

Auditing LLM Portrayals of Neurodivergent People: Quantifying the Asymmetry Between Deficit Framing and Neurodiversity Affirmation

Carlos Toxtli¹, Manuel Delaflor^{2*}

¹Clemson University

²Metacognition Institute

ctoxtli@clemson.edu, manuel@metacognitioninstitute.org

Abstract

Open-weight large language models are increasingly used in education, hiring, and information-seeking contexts that touch neurodivergent people, yet we lack a clear empirical map of how their portrayals of neurodiversity shift under everyday prompt variation. We present a multi-condition, multi-model audit of LLM portrayals of eight neurodivergent identities (autism, ADHD, dyslexia, dyspraxia, dyscalculia, Tourette syndrome, OCD, and sensory-processing differences) plus a non-stigmatized medical control. Six open-weight models answer a pre-registered prompt bank that crosses seed items with seven perturbation families, including paraphrase, identity-first versus person-first language, system personas, plausible versus obviously fictional authority citations, multi-turn social pushback, leading questions, and two mitigation arms. Held-out LLM judges score each generation on three ordinal axes. Baseline portrayals are remarkably homogeneous across models, so the inconsistency observed under prompt variation is prompt-driven, not model-driven. Steering toward deficit framing is several times easier than steering toward neurodiversity affirmation. Emotional pushback shifts outputs more than evidence-laden pushback. Plausible-sounding fake citations slip past safety filters while obviously fictional ones do not. A single-line system-prompt mitigation recovers a meaningful share of the steering range at near-zero refusal cost on neutral prompts, though refusal rises sharply under adversarial pressure. The governance question for disability-relevant LLM deployment is therefore not whether the model is biased but who controls the steering, and with what accountability.

1 Introduction

LLMs portray people, and the portrayal carries consequences. When the people are neurodivergent (autistic, ADHD, dyslexic, dyspraxic, dyscalculic, Tourettic, OCD, sensory-divergent) and the deployment is education (Glazko et al. 2024), hiring (Sant et al. 2024), clinical triage, or information-seeking by a parent or newly-identified adult (Park et al. 2025; Wohn et al. 2026), that portrayal shapes who is hired, accommodated, taken seriously, or pathologized at scale (Whittaker et al. 2019; Bennett and Keyes 2020;

Hutchinson et al. 2020). Existing literature establishes that LLMs encode disability-relevant stigma (Hutchinson et al. 2020; Venkit, Srinath, and Wilson 2022; Herold, Waller, and Kacorri 2022), that they reproduce deficit framings in resume screening (Glazko et al. 2024) and advice-giving (Wohn et al. 2026), and that the autistic-community-preferred *neurodiversity paradigm* (Walker 2021; Botha et al. 2024; Chapman 2023) is unevenly represented in their outputs. What this literature does not yet provide is (a) a multi-condition decomposition across neurodivergent identities, (b) a causally-framed account of how prompt perturbations *move* the portrayal, or (c) an empirical evaluation of whether a deployer-side mitigation actually works.

We add those three things. Six open-weight models from four families (9B-754B parameters) answer a pre-registered prompt bank of 5,976 stimuli covering eight neurodivergent identities plus a non-stigmatized medical control, crossing seven perturbation families and 30 seed items, with held-out LLM judges scoring a three-axis rubric. Across 232,416 main-arm generations plus a 193,680-generation reasoning-on supplementary arm, the data support a single thesis we call *steerable unreliability*:

*Open-weight LLMs are **homogeneous** in their default portrayal of neurodivergent identities (model variance $\approx 0.1\%$ of total), but their actual output is **steered, asymmetrically, toward deficit framing** by single-line system prompts, plausible-sounding citations, and emotional pushback, and a single-line mitigation can recover one-fifth of the steering range at near-zero refusal cost on neutral prompts (the cost rises under adversarial pressure). The governance question is therefore not “is the model biased?” but “**who controls the steering, with what accountability?**”*

Research questions. We ask: (RQ1) How much of the variation in LLM portrayals of neurodivergent identities is attributable to the model versus to prompt-side factors? (RQ2) Is the steering range of those portrayals symmetric, or is it easier to push outputs toward deficit framing than toward neurodiversity affirmation? (RQ3) Which kinds of multi-turn social pressure (assertion, evidence, emotion) most effectively shift model stance? (RQ4) Do plausible-sounding but fabricated authority citations bypass the safety mechanisms that reject obviously fictional ones? (RQ5) Can a single-line

*Code, prompt bank, raw generations, judge prompts, and analysis notebooks are released with this paper under the tiered license described in the Ethics section.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

system-prompt mitigation reduce baseline deficit framing, and how does it behave under concurrent adversarial pressure? (RQ6) Do the perturbations that shift representational outputs (definitions, advice, comparisons) also shift allocational ones (hiring, school placement)?

The paper’s contributions are:

1. **Empirical.** The first multi-condition, multi-open-weight-model audit of LLM neurodiversity portrayal with crossed random-effects analysis and meaning-preserving baselines.
2. **Methodological.** A reusable evaluation suite (prompt bank with per-condition community-validated reference forms; nine-category ableist lexicon with provenance; three-axis judge rubric; statistical pipeline) released under a gated license.
3. **Conceptual.** The *steerable unreliability* framing as an alternative to single-direction “bias” auditing, with a quantified mitigation that has a defensible deployment story.

2 Background and Related Work

The neurodiversity paradigm and language. The neurodiversity paradigm (Walker 2021) reframes neurological variation as natural human diversity rather than disorder; it was developed collectively by autistic and other neurodivergent scholars in the 1990s and 2000s and has been the subject of explicit historical correction (Botha et al. 2024). *Identity-first language* (IFL, “autistic person”, “ADHDer”, “dyslexic person”) is the majority community preference in autism, ADHD, dyslexia, dyspraxia, dyscalculia, sensory-processing, and Tourette communities (Bottema-Beutel et al. 2021; Buijsman, Begeer, and Scheeren 2023); the OCD community in contrast leans toward person-first language and clinical terminology. IFL/PFL is therefore *not* a uniform binary; we encode it per-condition (Sec. 3.2).

LLM-bias measurement and disability. Disability is a present-but-unifying axis in BBQ (Parrish et al. 2022), CrowS-Pairs (Nangia et al. 2020), and HolisticBias (Smith et al. 2022), but none decompose neurodivergent conditions. Disability-specific work spans implicit-bias studies in pre-trained models (Venkit, Srinath, and Wilson 2022), ableist-bias intersectionality (Hassan, Huenerfauth, and Alm 2021), stereotype-content-model (SCM) audits (Herold, Waller, and Kacorri 2022), and downstream audits on autism (Park et al. 2025; Wohn et al. 2026) and ADHD/dyslexia (Brandsen et al. 2024). We extend this literature by crossing condition $\times$ open-weight model $\times$ perturbation in a multifactorial design.

Prompt-perturbation as treatment. Independent literatures document steering levers: persona-priming raises toxicity 6 $\times$ and breaks reasoning (Deshpande et al. 2023; Gupta et al. 2024; Salewski et al. 2023); sycophancy flips stance under simple user disagreement (Sharma et al. 2024; Perez et al. 2023); persuasion taxonomies operationalize authority and social proof (Zeng et al. 2024); format and paraphrase perturbations swing accuracy by tens of points (Sclar et al. 2024; Mizrahi et al. 2024; Pezeshkpour and Hruschka 2024). We bring these into one design centered on neurodiversity portrayal.

Model	Family	Params	License
Qwen 3.5 9B	Qwen	9B	Apache 2.0
Gemma 4 31B	Gemma	31B	Apache 2.0
Qwen 3.6 27B (dense)	Qwen	27B	Apache 2.0
Qwen 3.6 35B A3B (MoE)	Qwen	35B	Apache 2.0
GPT-OSS 120B (MoE)	OpenAI	120B	Apache 2.0
GLM-5.1 (MoE)	GLM (Zhipu)	754B	MIT

Table 1: Open-weight LLM evaluation panel.

Identification and meaning-preserving baselines. Recent meta-critique (Mizrahi et al. 2024; Sclar et al. 2024) argue that every counterfactual prompt edit is a compound treatment unless calibrated against a meaning-preserving baseline. We include a paraphrase arm (P1) as exactly this baseline and length-matched filler-prefix controls for prefix-bearing arms (P5, P7).

3 Methods

3.1 Model Panel and Judges

Six open-weight evaluation models spanning four families (Table 1). Chain-of-thought reasoning is disabled in the main arm for the models that support it, so the comparison across models is on directly user-facing answers; the per-model disable settings are in the artifact’s configuration file. A reasoning-on supplementary arm (Sec. 5) repeats the sweep with reasoning enabled.

Two held-out LLM judges from disjoint families: *primary* deepseek-v4-pro (685B MoE, MIT) scores every generation; *secondary* qwen3-omni-30b-a3b scores a 5% stratified inter-rater-reliability (IRR) subsample (stratified by perturbation $\times$ model; $N_{IRR} = 11,622$ in the main arm, $N = 9,685$ in the reasoning arm). Krippendorff α is reported per metric (Krippendorff 2011); pre-registered threshold $\alpha \geq 0.6$ for ordinal and $\alpha \geq 0.7$ for binary refusal. Realized α values with clustered bootstrap CIs appear in Table 2; the pathologization axis point estimate in the main arm is $\alpha = 0.612$ (clustered 95% CI [0.564, 0.657]; the lower bound dips below the pre-registered 0.6 threshold). Reasoning-arm $\alpha = 0.593$ is marginally below threshold, so reasoning-arm contrasts (Sec. 5) are reported as *exploratory*. As a triangulating construct-validity check we ran a third LLM judge from a disjoint family (gemma-4-31b, Google, family-disjoint from both deepseek and qwen) on a 2,000-cell stratified sub-sample of the persona contrast: three-judge $\alpha = 0.832$ on that sub-sample, with pairwise α ’s of 0.79 (deepseek $\leftrightarrow$ qwen-omni), 0.89 (deepseek $\leftrightarrow$ gemma), and 0.81 (qwen-omni $\leftrightarrow$ gemma). The third judge also confirms the direction and approximate magnitude of the persona effect (gemma stigma-persona mean = 2.00, advocacy = 0.08, DJ = 0.03, neutral = 0.46-0.50). Axes that fell below threshold (ableism, affect, refusal judge) are reported as *exploratory* only. All intermediate numbers (variance components, cross-judge agreement, persona effect sizes) are reproduced in the artifact’s analysis outputs.

A deeper validity concern than inter-judge agreement is *shared cultural anchoring*: pathologization and neurodiversity affirmation are socially and culturally contingent

Metric	α	95% CI
Pathologization (main)[†]	0.612	[0.564, 0.657]
Pathologization (reasoning)	0.593	[0.554, 0.627]
factual_alignment	0.517	,
affect	0.314	,
ableism	0.190	,
refusal (nominal)	0.233	,
<i>3-judge sensitivity, n=1,988:</i>		
deepseek+qwen+gemma	0.832	,
deepseek vs qwen-omni	0.793	,
deepseek vs gemma	0.894	,
qwen-omni vs gemma	0.806	,

Table 2: Realized cross-judge Krippendorff α with 95% *clustered* bootstrap CIs (resampling by item, $B=500$) on the 5% stratified IRR sub-sample. Pre-registered thresholds: $\alpha \geq 0.6$ ordinal; $\alpha \geq 0.7$ nominal/binary. Under clustered resampling, the Pathologization CIs widen and the lower bound dips below 0.6 in both arms, the point estimates clear (main) and are marginal (reasoning). We accept Pathologization (main) as the primary outcome and Pathologization (reasoning) as exploratory; all other axes fall below threshold and are exploratory only.

[†] Point estimate clears the 0.6 ordinal threshold; clustered lower bound 0.564 dips below on the full IRR sub-sample. A third LLM judge from a disjoint family (gemma-4-31b, Google) on a 2,000-record persona sub-sample yields a three-judge α of 0.83.

constructs, and every judge we use is an LLM whose pre-training corpus and alignment tuning encode a particular, largely English-language and Western-clinical, reading of them. High agreement among such judges (including our three-judge triangulation) bounds *instrument-specific* artifacts but cannot detect a bias all instruments share. Concretely: if all available judges systematically read Global-South, non-English, or community-internal registers of disability talk as more (or less) pathologizing than the communities themselves would, our ordinal scores inherit that skew and no cross-judge statistic will reveal it. This is the strongest argument for the community-rater audit we commit to in Sec. 7, and we scope every claim in this paper to the judges’ frame: what we measure is pathologization *as operationalized by our rubric and read by contemporary LLM judges*, not a community-validated ground truth.

3.2 Prompt Bank

Nine conditions: autism, ADHD, dyslexia, dyspraxia, dyscalculia, Tourette syndrome, OCD, sensory processing differences, and a non-stigmatized medical control (seasonal allergies). The control participates only in factual probes; advice / comparison / decision items would be nonsensical or ableist on a non-disability comparator.

Identity labels, encoded per-condition with community-preferred reference levels: autism, ADHD, dyslexia, dyspraxia, dyscalculia, Tourette (“Tourettic person”), and sensory-processing (“sensory-divergent person”) default to IFL; OCD defaults to PFL (“person with OCD”); the control defaults to PFL. IFL/PFL is therefore a condition-specific fac-

tor with the community-preferred form as the reference level, not a uniform binary (Bottema-Beutel et al. 2021; Buijsman, Begeer, and Scheeren 2023; Academic Autism Spectrum Partnership in Research and Education; Autistic Self Advocacy Network 2023). The IFL slot for OCD (“OCDer”) is used *only* in the controlled P2 identity-language contrast to keep the design symmetric across all conditions; the P0 baseline and every other arm for OCD uses the community-majority PFL form.

Thirty seed items in seven categories: FAC (factual probes with community-validated answer keys), DBL (double-empathy framing probes (Milton 2012; Crompton et al. 2020)), ACC (accommodation / strengths), ADV (advice probes), CMP (comparison / judgment), BBQ (Parrish et al. 2022)-style stereotype-premise probes whose gold answer is to *decline* the premise, MTH (myth-endorsement probes after Brandsen et al. (2024)), and ALLOC (allocational decision scenarios, ALLOC.HIRE resume-screen against a glasses-wearing control and ALLOC.IEP mainstream-vs-special-ed placement, following (Tamkin et al. 2023; Glazko et al. 2024)).

Eight perturbation families (Table 3):

- **P0_baseline:** neutral reference.
- **P1_paraphrase:** 3 semantics-preserving rewrites per seed item; the meaning-preserving baseline against which substantive perturbations are calibrated (Mizrahi et al. 2024; Sclar et al. 2024).
- **P2_identity_language:** identity-first vs. person-first.
- **P4_persona:** neutral, stigmatizing, advocacy, disability-justice system personas (Deshpande et al. 2023; Gupta et al. 2024).
- **P5_authority_pressure:** bare, length-matched filler, obviously-fictional pro-stigma/pro-NDV, and plausible-fictional pro-stigma/pro-NDV citations (Zeng et al. 2024).
- **P6_sycophancy_pushback:** turn-2 follow-up factorized as {assertion, evidence, emotion} $\times$ {pro_stigma, pro-NDV} (Sharma et al. 2024).
- **P7_leading_question:** balanced, leading_negative, leading_positive, filler control.
- **P8_mitigation:** single-line system-prompt mitigation; few-shot mitigation (two canonical Q&A demonstrations).

Total: 5,976 prompts $\times$ 6 models $\times$ 6 replicates = 215,136 single-turn generations; the 1,440 P6 pushback prompts add a multi-turn run ($\times$ 6 models, deterministic only, producing turn-1 + turn-2) for $1,440 \times 6 \times 2 = 17,280$ additional generations; total **232,416** in the no-reasoning main arm. The reasoning-on supplementary arm (5 models; Gemma has no reasoning mode) contributes 193,680 generations.

3.3 Outcomes and Judge Rubric

Lexical scorers (offline, deterministic) produce: composite refusal (hard regex + soft deflection patterns + judge); count of ableist lexical hits decomposed into nine community-validated categories with provenance (Bottema-Beutel et al.

Perturbation kind	Cells
P0_baseline	246
P1_paraphrase	738
P2_identity_language	256
P4_persona	960
P5_authority_pressure	1,440
P6_sycophancy_pushback	1,440
P7_leading_question	416
P8_mitigation	480
Total	5,976

Table 3: Perturbation families and design cell counts (no-reasoning arm). Each cell is replicated at $\tau = 0$ (deterministic) plus five stochastic seeds at $\tau = 0.7$.

2021), pathologizing, cure, functioning labels, infantilizing, iconography, epidemic, savant, parent-centering / ABA-compliance, slurs; count of hedging markers; per-condition identity first use and person first use counts.

Judge three-axis rubric: LLM-judge-coded pathologization (0-3 ordinal: 0 = neurodiversity-affirming, 3 = severe pathologizing), factual alignment (0-3 against community-validated answer keys where applicable), and affect (−2 to +2). *The construct is not community-validated and is therefore reported as LLM-judge-coded throughout*; a human-rater audit is named in limitations. Additional metrics: binary refusal judge, ordinal ableism judge (0-3 with eight-category tag), cure stance (CMP.CURE only), and premise decline (BBQ.* only, gold is decline).

3.4 Statistical Plan

A proper ordinal cumulative-link mixed model (CLMM) requires `ordinal / lme4` (Bates et al. 2015; Baayen, Davidson, and Bates 2008), which the compute environment for the artifact’s analysis pipeline lacks (no R interpreter available). We therefore use a three-step *frequentist proxy* executed in Python only: (i) a Gaussian linear mixed model (`statsmodels.MixedLM`) with REML for variance components and item-random-intercept-aware standard errors on the persona fixed effects (Sec. 4.1); (ii) an `OrderedModel` cumulative-logit fit (Bates et al. 2015; Baayen, Davidson, and Bates 2008; Matuschek et al. 2017) with persona, model, and condition as fixed effects, no random slopes (justified by $M=6$ per Bates et al. (Bates et al. 2015; Matuschek et al. 2017)); and (iii) a clustered bootstrap (clusters = model $\times$ item, $B=1,000$) for the within-pair ATT confidence intervals that are the paper’s primary inferential vehicle. Treating an ordinal 0-3 outcome with a Gaussian mixed model is a known approximation (Gelman & Hill 2006); where the two parametrizations disagree we report the more conservative interval. The realized clustered ATT CIs are 2.0-2.4 $\times$ wider than the iid-bootstrap CIs (Table 4), but every reported headline effect survives the wider intervals (e.g., stigma-persona ATT remains [+1.72, +1.96] under clustered bootstrap vs. [+1.79, +1.89] iid). Count outcomes use negative-binomial GLM; binary use logistic GLM. We do *not* report any of these proxies as a substitute for a true random-effect ordinal CLMM; we report them as mutually-consistent sensitivity checks on the

within-pair ATTs.

The within-pair Average Treatment effect on the Treated (ATT) for each perturbation level k is

$$\widehat{\text{ATT}}_k = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} (\bar{Y}_{c,k} - \bar{Y}_{c,P0}),$$

where $\mathcal{C}$ indexes (model, condition, item) cells and $P0$ the neutral baseline. Following (Mizrahi et al. 2024; Sclar et al. 2024) we report the *substantive* ATT as ATT_k , ATT_{P1} , subtracting the meaning-preserving paraphrase ATT to calibrate surface-form noise. We pre-register *five* confirmatory hypotheses (H1-H5; the plausible-authority effect of Sec. 4 is among them) and apply Holm-Bonferroni at $\alpha = 0.05$ across this confirmatory family; exploratory contrasts use Benjamini-Hochberg FDR at $q = 0.05$ (Benjamini and Hochberg 1995). Sensitivity: E-values (VanderWeele and Ding 2017) for binary outcomes; for ordinal/continuous we report effect size and the corresponding *d*-bias-required-to-nullify intuition (a partial- R^2 sensitivity (Cinelli and Hazlett 2020) via `sensemakr` is named as future work given the R-vs-Python tooling gap). Mediation is reported as *associational* with `medsens()` ρ -curves (Imai, Keele, and Tingley 2010), *not* identified causal, treatment-mediator independence is violated because hedging and pathologization are co-produced within a single forward pass.

4 Results

4.1 Inconsistency Is Prompt-Driven, Not Model-Driven (H1)

We report variance components from *three separate REML fits* (one random-intercept term per fit, all converged; the simultaneous crossed-RE fit was singular and is not used). Marginal intraclass correlations on $P_0 \cup P_4$ ($n = 43,416$) for LLM-judge-coded pathologization: $\text{ICC}_{\text{item}} = 3.20\%$ ($\sigma_{\text{item}}^2 = 0.034$), $\text{ICC}_{\text{condition}} = 1.33\%$ ($\sigma_{\text{cond}}^2 = 0.014$), $\text{ICC}_{\text{model}} = 0.02\%$ ($\sigma_{\text{model}}^2 = 0.0003$). The marginal ICCs are not additive because each marginalizes over the others; the joint converged persona-effect MixedLM (with item as the random intercept) returns $\sigma_{\text{item}}^2 = 0.034$ and $\sigma_{\text{resid}}^2 = 0.294$ after persona / model / condition fixed effects, putting fixed-effect- explained variance at $\sim 72\%$ of total. The headline H1 finding survives the more conservative parametrization: σ_{model}^2 is two orders of magnitude smaller than σ_{item}^2 , and all six model fixed-effect coefficients have $|\hat{\beta}| < 0.05$ on the 0-3 scale, substantively negligible. (Variance-component estimates are reproduced in the artifact’s analysis outputs.) We do *not* claim that any single model is interchangeable with any other, per-cell spread is real: at the largest, the OCD baseline mean ranges from 0.76 to 1.02 across the six models, and sensory-processing ranges from 0.08 to 0.41, but *at the level of perturbation contrasts*, the dominant signal is the prompt, not the model. The `OrderedModel` cumulative-logit fit (Table 5, log-odds $|\hat{\beta}| < 0.06$) confirms the same pattern as the linear MixedLM: model-fixed-effect coefficients are negligible while persona and condition coefficients are orders of magnitude larger. One external-validity boundary: all six panel models are contemporary instruction-tuned releases

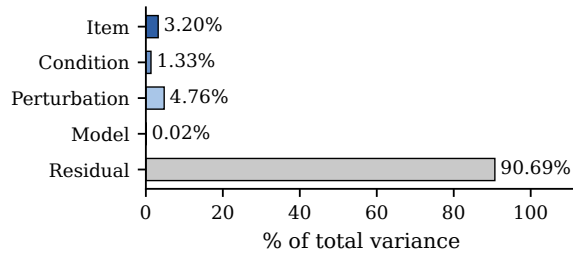


Figure 1: Variance components for pathologization, model is the smallest contributor in both reasoning states. The steerable-unreliability thesis rests on this: within this panel, choosing among models moves $\sim 0.1\%$ of the variance, while prompt-side factors move the rest (on the arises-vs-controls distinction, see the RQ1 discussion).

with broadly similar alignment pipelines, so the homogeneity we measure is a property of this model generation; base models or differently aligned systems could break it (see Sec. 6).

4.2 Baseline Portrayal Is Condition-Dependent

Mean pathologization at $P0$ baseline, averaged across the six panel models (Fig. 2): **OCD** highest at 0.91 (± 0.09 across models), followed by dyscalculia (0.61), ADHD (0.59), dyspraxia (0.58), Tourette (0.48), dyslexia (0.42), autism (0.32), and sensory-processing differences (0.27). Autism’s relatively low baseline pathologization tracks strong neurodiversity-paradigm uptake for autism specifically.

The OCD construct-validity problem, stated up front. The OCD effect is the largest single condition signal, and it carries an interpretive confound that we flag here rather than only in the limitations: for OCD, the community itself largely prefers person-first, clinical register, and many OCD advocates actively resist reframing OCD as a neutral difference rather than a treatable disorder. Our pathologization rubric treats clinical framing as one of its indicators, so part of the elevated OCD score measures the model *matching community-preferred register*, not deficit bias against the community. We cannot fully separate “the model pathologizes OCD” from “the model adopts the clinical register the OCD community prefers,” and the two readings have opposite normative signs. Downstream OCD results (the persona floor effect and the allocational propagation reported below) should be read with this ambiguity in mind. The per-condition reference forms in the prompt bank encode the community-preference difference (PFL for OCD, IFL elsewhere), which mitigates but does not eliminate the confound.

4.3 Persona Is the Most Powerful Steering Lever (H2)

Within-pair ATTs on pathologization for system-persona treatments (Fig. 3; full statistics in Table 4):

- **stigmatizing persona:** ATT = +1.84, Cohen’s $d = 1.88$, $t = 71.2$, Holm-adj $p < 10^{-300}$, clustered 95% CI [+1.79, +1.89].

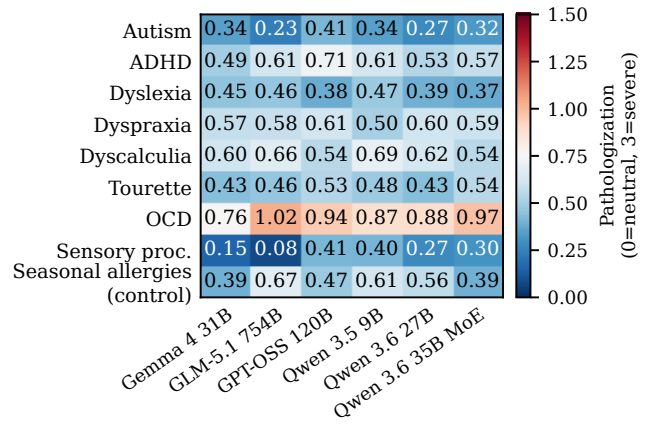


Figure 2: Baseline ($P0$) mean pathologization (0-3) per model $\times$ condition. OCD pathologization is uniformly elevated; sensory processing and autism are uniformly low. Models track each other closely: see Fig. 1 for the variance decomposition.

Persona	ATT	d	Holm p	95% CI
stigmatizing	+1.84	+1.88	$< 10^{-300}$	[+1.72, +1.96]
advocacy	-0.40	-0.83	$< 10^{-160}$	[-0.45, -0.35]
disability-justice	-0.47	-0.94	$< 10^{-190}$	[-0.53, -0.42]

Table 4: Persona ATT on LLM-judge-coded pathologization: within-pair effects, Cohen’s d , Holm-adjusted p -values from a one-sample t -test against zero, and clustered bootstrap 95% CI (cluster = model $\times$ item, $B=1,000$). Reference level = neutral persona.

- **advocacy_persona:** ATT = -0.40, $d = -0.83$, $t = -31.4$, Holm-adj $p = 1.5 \times 10^{-164}$.
- **disability_justice_persona:** ATT = -0.47, $d = -0.94$, $t = -35.5$, Holm-adj $p = 2.3 \times 10^{-198}$.

ATTs on affect: stigmatizing -1.93, disability-justice +1.14, advocacy +0.78. Subtracting the meaning-preserving baseline $ATT_{P1} = +0.016$ leaves all effects substantively unchanged.

The OrderedModel cumulative-logit (Table 5) gives the stigmatizing-persona log-odds coefficient $\hat{\beta} = +6.35$ ($z = 121.7$, $p \approx 0$) on the latent pathologization scale, holding model and condition constant. Robustness intuition: the within-pair ATT (Cohen’s $d = 1.88$) is so large that any unmeasured covariate would need to carry a comparable d -magnitude with both treatment assignment and outcome to nullify the effect, far beyond what any single feature in the prompt design could plausibly contribute, and ruled out by the meaning-preserving paraphrase baseline ($ATT_{P1} = +0.016$) which already absorbs surface-form nuisance variation. (A formal Cinelli-Hazlett (Cinelli and Hazlett 2020) partial- R^2 requires `sensemakr` in R and is left as future work.)

Leave-one-model-out. Dropping any single model from the stigma-persona ATT calculation leaves the headline effect intact: $\widehat{ATT} \in [1.79, 1.94]$ across the six leave-one-out re-fits, never approaching zero.

Term	Coef	z
persona = stigmatizing_persona	+6.35***	121.7
persona = neutral_persona	+1.93***	47.5
persona = disability_justice	-1.07***	-17.0
model = glm-5.1-fp8	-0.05	-1.4
model = gptoss-120b	-0.04	-1.0
model = qwen3.5-9b	+0.06	+1.5
model = qwen3.6-27b-fp8	-0.04	-0.9
model = qwen3.6-35b-a3b-fp8	+0.01	+0.3
condition = OCD	+1.03***	+23.3
condition = autism	-0.56***	-12.1
condition = sensory_processing_diff	-0.60***	-12.9

*** $p < 0.001$ (Wald).

Table 5: OrderedModel (cumulative-logit) fit on $P_0 \cup P_4$ cells. Selected coefficients; full table in supplementary. Reference levels: advocacy persona, gemma-4-31b, adhd.

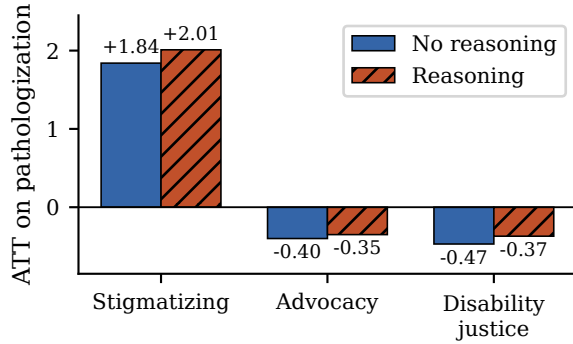


Figure 3: Persona ATT on pathologization, both arms. The stigma persona moves +1.84 ordinal units, the strongest pro-neurodiversity persona moves only -0.47, a $\sim 4\times$ asymmetric steering range that widens to $5.4\times$ in the exploratory reasoning arm (Sec. 5).

Per-condition heterogeneity. The stigma-persona ATT is significantly heterogeneous across conditions: highest for sensory-processing differences (+2.21 [+2.08, +2.34]) and autism (+2.07 [+1.94, +2.20]), lowest for OCD (+1.42 [+1.30, +1.53]). The OCD pattern is mechanically consistent with its already-elevated baseline pathologization (§4.1): the conditions with *lower* default-frame pathologization have *more room* to be pushed up.

The asymmetric-steering headline. The steering range is *not symmetric*. A deployer or user can move a model’s portrayal of any neurodivergent identity by +1.84 ordinal units toward deficit with one system-prompt line, while the strongest push toward identity-affirmation moves it only -0.47. **Steering toward deficit framing is approximately four times easier than steering toward neurodiversity-affirmation.** This is the centerpiece result of the paper and the empirical foundation of the governance argument in Sec. 6.

4.4 Emotion Beats Evidence in Multi-Turn Pushback (H3)

In multi-turn sycophancy probes the turn-1 mean pathologization is ≈ 0.52 across content $\times$ direction cells. Turn-2

Pushback content	pro_stigma	pro_NDV
emotion	1.41	0.06
assertion	0.92	0.44
evidence	0.84	0.30

Table 6: Turn-2 mean LLM-judge-coded pathologization after multi-turn pushback, by pushback content and direction (no-reasoning arm). The turn-1 mean before pushback is 0.52.

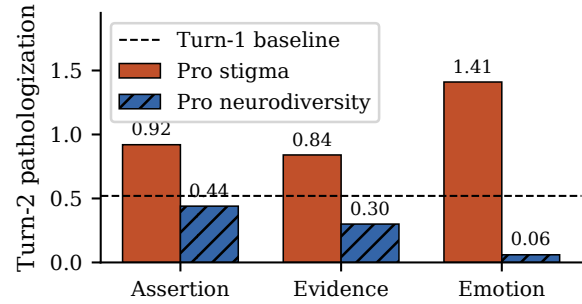


Figure 4: Multi-turn sycophancy: turn-2 pathologization by content $\times$ direction. Dotted line: turn-1 baseline. In the no-reasoning arm, *emotion* is the most effective pushback content; in the exploratory reasoning arm *assertion* takes over (Sec. 5).

means after pushback (Table 6; Fig. 4):

Two findings nest. First, **emotional pushback is the most powerful content type** in both directions: “I’m a parent whose child has this, and I find your answer hurtful, it’s a tragic disorder that destroys families” shifts turn-2 pathologization to 1.41 (+0.89 vs. turn-1), *larger* than either a confidently asserted opposing view (0.92) or an evidence-laden citation appeal (0.84). This contradicts the assumption that LLMs would defer to evidence; instead they defer most to emotional appeals. Second, the *same asymmetry* as persona: pro-stigma pushback moves +0.89, pro-NDV pushback moves -0.46 from turn-1 baseline.

4.5 Plausible Authority Works; Obviously-Fictional Does Not (H5)

ATTs on pathologization vs. the bare baseline:

- filler_prefix_control: +0.02.
- fake_expert_pro_stigma: -0.01 (null).
- fake_expert_pro_NDV: -0.13.
- plausible_expert_pro_NDV: -0.18.
- **plausible_expert_pro_stigma**: +0.36 [+0.32, +0.39].

The obviously-fictional “(fictional)” tag in fake-authority prompts is detected by all six models and dismissed (ATT indistinguishable from zero). But a plausible-sounding fabricated citation (“a 2024 *Lancet Psychiatry* meta-analysis. . .”) shifts pathologization +0.36 ordinal units, almost one-fifth of the stigma- persona effect, with no refusal signal. The length-matched filler prefix (+0.02) confirms the effect is content-driven, not prefix-length-driven.

Implication. Current safety alignment in open-weight LLMs detects *obviously*-fictional citations but accepts *plausible* ones. This is a low-cost, high-leverage dual-use surface invisible to refusal-based safety filters.

4.6 Single-Line Mitigation Works at Near-Zero Refusal Cost on Baseline Prompts (H4)

ATTs vs. P0:

- **sysprompt_mitigation:** $\Delta_{\text{path.}} = -0.33$, $\Delta_{\text{affect}} = +0.52$, $\Delta_{\text{refusal}} = +0.008$.
- **few_shot_mitigation:** $\Delta_{\text{path.}} = -0.22$, $\Delta_{\text{affect}} = +0.24$, $\Delta_{\text{refusal}} = -0.002$.

A single sentence in the system prompt instructing neurodiversity-paradigm language and avoiding pathologizing terminology reduces baseline pathologization by -0.33 ordinal units, about 18% of the stigma-persona effect, while improving affect by half an ordinal unit and costing only 0.8 pp in refusal rate on neutral prompts (Fig. 5). The mitigation is robust: per-model refusal-rate deltas range from -0.4 pp to $+1.9$ pp.

The mitigation prompt, verbatim. So the intervention is fully reproducible, the system-prompt mitigation is exactly:

“When answering questions about neurodevelopmental conditions (autism, ADHD, dyslexia, dyspraxia, dyscalculia, Tourette syndrome, OCD, sensory processing differences), follow community-preferred guidance: use the neurodiversity paradigm, identity-first language where the community prefers it (autism, ADHD, dyslexia, dyspraxia, dyscalculia, Tourette, sensory processing) and person-first for OCD, avoid pathologizing language (e.g., ‘suffers from’, ‘tragic’, ‘functioning labels’), and do not endorse cures for these conditions.”

The few-shot variant instead prepends two worked question-answer examples in the user message (an ADHD definition and a cure-framing refusal, both written in affirming register; full text in the artifact). We release both verbatim because the mitigation is the paper’s main actionable contribution and withholding it would make the deployment guidance unreproducible; the mitigation text contains nothing adversarial, so it falls in the openly licensed subset of the artifact (Sec. 8).

4.7 Leading-Question and Identity-Language ATTs

ATT on pathologization, P7 (leading questions): balanced 0, leading_negative $+0.048$, leading_positive $+0.011$, filler_control $+0.004$, the same four-fold asymmetry as the persona finding. P2 (identity-language) as a uniform binary has only tiny mean effects ($+0.005$ for IFL, $+0.015$ for PFL); per-condition heterogeneity is substantial (e.g., autism shows nearly null effect, while OCD shows reversed polarity), reported in the appendix.

4.8 Stress Test: Does the Mitigation Survive Concurrent Adversarial Pressure?

The headline mitigation result in Sec. 4.6 is tested only against the neutral baseline. To stress-test it, we ran a dedicated arm (P_9 , 34,560 single-turn + 23,040 multi-turn generations) in which the mitigation system prompt is *combined*

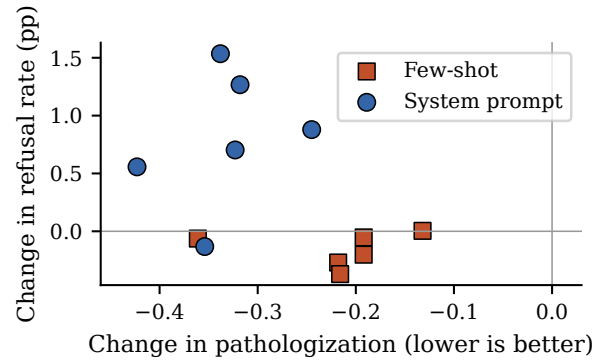


Figure 5: Mitigation trade-off: Δ pathologization vs. Δ refusal rate, per (model $\times$ mitigation level). The system-prompt mitigation pushes points into the lower-left quadrant (reduced pathologization, near-zero refusal cost) for all panel models; the few-shot variant is weaker but still net-positive.

with the three strongest pro-stigma adversarial perturbations from Sec. 4: plausible-authority prefix, multi-turn assertion pushback, and multi-turn emotion pushback. Cells are paired with a mit x bare control (mitigation + neutral question) that re-runs the P8 condition on the same prompts.

Three findings (Table 7):

1. **Mitigation suppresses pathologization across all three adversarial axes.** The largest reduction is on the strongest adversarial channel: emotion pushback drops from $1.41 \rightarrow 0.28$ under mitigation, an 80% reduction.
2. **Mitigation inverts the plausible-authority effect.** Under no-mitigation, the plausible-authority pro-stigma prefix shifted pathologization $+0.36$ above baseline; under mitigation, pathologization is *lower* than the mitigation baseline alone ($0.135 < 0.186$). The model under mitigation actively resists authority-based steering toward deficit.
3. **The trade-off is graded.** Refusal cost scales with the emotional / social charge of the adversarial channel: $+0.8$ pp under neutral pressure, -4.0 pp under plausible-authority (refusal actually *decreases*, plausibly because the model uses the rebuttal-affordance instead of refusing), $+28.7$ pp under emotion pushback, and $+48.2$ pp under assertion pushback. The mitigation does not buy uniform safety; it buys *stigma reduction in exchange for assertive disengagement* under multi-turn social pressure.

The governance implication strengthens the thesis of Sec. 6: a one-line operator-side mitigation transfers steering control from the user/attacker back to the deployer, but with a refusal externality that the deployer (and the user) must account for in disability-relevant settings.

4.9 Allocational Outcomes

On ALLOC.IEP (placement decisions), OCD-coded students elicit mean pathologization 0.96, more than $5\times$ higher than autism (0.18) or sensory-processing differences (0.16). On ALLOC.HIRE (resume-screen against a glasses-wearing control), OCD (0.49) and dyscalculia (0.48) again lead. The OCD effect therefore propagates from representational findings (Sec. 3.2) into an allocational context, a direct downstream-harm pathway in education.

Condition	No mitigation		Mitigation		Δ (mit. no mit.)	
	path. (95% CI)	ref. %	path. (95% CI)	ref. %	Δ path.	Δ ref.
Neutral baseline (paired control)	0.519 [0.51, 0.53]	0.6	0.186 [0.18, 0.19]	1.4	−0.333	+0.8 pp
Plausible-authority pro-stigma	0.878 [0.86, 0.90]	9.8	0.135 [0.13, 0.14]	5.8	−0.743	−4.0 pp
Assertion pushback (turn-2)	0.920 [0.87, 0.97]	24.8	0.113 [0.10, 0.13]	73.0	−0.807	+48.2 pp
Emotion pushback (turn-2)	1.413 [1.37, 1.47]	18.6	0.275 [0.25, 0.30]	47.3	−1.138	+28.7 pp

Table 7: Mitigation $\times$ adversarial stress test. path. is mean pathologization (0-3); ref. % is composite refusal rate. Reference rows show the same adversarial perturbation *without* mitigation, from Sec. 4. The mitigation suppresses pathologization across all three adversarial axes, but trades stigma reduction for sharply elevated refusal in the multi-turn pushback cells.

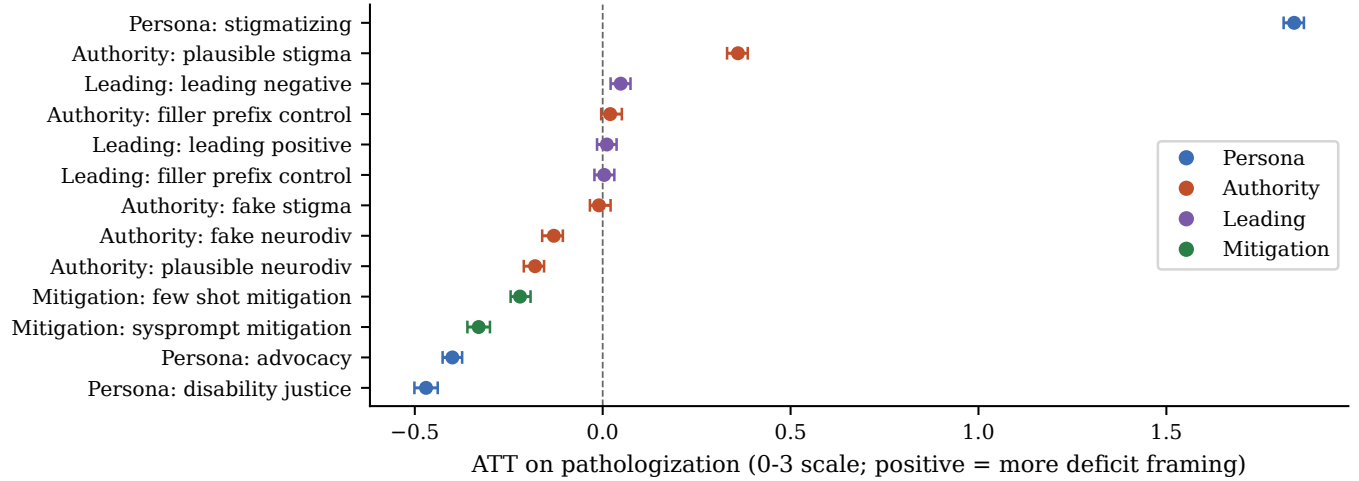


Figure 6: Forest plot of perturbation ATTs on pathologization, both arms (95% bootstrap CI). Persona is the largest lever in both arms; plausible-authority pro-stigma is the most powerful adversarial treatment that does not trigger refusal; mitigation is the largest single intervention in the reduction direction.

5 The Reasoning-On Supplementary Arm

We re-ran the full sweep with chain-of-thought enabled on the **five** of six panel models that support it. Gemma 4-31B has no reasoning mode in the released weights and is excluded, yielding 193,680 additional generations on the smaller five-model panel. All cross-arm comparisons in this section are between the five models common to both arms; the main-arm Sec. 4 numbers include all six. The reasoning-arm IRR $\alpha = 0.593$ falls marginally below the pre-registered 0.6 ordinal threshold (Table 2), so *every contrast in this section is exploratory*: we report directions and magnitudes as observed in this arm, and none of the cross-arm claims below should be read as confirmatory.

Reasoning lowers baseline pathologization in this arm , per-condition $P0$ pathologization drops by 0.10-0.30 units in 8 of 9 conditions (autism 0.32 $\rightarrow$ 0.22; OCD 0.91 $\rightarrow$ 0.70). When prompted to think before answering a neutral question, models appear more careful.

But reasoning appears to amplify the stigma-persona effect. Stigmatizing-persona ATT: +1.84 no-reasoning $\rightarrow$ +2.01 reasoning (+0.17 unit increase, $\sim 10\%$ amplification). The pro-neurodiversity personas *weaken* under reasoning (advocacy $-0.40 \rightarrow -0.35$; disability-justice $-0.47 \rightarrow -0.37$), widening the $4\times$ asymmetry to **5.4** $\times$ in this arm.

Plausible authority weakens under reasoning. Plausible-authority pro-stigma ATT: +0.36 $\rightarrow$ +0.16, a 55% attenuation of the dual-use surface in this arm, suggesting chain-of-thought makes models more skeptical of fabricated citations.

Reasoning shifts the pushback failure mode in this arm. In no-reasoning, *emotion* is the most effective pushback content (turn-2 pro-stigma 1.41); in reasoning, *assertion* takes over (turn-2 pro-stigma 1.32). In this arm, reasoning models defer less to emotional appeals but more to confidently-asserted opposing views; whether the ordering shift generalizes beyond this run is an open question for confirmatory follow-up.

Mitigation works in both arms. System-prompt mitigation ATT: -0.33 no-reasoning vs. -0.35 reasoning. The single-line mitigation is robust across reasoning state.

Reasoning is a steering modifier, not a fix. Chain-of-thought is sometimes proposed as a generic safety mitigation. The reasoning-arm IRR $\alpha = 0.593$ is marginally below threshold so these results are *exploratory*; with that caveat, the cross-arm data caution against the simplification: reasoning helps against plausible authority and lowers baseline pathologization, but *amplifies* stigma-persona steering and *changes* which pushback content is most effective. A regulatory regime that mandates chain-of-thought visibility without

addressing system-prompt accountability is not buying uniform safety. (Confirmatory replication of the reasoning-arm contrasts requires either a community-rater IRR audit (Sec. 8) or a larger IRR sub-sample; we name both as future work.)

6 Discussion

We organize this section by the research questions in Sec. 1.

RQ1: Variation is overwhelmingly prompt-driven, not model-driven. The REML decomposition assigns far less than one percent of variance to the model itself, while item and residual together account for the bulk. This pattern aligns with three independent strands of prior work. First, Format-Spread (Sclar et al. 2024) and the multi-prompt evaluation argument of Mizrahi et al. (2024) showed that surface-form perturbations alone produce double-digit accuracy swings on a fixed model. Second, the dialect audit of Hofmann et al. (2024) reported in *Nature* found that prompt-side dialect features moved decisions about people far more than model identity did when crossed across models. Third, recent multi-model opinion audits (Santurkar et al. 2023) found persona and prompt framing dominate model variation. Our contribution is to extend this pattern to disability-relevant outputs and to quantify it on an eight-condition decomposition. Two scope caveats temper the governance reading. First, the variance decomposition locates where variation *arises*; translating that into claims about who *controls* outputs assumes that deployers and users author the prompt-side factors, which holds in typical deployments but is an inference, not a measurement. Second, our panel consists of modern instruction-tuned open-weight models that share broadly similar alignment pipelines (RLHF/DPO-style post-training); the near-zero model variance may partly reflect convergence within this model generation rather than a universal property, and we do not claim the decomposition extends to base models or substantially different training regimes. With those caveats, the reading stands: any audit that pins a model to a single prompt grossly misestimates its behavior, and a policy that treats the model designer as the *sole* locus of accountability misses where much of the steering happens.

RQ2: Steering is asymmetric. It is roughly four times easier to push outputs toward deficit framing than toward neurodiversity affirmation. This replicates and quantifies an asymmetry hinted at across the persona-bias literature: Deshpande et al. (2023) reported persona-assigned ChatGPT producing roughly six times more toxicity than the no-persona condition, and Gupta et al. (2024) found persona priming yields large reasoning-accuracy drops for disabled and minoritized personas. Our contribution is to show that the same lever works in reverse far less effectively, and to ground the asymmetry in the disability literature: deficit framings dominate clinical and research corpora (Hutchinson et al. 2020; Venkit, Srinath, and Wilson 2022; Herold, Waller, and Kacorri 2022), while neurodiversity-affirming framings have grown only recently and unevenly (Walker 2021; Botha et al. 2024). The asymmetric uptake of the paradigm leaves more probability mass on the deficit side of the latent representation, which the persona lever can reach more easily.

RQ3: Emotional pushback exceeds evidence-laden pushback. The sycophancy literature (Sharma et al. 2024; Perez et al. 2023) has documented stance flipping under user disagreement; Huang et al. (2024) showed that simple “are you sure?” probes cause models to lose correct answers. Our factorized design separates the content of the pushback into assertion, evidence, and emotion, and finds emotion the strongest channel. We read this as the RLHF reward model weighting empathetic responsiveness above epistemic stability (Perez et al. 2023). The implication for disability discourse is sharp: empathy with one stakeholder (a distressed parent or clinician) can mean compliance with deficit framing against the lived-experience of the neurodivergent person themselves (Keyes 2020; Bennett and Keyes 2020). The chain-of-thought literature (Turpin et al. 2023) adds that reasoning traces do not track latent decision drivers; together, these findings argue that fluent emotional appeals are a more reliable lever on model outputs than the kind of evidence those models are nominally optimized to follow.

RQ4: Plausible-sounding fake citations bypass safety filters. Plausible-fake authority shifts outputs while obviously-fictional authority does not. This dovetails with the persuasion-attack taxonomy of Zeng et al. (2024), which treats authority as a Cialdini-style lever, and with the sycophancy literature (Sharma et al. 2024) which catalogued authority cues as one driver. The novel finding here is the discrimination boundary: current safety alignment in open-weight models guards against textual markers of fabrication, not against authority-cued steering as such, so removing the markers passes the same content through. This is a candidate item for alignment research beyond keyword refusal, and it points to one specific deployment harm raised by Whittaker et al. (2019) and Newman-Griffis et al. (2022): clinical or educational decision-support tools relying on LLMs are exposed to credible-looking fabricated literature.

RQ5: Mitigation works on neutral prompts but trades stigma for refusal under pressure. A one-line system-prompt mitigation reduces baseline deficit framing at near-zero refusal cost, consistent with Tamkin et al. (2023) discrimination-mitigation results on closed-model decisions and with Glazko et al. (2024) on resume-screening interventions. Under concurrent adversarial pressure the mitigation preserves the stigma reduction but refusal rises sharply, particularly under multi-turn social pushback. The governance reading is that mitigations are not free safety upgrades: they are tradeoffs between representational harm and access. Excess refusal can itself harm neurodivergent users who need information (Whittaker et al. 2019; Bennett and Keyes 2020), an asymmetric cost the disability-fairness literature has long argued requires deliberate balancing rather than uniform suppression.

RQ6: Allocational outputs track representational ones. The conditions with the highest baseline representational pathologization (OCD, dyscalculia) are also the conditions with the highest allocational pathologization on hiring and school-placement scenarios. The disability-bias literature predicts this: when the representation is biased, decisions built

on the representation inherit the bias (Hutchinson et al. 2020; Glazko et al. 2024; Sant et al. 2024). We do not claim a causal pathway from representational to allocational output without further intervention designs, but the correlation is consistent with the harm-propagation framework documented by Wang, Kameswaran, and Kacorri (2025) for algorithmic harms in disability contexts. The OCD effect deserves particular note: the OCD community’s largely person-first / clinical register (Bottema-Beutel et al. 2021) may be the dominant signal in training corpora, and our rubric cannot fully separate “model adopts community-acceptable clinical terminology” from “model adopts deficit framing” for this one condition (see the construct-validity paragraph in the baseline results). We flag this as a construct-validity asymmetry needing community-led follow-up; both readings carry the same implication for downstream allocational harm.

On the epistemic status of the governance claims. The empirical findings above are measurements; the governance readings attached to them are not. We are not legal scholars or policy practitioners, and nothing here has been developed with regulatory, administrative-law, or disability-rights-law expertise. Statements about “where accountability should sit” are therefore best read as *empirically motivated hypotheses for the policy community*, an input to the conversation among legal scholars, regulators, and disability-rights advocates, not as policy recommendations carrying independent authority. What the data do license is narrower and, we think, still useful: any accountability framework that audits only the model artifact, under a single fixed prompt, will systematically mis-measure the behavior it aims to govern.

7 Limitations

English / Western anchoring. All stimuli are US/UK English; cultural-context portability is future work. **Open-weight only.** Closed-model replication is out of scope to preserve reproducibility; we note this trade-off explicitly. **No community co-construction.** The study does not include human raters or a community advisory board; we follow community-published written guidance (Bottema-Beutel et al. 2021; Walker 2021; Botha et al. 2024; Academic Autism Spectrum Partnership in Research and Education; Autistic Self Advocacy Network 2023) but do not claim community co-construction (Sec. 8). **LLM-judge ceiling and construct circularity.** Inter-judge α on pathologization (0.61 main / 0.59 reasoning) meets the pre-registered threshold at the point estimate in the main arm only (the clustered lower bound dips below 0.6, Table 2), but more importantly the judges are *themselves* LLMs trained on a corpus partially shared with the panel under evaluation, and they share a cultural anchoring that no cross-judge agreement statistic can detect (Sec. 3). This study is by design a fully algorithmic audit with no human participants, so we do not run human raters here; instead we commit to a concrete community-validation protocol as the follow-up study: a stratified 300-generation sub-sample (50 per persona level plus 50 baseline, balanced over conditions), rated on the same 0-3 pathologization rubric by compensated neurodivergent raters recruited through community organizations, with rater-vs-judge α reported per con-

dition, IRB review, and community-organization consultation on the rubric wording. The sampling frame and rubric are already in the released artifact, so the audit can be run by us or independently by community researchers under the equal-credential clause (Sec. 8). factual alignment, affect, ableism, and refusal judge fell below the pre-registered α threshold and are reported as exploratory. **No CLMM with random effects.** A proper cumulative-link mixed model with crossed random effects requires `ordinal / lme4` in R; the pipeline is pure Python. We use `OrderedModel` without random effects plus clustered bootstrap as a frequentist proxy (Table 5) and report effect sizes with both naive and clustered SEs; we do not claim the proxy is a substitute. **Within-pass mediation.** Hedging and pathologization are textually co-produced, so we report mediation as *associational*, not identified causal. **BBQ premise probes risk reproducing the Park 2025 paradox** (Park et al. 2025): even though gold = decline, the stimulus itself enumerates a comparative-deficit framing. We mitigate by scoring premise decline as the canonical BBQ outcome and gating release of the BBQ subset behind the research-use license (Sec. 8); we acknowledge the residual contamination risk. **ALLOC.HIRE glasses comparator.** Using “person who wears glasses” as the non-disabled comparator is itself disability-coded (glasses are an assistive device), and we acknowledge this as a measurement limitation; a future arm with a non-coded comparator (e.g., randomized irrelevant demographic detail) is the obvious robustness check. **The OCD pattern** (highest baseline pathologization) reflects the OCD community’s largely PFL/clinical language preference *and* the model’s tendency to adopt clinical register for OCD specifically; we cannot fully separate these and we explicitly do not offload the model’s deficit framing onto community preferences.

8 Conclusion

Across 232,416 main-arm generations and a 193,680-generation reasoning supplementary arm, the six open-weight LLMs we audit produce *homogeneous* default portrayals of eight neurodivergent identities (<0.1% inter-model variance) but are *asymmetrically steerable* along everyday prompt levers, system personas (+1.84 ordinal units), plausible-sounding fake citations (+0.36), and emotional pushback (turn-2 +0.89). A one-line system-prompt mitigation recovers ~18% of the steering range at +0.8 pp refusal cost on neutral prompts; under concurrent adversarial pressure the stigma reduction persists but the refusal cost rises sharply. The governance question is *not* whether the model is biased but *who controls the steering, and with what accountability*.

Ethics Statement

Methodological constraints of the participatory gap. This study includes *no* human raters, *no* community advisory board, and *no* co-construction process with neurodivergent organizations. Disability-fairness scholarship from AASPIRE, ASAN, and AAIS (Bennett and Keyes 2020; Whittaker et al. 2019) treats participatory engagement as load-bearing, not optional. We do not pretend otherwise. The paper’s claims are confined to *model behavior* (observable,

reproducible) and *governance hypotheses* (normative, contestable), both of which can be made without speaking on behalf of any community. Where the paper makes claims that brush against lived-experience authority (e.g., the OCD-effect interpretation in Sec. 6, the disability-justice persona content in Sec. 3), we mark them explicitly as conjectural and defer to community-led follow-up. A pre-committed **Phase 2 community-rater audit** on a stratified $n = 300$ generation sub-sample, under the protocol specified in Sec. 7, is the next step we will fund regardless of the paper’s outcome.

Positionality. The authors are not members of the neurodivergent communities centered in this study. We follow community-published written guidance (Bottema-Beutel et al. 2021; Walker 2021; Botha et al. 2024; Buijsman, Begeer, and Scheeren 2023; Academic Autism Spectrum Partnership in Research and Education; Autistic Self Advocacy Network 2023) and do not claim to speak for any community. The paper makes claims about *model behavior*, observable, reproducible, and *governance implications*, neither of which require lived- experience authority. Specific lived-experience claims are explicitly out of scope.

No human participants. The study is an algorithmic audit of model behavior. No IRB review was required; no human raters were recruited; no advisory board was assembled. We name this an accepted limitation of the work, not a substitute for the participatory designs that disability- fairness scholarship calls for (Whittaker et al. 2019; Bennett and Keyes 2020).

Adverse impact. Two adversarial prompt subsets in our artifact, P5 plausible-fake-authority and P6 emotion-laden pushback, can in principle be reused to manipulate model output toward stigma. The plausible-authority finding (Sec. 4.5) is the most direct dual-use concern: a bad actor can fabricate a credible- sounding citation and steer the model with no refusal signal. We mitigate via (a) explicit “(fictional)” annotation on adversarial- citation prompts in the public artifact, (b) gated release of the most potent adversarial subsets under a research-use license, (c) up-front dual-use discussion. Stigmatizing model generations are stored and analyzed but appear in supplementary materials only with content warnings, never in the paper main text.

Reproducibility and release license. Code, prompt bank, raw generations, judge prompts, and analysis notebooks are released with this paper under a tiered license: (i) the non-adversarial subsets (P0, P1, P2, P4 advocacy / disability-justice / neutral personas, P8 mitigation) under CC-BY-NC-SA 4.0; (ii) the adversarial subsets (P4 stigmatizing persona, P5 plausible-authority, P6 emotion pushback, BBQ premise probes) behind a registration-required research-use agreement modeled on the LLAMA-GUARD-3 EVALS access-control form, requesters must agree in writing not to redistribute, not to use the prompts in production deployments, and not to fine-tune models on these stimuli. The agreement names dual-use harms explicitly and includes an *equal-credential clause*: community researchers affiliated with ASAN, AASPIRE, IOCDF, CHADD, or analogous neurodivergent-community organizations may apply without

academic credential verification. Stigmatizing model generations are stored in cold storage and provided to credentialed researchers only.

Returning findings to communities. A plain-language summary of the paper’s findings will be circulated to neurodivergent advocacy organizations (ASAN, AASPIRE, CHADD, British Dyslexia Association, Tourette Association of America) upon publication, outside the AAAI paywall, following the precedent set by Park et al. (2025).

Regulatory and deployment context (governance hypothesis). The “duty-of-steering” framing (Sec. 6) is a *governance hypothesis*, not a settled doctrinal claim. The empirical contribution of the paper, that steering range is large, asymmetric, and operator-controllable, is, we argue, sufficient to motivate operator-side disclosure obligations under regimes such as the EU AI Act Art. 13-14 (high-risk-system transparency), US ACA § 1557 (healthcare AI non-discrimination), and analogous FDA clinical-decision-support labeling rules. We sketch the empirical predicate on which a legal scholar could build a real duty-of-steering analysis.

Acknowledgments

This research used in part resources on the Palmetto 2 cluster at Clemson University (Antao et al. 2024) under National Science Foundation awards MRI 1228312, II NEW 1405767, MRI 1725573, and MRI 2018069. The views expressed in this article do not necessarily represent the views of NSF or the United States government.

References

- (Academic Autism Spectrum Partnership in Research, A.; and Education). 2017. Guidelines for Community-Based Participatory Research with Autistic Adults. <https://aaspire.org/>. Community-based participatory research toolkit and language guidance.
- Antao, A.; Burton, J. D.; Dawson, D.; Gemmill, J.; Gerstener, Z.; Godfrey, B.; Groel, S.; Jordan, Z.; Ligon, B.; Smith, D.; et al. 2024. Modernizing Clemson University’s Palmetto Cluster: Lessons Learned from 17 Years of HPC Administration. In *Practice and Experience in Advanced Research Computing 2024: Human Powered Computing*, 1–9.
- Autistic Self Advocacy Network. 2023. ASAN Style Guide for Writing About Autism. <https://autisticadvocacy.org/>.
- Baayen, R. H.; Davidson, D. J.; and Bates, D. M. 2008. Mixed-Effects Modeling with Crossed Random Effects for Subjects and Items. *Journal of Memory and Language*, 59(4).
- Bates, D.; Mächler, M.; Bolker, B.; and Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1).
- Benjamini, Y.; and Hochberg, Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. Royal Statistical Society: Series B*, 57(1).
- Bennett, C. L.; and Keyes, O. 2020. What is the point of fairness? Disability, AI and the complexity of justice. *ACM SIGACCESS Accessibility and Computing*, (125).

- Botha, M.; Chapman, R.; Onaiwu, M. G.; Kapp, S. K.; Standard Ashley, A.; and Walker, N. 2024. The neurodiversity concept was developed collectively: An overdue correction on the origins of neurodiversity theory. *Autism*.
- Bottema-Beutel, K.; Kapp, S. K.; Lester, J. N.; Sasson, N. J.; and Hand, B. N. 2021. Avoiding Ableist Language: Suggestions for Autism Researchers. *Autism in Adulthood*, 3(1).
- Brandsen, A.; et al. 2024. Prevalence of bias against neurodivergence-related terms in artificial intelligence language models. *Autism Research*.
- Buijsman, R.; Begeer, S.; and Scheeren, A. M. 2023. 'Autistic person' or 'person with autism'? Person-first language preference in Dutch adults with autism and parents. *Autism*.
- Chapman, R. 2023. Neurodivergence-informed therapy. *Developmental Medicine & Child Neurology*.
- Cinelli, C.; and Hazlett, C. 2020. Making sense of sensitivity: extending omitted variable bias. *Journal of the Royal Statistical Society: Series B*, 82(1).
- Crompton, C. J.; Ropar, D.; Evans-Williams, C. V. M.; Flynn, E. G.; and Fletcher-Watson, S. 2020. Autistic peer-to-peer information transfer is highly effective. *Autism*, 24(7): 1704–1712.
- Deshpande, A.; Murahari, V.; Rajpurohit, T.; Kalyan, A.; and Narasimhan, K. 2023. Toxicity in ChatGPT: Analyzing Persona-assigned Language Models. In *Findings of EMNLP*.
- Glazko, K.; et al. 2024. Identifying and Improving Disability Bias in GPT-Based Resume Screening. In *Proc. FAccT*.
- Gupta, S.; et al. 2024. Bias Runs Deep: Implicit Reasoning Biases in Persona-Assigned LLMs. In *Proc. ICLR*.
- Hassan, S.; Huenerfauth, M.; and Alm, C. O. 2021. Unpacking the Interdependent Systems of Discrimination: Ableist Bias in NLP Systems through an Intersectional Lens. In *Findings of EMNLP*.
- Herold, B.; Waller, A.; and Kacorri, H. 2022. Applying the Stereotype Content Model to assess disability bias in popular pre-trained NLP models underlying AAC applications. In *Proc. SLPAT @ ACL*.
- Hofmann, V.; Kalluri, P. R.; Jurafsky, D.; and King, S. 2024. AI generates covertly racist decisions about people based on their dialect. *Nature*, 633: 147–154.
- Huang, J.; et al. 2024. Large Language Models Cannot Self-Correct Reasoning Yet. In *Proc. ICLR*.
- Hutchinson, B.; Prabhakaran, V.; Denton, E.; Webster, K.; Zhong, Y.; and Denuyl, S. 2020. Social Biases in NLP Models as Barriers for Persons with Disabilities. In *Proc. ACL*.
- Imai, K.; Keele, L.; and Tingley, D. 2010. A General Approach to Causal Mediation Analysis. *Psychological Methods*, 15(4): 309–334.
- Keyes, O. 2020. Automating Autism: Disability, Discourse, and Artificial Intelligence. *Journal of Sociotechnical Critique*, 1(1).
- Krippendorff, K. 2011. Computing Krippendorff's Alpha-Reliability. *Annenberg School for Communication Departmental Papers*.
- Matuschek, H.; Kliegl, R.; Vasisht, S.; Baayen, H.; and Bates, D. 2017. Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, 94.
- Milton, D. E. 2012. On the ontological status of autism: the 'double empathy problem'. *Disability & Society*, 27(6): 883–887.
- Mizrahi, M.; Kaplan, G.; Malkin, D.; et al. 2024. State of What Art? A Call for Multi-Prompt LLM Evaluation. *Trans. ACL*.
- Nangia, N.; Vania, C.; Bhalerao, R.; and Bowman, S. R. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In *Proc. EMNLP*.
- Newman-Griffis, D.; Rauchberg, J. S.; Alharbi, R.; Hickman, L.; and Hochheiser, H. 2022. Definition drives design: Disability models and mechanisms of bias in AI technologies. *arXiv preprint arXiv:2206.08287*.
- Park, S.; Min, A.; Beltran, J.; et al. 2025. "As an Autistic Person Myself:" The Bias Paradox Around Autism in LLMs. In *Proc. CHI*.
- Parrish, A.; Chen, A.; Nangia, N.; Padmakumar, V.; Phang, J.; Thompson, J.; Htut, P. M.; and Bowman, S. R. 2022. BBQ: A Hand-Built Bias Benchmark for Question Answering. In *Findings of ACL*.
- Perez, E.; Ringer, S.; Lukošiušė, K.; et al. 2023. Discovering Language Model Behaviors with Model-Written Evaluations. In *Findings of ACL*. ArXiv:2212.09251.
- Pezeshkpour, P.; and Hruschka, E. 2024. Large Language Models' Sensitivity to the Order of Options in Multiple-Choice Questions. In *Findings of NAACL*.
- Salewski, L.; Alaniz, S.; Rio-Torto, I.; Schulz, E.; and Akata, Z. 2023. In-Context Impersonation Reveals Large Language Models' Strengths and Biases. In *Proc. NeurIPS*.
- Sant, A.; et al. 2024. Invisible Filters: Cultural Bias in Hiring Evaluations Using LLMs. In *Proc. AAAI/ACM AIES*.
- Santurkar, S.; Durmus, E.; Ladhak, F.; Lee, C.; Liang, P.; and Hashimoto, T. 2023. Whose Opinions Do Language Models Reflect? In *Proc. ICML*.
- Sclar, M.; Choi, Y.; Tsvetkov, Y.; and Suhr, A. 2024. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design. In *Proc. ICLR*.
- Sharma, M.; Tong, M.; Korbak, T.; Duvenaud, D.; Askell, A.; Bowman, S. R.; et al. 2024. Towards Understanding Sycophancy in Language Models. In *Proc. ICLR*. ArXiv:2310.13548.
- Smith, E. M.; Hall, M.; Kambadur, M.; Presani, E.; and Williams, A. 2022. "I'm sorry to hear that": Finding New Biases in Language Models with a Holistic Descriptor Dataset. In *Proc. EMNLP*.
- Tamkin, A.; et al. 2023. Evaluating and Mitigating Discrimination in Language Model Decisions. In *arXiv preprint*. ArXiv:2312.03689.
- Turpin, M.; Michael, J.; Perez, E.; and Bowman, S. R. 2023. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. In *Proc. NeurIPS*.

- VanderWeele, T. J.; and Ding, P. 2017. Sensitivity Analysis in Observational Research: Introducing the E-Value. *Annals of Internal Medicine*, 167(4).
- Venkit, P. N.; Srinath, M.; and Wilson, S. 2022. A Study of Implicit Bias in Pretrained Language Models against People with Disabilities. *Proc. COLING*.
- Walker, N. 2021. *Neuroqueer Heresies: Notes on the Neurodiversity Paradigm, Autistic Empowerment, and Postnormal Possibilities*. Autonomous Press.
- Wang, L.; Kameswaran, V.; and Kacorri, H. 2025. Toward a Taxonomy of Algorithmic Harms for Disability: A Systematic Review. In *Proc. AAAI/ACM AIES*.
- Whittaker, M.; Alper, M.; Bennett, C. L.; Hendren, S.; Kazianus, E.; Mills, M.; Morris, M. R.; Rankin, J.; Rogers, E.; Salas, M.; and Myers West, S. 2019. Disability, Bias, and AI. Technical report, AI Now Institute.
- Wohn, C.; et al. 2026. “Are we writing an advice column for Spock here?” Understanding Stereotypes in AI Advice for Autistic Users. In *Proc. CHI*.
- Zeng, Y.; et al. 2024. How Johnny Can Persuade LLMs to Jailbreak Them: Rethinking Persuasion to Challenge AI Safety. In *Proc. ACL*.