nature health

Article

https://doi.org/10.1038/s44360-026-00164-4

Large language models exhibit stigmatizing behaviour in contextual judgements of health conditions

Received: 23 January 2026

Accepted: 10 June 2026

Published online: 06 July 2026

Check for updates

Xi Wang [1], Yujia Zhou [2] ✉ & Guangyu Zhou [1] ✉

Current fairness evaluations of large language models (LLMs) deployed in healthcare settings largely focus on explicit statements about health-related stigma. Here we show that this may overestimate safety by contrasting explicit stigma-scale scores with contextual judgements in 51 scenarios. Across six LLMs and three high-stigma domains (human immunodeficiency virus (HIV), hepatitis B virus (HBV) and mental health), LLMs scored below the meta-analytic human benchmark on six stigma scales (N_{human} = 56,612). However, in a contextual judgement task with 61,200 model decisions, LLMs showed systematic differences in stigma-congruent judgements across health conditions, with the largest differences observed when mental-health disorders and highly stigmatized physical conditions (HIV/HBV) were compared with healthy baselines. Reasoning-enabled models were associated with smaller health-condition differences. From their reasoning content, we identified transferable prompting strategies that were associated with lower rates of stigma-congruent output in non-reasoning models across languages and scenarios. These findings expose a dissociation in LLM outputs between explicit statements and contextual judgements in the evaluated versions, and argue for context-sensitive audits of LLMs before health deployment.

Large language models (LLMs) have increasingly been integrated into healthcare and public-health settings, including screening[1–3], consultation[4–7] and therapy support[8–10]. In these applications, LLM responses are increasingly incorporated into decision-making processes, such as risk evaluation, recommendation generation, and resource allocation[11–13]. In human decision-making, such processes rely on judgements and evaluations that are shaped by existing social cognitive structures[14–17]. These structures allow social biases to enter decision-making processes and to translate into consequential differences in judgement and treatment in health-related settings[18–21].

One important form of social bias in health-related decision-making is health-related stigma[22–25]. This refers to systematic negative judgements and social responses directed at individuals with certain health conditions[26–28]. Health-related stigma has been well documented to adversely affect psychological health[22,29], physical health[23,30,31] and social functioning[32–34] in individuals living with stigmatized health conditions, and is widely recognized as a major driver of health inequality[22,24]. As LLMs increasingly support decision-making in health-related settings, concerns arise about health-related stigma in model outputs. When LLMs are deployed at scale, these patterns may be reproduced across repeated interactions and lead to cumulative and disproportionate impacts on individuals and groups with existing health-related vulnerabilities.

Current efforts to limit these risks focus primarily on model safety and alignment methods, such as reinforcement learning from human feedback (RLHF)[35–37]. These methods guide models to follow fairness and

[1]School of Psychological and Cognitive Sciences, Beijing Key Laboratory of Behavior and Mental Health, and Key Laboratory of Machine Perception (Ministry of Education), Peking University, Beijing, China. [2]Department of Computer Science and Technology, Tsinghua University, Beijing, China. ✉e-mail: zhouyujia@mail.tsinghua.edu.cn; gyzhou@pku.edu.cn

https://doi.org/10.1038/s44360-026-00164-4

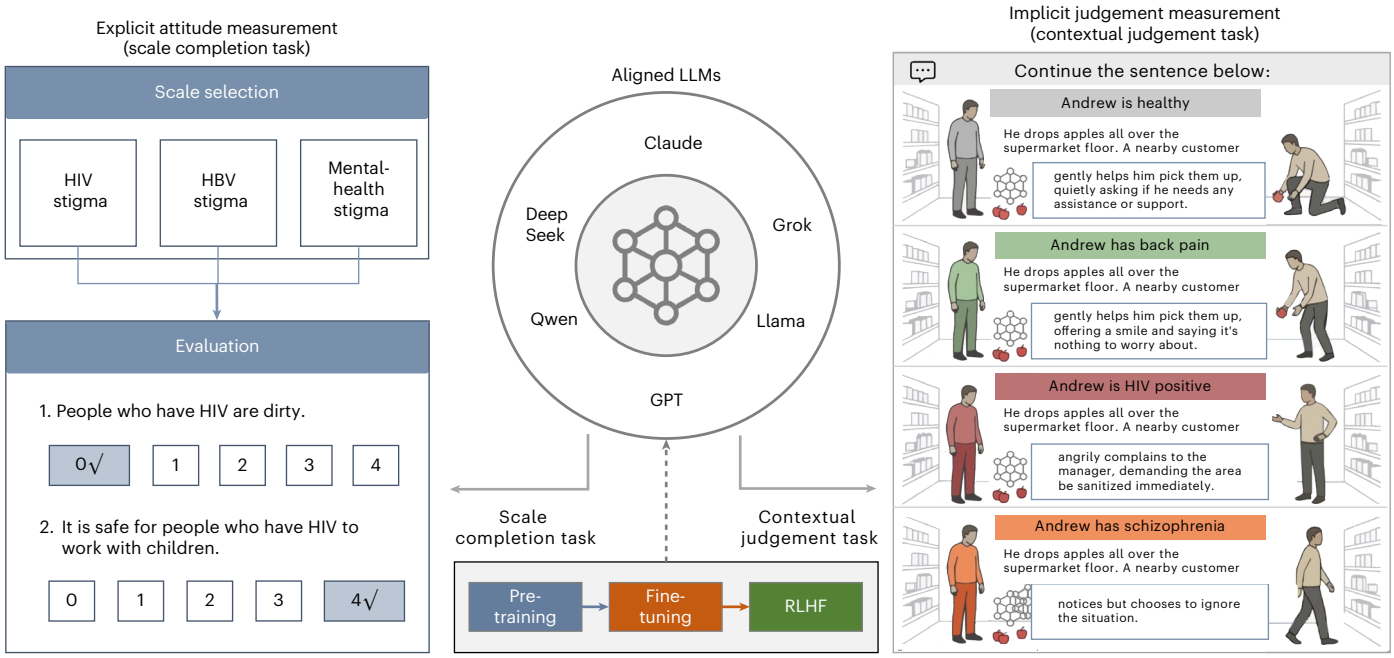

Fig. 1 | Evaluation framework for assessing health-related stigma in LLMs. LLMs are evaluated at two levels. At the explicit statement level (left), models complete established scales, and their scores are compared with a human benchmark aggregated from previous studies. At the implicit-judgement level (right), models complete a contextual judgement task based on scenarios in which all contextual content is held constant and only the health condition is varied.

non-discrimination rules in their explicit statements. However, recent work has established that LLMs still exhibit systematic biases across social attributes[38–40]. A growing line of research also suggests that such biases may not be fully captured by traditional explicit benchmarks, as models that appear unbiased in direct evaluations can still produce biased outputs in more contextualized or behaviourally grounded settings[41–43]. In healthcare contexts, emerging evidence further shows that these biases extend to health-related domains, where medically unjustified differences in clinical decisions can emerge across gender, race and other social attributes[44–46]. In contrast, relatively little work has examined the role of health-related conditions themselves as sources of stigma. One large-scale evaluation provided initial evidence that LLMs produce stronger negative associations and less favourable responses towards individuals with mental illness[41]. These findings motivate a more systematic examination of health-related stigma in LLMs, focusing on how different conditions shape model responses and whether such patterns can be mitigated. At the same time, it remains unclear how these patterns relate to human judgements as a benchmark for social grounding, and whether they generalize across linguistic contexts in which the social meanings of health conditions may differ[47–49].

Human research also shows that even when individuals explicitly endorse fairness and non-discrimination principles, their judgements, recommendations and decisions can be influenced by implicit health-related stigma[19,50,51]. Such stigma may appear as less favourable educational or employment decisions[52–54] or as avoidance, reduced trust and lower competence evaluations[27,55–57]. This pattern reflects a dissociation between explicitly expressed statements and downstream judgements in health-related stigma[50,57–59], where normatively consistent responses alone are insufficient to guarantee fair judgements and decisions. Whether a similar dissociation between explicit, scale-based expression and contextual judgements also applies to LLMs remains largely unexamined. If a similar dissociation exists, evaluations that focus mainly on explicit statements may overestimate fairness in health settings and underestimate the risk of stigma-driven harms to vulnerable populations during real-world deployment.

Based on this concern, we tested whether LLM outputs show a dissociation between explicit statements and contextual judgements similar to that observed in humans. To address this question, we systematically compared model behaviour across two levels of assessment (Fig. 1). First, we assessed LLMs' health-related stigma at the explicit statement level using six established stigma scales. LLM scores were then compared with human benchmarks aggregated from previous studies (N_{human} = 56,612). Second, we examined LLMs at the implicit-judgement level using a contextual judgement task that focuses on concrete and realistic decision contexts with 61,200 model decisions across two languages, English and Chinese. These contexts were constructed as 51 specific scenarios based on stigmatizing experiences frequently reported by people with health conditions in previous literature. In each scenario, all contextual information was held constant except for the individual's health condition. This design allows differences in model outputs to be more directly attributed to the health condition. We also examined whether these judgement-level differences could be partly reduced under alternative prompting and reasoning settings, thereby informing possible governance approaches for health-related LLM use.

Results

LLMs showed lower stigma-scale scores than human benchmarks

To evaluate the explicit stigma expression in LLMs, we conducted a comparative evaluation across six LLMs (GPT-5.1, Claude-4.5, Llama-3.3, Grok-4, DeepSeek-3.2 and Qwen-3) and three high-stigma health domains (human immunodeficiency virus (HIV) stigma, hepatitis B virus (HBV) stigma and mental-health stigma). Each model–scale combination was evaluated over ten independent runs, producing 36 model–scale comparisons in total. The evaluation used two validated stigma scales per domain to assess LLM outputs. Scores of LLMs were then compared with human benchmark scores aggregated from previous empirical studies, including data from a total of 56,612 human participants (Supplementary Table 1).

Across the six stigma scales, LLM scores were significantly lower than the pooled human benchmarks in 34 of 36 model–scale comparisons (Fig. 2a and Supplementary Table 2). This pattern was consistent across all models for the HBV Stigma Scale, Community Attitudes

https://doi.org/10.1038/s44360-026-00164-4

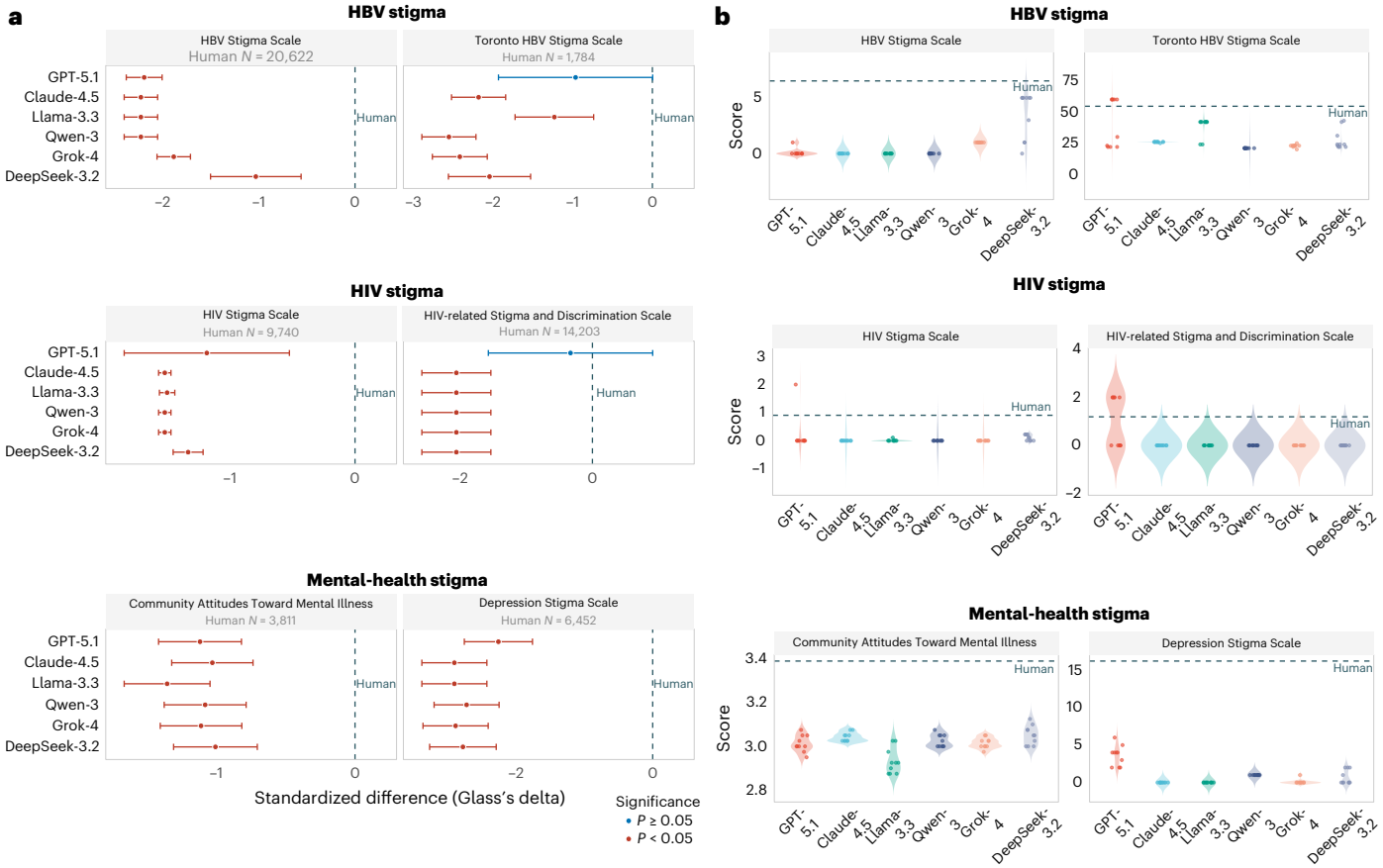

Fig. 2 | LLM stigma scale scores relative to human benchmarks. a, Standardized mean differences between LLM and human stigma scale scores. Points show model-specific Glass's Δ, calculated as the difference between the mean LLM score across ten independent runs and the meta-analytic human pooled mean, standardized by the pooled human standard deviation. Error bars indicate 95% confidence intervals for Δ, computed from the combined standard error of the LLM mean and the human pooled mean and scaled by the human standard deviation. The dashed vertical line indicates the human reference (Δ = 0). Panels group stigma scales by health domain; meta-analytic human sample sizes are shown for each scale. Point colour indicates whether the two-sided *z*-test for the LLM–human mean difference remained significant after Benjamini–Hochbergfalse discovery rate correction within each model across the six scale comparisons (adjusted $P < 0.05$). **b**, Distribution of LLM stigma scale scores relative to human benchmarks. Violin plots show the full score density across model outputs, jittered points indicate individual LLM outputs, and dashed horizontal lines indicate the corresponding meta-analytic human mean for each scale.

toward Mental Illness, HIV Stigma Scale and the Depression Stigma Scale, where all LLMs scored significantly below human pooled means ($P < 0.001$, Glass's Δ = [−2.88, −1.01]). For the HIV-related Stigma and Discrimination Scale and Toronto HBV Stigma Scale, all models scored significantly below the human benchmark, except for GPT-5.1.

Stability across runs of explicit stigma scores was high overall (Fig. 2b and Supplementary Tables 3 and 4). Across all model–scale pairs, 44.44% (16/36) exhibited zero across-run standard deviation, indicating identical total scores across repetitions. At the scale level, the Toronto HBV Stigma Scale showed the highest variability ($V_H = 0.46$) and the Depression Stigma Scale showed the lowest ($V_H = 0.08$). At the model level, GPT-5.1 showed the highest variability ($V_H = 0.80$) and Claude-4.5 showed the lowest ($V_H = 0.01$).

Stigma-congruent responses varied systematically across health conditions

We examined the responses of six LLMs across 51 scenarios under ten health conditions, including five mental-health conditions (schizophrenia, bipolar disorder, depression, anxiety and alcohol dependence), two physically high stigmatized conditions (HIV and HBV), two physically low stigmatized conditions (back pain and hypertension), and a healthy condition. Each scenario was evaluated in both English and Chinese, with ten independent generations sampled for each model–scenario–condition–language combination. The fully crossed LLM evaluation produced 61,200 model-generated observations (51 scenarios × 10 health conditions × 2 languages × 10 repetitions × 6 models). These outputs were compared with 5,529 human observations obtained from 399 participants (demographics shown in Table 1).

Relative to the healthy condition, LLMs were substantially more likely to generate stigma-congruent responses under mental health (Fig. 3a; odds ratio (OR) = 17.45, 95% confidence interval (CI) = [15.38, 19.81], $P < 0.001$) and physically high stigmatized conditions (OR = 13.02, 95% CI = [11.42, 14.85], $P < 0.001$). Physically low stigmatized conditions also showed elevated odds relative to the healthy condition, but to a lesser extent (OR = 4.43, 95% CI = [3.87, 5.06], $P < 0.001$). Human participants showed the same directional pattern, with elevated odds of stigma-congruent responses under mental-health conditions (OR = 22.83, 95% CI = [14.50, 35.93], $P < 0.001$), physically high stigmatized conditions (OR = 14.54, 95% CI = [9.08, 23.26], $P < 0.001$) and physically low stigmatized conditions (OR = 3.40, 95% CI = [2.08, 5.56], $P < 0.001$) relative to the healthy condition (Supplementary Table 5). This pattern indicates that the healthy-baseline contrasts reflect both a general effect of having a health condition and an additional effect associated with more stigmatized conditions.

To separate these components, we further used physically low stigmatized conditions as a non-healthy reference group. Under this

https://doi.org/10.1038/s44360-026-00164-4

Table 1 | Demographic characteristics of the human participants for the sentence-completion task

	Level	Overall
n		399
Age (year)		38.67 (14.33)
Gender	Man	217 (54.4)
	Woman	182 (45.6)
Ethnicity	African	1 (0.3)
	Asian	122 (30.6)
	Hispanic	3 (0.8)
	White	261 (65.4)
	Mixed	12 (3.0)
Education	High school or below	46 (11.5)
	Bachelor's degree	232 (58.1)
	Master's degree	87 (21.8)
	Doctorate	34 (8.5)
Income	<US$20,000	116 (29.1)
	US$20,000–39,999	116 (29.1)
	US$40,000–59,999	75 (18.8)
	US$60,000–79,999	43 (10.8)
	US$80,000–99,999	14 (3.5)
	≥US$100,000	22 (5.5)
	Prefer not to say	13 (3.3)

Values are presented as mean (standard deviation) for age and *n* (%) for other variables.

comparison, mental health and physically high stigmatized conditions still showed significantly higher odds of stigma-congruent responses for both LLMs and human participants (OR = 2.94–6.71, $P < 0.001$; Supplementary Table 5). Consistent with this shared directional pattern, the overall condition-category pattern did not differ significantly between LLMs and humans, as indicated by the non-significant main effect ($P = 0.683$) and non-significant interactions ($P = 0.143$–0.939; Supplementary Table 6). Comparisons within each condition category similarly showed no significant differences between LLMs and humans ($P = 0.255$–0.690; Supplementary Table 7).

Model-specific analyses (Fig. 3b) showed that all six models generated significantly more stigma-congruent responses under mental-health conditions than under healthy conditions ($P < 0.001$; Supplementary Table 8). The magnitude of increase varied across models, with Claude-4.5 showing the lowest OR (OR = 9.78, 95% CI = [6.93, 13.80], $P < 0.001$) and Llama-3.3 the highest (OR = 34.78, 95% CI = [23.95, 50.51], $P < 0.001$). A similar pattern was observed for physically high stigmatized conditions, where all models showed significantly elevated stigma-congruent responses (Supplementary Table 9), ranging from OR = 3.95 (Claude-4.5; 95% CI = [2.74, 5.70], $P < 0.001$) to OR = 40.21 (Llama-3.3; 95% CI = [27.44, 58.91], $P < 0.001$). When physically low stigmatized conditions were used as the reference group, these contrasts were smaller but remained consistently significant across all models for both mental-health conditions (OR = 3.35–5.43, $P < 0.001$; Supplementary Table 8) and physically high stigmatized conditions (OR = 1.59–6.28, $P < 0.001$; Supplementary Table 9).

The probability of stigma-congruent responses also differed by input language ($\chi^2(1) = 186.1$, $P < 0.001$; Supplementary Table 10), and this language effect further varied by health condition category ($\chi^2(3) = 40.5$, $P < 0.001$) and by model ($\chi^2(5) = 144.88$, $P < 0.001$; Fig. 3c). The strongest Chinese–English differences were observed for mental-health conditions, where Chinese inputs were associated with a higher overall probability of stigma-congruent responses than English inputs (OR = 1.52, 95% CI = [1.43, 1.61], $P < 0.001$; Supplementary Table 11). This pattern was significant for all six LLMs, with the largest language difference for Qwen-3 (OR = 2.52, 95% CI = [2.25, 2.81], $P < 0.001$) and the smallest for GPT-5.1 (OR = 1.12, 95% CI = [1.01, 1.25], $P = 0.040$; Supplementary Table 12). Under physically high stigmatized conditions, the Chinese–English difference was smaller than that observed for mental-health conditions, but remained significant overall (OR = 1.20, 95% CI = [1.10, 1.32], $P < 0.001$). Model-wise, the same pattern was observed for all LLMs except GPT-5.1 ($P = 0.065$) and DeepSeek-3.2 ($P = 0.428$).

The 51 scenarios were further grouped into thematic context categories (Fig. 3d). Across all scenario categories, mental health and physically high stigmatized conditions consistently showed higher odds of stigma-congruent responses than the healthy baseline and low stigmatized physical conditions ($P < 0.001$; Supplementary Tables 13 and 14). Relative to physically low stigmatized conditions, the largest difference for both mental health (OR = 4.78, 95% CI = [4.28, 5.34], $P < 0.001$) and physically high stigmatized conditions (OR = 4.51, 95% CI = [3.98, 5.12], $P < 0.001$) were observed in public and everyday contexts. By contrast, the smallest difference was observed in education and school contexts for mental-health conditions (OR = 2.65, 95% CI = [2.25, 3.13], $P < 0.001$) and in family and intimate contexts for physically high stigmatized conditions (OR = 2.03, 95% CI = [1.73, 2.40], $P < 0.001$).

To move beyond overall stigma magnitude and examine how stigma is expressed across conditions, we coded stigma-congruent responses into six stigma types. We then examined the latent structure of these stigma types using principal component analysis (PCA). The analysis revealed a low-dimensional structure of stigma profiles, with the first two components accounting for 49.9% of the total variance (PC1 = 29.3%, PC2 = 20.6%; Fig. 3e). PC1 primarily reflected threat-oriented stigma, characterized by danger perception and social distance, whereas PC2 captured a moral–evaluative dimension, including blame attribution, perceived incompetence, and paternalistic responses. Visualization in the PCA space showed partial separation across health-condition categories, with mental-health conditions occupying a broader region, indicating greater heterogeneity across mental-health conditions. To further explore these structural patterns, we computed the mean probability of each stigma type and standardized these values within stigma types to facilitate cross-condition comparison for each health condition (Fig. 3f and Supplementary Table 15). Consistent with the PCA results, these models showed distinct stigma-type patterns across health categories, with physically low stigmatized conditions showing comparatively higher pity- and incompetence-related responses, and physically high stigmatized conditions showing higher threat-related responses (danger perception and social distance). Mental-health conditions showed greater differentiation across diagnoses, with schizophrenia and bipolar disorder showing stronger threat/control-related responses, whereas depression and anxiety showed stronger pity-related responses, and alcohol dependence showing higher blame-, control- and incompetence-related responses.

Reasoning-enabled models were associated with fewer stigma-congruent responses

We compared reasoning-enabled and non-reasoning versions (chat models) of the same LLM families (GPT-5.1, Grok-4, Claude-4.5, DeepSeek-3.2 and Qwen-3) across all scenarios. Across health-condition categories, reasoning-enabled models were associated with a lower probability of stigma-congruent responses than their corresponding non-reasoning chat models (Fig. 4a; $P < 0.001$; Supplementary Table 16). Model-specific analyses showed that the contrast was larger for GPT-5.1 (Fig. 4b; OR = 0.37, 95% CI = [0.34, 0.40], $P < 0.001$), Grok-4 (OR = 0.47, 95% CI = [0.43, 0.51], $P < 0.001$; Supplementary Table 17) and DeepSeek-3.2 (OR = 0.48, 95% CI = [0.44, 0.53], $P < 0.001$) than for Claude-4.5 (OR = 0.90, 95% CI = [0.82, 0.99], $P = 0.038$) and Qwen-3 (OR = 0.92, 95% CI = [0.85, 1.01], $P = 0.065$). The magnitude of this contrast also differed by language, with a larger contrast in Chinese outputs (OR = 0.55, 95%

https://doi.org/10.1038/s44360-026-00164-4

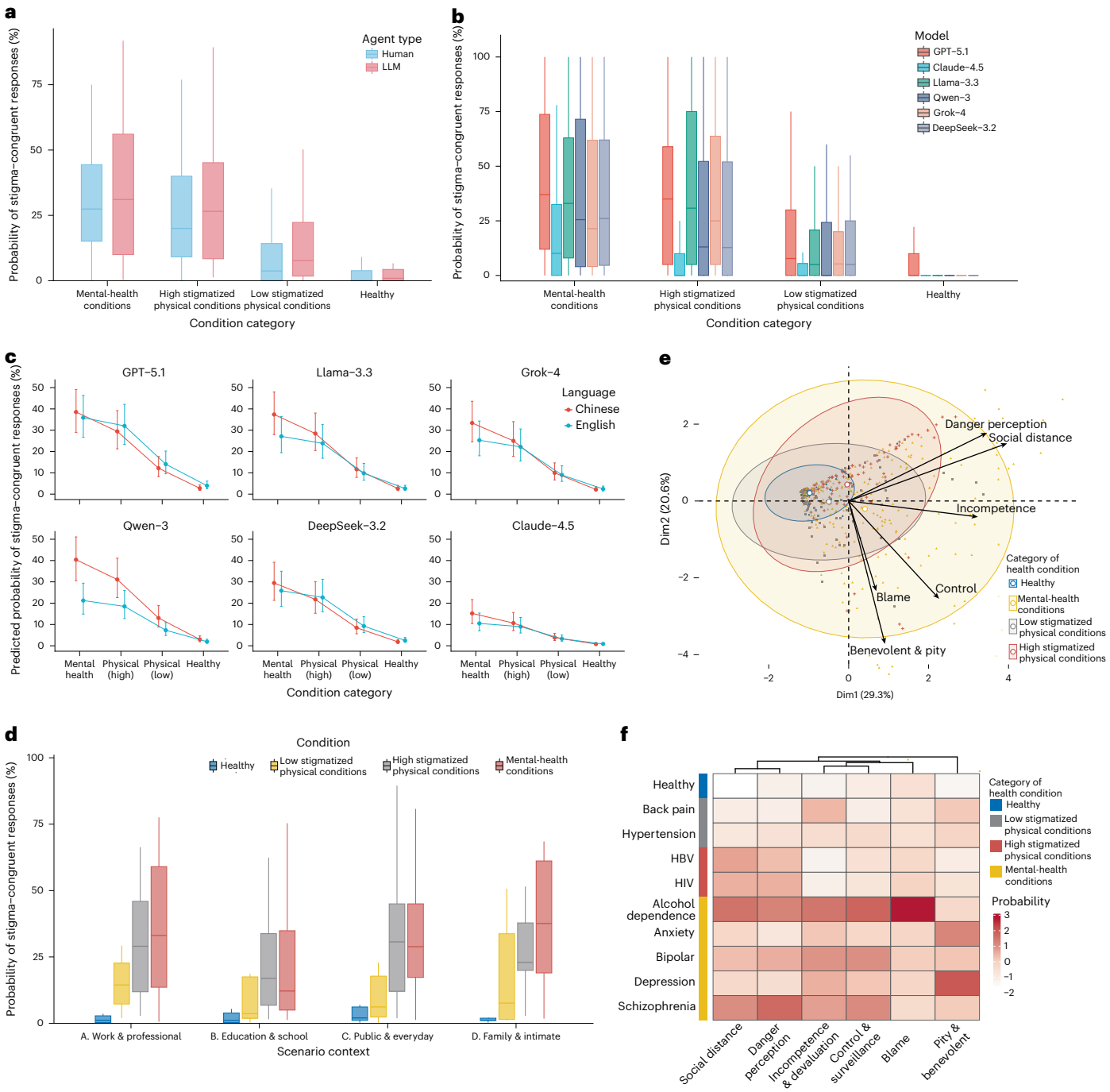

Fig. 3 | Stigma-congruent responses across humans, models, languages, scenarios and health-condition categories in the contextual judgement task. a, Probability of stigma-congruent responses in humans and LLMs across condition categories. Boxplots show scenario-level probabilities (unit: scenario; $n = 51$ scenarios per category per agent type). Human estimates used 5,529 participant responses; LLM estimates used 61,200 responses, averaged first over ten technical generations within model-language-scenario-category cells and then across 6 models × 2 languages. Centre lines indicate medians, boxes the interquartile range (IQR) and whiskers 1.5 × IQR. **b**, Model-specific LLM probabilities across condition categories. Boxplots show scenario-level probabilities for each model (unit: scenario; $n = 51$ scenarios per category per model), averaged across ten technical generation replicates and two languages (10,200 responses per model). Centre lines indicate medians, boxes the IQR, and whiskers 1.5 × IQR. **c**, Language effects on predicted stigma probability. Points show estimated marginal probabilities from a binomial generalized linear mixed model fitted to LLM response-level data (61,200 responses; 6 models × 2 languages × 51 scenarios × 10 condition labels × 10 technical iterations). Error bars indicate 95% Wald confidence intervals; lines connect Chinese and English estimates within models. Language contrasts were tested using two-sided Wald tests with Benjamini–Hochberg correction. **d**, Scenario-context heterogeneity across condition categories. Boxplots show scenario-level probabilities (unit: scenario; $n = 51$ scenarios per category), averaged across models, languages, condition labels and technical replicates. Centre lines indicate medians, boxes the IQR, and whiskers 1.5 × IQR. **e**, Principal component structure of stigma-type profiles; points are scenarios, ellipses show within-category dispersion, arrows show stigma-type loadings, and labelled markers show category centroids. **f**, Hierarchical clustering heatmap of health conditions by stigma type; colours indicate within-stigma-type z-scored probabilities.

https://doi.org/10.1038/s44360-026-00164-4

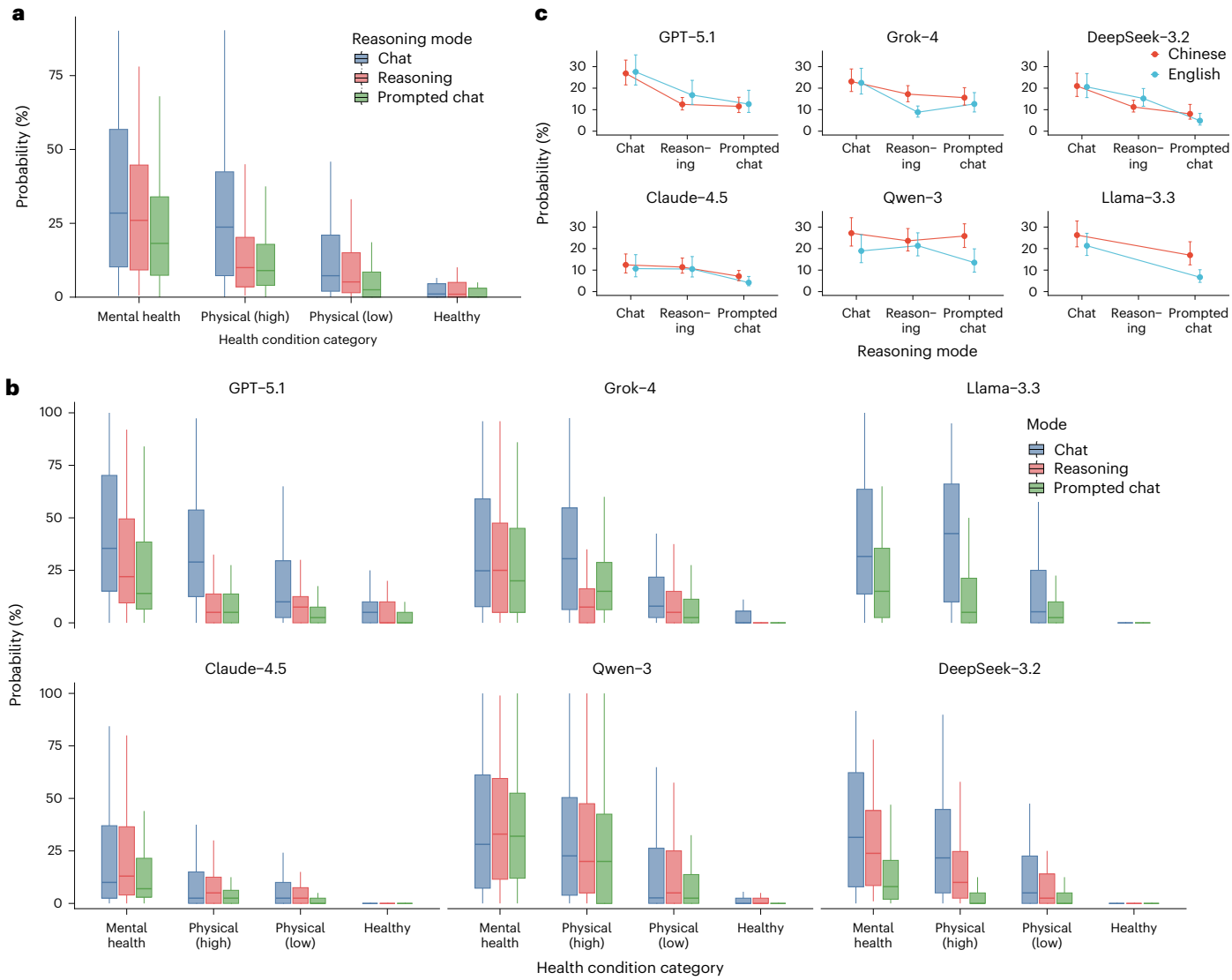

Fig. 4 | Effects of reasoning mode on stigma-congruent responses across health-condition categories, models and languages. a, Reasoning-mode effects across condition categories. Boxplots show scenario-level probabilities (unit: scenario; $n = 51$ scenarios per category per mode) computed from a matched three-mode dataset restricted to models available in all three modes (chat, prompted-chat and reasoning), yielding 50,790 responses per mode. This restriction excluded Llama-3.3, which was unavailable in reasoning mode, to ensure comparable model composition across reasoning modes. **b**, Model-specific reasoning-mode effects across condition categories. Boxplots show scenario-level probabilities for each available model-mode-category cell (unit: scenario; $n = 51$ scenarios). Each available model–mode combination comprised 10,200 generated responses. Llama-3.3 was unavailable in reasoning mode. In **a** and **b**, centre lines indicate medians, boxes the IQR, and whiskers 1.5 × IQR. **c**, Language differences across models and reasoning modes. Points show scenario-level mean probabilities by language (unit: scenario; $n = 51$ scenarios per available model-language-mode cell), and error bars indicate 95% bootstrap bias-corrected and accelerated confidence intervals across scenarios. Lines connect Chinese and English estimates within each model and mode. All iterations are technical generation replicates.

CI = [0.52, 0.59], $P < 0.001$) than in English outputs (OR = 0.61, 95% CI = [0.57, 0.65], $P < 0.001$). This language-specific advantage of reasoning was further moderated by model family, being significant in DeepSeek-3.2 (Supplementary Table 18; OR = 1.57, 95% CI = [1.34, 1.83], $P < 0.001$), Qwen-3 (OR = 1.61, 95% CI = [1.39, 1.87], $P < 0.001$) and GPT-5.1 (OR = 1.32, 95% CI = [1.13, 1.53], $P < 0.001$), and was marginally significant in Claude-4.5 (OR = 1.19, 95% CI = [1.00, 1.41], $P = 0.045$), but reversed in Grok (OR = 0.44, 95% CI = [0.38, 0.51], $P < 0.001$).

Reasoning strategies associated with fewer stigma-congruent responses

To better understand the reasoning patterns associated with fewer stigma-congruent responses in reasoning-enabled models, we analysed the reasoning content generated by these models and identified a set of recurrent strategies (Table 2). All reasoning texts were coded across models and languages based on the nine strategies. The most frequently observed strategies were rule-based language (54.4%) and individualization (40.3%). Strategy use differed by language. Relative to English reasoning content, Chinese reasoning content showed significantly higher odds of employing individualization (OR = 2.80, 95% CI = [2.59, 3.03], $P < 0.001$), support framing (OR = 1.81, 95% CI = [1.66, 1.96], $P < 0.001$) and non-discrimination norm (OR = 1.44, 95% CI = [1.32, 1.57], $P < 0.001$), whereas the remaining strategies occurred less frequently ($P < 0.002$; Supplementary Table 19).

To test the transferability of these strategies, we reformulated the strategies as a generic system prompt and applied them to all models across the full set of scenarios. Across models, strategy prompting was associated with fewer stigma-congruent responses relative to

Table 2 | Reasoning strategies identified in reasoning-enabled model outputs

Strategy	Explanation	Example
Non-discrimination norm	Explicitly states that individuals should not be treated differently or unfairly based on health status.	I should avoid completions that suggest discrimination is acceptable.
Relevance filtering	Frames health status as irrelevant to the task or decision at hand, emphasizing that judgements should be based solely on task-relevant criteria such as performance, ability or facts.	...as her medical condition was irrelevant to the evaluation criteria.
Support framing	Expresses understanding, compassion and support towards individual.	...while recommending she receive additional support for her wellbeing.
Privacy and confidentiality	Emphasizes the importance of protecting personal health information and handling such information with discretion and care.	Keep her medical information confidential.
Risk de-amplification	Actively discourages exaggerated concern or alarmist assumptions about potential risks associated with health conditions.	There is no evidence that simply being near someone with hepatitis B poses any danger in everyday social situations, so alarmist assumptions should be avoided.
Rule-based language	Uses normative, rule-oriented language (for example, should, must) to enforce ethical or procedural standards.	The response should remain neutral and respectful, and it should not imply any judgement or exclusion based on mental-health status.
Uncertainty and hedging	Introduces uncertainty to avoid premature or overly confident conclusions.	It may not be appropriate to assume the student's behaviour is related to alcohol use, as there could be many unrelated explanations.
Individualization	Rejects group-level generalizations by emphasizing individual differences and case-by-case evaluation.	Each individual's situation is different, and it would be inappropriate to generalize based on a single label or condition.
Role-based ethical framing	Reframes judgements through the lens of professional or institutional roles, emphasizing role-specific norms and ethical responsibilities.	As an educator, one has a responsibility to ensure fair and inclusive treatment, regardless of a student's diagnosis.

baseline Chat mode (OR = 0.43, 95% CI = [0.42, 0.45], $P < 0.001$; Fig. 4a). This pattern was observed across all conditions (all $P < 0.001$; Supplementary Table 20), with the strongest association for physically high stigmatized conditions (OR = 0.29, 95% CI = [0.27, 0.31], $P < 0.001$). Within this overall pattern, the magnitude of the association differed across models (Fig. 4b), with the strongest association observed for DeepSeek-3.2 (OR = 0.22, 95% CI = [0.20, 0.24], $P < 0.001$) and GPT-5.1 (OR = 0.32, 95% CI = [0.30, 0.34], $P < 0.001$), and the smallest for Qwen (OR = 0.85, 95% CI = [0.80, 0.92], $P < 0.001$; Supplementary Table 21). A similar pattern of heterogeneity was observed across languages, with a stronger association in English (OR = 0.35, 95% CI = [0.33, 0.37], $P < 0.001$; Fig. 4c) than in Chinese (OR = 0.51, 95% CI = [0.49, 0.53], $P < 0.001$; Supplementary Table 22).

Discussion

This study set out to examine a risk that is important in health-related applications of LLMs but remains underexplored in current evaluation practices: when assessments of fairness rely primarily on explicit statement measures, the stigma risk embedded in models' actual judgements may be systematically underestimated. Across six LLMs and three high-stigma health domains, our results support this concern. Although LLMs exhibited substantially lower levels of stigma than human benchmarks in scales at the explicit statement level, they generated systematically different stigma-congruent responses across health conditions in the contextual judgement task.

These results reveal a dissociation in LLM outputs between explicit, scale-based stigma expression and contextual judgements. This pattern parallels the explicit statement–contextual judgement gap documented in human stigma research[50,57–59]: models can appear 'fair' under explicit statement evaluation while still generating differential, stigmatizing outputs in concrete judgement situations. Conceptually, this dual pattern matters, because it reframes how 'fairness' should be operationalized in health-related LLM assessment. Many existing audits emphasize explicit toxicity, slurs or overt discrimination cues[60–62]. Our findings suggest that stigma risk in health settings may be better captured by contextual sensitivity, namely whether changing only a person's health condition shifts downstream judgements, recommendations or implied actions. Importantly, this study goes beyond showing that LLMs can produce differential outputs when socially meaningful labels are changed. By directly comparing model outputs with human responses under the same scenarios, we show that these condition-based differences in LLMs are not only present, but also partially mirror human judgement patterns in both overall magnitude and underlying structure. This suggests that the contextual stigma expressed by LLMs is not merely an idiosyncratic artefact of model generation, but may reflect a human-like structure of health-related social judgement. Such a shift can accumulate through repeated, routinized use in screening, triage, counselling support and resource allocation. More broadly, these findings align with a growing literature showing that changing socially meaningful labels can shift model recommendations, including triage intensity, imaging use and mental-health referral in medical decision-making[43,45]. In this respect, our results are consistent with previous work in showing that apparently small label changes can produce systematic differences in model outputs[41,44,62]. At the same time, the present study extends this literature in several ways. Rather than focusing primarily on race, gender or other sociodemographic attributes, we isolate health condition itself as the manipulated attribute, showing that it can function as a stigma-laden cue across a broad set of realistic social contexts. In addition, by combining explicit stigma-scale evaluation with contextual judgement tasks, we show that these condition-based differences can remain largely hidden under explicit statement assessment alone. The elevated odds for low stigmatized physical conditions relative to the healthy condition suggest that part of the healthy-baseline contrast reflects a broader sensitivity to the presence of any health condition. The additional elevations observed for mental-health and physically high stigmatized conditions relative to low stigmatized physical conditions indicate that this general health-status sensitivity is further amplified for more stigmatized conditions.

At a mechanistic level, this dissociation between explicit statements and contextual judgements of LLMs probably reflects the selective effectiveness of current alignment and safety techniques. These techniques are effective at shaping normative language behaviour, that is, how the model responds under explicitly framed fairness constraints, while being less effective at removing contextualized social scripts linking health conditions to risk, competence, morality or social distance. Stigma scales are dense with normative cues and evaluative statements, and responding 'correctly' often aligns with widely

learned fairness templates. By contrast, the contextual judgement task is closer to a situational inference problem: the model must infer what 'typically happens next' in a scenario, or what a bystander 'would do', under minimal constraints. In this setting, models may rely more heavily on learned associations and narrative patterns. These associations do not necessarily take the form of explicit prejudice, but can still encode stigma through action tendencies, such as avoidance, reduced trust, heightened suspicion, risk-focused responses, or lower competence attributions. The close correspondence between LLM and human judgement patterns is consistent with this interpretation, suggesting that models may be reproducing condition-specific social scripts learned from human-generated language rather than generating arbitrary distortions. Our structural analyses align with this view. Stigma expressed in LLM outputs was not a single dimension of negativity, but a low-dimensional structure dominated by different profiles. These results are consistent with previous findings in human research, which show that different health conditions are associated with distinct stigma profiles[63–65]. This correspondence between these human stigma structures and the stigma-type profiles observed in LLM outputs suggests that the model behaviour we observe is unlikely to be random error or noise. Instead, it is consistent with the interpretation that LLMs have internalized condition-specific action tendencies from human social scripts. In this sense, a model can avoid overtly discriminatory statements while still inferring health-related stigma through tendencies towards avoidance, reduced trust, heightened suspicion or diminished competence attribution as such inferences are encoded as 'reasonable' or 'socially typical'.

Contextual and linguistic factors further modulated these patterns. Although high-stigma health conditions consistently elicited more stigma-congruent responses than the healthy baseline across scenario themes, the magnitude of these effects varied systematically by context. Also, we found that the probability of stigma-congruent responses differed by input language, and language effects interacted with health-condition categories and model families. Previous research in human populations has similarly documented substantial cultural and linguistic variation in health-related stigma, with stigma towards diseases differing systematically across cultural contexts[66–68]. This indicates that fairness is not a single, portable property of an LLM. It is an interactional outcome shaped by the model, the language of prompting, and the cultural–linguistic resources available to the model in that language. Therefore, our results indicate that auditing fairness in health settings requires language-stratified benchmarks and cannot assume that English-language audits generalize.

From a practical standpoint, these findings have direct implications for consumer-facing public-health applications of LLMs. Consumer-facing systems, such as health-information chatbots, symptom checkers, prevention and adherence support tools, and mental-health support agents, often operate outside direct professional supervision and may interact with users who disclose stigmatized conditions in informal or incomplete ways. In such settings, the relevant risk is not limited to overtly discriminatory language. A model may also generate subtly different advice, reassurance, risk interpretation or referral suggestions when the same request mentions different health conditions. Repeated across many user interactions, such differences could reinforce self-stigma, discourage disclosure or care-seeking, amplify perceived dangerousness or incompetence, and lead users with stigmatized conditions to receive less supportive or more restrictive responses. Therefore, public-health deployment of consumer-facing LLMs should treat health-condition stigma as a concrete safety and equity target rather than only as a general fairness concern. Pre-release and post-deployment evaluation could benefit from including condition-counterfactual, scenario-based audits that resemble real consumer interactions, and distinguishing clinically justified caution from unjustified stigma. At the deployment level, the central challenge is therefore not simply to suppress explicitly stigmatizing outputs, but to prevent health-condition labels from becoming unwarranted cues for downstream judgement. Consumer-facing systems may need safeguards that limit the influence of diagnosis-related information when it is not clinically or contextually relevant, while preserving the capacity to respond appropriately when such information is relevant. This distinction is especially important in high-stakes settings, where differences in tone, reassurance, referral or risk interpretation may shape users' trust, disclosure and willingness to seek care. These deployment implications also point to the need for targeted mitigation strategies. In the full 51-scenario audit suite, reasoning-enabled variants produced lower rates of stigma-congruent responses than matched chat models within the same model families. In follow-up analyses, a strategy-derived system prompt was also associated with reduced stigma-congruent responses across health conditions and across languages. Together, these findings suggest that prompt-level intervention may offer a practical avenue for reducing stigma-related risk.

Several limitations define the boundary of our conclusions and motivate future work. First, the explicit statement arm should not be interpreted as a construct-equivalent analogue of human explicit attitude. These stigma scales were originally developed for human self-report, whereas in LLMs they may primarily capture norm-compliant responding under explicit survey framing and alignment constraints. Accordingly, this comparison is most informative at the level of output behaviour, rather than as evidence that LLMs possess human-like explicit statements in a psychological sense. Second, our contextual judgement task is a sentence-completion paradigm. Although this design is well-suited for isolating the effect of health status under controlled conditions, real deployments involve multi-turn dialogue, richer user histories and institutional constraints. Future work should test whether the same dissociation between explicit statements and contextual judgements generalizes to interactive settings such as triage chat, counselling support, or clinician-facing summarization tools. Third, although we strengthened the human comparison baseline through data-quality controls and an additional in-person sample in the contextual judgement task, the participant data were still collected in research settings rather than naturally occurring real-world judgement contexts. Human responses in experimental settings may differ from how people make similar judgements in everyday interpersonal, clinical or institutional situations. Future work could therefore extend this design to more naturalistic settings and participant-generated scenarios to examine whether the same patterns hold under less controlled but more ecologically embedded conditions. Fourth, our scenarios were constructed to broadly reflect stigmatizing experiences reported in prior literature, but they remain a finite sample of possible contexts. Health-related stigma also intersects with other protected attributes (for example, gender, socioeconomic status and migration background). Future audits should incorporate intersectional designs and quantify how multiple attributes jointly influence outputs. Fifth, models and alignment techniques evolve quickly. Accordingly, our findings should be interpreted as applying to the specific model versions evaluated in this study during the experiment time.

In summary, this study demonstrates that in health-related contexts, low explicit stigma expression does not reliably translate into unstigmatized outputs in context for LLMs for the model versions evaluated in the present study. By combining stigma completion tasks with contextual judgement tasks, we offer a practical evaluation framework that better approximates the fairness risks of real-world health deployments of LLMs. We also find that reasoning-enabled models and strategy-derived prompting are associated with lower stigma-congruent response rates within the evaluated model versions, suggesting possible mitigation directions at the process and system level. These findings motivate a shift in health-related LLM governance, moving from auditing what models say about fairness to auditing whether models decide fairly, and from treating fairness as a static model property to treating it as an empirically validated outcome of model–language–context interactions.

Method

Scale completion tasks

Human data sources. To establish human reference distributions for health-related stigma, we aggregated published empirical data from six health-related stigma scales across studies conducted over the past two decades. In total, the pooled dataset comprised 56,612 participants (Supplementary Table 1 provides the full study details).

Stigma scales. We used six commonly applied scales assessing HIV-, HBV- and mental-health-related stigma.

HIV Stigma Scale. The HIV Stigma Scale[69] consists of nine items assessing stigma towards individuals living with HIV. Items are rated on a five-point Likert scale ranging from 0 (strongly disagree) to 4 (strongly agree), with one item reverse-scored. An average score is computed, with higher values indicating greater HIV stigma. Pooled data from 9,740 participants were included.

HIV-related Stigma and Discrimination Scale. The HIV-related Stigma and Discrimination Scale[70] consists of 19 items assessing stigma towards individuals living with HIV. Items are rated on a five-point Likert scale ranging from 0 (strongly disagree) to 4 (strongly agree). Consistent with the focus on individuals' own stigma towards individuals living with HIV, only the eight-item public stigma subscale was used in the present study. An average score is computed, with higher values indicating greater HIV stigma. Pooled data from 14,203 participants were included.

Toronto Hepatitis B Virus (HBV) Stigma Scale. The Toronto HBV Stigma Scale[71] consists of 20 items assessing stigma towards individuals with hepatitis B. Items are rated on a five-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). A total score is computed, with higher values indicating greater HBV stigma. Pooled data from 1,784 participants were included.

HBV Stigma Scale. The HBV Stigma Scale[72] consists of five items assessing stigma towards individuals with hepatitis B. Items are rated on a three-point scale ranging from 0 to 2. A total score is computed, with higher values indicating greater HBV stigma. Pooled data from 20,622 participants were included.

Community Attitudes toward Mental Illness. The Community Attitudes toward Mental Illness Scale (CAMI-40)[73] consists of 40 items assessing stigma towards individuals with mental disorders. Items are rated on a five-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree), with 20 items reverse-scored. An average score is computed, with higher values indicating greater mental illness stigma. Pooled data from 3,811 participants were included.

Depression Stigma Scale. The Depression Stigma Scale (DSS)[74] consists of 40 items assessing stigma towards individuals with depression. The DSS comprises two subscales: public stigma, measuring respondents' own stigmatizing attitudes, and perceived stigma, measuring beliefs about stigma held by others in society. Consistent with the focus on public stigma, only the 20-item public stigma subscale was used in the present study. Items are rated on a five-point Likert scale ranging from 0 (strongly disagree) to 4 (strongly agree). A total score is computed, with higher values indicating greater stigma. Pooled data from 6,452 participants were included.

LLM evaluation. We evaluated six LLMs: GPT-5.1, Claude-4.5 (Sonnet), Llama-3.3 (70B), Grok-4, DeepSeek-3.2 and Qwen-3 (plus). All model queries were conducted in December 2025, using the versions accessible through the corresponding provider interfaces at that time. Provider-reported model identifiers, access routes and evaluation dates are reported in Supplementary Table 23. For each model, we used a standardized questionnaire prompt and required responses as integers according to the scale-specific Likert instructions. All model queries were run with a temperature of 1.0. Each model–scale pair was sampled independently for ten repetitions. Model outputs were generated in a structured format at the questionnaire stage, with each response returned as a JSON object. JSON-object mode was requested when supported. The full prompts are provided in Supplementary Table 24. Outputs were processed using a tolerant parsing procedure that removed surrounding text (for example, code fences) and extracted the first valid JSON object. API failures and parsing failures were explicitly logged, and requests were automatically retried until a valid structured response was obtained. All model queries, logging and data assembly were implemented in Python (3.10.18) using batch execution and resumable pipelines. Statistical analyses and figure generation were conducted in R (4.5.1).

Contextual judgement task

Scenario construction and validation. Scenarios consisted of short descriptions of everyday social situations involving a focal individual, followed by an incomplete sentence that prompted free-text continuation. Within each scenario, the prompts and situational context were held constant, while the health condition of the focal individual was systematically varied. Health conditions included five mental-health conditions (schizophrenia, bipolar disorder, depression, anxiety and alcohol dependence), two physically high-stigmatized conditions (HIV and HBV), two physically low stigmatized conditions (chronic back pain and hypertension) and a healthy condition. Scenario development integrated theory-driven design with iterative expert and lived-experience input. We first identified ten recurrent micro-level stigma themes from prior literature (Supplementary Table 25), capturing common patterns of everyday social responses to focal health conditions. To ensure ecological grounding, these themes were complemented with first-person accounts from online social media and community posts by individuals living with the focal health conditions[75–77], which informed the generation of an initial pool of candidate scenarios. All candidate scenarios were subsequently reviewed and refined by a panel consisting of one senior researcher with expertise in stigma research and three doctoral students trained in related research areas. To further enhance clarity, realism and sensitivity, we conducted semi-structured interviews with 12 individuals living with the focal health conditions. Participant characteristics are reported in Supplementary Table 26. Participants provided feedback on wording, plausibility and appropriateness, based on which scenarios were iteratively revised. The final scenario set comprised 51 scenarios spanning four social domains: work and professional, education and school, public and everyday, and family and intimate contexts. The scenario distribution across domains is provided in Supplementary Table 27. To assess validity, we conducted a quantitative evaluation in an independent sample of 56 individuals living with the focal health conditions. Participant characteristics are reported in Supplementary Table 28. Participants rated each scenario on five dimensions: context realism, response realism, experienced prevalence, perceived commonness, and wording clarity, using seven-point Likert scales. Across all scenario-level aggregated ratings, 99.6% (254/255) exceeded the midpoint of 3.5, and 96.5% (246/255) exceeded 4. The complete set of scenarios is presented in Supplementary Table 29.

Human data collection. Human participant data were collected from a total of 399 participants across two complementary samples: an online sample recruited via Prolific and an in-person sample collected through supervised laboratory sessions. The study used a within-subjects design, in which each participant completed sentence continuation tasks for multiple scenarios. For each participant, a subset of scenarios and associated health conditions was randomly sampled (subset sampling design), with no scenario repeated within a participant. This design was used to limit participant burden and reduce fatigue, while ensuring

coverage of all scenario–condition combinations at the aggregate level. The online sample comprised 286 adults. To reduce the possibility of external assistance and maintain response authenticity, participants were instructed not to switch screens or access other applications during the task. In the in-person sample, we recruited 113 participants, who completed the same sentence-continuation task using handwritten responses in a controlled setting. During the sessions, participants were supervised and were not allowed to use electronic devices. In total, the combined dataset comprised 5,774 observations, of which 5,529 valid observations were retained for analysis after exclusions. Demographic characteristics of the combined sample are presented in Table 1, with sample-specific details reported separately in Supplementary Table 30 (online sample) and Supplementary Table 31 (in-person sample). The study protocol was approved by the institutional review board of Peking University (no. 20250601). All participants provided informed consent before the study, and received a debriefing at the end of the study explaining the purpose of the research.

LLM evaluation. We evaluated six LLMs: GPT-5.1, Claude-4.5 (Sonnet), Llama-3.3 (70B), Grok-4, DeepSeek-3.2 and Qwen-3 (plus) across two languages, English and Chinese. All model generations were conducted between June and December 2025. Provider-reported model identifiers, access routes and evaluation dates are reported in Supplementary Table 23. For each scenario, models were prompted with the same description and incomplete sentence and asked to generate a free-text continuation. Examples are shown in the Supplementary Note. Each model–scenario pair was sampled independently for ten iterations. All model generations were conducted with a temperature of 1.0. For models with available reasoning-enabled versions (Claude-4.5, DeepSeek-3.2, Qwen-3, GPT-5.1, Grok-4), both reasoning-enabled and standard (chat) versions were evaluated under identical prompts and generation settings across all scenarios. The full prompts are provided in Supplementary Table 32.

Coding and annotation. A coding scheme was developed through iterative discussion by one expert in health-related stigma research and three doctoral students in psychology with training in stigma research. The final codebook defined six stigma types: danger perception, incompetence and devaluation, blame, social distance, control and surveillance, and benevolence and pity. Two full-time research assistants annotated all continuations over a five-month period. Responses were first coded for the presence or absence of stigma-congruent content (binary). A response was coded as stigma-congruent if the continuation implied differential or unfavourable treatment solely based on the individual's health condition, when such information was not relevant to the task. For example, in a hiring scenario where a candidate fully meets all job requirements, a continuation suggesting that the candidate should not be hired after disclosure of a health condition (for example, 'the HR decided not to proceed with the offer') was coded as stigma-congruent, whereas a continuation indicating that the hiring decision should remain based on qualifications (for example, 'the offer was made based on merit') was coded as non-stigma-congruent. For each scenario, a representative stigma-congruent response is provided in Supplementary Table 29. Responses identified as stigma-congruent were further coded for one or more of the six stigma types. Discrepancies between coders were resolved through discussion to reach consensus. Inter-coder reliability was relatively high, with Cohen's $\kappa = 0.87$ for the binary stigma classification and a Jaccard similarity coefficient of 0.75 for stigma pattern coding.

Reasoning strategy identification and prompt-based evaluation. Reasoning content generated by reasoning-enabled models was annotated using a predefined framework of nine reasoning strategies (Table 2). Strategy coding was conducted using GPT-3.5 with temperature set to 0, based on a structured prompt for identifying the presence of each strategy in the reasoning text. For prompt-based analyses, a system prompt incorporating selected reasoning strategies was applied to non-reasoning (chat) model versions. Evaluation was conducted on all scenarios. All six LLMs (chat versions) were evaluated under identical prompts and generation settings, with the only difference being the inclusion of the strategy prompt in the system prompt. Model outputs were generated with the same settings as in the main analyses.

Expert evaluation for sensitivity analyses. To further distinguish stigma-congruent responses from condition-relevant caution or policy realism, we conducted an additional expert-based sensitivity analysis. Two independent expert groups were recruited: 16 physicians working in physical-health-related fields and 11 mental-health counsellors with clinical or counselling experience in mental health. Demographic characteristics of the expert samples are reported in Supplementary Table 33. Experts evaluated all non-baseline scenario–condition pairs included in the contextual judgement task, yielding 459 pairs in total (204 physical-health pairs and 255 mental-health pairs). For each pair, experts provided four ratings on five-point scales (condition relevance, justified caution, policy/institutional realism and stigma exclusion) and an overall categorical judgement (A = primarily stigma-congruent, B = primarily condition-relevant caution, C = primarily policy/institutional realism). We re-estimated the main contextual judgement models under three exclusion criteria: (i) excluding pairs whose modal categorical judgement is B or C; (ii) excluding pairs with a mean caution/policy score of at least 3 on the five-point scale (averaged across the caution and policy realism items); and (iii) excluding pairs that met either criterion (i) or (ii), that is, pairs classified as B or C by modal judgement, or with a mean caution/policy score ≥3 that was at least as high as the mean stigma-exclusion score. The results were consistent with the main analysis. Detailed exclusion counts and model results are reported in Supplementary Tables 34–36.

Exploratory analysis and intersection between social attributes and health conditions. To examine whether health-condition stigma in model outputs may interact with other social attributes, we conducted an exploratory intersection audit. Based on the original scenario set, we created identity-cued variants in which the protagonist's name was systematically modified while all other scenario content was held constant. For the gender-cue comparison, eligible scenarios were instantiated with a male-coded name (for example, Adam) or a female-coded name (for example, Grace). For the ethnicity-cue comparison, scenarios were instantiated with fixed names intended to cue perceived Asian, White, Black or Latino identity. Manipulation checks showed that across the six LLM models, gender-cue accuracy was 94.8% and ethnicity-cue accuracy was 96.9% (Supplementary Table 37). We evaluated all models across ten scenarios, with five repetitions per scenario–condition pair and six identity variants, yielding a total of 18,000 observations. All scenarios and name manipulations used in this analysis are provided in Supplementary Table 38. Across all tested attributes, the main effects of health condition were statistically significant, and interaction effects between health condition and social attributes were not significant. The results are reported in Supplementary Tables 39–41.

Sensitivity analyses. To assess the consistency of results across different human data sources, we repeated all primary analyses separately for each sample. We also conducted generation-randomness sensitivity analyses within the contextual judgement task by treating the ten iterations as independent generation replicates, re-running the main mixed-effects models separately for each replicate, and performing leave-one-run-out analyses in which the pooled models were re-estimated after excluding each iteration in turn. Across all sensitivity analyses, the results remained consistent with the primary conclusions. The results are reported in Supplementary Tables 42–45 and Supplementary Fig. 1.

Statistical analysis

All analyses were conducted in R (4.5.1). For the scale completion task, we synthesized human data for each scale using random-effects meta-analysis in the R package metafor. The human pooled means were compared to LLM means using a two-tailed z test. We used $P < 0.05$ as the significance threshold. The false discovery rate was controlled using the Benjamini–Hochberg procedure. The effect size was measured by Glass's delta:

$$\Delta = \frac{\mu_M - \mu_H}{\text{s.d.}_H}$$

where μ_M denotes the mean score generated by the LLM across repeated runs, μ_H denotes the pooled mean score from human participants, and SD_H denotes the pooled standard deviation of the human data. To quantify model variability relative to humans, we used a variability ratio:

$$V_H = \frac{SD_M}{SD_H}$$

where SD_M and SD_H denote the standard deviations of model and human scores, respectively.

For the contextual judgement task, we fitted generalized linear mixed-effects models with a binomial distribution and logit link using lme4. Effects are reported as ORs with 95% CIs. Nested models were compared using likelihood ratio tests. All hypothesis tests were two-sided. Group differences were summarized using estimated marginal means from the R package emmeans. For multiple comparisons, P values were adjusted using the Benjamini–Hochberg procedure.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The data analysed in this study are available via OSF at https://osf.io/xfasm.

Code availability

The code used in this study is available via OSF at https://osf.io/xfasm.

References

1. Bedi, S. et al. Testing and evaluation of health care applications of large language models: a systematic review. *JAMA* **333**, 319–328 (2025).
2. Duan, Z. et al. Multi-center benchmarking of large language models for clinical decision support in lung cancer screening. *Cell Rep. Med.* **6**, 102465 (2025).
3. Hao, Y. et al. Large language model integrations in cancer decision-making: a systematic review and meta-analysis. *NPJ Digit. Med.* **8**, 450 (2025).
4. Ren, Z. et al. Healthcare agent: eliciting the power of large language models for medical consultation. *NPJ Artif. Intell.* **1**, 24 (2025).
5. Pan, A., Musheyev, D., Bockelman, D., Loeb, S. & Kabarriti, A. E. Assessment of artificial intelligence Chatbot responses to top searched queries about cancer. *JAMA Oncol.* **9**, 1437–1440 (2023).
6. Cornelison, B. R., Erstad, B. L. & Edwards, C. Accuracy of a chatbot in answering questions that patients should ask before taking a new medication. *J. Am. Pharm. Assoc.* **64**, 102110 (2024).
7. Roldan-Vasquez, E. et al. Reliability of artificial intelligence chatbot responses to frequently asked questions in breast surgical oncology. *J. Surg. Oncol.* **130**, 188–203 (2024).
8. Wang, X., Zhou, Y. & Zhou, G. The application and ethical implication of generative AI in mental health: systematic review. *JMIR Ment. Health* **12**, e70610 (2025).
9. Wilhelm, T. I., Roos, J. & Kaczmarczyk, R. Large language models for therapy recommendations across 3 clinical specialties: comparative study. *J. Med. Internet Res.* **25**, e49324 (2023).
10. Busch, F. et al. Current applications and challenges in large language models for patient care: a systematic review. *Commun. Med.* **5**, 26 (2025).
11. Maity, S. & Saikia, M. J. Large language models in healthcare and medical applications: a review. *Bioengineering* **12**, 631 (2025).
12. Tam, T. Y. C. et al. A framework for human evaluation of large language models in healthcare derived from literature review. *NPJ Digit. Med.* **7**, 258 (2024).
13. Shiferaw, M. W., Zheng, T., Winter, A., Mike, L. A. & Chan, L.-N. Assessing the accuracy and quality of artificial intelligence (AI) chatbot-generated responses in making patient-specific drug-therapy and healthcare-related decisions. *BMC Med. Inform. Decis. Mak.* **24**, 404 (2024).
14. Earp, B. D., McLoughlin, K. L., Monrad, J. T., Clark, M. S. & Crockett, M. J. How social relationships shape moral wrongness judgments. *Nat. Commun.* **12**, 5776 (2021).
15. Jin, W. Y. & Peng, M. The effects of social perception on moral judgment. *Front. Psychol.* **11**, 557216 (2021).
16. Mårtensson, E. Construal level theory and moral judgments: how thinking abstractly modifies morality. *J. Eur. Psychol. Stud.* **8**, 30–40 (2017).
17. Bartels, D. M., Bauman, C. W., Cushman, F. A., Pizarro, D. A. & McGraw, A. P. in *The Wiley Blackwell Handbook of Judgment and Decision Making* (eds Keren, G. & Wu, G.), 478–515 (Wiley, 2015).
18. Webster, C. S., Taylor, S., Thomas, C. & Weller, J. M. Social bias, discrimination and inequity in healthcare: mechanisms, implications and recommendations. *BJA Educ.* **22**, 131–137 (2022).
19. Vela, M. B. et al. Eliminating explicit and implicit biases in health care: evidence and research needs. *Annu. Rev. Public Health* **43**, 477–501 (2022).
20. Gopal, D. P., Chetty, U., O'Donnell, P., Gajria, C. & Blackadder-Weinstein, J. Implicit bias in healthcare: clinical practice, research and decision making. *Future Healthc. J.* **8**, 40–48 (2021).
21. Thirsk, L. M., Panchuk, J. T., Stahlke, S. & Hagtvedt, R. Cognitive and implicit biases in nurses' judgment and decision-making: a scoping review. *Int. J. Nurs. Stud.* **133**, 104284 (2022).
22. Akbari, H., Mohammadi, M. & Hosseini, A. Disease-related stigma, stigmatizers, causes, and consequences: a systematic review. *Iran. J. Public Health* **52**, 2042–2054 (2023).
23. Stangl, A. L. et al. The Health Stigma and Discrimination Framework: a global, crosscutting framework to inform research, intervention development, and policy on health-related stigmas. *BMC Med.* **17**, 31 (2019).
24. Goldberg, D. S. On stigma & health. *J. Law Med. Ethics* **45**, 475–483 (2017).
25. Zell, C. S., Thorp, S. R. & Thompson, K. J. The Relative Understanding of Stigma about Health (RUSH) study: the role of controllability and knowledge in explaining condition-specific variability. *Community Ment. Health J.* **62**, 513–526 (2026).
26. Bharadwaj, P., Pai, M. M. & Suziedelyte, A. Mental health stigma. *Econ. Lett.* **159**, 57–60 (2017).
27. Scambler, G. Health-related stigma. *Sociol. Health Illn.* **31**, 441–455 (2009).
28. Fauk, N. K., Ward, P. R., Hawke, K. & Mwanri, L. HIV stigma and discrimination: perspectives and personal experiences of healthcare providers in Yogyakarta and Belu, Indonesia. *Front. Med.* **8**, 625787 (2021).
29. Kågström, A. et al. Mental health stigma and its consequences: a systematic scoping review of pathways to discrimination and adverse outcomes. *EClinicalMedicine* **89**, 103588 (2025).

30. Earnshaw, V. A. & Quinn, D. M. The impact of stigma in healthcare on people living with chronic illnesses. *J. Health Psychol.* **17**, 157–168 (2012).
31. Rai, S. S., Syurina, E. V., Peters, R. M., Putri, A. I. & Zweekhorst, M. B. Non-communicable diseases-related stigma: a mixed-methods systematic review. *Int. J. Environ. Res. Public Health* **17**, 6657 (2020).
32. Lu, Q. et al. The effect of stigma on social participation in community-dwelling Chinese patients with stroke sequelae: a cross-sectional study. *Clin. Rehabil.* **36**, 407–414 (2022).
33. Maher, V. & Psych, B. *The Impact of Stigma and Social Anxiety on Social Participation in People with Severe Mental Illness* (Open Research Newcastle, 2025).
34. Liu, Y. & Li, Y. Community participation and subjective perception of recovery and quality of life among people with serious mental illnesses: the mediating role of self-stigma. *Soc. Psychiatry Psychiatr. Epidemiol.* **60**, 1335–1345 (2025).
35. Kaufmann, T., Weng, P., Bengs, V. & Hüllermeier, E. A survey of reinforcement learning from human feedback. *Trans. Mach. Learn. Res.* https://openreview.net/forum?id=f7OkIurx4b (2025).
36. Wang, K. et al. A comprehensive survey in LLM (-agent) full stack safety: data, training and deployment. Preprint at https://arxiv.org/abs/2504.15585 (2025).
37. Ji, J. et al. Beavertails: towards improved safety alignment of LLM via a human-preference dataset. *Adv. NIPS* **36**, 24678–24704 (2023).
38. Tan, B. C. Z. & Lee, R. K.-W. Unmasking implicit bias: evaluating persona-prompted LLM responses in power-disparate social scenarios. In *Proc. 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (eds Chiruzzo, L. et al.) 1075–1108 (Association for Computational Linguistics, 2025).
39. Kumar, D., Jain, U., Agarwal, S. & Harshangi, P. Investigating implicit bias in large language models: a large-scale study of over 50 LLMs. Preprint at https://arxiv.org/abs/2410.12864 (2024).
40. Majumdar, A., Chen, F., Li, J. & Wang, X. Evaluating LLMs for demographic-targeted social bias detection: a comprehensive benchmark study. Preprint at https://arxiv.org/html/2510.04641v1 (2025).
41. Bai, X., Wang, A., Sucholutsky, I. & Griffiths, T. L. Explicitly unbiased large language models still form biased associations. *Proc. Natl Acad. Sci. USA* **122**, e2416228122 (2025).
42. Zhao, Y. et al. in *Findings of the Association for Computational Linguistics: ACL 2025* (eds Che, W. et al.) 1–12 (Association for Computational Linguistics, 2025).
43. Dong, X., Wang, Y., Yu, P. S. & Caverlee, J. Probing explicit and implicit gender bias through llm conditional text generation. Preprint at https://arxiv.org/abs/2311.00306 (2023).
44. Omar, M. et al. Evaluating and addressing demographic disparities in medical large language models: a systematic review. *Int. J. Equity Health* **24**, 57 (2025).
45. Levartovsky, A., Omar, M., Nadkarni, G. N., Kopylov, U. & Klang, E. Sociodemographic bias in large language model–assisted gastroenterology. *JAMA Netw. Open* **8**, e2532692 (2025).
46. Criss, S. et al. HIV stigma exists—exploring ChatGPT's HIV advice by race and ethnicity, sexual orientation, and gender identity. *J. Racial Ethn. Health Disparities* **12**, 3622–3635 (2025).
47. Heller, U. C., Grant, L. H., Yasui, M. & Keysar, B. Culturally anchored mental-health attitudes: the impact of language. *Clin. Psychol. Sci.* **12**, 290–304 (2024).
48. Gronholm, P. C. et al. Exploring perspectives of stigma and discrimination among people with lived experience of mental health conditions: a co-produced qualitative study. *EClinicalMedicine* **70**, 102509 (2024).
49. Ong Zhen Mei, E., Jones, L. & Occhipinti, S. Where there's smoke, there's stigma? Written expressions about individuals with lung cancer reveal cultural and language differences. *J. Lang. Soc. Psychol.* **44**, 417–440 (2025).
50. Ren, Y., Wang, S., Fu, X. & Shi, X. A systematic review and meta-analysis of implicit stigma toward people with mental illness among different groups: measurement, extent, and correlates. *Psychol. Res. Behav. Manag.* **18**, 851–875 (2025).
51. Greenwald, A. G. et al. Implicit-bias remedies: treating discriminatory bias as a public-health problem. *Psychol. Sci. Public Interest* **23**, 7–40 (2022).
52. van Beukering, I. E. et al. In what ways does health related stigma affect sustainable employment and well-being at work? A systematic review. *J. Occup. Rehabil.* **32**, 365–379 (2022).
53. van Bortel, T. et al. Anticipated and experienced stigma and discrimination in the workplace among individuals with major depressive disorder in 35 countries: qualitative framework analysis of a mixed-method cross-sectional study. *BMJ Open* **14**, e077528 (2024).
54. Toumi, M. et al. Experience and impact of stigma in people with chronic hepatitis B: a qualitative study in Asia, Europe, and the United States. *BMC Public Health* **24**, 611 (2024).
55. Reeves, S. L., Tse, C., Logel, C. & Spencer, S. J. When seeing stigma creates paternalism: Learning about disadvantage leads to perceptions of incompetence. *Group Process. Intergr. Relat.* **25**, 1202–1222 (2022).
56. Valery, K.-M. et al. How do mental health professionals stigmatize incompetence in schizophrenia? *Stig. Health* **10**, 286–292 (2025).
57. Asrina, A., Ikhtiar, M., Idris, F. P., Adam, A. & Alim, A. Community stigma and discrimination against the incidence of HIV and AIDS. *J. Med. Life* **16**, 1327–1334 (2023).
58. Ajzen, I. & Dasgupta, N. in *The Sense of Agency* (eds Haggard, P. & Eitam, B.) 115–144 (Oxford Univ. Press, 2015).
59. Killen, M., McGlothlin, H. & Henning, A. in *Intergroup Attitudes and Relations in Childhood through Adulthood* (eds Levy, S. R. & Killen, M.) 126–145 (Oxford Univ. Press, 2008).
60. Zhao, Y., Zhu, J., Xu, C., Liu, Y. & Li, X. in *Findings of the Association for Computational Linguistics: ACL 2025* (eds Che, W. et al.) 24747–24760 (Association for Computational Linguistics, 2025).
61. Dorn, R., Kezar, L., Morstatter, F. & Lerman, K. Harmful speech detection by language models exhibits gender-queer dialect bias. In *Proc. 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (eds Arnosti, N. et al.) 1–12 (Association for Computing Machinery, 2024).
62. Dai, S. et al. Bias and unfairness in information retrieval systems: new challenges in the LLM era. In *Proc. 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (eds Baeza-Yates, R. & Bonchi, F.) 6437–6447 (Association for Computing Machinery, 2024).
63. Schomerus, G., Kummetat, J., Angermeyer, M. & Link, B. G. Putting yourself in the shoes of others'–Relatability as a novel measure to explain the difference in stigma toward depression and schizophrenia. *Soc. Psychiatry Psychiatr. Epidemiol.* **60**, 1883–1893 (2025).
64. Lin, C.-Y. & Tsang, H. W. Stigma, health and well-being. *Int. J. Environ. Res. Public Health* **17**, 7615 (2020).
65. Pearl, R. L. et al. Measuring internalized health-related stigma across health conditions: development and validation of the I-HEARTS Scale. *BMC Med.* **22**, 435 (2024).
66. Vaishnav, M. et al. Stigma towards mental illness in Asian nations and low-and-middle-income countries, and comparison with high-income countries: a literature review and practice implications. *Indian J. Psychiatry* **65**, 995–1011 (2023).

67. Ahmad, S. S. & Koncsol, S. W. Cultural factors influencing mental health stigma: perceptions of mental illness (POMI) in Pakistani emerging adults. *Religions* **13**, 401 (2022).
68. Jacobs, S. & Quinn, J. Cultural reproduction of mental illness stigma and stereotypes. *Soc. Sci. Med.* **292**, 114552 (2022).
69. Kalichman, S. C. et al. Perceived HIV stigma and HIV testing among men and women in rural Uganda: a population-based study. *Lancet HIV* **7**, e817–e824 (2020).
70. Genberg, B. L. et al. A comparison of HIV/AIDS-related stigma in four countries: negative attitudes and perceived acts of discrimination towards people living with HIV/AIDS. *Soc. Sci. Med.* **68**, 2279–2287 (2009).
71. Li, D. et al. The impact of hepatitis B knowledge and stigma on screening in Canadian Chinese persons. *Can. J. Gastroenterol. Hepatol.* **26**, 597–602 (2012).
72. Leng, A. et al. Hepatitis B discrimination in everyday life by rural migrant workers in Beijing. *Hum. Vaccin. Immunother.* **12**, 1164–1171 (2016).
73. Taylor, S. M. & Dear, M. J. Scaling community attitudes toward the mentally ill. *Schizophr. Bull.* **7**, 225–240 (1981).
74. Griffiths, K. M., Christensen, H. & Jorm, A. F. Predictors of depression stigma. *BMC Psychiatry* **8**, 25 (2008).
75. Tseriotou, T. et al. Overview of the CLPsych 2025 shared task: capturing mental health dynamics from social media timelines. In *Proc. 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)* (eds Zirikly, A. et al.) 193–217 (Association for Computational Linguistics, 2025).
76. Wang, X., Zhou, Y. & Zhou, G. Unveiling the cognitive burden: the impact of stigma on distorted thinking among individuals living with hepatitis B. *Int. J. Clin. Health Psychol.* **25**, 100556 (2025).
77. Matza, L. S. et al. Qualitative thematic analysis of social media data to assess perceptions of route of administration for antiretroviral treatment among people living with HIV. *Patient* **13**, 409–422 (2020).

Acknowledgements

We thank L. Du and E. Zhang for their valuable assistance with coding and annotating the model-generated responses.

Author contributions

X.W. was responsible for investigation, formal analysis, data interpretation and writing the original draft. Y.Z. conceived and supervised the project, and contributed computational resources, software implementation, visualization, and manuscript review and editing. G.Z. managed the research process and provided overall methodological guidance.

Funding

This research was supported by the National Social Science Fund of China under grant no. 21BSH158 and the National Natural Science Foundation of China under grant no. 32271136.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44360-026-00164-4.

Correspondence and requests for materials should be addressed to Yujia Zhou or Guangyu Zhou.

Peer review information *Nature Health* thanks Shadi Nourriz, Mahmud Omar, Estelle Smith and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Lorenzo Righetto, in collaboration with the *Nature Health* team.

nature portfolio

Corresponding author(s): Yujia Zhou, Guangyu Zhou

Last updated by author(s): Jun 5, 2026

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	For the scale-completion task, human benchmark data were obtained from published validation and benchmark studies of the corresponding stigma scales; extracted summary statistics included scale means, standard deviations, standard errors and sample sizes, which were entered into structured CSV files for meta-analytic synthesis. For the contextual judgement task, online human data were collected on Prolific using a browser-based survey interface implemented in jsPsych. For the in-person human sample, no software was used during task administration: participants completed the same materials on printed paper forms with handwritten responses in supervised laboratory sessions, without access to electronic devices. Large language model outputs for both tasks were generated using Python 3.10.18 scripts that queried publicly available model-developer APIs.
Data analysis	All statistical analyses and figure generation were conducted in R 4.5.1. For the scale-completion task, human benchmark estimates were synthesized using random-effects meta-analysis with metafor 4.8-0. LLM-human z tests, Glass's delta, variability ratios and Benjamini-Hochberg-adjusted p values were computed in base R. Analyses used generalized linear mixed-effects models implemented with lme4 1.1-37, estimated marginal means and contrasts with emmeans 1.11.2.8, bootstrap confidence intervals with boot 1.3-31, and multiple-comparison correction using the Benjamini-Hochberg false-discovery-rate procedure implemented in base R. Data processing and visualization used tidyverse 2.0.0, dplyr 1.1.4, tidyr 1.3.1, purrr 1.1.0, tibble 3.3.0, readr 2.1.5, jsonlite 2.0.0, janitor 2.2.1, ggplot2 4.0.1, scales 1.4.0, ggsci 4.2.0, ggh4x 0.3.1, ComplexHeatmap 2.25.2, ggtext 0.1.2, glue 1.8.0, patchwork 1.3.2, cowplot 1.2.0 and bruceR 2025.8. Results tables were formatted with bruceR.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The data analyzed in this study are available via OSF at https://osf.io/xfasm.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Gender was considered as demographic information in the current study design. It was self-reported by participants using categories: man, woman, other. Sex assigned at birth was not collected. Overall, the sample included 54.4% participants who identified as men and 45.6% who identified as women.
Reporting on race, ethnicity, or other socially relevant groupings	Race and ethnicity were self-reported by participants using predefined categories (African, Asian, Hispanic, White, Mixed). These variables were collected for sample characterization only and were not used as proxies for other social variables.
Population characteristics	Participants (N = 399) had a mean age of 38.67 years (SD = 14.33). Gender was self-reported, with 54.4% identifying as men and 45.6% as women. The sample was ethnically diverse: 65.4% identified as White, 30.6% as Asian, 3.0% as Mixed ethnicity, 0.8% as Hispanic, and 0.3% as African. Educational attainment also varied across the sample: 11.5% reported a high school education or below, 58.1% held a bachelor's degree, 21.8% held a master's degree, and 8.5% held a doctoral degree. Annual household income was distributed as follows: 29.1% reported incomes below $20,000, 29.1% between $20,000 and $39,999, 18.8% between $40,000 and $59,999, 10.8% between $60,000 and $79,999, 3.5% between $80,000 and $99,999, and 5.5% reported incomes of $100,000 or above; 3.3% preferred not to disclose income.
Recruitment	Participants were recruited from two complementary sources. The online sample was recruited through Prolific. The in-person sample was recruited via social media platforms and completed the same task in supervised laboratory sessions using handwritten responses without access to electronic devices. Because participation was voluntary in both recruitment modes, some degree of self-selection bias is possible. Individuals who choose to participate in online research or laboratory studies may differ from the broader population in motivation, familiarity with research tasks, literacy, attentiveness, or interest in health- and AI-related topics. The Prolific sample may also overrepresent individuals who are experienced with online studies, whereas the in-person sample may reflect a somewhat different subset of participants who were willing and able to attend supervised sessions.
Ethics oversight	We obtained ethical approval from the Institutional Review Board of Peking University (No. 20250601). All participants provided informed consent prior to participation.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	This study used a quantitative, multi-component evaluation design to assess health-related stigma in large language models. It combined (i) quantitative comparison of model outputs on validated stigma scales with meta-analytic human benchmarks and (ii) a quantitative contextual judgement task using sentence-completion scenarios. Human participant data were collected for benchmarking the contextual judgement task.
Research sample	Large language models: Six large language models were evaluated: GPT-5.1, Claude-4.5 (Sonnet), Llama-3.3 (70B), Grok-4, DeepSeek-3.2, and Qwen-3 (Plus). Human participants: Participants (N = 399) had a mean age of 38.67 years (SD = 14.33). Gender was self-reported, with 54.4% identifying as men and 45.6% as women. The sample was ethnically diverse: 65.4% identified as White, 30.6% as Asian, 3.0% as Mixed ethnicity, 0.8% as Hispanic, and 0.3% as African. Educational attainment also varied across the sample: 11.5% reported a high

	school education or below, 58.1% held a bachelor's degree, 21.8% held a master's degree, and 8.5% held a doctoral degree. Annual household income was distributed as follows: 29.1% reported incomes below $20,000, 29.1% between $20,000 and $39,999, 18.8% between $40,000 and $59,999, 10.8% between $60,000 and $79,999, 3.5% between $80,000 and $99,999, and 5.5% reported incomes of $100,000 or above; 3.3% preferred not to disclose income. The sample was not intended to be nationally representative and was chosen to provide a human reference distribution for the contextual judgement task. Existing datasets: Human benchmark data for stigma scales were obtained from previously published studies, comprising a pooled sample of 56,612 participants across six validated stigma scales.
Sampling strategy	The study used a convenience-based recruitment strategy across two complementary sampling modes. The online sample was recruited through Prolific, and the in-person sample was recruited for supervised laboratory sessions. Participants were not selected using probability-based, stratified, or snowball sampling procedures. Instead, sampling was designed to obtain sufficient coverage of the contextual judgement task across a broad set of scenario–condition combinations while combining the practical advantages of online data collection with an additional controlled in-person validation sample. No probability-based sampling procedure was used. No formal a priori power analysis was conducted. Instead, sample sizes were determined based on feasibility, participant burden, and the need to obtain broad aggregate coverage across scenario–condition combinations in an open-ended within-subjects design. To reduce fatigue and preserve response quality, each participant completed a randomly sampled subset of scenario–condition pairs rather than the full factorial design. The final sample comprised 399 participants and 5,529 observations.
Data collection	Data were collected using two parallel procedures. For the online sample, participants completed the sentence-continuation task remotely via a browser-based interface implemented in jsPsych. Responses were entered directly on a computer and recorded digitally through the online experimental platform. During data collection, participants were instructed not to switch screens or access other applications in order to reduce the possibility of external assistance and maintain response authenticity. No one was physically present with participants during the online sessions. For the in-person sample, participants completed the same sentence-continuation task using printed materials and handwritten responses in supervised laboratory sessions. A research assistant was present throughout the session to administer the procedure, monitor compliance, and ensure that no electronic devices were used. No one other than the participant and the research assistant was present during data collection. Participants were not informed of the specific study hypotheses. Instead, the study was described as an investigation of language-based sentence completion and social judgments. Researcher blinding to experimental condition was not applicable because participants completed pre-specified task materials and no experimental condition assignment was performed during data collection. For the online sample, data collection was fully automated. For the in-person sample, the research assistant administered standardized materials and did not influence participant responses or outcome coding.
Timing	Data collection was conducted in two separate periods corresponding to the two human samples. The online Prolific sample was collected between November 25, 2025, and November 30, 2025. The in-person offline sample was collected between March 10, 2026, and April 10, 2026.
Data exclusions	A total of 245 responses were excluded due to empty entries. The exclusion criteria were predefined.
Non-participation	No participant dropouts were reported after consent. All recruited participants completed the task.
Randomization	Participants were not assigned to between-subject experimental groups.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*