Dear Editor and Reviewers,

Thank you for taking the time to read our paper and for the helpful feedback. We really appreciate the minor-revision decision. We have updated the manuscript to address the remaining concerns.

Below is our point-by-point response.

Comment 1: Separate long-context retrieval vs. cross-lingual IR in the framing

Reviewer:
It would help to discuss the challenges of cross-lingual information retrieval (CLIR) and retrieval from long documents separately first, since they are different problems with different lines of work. The Introduction and Related Work should be more structured (e.g., a paragraph for each) before discussing the combination.

Our response:
We agree. In the previous version, we mixed the two problems too early, which made it look like they are a single problem with one line of prior work. In the revision, we reorganized both the Introduction and Related Work to make the structure much clearer.

What we changed:

- In the **Introduction**, we now clearly explain:
    - the long-document retrieval problem (lost-in-the-middle / needle-in-a-haystack), and
    - the cross-lingual retrieval problem (query in one language, evidence in another), as **two separate challenges**.
      After that, we explain why they become harder when they happen together.
- In **Related Work**, we split the section into separate parts:
    - work on long-context retrieval and long-context QA
    - work on CLIR
    - work on RAG design choices (including chunking / sentence-level retrieval)

- work on neurosymbolic / code-based reasoning
  Then we summarize how our paper fits into the combined setting.

We believe this directly addresses the reviewer's concern that we were "tying the two problems together" too early and not acknowledging the rich literature on each separately.

Comment 3: What is actually new about CROSS? And evaluate retrieval before generation

Reviewer:
CROSS sounds very similar to existing cross-lingual RAG systems (embed chunks and retrieve top-k). If CROSS is a contribution, it should be clearer what is novel. Also, retrieval should be evaluated at the retrieval stage (before generation), not only by comparing to LLM-only.

Our response:
We understand this concern and we agree with the key point: **CROSS is not meant to be a brand-new retrieval algorithm.** At a high level, it follows the standard dense-retrieval RAG pipeline (segment → embed → retrieve top-k). The main contribution of our paper is **NSAR (the reasoning layer)** and the combined evaluation setting. In the revision, we rewrote parts of the paper to make this explicit and avoid over-claiming.

What we changed:

1. **We clarified the positioning of CROSS.**
   In the CROSS section, we added a clear paragraph explaining that CROSS is aligned with standard dense cross-lingual RAG methods. We explain that CROSS is best understood as a **practical retrieval backbone** designed for our specific regime (very long single documents + cross-lingual queries + strict LLM budget via a sentence cap).

2. **We added more related work on RAG chunking / sentence-level retrieval.**
   We agree that sentence-level chunking alone is not a new idea. The revised Related Work now discusses prior work on chunking/granularity and sentence-level retrieval, and we cite representative references. This addresses the concern that we were not placing this choice in context.

3. **We made the "retrieval before generation" evaluation clearer.**
   We already had an embedding failure vs. LLM failure breakdown. In the revision, we explicitly frame this as a standard retrieval-stage metric (whether the gold needle is included in the top-k retrieved sentences before the LLM answers). In other words, we now clearly report:

- retrieval-stage success (needle appears in retrieved top-k), and
- end-to-end success (final answer correctness).
  This isolates retrieval behavior from reasoning behavior without requiring new experiments.

Follow-up (a): “If the main idea is sentence-level chunks, novelty is limited”

Reviewer:
If the key novelty is sentence-level chunking, that is limited and needs more chunking-related discussion.

Our response:
We agree. In the revision, we do **not** present sentence-level chunking as the main novelty. We acknowledge prior sentence-level RAG work and cite it. We also adjust our language so CROSS is not framed as a fundamentally new algorithm. Instead, CROSS is described as a standard RAG-style retrieval backbone that is tuned and evaluated for our specific combined setting, while NSAR is the main reasoning contribution.

Follow-up (b): What does “massive scale (512k tokens)” mean?

Reviewer:
What does 512k tokens refer to?

Our response:
Thank you for pointing out that this was unclear. We corrected and clarified this throughout the paper. The “512k” refers to the **haystack length in our dataset**, measured as **512,000 words** (not tokenizer tokens). We made the wording consistent everywhere to avoid confusion.

Thank you again for the detailed feedback—it helped us improve the clarity and positioning of the paper.

Sincerely,
Sina Bagheri Nezhad and Ameeta Agrawal