

Peer Review

Review of: "Enhancing Annotated Bibliography Generation with LLM Ensembles"

Evgenii Arsentev¹

1. Independent researcher

Summary

The paper proposes a three-tier pipeline for producing annotated bibliographies with LLMs: a generation tier that varies temperature, top_k, and top_p to get diverse candidate annotations (Section 2, Level 1), an LLM-as-a-judge tier that scores those candidates on relevance, accuracy, and coherence (Level 2), and a summarization tier that merges the selected candidates and strips redundancy using sentence-similarity techniques (Level 3). Two selection rules are defined in Level 3: "top temperature," which keeps responses generated at the temperature with the highest average rating, and "top M responses," which keeps the M highest-rated responses outright. The empirical support is a single table. Section 3, Table 1 reports four rows — Baseline (Individual), Mean Individual, Top M Responses, Top Temperature — against two columns, Avg. Sentence Length and Readability Score, obtained with Gemini 1.5 flash and Gemini 1.5 pro. The Abstract turns that table into the headline pair: "a 38% improvement in annotation quality and a 51% reduction in content redundancy." The architecture is reasonable, and the author is candid that the validation is "preliminary" (Abstract; Section 3, first sentence). My concerns are almost entirely about the gap between what Table 1 measures and what the Abstract and Conclusions say it shows.

Does the claim hold on the data

Not as stated. Start with the two headline numbers and trace them back. The 38% is the Top M Responses readability score of 31.41 against the Baseline (Individual) readability score of 22.71 (Table 1, rows 3 and 1, column "Readability Score"), which is 38.3% — the arithmetic is right. But the Abstract calls this "improvement in annotation quality." Readability score is not annotation quality. It is a surface statistic. The 51% is the Top Temperature average sentence length of 19.11 against the baseline 39.00

(Table 1, rows 4 and 1, column "Avg. Sentence Length"), which the Abstract relabels as "reduction in content redundancy." Sentence length is not redundancy either. No redundancy metric appears anywhere in the paper, even though redundancy removal is a named component of Level 3 (Section 2). So both headline quantities are renamed proxies, and the names they are given are the two things a reader most wants measured.

There is also a splice. The 38% comes from Top M Responses, and the 51% comes from Top Temperature — Section 3, paragraph 2 says so explicitly ("Top Temperature outperforms Top M in conciseness, achieving a 51% (vs. 44%) reduction"). The Abstract presents both as properties of "the LLM ensemble," as though one configuration delivered them together. Per Table 1, no single row does. Top M gets 41.5% on length, Top Temperature gets 17.6% on readability. The best-of-both framing is a composite of two different systems.

Then the metric-independence problem, which I think is the most serious technical issue. Readability formulas in the Flesch/Flesch-Kincaid/ARI family are, almost entirely, a function of words per sentence and syllables per word. Table 1's two columns, therefore, are not two pieces of evidence — they move together by construction. Look at the ordering: sentence length descends 39.00, 34.80, 22.80, 19.11 down the column, and readability climbs 22.71, 25.01, 31.41, 26.71 with only Top Temperature breaking rank. Shortening sentences mechanically raises the readability score. Section 3, paragraph 1 nonetheless treats the two as jointly confirming that the ensembles "far outperform the baseline ... in terms of both average sentence length and readability score." One effect reported twice is not two effects. And since the summarization tier's explicit job (Section 2, Level 3) is to shorten and compress, the pipeline is being credited for producing exactly the artifact its final stage is designed to produce.

Next, the claim the Abstract makes about "coherence and relevance" — "the combined outputs from the LLM ensemble improve coherence and relevance compared to individual responses." Table 1 has no coherence column and no relevance column. The LLM judge in Level 2 is described as assigning "numerical ratings" on relevance, accuracy, and coherence (Section 2, Level 2), and Section 3, paragraph 4 says the judge "assigns numerical ratings that reflect their quality, relevance, coherence, and factual accuracy" — but not one of those ratings is reported. The judge is the load-bearing component of the whole architecture, and its output never reaches the reader. Coherence and relevance are asserted, not shown.

On sample size and inference: Section 3, paragraph 1 says "several preliminary experiments." That is the only statement of scale in the paper. How many topics, how many source references per bibliography,

how many generations per hyperparameter setting, how many repeats — none of it is given. Table 1 carries no dispersion, no interval, no n, and no test, yet Section 3, paragraph 1 says the ensembles “far outperform” the baseline. With sampling temperature in play, run-to-run variance is exactly what needs bounding, and the paper cites the relevant work itself (reference [13], Atil et al., “LLM stability”). A gap of 22.71 to 31.41 may well be real; on the evidence in the file, there is no way to tell whether it exceeds the spread of repeated runs at fixed settings.

The baseline choice needs defending too. Baseline (Individual) has an average sentence length of 39.00, while Mean Individual has 34.80 (Table 1, rows 1 and 2) — the baseline is worse than the mean over individual configurations on that column, and worse than it on readability as well (22.71 vs. 25.01). Neither row is defined anywhere in the text. If “Baseline” is a single default configuration, then the headline 38% is measured against a below-average single point, and the honest comparator for an ensemble is the best individual configuration, not a default one. That comparison is absent. The Mean Individual row is the more informative one, and it shrinks the effect considerably.

Finally, compute budget. The ensemble generates many candidates, runs a judge pass over them, then runs a summarizer pass. The baseline generates one response. They are compared under one table with no accounting of tokens, calls, latency, or cost anywhere in the paper. Some of the observed gain is simply more inference, and the paper cannot separate “ensembling helps” from “more sampling plus a compression step helps.”

Reproducibility

As the manuscript stands, nobody can rerun this. The missing items are specific, and most are one line each:

Section 3, paragraph 1 names “Gemini 1.5 flash and Gemini 1.5 pro” with no snapshot identifier, no API version, and no access date — both are versioned model families whose behavior differs between releases. Which model occupies which tier is never stated: the reader cannot tell whether pro is the judge and flash the generator, whether the judge and the generators are the same model (which would raise self-preference bias, a known failure mode documented in the very survey cited as [17]/[26]), or whether the summarizer in Level 3 is a third distinct model despite the text calling it “a third LLM.” Section 2, Level 1 says temperature, top_k, and top_p were varied and that “a particular optimization of the LLM hyperparameters is additionally done,” but no grid, no ranges, no optimizer, and no resulting values appear. M in “top M responses” (Section 2, Level 3) is never given a value. The readability formula is never

named — Flesch Reading Ease, Flesch-Kincaid, SMOG, and ARI produce entirely different numbers on the same text, and without the formula and library, the 22.71 and 31.41 in Table 1 cannot be recomputed or even interpreted directionally with confidence. The "sentence similarity techniques" for redundancy removal (Section 2, Level 3) are named only as a category: no embedding model, no similarity threshold. No seeds, no prompts, no judge rubric or rating scale, no topic list, no source corpus, no code or data availability statement, and no cost figures.

One structural point belongs here. Section 1 promises that "Section 3 details the experimental setup, results, and analysis." Section 3 is titled "Results and discussion" and contains no experimental setup at all. The manuscript has no methods section in the usual sense — Section 2 describes the architecture in the abstract, and the empirical protocol is simply absent. That single omission accounts for most of the list above.

Specific points

1. **Abstract, headline claims vs Table 1.** "38% improvement in annotation quality" is the Readability Score column ($22.71 \rightarrow 31.41$, Table 1 rows 1 and 3) and "51% reduction in content redundancy" is the Avg. Sentence Length column ($39.00 \rightarrow 19.11$, Table 1 rows 1 and 4). Neither quantity measures what the Abstract names. Either rename them to what they are (readability, sentence length) or add measurements of annotation quality and of redundancy.
2. **Abstract: "improve coherence and relevance" is unmeasured.** No coherence or relevance number appears in Table 1 or anywhere else. The Level 2 judge produces exactly these ratings (Section 2, Level 2) — report them. A table of per-criterion judge ratings by configuration would do more for the paper than anything else on this list.
3. **Section 3, paragraph 2: the "23%" does not follow from Table 1.** The text says Top M achieves "a 23% improvement over the mean individual readability score of 25.01." From Table 1, $(31.41 - 25.01)/25.01 = 25.6\%$. Under the reverse convention, $(31.41 - 25.01)/31.41 = 20.4\%$. Neither is 23%. The paper's other figures in the same paragraph use the first convention and truncate ($38.3 \rightarrow "38\%"$, $17.6 \rightarrow "17\%"$, $6.8 \rightarrow "6\%"$), which would give "25%" here. Please recheck.
4. **Section 3, paragraph 2: the "44%" does not follow from Table 1.** The text gives Top M a "44%" length reduction against the baseline. From Table 1, $(39.00 - 22.80)/39.00 = 41.5\%$. The paired figure for Top Temperature in the same sentence, 51%, is exactly right $((39.00 - 19.11)/39.00 = 51.0\%)$, which makes the 44% look like a transcription slip rather than a different convention. Same sentence:

"35%" against the mean individual result computes to 34.5% $((34.80 - 22.80)/34.80)$, a rounding inconsistency with the truncation used elsewhere. Minor, but in a paper resting on eight percentages, two of which do not reproduce, every one of them now has to be rechecked.

5. **Table 1: the two columns are not independent evidence.** Standard readability indices are dominated by words per sentence. Since the Level 3 summarizer shortens text by design, a rise in readability is close to guaranteed by a fall in sentence length. Section 3, paragraph 1 presents both columns as jointly demonstrating superiority. Add at least one metric that is not a function of surface length — judge ratings, a human rating, factual-support checking against the cited sources, or coverage of the source's key claims.
6. **Table 1, rows 1 and 2: "Baseline (Individual)" and "Mean Individual" are never defined.** The text does not say what configuration the baseline is, nor what population the mean is taken over. Note the baseline is worse than the mean on both columns (39.00 vs 34.80; 22.71 vs 25.01), so the 38% headline is measured against a below-average single configuration. Add a best-individual-configuration row; that is the comparator the ensemble claim needs to beat.
7. **Section 3: no sample size, no dispersion, no test.** "Several preliminary experiments" (paragraph 1) is the only quantification of scale, and "far outperform" (same sentence) is a strength claim Table 1 cannot support without n , variance across repeated runs at fixed settings, and some interval. Temperature sampling makes this non-optional. Also state whether any generation failed to produce a parseable annotation or a parseable judge rating, and how those cases were handled — if failures were dropped, the reported means are over survivors.
8. **Section 2, Level 2: the judge is not validated.** The text argues that "the LLM-as-a-judge approach can achieve greater objectivity than that obtained using traditional metrics" citing [27]. That is stronger than the literature supports, and the survey the paper itself cites as [17] and [26] catalogues position, verbosity, and self-preference biases. Verbosity bias matters here specifically because the pipeline's selection step and its length-based metrics interact. Some agreement check against human ratings on even a small subset would change the standing of every downstream number.
9. **Section 2, Level 1 and Section 3: the "ensemble of multiple LLMs" is one model family.** The Abstract and Section 1 describe "multiple LLMs in different roles," but the only models named anywhere are Gemini 1.5 flash and Gemini 1.5 pro (Section 3, paragraph 1), and the diversity mechanism described in Level 1 is hyperparameter variation, not model variation. Either soften the framing to hyperparameter-diversity within a family, or add a genuinely heterogeneous model to the pool. If the judge is also Gemini 1.5, say so and discuss self-preference.

10. **Compute budget is not reported.** Nowhere in Section 3 or Table 1 is there a count of generations per condition, tokens, API calls, latency, or cost. The ensemble rows and the baseline row are not matched on inference budget, so the table cannot separate the contribution of the architecture from the contribution of spending more. A cost column, or a budget-matched baseline (for example, best-of-N under the same token spend), would settle it.
11. **Section 2, novelty positioning against [24] and [25].** The three-tier design is introduced as proposed here while citing Farr et al. [24] for "LLM chain ensembles" in the same sentence, and the diversity-maximizing generation idea sits close to LLM-TOPLA [25], cited in Level 1 and again in Level 3. Neither is contrasted with the present work. One paragraph saying what is new relative to those two would protect the contribution claim.
12. **Section 1, paragraph 2 vs reference [13].** The text asserts that LLMs deliver "improved consistency and objectivity compared to human experts," unsupported, while paragraph 3 cites Atil et al. [13], "LLM stability: A detailed analysis with some surprises," which documents the opposite tendency. Either qualify the consistency claim or drop it.
13. **No example output.** For a paper about annotation quality, there is not a single generated annotation shown — not a baseline one, not an ensemble one. An appendix with one input reference and the outputs from each of the four Table 1 conditions, plus the judge's ratings for them, would let readers see what a 38% readability gain actually looks like on the page.
14. **Reference list: [17] and [26] are the same paper.** Gu J, Jiang X, Shi Z, Tan H, Zhai X, Xu C, Li W, Shen Y, Ma S, Liu H, Wang Y, Guo J (2024), "A survey on LLM-as-a-judge" (arXiv:2411.15594) appears as two separate numbered entries. Merge them.
15. **Table 1 caption and presentation.** The caption describes the comparison but gives no units, no readability formula, no n, and no indication of what is averaged over. Everything a reader needs to interpret 22.71 and 31.41 is outside the table and, in the case of the formula, outside the paper.
16. **Section 3, paragraph 2: the explanation offered for Top M beating Top Temperature is speculative and flagged as such ("presumably because it selected responses based on their individual ratings").** With the judge ratings in hand, this is testable rather than presumed — correlate the judge's coherence/organization ratings with the readability score and report it.

On the extraction I worked from: Table 1 came through cleanly in the version I received — four rows, two numeric columns, values as quoted above — so I have treated those figures as authoritative. There are no figures in the file, and no appendix or supplementary material accompanies it.

Recommendation

Major revisions. The architecture is sensible, the writing is clear, and the author is upfront that the validation is preliminary — but the paper’s stated conclusions are about annotation quality, coherence, relevance, and redundancy, and the only evidence offered is two mutually dependent surface statistics over an unreported number of runs, with two of the eight percentages in Section 3 failing to reproduce from Table 1 and the two headline figures drawn from two different systems. Reporting the Level 2 judge ratings that the pipeline already produces, stating n with some measure of spread, adding a best individual and a budget-matched baseline, naming the models, the readability formula, and M , and rechecking the arithmetic would put the contribution on solid footing; none of that requires a new architecture, and the claims may well survive it.

Declarations

Potential competing interests: No potential competing interests to declare.