

Peer Review

Review of: "LLM Confidence Evaluation Measures in Zero-Shot CSS Classification"

Evgenii Arsentev¹

1. Independent researcher

1. What the work claims

The paper asks a practical question that anyone who has labelled social-science text with an LLM will recognise: once the model has produced its labels, which of them should a human look at first? The authors compare five zero-shot confidence signals on three models (Llama-3.1 8B Instruct, Flan UL2, and GPT-4o, §4) and three annotation tasks (SemEval-2016 stance, IBC ideology with sub-sentential annotations, and GVFC gun-violence frames). Two signals are verbalised by the model: a number from 0 to 100, and a word from "no" to "absolute" mapped onto 0, 0.25, 0.50, 0.75, and 1 (§3.1). Two come from token log probabilities over the constrained label set: the gap between the top two label log probabilities, called *C_score* (§3.2, Eq. 1), and the exponentiated log probability of the top label, called log inverse (§3.3). The fifth, *C_ensemble* (§3.4, Eq. 2), sums the *C_score* values of all models that return the same top label.

To compare them, §4.1 sorts each model's annotations from least to most confident, plots the share of all wrong labels recovered against the share of data inspected, and reports the area under that curve. Table 1 is the whole of the quantitative evidence. In its AVG column, *C_Ensemble* scores 66.8, *C_Score* 64.2, Log Inverse 57.9, Qual. 55.6, and Quant. 55.0. From this, the abstract concludes that the aggregation strategy "improves upon existing methods and can be used to significantly improve human-in-the-loop data annotation processes", and §6 adds that, with a single model, the top-two log-probability gap is also effective. Appendix A gives labelling accuracy per model and task.

I like the framing. The evaluation target is the thing a practitioner actually pays for (human review hours), the signals chosen are cheap, and the result that verbalised confidence from small open models is close to useless while the logit gap is consistently useful is worth having on record. The paper is also

honest in §8 about its small scope. My concerns are about how far Table 1, on its own, can carry the claims made in the abstract and §5.

2. Whether the conclusion holds on the data

The averages are right; the word "robust" is not supported by them. I recomputed the AVG column from the nine cells of each row in Table 1. C_Ensemble: $71.4 + 65.0 + 66.6 + 73.2 + 54.3 + 69.3 + 73.6 + 58.4 + 69.1 = 600.9$, and $600.9 / 9 = 66.77$, which rounds to the reported 66.8. The other four rows also reproduce (C_Score $577.5 / 9 = 64.17$; Log Inverse $521.5 / 9 = 57.94$; Qual. $500.7 / 9 = 55.63$; Quant. $495.15 / 9 = 55.02$). So the arithmetic is fine. What the averages hide is that C_Ensemble is the best row in only five of the nine columns. As rendered on Qeios, the bold marking promised in the Table 1 caption is lost, so I identified the column winners myself:

- GPT, GVFC: C_Score 67.2 and Log Inverse 66.7 both beat C_Ensemble 66.6.
- Flan, IBC: Log Inverse 60.8 and C_Score 60.3 beat C_Ensemble 54.3, a gap of 6.5 points.
- Llama, IBC: C_Score 62.7 beats C_Ensemble 58.4.
- GPT, IBC: C_Ensemble 65.0 wins, but only by 0.4 over Quant. 64.6.

On IBC, the ensemble is below the single-model C_Score for two of the three models. That is a systematic pattern on one task, not noise scattered across the table, and it deserves an explanation (IBC accuracies in Appendix A sit between 62.3 and 65.2 for all three models, so pooling three weak voters may simply add correlated error). The claim in §5 that the ensemble is "proving to be effective across all model types" is true per model on average, but "most robust" in the caption and in §5 is not true per task.

There are no uncertainty estimates at all. The paper never states how many items were annotated in each dataset, so I cannot even estimate how much an AUC value would move under resampling. There are no confidence intervals, no bootstrap over items, no repeated runs, and no test of any difference. The C_Ensemble vs C_Score gap on the AVG column is $66.8 - 64.2 = 2.6$ points, averaged over cells that disagree in sign. Whether that is distinguishable from zero depends entirely on the unreported sample sizes. The abstract's "significantly improve" therefore has no statistical basis in the text; either run a paired bootstrap on items (the ranking of the same items under two signals is naturally paired) or drop the word.

The metric needs its reference points. A signal that ranks items at random recovers, in expectation, x% of the errors after inspecting x% of the data, so its AUC is 50. The paper never says this, and it matters for

reading Table 1: three cells are below random. Table 1, row Log Inverse, column GPT/IBC is 42.4; row Quant., column Flan/GVFC is 46.25; row Quant., column Flan/Stance is 49.8. The first of these means that for GPT-4o on IBC, the most confident labels are more likely to be wrong than the least confident ones. That is a striking finding, and §5 passes over it with a sentence about GPT returning "less deterministic responses". I would want to see it confirmed and discussed.

The ceiling matters just as much. A perfect ranker places all errors first, reaches 100% recall after inspecting a fraction e of the data (e is the error rate), and stays there, so its AUC is $1 - e/2$. Using the Appendix A accuracies, that ceiling differs across columns. For GPT-4o on Stance, accuracy 77.4 gives $e = 0.226$ and a ceiling of $1 - 0.113 = 88.7$. For Flan UL2 on GVFC, accuracy 58.7 gives $e = 0.413$ and a ceiling of $1 - 0.2065 = 79.35$. Averaging raw AUCs across columns with ceilings nine points apart treats unlike quantities as like. A normalised score, $(\text{AUC} - 50) / (\text{ceiling} - 50)$, would be more honest. As an illustration of why: Flan/IBC, row C_Ensemble, is 54.3 against a ceiling of $1 - 0.377/2 = 81.15$, so it captures $(54.3 - 50) / (81.15 - 50) = 4.3 / 31.15 = 0.14$ of the achievable gain. The headline differences may survive normalisation. I simply could not tell from the text.

The novelty of the metric is overstated. Contribution (1) in the abstract presents the measure as new. The curve in §4.1 is, as far as I can see, the cumulative-gains (lift) curve for the binary task "is this label wrong", scored by confidence; the misclassification-detection literature usually reports AUROC or AUPR for the same ranking (Hendrycks and Gimpel's 2017 baseline paper is the usual reference), and the selective-classification literature uses the area under the risk-coverage curve. I do not object to the choice. It is arguably more readable for annotation managers. But §2 should position it against these, and ideally Table 1 should report AUROC as well, so that readers can compare with prior UQ work.

The ensemble comparison is not resource-matched. §3.4 concedes that C_ensemble is "more resource intensive", requiring three models, but it is compared only with single-model signals. The obvious question is how much of its advantage comes from the log probabilities and how much from plain agreement. A zero-cost ablation using the same runs would settle it: rank items by the number of models agreeing with the target model's label (0, 1, or 2 others), breaking ties by that model's own C_score. If agreement count alone gets most of the way to 66.8, the paper's recommendation changes: practitioners without log-probability access (§8 names this constraint) could still use it.

Ties are unaddressed, and they matter for two of the five rows. AUC here depends only on the ranking of items. The qualitative self-report has five possible values (§3.1), so large groups of items share a rank, and a 0–100 score from instruction-tuned models often clusters on a few round numbers. How items

with equal confidence are ordered within a percentile threshold can change the AUC substantially, and a favourable or unfavourable tie-break (for example, by dataset order) could explain part of the gap between the verbalised rows and the logit rows. The paper should state the tie-breaking rule and, preferably, report the expected AUC under random tie-breaking.

Two smaller interpretive points. §4 and §5 split the models into the RLHF model (GPT-4o) and "non-RLHF models, Flan and Llama". Llama-3.1 8B Instruct is preference-tuned, so the dichotomy that the self-report findings are hung on is at best partly true and should be described more carefully. And the caption of Figure 1 says roughly half of the errors are found in the bottom 20% for Flan UL2 stance. As rendered on Qeios, Figure 1 and all Appendix B plots are not legible, so I could not check that statement against the curve.

3. Reproducibility

The good news first. §7 links a code and data repository, gives the local compute (two NVIDIA A6000 GPUs for 18 hours), and even discloses that GPT was used to debug plots. The datasets are public benchmarks with citations. That is more than many short papers offer. I could not open or audit the repository from the review text, so everything below concerns what the paper itself states.

- **Model identity is inconsistent.** §4 names Llama-3.1 8B Instruct. The Appendix A table has columns FLAN UL2, GPT-4o and *Mistral 8b*, with no Llama column. Either the appendix reports a different model or the header is wrong; readers cannot tell which accuracies belong to the Llama column of Table 1.
- **No snapshot for GPT-4o.** "GPT-4o" is an alias that has pointed to several dated releases. No date or snapshot is given, nor the access period.
- **No decoding parameters.** Temperature, top-p, and maximum tokens are not reported for any model. This is not a side detail here, since §5 explains a result by GPT "return[ing] less deterministic responses", which is exactly what temperature controls.
- **Log-probability extraction is underspecified.** The OpenAI API returns only the top-k alternatives at each position. How were label log probabilities obtained when a label was absent from the top-k? How were multi-token labels (stance labels, frame names) scored: first token, sum over tokens, or something else? Did Flan UL2 and Llama use full-vocabulary logits restricted to the label set, or generation with top-k? Eq. 1 in §3.2 also does not render cleanly on Qeios; its definition of the "second-highest" term is hard to read.

- **The ensemble target label is ambiguous.** Eq. 2 sums over models whose top label equals t^* , but Table 1 reports a separate C_{Ensemble} value per model column. I assume that, for the GPT column, t^* is GPT's label and the sum runs over the other models that agree with it, but this is not stated. Nor is it said whether the summed C scores are on comparable scales across models (they are unnormalised log-probability gaps, so the sharpest model will dominate).
- **Dataset slices and sizes are missing.** SemEval-2016 Task 6 has several targets and a train/test split; IBC and GVFC have their own label schemes and partitions. The number of items, the number of classes, and the split used for each task are needed to interpret Appendix A accuracies (58.3 on GVFC means little without the class count) and to judge sampling error in Table 1.
- **Prompts.** §3.1 says prompts are in the repository. The paper should still reproduce at least one full prompt per signal in an appendix and say whether the confidence was elicited in the same call as the label or in a follow-up turn, since that changes what the self-report measures.
- **AUC computation.** The percentile grid (1%? 5%?) and the integration rule are not stated.
- **Link permanence.** The anonymous.4open.science link was appropriate for blind review; on a published Qeios version, it should be replaced with a permanent, versioned archive.

No seeds are mentioned. For locally run models with greedy decoding, that may be fine, but the paper should say so.

4. What specifically to fix

1. **Abstract and §5, first paragraph; Table 1 caption.** Replace "significantly improve" and "most robust" with statements the table supports: C_{Ensemble} has the highest average (66.8 vs 64.2 for C_{Score}) and wins five of nine task-model combinations, losing to single-model signals on GPT/GVFC, Flan/IBC and Llama/IBC.
2. **§4.** Report item counts, number of classes, and the split used for each of SemEval-2016, IBC, and GVFC.
3. **Table 1.** Add 95% confidence intervals from a bootstrap over items, and a paired bootstrap test for C_{Ensemble} vs C_{Score} per column. Restore the bold markings so the caption matches the table.
4. **§4.1.** State explicitly that a random ranking scores 50 and that the oracle scores $1 - e/2$; consider adding a normalised column.
5. **§5, discussion of GPT-4o.** Address the three sub-random cells: Log Inverse on GPT/IBC (42.4), Quant. on Flan/GVFC (46.25), and Quant. on Flan/Stance (49.8). Inverted confidence on GPT/IBC

should be checked for a sign or label-mapping error before it is interpreted.

6. Appendix A header: Mistral 8b or Llama-3.1 8B Instruct? Fix it, and align the column order with Table 1 (GPT, Flan, Llama).
7. §3.4, Eq. 2. Define t^* per column, say whether C scores are normalised before summing, and add the agreement-count ablation described in Section 2 of this review so the contribution of log probabilities is isolated from simple model agreement.
8. §3.1. State how ties are broken when ranking the qualitative (five-level) and quantitative scores, and report AUC averaged over random tie-breaks.
9. §3.2 and §3.3. Describe exactly how label log probabilities were obtained for each model (top-k limits, multi-token labels), and re-typeset Eq. 1, which does not render legibly on Qeios.
10. §7 needs the GPT-4o snapshot, access dates, temperature, top-p, and max tokens for all three models.
11. §2. Relate the proposed measure to AUROC/AUPR for misclassification detection and to risk-coverage curves, and temper "we propose an uncertainty quantification (UQ) performance measure" in the abstract accordingly. Reporting AUROC alongside the curve AUC would help comparison.
12. §4 and §5. Reword the RLHF vs non-RLHF split; Llama-3.1 8B Instruct is preference-tuned.
13. Minor: Table 1, row Quant., column Flan/GVFC reports 46.25 with two decimals while every other cell has one; §1 mentions "multi-modal systems," which the study does not touch; typo "misclassified" in §4.1.

5. Recommendation

Major revisions. The question is useful, the design is sensible, and the one table the paper offers is internally consistent. But that table alone, with no sample sizes, no intervals, no random or oracle reference, an unresolved model-identity mismatch in Appendix A, and a winner that loses on four of nine columns, does not support the abstract's claim of significance or the caption's "most robust." None of the required fixes needs new models or new datasets; almost all can be computed from the runs the authors already have. With those added, I would expect this to become a solid, practical short paper.

Declarations

Potential competing interests: No potential competing interests to declare.