

Peer Review

Review of: "MTRAG: A Multi-Turn Conversational Benchmark for Evaluating Retrieval-Augmented Generation Systems"

Evgenii Arsentev¹

1. Independent researcher

1. What the work claims

The authors present **mtRAG**, a human-generated benchmark for evaluating retrieval-augmented generation over multi-turn conversations, together with a synthetic companion set, **mtRAG-S**. The benchmark consists of 110 conversations (29 CLAPnq, 27 FiQA, 28 Govt, 26 Cloud; Section 4.4) averaging 7.7 turns, yielding 842 evaluation tasks over four document corpora (Table 2). Its distinguishing feature is the construction procedure (Section 4.3): annotators conversed with a live RAG agent (ELSERv1 retriever plus Mixtral 8x7B Instruct generator), adjusted the retrieved passage set at every turn, and repaired the generated response, so that retrieval is *active* throughout the conversation rather than fixed at the start (Table 1). The authors then use mtRAG to evaluate three retrievers (BM25, BGE-base 1.5, Elser) under last-turn and query-rewrite strategies (Table 3), and nine LLM generators under three retrieval settings - Reference, Reference+RAG, full RAG (Table 5) - using two reference-based metrics (RB_{alg} , RB_{llm}) and one reference-less faithfulness judge (RL_F), all gated by an "IDK" abstention judge (Section 6.3.3). A human evaluation on 159 tasks over two frontier systems (Section 7, Table 6) is used both to show that the human reference answers are preferred and to select which automatic metrics to report (Section 8, Table 7, Figure 4). The headline claims are that state-of-the-art RAG systems struggle on mtRAG, that they struggle specifically on later turns, unanswerable questions, and non-standalone questions, and that current automatic metrics correlate only weakly with human judgement.

2. Whether the conclusion holds on the data

The central artefact is real, and the qualitative story is plausible, but several headline claims are stated with more confidence than the reported numbers support.

Scale and clustering. The benchmark contains 842 tasks, but they are drawn from only 110 conversations, each contributed by a single annotator over a single topic and a shared passage set. Tasks within a conversation are not independent observations, yet every subgroup number in Tables 3-6 and 15-16 is reported as if n were the task count. The effective sample size for most comparisons is closer to the number of conversations (110 overall, 26-29 per domain, 20 in the human evaluation) than to the task count. No confidence intervals, standard errors, clustered errors, or significance tests appear anywhere in the paper, yet Table 5 bolds the best and underlines the second-best value per column, which invites exactly the rankings the data cannot support. In Table 5, Reference setting, RB_{alg} , GPT-4o scores 0.46, and Llama 3.1 405B Instruct scores 0.47; the paper's own text (Section 6.4) reads this as "GPT-4o and Llama 3.1 405B Instruct perform the best," which is fine, but nothing in the paper establishes that a one-point gap on 426 clustered tasks is distinguishable from noise.

The most direct evidence of run-to-run variance is in the paper itself. In Table 5, the "Reference" row is a fixed human-written text that does not change between retrieval settings. Its answerability accuracy is nevertheless reported as 0.98 (Reference), 0.97 (Reference+RAG), and 0.98 (full RAG). If the IDK judge sees only the response, this quantity is constant by construction, and the 0.01 wobble is judge non-determinism; if the judge also sees the passages, the paper does not say so. Either way, the wobble is of the same magnitude as many of the inter-model differences in the same column (for example, GPT-4o 0.89 against Qwen 2.5 7B 0.88 in the Reference setting).

Missing controls. Two controls that would isolate the effects the benchmark is built to measure are absent. (a) There is no closed-book condition. CLAPnq is derived from Natural Questions, FiQA from a 2018 WWW challenge, and the Govt corpus from public .gov and .mil pages – all web-public and plausibly inside the pretraining data of every model in Section 6.2. Without a no-passage baseline, it is impossible to say how much of the Reference-setting score in Table 5 comes from the supplied passages rather than from parametric memory, and therefore how much of the benchmark actually tests retrieval-augmented generation. (b) There is no last-question-only generation condition. The retrieval experiments do isolate the effect of context (last turn against query rewrite, Table 3), but the generation experiments always supply the full preceding conversation. The claim in Section 6.4 that the first-turn/late-turn gap in

Figure 3(b) "proves our conjecture that answering questions in a multi-turn setting is more challenging" therefore rests on a comparison between two different question populations, not on an ablation of the conversational context.

Confounding of the turn effect with the seeding procedure. Section 4.2 states that first turns were bootstrapped from seed questions – existing QA-dataset questions for CLAPnq and FiQA, and annotator-written questions over a selected seed document for Govt and Cloud. First-turn questions are thus, by construction, anchored to a known document. Table 4 reports Elser R@5 of 0.89 for Turn 1 (n=102) against 0.47 for later turns (n=675). A large part of that 42-point gap is attributable to how the first turn was created, not to multi-turn difficulty. The paper should say so, or provide a first-turn subset not derived from seeds.

The non-standalone claim is not supported at the reported sample size. Section 5.3 concludes that "non-standalone questions continue to pose a challenge". Table 4 gives Standalone R@5 = 0.48 (n=555) against Non-Standalone R@5 = 0.42 (n=120). A crude binomial standard error at $p=0.42$, $n=120$ is about 0.045, so the non-standalone estimate is roughly 0.42 plus or minus 0.09 before any correction for within-conversation clustering. The 6-point difference is not separable from noise. The same applies to Section 3.2, which reports on average 1.3 co-referential questions per conversation, i.e., about 143 tasks in the whole benchmark: one of the benchmark's four advertised challenge dimensions is carried by roughly 14 percent of post-first turns.

Unanswerability is carried by 55 tasks, and for those tasks the three metrics collapse into one binary. Table 13 is computed "on 55 unanswerable questions", which is 6.5 percent of the 842 tasks. More importantly, the conditioning rule in Section 6.3.3 assigns, for A = no, a score of 1 when IDK = yes and 0 otherwise. All three metrics therefore reduce to a single abstention indicator on unanswerable tasks. Table 13 confirms this exactly: after IDK conditioning, GPT-4o scores 0.20 on RL_F, 0.20 on RB_{llm}, and 0.20 on RB_{alg}; Llama 3.1 405B scores 0.20/0.20/0.20; the Reference scores 0.87/0.87/0.87. The unanswerable column of Table 16a repeats the same single number per row for the same reason. This is a defensible design choice, but it means the dramatic drop on unanswerables in Figure 3(a) is not a measure of response quality at all – it is abstention frequency, reported three times. The text never states this.

The same table bounds the IDK judge far more tightly than the claimed accuracy. Section 6.3.3 and Appendix F.3 report IDK judge accuracy "over 97%". But in Table 13, the Reference row – the gold human answer to an unanswerable question – scores 0.87 after conditioning, meaning that on roughly 13 percent of the 55 unanswerable tasks, the gold answer is not recognised as an abstention (or the gold answer does

not abstain). On the one class where the IDK judge does all the work, the implied error rate is four times the headline figure. Since every unanswerable score in Table 5, Figure 3(a), Table 13, and Table 16a is the output of this judge, a 13-point ceiling on the gold answer must be stated alongside the model numbers - GPT-4o's 0.20 is measured against a 0.87 ceiling, not 1.00.

The headline generation table is computed on a non-random half of the benchmark. Table 5 is reported on $n=426$, and footnote 6 explains that this is the Reference+RAG task set, which Section 6.1 defines as "the 426 tasks that have at most two reference passages". This is 50.6 percent of the 842 tasks, selected on a property (number of supporting passages) that plausibly correlates with task difficulty. The effect is visible in the paper: Command-R+ scores $RL_F = 0.69$ in the Reference setting of Table 5 but 0.76 "Overall" in Table 16a; GPT-4o scores 0.69 against 0.75. Footnote 6 says "the trends persist throughout the benchmark", and the ordering largely does, but the levels move by up to 7 points, and it is the Table 5 levels that the abstract and Section 10 generalise from.

One reference-less comparison in Table 5 is an artefact rather than a result. Section 6.4 reads the ordering across settings as showing "the noise has an impact on generation". For RB_{llm} and RB_{alg} , this is reasonable. For RL_F , it is not: the Reference row falls from 0.86 (Reference) and 0.87 (Reference+RAG) to 0.67 (full RAG) even though the text being scored is identical in all three columns. What changes is the passage set the faithfulness judge scores it against. In the full-RAG column, the Reference row therefore measures the retriever, and 0.67 is the effective ceiling for that column - against which Command-R+ scores 0.66, i.e., within 0.01 of the human answer. RL_F has essentially no headroom in the full-RAG setting, and this should be stated rather than left for the reader to derive.

The domain conclusion is contradicted by the authors' own appendix. Section 6.4 states that "model performance is similar across domains except for FiQA where the results tend to be lower" and attributes this to the corpus being short, informal, and subjective. Table 16c permits a sharper test because Figure 3 and Table 16 are in the Reference setting, where passages are gold and retrieval cannot confound. On RB_{alg} , FiQA is the lowest of the four domains for all nine models (Command-R+ 0.39 against 0.47/0.43/0.45; GPT-4o 0.38 against 0.48/0.47/0.45; GPT-4o-mini 0.37; Llama 405B 0.41; Llama 70B 0.37; Llama 8B 0.32; Mixtral 0.34; Qwen 72B 0.37; Qwen 7B 0.35). On RL_F , however, FiQA is frequently *not* the lowest: GPT-4o scores 0.74 on FiQA against 0.73 on CLAPnq; Llama 3.1 405B scores 0.76 on FiQA against 0.72 on CLAPnq; Qwen 2.5 72B scores 0.70 on FiQA against 0.68 on CLAPnq; Command-R+ scores 0.76 on FiQA against 0.73 on Cloud. A deficit that appears only in the metric built on Rouge-L and BERTScore against a reference answer, and disappears in the reference-less faithfulness judge, is better explained as

an answer-style and lexical-overlap artefact of RB_{alg} than as lower model performance. The Reference row of Table 16c is flat across domains (RB_{alg} 0.88/0.88/0.88/0.87), so the ceiling is not the explanation.

Two claims of consistency are overstated. Section 5.3 states that "the query rewriting strategy consistently outperforms using only the last turn (without rewriting), across all metrics, for all models". That is true of every cell of Table 3, which I verified. It is not true of Table 15: on the Cloud domain, Elser $R@5$ falls from 0.48 (Last Turn) to 0.47 (Query Rewrite). Similarly, Section 6.4's "in almost all cases the models perform better on first turn vs subsequent turn questions" is accurate, but Section 10's "proves our conjecture" is not, and Table 16b contains three counterexamples on RB_{llm} : Llama 3.1 8B Instruct scores 0.56 on Turn 1 against 0.59 on later turns, Qwen 2.5 7B scores 0.71 against 0.72, and Llama 3.1 405B ties at 0.74/0.74.

The synthetic-data comparison has an unexamined self-generation confound. Section 9 and Appendix G.2 report that models score higher on mtRAG-S than on mtRAG. Table 14(a) shows RL_F gains of +0.07 (Command-R+ 0.76 to 0.83), +0.10 (GPT-4o-mini 0.71 to 0.81), +0.10 (Llama 405B 0.75 to 0.85), +0.10 (Qwen 72B 0.72 to 0.82) – and +0.18 for Mixtral 8x22B Instruct (0.61 to 0.79). Appendix G.1 states that mtRAG-S was generated with Mixtral 8x22B Instruct v0.1. The model that wrote the synthetic conversations gains roughly twice as much as any other model, and the pattern repeats in Table 14(b) with BERT-K-Prec (Mixtral +0.12, all others +0.05 to +0.08). Part of this may be regression to the mean, since Mixtral starts lowest on mtRAG in both panels, but the paper lists "idiosyncrasies of synthetic data generation approaches" as a candidate explanation without noticing that its own table contains a direct test of it. Table 14 also silently drops four of the nine models of Table 5 (GPT-4o, Llama 70B, Llama 8B, Qwen 7B) with no stated reason, and does not say which retrieval setting it uses.

Arithmetic and internal consistency. Most totals reconcile. 842 tasks over 110 conversations is 7.65, consistent with the stated 7.7; the domain counts sum to 110; Table 2's passages-per-document column is correct to the reported precision ($61,022/57,638 = 1.06$; $183,408/4,293 = 42.7$; $49,607/7,661 = 6.5$; $72,442/8,578 = 8.4$); Table 8's synthetic set reconciles ($1,181/200 = 5.9$); and Table 4's subsets are mutually consistent ($102 + 675 = 777$; $555 + 120 = 675$; $208 + 180 + 201 + 188 = 777$), with the weighted averages reproducing the 0.52 and 0.47 figures. Three things do not reconcile, and are listed as items 1, 2 and 3 of Section 4 below.

LLM-as-a-Judge: agreement, family overlap, and self-validation. This is where the paper is weakest, and the authors deserve credit for being more candid than most about it (Table 7 reports that RL_F correlates

with win-rate at 0.01 and RL_{MTB} at -0.17). But three structural problems remain.

First, *the judges are the systems*. Appendix F.1 states that RB_{llm} is the median of four judges: GPT-4o-mini (2024-07-18), Qwen 2.5 72B, Mixtral 8x22B Instruct, and Llama 3.1 405B Instruct. All four are rows of Table 5. Mixtral 8x22B Instruct scores itself. GPT-4o-mini scores itself and its larger sibling GPT-4o. Llama 3.1 405B scores itself and the 70B and 8B models. Qwen 2.5 72B scores itself and the 7B model. Command-R+ is the only evaluated family with no representative on the panel, and it is also the model whose RB_{llm} rank is furthest below its RL_F rank: in Table 16a, Command-R+ leads all systems on RL_F Overall (0.76, above GPT-4o's 0.75 and Llama 405B's 0.75) while trailing GPT-4o on RB_{llm} Overall by seven points (0.69 against 0.76). Conversely, Mixtral, which sits on the panel, has the second-lowest RL_F Overall (0.61) but an RB_{llm} Overall of 0.69, above Llama 70B's 0.66. This is exactly the signature of self-preference that the panel was meant to suppress. Taking a median across four judges reduces variance but does not remove a bias shared by the panel members toward their own family, and with an even panel size, the "median" is not even uniquely defined – the tie-breaking rule is not stated.

Second, *one model occupies three roles*. GPT-4o-mini (2024-07-18) is a system under test (Table 5), a member of the RB_{llm} judge panel (Appendix F.1), and the sole IDK judge (Appendix F.3). Since the IDK judge gates every score in Tables 5, 13, and 16, GPT-4o-mini's abstention decisions determine whether each of the nine systems, including GPT-4o-mini itself, receives its metric score or a hard zero.

Third, *the benchmark validates the judge that scores the benchmark, with no held-out split*. The IDK judge's "over 97%" accuracy is measured on the human evaluation results (Appendix F.3), the metric-to-human correlations of Table 7 and Figure 4 are computed on the same human evaluation, and Section 8 states that "these findings drive our decision to report RB_{alg} , RB_{llm} , and RL_F as our main metrics in Section 6". Metric selection, judge validation, and judge deployment all happen on the same 159 tasks drawn from 20 conversations and two systems. Nothing is held out. And the two systems used for validation (GPT-4o and Llama 3.1 405B) are near-indistinguishable from one another in the human data (win-rate 47.8 against 47.4 in Table 6), so the correlations in Table 7 are estimated on a pair of systems with almost no true quality separation and are then used to license a nine-way ranking spanning 8B to 405B parameters. A Spearman correlation of 0.33 (RB_{llm} , Table 7) obtained under those conditions, with no confidence interval, no p-value, and no stated n, is a thin basis for the ranking in Table 5.

The human-agreement figures should also be read with care. Appendix E defines agreement as the percentage of tasks where the three annotators either agreed exactly or where two agreed and the third

was within one point. On a four-point scale whose observed values are compressed into 3.5–4.0 (Table 6), a tolerance of plus or minus one point covers most of the usable range, so the reported {F: 89.6, A: 92.1, N: 95.6, C: 90.3, WR: 87.4} are not chance-corrected and are not comparable to the kappa or alpha values readers will assume. The compression is severe enough to matter on its own: in Table 6, Naturalness is 4.0 for both the Reference and GPT-4o, and Llama 3.1 scores 3.9 on Appropriateness against the Reference's 3.8, so two of the four FANC dimensions have no discriminative power at all.

3. Reproducibility

The artefact release is a genuine strength: the benchmark is available at the stated GitHub repository (footnote 8), the corpora are described with document and passage counts (Table 2), the chunking is specified (512-token passages, stride 100, Section 4.2), the judge prompts are given (Figures 5, 6, 7, 8), and one judge is pinned to a dated snapshot (GPT-4o-mini 2024-07-18, Appendices F.1 and F.3). That is more than most papers of this kind provide. What is missing, however, prevents an independent party from reproducing any number in Tables 3, 5, 6, 7, 13, 14, 15, or 16.

Model versions. Only GPT-4o-mini, and only in its judge role, carries a snapshot date. The GPT-4o and GPT-4o-mini *systems under test* in Table 5 have no snapshot identifier; footnote 4 links the system card, which is not a version. Command R+ is identified only by a HuggingFace collection URL (footnote 5) with no revision hash. Llama 3.1 8B/70B/405B, Qwen 2.5 7B/72B, BGE-base 1.5, and Mixtral 8x22B Instruct carry no revision hashes either; Section 6.2 says "Mixtral 8x22B" while Appendix G.1 says "Mixtral 8x22B Instruct v0.1", so even within the paper the versioning is inconsistent. Elser is pinned in Section 4.3 (ELSERv1, ElasticSearch 8.10) for data creation, but Section 5.1 says only "Elser" for the retrieval experiments, and the reader cannot tell whether the same version was used.

Decoding parameters and seeds. Temperature, top-p, top-k, repetition penalty, maximum generation length, stop sequences, and random seeds appear nowhere in the paper. Appendix D.2 describes how the prompt is composed via HuggingFace ChatTemplate and states that no truncation was needed, but says nothing about sampling. This is true both for the nine systems under test and for the five judge calls (four RB_{llm} judges, the IDK judge, RL_F , RL_{MTB} , and the Mixtral 8x7B query rewriter of Appendix C.1). The number of samples per task is not stated; the reader must assume a single generation per task, with no repeat runs and therefore no estimate of generation variance. As noted in Section 2 above, the 0.98/0.97/0.98 wobble on the fixed Reference row of Table 5 is evidence that this variance is non-zero and of the same order as several reported differences.

Judge models that are never named. RL_F is described (Section 6.3.2) as "the Faithfulness LLM judge from RAGAS", and RL_R as RAGAS Answer Relevance, but the underlying LLM, its version, and its decoding settings are never given. The same holds for RL_{MTB} : Appendix F.2 gives the adaptation and the prompt (Figure 7) but not the judge model. RL_F is one of the paper's three headline metrics, and it is also the metric used to carry the entire mtRAG versus mtRAG-S comparison in Table 14(a). The RAGAS library version is also unspecified, which matters because its faithfulness implementation has changed materially across releases.

Serving stack and numerics. The paper does not state how the open-weight models were served (vLLM, TGI, HuggingFace transformers, a hosted API), nor the precision or quantisation used. For Llama 3.1 405B and Mixtral 8x22B, this is not a detail: bf16 against an 8-bit or 4-bit quantised serving path can move generation quality by more than the inter-model gaps reported in Table 5.

Retrieval configuration. BM25 is reported without its k_1 and b parameters, without the analyser, tokenizer, or stopword list, and without stating the implementation. BGE-base 1.5 is reported without query and passage instruction prefixes, which the model requires and which materially change its scores. The Elasticsearch index mapping and the Elser inference configuration are not given. Since Table 3's BM25 row ($R@5 = 0.20$ last turn) is used to argue that mtRAG is hard for lexical retrieval, these settings should be pinned.

Corpus snapshots. The Govt corpus is a crawl of .gov and .mil pages, and the Cloud corpus is a crawl of "the public technical documentation of a major cloud provider" (Appendix B.1). No crawl date is given for either, no URL or seed list is published, the cloud provider is not named, and no licence or redistribution terms are stated. Government pages and cloud documentation both change continuously; without a snapshot date, the Govt and Cloud results in Tables 4, 15, and 16c are not reproducible even in principle, and the anonymous provider makes the Cloud corpus impossible to cite or to extend. CLAPnq and FiQA are properly cited, but their versions or commit hashes are not.

Data-creation leakage. Footnote 3 explains, correctly and commendably, that Mixtral 8x7B was excluded from the systems under test because it generated the responses that annotators repaired. But Appendix C.1 then uses Mixtral 8x7B Instruct as the query rewriter for the retrieval experiments and, by extension, for the Reference+RAG and full-RAG settings of Table 5. The same model that shaped the reference answers therefore sits inside the evaluation pipeline of every system. The effect is common to all systems and so does not distort the ranking, but it should be acknowledged alongside the Elser bias already acknowledged in Section 5.1.

Annotation process. The number of annotators who created the 126 conversations is never given, nor the number per domain, nor whether the same people performed creation and review, nor the number who participated in the Appen human evaluation beyond "three skilled annotators" per task. Annotator identity is a random effect at the conversation level, and with only 110 conversations, a small pool would substantially limit the generality of the benchmark. The annotation guidelines are referred to in Appendix B.2, but the text of the instructions does not appear in the article body as rendered.

Aggregation rules. Two aggregations cannot be reproduced from what is printed. The "All" column of Table 6 is described as the harmonic mean of FANC, but the harmonic mean of the Reference row's printed 3.8/3.8/4.0/3.9 is 3.87, not the reported 3.8; for GPT-4o, it is 3.74 against a reported 3.6; for Llama 3.1, it is 3.74 against a reported 3.6. The gap of roughly 0.14 for the two models is larger than rounding and is consistent with the harmonic mean being computed per task and then averaged, but the paper does not say which order is used. Similarly, Table 7 reports a "weighted" Spearman correlation "to correct for data imbalance" without giving the weights, the weighting variable, the unit of observation, or n.

4. What specifically to fix

1. **Reconcile the retrieval subset sizes between Table 4 and Table 15.** Table 4 reports Govt (201) and Cloud (188), summing with CLAPnq (208) and FiQA (180) to 777, which matches Table 4's own turn split ($102 + 675 = 777$). Table 15 reports Govt (203), Cloud (189) and All (780). The two tables report identical metric values for the same runs but different denominators. 777 appears to be the correct figure, since 842 tasks minus the 10 conversational tasks excluded in Section 6.3.3 minus the 55 unanswerable tasks of Table 13 gives exactly 777, and retrieval is computed only on answerable and partially answerable tasks (Section 5.1). Please correct Table 15 or explain the three-task difference.
2. **Correct the repair rate in Section 3.2.** Section 3.2 states "Our conversations contain repairs on 92% of the turns". Appendix B.3 states that the number of repaired responses rose "from 778 repaired responses before review to 799 responses after review". $778/842 = 92.4$ percent and $799/842 = 94.9$ percent, so the 92 percent quoted as a property of the released benchmark is the pre-review figure. Table 8's "Avg # edited responses per conv." of 7.3 corresponds to $799/110 = 7.26$, confirming that 94.9 percent is the correct figure for the benchmark as released.
3. **Define the edit score and reconcile Section 4.3 with Appendix B.3.** Section 4.3 gives an edit score (Rouge-L similarity between the original and the repaired response) of 60.7 for the final benchmark, described as "indicating significant amount of repair". Appendix B.3 states that for the 264

responses repaired during review the average edit score "increased from 59.5 to 69.2". Under the stated definition a higher score means greater similarity to the original model output, i.e. *less* repair, so an increase during a repair pass is counterintuitive. State explicitly whether higher means more or less repair, and whether 60.7 and 69.2 are computed over the same population.

4. **State the aggregation order for the "All" column of Table 6** and reconcile it with the printed FANC values. As computed above, the harmonic mean of the printed per-property medians gives 3.87 / 3.74 / 3.74 against the reported 3.8 / 3.6 / 3.6.
5. **Report the composition of the n=426 subset used in Table 5.** Section 6.1 defines it by "at most two reference passages"; state how many of the 426 are answerable, partially answerable and unanswerable, since answerability accuracy and every unanswerable conclusion depend on that breakdown, and since unanswerable tasks have zero reference passages and so automatically satisfy the selection rule.
6. **Add a table with the exact task counts behind Figure 2.** Figure 2 is the only presentation of the benchmark's distribution over question type, multi-turn type, answerability and domain, and it is a figure. Every subgroup claim in Sections 5.3, 6.4 and 7 and in Figures 3, 9 and 10 depends on these counts; they should be readable as numbers.
7. **State explicitly in Section 6.3.3 and in the caption of Figure 3(a) that for A = no the three metrics collapse to a single binary abstention indicator.** Table 13's post-conditioning columns (GPT-4o 0.20/0.20/0.20, Llama 3.1 0.20/0.20/0.20, Reference 0.87/0.87/0.87) already show this; presenting the same quantity as three metrics in Figure 3(a) and in the "Unans." column of Table 16a overstates the evidence. Consider reporting it plainly as an abstention rate.
8. **Report the IDK judge's accuracy separately by answerability class, not only in aggregate.** The "over 97%" of Appendix F.3 is dominated by answerable tasks. The Reference row of Table 13 implies roughly 13 percent error on the 55 unanswerable tasks – the one class where the judge carries the whole measurement. Give the confusion matrix of the judge against gold answerability and state the 0.87 ceiling wherever unanswerable model scores are reported.
9. **Replace answerability accuracy in Table 5 with a two-way breakdown.** Section 6.4 argues that Llama 3.1 70B and 8B "say 'I don't know' too often. This can also be seen by their low answerability accuracy in Table 5". A single accuracy figure conflates over-abstention with hallucination and cannot show that. Report the false-abstention rate on answerable tasks and the failure-to-abstain rate on unanswerable tasks separately.

10. **Remove the four judge models from the pool of systems scored by RB_{llm} , or quantify the self-preference bias.** Appendix F.1's panel (GPT-4o-mini 2024-07-18, Qwen 2.5 72B, Mixtral 8x22B Instruct, Llama 3.1 405B Instruct) consists entirely of Table 5 rows, and Command-R+ – the only family with no panel seat – shows the largest RL_F -versus- RB_{llm} rank discrepancy in Table 16a (RL_F Overall 0.76, highest of all systems; RB_{llm} Overall 0.69, seven points below GPT-4o's 0.76). At minimum, report each judge's scores separately for each system so readers can see the own-family deltas, and add a leave-one-family-out RB_{llm} variant.
11. **Disclose the median rule for an even-sized judge panel.** Section 6.3.1 and Appendix F.1 specify a median over four judges; state whether this is the mean of the two central values, the lower, or something else, and report the inter-judge spread (for example the interquartile range per system in Table 5).
12. **Break the circularity between judge validation and judge deployment.** The IDK judge is validated on the human evaluation (Appendix F.3), the metrics are selected on the same human evaluation (Section 8, "These findings drive our decision to report RB_{alg} , RB_{llm} and RL_F "), and those metrics then score the full benchmark. Split the 159 human-evaluated tasks into a metric-selection set and a held-out validation set, and report Table 7's correlations on the held-out portion.
13. **Report n, confidence intervals and p-values for every correlation in Table 7 and Figure 4, and state the weighting scheme behind the "weighted Spearman".** As reported, the best correlation with win-rate is 0.33 (RB_{llm}) and RL_F is 0.01, yet RL_F is retained as a headline metric on the strength of Figure 4 alone. If the justification for RL_F is its correlation with the human F score specifically rather than with win-rate, say so in Section 6.3.2 and give that number in the text.
14. **Extend the human evaluation beyond two near-identical systems, or restrict the claims it licenses.** The validation set comprises GPT-4o and Llama 3.1 405B, which in Table 6 differ by 0.4 points of win-rate (47.8 against 47.4). Correlations estimated on a pair with almost no separation cannot justify the nine-way ranking of Table 5. Adding at least one clearly weaker system (Llama 3.1 8B, which scores lowest on almost every column of Table 16a) would give the correlation estimates real range.
15. **Report chance-corrected inter-annotator agreement.** Replace or supplement the Appendix E figures ({F: 89.6, A: 92.1, N: 95.6, C: 90.3, WR: 87.4}) with Krippendorff's alpha or Fleiss kappa. With a four-point scale whose observed values occupy 3.5–4.0 (Table 6) and a plus-or-minus-one-point

tolerance, the reported percentages are close to what chance would produce, and Naturalness in particular is at ceiling (4.0 for both Reference and GPT-4o).

16. **Add clustered or bootstrap confidence intervals at the conversation level to Tables 3, 4, 5, 6, 15 and 16**, and remove the bolding and underlining in Table 5 unless the differences survive a test. Tasks within a conversation share a topic, a passage set and an annotator, so the nominal task counts overstate the effective sample size throughout.
17. **Qualify the turn and standalone claims in Section 5.3 in light of the construction procedure.** Table 4's Turn 1 R@5 of 0.89 (n=102) is inflated by the seeding process of Section 4.2, under which first-turn questions were derived from existing QA datasets or written against a chosen seed document. Likewise, state that the Standalone 0.48 (n=555) against Non-Standalone 0.42 (n=120) comparison is within sampling error and does not on its own support "non-standalone questions continue to pose a challenge".
18. **Soften "consistently" in Section 5.3 or restrict it to Table 3.** Table 15 contains a counterexample: Elser on Cloud (n as printed, 189) scores R@5 = 0.48 with Last Turn against 0.47 with Query Rewrite. Similarly, replace "proves our conjecture" in Section 6.4 with the weaker claim the data support, and note the three RB_{llm} counterexamples in Table 16b (Llama 3.1 8B 0.56 to 0.59, Qwen 2.5 7B 0.71 to 0.72, Llama 3.1 405B 0.74 to 0.74).
19. **Revise the domain explanation in Section 6.4.** The FiQA deficit is metric-specific: it appears in RB_{alg} for all nine models of Table 16c but not in RL_F, where GPT-4o (0.74 FiQA against 0.73 CLAPnq), Llama 3.1 405B (0.76 against 0.72), Qwen 2.5 72B (0.70 against 0.68) and Command-R+ (0.76 FiQA against 0.73 Cloud) do not show FiQA as lowest. Either present the FiQA result as a property of the reference-based metric under informal short reference answers, or explain why RL_F fails to detect a genuine quality drop.
20. **Explain the Reference row of the full-RAG column of Table 5 and state the resulting ceiling.** The human reference answer is a fixed text, yet its RL_F falls from 0.86 (Reference) and 0.87 (Reference+RAG) to 0.67 (full RAG). That column's Reference value measures the retriever, and Command-R+ at 0.66 is effectively at the gold ceiling, so the RL_F full-RAG column cannot discriminate. Section 6.4's conclusion about degradation across settings should be limited to RB_{llm} and RB_{alg}.
21. **Report the three components of RB_{alg} per model.** Section 6.3.1 defines RB_{alg} as the harmonic mean of Bert-Recall, Bert-K-Precision and Rouge-L; a harmonic mean is dominated by its smallest term,

and Table 14(b) shows Bert-K-Prec values of 0.27-0.33 on mtRAG, far below the composite. RB_{alg} may therefore be largely a rescaling of Bert-K-Prec, which Section 8 also evaluates as an independent reference-less metric – so Table 7 and Figure 4 are not comparing independent quantities. Publishing the components would settle this. It would also make clear that RB_{alg} is not purely reference-based, since Bert-K-Prec compares the response to the passages.

22. **Add a closed-book control condition.** Run each of the nine models of Section 6.2 with the conversation but no passages, and report it as a fourth column in Table 5. Three of the four corpora are public and pre-date the models' training cutoffs, and without this baseline the Reference-setting numbers cannot be separated from parametric recall. This is the single most valuable addition the paper could make.
23. **Add a no-history generation ablation.** Supply only the final user question plus the same passages, and compare against the full-conversation condition of Table 5. Without it, the first-turn versus later-turn gap in Figure 3(b) and Table 16b cannot be attributed to multi-turn context, which is the paper's central motivation.
24. **Test the self-generation confound in Section 9 and Table 14.** Mixtral 8x22B Instruct generated mtRAG-S (Appendix G.1) and shows the largest mtRAG-to-mtRAG-S gain in both panels (+0.18 on RL_F , +0.12 on BERT-K-Prec, against +0.05 to +0.10 for every other model). Regenerate a portion of mtRAG-S with a different generator, or at minimum note the confound and the competing regression-to-the-mean explanation (Mixtral has the lowest mtRAG baseline in both panels). Also state which retrieval setting Table 14 uses, and why Table 14 covers five of the nine models of Table 5.
25. **Pin every model, judge and retriever used.** Specifically: dated snapshots for the GPT-4o and GPT-4o-mini systems under test in Table 5 (currently only footnote 4's system-card link); HuggingFace revision hashes for Command R+ (footnote 5), Llama 3.1 8B/70B/405B, Qwen 2.5 7B/72B, Mixtral 8x22B Instruct and BGE-base 1.5; the Elser version used in Section 5.1, which is stated only for data creation in Section 4.3; and the model, version and settings for the RL_F and RL_R RAGAS judges (Section 6.3.2) and the RL_{MTB} judge (Appendix F.2), none of which are named anywhere.
26. **Publish the run configuration.** Add a table giving, for every generation and every judge call: temperature, top-p, maximum new tokens, stop sequences, random seed, number of samples per task, serving stack and numerical precision or quantisation. Appendix D.2 currently covers only

prompt composition. Include the same for the Mixtral 8x7B query rewriter of Appendix C.1, since it feeds the retrieval results of Table 3 and both RAG settings of Table 5.

27. **Publish the retrieval configuration.** Give BM25's k_1 and b , the analyser and stopword list, and the implementation; give the query and passage instruction prefixes used for BGE-base 1.5; and give the Elasticsearch index mapping and Elser inference settings (Section 5.1, Table 3).
28. **Give crawl dates, seed URLs and licensing for the Govt and Cloud corpora (Section 4.2, Appendix B.1), and name the cloud provider.** Without a snapshot date the Govt and Cloud rows of Tables 4, 15 and 16c are not reproducible; without a named provider and a licence the Cloud corpus cannot be cited, extended or lawfully redistributed. State the release status and licence of mtRAG-S as well, which Section 9 introduces but footnote 8 does not cover.
29. **Report the annotation workforce.** Section 4.1 and Section 7 should state how many annotators created the 126 conversations, how many per domain, whether creators and reviewers were disjoint, and how many annotators took part in the Appen evaluation. With 110 conversations, annotator identity is a first-order source of variation in the benchmark.
30. **Add size columns to Table 1.** The comparison to QuAC, Or-QuAC, CoQA, ShARC, MD2Dial, FaithDial, iKAT and RAD-Bench is a matrix of ticks with no criteria, no citations for the individual judgements and no scale information. Adding conversations, turns and domains per benchmark would let readers place mtRAG's 110 conversations and 842 tasks in context, and would support or qualify the priority claim in Section 1 and Section 10 that mtRAG is "the first end-to-end human-generated multi-turn RAG benchmark" – a claim that currently rests on an unstated definition of "end-to-end" that happens to exclude iKAT.
31. **Reconcile the passage statistics.** Section 3.2 reports 20.9 relevant passages in total per conversation, which over 110 conversations and the 777 tasks that carry reference passages implies about 2.96 per task; Table 8 reports 2.8. The difference may be definitional, but the two quantities should be defined precisely enough to reconcile.
32. **Fix the caption of Table 6.** Section 7 states that the results are on "the answerable/partial subset (152 tasks)" while the caption states "All results are reported on the Answerable subset". Also give n for the unanswerable analysis in the paragraph following Table 6: 159 total evaluation tasks minus 152 leaves at most 7 tasks, on which the observation that "Llama 3.1 405B Instruct is the least preferred by win-rate" cannot be more than anecdotal.
33. **Define win-rate precisely in Section 7.** With three compared systems, pairwise comparisons and ties split evenly, the mean win-rate across systems should be 50.0; the values in the WR column of

Table 6 (Reference 59.3, GPT-4o 47.8, Llama 3.1 47.4) sum to 154.5 and average 51.5. State whether ties are dropped, split, or counted, and whether win-rate is computed per pair or pooled.

34. Acknowledge the Mixtral 8x7B query-rewriting path as a second creation-time dependency.

Section 5.1 properly flags the Elser bias; Appendix C.1 uses the same generator model that produced the responses annotators repaired (Section 4.3) as the query rewriter for the retrieval results of Table 3 and the Reference+RAG and full-RAG settings of Table 5. Footnote 3 shows the authors are alert to this class of issue for generation; the same reasoning applies here.

5. Recommendation

Major revisions. mtRAG is a valuable and carefully built resource, and its construction procedure – annotators driving a live RAG agent and repairing both the passage set and the response at every turn – is a genuine methodological advance over benchmarks that fix retrieval at the start of a conversation. The data, the corpora, and the released artefact will be useful to the community regardless of how this review is resolved, and the authors are unusually forthright about the weakness of current automatic metrics (Table 7) and about the Elser bias (Section 5.1).

The revisions needed are nevertheless substantial and go beyond presentation. Three internal inconsistencies need correcting (items 1-3); the evaluation apparatus is circular in a way that affects every headline number, since the four RB_{llm} judges are themselves evaluated systems, GPT-4o-mini simultaneously plays three roles, and the judges are validated and selected on the same 159 tasks they then score (items 10-14); the paper reports no uncertainty anywhere despite drawing subgroup conclusions from as few as 55 unanswerable tasks, 120 non-standalone tasks, and 7 human-evaluated unanswerable tasks (items 16-18, 31); two controls that would isolate the effects the benchmark exists to measure are missing (items 22-23); and the run configuration is under-specified to the point that no table in the paper can be reproduced (items 25-27). None of these require new data collection – items 22, 23, and 24 need additional runs over the existing benchmark, and the remainder are matters of disclosure, recomputation, and more careful claims. With those addressed, this would be a strong and durable contribution.

Declarations

Potential competing interests: No potential competing interests to declare.