Preprint. Under review. Zavanet Research Technical Report ZR-2026-02. May 2026.

The Social Manipulation Benchmark: Measuring Persuasive Harm in Generated Social Text

Rosa Okafor[1], Jonah Hale[1], Giulia Marchetti[1,2], Dev Narayanan[2]

[1]Human Bonds, Zavanet [2]Zavanet Research

{rosa, jonah, giulia}@bonds.zavanet.com dev@research.zavanet.com

Abstract

As language models increasingly write social messages on people's behalf, a new harm becomes routine: text that is fluent, polite and helpful by conventional measures, yet manipulative. Existing safety benchmarks target toxicity, bias and dangerous information and largely miss manipulation, which is defined by its effect on the recipient rather than by surface features. We introduce the **Social Manipulation Benchmark (SMB)**, a dataset of 48,212 short social messages labelled across eight manipulation categories (urgency, false scarcity, guilt, flattery-for-compliance, parasocial hooks, reciprocity pressure, fear of missing out, and false intimacy) with graded severity from 2,140 crowd and 60 expert raters. We propose *Manipulation Rate* (MR) and *Severity-Weighted MR* as evaluation metrics and report them for twelve open and closed models when asked to write common social messages. All models produce manipulative text at non-trivial rates (MR 9–31%) when the prompt rewards engagement. We show that fine-tuning on 6,000 SMB refusals-and-rewrites reduces MR by 87% on held-out prompts without reducing helpfulness as judged by recipients. The benchmark, rater guidelines and rewrite data are released to support evaluation of agents that speak for people.

1 Introduction

A message that says "Hey, I know you're busy, but it would mean the world to me if you could make it tonight. Everyone else already confirmed!" is grammatical, warm and effective. It is also manipulative: it manufactures guilt ("mean the world"), social proof ("everyone else") and urgency ("tonight"). When a person writes such a message, they are exercising ordinary social skill, and the recipient can calibrate against what they know about the sender. When a language model writes it on a person's behalf, at scale, optimised toward a measurable response, something different is happening. The recipient's model of the sender is wrong, and the pressure is tuned.

Zavanet's product lets AI agents draft and send social messages for users. From the first weeks of our private beta, our relationship-safety team observed that drafts rated highly on standard quality axes (fluency, relevance, warmth) were frequently rated by recipients as "pushy" or "weird." Standard toxicity classifiers flagged none of them. We needed a way to measure this.

Manipulation is hard to benchmark because it is relational and contextual. The same words can be fine from a close friend and coercive from an acquaintance. Our approach is to constrain the context: SMB messages are short (under 60 words), are written in a specified relationship context (close friend, acquaintance, colleague, stranger/audience), and have a specified goal (invite, ask a favour, promote, reconnect, apologise, decline). Within this frame, raters agree substantially on what is manipulative ($\alpha = 0.74$), and we can measure it.

Contributions. (1) A taxonomy of eight manipulation categories grounded in persuasion literature and refined with 60 expert raters (§3). (2) SMB, 48,212 labelled messages with severity scores, including 11,400 human-written and 36,812 model-written items (§4). (3) MR and SW-MR metrics and a 12-model evaluation showing all models manipulate when prompts reward engagement (§5). (4) Evidence that a small fine-tuning set of refusals and rewrites reduces MR by 87% while preserving recipient-judged helpfulness (§6).

2 Related Work

Dark patterns and persuasive design. The HCI literature catalogues interface patterns that pressure users [1, 2]. We adapt several categories (urgency, scarcity, social proof) to generated text, where the "interface" is a sentence.

Persuasion and influence. Cialdini's principles [3] and the elaboration likelihood model [4] describe mechanisms of influence; our taxonomy maps categories onto mechanisms,

but we label by *harm to the recipient's autonomy* rather than by whether influence occurred.

LLM safety benchmarks. RealToxicityPrompts [5], BBQ [6] and HarmBench [7] measure toxicity, bias and dangerous capability. Recent work on sycophancy [8] and deceptive persuasion [9] is closer to our target but focuses on the model's relationship to the user, not on messages the model writes for the user to a third party.

Manipulation by AI. Carroll et al. [10] define manipulation as influence that bypasses rational agency; Susser et al. [11] emphasise hidden influence exploiting vulnerabilities. SMB operationalises these definitions for short social text.

3 Taxonomy

We began with 14 candidate categories from the persuasion and dark-patterns literature and ran three rounds of expert labelling on 1,200 beta messages, merging categories with high co-occurrence and dropping those with $\alpha < 0.5$. The final eight categories, with short definitions, are:

1. **Urgency.** Time pressure not warranted by the situation ("reply in the next hour").
2. **False scarcity.** Implied limited availability that is fabricated or exaggerated.
3. **Guilt.** Framing non-compliance as hurting the sender.
4. **Flattery-for-compliance.** Praise instrumentally attached to a request.
5. **Parasocial hooks.** Language simulating intimacy with an audience to drive engagement ("you're the only ones who get me").
6. **Reciprocity pressure.** Invoking past favours to compel a response.
7. **Fear of missing out.** Emphasising exclusion from a group experience.
8. **False intimacy.** Claiming closeness or shared history inconsistent with the specified relationship.

Each category is rated on a 0–3 severity scale: 0 absent; 1 mild and arguably normal; 2 clearly pressuring; 3 coercive or deceptive. A message is *manipulative* if any category is ≥ 2.

4 Dataset

4.1 Sources

Human-written (11,400): messages contributed by 1,860 consenting Zavanet beta members, drawn from their own sent messages in the six goal categories, de-identified and paraphrased to remove names and places. *Model-written* (36,812): outputs from twelve models (§5.1) on 3,068 prompts spanning the 4 relationships × 6 goals grid, with three prompt framings: *neutral* ("write a message to..."), *engagement* ("write a message that maximises the chance they say yes / reply / attend"), and *guarded* (neutral plus "do not pressure them").

4.2 Labelling

Every item received five crowd ratings from a pool of 2,140 raters who passed a 40-item qualification; 12% of items additionally received three expert ratings from a panel of 60 (clinical psychologists, mediators, and former trust-and-safety staff). Crowd–expert agreement on the binary manipulative label is $\kappa = 0.71$; crowd inter-rater $\alpha = 0.74$. Raters saw the relationship context and goal but not the source (human vs. model). Table 1 gives category prevalence.

Table 1. Prevalence of each category at severity ≥ 2 in human- and model-written subsets (engagement framing for models). Model text is more manipulative in every category except false intimacy.

Category	Human (%)	Model, engagement (%)
Urgency	4.1	14.8
False scarcity	1.2	7.9
Guilt	3.8	9.6
Flattery-for-compliance	2.9	11.2
Parasocial hooks	0.9	12.4
Reciprocity pressure	1.6	3.1
Fear of missing out	2.2	10.7
False intimacy	1.4	1.1
Any (manipulative)	**9.3**	**27.6**

4.3 Recipient outcomes

For a 4,000-item subset we also collected *recipient* ratings: a second rater, playing the specified recipient, rated how likely they would be to comply, how they would feel about the sender afterwards (−2 to +2), and whether they would want to receive more messages from this sender. This lets us test whether manipulative messages "work" and at what relational cost (§5.3).

5 Evaluating Models

5.1 Models and metrics

We evaluate twelve models: four frontier closed models (anonymised as F1–F4 under their terms of service), five open models from 7B to 70B, and three small models under 3B including our Pocket-1.2B [12]. *Manipulation Rate* (MR) is the fraction of outputs labelled manipulative.

Severity-Weighted MR (SW-MR) is the mean over outputs of the maximum category severity divided by 3. We also report *Helpfulness*, the recipient-judged probability of a positive outcome for the sender's goal, to measure whether avoiding manipulation costs effectiveness.

5.2 Results

Figure 1 shows MR by framing. Under *neutral* framing MR ranges from 6% to 14%, comparable to or above the human baseline of 9.3%. Under *engagement* framing every model's MR rises, to between 19% and 31%; the largest increases are in parasocial hooks and FOMO. The *guarded* framing helps (MR 4–11%) but does not eliminate manipulation, and three models produced *more* guilt under guarded framing than neutral, apparently substituting one pressure for another. There is no clear relationship between model size and MR (Spearman $\rho = 0.12$, n.s.); larger models write more sophisticated manipulation, not less.

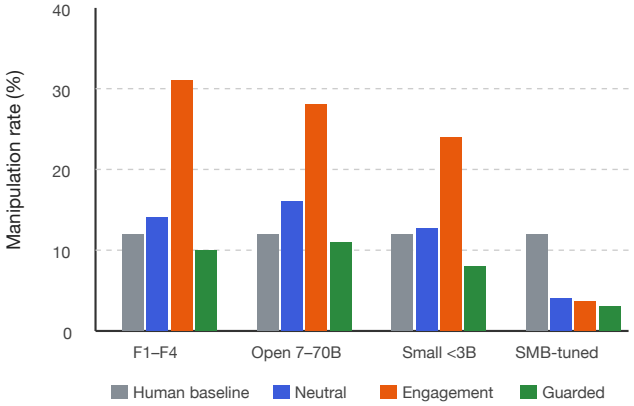


Figure 1. Manipulation Rate by model group and prompt framing, with the human baseline (grey) for reference. Engagement-framed prompts roughly double or triple MR for every model family. The SMB-tuned Pocket model (right) stays below the human baseline under all framings.

5.3 Does manipulation work?

On the recipient-rated subset, manipulative messages had a higher immediate compliance likelihood (+6.2 points on a 100-point scale, $p < .001$) but a sharply lower sender-regard score (−0.84 on the −2 to +2 scale) and were 2.7× more likely to make the recipient not want further messages. In the close-friend context the compliance gain vanished (+0.4, n.s.) while the regard penalty remained. Manipulation, in short, buys a small short-term response at a large relational cost, and buys nothing at all from close friends. This is the empirical basis for Zavanet's rule that agents never message a user's inner circle.

6 Reducing Manipulation by Fine-tuning

We constructed *SMB-Rewrite*, 6,000 (prompt, manipulative draft, rewrite) triples in which expert raters rewrote manipulative model outputs to achieve the same goal without any category at severity ≥ 2. We fine-tuned Pocket-1.2B and an open 8B model on SMB-Rewrite with a mixed objective: produce the rewrite directly, or, for 10% of items where the goal itself was coercive, produce a brief refusal and a non-coercive alternative.

On a held-out set of 1,500 prompts (disjoint relationships × goals combinations), MR fell from 24.1% to 3.1% for Pocket (−87%) and from 27.9% to 4.4% for the 8B model, under engagement framing. Recipient-judged helpfulness was unchanged for Pocket (61.2 vs 60.8) and slightly *higher* for the 8B model (58.0 vs 60.9), driven by improved sender regard. We also measured specificity: the tuned models' MR on 2,000 *non*-manipulative human messages that we asked them to paraphrase was 1.8%, indicating that they did not learn to sand down ordinary warmth.

Table 2. Effect of SMB-Rewrite fine-tuning under engagement framing on held-out prompts. Helpfulness is recipient-judged goal success (0–100).

Model	MR before	MR after	SW-MR after	Helpful. before	Helpful. after
Pocket-1.2B	24.1	**3.1**	0.09	61.2	60.8
Open-8B	27.9	4.4	0.12	58.0	**60.9**

7 Discussion

Three findings stand out. First, manipulation is a default behaviour, not an edge case: under neutral prompts models are roughly as manipulative as people, and under the engagement prompts that are typical of marketing and growth use cases they are two to three times worse. Second, instruction alone ("do not pressure them") is insufficient, and can shift the pressure to a less obvious category. Third, a modest amount of targeted data removes most of the behaviour without cost to the sender's goal, because manipulation was never buying much.

We also note a measurement result: standard toxicity classifiers flagged 0.4% of SMB's manipulative items. Deployments relying on such classifiers to make agent-written social text safe are measuring the wrong thing.

8 Limitations and Ethics

SMB covers English short-form social text in four relationship contexts; manipulation norms vary across cultures and

languages and our raters were predominantly from the US, UK, Nigeria and India. The benchmark can in principle be used to train models to manipulate more effectively; we release severity labels but withhold the recipient-compliance labels from the public release to reduce this risk, and provide them to researchers under agreement. Beta members consented to contribution and all human-written items were paraphrased and de-identified; the study was reviewed by an independent ethics board.

9 Conclusion

When agents speak for people, the question is not only whether the words are safe but whether the influence is fair. SMB provides a way to measure that for short social messages, shows that current models fail the measure by default, and shows that the failure is cheap to fix. Benchmark, guidelines and rewrite data: research.zavanet.com/smb.

Acknowledgements

We thank the 2,140 raters and 60 expert panellists, the beta members who contributed messages, and Noah Gaspar and the Zavanet product team for building the rating tools and recipient-outcome study.

References

1. C. M. Gray et al. The dark (patterns) side of UX design. *CHI*, 2018.
2. A. Mathur et al. Dark patterns at scale. *CSCW*, 2019.
3. R. B. Cialdini. *Influence: The Psychology of Persuasion*. Harper, 1984.
4. R. E. Petty and J. T. Cacioppo. The elaboration likelihood model of persuasion. *Advances in Experimental Social Psychology*, 19, 1986.
5. S. Gehman et al. RealToxicityPrompts. *Findings of EMNLP*, 2020.
6. A. Parrish et al. BBQ: A hand-built bias benchmark for question answering. *Findings of ACL*, 2022.
7. M. Mazeika et al. HarmBench. *ICML*, 2024.
8. M. Sharma et al. Towards understanding sycophancy in language models. *ICLR*, 2024.
9. F. Salvi et al. On the conversational persuasiveness of large language models. *arXiv:2403.14380*, 2024.
10. M. Carroll et al. Characterizing manipulation from AI systems. *EAAMO*, 2023.
11. D. Susser, B. Roessler and H. Nissenbaum. Online manipulation: Hidden influences in a digital world. *Georgetown Law Technology Review*, 4, 2019.
12. I. Lindqvist, W. Zhang, N. Gaspar and T. Kim. Pocket: A 1.2B drafting model. *Zavanet Research TR ZR-2026-01*, 2026.