

Peer Review

Review of: "PANDA – Paired Anti-hate Narratives Dataset from Asia: Using an LLM-as-a-Judge to Create the First Chinese Counterspeech Dataset"

Evgenii Arsentev¹

1. Independent researcher

1. What the work claims

The authors build PANDA, a paired hate-speech/counterspeech corpus in Mandarin Chinese, and present it as the first such resource for an East Asian language. They pool the hateful rows of three open corpora (COLD, SWSR, and CHSD; Table 1, row "Used In Preprocessing", 26,420 instances), score them with a Llama-3.1 70B Instruct model calibrated on 19 hand-scored examples, and keep the subset that maximises $\log(\text{average hate score}) \times \log(\text{average text length}) \times \text{number of instances}$, which lands at a minimum length of 53 characters and a minimum hate score of 51 (§3.3, Figure 2). For every remaining item, a simulated annealing loop (§3.4.1) asks a random choice of six open LLMs for candidate replies, prunes near-duplicates by Hamming distance, scores candidates pairwise with JudgeLM, and keeps the top four, which are then ordered by a round-robin tournament. Four annotators label the source items as hate speech (1), counterspeech (-1), or neither (0), pick the best of the four candidates, and edit it (§3.5, Appendix C.1).

Two findings are reported. On the first 785 annotated items, only 41.3% were confirmed as hate speech, 31.0% were counterspeech, and 27.7% neither (§3.6, Figure 3). And JudgeLM ranks the human-chosen or human-edited reply below first place more often than chance would suggest, which the authors read as a systematic preference for AI-style text that restates the hateful message (§3.6, Table 3, Table 4, §4.2).

2. Whether the conclusion holds on the data

I think the paper has two genuinely useful things in it, and I want to name them before anything else. The first is the corpus itself: a paired Chinese resource with a public repository (footnote 1) fills a real gap, and Table 1 is a helpful map of what exists and under which terms. I checked every row of Table 1, and the parts add up exactly (for example, COLD: $18,041 + 19,439 = 37,480$; the preprocessing row: $18,041 + 894 + 7,485 = 26,420$). The second is the honest negative result. Many dataset papers hide the fact that their filter lets through a lot of non-hateful text; this one puts the 41.3 / 31.0 / 27.7 split up front, and the Table 4 example makes the judge problem visible in a way a single number cannot.

Where I would ask for more care is the distance between those findings and the wording of the abstract and §4.

2.1 Size of the verified corpus. §4.1 speaks of “~12,000 HS–CS pairs” and, a few lines later, of 785 human-processed pairs out of “the proposed 2,974”. If 2,974 is the size of the filtered subset, then $2,974 \times 4 = 11,896$, so the ~12,000 figure appears to count the four machine candidates per item rather than verified pairs. That should be said directly. And of the 785, only 41.3% were confirmed as hate speech: $0.413 \times 785 \approx 324$. So the part of PANDA that is both human-checked and truly hate speech is, as far as I can tell from the text, about 324 pairs. That is still worth publishing. It is simply a different claim from 12,000.

2.2 Who mislabeled the data. The abstract says the analysis gives “strong evidence for the lack of open-source, properly labeled Chinese hate speech data”. But the 785 items passed two filters: the source label and the Llama-3.1 70B hate score of at least 51 (§3.3). The 58.7% that annotators rejected (31.0 + 27.7) could come from either filter. Without annotating a random sample of the *unfiltered* source rows, the paper cannot separate source-label error from LLM-scorer error. I would soften the abstract or add that sample. It would also help to report the split per source corpus, because CHSD is itself built from COLD, CDIAL, and SWSR (§3.2), so the same sentence may have entered the pool twice. I found no deduplication step in §3.3.

2.3 The JudgeLM test. This is the part I would change most. Table 3 reports a one-tailed one-sample t-test of the mean rank of the human-preferred answer against 1.5. A mean rank above 1.5 means only that the human answer is not almost always first. Under a judge with no preference at all, and four candidates, the expected rank is $(1 + 2 + 3 + 4) / 4 = 2.5$; with a fifth, human-edited candidate, it would be 3. So the test as run cannot distinguish “JudgeLM is biased against humans” from “JudgeLM is indifferent”. I would test against the no-preference expectation, or, more simply, report the share of items where the human

answer ranks first and compare it with 1/4 (or 1/5) by a binomial test. Ranks are ordinal, so a Wilcoxon signed-rank test is a better fit than a t-test. Table 3 also gives no per-annotator n, no means, and no degrees of freedom, so I could not check any of the t values (Annotator 1, $t = 13.3$; Combined, $t = 18.7$).

There is a design issue under the test as well. The four candidates were chosen by *JudgeLM* as its own top four (§3.4.1, final step). If an annotator picks one of them unedited, the human choice is already a *JudgeLM* favourite; if the annotator edits it, the edited text competes against three *JudgeLM*-selected rivals. Either way, the comparison is partly circular. It would help to separate edited from unedited selections in Table 3 and to report edit length, because a judge that likes longer text will penalise the short, direct edits that Table 4 shows.

§4.2 also states that the average rank "was significantly lower than first place". Given that rank 1 is best (§3.6), the intended meaning is numerically higher, i.e., worse. Worth fixing, since readers will quote it.

2.4 Contributions that the results do not yet show. The introduction lists "novel metrics for assessing the quality and effectiveness of generated Chinese counterspeech", and §2.5 says "the current study uses" *JudgeLM*, BLEU, ROUGE-L, BERTScore, Novelty, and Genlen. I could not find a single BLEU, ROUGE-L, BERTScore, Novelty, or Genlen value anywhere in the paper. None of the listed metrics is new either. The third contribution, "a comprehensive annotation scheme based on established CS strategies", also does not match §3.5: annotators assign 1 / 0 / -1 and choose a reply, but the eight strategies of §2.3 are never annotated. Either add these results or narrow the contribution list; both are fair options.

2.5 The Table 4 example. I agree with the authors that it is telling, and I would push the reading further. The AI response does not only rephrase the hateful text. It invents facts about a named person (that she "actively participates in community activities"), which is fabrication, not style. That is a stronger and more specific failure than "superficial alignment", and it matters for anyone who reuses the unverified candidates.

3. Reproducibility

The public repository (footnote 1) and the full annotator instructions (Appendix C.1) are real strengths. I could not inspect the repository from the Qeios text, so my comments are about what the paper itself documents.

- **Model versions.** The generator list in §3.4.1 gives names, not checkpoints. "Meta-Llama-Large" does not identify a model, and "Nous-Hermes-Mixtral" matches several releases. *JudgeLM* is cited through

the ranking paper [20], but the JudgeLM model size used (it exists in more than one size) is not stated. Nor is the exact Llama-3.1 70B checkpoint or its quantisation.

- **Decoding.** No temperature, top-p, maximum length, or number of samples per call is reported for any of the six generators, the hate scorer, or JudgeLM. No seeds.
- **Prompts.** None are given: not the hate-scoring prompt, not the generation prompt, not the JudgeLM prompt that scores "relevance, fluency, and effectiveness". The 19 anchor examples of §3.3 are also not listed.
- **Annealing hyperparameters.** The base B of the Boltzmann term, the Hamming threshold d, the Latin-character ratio cut-off, the size of the neighbour set, the iteration count, and the stopping threshold are all unstated. The word list used to perturb candidates is not given. The acceptance probability in §3.4.1 is normalised over the candidate set C, which is good, but without B its behaviour cannot be reproduced.
- **Initialisation.** " $c_0 = h$ or an empty string" leaves open which one was used, and for how many items.
- **Hamming distance.** It is defined only for strings of equal length; the paper should say how unequal-length candidates were compared.
- **Compute and annotation effort.** No GPU type or hours. "Several hours" of annotator time (§4.1) could be given as a number, together with how items were split among the four annotators and whether any item was labelled twice.
- **Inter-annotator agreement.** None is reported. With four annotators and a three-way label, this is the single most useful number a reader of a dataset paper wants.
- **Figures.** Figures 1–4 are images, and as rendered on Qeios I could read only their captions, so my comments on Figures 2–4 rely on the numbers given in the text.

4. What specifically to fix

1. **§4.1 and the abstract.** State three numbers side by side: items in the filtered subset (2,974?), machine candidate pairs (about 12,000, if that is $2,974 \times 4$), and human-verified pairs (785, of which roughly 324 are confirmed hate speech). Call only the last one the corpus, or label the tiers clearly in the release.
2. Rerun the Table 3 analysis against the no-preference expectation (2.5 for four candidates) or as a first-place binomial test, and add n, mean rank, and df per annotator.

3. In Table 3, split human selections into unedited and edited ones. The unedited case is a JudgeLM candidate by construction.
4. §4.2, "significantly lower than first place": reverse the direction of the wording.
5. **§2.5 and Contributions.** Report BLEU, ROUGE-L, BERTScore, Novelty, and Genlen for the AI candidates versus the human-edited replies, or remove the list and the claim of novel metrics. The Novelty definition refers to "training data"; say what the training data is here, since no model is trained.
6. Contribution 3 should match §3.5. Either annotate the eight §2.3 strategies on at least a subset, or describe the scheme as the three-way label plus selection.
7. **§3.6 / Figure 3.** Annotate a random sample of unfiltered source rows so that source-label error and Llama-scorer error can be told apart, and give the label split per source corpus.
8. §3.3: add a deduplication step (CHSD overlaps COLD and SWSR by construction) and report how many duplicates were removed.
9. Inter-annotator agreement (Cohen's or Fleiss' κ) on a doubly labeled subset.
10. §3.4.1: publish all hyperparameters, the word list, every prompt, and exact model checkpoints with decoding settings; replace "Meta-Llama-Large" with a real identifier.
11. **§2.2 against §3.2 and Table 1.** §2.2 says four of the six datasets are "readily accessible", Table 1 marks three as open source, and §3.2 says three were used. Please reconcile.
12. §3.1 calls GPL a "permissive" license. GPL is a copyleft software license. For a dataset, a data license (for example CC BY or CC BY-SA) is clearer, and the release terms must also respect the licenses of COLD, SWSR, and CHSD.
13. Appendix A describes BLEU as incorporating "semantic understanding". BLEU is n-gram overlap; BERTScore or a sentence-embedding distance would be the semantic option.
14. **Appendix B and Table 4.** Appendix B promises de-identification, yet the Table 4 example keeps two personal names. I would mask them in the paper and check the release for the same issue.
15. Flag in the release which AI candidates were never human-checked, given the fabricated details visible in Table 4.
16. The claim of being the first East Asian counterspeech corpus would be stronger with a sentence on how the search for Japanese and Korean resources was done.

5. Recommendation

Major revisions. The resource is real, openly released, and needed, and the finding that an off-the-shelf judge favours paraphrase (and even invented detail) over direct human rebuttal is worth reporting. The fixes are larger than cosmetic, though: the verified corpus size has to be stated plainly, the JudgeLM test needs the right null, the promised metric results and agreement figures are missing, and the generation pipeline cannot yet be rerun from the text. None of this needs a new study. It needs the numbers the authors very likely already have.

Declarations

Potential competing interests: No potential competing interests to declare.