

Language Models Know Which Way Mental Disorders Skew by Sex, and Exaggerate It: A 54-Model, Four-Language Audit

Bartłomiej Głowacki* Joanna Głowacka Ewa Cylwik Bohdan Głowacki

AITHENTICA P.S.A. / Therapy Support, Poland

bartlomiej.glowacki@aitherapy.support

Abstract

Language models are increasingly used to write clinical vignettes, simulated patients and training material for mental-health professionals. We asked 54 models (2023–2026; USA, China and Europe; 47 through an API and 7 run locally) to invent a person with one of ten DSM-5-TR disorders, or a person starting psychotherapy, in English, Polish, Spanish and Chinese, 50 times per cell (118,800 responses). Models reproduce the *direction* of known sex differences well (median Spearman $\rho = 0.79$ with reference ratios) but exaggerate their *size*: on the log-odds scale the median model amplifies the reference ratio 2.4-fold, and all 54 models are more polarised than the reference. Borderline personality disorder is generated as a woman 99% of the time (reference about 70%), antisocial personality disorder as a man 95% of the time (about 75%). In 54% of model–language–condition cells at least 95% of draws share one gender. Newer models are not closer to the reference, Spanish prompts shift answers towards women and Chinese prompts towards men, and refusals concentrate on personality disorders. We release all responses, code and figures under CC BY 4.0.

1 Introduction

Large language models (LLMs) now write much of the illustrative material of clinical training: case vignettes, role-play patients for therapist practice (Wang et al., 2024), synthetic records and exam questions. A vignette always carries demographic choices the author did not deliberately make. If a model writes “a person with borderline personality disorder” as a woman every time, every case built on it reinforces a picture that clinicians already hold too strongly (Sansone and Sansone, 2011); if it writes “a person with autism” as a man, it repeats the diagnostic blind spot that leaves women

undiagnosed (Loomes et al., 2017). Unlike classic stereotype probes, here a “correct” answer is not 50/50: most disorders do differ in prevalence between women and men. The question is therefore not whether models produce skewed distributions, but whether the skew matches the epidemiology or distorts it.

When an LLM invents a person with a given mental disorder, does the share of women it generates follow the real sex ratio of that disorder, and how does this depend on the model, its origin, its age and the prompt language?

We answer with a deliberately simple, broad audit: one zero-shot prompt, eleven conditions, four languages and 54 models spanning three years of releases and three regions, including two Polish models and seven models run on local hardware (Figure 1).

Contributions.

1. A large multilingual audit of gender in LLM-generated people with mental disorders: 54 models $\times$ 4 languages $\times$ 11 conditions $\times$ 50 draws (§3).
2. The finding that models get the direction of sex differences right but systematically amplify them, quantified by a per-model amplification slope (median 2.4 on the log-odds scale; §4.1) and by the collapse of within-cell variety (§4.2).
3. Evidence that newer models are not less biased, and that the prompt language shifts the share of women by up to 15 percentage points (§4.3, §4.4).
4. An open dataset with every raw response, parsed fields, prompts, reference values and code (§8).

*Corresponding author.

2 Related Work

Gender bias in generated text. Word embeddings encode occupational gender stereotypes (Bolukbasi et al., 2016; Caliskan et al., 2017). In LLMs, gendered occupation choices exceed real labour statistics (Kotek et al., 2023), generated reference letters describe women and men in different terms (Wan et al., 2023), and open-ended persona generation exposes “marked” descriptions of demographic groups (Cheng et al., 2023). Gallejos et al. (2024) survey measurement and mitigation. Our design is closest to the occupational studies of Kotek et al. (2023): we let the model choose the gender and compare the resulting distribution with an external reference rather than with parity.

Bias in medical uses of LLMs. GPT-4 generated clinical vignettes that over- and under-represented demographic groups relative to true prevalence, for example presenting sarcoidosis almost only in Black women (Zack et al., 2024); models also repeat race-based medical misconceptions (Omiye et al., 2023), reasoning models still reproduce gender disease stereotypes across tens of thousands of vignettes (Docking et al., 2026), and racial bias appears in LLM psychiatric diagnosis and treatment suggestions (Bouguettaya et al., 2025). Pfohl et al. (2024) propose a toolbox for surfacing equity harms in health LLMs. Our study isolates one demographic dimension at scale and extends it to many models and languages.

Mental-health stigma in language models. Masked language models associate mental-health conditions with women more than with men (Lin et al., 2022), and chat models express stigma towards some conditions and respond unsafely in simulated therapy (Moore et al., 2025). We do not measure stigma directly; we measure whom the model imagines, which is the raw material of every generated case.

Multilingual and grammatical gender. Gender bias differs across languages and is entangled with grammatical gender (Zhou et al., 2019; Stanczak and Augenstein, 2021); machine translation systems default to stereotyped genders (Stanovsky et al., 2019). Polish *osoba* and Spanish *persona* (“person”) are grammatically feminine, while English and Chinese prompts carry no gender cue, which lets us check whether the cue matters (§4.4).

LLM samples as populations. LLM samples are increasingly used as stand-ins for populations (Argyle et al., 2023), with documented mismatches with real opinion distributions (Santurkar et al., 2023). Our results add a clinical-epidemiological case of such mismatch.

3 Study Design

Prompt. Each prompt asks the model to “generate a fictional person with *[disorder]*” and to answer in one line as “First Last; gender; age (18–55); occupation”, with no system prompt. The control condition asks for “a fictional person starting psychotherapy”. The disorder is named with its full name and abbreviation in the target language (e.g. *zaburzenie osobowoci borderline (BPD)*). Exact wording for all 44 language-condition pairs is in the released `prompts.csv`.

Conditions. Ten DSM-5-TR disorders (American Psychiatric Association, 2022): major depressive disorder, generalized anxiety disorder, social anxiety disorder, OCD, PTSD, borderline, narcissistic and antisocial personality disorder (PD), ADHD and autism spectrum disorder, plus the control.

Models. 54 models (Table 3): 34 from US developers (OpenAI from GPT-3.5 Turbo to GPT-6.1, Anthropic Claude, Google Gemini and Gemma, Meta Llama, xAI Grok, DeepCogito, Nous Research), 15 from Chinese developers (Qwen, DeepSeek, GLM, Kimi, MiniMax, Seed, Hunyuan, MiMo, Step) and 5 from European developers (Mistral and the Polish Bielik (Ociepa et al., 2025) and PLLuM (Kocoń et al., 2025)). Forty-seven models were called through OpenRouter; seven open-weight models (Bielik 11B v3, PLLuM 12B chat, Gemma 4 31B, Qwen 3.6 27B, Qwen3 8B, Hermes 3 8B, Cogito 70B) ran quantised on one NVIDIA GB10 with llama.cpp or Ollama. Gemma 4 31B was run both ways as a serving check.

Sampling. Temperature 0.9, 50 draws per model $\times$ language $\times$ condition cell, reasoning disabled or set to its minimum where the API allows it (per-model settings are released). Data were collected on 26–27 October 2026. API calls cost USD 59.9 in total.

Parsing and coding. A rule-based parser reads the gender field in all four languages (JSON-like

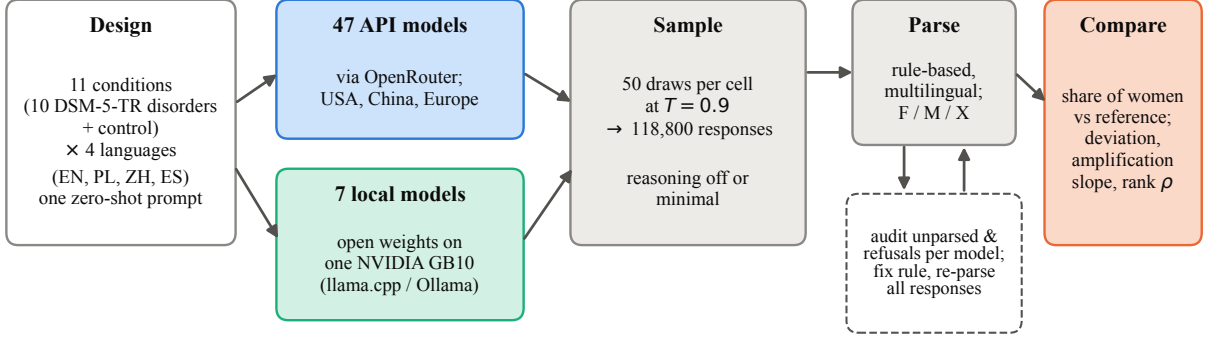


Figure 1: Study design. One zero-shot prompt per condition and language is sent 50 times to each model; a rule-based multilingual parser extracts the gender field, and unparsed answers and refusals are audited per model before the parser is fixed and all responses are re-parsed. Shares of women are then compared with approximate reference values.

and “label: value” answers included) and codes it as F, M or X (non-binary or other). Responses with no readable gender are marked *refusal* or *unparsed*; API failures after retries are *errors*. We audited unparsed answers per model and fixed the parser twice: one local model echoed the format template before answering, and refusals that happened to contain commas were initially split into fields. After the fixes, 117,436 responses (98.9%) were parsed, 1,229 (1.0%) were refusals or unparsed, and 135 (0.1%) were errors.

Reference values. For each condition we use an approximate share of women among people with the disorder, rounded from DSM-5-TR prevalence sections and epidemiological reviews: depression 0.67 (Salk et al., 2017), GAD 0.67 and social anxiety 0.58 (McLean et al., 2011), OCD 0.52 (Ruscio et al., 2010), PTSD 0.68 (Olff, 2017; Kessler et al., 2017), borderline PD 0.70 (clinical samples; community surveys are closer to parity, Grant et al., 2008), narcissistic PD 0.32 (Stinson et al., 2008), antisocial PD 0.25 (Compton et al., 2005), adult ADHD 0.38 (American Psychiatric Association, 2022; Fayyad et al., 2007) and autism 0.25 (Loomes et al., 2017). For the control we use 0.65, the approximate share of women among adults receiving psychotherapy (Olsson and Marcus, 2010). These are order-of-magnitude reference points, not ground truth; our main conclusions depend only on their direction and rough size.

Measures. Let $p_{m,\ell,c}$ be the share of women among F/M answers of model m in language ℓ and condition c (cells with fewer than 5 such answers

are dropped: 16 of 2,376, nearly all refusals), and $p_{m,c}$ its mean over languages; r_c is the reference. We report:

- **deviation** $\text{MAD}_m = \frac{1}{11} \sum_c |p_{m,c} - r_c|$;
- **amplification slope** β_m from the per-model regression $\text{logit } p_{m,c} = \alpha_m + \beta_m \text{logit } r_c$ over the 11 conditions (shares clipped to $[0.01, 0.99]$); $\beta_m = 1$ means sex ratios are reproduced in proportion, $\beta_m > 1$ that they are exaggerated;
- **rank agreement** ρ_m , the Spearman correlation between $p_{m,\cdot}$ and $r_\cdot$;
- **polarisation** $\frac{1}{11} \sum_c |p_{m,c} - 0.5|$, compared with the same quantity for r .

Models are the unit of inference: we test per-condition shifts with Wilcoxon signed-rank tests over models and give bootstrap 95% CIs over models. With 50 draws, the standard error of a single cell is at most 7 percentage points (pp).

4 Results

4.1 Right direction, exaggerated size

Table 1 and Figure 2 give the central result. Model shares follow the reference ordering closely (median $\rho = 0.79$, range 0.23–0.95), so models “know” that anxiety and borderline PD are diagnosed more often in women and antisocial PD more often in men. But they overshoot in both directions. The control prompt, which names no disorder, yields 86% women against a reference of about 65%; borderline PD 99% (range across models 84–100%) against about 70%; GAD 93% against 67%. At the other end, antisocial PD yields 5% women against about 25%, and narcissistic PD 21% against 32%. All 54 models

Condition	Ref.	Models [CI]	Range	>ref	<i>p</i>
<i>Female-skewed in the reference</i>					
Control	65	86 [82–89]	34–100	50/54	<0.001
Borderline PD	70	99 [98–99]	84–100	54/54	<0.001
Depression	67	79 [74–83]	12–100	44/54	<0.001
GAD	67	93 [91–95]	66–100	53/54	<0.001
PTSD	68	62 [56–68]	1–96	28/54	0.50
Social anxiety	58	80 [75–84]	6–100	51/54	<0.001
<i>Near parity</i>					
OCD	52	69 [63–75]	2–98	45/54	<0.001
<i>Male-skewed in the reference</i>					
ADHD	38	50 [43–56]	0–91	36/54	0.003
Narcissistic PD	32	21 [16–26]	0–78	13/54	<0.001
Autism	25	45 [38–51]	0–86	42/54	<0.001
Antisocial PD	25	5 [3–8]	0–49	4/53	<0.001

Table 1: Share of women (%) per condition: approximate reference value, mean of per-model shares (each the mean of 4 languages $\times$ 50 draws) with a bootstrap 95% CI over models, the range across models, the number of models above the reference, and a Wilcoxon signed-rank test of model shares against the reference. Every female-skewed condition except PTSD is pushed further towards women, and both male-skewed personality disorders further towards men. ADHD and autism are moved towards parity instead.

are more polarised than the reference (mean distance from parity 0.33 vs 0.16), and the amplification slope exceeds 1 for all 54 models (median $\beta = 2.39$, IQR 2.09–2.95, bootstrap 95% CI of the median 2.17–2.59).

Two conditions break the pattern. PTSD is the only female-skewed condition that is not pushed further towards women (62% vs 68%, $p = 0.50$), with very wide disagreement between models (1–96%); in English, paramedic is the most frequent occupation for both genders. ADHD and autism, which are diagnosed more often in men, are moved towards parity (50% and 45% women vs 38% and 25%), possibly echoing recent public discussion of under-diagnosed women.

4.2 Collapse of variety

The amplification is not only a shift of the mean. In 54% of the 2,360 model–language–condition cells at least 95% of the 50 draws have the same gender, and in 40% all of them do, even at temperature 0.9. The same collapse shows in names and occupations: across models the median share of distinct first names per model and language is 6%; *Maya* is 34% of all English women and *Marcus* 28% of English men, *Anna* 37% of Polish women and *Marek* 38% of Polish men, *Lucía* 37% of Spanish women. Occupations follow stereotypes: the most frequent English answers are a fe-

male graphic designer for borderline PD, depression, GAD and the control, a male corporate attorney for narcissistic PD, a used-car salesman for antisocial PD and a male software developer for autism. A generator built on such defaults produces a narrow, repetitive set of patients unless diversity is enforced explicitly.

4.3 Newer is not better

Figure 3 plots each API model’s deviation against the date it was listed on OpenRouter. Deviation does not decrease over time; it slightly increases (Spearman $\rho = 0.36$, $p = 0.013$, $n = 47$). The closest models to the reference are Qwen 2.5 72B (MAD 12.1 pp), GPT-5.4 (13.2) and GPT-4o mini (15.0); the farthest are Gemini 3.1 Pro (36.8), GPT-6 Sol (32.8) and GPT-6 Luna (32.4). Within the OpenAI series (Figure 4), the share of women for narcissistic PD fell from 23–55% (GPT-4 to GPT-5) to 6–17% (GPT-5.1 to GPT-5.6 Sol), rose to 48–49% in GPT-6 Luna and Sol (antisocial PD: 32–49%), and fell back to 2–4% for both in GPT-6.1 Sol, which suggests that this behaviour is a by-product of post-training rather than a stable property of a model family.

Gemini 3.1 Pro shows the opposite default to most models: it generates mostly men for depression (12% women), social anxiety (6%) and PTSD (1%), and no women at all for ADHD and autism, while still writing 100% women for borderline PD. Gemini 3.8 Flash and Gemini 4 Aragon share this pattern in weaker form, so a male default is not a guarantee of less bias; it is a different bias. Differences between regions are small (mean MAD: China 20.4, Europe 23.4, USA 24.4 pp; Kruskal–Wallis $p = 0.07$). The seven local models (mean MAD 20.9) are not worse than API models (23.6); the Polish PLLuM 12B (16.9) is among the closest, and Bielik 11B is the only model that matches the reference for antisocial PD (25% women). Gemma 4 31B served locally and through the API gives almost the same profile (Pearson $r = 0.995$ over conditions, mean difference 2.9 pp), so quantisation and serving had little effect for this model.

4.4 Prompt language

The same question asked in Spanish yields 12.5 pp more women than in English (43 of 54 models higher, $p < 0.001$); asked in Chinese, 14.5 pp fewer (41 of 54 lower, $p < 0.001$; Table 2, Figure 5). Polish is not different from English on

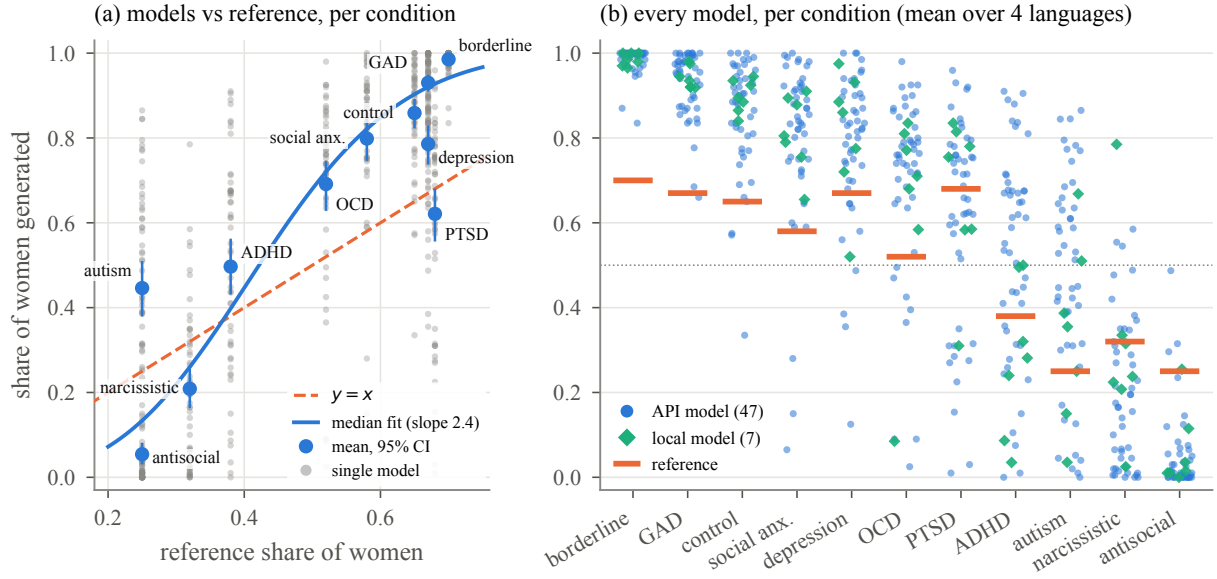


Figure 2: (a) Mean share of women generated (blue, 95% CI over models) against the reference share for each condition; grey points are single models. The dashed line is perfect calibration; the blue curve is the median per-model logistic fit, with slope 2.4 on the log-odds scale: female-skewed conditions are pushed up and male-skewed ones down. (b) Every model in every condition (mean of four languages), with the reference in orange.

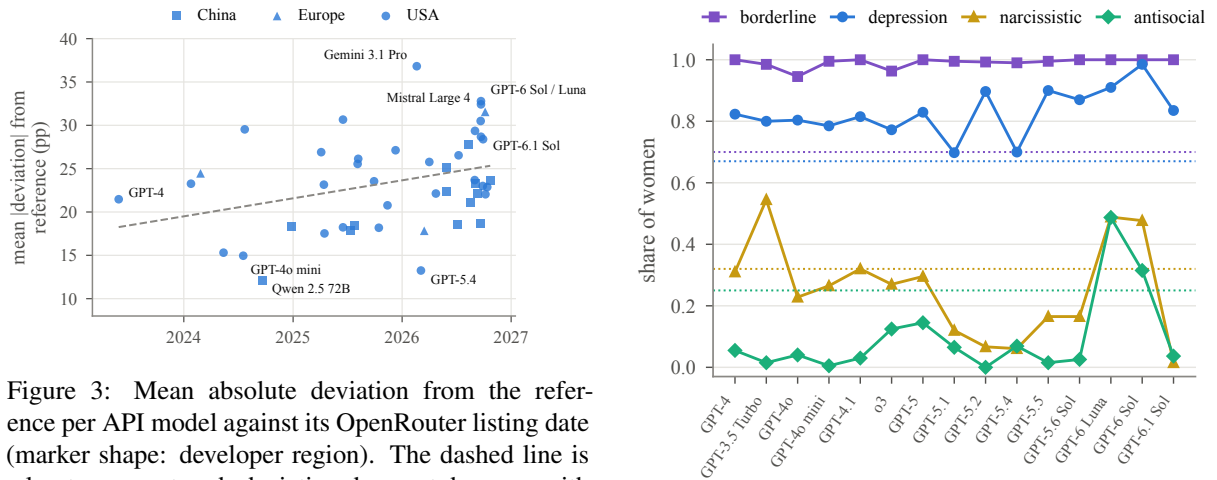


Figure 3: Mean absolute deviation from the reference per API model against its OpenRouter listing date (marker shape: developer region). The dashed line is a least-squares trend: deviation does not decrease with newer models (Spearman $\rho = 0.36$, $p = 0.013$).

Figure 4: OpenAI models in order of OpenRouter listing: share of women for borderline PD, depression, narcissistic and antisocial PD; dotted lines are the references. The personality-disorder defaults move back and forth between consecutive releases.

average (+1.5 pp, $p = 0.70$), although the control prompt, which names no disorder, gives 92% women in Polish and 95% in Spanish against 83% in English. Grammatical gender therefore cannot be the whole explanation: Polish and Spanish share the feminine *osoba/persona* but diverge, while the language effect in Chinese runs towards men for nearly every condition. The language effect is largest where models disagree most (PTSD, OCD, ADHD, autism) and nearly absent at the extremes (borderline PD, antisocial PD), where all languages saturate. Non-binary or other genders were rare (1.2% of answers) and almost only in

English (4.8% of English answers vs $\leq 0.1\%$ in the other languages), most often for ADHD and autism.

4.5 Refusals

Refusals are rare overall (1.0%) but concentrate on personality disorders: 3.7% of answers for antisocial PD and 2.3% for narcissistic PD, against 0.1% for the control. Mistral Large 4 refused 32% of all prompts and 99% of antisocial PD prompts,

Lang.	All	Contr.	ASPD	Δ vs EN	p
English	63	83	8	—	—
Polish	65	92	1	$+1.5 \pm 16.8$	0.70
Spanish	75	95	11	$+12.5 \pm 14.6$	<0.001
Chinese	48	73	2	-14.5 ± 22.5	<0.001

Table 2: Share of women (%) by prompt language: all conditions, the control prompt and antisocial PD. Δ : per-model difference from English (pp, mean $\pm$ SD over 54 models, averaged over conditions) with a Wilcoxon test. Spanish prompts yield more women and Chinese prompts fewer; Polish does not differ from English on average despite the grammatically feminine *osoba*.

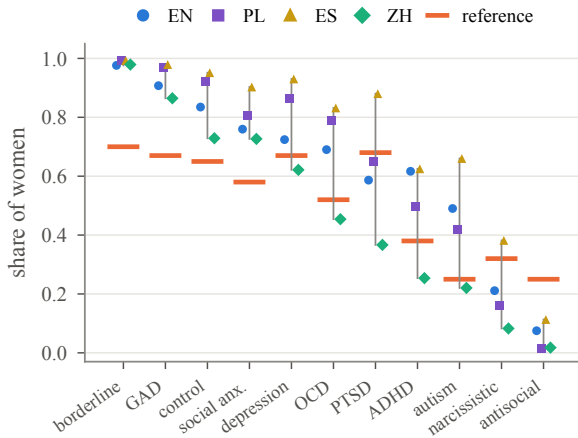


Figure 5: Share of women by prompt language and condition (mean over 54 models), with the reference in orange. Spanish prompts sit highest and Chinese prompts lowest in almost every condition.

explaining that a character “defined by” the disorder would reinforce stigma; GPT-5.2 refused 6% overall and 29% for antisocial PD. Refusing removes the most male-skewed cells from the data of exactly those models, which is why we report refusals separately rather than dropping them.

5 Discussion

What the models seem to learn. The high rank agreement and the slope well above 1 fit a simple story: models learn the direction of sex differences from text, where they are stated as facts (“more common in women”), and then generate the *modal* person rather than sampling from the conditional distribution. This is the same mode-seeking that shows in names (a few names dominate) and occupations, and it survives temperature 0.9. A ratio of 7:3 in the population becomes 99:1 in the output.

Why it matters for clinical material. Generated vignettes and simulated patients are used to train attention. A trainee who only ever meets women with borderline PD and men with antisocial PD, or a decision-support evaluation set built the same way, inherits a stronger prior than the epidemiology supports, in conditions where over-diagnosis in women (borderline) and under-diagnosis in women (autism, ADHD, antisocial PD) are documented clinical problems. Developers of such tools should specify gender (and other demographics) explicitly from a target distribution instead of leaving it to the model, and should check generated cohorts against reference ratios.

No free lunch from newer or local models. Model choice moves deviation from 12 to 37 pp, but neither recency, region, size nor local deployment predicts it. Post-training appears to move these defaults back and forth between consecutive releases (§4.3), so the measurement has to be repeated for each model version that is deployed.

6 Limitations

- **One prompt template.** Results describe default behaviour for a short zero-shot request without system prompt or clinical context; deployed applications may differ. We did not vary wording within a language.
- **Approximate reference values.** Sex ratios differ by country, setting and diagnostic practice; borderline PD in particular is female-skewed in clinical samples but closer to parity in community surveys. Using a parity reference for it would make the observed 99% look even more extreme.
- **Grammatical gender and translation.** The Polish and Spanish prompts contain a feminine noun for “person”; disorder names were translated by us and may carry different connotations in each language.
- **API routing.** OpenRouter may route a model to different providers; the provider of every call is recorded in the dataset. Model names and listing dates are those of OpenRouter, which may differ from official release dates.
- **Binary analysis.** Shares are computed over F/M answers; X answers (1.2%) are reported separately.
- **Rule-based parsing.** The parser was audited per model, but a small number of misread fields

cannot be excluded.

7 Conclusion

Across 54 LLMs and four languages, models know which way mental disorders skew by sex but exaggerate the skew about 2.4-fold on the log-odds scale, collapse within-condition variety, and do not improve with newer releases. Anyone generating clinical cases or simulated patients with LLMs should set demographics explicitly rather than inherit these defaults.

8 Data and Code Availability

All 118,800 responses with prompts, parsed fields, provider and timestamps, per-model settings, reference values, summaries, figures and the full collection and analysis code are released under CC BY 4.0 at <https://doi.org/10.5281/zenodo.23676720> and <https://huggingface.co/datasets/therapy-support/llm-gender-bias-mental-health>.

Ethics Statement

No human participants or personal data were involved; all people in the dataset are fictional and generated by models. Some generated names may coincide with real people by chance. The study measures a bias in order to help developers avoid it; the released outputs contain stereotyped content and should not be used as training data without filtering. Writing assistance: LLM tools were used for code and for editing the text; all analyses and claims were checked by the authors.

References

- American Psychiatric Association. 2022. *Diagnostic and Statistical Manual of Mental Disorders*, 5th, text revision edition. American Psychiatric Association Publishing, Washington, DC.
- Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. *Out of one, many: Using language models to simulate human samples*. *Political Analysis*, 31(3):337–351.
- Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.
- Ayoub Bouguettaya, Elizabeth M. Stuart, and Elias Aboujaoude. 2025. *Racial bias in AI-mediated psychiatric diagnosis and treatment: a qualitative comparison of four large language models*. *npj Digital Medicine*, 8(1).
- Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. *Semantics derived automatically from language corpora contain human-like biases*. *Science*, 356(6334):183–186.
- Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. *Marked Personas: Using natural language prompts to measure stereotypes in language models*. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1504–1532. Association for Computational Linguistics.
- Wilson M. Compton, Kevin P. Conway, Frederick S. Stinson, James D. Colliver, and Bridget F. Grant. 2005. *Prevalence, correlates, and comorbidity of DSM-IV antisocial personality syndromes and alcohol and specific drug use disorders in the United States: Results from the National Epidemiologic Survey on Alcohol and Related Conditions*. *The Journal of Clinical Psychiatry*, 66(06):677–685.
- Joshua J Docking, Lee X Li, Bradley D Menz, Stephen Bacchi, Ashley M Hopkins, and Michael J Sorich. 2026. *Evaluating the potential of reasoning large language models to perpetuate racial and gender disease stereotypes in Health Care*. *Journal of Medical Internet Research*, 28:e82256.
- John Fayyad, Ron De Graaf, Ronald Kessler, Jordi Alonso, Matthias Angermeyer, Koen Demyttenaere, Giovanni De Girolamo, Josep Maria Haro, Elie G. Karam, Carmen Lara, Jean-Pierre Lépine, Johan Ormel, José Posada-Villa, Alan M. Zaslavsky, and Robert Jin. 2007. *Cross-national prevalence and correlates of adult attention-deficit hyperactivity disorder*. *British Journal of Psychiatry*, 190(5):402–409.
- Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Deroncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. *Bias and fairness in large language models: A survey*. *Computational Linguistics*, 50(3):1097–1179.
- Bridget F. Grant, S. Patricia Chou, Rise B. Goldstein, Boji Huang, Frederick S. Stinson, Tulshi D. Saha, Sharon M. Smith, Deborah S. Dawson, Attila J. Pulay, Roger P. Pickering, and W. June Ruan. 2008. *Prevalence, correlates, disability, and comorbidity of DSM-IV borderline personality disorder: Results from the wave 2 National Epidemiologic Survey on Alcohol and Related Conditions*. *The Journal of Clinical Psychiatry*, 69(4):533–545.
- Ronald C. Kessler, Sergio Aguilar-Gaxiola, Jordi Alonso, Corina Benjet, Evelyn J. Bromet, Graça Cardoso, Louisa Degenhardt, Giovanni de Girolamo, Romyana V. Dinolova, Finola Ferry, Silvia

- Florescu, Oye Gureje, Josep Maria Haro, Yueqin Huang, Elie G. Karam, Norito Kawakami, Sing Lee, Jean-Pierre Lepine, Daphna Levinson, and 13 others. 2017. [Trauma and PTSD in the WHO World Mental Health Surveys](#). *European Journal of Psychotraumatology*, 8(sup5).
- Jan Kocoń, Maciej Piasecki, Arkadiusz Janz, Teddy Ferdinan, Łukasz Radliński, Bartłomiej Koptyra, Marcin Oleksy, Stanisław Woźniak, et al. 2025. PLLuM: A family of Polish large language models. *arXiv preprint arXiv:2511.03823*.
- Hadas Kotek, Rikker Dockum, and David Sun. 2023. [Gender bias and stereotypes in large language models](#). In *Proceedings of The ACM Collective Intelligence Conference*, pages 12–24. ACM.
- Inna Lin, Lucille Njoo, Anjalie Field, Ashish Sharma, Katharina Reinecke, Tim Althoff, and Yulia Tsvetkov. 2022. [Gendered mental health stigma in masked language models](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2152–2170. Association for Computational Linguistics.
- Rachel Loomes, Laura Hull, and William Polmear Locke Mandy. 2017. [What is the male-to-female ratio in Autism Spectrum Disorder? a systematic review and meta-analysis](#). *Journal of the American Academy of Child & Adolescent Psychiatry*, 56(6):466–474.
- Carmen P. McLean, Anu Asnaani, Brett T. Litz, and Stefan G. Hofmann. 2011. [Gender differences in anxiety disorders: Prevalence, course of illness, comorbidity and burden of illness](#). *Journal of Psychiatric Research*, 45(8):1027–1035.
- Jared Moore, Declan Grabb, William Agnew, Kevin Klyman, Stevie Chancellor, Desmond C. Ong, and Nick Haber. 2025. [Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers](#). In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 599–627. ACM.
- Krzysztof Ociepa, Łukasz Flis, Krzysztof Wróbel, Adrian Gwoździej, and Remigiusz Kinas. 2025. Bielik 11B v2 technical report. *arXiv preprint arXiv:2505.02410*.
- Miranda Olf. 2017. [Sex and gender differences in post-traumatic stress disorder: an update](#). *European Journal of Psychotraumatology*, 8(sup4).
- Mark Olfson and Steven C. Marcus. 2010. [National trends in outpatient psychotherapy](#). *American Journal of Psychiatry*, 167(12):1456–1463.
- Jesutofunmi A. Omiye, Jenna C. Lester, Simon Spichak, Veronica Rotemberg, and Roxana Daneshjou. 2023. [Large language models propagate race-based medicine](#). *npj Digital Medicine*, 6(1).
- Stephen R. Pfohl, Heather Cole-Lewis, Rory Sayres, Darlene Neal, Mercy Asiedu, Awa Dieng, Nenad Tomasev, Qazi Mamunur Rashid, Shekoofeh Azizi, Negar Rostamzadeh, Liam G. McCoy, Leo Anthony Celi, Yun Liu, Mike Schaekermann, Alanna Walton, Alicia Parrish, Chirag Nagpal, Preeti Singh, Akeiyah Dewitt, and 11 others. 2024. [A toolbox for surfacing health equity harms and biases in large language models](#). *Nature Medicine*, 30(12):3590–3600.
- A M Ruscio, D J Stein, W T Chiu, and R C Kessler. 2010. [The epidemiology of obsessive-compulsive disorder in the National Comorbidity Survey Replication](#). *Molecular Psychiatry*, 15(1):53–63.
- Rachel H. Salk, Janet S. Hyde, and Lyn Y. Abramson. 2017. [Gender differences in depression in representative national samples: Meta-analyses of diagnoses and symptoms](#). *Psychological Bulletin*, 143(8):783–822.
- Randy A. Sansone and Lori A. Sansone. 2011. Gender patterns in borderline personality disorder. *Innovations in Clinical Neuroscience*, 8(5):16–20. PMID: 21686143; PMCID: PMC3115767.
- Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cino Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR.
- Karolina Stanczak and Isabelle Augenstein. 2021. A survey on gender bias in natural language processing. *arXiv preprint arXiv:2112.14168*.
- Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. [Evaluating gender bias in machine translation](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684. Association for Computational Linguistics.
- Frederick S. Stinson, Deborah A. Dawson, Rise B. Goldstein, Patricia Chou, Boji Huang, Sharon M. Smith, W. June Ruan, Attila J. Pulay, Tulshi D. Saha, Roger P. Pickering, and Bridget F. Grant. 2008. [Prevalence, correlates, disability, and comorbidity of DSM-IV narcissistic personality disorder: Results from the wave 2 National Epidemiologic Survey on Alcohol and Related Conditions](#). *The Journal of Clinical Psychiatry*, 69(7):1033–1045.
- Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. [“Kelly is a warm person, Joseph is a role model”: Gender biases in LLM-generated reference letters](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3730–3748. Association for Computational Linguistics.

- Ruiyi Wang, Stephanie Milani, Jamie C. Chiu, Jiayin Zhi, Shaun M. Eack, Travis Labrum, Samuel M Murphy, Nev Jones, Kate V Hardy, Hong Shen, Fei Fang, and Zhiyu Chen. 2024. [PATIENT- \$\psi\$: Using large language models to simulate patients for training mental health professionals](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12772–12797. Association for Computational Linguistics.
- Travis Zack, Eric Lehman, Mirac Suzgun, Jorge A Rodriguez, Leo Anthony Celi, Judy Gichoya, Dan Jurafsky, Peter Szolovits, David W Bates, Raja-Elie E Abdounour, Atul J Butte, and Emily Alsentzer. 2024. [Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study](#). *The Lancet Digital Health*, 6(1):e12.
- Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. 2019. [Examining gender bias in languages with grammatical gender](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5275–5283. Association for Computational Linguistics.

A Additional Results

#	Model	Developer	Reg.	Listed	MAD	Slope	ρ	Refusal	Control	Depr.	Borderline	Antisocial
1	Qwen 2.5 72B	Alibaba	CN	2024-09-19	12.1	2.16	0.78	0.1	69	74	87	1
2	GPT-5.4	OpenAI	US	2026-03-05	13.2	2.32	0.85	1.6	92	70	99	7
3	GPT-4o mini	OpenAI	US	2024-07-18	15.0	2.40	0.90	0.0	81	78	100	0
4	GPT-4o	OpenAI	US	2024-05-13	15.3	1.83	0.81	0.4	74	80	95	4
5	PLLuM 12B chat [†]	PLLuM (CYFRAGOVPL)	EU	–	16.9	2.38	0.83	0.0	84	78	97	2
6	Cogito 70B [†]	DeepCogito	US	–	17.4	2.78	0.84	0.0	86	86	100	1
7	o3	OpenAI	US	2025-04-16	17.5	1.42	0.68	2.4	76	77	96	12
8	Mistral Small 26.03	Mistral	EU	2026-03-16	17.8	1.95	0.64	0.1	57	49	97	0
9	Kimi K2	Moonshot	CN	2025-07-11	17.9	1.61	0.86	0.7	76	84	95	12
10	Qwen 3.6 27B [†]	Qwen	CN	–	18.1	2.17	0.82	0.1	94	88	100	12
11	Claude Haiku 4.5	Anthropic	US	2025-10-15	18.2	3.38	0.78	2.0	66	58	100	0
12	Gemini 2.5 Pro	Google	US	2025-06-17	18.2	2.61	0.67	0.0	77	64	100	0
13	DeepSeek V3	DeepSeek	CN	2024-12-26	18.3	2.39	0.82	0.0	86	84	100	0
14	GLM 4.5	Zhipu	CN	2025-07-25	18.5	3.18	0.86	0.0	70	75	100	0
15	Hunyuan 3	Tencent	CN	2026-07-06	18.6	3.01	0.75	0.0	98	74	100	0
16	MiMo 2.6 Pro	Xiaomi	CN	2026-09-21	18.7	2.10	0.82	0.0	85	82	98	5
17	Qwen3 8B [†]	Qwen	CN	–	19.0	3.13	0.84	0.5	92	72	98	0
18	GPT-5.1	OpenAI	US	2025-11-13	20.8	1.88	0.68	0.0	83	70	100	6
19	GLM 5.3	Zhipu	CN	2026-08-18	21.1	1.88	0.68	0.2	77	84	100	8
20	Bielik 11B v3 [†]	SpeakLeash	EU	–	21.3	1.70	0.85	4.6	89	93	97	25
21	GPT-4	OpenAI	US	2023-05-28	21.5	2.09	0.76	0.1	82	82	100	6
22	Claude Haiku 5.5	Anthropic	US	2026-10-07	22.0	2.14	0.84	0.0	90	96	100	24
23	DeepSeek V4.1	DeepSeek	CN	2026-09-10	22.1	2.30	0.59	0.0	97	56	97	0
24	GPT-5.5	OpenAI	US	2026-04-24	22.1	2.80	0.88	0.0	94	90	100	2
25	Step 3.7 Flash	StepFun	CN	2026-05-28	22.4	2.17	0.74	0.5	97	89	97	8
26	Gemini 4 Argon	Google	US	2026-10-13	22.9	2.09	0.55	0.0	57	38	96	1
27	Claude Sonnet 5.5	Anthropic	US	2026-09-28	23.0	2.64	0.59	0.1	82	64	100	0
28	GPT-4.1	OpenAI	US	2025-04-14	23.2	2.17	0.74	0.0	90	82	100	3
29	GPT-3.5 Turbo	OpenAI	US	2024-01-25	23.3	1.83	0.78	0.0	80	80	98	2
30	Qwen 3.8 Max	Alibaba	CN	2026-09-03	23.3	3.67	0.83	0.0	84	76	100	0
31	Claude Sonnet 4.5	Anthropic	US	2025-09-29	23.5	3.91	0.89	0.0	92	92	100	0
32	Kimi K3.1 Pro	Moonshot	CN	2026-10-23	23.6	3.08	0.74	0.8	90	86	100	0
33	Claude Fable 5.1	Anthropic	US	2026-09-01	23.7	1.75	0.57	0.0	94	68	84	0
34	Mistral Large (2024)	Mistral	EU	2024-02-26	24.5	3.65	0.82	0.0	100	100	100	0
35	MiniMax M3	MiniMax	CN	2026-05-31	25.1	2.11	0.90	0.0	88	86	100	6
36	Claude Opus 4.1	Anthropic	US	2025-08-05	25.5	3.70	0.76	0.2	97	100	100	0
37	Gemma 4 31B	Google	US	2026-04-02	25.8	3.34	0.79	0.0	91	64	100	0
38	GPT-5	OpenAI	US	2025-08-07	26.1	1.72	0.59	0.0	88	83	100	14
39	Gemma 4 31B [†]	Google	US	–	26.4	3.22	0.80	0.0	94	52	100	1
40	GPT-5.6 Sol	OpenAI	US	2026-07-09	26.5	2.65	0.85	0.0	94	87	100	3
41	Llama 4 Maverick	Meta	US	2025-04-05	26.9	3.28	0.79	0.0	100	100	100	2
42	GPT-5.2	OpenAI	US	2025-12-10	27.1	3.00	0.78	6.2	91	90	99	0
43	Hermes 3 8B [†]	Nous	US	–	27.1	2.40	0.85	0.1	88	98	98	4
44	Seed 2.1 Turbo	ByteDance	CN	2026-08-12	27.8	3.34	0.87	1.0	99	96	100	1
45	GPT-6.1 Sol	OpenAI	US	2026-09-29	28.4	2.47	0.66	0.0	88	84	100	4
46	Claude Opus 5.5	Anthropic	US	2026-09-22	28.7	2.52	0.62	0.0	96	74	96	0
47	Gemini 3.8 Flash	Google	US	2026-09-02	29.4	2.49	0.67	0.0	65	36	100	0
48	Llama 3.1 70B	Meta	US	2024-07-23	29.5	1.99	0.95	2.0	93	94	100	30
49	Grok 4.7	xAI	US	2026-09-21	30.5	2.79	0.74	0.0	96	94	100	1
50	Gemini 2.5 Flash	Google	US	2025-06-17	30.7	2.35	0.81	0.0	98	94	99	3
51	Mistral Large 4	Mistral	EU	2026-10-06	31.6	1.29	0.54	32.0	98	90	100	–
52	GPT-6 Luna	OpenAI	US	2026-09-22	32.4	1.47	0.84	0.0	94	91	100	49
53	GPT-6 Sol	OpenAI	US	2026-09-22	32.8	2.26	0.82	0.0	100	98	100	32
54	Gemini 3.1 Pro	Google	US	2026-02-19	36.8	2.56	0.82	0.0	33	12	100	0

Table 3: All 54 models, ordered by mean absolute deviation from the reference (MAD, pp; lower is closer). Listed: date the model was listed on OpenRouter (local models: –). Slope: per-model regression slope of logit(model share) on logit(reference share) across the 11 conditions (1 = no amplification). ρ : Spearman rank correlation with the reference ordering. Refusal: share of responses with no readable gender (%). Last four columns: share of women (%). [†]Run locally (quantised, Ollama/llama.cpp).

	control	depression	GAD	social anx.	OCD	PTSD	borderline	narcissistic	antisocial	ADHD	autism
reference	65	67	67	58	52	68	70	32	25	38	25
Qwen 2.5 72B	69	74	84	74	74	62	87	21	1	36	17
GPT-5.4	92	70	88	65	53	62	99	6	7	45	26
GPT-4o mini	81	78	84	75	72	75	100	26	0	41	39
GPT-4o	74	80	86	82	68	66	95	23	4	54	39
PLLuM 12B (local)	84	78	92	80	68	76	97	21	2	28	39
Cogito 70B (local)	86	86	94	79	84	78	100	32	1	32	25
o3	76	77	87	80	70	53	96	27	12	61	53
Mistral Small 26.03	57	49	78	59	42	15	97	6	0	26	31
Kimi K2	76	84	90	83	72	74	95	20	12	47	61
Qwen 3.6 27B (local)	94	88	94	76	71	58	100	22	12	50	36
Claude Haiku 4.5	66	58	66	59	47	31	100	1	0	1	2
Gemini 2.5 Pro	77	64	86	76	40	28	100	0	0	44	31
DeepSeek V3	86	84	84	83	78	76	100	35	0	50	42
GLM 4.5	70	75	98	60	50	31	100	4	0	23	4
Hunyuan 3	98	74	100	88	76	72	100	37	0	31	32
MiMo 2.6 Pro	85	82	89	77	77	69	98	14	5	55	45
Qwen3 8B (local)	92	72	92	88	58	58	98	34	0	9	4
GPT-5.1	83	70	92	81	74	62	100	12	6	62	64
GLM 5.3	77	84	96	85	74	60	100	20	8	41	77
Bielik 11B v3 (local)	89	93	92	89	77	81	97	24	25	50	67
GPT-4	82	82	95	91	82	75	100	31	6	65	51
Claude Haiku 5.5	90	96	98	89	86	78	100	35	24	66	44
DeepSeek V4.1	97	56	95	75	64	62	97	10	0	68	58
GPT-5.5	94	90	96	86	76	80	100	16	2	53	40
Step 3.7 Flash	97	89	95	92	86	76	97	29	8	44	59
Gemini 4 Argon	57	38	87	28	36	22	96	0	1	28	38
Claude Sonnet 5.5	82	64	98	64	72	28	100	1	0	68	45
GPT-4.1	90	82	98	91	78	62	100	32	3	71	59
GPT-3.5 Turbo	80	80	84	84	80	78	98	55	2	70	66
Qwen 3.8 Max	84	76	94	71	66	27	100	0	0	15	1
Claude Sonnet 4.5	92	92	96	87	64	66	100	0	0	10	4
Kimi K3.1 Pro	90	86	99	92	75	52	100	4	0	59	30
Claude Fable 5.1	94	68	94	82	66	35	84	16	0	81	61
Mistral Large (2024)	100	100	100	99	76	74	100	38	0	61	13
MiniMax M3	88	86	90	72	87	92	100	18	6	68	70
Claude Opus 4.1	97	100	100	100	79	60	100	2	0	34	43
Gemma 4 31B	91	64	100	72	9	32	100	2	0	8	14
GPT-5	88	83	98	93	85	75	100	30	14	84	78
Gemma 4 31B (local)	94	52	98	66	8	31	100	2	1	4	15
GPT-5.6 Sol	94	87	100	90	79	84	100	16	3	67	63
Llama 4 Maverick	100	100	100	100	98	72	100	58	2	31	42
GPT-5.2	91	90	99	98	88	71	99	7	0	68	55
Hermes 3 8B (local)	88	98	98	91	81	84	98	78	4	24	51
Seed 2.1 Turbo	99	96	100	92	80	82	100	7	1	65	53
GPT-6.1 Sol	88	84	100	84	74	76	100	2	4	91	74
Claude Opus 5.5	96	74	97	91	79	84	96	2	0	67	86
Gemini 3.8 Flash	65	36	87	15	9	3	100	0	0	24	5
Llama 3.1 70B	93	94	98	94	92	96	100	55	30	72	68
Grok 4.7	96	94	99	97	92	80	100	21	1	90	64
Gemini 2.5 Flash	98	94	100	87	90	87	99	35	3	84	84
Mistral Large 4	98	90	97	97	86	57	100	42	–	83	84
GPT-6 Luna	94	91	94	98	90	92	100	49	49	88	80
GPT-6 Sol	100	98	100	99	92	93	100	48	32	89	77
Gemini 3.1 Pro	33	12	68	6	2	1	100	0	0	0	0

Figure 6: Share of women (%) for every model and condition (mean of four languages), ordered by deviation from the reference (top row). Pink: more women; blue: more men; “–”: fewer than 5 gendered answers in every language (refusals).