Qeios

Peer Review

Review of: "LFOSum: Summarizing Long-form Opinions with Large Language Models"

Evgenii Arsentev[1]

1. Independent researcher

What the paper claims

Three contributions are offered. A dataset, LFOSum, pairing 500 hotels with roughly 1.5K TripAdvisor reviews each, about 207K GPT-4o tokens per entity, against structured PROS and CONS summaries taken from Oyster, averaging 12.29 sentences split 8.45 positive and 3.84 negative (Table 5). Two training-free summarisation routes: a long-context critic that reads everything at once with sentiment and length control, and a RAG pipeline that mines frequent explicit aspects as query terms, retrieves sentences per term with BM25 or a dense retriever, and has the model rerank or rewrite them (Sections 3.1 and 3.2). And a set of reference-free metrics built on Aspect-Opinion-Sentiment triplets that score a summary against the retrieved evidence rather than against the gold text (Section 3.2.4).

The dataset is the part that will outlive the paper: book-length inputs with structured expert references and explicit control dimensions do not otherwise exist in this area, as Table 1 shows honestly. The Sketch-Fetch-Fill recovery for malformed JSON in Section 3.1 is a small practical contribution that many people will quietly reuse. My main concern is with what the gold summaries are.

Does the conclusion hold on the data?

1. The reference is not a summary of the input. Section 2 states plainly that Oyster summaries come from first-hand visits by professional reviewers, while the inputs are TripAdvisor user reviews for the same hotel. These are two independent observations of one object, not a document and its summary. A system is therefore rewarded for reproducing judgements that may appear nowhere in its input, and penalised for faithfully summarising opinions the Oyster reviewer never raised. That does not invalidate the dataset, but it changes what every score in Tables 2 and 3 means. The paper already has the

 doi.org/10.32388/0WXOBG

machinery to quantify it: run the AOS extractor over the gold summaries and the input reviews, and report what share of gold PROS and CONS sentences have supporting evidence anywhere in the 1.5K reviews. If that share is high, the concern dissolves and the dataset is stronger for having measured it; if it is low, the Oracle ceiling in Section 4.2 is partly a measure of reference mismatch rather than of model capability.

2. The explanation for weak CONS performance is confounded with the reference itself. Section 4 attributes it to negative opinions being rarer in the input, a needle in a haystack. But Table 5 shows the reference also carries less than half as many negative sentences, 3.84 against 8.45, so the CONS score is computed against a much smaller target where each miss costs more. Separate the two explanations by reporting CONS performance normalised by reference length, or by subsampling PROS references to 3.84 sentences and re-scoring. Only then can the claim be about the input distribution rather than about the metric denominator.

3. Models that fail to produce output are scored on the cases they survived. Section 4 reports Gemini-1.5-Flash generating 372 valid summaries out of 500 in the basic setting, and Table 6 shows it evaluated over 495 entities in length control while the others use 500; Claude-3-Haiku adheres to both lengths in 450 of 500. Averaging quality only over produced summaries favours a model that declines the hard entities, and the hard entities here are exactly the ones with offensive or unusual content, as Section 5 describes. Report every table over the full 500 with failures scored as zero, next to the valid-only figures, and state the denominator in each caption.

4. The query terms are constrained by a fixed aspect lexicon. Section 3.2.1 ranks frequent unigrams and skip bigrams and then refines them against the gold hotel term list of Pontiki et al. That keeps the terms clean, but it also means the retrieval stage cannot surface any aspect outside a predefined SemEval inventory, which is precisely where idiosyncratic complaints live. Report the share of gold CONS sentences whose aspect falls outside that list; if it is substantial, the ceiling on the RAG variant is set by the lexicon rather than by the retriever or the model.

5. The new metrics rest on an extractor whose error rate is never measured. All of Table 4 depends on a pre-trained AOS triplet extractor. If it misses opinions or flips sentiment at some rate, every downstream number inherits that error, and there is no reason to assume it behaves the same on professional editorial prose and on user review sentences. Hand-annotate a sample, two hundred sentences would do, and report extraction precision and recall separately for the two text types.

Reproducibility

Section A.1 gives decoding settings for the RAG variants: max_new_tokens 256, temperature 0.7, top_p 0.9 across five models, with max_tokens 512 for the critic, and the prompts are in the appendix. GPT-4o-mini is dated to its August 2024 release. That is a reasonable base.

What is missing: snapshot identifiers and access dates for Claude-3-Haiku and Gemini-1.5-Flash, which changed during the period this work spans; the dense retriever checkpoint, referenced by a footnote rather than named in text; the frequency threshold and window used in candidate term extraction; the number of runs per configuration, since at temperature 0.7 a single generation per entity is one sample from a distribution and no seeds or repetitions are reported; and whether the dataset itself will be released, which for a dataset paper is the central question.

That last point needs care rather than a link. The corpus is crawled from two commercial platforms whose terms restrict redistribution, and Section 5 shows it contains personal data and offensive statements about identifiable staff, including a phone number. The paper should state the licence position, the scrape dates, and what PII filtering was applied before release.

What to fix

1. Section 2: measure and report the evidence overlap between Oyster gold sentences and the TripAdvisor input, using the AOS pipeline already built, and discuss how it bounds the Oracle in Section 4.2.
2. Section 4: re-report CONS results normalised for the 3.84 versus 8.45 reference imbalance in Table 5, so the needle-in-a-haystack explanation can be separated from the metric denominator.
3. Sections 4 and Table 6: score all models over the full 500 entities with failures as zero, and give the denominator in every table caption.
4. Section 3.2.1: report the share of gold CONS aspects outside the Pontiki term list, and consider an open-vocabulary variant of query term extraction.
5. Section 3.2.4 and Table 4: validate the AOS extractor on a hand-labelled sample, reporting precision and recall separately for expert summaries and user reviews.
6. Appendix A.1: name the Claude and Gemini snapshots with dates, name the dense retriever checkpoint in text, give the frequency threshold and window, and report the number of generations per entity with seeds.

7. Add a data statement: licence position for TripAdvisor and Oyster content, crawl dates, PII handling, and how the offensive content documented in Section 5 is treated on release.
8. Tables 2, 3, and 4 are discussed entirely in prose; put the headline numbers in the running text as well, and do not rely on green and red alone to mark the deltas in Table 3, since that is lost in print and for colour-blind readers.
9. Section 4.2: state how ties are handled in the Oracle baseline and whether the Oracle may select the same sentence for several gold sentences, since that affects how tight the reported ceiling really is.

Recommendation

Major revisions. The dataset is a genuine contribution, and the two pipelines are described well enough to reimplement, but the paper currently treats an independent expert review as if it were a summary of the user reviews, and that assumption propagates into every number, including the Oracle ceiling that frames the difficulty of the task. Measuring the overlap would either remove the objection or turn it into one of the more interesting findings here, namely how much of what experts say about a hotel is recoverable from a thousand guests talking about it. The other items are smaller: report over full denominators, normalise the CONS comparison, and validate the extractor on which the new metrics depend.

Evgenii Arsentev, PhD

Declarations

Potential competing interests: No potential competing interests to declare.