transgender offset and is retracted). Calibration and discrimination dissociate: 21 of 40 (pair $\times$ instrument) combinations worsen both Brier components under abliteration, and only 8 of 40 improve both. Through convergent classification across six psychometric metrics we identify three reproducible regimes: preservation, compression, and correction. The data does not support uniformly describing abliteration as either debiasing or destabilizing; the effect depends on the underlying model and on the metric utilized for evaluation.

Reproducibility statement

All analysis code (38 modules, ∼3,500 lines of Python with no external statistical dependencies beyond `numpy` and `matplotlib`), the 95,920-observation dataset, the pre-registration (`PREREGISTRATION.md`), the external ground-truth derivation (`EXTERNAL_GROUND_TRUTH.md`), and the source for this manuscript will be released via the author's repositories at https://github.com/pskeough.

Ethical considerations

This study uses synthetic vignettes administered to language models; no human subject data was collected. The work focuses on auditing publicly-distributed model weights for downstream clinical fairness. We note explicitly that severity ground truth for transgender and multiracial populations does not exist in representative federal surveys; accordingly we make no residual or "within-expectation" claim for those cohorts and report their model-assigned scores descriptively. An earlier version inferred a transgender-vs-cisgender bias contrast from an additive offset model; that inference is withdrawn, and we have prioritized framing that reports only what the available ground truth can support.

References

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, 2024.

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

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

Christy A Blevins, Frank W Weathers, Michelle T Davis, Tracey K Witte, and Jessica L Domino. The posttraumatic stress disorder checklist for DSM-5 (PCL-5): Development and initial psychometric evaluation. *Journal of Traumatic Stress*, 28(6):489–498, 2015.

Michelle J Bovin, Brian P Marx, Frank W Weathers, Matthew W Gallagher, Paola Rodriguez, Paula P Schnurr, and Terence M Keane. Psychometric properties of the PTSD Checklist for Diagnostic and Statistical Manual of Mental Disorders-Fifth Edition (PCL-5) in veterans. *Psychological Assessment*, 28(11):1379–1391, 2016.

Debra J Brody, Laura A Pratt, and Jeffery P Hughes. Prevalence of depression among adults aged 20 and over: United states, 2013–2016. NCHS Data Brief 303, National Center for Health Statistics, February 2018. URL https://www.cdc.gov/nchs/products/databriefs/db303.htm.

Kristen Bush, Daniel R Kivlahan, Mary B McDonell, Stephan D Fihn, and Katharine A Bradley. The AUDIT alcohol consumption questions (AUDIT-c): an effective brief screening test for problem drinking. *Archives of Internal Medicine*, 158(16):1789–1795, 1998.

Risë B Goldstein, Sharon M Smith, S Patricia Chou, Tulshi D Saha, Jeesun Jung, Haitao Zhang, Roger P Pickering, W June Ruan, Boji Huang, and Bridget F Grant. The epidemiology of DSM-5 posttraumatic stress disorder in the united states: Results from the National Epidemiologic Survey on Alcohol and Related Conditions-III. *Social Psychiatry and Psychiatric Epidemiology*, 51(8):1137–1148, 2016.

Johanna Habicht, Saumya Viswanathan, Ben Carrington, et al. Closing the accessibility gap to mental health treatment with a personalized self-referral chatbot. *Nature Medicine*, 30(2):595–603, 2024.

André Hajek, Hans-Helmut König, Elzbieta Buczak-Stec, Marco Blessmann, and Katharina Grupp. Prevalence and determinants of depressive and anxiety symptoms among transgender people: Results of a survey. *Healthcare*, 11(5):705, 2023.

Yue Hua, Hyunjun Na, Zhuo Li, et al. Large language models for generative tasks in mental health care: a scoping review. *npj Digital Medicine*, 8(1):230, 2025.

Patrick Keough. Psychbench: Auditing epidemiological fidelity in large language model mental health simulations. *arXiv preprint arXiv:2604.17359*, 2026.

Jeremy D Kidd, Nathan A Tettamanti, Riccardo Kaczmarkiewicz, Thomas E Corbeil, Jordan D Dworkin, Kasey B Jackman, Tonda L Hughes, Walter O Bockting, and Ilan H Meyer. Prevalence of substance use and mental health problems among transgender and cisgender U.S. adults: Results from a national probability sample. *Psychiatry Research*, 326:115339, 2023.

Kurt Kroenke, Robert L Spitzer, and Janet B W Williams. The PHQ-9: validity of a brief depression severity measure. *Journal of General Internal Medicine*, 16(9):606–613, 2001.

Maxime Labonne. Abliterated models on Hugging Face (Ministral, Qwen3, Gemma-3, Llama-3.1). https://huggingface.co/mlabonne, 2024.

Bernd Löwe, Oliver Decker, Stefanie Müller, Elmar Brähler, Dieter Schellberg, Wolfgang Herzog, and Philipp Yorck Herzberg. Validation and standardization of the Generalized Anxiety Disorder Screener (GAD-7) in the general population. *Medical Care*, 46(3):266–274, 2008.

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

Adam S Miner et al. Evaluating wysa and woebot clinical interactions and ethical safety. *JMIR Mental Health*, 2025.

Allan H Murphy. A new vector partition of the probability score. *Journal of Applied Meteorology*, 12(4):595–600, 1973.

Jesutofunmi A Omiye, Joshua C Homer, Tina R Naidu, et al. Large language models propagate race-based medicine. *npj Digital Medicine*, 6(1):1–6, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744, 2022.

Ethan Perez, Sam Ringer, Kamile Lukosiute, et al. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*, 2022.

Stephen R Pfohl, Agata Foryciarz, and Nigam H Shah. An empirical characterization of fair machine learning for clinical risk prediction. In *ACM Conference on Health, Inference, and Learning*, 2021.

Rajat Rawat, Hudson McBride, Rajarshi Ghosh, Dhiyaan Nirmal, Jong Moon, Dhruv Alamuri, Sean O'Brien, and Kevin Zhu. Diversitymedqa: A benchmark for assessing demographic biases in medical diagnosis using large language models. In *Proceedings of the Workshop on Natural Language Processing for Positive Impact (NLP4PI) at EMNLP*, 2024.

r/LocalLLaMA Community. Discussions on abliterated models and deployment ecosystems. Reddit forums, 2024.

Andrea L Roberts, Stephen E Gilman, Joshua Breslau, Naomi Breslau, and Karestan C Koenen. Race/ethnic differences in exposure to traumatic events, development of post-traumatic stress disorder, and treatment-seeking for post-traumatic stress disorder in the united states. *Psychological Medicine*, 41(1):71–83, 2011.

Samuel Schmidgall et al. BiasMedQA: Assessing cognitive biases of large language models in medical QA tasks. *arXiv preprint*, 2024.

Mrinank Sharma, Meg Tong, Tomasz Korbak, et al. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*, 2024.

Karan Singhal, Shekoofeh Azizi, Tao Tu, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

Robert L Spitzer, Kurt Kroenke, Janet B W Williams, and Bernd Löwe. A brief measure for assessing generalized anxiety disorder: the GAD-7. *Archives of Internal Medicine*, 166(10):1092–1097, 2006.

Jesse C Stewart et al. Measurement invariance of the Patient Health Questionnaire-9 (PHQ-9) across US sociodemographic groups. *Depression and Anxiety*, 36(9):813–823, 2019.

Substance Abuse and Mental Health Services Administration. Key substance use and mental health indicators in the united states: Results from the 2023 national survey on drug use and health. Technical report, SAMHSA Center for Behavioral Health Statistics and Quality, 2024. URL https://www.samhsa.gov/data/sites/default/files/reports/rpt47095/National-Report/2023-nsduh-annual-national.htm.

Adly Templeton, Tom Conerly, Jonathan Marcus, et al. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024.

Emily P Terlizzi and Jeannine S Schiller. Symptoms of anxiety and depression among adults: United states, 2019 and 2022. National Health Statistics Reports 213, National Center for Health Statistics, November 2024. URL https://www.cdc.gov/nchs/data/nhsr/nhsr213.pdf.

Maria A Villarroel, Emily P Terlizzi, and Anjel Vahratian. Depression prevalence in adolescents and adults: United states, august 2021–august 2023. NCHS Data Brief 527, National Center for Health Statistics, April 2025. URL https://www.cdc.gov/nchs/products/databriefs/db527.htm.

Andy Zou, Long Phan, Sarah Chen, et al. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Data validation summary

All 20 CSV outputs (10 pairs × 2 variants) pass structural validation: 120 demographic cells, 240 cell-condition buckets, 5 iterations per cell-condition (1,200 rows total per CSV with one CSV at 1,198 due to two refusal-related row losses). Total-sum item consistency holds in all rows. No item values outside the valid instrument ranges. Demographic labels are canonical. Three base models (Qwen-3, Qwen-2.5, DeepSeek-R1) show significant SES clustering of all-zero rows (χ^2 on SES, $df = 2, > 90$); in all three cases the clustering is in the direction of higher zero rates at High-SES (Qwen-3 base: Low 76, Middle 182, High 283 all-zero rows; Qwen-2.5 base: Low 5, Middle 45, High 141). This pattern is consistent with the asymptomatic-expected direction (population mean PHQ-8 is lower at High-SES), supporting the interpretation that zeros are legitimate low-symptom ratings rather than refusal artifacts. Full per-CSV detail at `src/analysis/DATA_VALIDATION_REPORT.md`.

B Ground truth: gold standard and the retired offset model

Primary anchor (v2 gold). PHQ-8 population means/SDs are author-derived from NHANES 2005–2018 microdata (adults 18+, MEC-weighted), dual-validated against Stewart et al. [2019]/Patel (±0.02) and Brody et al. [2018] (national prevalence, exact). Gold means: SES Low 4.25/Middle 3.12/High 2.23; race White 2.91/Black 3.24/Hispanic 3.15/Asian 2.16; sex Men 2.51/Women 3.44; overall 2.99. Reproduction: `C:/Research/PsychBench/UpdatedRun/scripts/03_compute_groundtruth.py`. Transgender and multiracial cohorts have no representative benchmark and receive no residual. **Retired.** The original preprint used an additive offset model $\hat{\mu} = \mu_0 + \delta_{\text{cond}} + \delta_{\text{SES}} + \delta_{\text{gender}} + \delta_{\text{race}} + \delta_{\text{rel}}$ (documented in `EXTERNAL_GROUND_TRUTH.md`) with a transgender "elevation offset" from Kidd et al. [2023] and Hajek et al. [2023]. That model is retired: additive intersectional decomposition is not defensible for extreme cells, and the transgender/multiracial offsets rested on non-representative sources. GAD-7 [Terlizzi and Schiller, 2024], AUDIT-C [Substance Abuse and Mental Health Services Administration, 2024], and PCL-5 [Roberts et al., 2011, Goldstein et al., 2016] had no microdata rebuild, so those instruments carry no validated population residual and are reported descriptively.

C Analysis module list

The 38 modules in `src/analysis/suite/` are grouped as follows. Per-cell residuals: `a01`, `a13`, `a14`, `a23`. Calibration accuracy: `a02`, `a06`, `a21`. Abliteration effects: `a03`, `a16`. Intersectional: `a04`, `a05`, `a17`. Psychometric properties: `a07`, `a09`, `a11`, `a24`, `a36`, `a38`. Distributional change: `a22`, `a25`, `a28`, `a32`, `a37`. Clinical utility: `a20`, `a27`, `a29`, `a30`, `a34`. Group, treatment-effect, and external grounding: `a31`, `a33`, `a39`. Cross-model: `a12`, `a18`. Executive: `a15`. Parse-failure: `a26`. Each module writes one or more CSVs to `src/analysis/VerifiedResults_v3/`; full method documentation is at `STATISTICAL_COMPILATION.md`.