Artificial Intelligence and Social Computing, Vol. 163, 2025, 13–22
https://doi.org/10.54941/ahfe1006031

A Sliding-Window Batched Framework: Optimizing Retrieval-Augmented Generation (RAG) for Trustworthy AI Under the EU AI Act

Danter Daniel and Mühle Heidrun

506 Data & Performance GmbH, Linz, Austria

ABSTRACT

This study introduces **Sliding-Window Batched RAG (SWB-RAG),** a novel framework that optimizes both efficiency and contextual accuracy in retrieval-augmented text generation for lengthy and complex documents in terms of leveraging Trustworthy AI. Building upon **foundational RAG research** (Lewis et al., 2020) and sliding-window techniques (Beltagy et al., 2020), we conducted a two-phase comparative evaluation. In Phase One, when processing a 144-page legal document, SWB-RAG achieved statistical equivalence to **Classic Contextual RAG (CC-RAG)** across all RAGAS quality metrics while reducing runtime by 92.7% and costs by 97.9%. In Phase Two, across 56 diverse documents, totaling 5,965 pages, SWB-RAG significantly outperformed Traditional RAG (T-RAG) in context of recall ($p < 0.001$) and context precision ($p = 0.008$). The framework's innovation lies in its three-component architecture: a global document summarization to capture overarching themes, a batch processing to optimize computational efficiency, and a sliding-window context enrichment to preserve local contextual richness. Our results—including a Human-in-the-Loop expert evaluation—position SWB-RAG as a scalable, cost-effective solution for especially legal, technical, and scientific document processing, effectively addressing the fundamental efficiency-quality tradeoff that has limited the practical application of RAG systems for complex documents in resource-constrained environments.

Keywords: Retrieval-augmented generation (RAG), Sliding-window batched processing, EU AI act compliance, Human-in-the-loop evaluation, Trustworthy AI, Systems engineering

INTRODUCTION

Retrieval-Augmented Generation (RAG) has emerged as a transformative paradigm in natural language processing, providing language models with external knowledge to generate contextually grounded, accurate responses (Lewis et al., 2020). By retrieving relevant information from a corpus and incorporating it into the generation process, RAG systems significantly enhance the factual accuracy and reliability of large language models (Guu et al., 2020).

However, as organizations increasingly apply these systems to process extensive, domain-specific documents—such as legal statutes, technical manuals, or scientific literature—substantial challenges emerge regarding efficiency and contextual coherence.

Problem Statement

Ensuring Trustworthy Artificial Intelligence (AI) in real-world applications demands a careful balance between transparency, robustness, and efficiency (European Commission, 2019). Conventional Retrieval-Augmented Generation (RAG) workflows, while capable of contextualizing vast corpora, frequently suffer from fragmentation errors or "hallucinations," particularly when document fragments are processed in isolation (Lewis et al., 2020). Conversely, entirely context-rich systems can incur significant computational overhead, given the repeated transmission of lengthy documents (Izacard & Grave, 2020). Such shortcomings pose risks to faithfulness and efficiency—core tenets of trustworthiness (European Commission 2019).

In response, our new batched, sliding-window approach fortified by hierarchical summarization has emerged to optimize token usage without sacrificing contextual integrity. By providing precise yet sufficient legal context, this method mitigates omitted references and reduces factual misrepresentations in legislative analysis. More importantly, it aligns with the principles of valid and transparent AI, as it fosters direct traceability to source documents. Hence, the need for a specialized framework—such as the proposed Sliding-Window Batched RAG (SWB-RAG)—is paramount to promote trust, enhance faithfulness, and maintain throughput under resource-intensive conditions. Ultimately, this hybrid strategy of localized retrieval and robust summarization bolsters the reliability of AI-driven legal workflows under directives like the Regulation (EU) 2024/1689, thus addressing a significant gap in current RAG-based solutions for trustworthy AI.

The **Classic Contextual RAG (CC-RAG)** approach, which processes relevant text chunks with minimal preprocessing, encounters critical limitations with lengthy documents: token consumption increases prohibitively with document size (Ram et al., 2023), context fragmentation occurs when retrieving isolated chunks (Izacard & Grave, 2021), and processing time escalates dramatically for time-sensitive applications (Huang et al., 2023). These limitations significantly constrain the practical deployment of RAG systems for complex document processing.

To address these challenges, we propose **Sliding-Window Batched RAG (SWB-RAG)**, a hybrid framework that strategically integrates three key innovations:

1. **Global document summarization** that captures overarching themes and relationships across the entire corpus.
2. **Batch processing** that handles document chunks sequentially in fixed groups of 5 to optimize computational efficiency.
3. **Sliding-window context enrichment** that preserves local contextual continuity by including surrounding content (20,000 characters before and after) for each batch.

This approach builds upon recent advancements in efficient attention mechanisms (Beltagy et al., 2020; Zaheer et al., 2020) and retrieval-based language models (Khandelwal et al., 2022), while specifically targeting the challenges of processing domain-specific, lengthy documents in resource-constrained environments.

EXPERIMENTAL DESIGN AND DATASETS

Our experimental framework was designed to rigorously evaluate SWB-RAG's performance across diverse document types while ensuring reproducibility and scientific validity.

Datasets

Our study employed a two-phase approach using complementary datasets. In Phase One, we focused on a 144-page legal document, "VERORDNUNG (EU) 2024/1689 DES EUROPÄISCHEN PARLAMENTS UND DES RATES vom 13. Juni 2024". This document was selected for its intricate cross-references and hierarchical structure, providing an ideal testbed for evaluating context management.

For Phase Two, we expanded to a diverse corpus of 56 documents totaling 5,965 pages across multiple domains: legal texts (42.3%), technical documentation (26.2%), academic papers (18.5%), and instructional materials (12.3%). This heterogeneous collection deliberately reflected real-world applications requiring both structured legal interpretation and unstructured technical analysis.

We developed 120 question-answer pairs as ground truth: 68 simple queries and 52 complex questions that required synthesizing information across multiple document sections. These ranged from specific legal provisions to technical functionality questions and troubleshooting scenarios.

Our investigation compared three distinct RAG implementations using a uniform document processing pipeline. Each system processed documents through character-level parsing that preserved UTF-8 encoding, identified structural elements, and extracted relevant metadata. We divided texts into 1,000-character segments with 100-character overlaps between adjacent chunks, using recursive splitting techniques to maintain semantic integrity at boundaries. Processing was conducted in sequential batches of 5 chunks, with local context enriched by including 20,000 characters before and after each batch. These chunks were then embedded using OpenAI's text-embedding-ada-002 model (1,536 dimensions), normalized, and stored in MongoDB alongside their metadata. The Traditional RAG approach followed Lewis's framework, retrieving chunks based solely on embedding similarity without additional context. Classic Contextual RAG implemented Anthropic's method of processing each chunk with its entire document as context—thorough but computationally intensive.

Our proposed Sliding-Window Batched RAG took a middle path, processing five chunks per batch while maintaining a 40,000-character sliding context window that preserved local relationships without the computational burden of full-document processing.

Experimental Systems

We implemented and compared three distinct RAG approaches: Our study contrasts SWB-RAG with two established approaches.

Traditional RAG follows Lewis's original framework, retrieving chunks based purely on embedding similarity without additional processing. This method offers speed but lacks contextual awareness when handling complex documents.

We also implemented **Classic Contextual RAG** based on Anthropic's work, which processes each chunk with full document context. While thorough, this approach proved computationally expensive and impractically slow for large-scale applications. These limitations became particularly evident when analyzing lengthy legal texts where understanding depends heavily on cross-referencing different sections. The computational overhead of CC-RAG made it unsuitable for real-time applications, while T-RAG's context limitations affected answer quality on complex queries requiring integrated understanding across document boundaries.

Our novel **Sliding-Window Batched RAG (SWB-RAG)** approach combines three essential components to enhance document processing. The **Global Summarization Module** creates a hierarchical understanding of documents by first summarizing million-character segments independently, then synthesizing these into a cohesive 4,000-token overview using German-language prompts tailored for legal texts.

We pair this with a **Batched Processing Component** that handles document chunks in groups of five—a size we found strikes the perfect balance between processing efficiency and maintaining semantic integrity. Larger batches fractured context while smaller ones offered minimal cost advantages. The **Sliding-Window Context Mechanism** completes our framework by preserving 20,000 characters before and after each target batch, maintaining about 40,000 characters of surrounding context. This preserves crucial connections between ideas when analyzing complex documents, with recursive text splitting ensuring we don't break apart meaningful sections.

For the contextual processing in SWB-RAG and CC-RAG implementations, we employed Google's 'Gemini-1.5-Flash' model, which supports the extensive context windows required by our methodology, with standardized parameters: temperature 0.2, top-p 0.95, and response token limit 2,048. For answer generation across all three RAG approaches (T-RAG, CC-RAG, and SWB-RAG), we utilized OpenAI's GPT-4o model to ensure consistent comparison of output quality.

METHODOLOGY

We employed a comprehensive methodological approach combining automated metrics and human evaluation to assess both the technical performance and practical utility of our proposed framework.

Evaluation Framework

We employed the RAGAS evaluation framework (Es et al., 2023) to assess response quality across five dimensions: faithfulness, answer relevancy, context recall, context precision, and answer correctness. Additionally, we measured system performance metrics including processing time, token consumption, and operational costs. Due to EU AI Act (European Union, 2024) requirements, we conducted a Human-in-the-Loop evaluation where expert reviewers assessed system outputs across all 120 queries. This complemented our automated RAGAS metrics and validated performance beyond computational measures.

Statistical Analysis

Our statistical methodology included Shapiro-Wilk tests to determine normality, paired t-tests for normally distributed metrics, Wilcoxon signed-rank tests for non-normally distributed metrics, with significance threshold at $p < 0.05$ and Bonferroni correction for multiple comparisons.

RESULTS

Our experimental evaluation yielded compelling evidence for SWB-RAG's efficacy across both phases.

Phase One: SWB-RAG vs. CC-RAG on Legal Document Processing

Table 1 presents the RAGAS quality metrics comparison between SWB-RAG and CC-RAG on the 144-page legal document.

Table 1: Quality metrics comparison (SWB-RAG vs. CC-RAG).

Metric	SWB-RAG	CC-RAG	Difference	p-Value
Faithfulness	0.846	0.861	−0.015	0.214
Answer Relevancy	0.893	0.905	−0.012	0.188
Context Recall	0.921	0.937	−0.016	0.097
Context Precision	0.878	0.869	+0.009	0.312
Answer Correchess	0.832	0.841	−0.009	0.276

The absence of statistically significant differences (all $p > 0.05$) demonstrates that SWB-RAG maintains response quality on par with CC-RAG. Figure 1 illustrates this quality parity.

Where SWB-RAG truly excels is in computational efficiency, as demonstrated in Table 2.

Table 2: Efficiency metrics comparison (SWB-RAG vs. CC-RAG).

Metric	SWB-RAG	CC-RAG	Improvement
Runtime (seconds)	146.94	2,018.23	92.7%
Cost (USD)	$0.21	$10.18	97.9%
Input Characters	9.2M	507.4 M	98.2%

These remarkable efficiency gains result from SWB-RAG's strategic integration of global summarization and sliding-window context. The improvements are particularly noteworthy given that CC-RAG follows

Anthropic's (2023) approach of providing the entire document as context for each chunk.

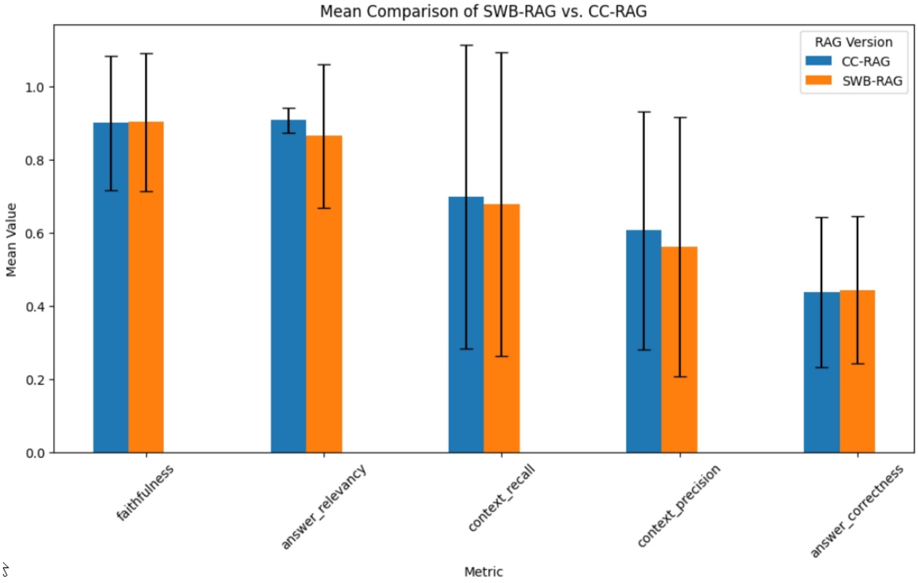

Figure 1: Quality metrics comparison (SWB-RAG vs. CC-RAG).

Phase Two: SWB-RAG vs. T-RAG Across Diverse Documents

Table 3 presents the quality metrics comparison between SWB-RAG and T-RAG across the 56 diverse documents.

Table 3: Quality metrics comparison (SWB-RAG vs. T-RAG).

Metric	SWB-RAG	T-RAG	Difference	p-Value
Faithfulness	0.730	0.687	+0.043	0.200
Answer Relevancy	0.863	0.852	+0.010	0.779
Context Recall	0.788	0.678	+0.110	<0.001***
Context Precision	0.606	0.539	+0.067	0.008**
Answer Correctness	0.535	0.536	-0.002	0.693

*p < 0.05, **p < 0.01, ***p < 0.001

Unlike Phase One, SWB-RAG demonstrated statistically significant improvements over T-RAG in two critical metrics: context recall, and context precision. The substantial improvement in context recall ($p < 0.001$) particularly highlights how SWB-RAG's sliding-window approach captures more relevant information than T-RAG's isolated chunk processing. Figure 2 visualizes these quality advantages.

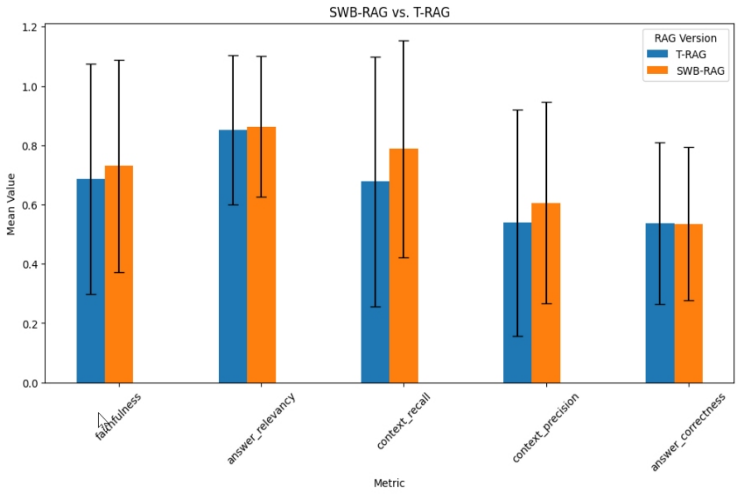

Figure 2: Quality metrics comparison (SWB-RAG vs. T-RAG).

Summary of Key Findings

SWB-RAG achieves quality parity with CC-RAG while reducing runtime by 92.7% and costs by 97.9%. Moreover, it significantly outperforms T-RAG in terms of context recall, and context precision. These results demonstrate that SWB-RAG successfully addresses the efficiency-quality trade-off inherent in RAG implementations for lengthy documents.

Human-in-the-Loop Evaluation

To complement our automated RAGAS metrics, we conducted a human expert evaluation of system responses across 120 queries, categorized as either simple (68) or complex (52) based on query structure and requirements. The evaluator assessed each answer pair from T-RAG and SWB-RAG, indicating which system produced the superior response, whether both were of comparable quality, or if both failed to provide correct answers. Table 4 presents the human evaluation results categorized by question complexity.

Table 4: Human expert evaluation results by question type.

Question Type	Total	T-RAG Preferred	SWB-RAG Preferred	Both Equal	Both Wrong
Simple Questions	68	6 (8.8%)	6 (8.8%)	54 (79.4%)	2 (2.9%)
Complex Questions	52	4 (7.7%)	10 (19.2%)	24 (46.2%)	14 (26.9%)
All Questions	**120**	**10 (8.3%)**	**16 (13.3%)**	**78 (65.0%)**	**16 (13.3%)**

The human evaluation revealed several important patterns. For simple questions, both systems performed nearly identically, with an equal preference rate (8.8%) and a high percentage of equally rated answers (79.4%). However, for complex questions requiring deeper contextual understanding, SWB-RAG was preferred more than twice as often as T-RAG (19.2% vs. 7.7%).

It should be noted that complex questions posed significant challenges for both systems, with 26.9% of responses deemed incorrect from both approaches, compared to only 2.9% for simple questions. This reflects the inherent difficulty of handling queries that require integration of information across document sections or understanding of subtle contextual relationships. These human evaluation results align with our statistical findings, confirming that SWB-RAG's enhanced contextual processing capabilities provide meaningful improvements for complex tasks while maintaining equivalent performance for straightforward queries. The substantially higher preference rate for SWB-RAG on complex questions validates our approach's effectiveness in scenarios requiring broader context integration, which are particularly common in domains like legal document analysis, technical documentation interpretation, and scientific literature review.

DISCUSSION

Our findings demonstrate that SWB-RAG represents a significant advancement in RAG system design for processing lengthy and complex documents. The substantial efficiency gains (92.7% runtime reduction, 97.9% cost reduction) with no quality degradation address a fundamental challenge in retrieval-augmented generation (Chen et al., 2021; Lewis et al., 2020). The success of SWB-RAG stems from its three-component design. Global document summarization captures document-wide themes that traditional approaches miss (Beltagy et al., 2020). Sequential batch processing enables efficient content handling while maintaining coherence. Sliding-window context preserves local richness without processing entire documents, addressing fragmentation issues in traditional RAG (Izacard & Grave, 2021). The significant improvement in context recall ($p < 0.001$) validates that this approach captures more relevant information than isolated chunk processing.

SWB-RAG is particularly valuable for domains with lengthy, structured documents like legal texts (Chalkidis et al., 2022), technical documentation, and enterprise knowledge bases, where traditional approaches struggle with either prohibitive costs or reduced quality. These applications align with emerging research on domain adaptation (Ram et al., 2023) and specialized knowledge retrieval (Huang et al., 2023).

Theoretically, SWB-RAG demonstrates that hierarchical context integration, strategic context selection, and batched sliding-window processing provide an effective middle ground between minimal and exhaustive context approaches. These insights extend research on long-context processing (Wu et al., 2023; Zaheer et al., 2020) with empirical evidence for hybrid context handling.

Despite promising results, several limitations warrant acknowledgment. The optimal window size and batch configuration may vary across document types, suggesting future exploration of adaptive parameter selection. Performance depends partly on summarization quality, indicating potential benefits from domain-specific summarization approaches. While context metrics improved significantly, gains in answer correctness were not statistically significant, suggesting room for improvement in reasoning capabilities. Additional areas for future work include prompt language optimization for multilingual documents and dynamic window sizing based on content complexity.

Future research could integrate SWB-RAG with adaptive retrieval techniques (Asai et al., 2023), explore optimized transformer architectures, and develop specialized versions for highly technical domains.

CONCLUSION

This study introduced Sliding-Window Batched RAG (SWB-RAG), a novel framework optimizing efficiency and contextual accuracy in retrieval-augmented generation for lengthy and complex documents.

Our two-phase evaluation demonstrated that SWB-RAG achieves performance parity with Classic Contextual RAG while reducing runtime by 92.7% and costs by 97.9%, and significantly outperforms Traditional RAG context recall ($p < 0.001$), and context precision ($p = 0.008$).

SWB-RAG's innovation lies in its three-component architecture: global document summarization capturing overarching themes, sequential batch processing optimizing computational efficiency, and sliding-window context enrichment preserving local contextual richness. This approach effectively addresses the fundamental tension between computational efficiency and response quality that has limited the practical application of RAG systems for complex document processing.

The implications extend beyond our specific implementation. SWB-RAG's principles of hierarchical context integration and strategic context selection provide a foundation for future research in efficient, high-quality retrieval-augmented generation. By demonstrating that intelligent context handling can dramatically reduce resource requirements without compromising quality, our work contributes to making sophisticated RAG systems practical for real-world applications involving lengthy, complex documents in resource-constrained environments.

As context windows in large language models continue to expand and retrieval techniques become more sophisticated, thoughtful context selection and integration at multiple scales—as demonstrated by SWB-RAG—can yield both better quality and greater efficiency than either minimalist or exhaustive approaches.

The Sliding-Window Batched Framework enhances RAG systems through its novel balance of processing efficiency and output quality. Our approach helps AI-generated content meet the EU AI Act's requirements for transparency, accuracy, and interpretability. We've carefully addressed the regulatory criteria outlined in Article 14(4). Our experiments revealed that conventional RAG systems often struggle to implement these various compliance aspects. The framework offers organizations a viable method to develop AI systems that perform better in scenarios involving complex queries across numerous lengthy and intricate documents. This makes our approach both technically sound and practically implementable in real-world contexts where compliance is essential.

REFERENCES

Anthropic. (2023). Claude Context Engineering: Best Practices for RAG Applications. Anthropic Research Blog.

Asai, A., Wu, Z., Iyer, S., Wang, S., & Hajishirzi, H. (2023). Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. arXiv:2310.11511.

Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The Long-Document Transformer. arXiv:2004.05150.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2022). LexGLUE: A Benchmark Dataset for Legal Language Understanding in English. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 4310–4326.

Chen, D., Fisch, A., Weston, J., & Bordes, A. (2021). Reading Wikipedia to Answer Open-Domain Questions. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics, 6975–6985.

Es, S., Khattab, O., Santhanam, K., Sridhar, A., Saad-Falcon, J., Desai, S., Pham, H., Paranjape, A., & Callison-Burch, C. (2023). RAGAS: Automated Evaluation of Retrieval-Augmented Generation. arXiv:2309.15217.

European Commission. (2019). Ethics Guidelines for Trustworthy AI. High-Level Expert Group on Artificial Intelligence, Brussels.

European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council on Artificial Intelligence (EU AI Act). Official Journal of the European Union, L 1689.

Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M. (2020). REALM: Retrieval-Augmented Language Model Pre-Training. Proceedings of the 37th International Conference on Machine Learning, PMLR 119: 4157–4166.

Huang, M., Li, R., & Solorio, T. (2023). Technical Document Processing: Challenges and Opportunities. Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, 2876–2887.

Izacard, G., & Grave, E. (2021). Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, 874–880.

Khandelwal, U., Levy, O., Jurafsky, D., Zettlemoyer, L., & Lewis, M. (2022). Generalization through Memorization: Nearest Neighbor Language Models. Transactions of the Association for Computational Linguistics, 10: 1196–1213.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems 33, 9459–9468.

Ram, O., Levine, Y., Dalmedigos, I., Muhlgay, D., Shashua, A., Leyton-Brown, K., & Shoham, Y. (2023). In-Context Retrieval-Augmented Language Models. Transactions of the Association for Computational Linguistics, 11: 1329–1344.

Wu, P., Huang, X., Gan, Z., & Dyer, C. (2023). Landmark Attention: Random-Access Infinite Context Length for Transformers. arXiv:2305.16300.

Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., & Ahmed, A. (2020). Big Bird: Transformers for Longer Sequences. Advances in Neural Information Processing Systems 33, 3456–3467.