

Peer Review

Review of: "DAFE: LLM-Based Evaluation Through Dynamic Arbitration for Free-Form Question-Answering"

Evgenii Arsentev¹

1. Independent researcher

1. What the work claims

The paper proposes DAFE, a cascade for judging free-form QA answers against reference answers. Two open-weight judges (Llama-3.1 70B and Mistral 7B) each return a binary True/False verdict; when they agree, that verdict is final, and when they disagree, GPT-3.5-turbo is called as a tiebreaker, and the final label is the majority of the three verdicts (§2.3, Figure 1). Five candidate models (DeepSeek-V3, Llama-3.1 70B, GPT-3.5-turbo, Mixtral 8x7B, Mistral 7B) answer 300 randomly sampled questions from each of AmbigQA, FreshQA, HotpotQA, NQ-Open, and TriviaQA (Appendix A, seed 42), giving 7,500 outputs (§4.2). Three graduate-student annotators label every output, and their majority (HM) is the ground truth. DAFE is compared with a containment variant of Exact Match, thresholded BERTScore, five single LLM judges, and (in the appendices) a three-judge majority vote, GPT-4o as a single judge, and a reference-free setting.

The headline claims are that DAFE reaches the highest agreement with HM in every reported cell (Table 1, Cohen's kappa; Table 2, Macro-F1, caption: "DAFE consistently achieves the highest Macro-F1 across all evaluated settings"), that arbitration produces large gains where the primary judges disagree (§4.4, AmbigQA Macro-F1 72.9% to 86.6%, kappa 0.467 to 0.773), and that selective arbitration is much cheaper than majority voting (Appendix C.2: 1,318 tiebreaker calls, "82–94%" fewer arbitration calls). I want to credit what is solid here. Labeling all 7,500 outputs with three annotators is a lot of careful work, and Tables 4 and 5 report both Fleiss' kappa and raw percent agreement, which lets a reader see the kappa paradox directly (DeepSeek on TriviaQA: kappa 0.456 in Table 4 with 90.0% agreement in Table 5). The error analysis in §4.2 contains specific, checkable failure cases (the Hermann Müller example, the

Patriots/Dolphins date). Table 7 is internally consistent: I summed the Tiebreaker Usage column per candidate (273, 268, 257, 243, 277) and got 1,318, matching the Total row, and each count equals the disagreement rate times 300 after rounding (for example, 44.3% of 300 = 132.9, reported as 133).

2. Whether the conclusion holds on the data

2.1 DAFE and the three-judge majority vote are the same decision rule. This is the point I would ask the authors to address first, because it changes how every result should be read. Appendix C.5 builds a majority vote from Llama, GPT-3.5, and Mistral, the same three models DAFE uses. Consider any instance. If Llama and Mistral agree, the majority of the three is their shared verdict whatever GPT-3.5 says, and DAFE outputs the same verdict. If they disagree, the majority is decided by GPT-3.5 alone, and DAFE also outputs GPT-3.5's verdict. For deterministic judges, the two procedures return identical labels on every instance. DAFE is therefore not an approximation to majority voting; it is an exact, cheaper way of computing it (a lazy evaluation of the third vote). That is a legitimate engineering point, but it has two consequences the paper does not draw.

First, Table 8 should show identical numbers in the Majority Voting and DAFE columns, and it does not. Row Mixtral / TriviaQA gives Macro F1 95.0 and Kappa 0.90 under majority voting against 88.2 and 0.76 under DAFE. Row Mixtral / NQ-Open gives 98.3 / 0.97 against 95.6 / 0.91; row GPT / TriviaQA 90.5 / 0.81 against 88.7 / 0.77; row Mixtral / AmbigQA 98.9 / 0.98 against 97.5 / 0.95. The explanation offered in C.5, “rare instances where both the primary judges and the arbitrator make errors”, cannot produce a difference, because such instances are mislabelled identically by both procedures. Non-determinism is also an unlikely explanation, since C.6.1 reports that GPT produced identical verdicts in 100% of five repeated trials. Something else differs between the two pipelines: the prompt given to GPT-3.5 in its arbitrator role, the parsing of its output, the sample, or a bookkeeping error. Until that is identified, I do not think either column of Table 8 can be taken at face value, and the DAFE columns of Tables 1 to 3 inherit the same uncertainty.

Second, the claimed accuracy gain belongs to the three-model ensemble, not to “dynamic arbitration”. The fair comparisons for DAFE's quality are other three-judge ensembles and single judges of comparable total cost; the fair comparison for its efficiency is the number of calls saved relative to the identical majority vote. Framing the method as something that “prioritizes evaluation reliability” beyond majority voting (Abstract) is not supported.

2.2 The efficiency numbers in C.2 do not match Table 7. Aggregate savings in third-judge calls are $1 - 1,318/7,500 = 82.4\%$. Per cell, they range from $100 - 44.3 = 55.7\%$ (GPT / FreshQA) to $100 - 4.7 = 95.3\%$ (Mixtral / HotpotQA). So the stated range “82–94%” mixes an aggregate with an unexplained upper value, and the bullet “Over 90% fewer third-judge inferences” is false in aggregate (82.4%). More importantly, the primary judges run on all 7,500 outputs. Total judge calls for DAFE are $2 \times 7,500 + 1,318 = 16,318$ against $3 \times 7,500 = 22,500$ for majority voting, a reduction of $1 - 16,318/22,500 = 27.5\%$ in calls. The bullet “Up to 95% cost savings” is only true if the Llama 70B and Mistral 7B inference is treated as free, and the paper reports no GPU type, hours, or serving cost for those models. C.2 also says the \$5.40 figure is shown in Table 7; as rendered on Qeios, Table 7 has no cost column.

2.3 No uncertainty is reported for any agreement score. Every cell in Tables 1 to 3 and 8 is computed on 300 instances, and many of the differences the text relies on are one or two points of Macro-F1. Example: Table 2, candidate Mixtral, HotpotQA, DAFE 0.987 against the Llama judge 0.970. I could not compute intervals myself because the confusion matrices behind these cells are only in Figures 3 and 8 to 11, which are not legible as rendered on Qeios. “Significant improvements” in the Abstract and “significantly enhanced” in §4.4 should either be backed by a paired test (McNemar on per-instance correctness against HM, or a paired bootstrap on Macro-F1 and kappa) or reworded.

2.4 The size of some DAFE gains needs per-instance evidence. A majority of three can beat its best member when errors are fairly independent, so I do not claim the numbers are impossible. Still, some jumps are large. Table 3 (FreshQA), candidate Mixtral: the three components score Llama 0.708, GPT 0.779, Mistral 0.703, and DAFE scores 0.936, 15.7 points above the best component. Table 1, candidate Mistral, TriviaQA: Llama 0.758, GPT 0.725, Mistral 0.470, DAFE 0.854. The paper’s own error analysis (§4.2) argues that on FreshQA, judges share a systematic bias, and shared errors are exactly what majority voting cannot fix. A small table of how often each judge is right when the other two are wrong, per dataset, would settle whether these gains come from genuinely complementary errors.

2.5 The FreshQA narrative does not match Table 6. §4.2 says LLM judges on time-sensitive questions “tend to hallucinate by consistently classifying candidate model responses as True, even when incorrect.” If that held, judge accuracy estimates on FreshQA would sit above HM. Table 6 shows the opposite for two judges. Candidate GPT, FreshQA: HM 70.7, DeepSeek judge 58.0, Llama judge 51.7. Candidate Mistral, FreshQA: HM 81.7, DeepSeek 60.7, Llama 59.0. Candidate Mixtral, FreshQA: HM 86.0, Llama 67.0. These judges accept fewer answers than the annotators do. Other judges (GPT, Mistral, Mistral) do sit above HM in some of these rows, so the accurate statement is that the direction of the error is judge-dependent.

2.6 Some prose claims overstate the tables. §4.1 says DeepSeek and GPT “achieve higher Macro-F1 scores (0.97-0.98) across AmbigQA, HotpotQA, and NQ-Open.” In Table 2, the DeepSeek column reaches 0.97 or above in one cell only (candidate GPT, AmbigQA, 0.976); its NQ-Open values are 0.898, 0.904, 0.912, and 0.898. The GPT column’s NQ-Open values are 0.896, 0.911, 0.863, and 0.919. The Llama 70B judge beats DeepSeek in several cells (candidate GPT, HotpotQA: 0.977 against 0.903), which weakens the “scaling law in evaluation capability” sentence in the same paragraph. The quoted EM/BERTScore example is correct: Table 6, candidate Llama, AmbigQA gives EM 42.3 and BS 63.0, and Table 2 gives 0.744 and 0.641.

2.7 The appendix ablations are too thin to support the conclusions drawn from them. C.4 states that DAFE with GPT-3.5 as arbitrator reached 0.984 Macro-F1 on HotpotQA for candidate GPT-3.5. Table 2 and Table 8 both give 0.993 / 99.3 for that cell; 0.984 is the AmbigQA value in Table 2. C.3 reports DeepSeek-as-arbitrator results on three different scales (“91.23”, “79.11”, “0.914”) without the GPT-3.5 arbitrator comparison for the same cells; for candidate DeepSeek on NQ-Open, that comparison is not in any table because Tables 1 and 2 omit candidate DeepSeek entirely. C.6 gives no sample sizes for C.6.2 to C.6.4, and C.6.3 reports kappa falling “from 0.95 to 0.72” without saying on which subset; for candidate Mistral / HotpotQA / GPT judge, the full-sample kappa in Table 1 is 0.902. C.6.2’s “2% increase in Macro-F1” from few-shot prompting also sits awkwardly with §3.3, which cites prior work that in-context examples do not help. Finally, C.6.1 (identical verdicts in 100% of trials) and C.6.3 (13% of verdicts changed between repeated verdict-only runs) describe very different stability at temperature 0, and the paper does not explain the gap.

2.8 Missing cells. §4.2 speaks of 7,500 evaluations per evaluator, but Tables 1 and 2 report 16 of 25 candidate-by-dataset cells: candidate DeepSeek is absent, and FreshQA is split into Table 3, which has no EM or BERTScore columns. The largest arbitration effect quoted in §4.4 (kappa 0.467 to 0.773) is for candidate DeepSeek on AmbigQA (per C.1), which I cannot find in any table. As rendered on Qeios, Figures 2 and 4 to 7 are not legible, so I could not verify the 72.9% to 86.6% change or the “up to 21.5 points” in C.1. I also noticed that the §3.4 definition of EM (a gold answer contained in the response) and the Appendix C wording (“requires a candidate’s response to exactly match one of the reference answers”) describe different metrics.

2.9 The human ground truth is reasonable but under-described. Inter-annotator agreement is high, and I have no reason to doubt it. Still, the annotators were volunteers from the authors’ network, were allowed to consult the internet (B.1), and saw the same reference answers as the judges. For FreshQA, whose

answers change over time, the date of annotation relative to the reference snapshot matters, and it is not given. How HM was formed for FreshQA false-premise questions is not described either.

2.10 Scope. The conclusion (§6) calls DAFE “a trustworthy and reliable framework for evaluating open-ended language generation tasks.” The evidence covers short-answer, English, reference-based factoid QA with binary labels, one zero-shot prompt, and one judge triple. That is a useful setting, but it is not open-ended generation in general, and the reference-free experiment in C.7 shows how strongly the results depend on having references. The sentence should be narrowed.

3. Reproducibility

Some things are well specified. Temperature is 0 for all models (§3.1). Hugging Face identifiers are given for Llama-3.1-70B-Instruct, Mistral-7B-Instruct-v0.3, Mixtral-8x7B-Instruct, and the DeBERTa model used for BERTScore (footnotes 1 to 4). Dataset splits and subsets are named for four of five datasets, and sampling uses seed 42 (Appendix A). The annotation guidelines are reproduced in full (B.1). The following is missing.

- **API model snapshots.** GPT-3.5-turbo, GPT-4o, and DeepSeek-V3/R1 are named only by family. OpenAI aliases have pointed to different snapshots over time, and GPT-3.5-turbo is the arbitrator on which every DAFE number depends. No run dates are given.
- **Decoding parameters beyond temperature.** Max tokens is not stated for the main runs; C.2 mentions a 2,048 max-token setting only as a hypothetical cost scenario. `top_p`, stop sequences, and system prompts for candidates are not listed in the text.
- **Serving stack.** Nothing on how the open-weight models were run: framework, precision or quantization, hardware, batch settings. Quantization, in particular, can change 7B and 70B judge verdicts.
- **Prompts.** The candidate and judge prompts are given only as images (Figures 12 to 15), which are not legible as rendered on Qeios. The arbitrator prompt, the majority-voting prompt (see 2.1), and the reference-free prompt of C.7 are not shown at all. Verdict parsing and the handling of unparseable outputs are not described.
- **FreshQA version.** FreshQA is updated regularly. Appendix A gives splits for the other four datasets but no version or date for FreshQA, and the example in §4.2 (reference “November 24, 2024”) shows that this matters.

- **BERTScore details.** The threshold $\tau = 0.5$ is fixed without tuning, and the paper does not say whether baseline rescaling was used, which layer, or whether P, R, or F1 was thresholded. A threshold chosen on held-out data would be a fairer baseline.
- **EM normalization.** C.9 mentions “character normalization” as a source of false positives, but the normalization is not specified.
- **Data and code.** I could not find a link to code, the 7,500 candidate outputs, the judge verdicts, or the human labels. Releasing the per-instance verdicts would let readers resolve 2.1 and 2.4 themselves.
- **Error-analysis protocol.** §4.2 samples 100 errors per evaluator, but no category counts, coders, or agreement for the categorization are reported, so the qualitative claims cannot be checked.

One small reporting issue: several percentages in Table 6 are not attainable at $n = 300$ under ordinary rounding. Column EM, FreshQA rows for candidates Llama, GPT, Mixtral, and Mistral give 25.6, 24.6, 18.6, and 15.6; row DeepSeek / HotpotQA gives 38.6. With $n = 300$, the attainable values nearest these are $77/300 = 25.67$, $74/300 = 24.67$, $56/300 = 18.67$, $47/300 = 15.67$, and $116/300 = 38.67$, which would appear as 25.7, 24.7, 18.7, 15.7, and 38.7 if rounded. The reported values match truncation, while other cells are rounded (Table 6, candidate DeepSeek / AmbigQA, Llama judge, $73.7 = 221/300 = 73.67$). Either the sample size differs for those cells, or two rounding conventions were used. By contrast, I checked the four identical kappas of 0.893 in Table 1 (candidate Mistral, AmbigQA) against the differing raw accuracies of 47.7 and 46.3 in Table 6, and they are arithmetically consistent with different confusion matrices, so that apparent coincidence is not a problem.

4. What specifically to fix

1. **§2.3 and Appendix C.5.** State explicitly that, for deterministic judges, DAFE returns exactly the majority vote of the same three models, and present the method as an exact, cheaper computation of that vote. Remove the C.5 sentence attributing differences to cases where “both the primary judges and the arbitrator make errors”.
2. **Table 8.** Find out why the columns differ (rows Mixtral / TriviaQA, Mixtral / NQ-Open, GPT / TriviaQA, Mixtral / AmbigQA, and others). Report the number of instances on which the two pipelines give different final labels, and the reason. If the arbitrator uses a different prompt from GPT-3.5 in the majority vote, say so and show both prompts.
3. **Appendix C.2, bullets under “This leads to”.** Replace “82–94%” with the aggregate 82.4% and the per-cell range 55.7% to 95.3% from Table 7. Delete “Over 90% fewer third-judge inferences”. Report

- total calls (16,318 against 22,500) and the compute cost of the open-weight judges before any claim about overall cost savings. Either add the cost figures to Table 7 or remove “(see Table 7)”.
4. Add confidence intervals (bootstrap over the 300 instances) to Tables 1, 2, 3, and 8, and a paired significance test for DAFE against the best single judge in each cell; drop “significant” wherever the test does not support it.
 5. **Tables 1 and 2.** Add the candidate DeepSeek rows, and add FreshQA with EM and BERTScore so all 25 cells use the same evaluators.
 6. Compare DAFE with at least two other three-judge combinations and with DeepSeek-V3 as a single judge on all cells, at matched cost. Right now, the only strong-single-judge test is GPT-4o on two cells (C.4).
 7. **§4.4 and C.1.** Define “before arbitration”. When the primary judges disagree, there is no single label, so the 72.9% Macro-F1 and 0.467 kappa must have been computed under some rule, and the reader needs to know which.
 8. **§4.2, FreshQA paragraph.** Rewrite the claim that judges “consistently” mark incorrect answers True; Table 6 shows the Llama and DeepSeek judges below HM on FreshQA for candidates GPT, Mistral, and Mixtral.
 9. **§4.1, second paragraph after Table 1.** Correct “(0.97-0.98) across AmbigQA, HotpotQA, and NQ-Open” to match Table 2, and soften the scaling-law sentence.
 10. **Appendix C.4.** The DAFE HotpotQA value for candidate GPT-3.5 is 0.993 in Table 2, not 0.984.
 11. **Appendix C.3.** Put all three results on one scale, give the matching GPT-3.5-arbitrator numbers next to them, and say which DAFE value each should be compared with.
 12. For C.6.1 to C.6.4, give n, dataset, candidate, and judge for every experiment, and explain how 100% verdict stability (C.6.1) coexists with 13% volatility in verdict-only mode (C.6.3) at temperature 0.
 13. A short per-dataset analysis of error overlap among Llama, Mistral, and GPT-3.5 (how often exactly one, two, or three are wrong), which is what explains or fails to explain the gains in Table 3.
 14. **Appendix C vs §3.4.** Make the EM description consistent (containment, not exact match).
 15. **Table 6.** Check the EM cells 25.6, 24.6, 18.6, 15.6, and 38.6 for the sample size or rounding issue described in Section 3.
 16. Give snapshot identifiers and run dates for GPT-3.5-turbo, GPT-4o, and DeepSeek; the serving stack and precision for the open-weight models; max tokens; the FreshQA version; and BERTScore settings.

17. Put the full prompts (candidate, judge, arbitrator, reference-free) in text, not images, and release code, outputs, verdicts, and human labels.
18. **§6 and the Abstract.** Narrow “open-ended language generation tasks” to reference-based short-answer QA in English with binary labels.
19. Typos: “reference-free mood” in the Table 9 caption; “Mistra 7B” in the Figure 6 caption; “Instructv0.1” in footnote 3.

5. Recommendation

Major revisions. The human-labeled dataset is valuable, and the practical idea of calling a third judge only on disagreement is sound. But the paper does not recognize that this procedure is identical to the three-judge majority vote, Table 8 reports different numbers for two procedures that should agree exactly, the efficiency claims in C.2 do not match Table 7, and none of the agreement scores carries an interval. I would be glad to see a revised version that reframes DAFE as an exact, cheaper majority vote, explains the Table 8 discrepancy, and adds the missing cells, intervals, and release materials.

Declarations

Potential competing interests: No potential competing interests to declare.