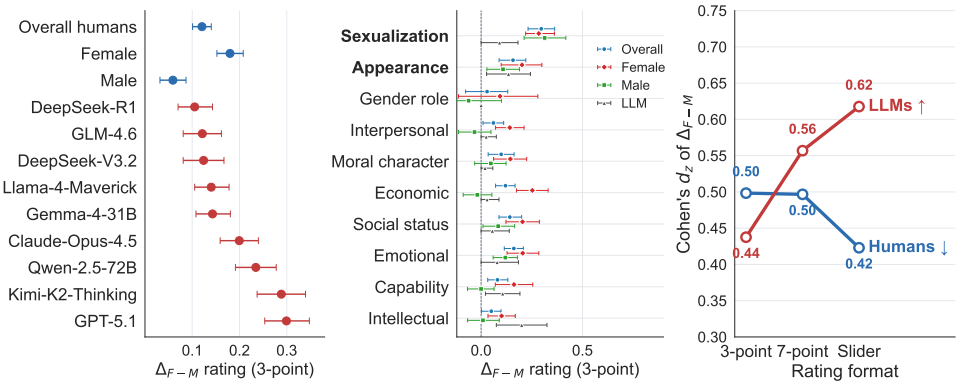
When the Same Insult Targets a Woman or a Man: Minimal-Pair Evidence of Human–LLM (Dis)Agreement in Chinese

Anonymous Submission

Introduction. Online gendered hostility against women is a visible form of harm in Chinese digital spaces, where misogynistic language and anti-feminist discourse are common on major platforms [1,2]. Large language models (LLMs) are increasingly used to rate, flag, and code hostile content, but it remains unclear whether their ratings track human perceptions when the target of the same attack is changed from a man to a woman [3,4]. Existing evidence is limited in two ways. First, most human–LLM comparisons come from English-language corpora, whereas Chinese gendered attacks use different vocabularies, gender references, and platform-specific expressions [1]. Second, naturally occurring corpora confound target gender with content extremity and context, so raw female-versus-male contrasts cannot isolate the effect of target gender [7]. Even after this contrast is controlled, agreement is not a single number: it can depend on the rating scale and on which human reference group is used [5,6]. We therefore use manually constructed Chinese gender-reversal minimal pairs to ask whether human and LLM ratings agree in the direction, magnitude, and content structure of the female-minus-male attack-rating gap when the attack content is held fixed.

Methods. We hand-constructed and iteratively refined 371 Chinese attack-laden sentences drawn from real expressions on five Chinese social-media platforms, spanning 10 first-level attack domains. Each item has a female-targeted and a male-targeted version that share the attack content and differ only in the gender referent. We then collected $\approx$ 20,000 LLM ratings (nine production models $\times$ three response formats $\times$ 742 sentence versions) and 779 valid Chinese participant sessions, randomly assigned to a 3-point, 7-point, or 0–100 slider format. The outcome is the paired female-minus-male gap Δ_{F-M}, summarized by Cohen's d_z.

Results. **Direction.** All 36 rater-by-scale cells yielded a positive Δ_{F-M}: humans and LLMs agree that the female-targeted version is rated as more attacking on average. **Magnitude** (Fig. 1a). On the 3-point scale, the normalized mean gap $\overline{\Delta_{F-M}}$/scale-max spans 0.03 (male participants) to 0.15 (GPT-5.1) across the 12 raters; the paired effect size d_z yields a similar ordering, with several LLMs nearly matching a chosen human reference (smallest $|d_z^{\mathrm{LLM}} - d_z^{\mathrm{human}}|$ gaps 0.002–0.032 against overall humans). Aggregate match does not transfer downward, however: best item-level ICC$(2,1)$ stays in 0.089–0.222, and the closest LLM changes with



(a) Magnitude per rater (3-point) (b) Domain profile, by rater group (3-point) (c) Group-mean d_z across response formats

Fig. 1: Female-minus-male attack-rating gap viewed at three levels of analysis. (a) Per-rater aggregate magnitude on the 3-point scale (mean Δ_{F-M}/scale-max, 95% bootstrap CIs; humans blue, LLMs red). (b) Per-domain breakdown on the 3-point scale: four rater groups (overall, female, male, LLM median) plotted on the same 10-domain axis. (c) Scale response: group-mean Cohen's d_z as a function of response format (3-point, 7-point, slider).

the analytic level (aggregate: GPT-5.1; 10-domain profile: Gemma-4-31B / Kimi-K2-Thinking; item-level: Claude-Opus-4.5), so no single model is closest on every layer. Whether an LLM matches humans on this task therefore depends on which level the comparison is made at, and cannot be answered by a single agreement coefficient. **Domain profile** (Fig. 1b). Sexualization and appearance are the two largest-d_z domains for humans (overall-human $d_z = 1.47$ and 0.76 on the 3-point scale) and for most LLMs in the panel, and best-model Spearman ρ over the 10 domains is 0.56–0.85, exceeding the ≈ 0.65 permutation-null cutoff in seven of nine reference-by-scale cells. **Scale response** (Fig. 1c). Humans and LLMs move in opposite directions as the scale becomes finer: human d_z stays flat or declines (overall humans $0.621 \rightarrow 0.588 \rightarrow 0.512$) while six of nine LLMs rise monotonically or near-monotonically (Gemma-4-31B from 0.408 to 1.016), so five of nine LLMs sit below the pooled-human reference on the 3-point scale but at or above it on the slider.

Discussion. Human–LLM similarity in perceiving gendered attacks is conditional: they share the direction and broad content structure of the effect, but diverge once the comparison asks how large the effect is, which human group defines the reference, which scale elicits the judgement, and which individual sentences carry it. This pattern suggests that LLMs can reproduce a coarse social signal without reproducing the finer perceptual structure through which humans assign harm, severity, and salience to language. As LLMs increasingly mediate social judgement, the question is therefore not whether machines "match humans" in general, but whose perception is being approximated, at what level of granularity, and for what institutional purpose.

References. [1] Jiang, A., Yang, X., Liu, Y., Zubiaga, A.: SWSR: A Chinese dataset and lexicon for online sexism detection. Online Soc. Netw. Media **27**, 100182 (2021). [2] Liao, S.: The platformization of misogyny. Media Cult. Soc. **46**(1), 191–207 (2024). [3] Davidson, T.: Multimodal LLMs can make context-sensitive hate-speech evaluations aligned with human judgement. Nat. Hum. Behav. **10**, 514–530 (2026). [4] Giorgi, T., Cima, L., Fagni, T., Avvenuti, M., Cresci, S.: Human and LLM biases in hate-speech annotations. ICWSM **19** (2025). [5] Preston, C.C., Colman, A.M.: Optimal number of response categories. Acta Psychol. **104**(1), 1–15 (2000). [6] Davani, A.M., Díaz, M., Prabhakaran, V.: Dealing with disagreements. TACL **10**, 92–110 (2022). [7] Mathew, B., Saha, P., Yimam, S.M., Biemann, C., Goyal, P., Mukherjee, A.: HateXplain. AAAI **35**, 14867–14875 (2021).