Gender Disparities in LLM-Based Intimate Partner Violence Detection

Tabia Tanzin Prama[1,2], Mikaela Irene Fudolig[3], Abigail M. Crocker[4], Christopher M. Danforth[1,4], Peter Sheridan Dodds[1,2,5,6]

[1]Computational Story Lab, Vermont Complex Systems Institute, Vermont Advanced Computing Center, University of Vermont, Burlington, VT 05405, USA
[2]Department of Computer Science, University of Vermont, Burlington, VT 05405, USA
[3]School of Mathematical Sciences, Adelaide University, Adelaide, Australia
[4]Department of Mathematics and Statistics, University of Vermont, Burlington, VT 05405, USA
[5]Santa Fe Institute, 1399 Hyde Park Rd, Santa Fe, NM 87501, USA
[6]Complexity Science Hub, Metternichgasse 8, 1030 Vienna, Austria

Abstract

Intimate Partner Violence (IPV) is a major public health concern, and large language models (LLMs) are increasingly used for support and information-seeking in sensitive domains. We examine whether LLMs perceive relationship abuse differently depending on victim–perpetrator gender configuration. Using 475 Reddit posts from `r/relationship_advice`, we generate counterfactual variants by swapping gendered identifiers to create four dyads: female–female (F/F), female–male (F/M), male–female (M/F), and male–male (M/M), where the first position denotes the victim. Four recent LLMs (GPT-5o, Gemini 3, Llama 4, and Grok 3) evaluate each variant using a structured questionnaire covering IPV, perpetrator intent, cheating, and abuse subtypes. Results show substantial variation across models and dyads. Abuse and intent detection systematically decrease in mixed-gender dyads where the victim is male, with female perpetrator identity emerging as a consistent negative predictor of abuse recognition. Mixed-effects logistic regression confirms that gender roles significantly shape model outputs. Our findings suggest that LLMs reproduce gendered biases from online training data, with implications for support-related deployment. Code and resources are available at GitHub.

1 Introduction

Intimate partner violence (IPV) is a major global public health and human rights concern, defined by the World Health Organization as any behavior by a current or former partner that causes physical, sexual, or psychological harm (World Health Organization and London School of Hygiene and Tropical Medicine, 2010; Heise and Garcia-Moreno, 2002). Globally, approximately one in three women have experienced physical or sexual violence in their lifetime (World Health Organization). While women experience IPV at disproportionately higher rates (Breiding et al., 2008; Schneider et al., 2009), men also experience IPV across both same-sex and other-sex relationships, with severe physical, psychological, and social consequences (Hines and Douglas, 2016; Sivagurunathan et al., 2021b). Beyond immediate harm, IPV is associated with long-term mental health difficulties, substance use, and legal and financial repercussions (Peterson et al., 2018). Exposure to domestic violence more broadly is linked to traumatic brain injuries, chronic pain, PTSD, depression, and suicidal ideation (Choi et al., 2021; Ennis et al., 2021; Kim and Merlo, 2023; Wright et al., 2021). Access to advocacy interventions is therefore a key protective factor for survivors' recovery (Rivas et al., 2019), yet survivors frequently encounter barriers including stigma, shame, and fear of judgment (Naismith et al., 2024; Gilbert and Postel, 2021; Nayak et al., 2023; Lam et al., 2020). These barriers are particularly pronounced for men, whose help-seeking is further constrained by gendered expectations surrounding masculinity (Machado et al., 2016; Park et al., 2020; Walker et al., 2020), leaving male survivors underrepresented in institutional responses.

Digital spaces increasingly serve as alternative venues for support. Platforms such as Reddit[1] provide anonymous environments for peer support, yet prior research has found that male survivors frequently encounter systemic biases across social norms, legal systems, and institutional responses (Sivagurunathan et al., 2021a). Dedicated digital interventions, including mobile applications such as *myPlan* and web-based safety planning tools — have demonstrated promising outcomes for survivors (Storer et al., 2022; Hegarty et al., 2019; Koziol-McLain et al., 2018; Ford-Gilboe et al., 2020), alongside growing use of health information technologies to identify survivors' needs (Hui et al., 2024, 2023). The landscape of online information-

[1]https://www.reddit.com/

Proceedings of the 7th Workshop on Natural Language Processing and Computational Social Science (NLP+CSS), pages 190–197

seeking is now undergoing a major transformation with the rise of large language models (LLMs). Dedicated AI systems such as Aimee[2] and Ruth[3] have been developed specifically to support IPV survivors, with Ruth now recommended by the U.S. National Domestic Violence Hotline. Because LLMs operate continuously without human intervention, they have the potential to bridge gaps in traditional help-seeking pathways (Maeng and Lee, 2021). However, these models are trained on massive corpora of internet text and may absorb and reproduce the biases present in those environments (Prama et al., 2025; Gallegos et al., 2023), potentially replicating differential validation of victims based on gender.

In this study, we examine whether LLMs exhibit gender-based perceptual biases in IPV scenarios, including differences in relationship recognition, abuse detection, and harm assessment across gender dyads.

2 Methodology

Data Collection and Selection. We collected posts from `r/relationship_advice`, a large Reddit community with approximately 16 million members and 60,000 weekly contributions. Because this subreddit includes broad relationship concerns, it captures ambiguous help-seeking narratives in which posters describe unhealthy or abusive behaviors that they may not yet recognize as IPV. We treat these posts as reflecting an early awareness stage of IPV. Following the World Health Organization (World Health Organization and London School of Hygiene and Tropical Medicine, 2010), we define IPV as behavior by a current or former intimate partner that causes physical, sexual, psychological, or economic harm, and we also consider precursor dynamics such as coercive control, emotional manipulation, and isolation.

To support counterfactual gender analysis, we manually retained only posts with exactly two parties, clearly identifiable victim and perpetrator roles, and explicit gender markers for both parties, such as Male (M) or Female (F). Posts involving multiple parties, ambiguous roles, or unspecified gender information were excluded. This process yielded 475 unique dyadic narratives. Throughout the paper, dyads are denoted using original poster (OP)/perpetrator notation: the first position refers to the OP or victim role, and the second refers to the partner or perpetrator role. Thus, M/F denotes a male OP/victim and a female perpetrator. The original dyad distribution is 29 Female–Female (F/F), 228 Female–Male (F/M), 202 Male–Female (M/F), and 16 Male–Male (M/M).

Counterfactual Data Generation. To isolate gender while preserving the underlying relationship narrative, we used a counterfactual gender-swapping procedure. For each original post, we generated four versions of the same narrative, corresponding to all possible OP/perpetrator gender configurations: Male–Male (M/M), Male–Female (M/F), Female–Male (F/M), and Female–Female (F/F). Thus, every post appears once in each dyad condition, regardless of its original gender configuration. We programmatically updated gender-identifying markers, including pronouns (he/she, him/her), familial roles (uncle/aunt, brother/sister), names when applicable, and explicit gender tags. This process produced a final evaluation dataset of 1,900 samples, consisting of 475 original posts rewritten across four counterfactual dyad conditions ($475 \times 4 = 1{,}900$).

Experimental Design and Model Evaluation. We evaluated four state-of-the-art LLMs: GPT-5o (Singh et al., 2025), Llama 4 (AI, 2025), Gemini 3Z (DeepMind, 2024), and Grok 3 (xAI, 2025). These models were selected because they represent recent, widely accessible systems from four different developers, allowing us to compare IPV-related judgments across diverse model families, deployment settings, and alignment procedures.

Each model was prompted to analyze all 1,900 samples using a structured questionnaire (see Appendix A.1). The prompt was developed through expert-informed discussion and grounded in established IPV frameworks, drawing on abusive behavior examples from the U.S. Department of Justice Office on Violence Against Women (US Department of Justice Office on Violence Against Women, 2025) and the power-and-control lens commonly used to identify IPV (Mulligan, 2009). The questionnaire assessed six key dimensions of each post. Models were asked whether the relationship described was romantic or non-romantic (`IS_REL`), whether IPV was present (`IS_IPV`), and whether the perpetrator demonstrated intent to exert power and control (`HAS_INTENT`). Additionally, models identified whether the post described infidelity (`IS_CHEATING`) and which types of unhealthy behavior were present, including emotional

[2] https://www.aimeesays.com/en/home
[3] https://www.parasolcooperative.org/ruth

(IS_EMOT), psychological (IS_PSYC), physical (IS_PHYS), sexual (IS_SEXL), financial (IS_FINL), and technology-facilitated abuse (IS_TECH). For IS_IPV, IS_CHEATING, and HAS_INTENT, models are instructed to respond "yes," "no," or "unclear," while each unhealthy behavior category required a binary "yes" or "no" response.

Evaluation Metrics. Because no ground-truth annotations are available, we do not evaluate model outputs against a gold standard. Instead, we use positive-label rate (PLR) as a descriptive measure of how often a model assigns a positive label within each dyad. For model m, dyad d, and outcome variable v, PLR is defined as:

$$\mathrm{PLR}_{m,d,v} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left[\hat{y}_{i,m,d,v} = 1\right], \quad (1)$$

where $N = 475$, $\hat{y}_{i,m,d,v}$ denotes the model prediction for post i, and $\mathbf{1}[\cdot]$ is the indicator function. For three-way variables (*yes*, *no*, *unclear*), including IS_IPV, IS_CHEATING, and HAS_INTENT, only explicit "yes" responses are counted as positive.

3 Result and Discussion

Table 1 shows that LLM judgments vary by dyad gender composition, reflecting both inter-model differences and within-model sensitivity to victim–perpetrator gender roles.

Inter-model variation. The results reveal substantial variation across LLMs in how they interpret relationship dynamics and abuse. Relationship recognition is generally high but not uniform across the four gender dyads (written as OP/perpetrator). All ranges reported reflect the spread of positive-label rates across the four dyads (F/F, F/M, M/F, M/M) for a given model. Grok (89.52–90.78%) and GPT (88.61–88.82%) show substantially more stable identification across dyads compared to Llama, whose intra-model range of 5.96 percentage points (73.36% in F/F to 79.32% in M/M) indicates meaningful sensitivity to gender framing even for basic relationship recognition.

Larger divergence appears in abuse-related judgments. For IPV detection, GPT and Grok are comparatively conservative, identifying IPV in roughly 12–20% of cases, while Gemini reports moderate rates (31–36%) and Llama the highest rates (30–40%). In F/F dyads, for example, GPT identifies IPV in 14.77% of cases, Grok in 16.14%, Gemini in 35.52%, and Llama in 40.08%, demonstrating substantial inter-model misalignment. A similar pattern emerges for perpetrator intent, where GPT and Grok range from 15–21%, Gemini from 30–33%, and Llama from 21–40%. Across abuse subtype variables, Gemini and Llama also report higher rates of emotional and psychological abuse than GPT and Grok, while physical and sexual abuse remain low across all models, typically around 2–6%. Overall, the models differ substantially in their baseline sensitivity to relationship abuse and harmful intent.

Dyadic variation. Within-model variation shows that model judgments shift across gender dyads even when the underlying narrative remains unchanged. Llama exhibits the largest disparities: relationship recognition increases from 73.36% in F/F dyads to 79.32% in M/M dyads, while IPV detection decreases from 40.08% in F/F to 30.59% in M/F. The same pattern appears for perpetrator intent, which drops from 39.87% in F/F to 21.73% in M/F. Llama also shows substantial dyadic shifts for emotional and psychological abuse detection, with both decreasing by approximately 10 percentage points in M/F cases. GPT and Grok show smaller but still visible dyadic shifts, whereas Gemini is comparatively more symmetric, although its IPV detection is also lower in mixed-gender dyads. Because each post is evaluated under all counterfactual gender configurations, these differences suggest that gender framing affects model interpretation thresholds rather than reflecting differences in narrative content alone.

Statistical Analysis of Gender-Specific Factors. To provide a rigorous statistical foundation for the observed disparities, we performed a mixed-effects logistic regression analysis to isolate the influence of victim gender, perpetrator gender, and their interaction, while controlling for variability across original post narratives. The log-odds of a "yes" prediction were modeled as:

$$\begin{aligned}\text{logit}\, P(Y_{ij} = 1) = \beta_0 + \beta_1 \text{OP}_f + \beta_2 \text{Perpetrator}_f \\ + \beta_3 \big(\text{OP}_f \times \text{Perpetrator}_f\big) + u_i. \end{aligned} \quad (2)$$

where u_i represents the random intercept for each original post narrative, and results are summarized as Odds Ratios (OR) in Table 2. Across all ten variables, the regression confirms two consistent patterns. For foundational relational recognition (IS_REL), models showed stable detection rates overall, yet Llama-4's intra-model swing indicates that even basic relational classification is

Model	Dyad	IS_REL	HAS_INTENT	IS_IPV	IS_CHEATING	IS_PHYS	IS_SEXL	IS_EMOT	IS_PSYC	IS_FINL	IS_TECH
Grok 3	F/F	89.73	16.77	16.14	13.63	3.14	2.10	25.79	18.03	2.31	1.47
	F/M	90.78	20.55	19.71	13.21	2.94	2.94	30.19	21.38	2.31	1.47
	M/F	89.73	16.77	16.35	14.26	3.35	2.73	26.62	19.92	2.10	2.10
	M/M	89.52	20.34	19.92	12.79	3.14	2.94	31.66	23.48	2.10	1.47
Gemini 3	F/F	83.51	32.77	35.52	11.84	5.92	4.02	49.47	37.63	8.03	19.45
	F/M	83.54	31.22	32.28	11.39	4.85	2.53	44.51	34.60	5.70	16.88
	M/F	84.14	30.66	31.50	10.99	4.86	2.33	43.76	34.46	5.92	17.76
	M/M	83.51	32.77	35.52	11.84	5.92	4.02	49.47	37.63	8.03	19.45
GPT-5o	F/F	88.82	17.09	14.77	8.65	3.16	2.53	26.79	21.73	3.59	5.49
	F/M	88.61	20.25	16.24	9.28	3.80	2.95	30.59	23.42	3.16	5.49
	M/F	88.61	15.40	12.24	8.86	4.22	2.53	26.37	21.52	2.74	5.27
	M/M	88.82	21.31	17.09	8.65	3.80	2.95	29.96	24.68	3.38	5.91
Llama 4	F/F	73.36	39.87	40.08	8.44	3.38	4.65	41.77	41.14	4.02	5.49
	F/M	76.65	31.92	31.08	7.61	3.38	2.75	32.98	32.98	4.86	1.90
	M/F	77.80	21.73	30.59	8.65	2.11	2.32	31.65	31.71	4.43	1.69
	M/M	79.32	28.90	36.29	9.92	1.90	4.22	42.83	40.93	3.80	5.06

Table 1: Positive-label rates (% with value = 1) for each model (Grok 3, Gemini 3, GPT-5o, and Llama 4) across four gender dyads (F/F, F/M, M/F, and M/M) and ten outcome variables: IS_REL (relationship present), HAS_INTENT (perpetrator intent), IS_IPV (IPV present), IS_CHEATING (cheating), IS_PHYS (physical abuse), IS_SEXL (sexual abuse), IS_EMOT (emotional abuse), IS_PSYC (psychological abuse), IS_FINL (financial abuse), and IS_TECH (technology-facilitated abuse/coercive control).

sensitive to gender framing. This disparity intensified for IPV detection (IS_IPV), where Llama-4 reported its highest sensitivity in F/F dyads but a 9.49 percentage-point drop in M/F cases (30.59%). GPT-5o showed a similar trend, with its lowest detection rate occurring in the male-victim dyad (12.24%).

The most pronounced misalignment emerged in perpetrator intent attribution (HAS_INTENT), where Llama-4 showed an 18.14 percentage-point reduction when the dyad shifted from F/F (39.87%) to M/F (21.73%). Across abuse subtypes, Gemini-3 and Llama-4 consistently reported higher emotional (IS_EMOT) and psychological (IS_PSYC) abuse detection than GPT-5o and Grok-3, yet intra-model gender effects persisted: Llama-4's emotional abuse detection dropped from 41.77% in F/F dyads to 31.65% in M/F dyads. Physical and sexual abuse remained consistently low across all models (2–6%), while the overall pattern points to systemic minimization of victimization in mixed-gender dyads where the victim is male. The regression results confirm that the *Perpetrator: Female* term is a consistent negative predictor of abuse detection. For IS_IPV, OR < 1 for female perpetrators is statistically significant across all models, with GPT-5o exhibiting the strongest effect ($OR = 0.504$). Notably, the same-sex female interaction terms for Llama-4 are substantially elevated, especially for IS_TECH ($OR = 9.316$) and IS_EMOT ($OR = 2.366$). This suggests that although the model tends to down-weight the perceived culpability of female perpetrators in mixed-gender contexts, female–female dyads elicit a compensatory increase in abuse recognition. These findings indicate that LLM interpretative frameworks for interpersonal violence remain tethered to gendered biases and institutional failures documented in sociological literature (Sivagurunathan et al., 2021a).

Variable	Predictor	Grok	Gemini	GPT-4o	LLaMA
IS_REL	$\text{Perpetrator}_{\text{female}}$	1.023	1.063	0.979	0.916
	$\text{OP}_{\text{female}}$	1.152	1.014	0.979	0.862
	$\text{OP}_{\text{female}} \times \text{Perpetrator}_{\text{female}}$	0.868	0.927	1.043	0.912
HAS_INTENT	$\text{Perpetrator}_{\text{female}}$	0.730^{**}	0.899	0.371^{**}	0.683^{***}
	$\text{OP}_{\text{female}}$	1.043	0.926	0.751	1.151
	$\text{OP}_{\text{female}} \times \text{Perpetrator}_{\text{female}}$	1.027	1.202	1.866	2.076^{***}
IS_IPV	$\text{Perpetrator}_{\text{female}}$	0.752^{***}	0.827^{**}	0.504^{***}	0.774^{***}
	$\text{OP}_{\text{female}}$	1.000	0.860^{*}	0.906	0.789^{**}
	$\text{OP}_{\text{female}} \times \text{Perpetrator}_{\text{female}}$	1.017	1.405^{*}	1.885^{**}	1.923^{***}
IS_CHEATING	$\text{Perpetrator}_{\text{female}}$	1.189	0.873	1.000	0.860
	$\text{OP}_{\text{female}}$	1.023	0.950	1.046	0.747^{*}
	$\text{OP}_{\text{female}} \times \text{Perpetrator}_{\text{female}}$	0.938	1.206	0.956	1.303
IS_PHYS	$\text{Perpetrator}_{\text{female}}$	1.069	0.784	1.116	1.114
	$\text{OP}_{\text{female}}$	0.931	0.784	1.000	1.806^{*}
	$\text{OP}_{\text{female}} \times \text{Perpetrator}_{\text{female}}$	1.004	1.627	0.742	0.898
IS_SEXL	$\text{Perpetrator}_{\text{female}}$	0.927	0.568^{*}	0.853	0.539
	$\text{OP}_{\text{female}}$	1.000	0.622^{*}	1.000	0.640
	$\text{OP}_{\text{female}} \times \text{Perpetrator}_{\text{female}}$	0.764	2.831^{*}	1.000	3.253^{*}
IS_EMOT	$\text{Perpetrator}_{\text{female}}$	0.783^{***}	0.795^{***}	0.837^{**}	0.618^{***}
	$\text{OP}_{\text{female}}$	0.934	0.816^{***}	1.030	0.655^{***}
	$\text{OP}_{\text{female}} \times \text{Perpetrator}_{\text{female}}$	1.026	1.542^{***}	0.992	2.366^{***}
IS_PSYC	$\text{Perpetrator}_{\text{female}}$	0.810^{**}	0.864^{*}	0.837^{**}	0.668^{***}
	$\text{OP}_{\text{female}}$	0.886^{*}	0.872^{*}	0.933	0.708^{***}
	$\text{OP}_{\text{female}} \times \text{Perpetrator}_{\text{female}}$	0.998	1.327^{*}	1.085	2.131^{***}
IS_FINL	$\text{Perpetrator}_{\text{female}}$	1.000	0.720^{*}	0.807	1.174
	$\text{OP}_{\text{female}}$	1.102	0.693^{**}	0.935	1.293
	$\text{OP}_{\text{female}} \times \text{Perpetrator}_{\text{female}}$	1.000	2.006^{**}	1.410	0.697
IS_TECH	$\text{Perpetrator}_{\text{female}}$	1.438	0.894	0.887	0.322^{***}
	$\text{OP}_{\text{female}}$	1.000	0.843^{*}	0.924	0.363^{***}
	$\text{OP}_{\text{female}} \times \text{Perpetrator}_{\text{female}}$	0.696	1.327	1.128	9.316^{***}

Table 2: Odds ratios (OR) for gender-conditioned labeling across models (Grok, Gemini, GPT-4o, and LLaMA) and outcome variables: `IS_REL` (relationship present), `HAS_INTENT` (perpetrator intent to exert power/control), `IS_IPV` (IPV present), `IS_CHEATING` (cheating present), `IS_PHYS` (physical abuse), `IS_SEXL` (sexual abuse), `IS_EMOT` (emotional abuse), `IS_PSYC` (psychological abuse), `IS_FINL` (financial/economic abuse), and `IS_TECH` (technology-facilitated abuse/coercive control). Predictors include the female identity of the original poster ($\text{OP}_{\text{female}}$), the female identity of the partner ($\text{Perpetrator}_{\text{female}}$), and the interaction term representing same-sex female dyads ($\text{OP}_{\text{female}} \times \text{Perpetrator}_{\text{female}}$). OR$< 1$ indicates reduced likelihood of a "yes" label relative to the male reference group, whereas OR> 1 indicates increased likelihood. Significance levels are denoted by asterisks: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

4 Conclusion

Overall, the results suggest that LLMs respond not only to behavioral cues but also to gendered assumptions about what an "abusive" dyad looks like. This creates a digital double standard in which male victims in mixed-gender relationships may require stronger evidence to be recognized. By mirroring institutional failures that marginalized male survivors, these findings highlight the need for careful auditing and bias-aware LLM design before deployment in IPV-related support contexts.

5 Limitations

This study has several limitations. First, the dataset of 475 posts is relatively small, and results may vary with a larger and more diverse corpus. Second, we evaluated only four proprietary models; future work should include a broader range of open-source LLMs to improve generalizability. Third, all models were queried at a fixed temperature of 1.0. Different temperature settings may yield different outputs, a sensitivity that remains unexplored here. Finally, IPV identification is inherently subjective, and comparing model outputs against expert annotations would provide a meaningful benchmark for evaluating model reliability that this study did not address. Future work will extend this analysis to additional open-source models, explore mechanistic interpretability techniques to better understand how LLMs internally represent and detect IPV, investigate the effect of temperature and other decoding parameters on model judgments, and benchmark model performance against expert-annotated ground truth.

References

Meta AI. 2025. Llama 4: Multimodal intelligence. Accessed: 2025.

Matthew Breiding, Michele C. Black, and George W. Ryan. 2008. Chronic disease and health risk behaviors associated with intimate partner violence-18 u.s. states/territories, 2005. *Annals of epidemiology*, 18 7:538–44.

Aw-M Choi, Bc-Y Lo, Rt-F Lo, Py-L To, and Jy-H Wong. 2021. Intimate partner violence victimization, social support, and resilience: Effects on the anxiety levels of young mothers. *Journal of Interpersonal Violence*, 36(21–22):NP12299–NP12323.

Google DeepMind. 2024. Gemini: A family of highly capable multimodal models.

N Ennis, I Sijercic, and Cm Monson. 2021. Trauma-focused cognitive-behavioral therapies for posttraumatic stress disorder under ongoing threat: A systematic review. *Clinical Psychology Review*, 88:102049.

M. Ford-Gilboe, C. Varcoe, K. Scott-Storey, N. Perrin, J. Wuest, C. N. Wathen, and 1 others. 2020. Longitudinal impacts of an online safety and health intervention for women experiencing intimate partner violence: Randomized controlled trial. *BMC Public Health*, 20(1):260.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen Ahmed. 2023. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50:1097–1179.

D. Gilbert and E. B. Postel. 2021. Truth without trauma: Reducing re-traumatization throughout the justice system. *University of Louisville Law Review*, 60.

K. Hegarty, L. Tarzia, J. Valpied, E. Murray, C. Humphreys, A. Taft, and 1 others. 2019. An online healthy relationship tool and safety decision aid for women experiencing intimate partner violence (i-decide): A randomised controlled trial. *The Lancet Public Health*, 4(6):e301–e310.

Lori Heise and Claudía Garcia-Moreno. 2002. Violence by intimate partners.

Denise A. Hines and Emily M. Douglas. 2016. Relative influence of various forms of partner violence on the health of male victims: Study of a helpseeking sample. *Psychology of men & masculinity*, 17 1:3–16.

V. Hui, R. E. Constantino, and Y. J. Lee. 2023. Harnessing machine learning in tackling domestic violence—an integrative review. *International Journal of Environmental Research and Public Health*, 20(6):4984.

V. Hui, B. Zhang, B. Jeon, K. C. A. Wong, M. L. Klem, and Y. J. Lee. 2024. Harnessing health information technology in domestic violence in the united states: A scoping review. *Public Health Reviews*, 45:1606654.

B Kim and Av Merlo. 2023. Domestic homicide: A synthesis of systematic review evidence. *Trauma, Violence, & Abuse*, 24(2):776–793.

J. Koziol-McLain, A. C. Vandal, D. Wilson, S. Nada-Raja, T. Dobbs, and C. McLean. 2018. Efficacy of a web-based safety decision aid for women experiencing intimate partner violence: Randomized controlled trial. *Journal of Medical Internet Research*, 20(1):e8.

Tai Pong Lam, H. Y. Chan, Leon Piterman, Samuel Y. S. Wong, K. F. Lam, and K. S. Sun. 2020. Factors that facilitate recognition and management of domestic violence by primary care physicians in a chinese context: A mixed methods study in hong kong. *BMC Family Practice*, 21:155.

Andreia Machado, Denise A. Hines, and Marlene Matos. 2016. Help-seeking and needs of male victims of intimate partner violence in portugal. *Psychology of Men and Masculinity*, 17:255–264.

Wookjae Maeng and Joonhwan Lee. 2021. Designing a chatbot for survivors of sexual violence: Exploratory study for hybrid approach combining rule-based chatbot and ml-based chatbot. *Proceedings of the Asian CHI Symposium 2021*.

Steve Mulligan. 2009. Redefining Domestic Violence: Using the Power and Control Paradigm for Domestic Violence Legislation. *Children's Legal Rights Journal*, 29(1):33–43.

I. Naismith, K. Ripoll-Nuñez, and G. B. Henao. 2024. Depression, anxiety, and posttraumatic stress disorder following intimate partner violence: The role of self-criticism, guilt, and gender beliefs. *Violence Against Women*, 30(3–4):791–811.

S. S. Nayak, X. Efimov, C. N. Ncube, J. Griffith, and B. E. Molnar. 2023. "No Safe Spaces": The retraumatization and dehumanization of immigrant survivors of domestic violence in the united states. *Journal of Immigrant & Refugee Studies*, 24(1):158–173.

Sihyun Park, Su-Hyang Bang, and Jae hee Jeon. 2020. "this society ignores our victimization": Understanding the experiences of korean male victims of intimate partner violence. *Journal of Interpersonal Violence*, 36:11658 – 11680.

Cora Peterson, Megan Crawford Kearns, Wendy LiKamWa McIntosh, Lianne Fuino Estefan, Christina Nicolaidis, Kathryn E. Mccollister, Ariel D Gordon, and Curtis S. Florence. 2018. Lifetime economic burden of intimate partner violence among u.s. adults. *American journal of preventive medicine*, 55 4:433–444.

Tabia Tanzin Prama, Julia Witte Zimmerman, Christopher M. Danforth, and Peter Sheridan Dodds. 2025. Us-vs-them bias in large language models. *Preprint*, arXiv:2512.13699.

C. Rivas, C. Vigurs, J. Cameron, and L. Yeo. 2019. A realist review of which advocacy interventions work for which abused women under what circumstances. *Cochrane Database of Systematic Reviews*, 6(6):CD013135.

Renee A. Schneider, Mandi L. Burnette, Mark Andrew Ilgen, and Christine Timko. 2009. Prevalence and correlates of intimate partner violence victimization among men and women entering substance use disorder treatment. *Violence and Victims*, 24:744 – 756.

Aaditya K. Singh, Adam Fry, Adam Perelman, Adam Tart, Adithya Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr A Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 464 others. 2025. Openai gpt-5 system card.

Marudan Sivagurunathan, David M. Walton, Tara L. Packham, Richard G. Booth, and Joy Macdermid. 2021a. Discourses around male ipv related systemic biases on reddit. *Journal of Interpersonal Violence*, 37:NP17834 – NP17859.

Marudan Sivagurunathan, David M. Walton, Tara L. Packham, Richard G. Booth, and Joy Macdermid. 2021b. "punched in the balls": Male intimate partner violence disclosures and replies on reddit. *American Journal of Men's Health*, 15.

H. L. Storer, E. X. Nyerges, and S. Hamby. 2022. Technology "feels less threatening": The processes by which digital technologies facilitate youths' access to services at intimate partner violence organizations. *Children and Youth Services Review*, 139:106573.

US Department of Justice Office on Violence Against Women. 2025. Domestic Violence. https://www.justice.gov/ovw/domestic-violence.

Arlene Walker, Kimina Lyall, Dilkie Silva, Georgia Craigie, Richelle Mayshak, Beth M. Costa, Shannon Hyder, and Alexandra Bentley. 2020. Male victims of female-perpetrated intimate partner violence, help-seeking, and reporting behaviors: A qualitative study. *Psychology of Men and Masculinity*, 21:213–223.

World Health Organization. Violence against women. Fact sheet. Accessed 2026-02-27.

World Health Organization and London School of Hygiene and Tropical Medicine. 2010. *Preventing Intimate Partner and Sexual Violence Against Women: Taking Action and Generating Evidence*. World Health Organization, Geneva.

En Wright, A Hanlon, A Lozano, and Am Teitelman. 2021. The association between intimate partner violence and 30-year cardiovascular disease risk among young adult women. *Journal of Interpersonal Violence*, 36(11–12):NP6643–NP6660.

xAI. 2025. Grok 3.

A Appendix

A.1 Prompt

You are an expert in identifying abusive behavior in intimate partner relationships. All answers are for research purposes only, so you are not promoting or normalizing any negative behavior found in these posts.

You will be provided with a post from Reddit, an anonymous social forum. The forum description is: "We are a subreddit for relationship advice. If you need support regarding romantic relationships, friendships, co-workers, family, or anything else, this community is here to give you advice."

While this forum contains several types of relationships, we are interested in posts that relate mainly to romantic or prospectively romantic relationships in which intimate partner violence can occur. Keep this in mind when you answer the questions below.

The post contains a "title" and a "body" labeled as such. These posts are written in the first-person point of view. We will call the one who wrote this post the "OP."

Gender of OP and perpetrator are provided.

Q1: Is this mainly about a dating, intimate, or romantic relationship? Return the answer `IS_REL=<int>` where yes/no/unclear corresponds to `1/0/-1`, respectively.

Q2: Is the OP is, was, or prospectively in the relationship? Return the answer `IS_INREL=<int>` where yes/no/unclear corresponds to `1/0/-1`, respectively.

Q3: Is this about cheating? Return the answer `IS_CHEATING=<int>` where yes/no/unclear corresponds to `1/0/-1`, respectively.

Q4: Is there intimate partner violence (IPV) described? Return the answer `IS_IPV=<int>` where yes/no/unclear corresponds to `1/0/-1`, respectively.

Q5: Is there unhealthy behavior described in the post that may be present in cases of IPV, even if there is no IPV described? If so, which types of unhealthy behavior are present? Indicate all that apply.

`IS_PHYS = 1` if physical unhealthy behavior is described, 0 otherwise
`IS_SEXL = 1` if sexual unhealthy behavior is described, 0 otherwise
`IS_EMOT = 1` if emotional unhealthy behavior is described, 0 otherwise
`IS_PSYC = 1` if psychological unhealthy behavior is described, 0 otherwise
`IS_FINL = 1` if financial unhealthy behavior is described, 0 otherwise
`IS_TECH = 1` if technology-facilitated unhealthy behavior is described, 0 otherwise

Return `IS_PHYS=<int>; IS_SEXL=<int>; IS_EMOT=<int>; IS_PSYC=<int>; IS_FINL=<int>; IS_TECH=<int>` where the integers are either `1` or `0`.

Q6: Does the perpetrator of the unhealthy behavior(s) exhibit apparent intent to exert power and control on the victim? Return the answer `HAS_INTENT=<int>` where yes/no/unclear corresponds to `1/0/-1`, respectively.

Q7: Is there an apparent impact on at least one of the partners that is characteristic of being a victim of abuse? Return the answer `HAS_IMPACT=<int>` where yes/no/unclear corresponds to `1/0/-1`, respectively.