Critical Reading Exam

Large Language Models for Historical Document Intelligence: Retrieval-Augmented Generation, Automated Evaluation, and Agentic Architectures

Stergios Konstantinidis
Faculty of Business and Economics (HEC), University of Lausanne
Submitted to the Doctoral Commission, June 2026

***Abstract*—This document presents five critical reviews written in preparation for the PhD Critical Reading Exam. The selected articles span four interconnected themes in historical document intelligence: (1) retrieval-augmented generation (RAG) applied directly to historical newspaper archives, examining a multilingual RAG pipeline over OCR-corrupted corpora and the foundational RAG framework from which it descends; (2) post-OCR correction using synthetic data generation and test-time adaptation with large language models (LLMs); (3) automated LLM-based evaluation through the LLM-as-a-judge paradigm and its reliability implications for archival assessment pipelines; and (4) multi-modal long-context document understanding through agentic frameworks. Across these themes, a consistent tension emerges: the architectural assumptions that make LLM-based document intelligence effective—clean text, monolingual content, well-structured layouts, and static knowledge bases—are systematically violated by the historical and archival corpora that most urgently require it.**

***Index Terms*—retrieval-augmented generation, post-OCR correction, LLM-as-a-judge, agentic document understanding, multilingual archival corpora, historical newspaper digitisation**

I. Motivation for Paper Selection

THE five papers selected for this exam reflect the core challenges of my doctoral research, which examines how large language models and retrieval-augmented generation can be applied to the processing and intelligent querying of historical and archival documents. The central thesis of this selection is architectural: the Retrieval-Augmented Generation (RAG) framework defines the paradigm within which all subsequent work in this selection operates, but its foundational assumptions—clean text, monolingual content, static knowledge bases—must be confronted by each of the four remaining papers in succession, each from a distinct angle.

The first paper [1] by Tran et al. is the direct practical application of RAG to historical newspaper archives, and the primary motivation anchor for this entire selection. Published at JCDL 2024—the premier ACM/IEEE venue for digital libraries—it proposes a multilingual RAG pipeline with hybrid dense retrieval and NER-enhanced reranking evaluated on the OCR-corrupted NewsEye corpus. This paper establishes the problem setting: multilingual, OCR-corrupted archival text is the domain; RAG is the chosen architectural approach; and ground-truth-free LLM-based evaluation metrics are the proposed solution to annotation scarcity. The remaining four papers each address a prerequisite, limitation, or extension of this system.

The second paper [2] by Guan et al. addresses the data scarcity challenge inherent in any LLM-based processing of historical text: obtaining aligned (OCR, ground-truth) training pairs is expensive. They propose a principled synthetic data pipeline and a novel test-time adaptation mechanism for proper nouns—directly relevant to the rare entities that dominate historical archives and are systematically degraded by OCR.

The third paper [3] by Gu et al. provides the systematic taxonomy of the LLM-as-a-judge paradigm that underlies the ground-truth-free evaluation metrics proposed by Tran et al. Understanding the biases and reproducibility limitations of LLM-based evaluation is essential for any archival RAG pipeline where human-annotated QA pairs are unavailable at scale.

The fourth paper [4] moves from retrieval to agentic document understanding, introducing DocAgent—a multi-agent framework with selective retrieval, multi-modal tool use, and a reviewer agent for answer verification. This paper projects the RAG paradigm forward into agentic architectures and directly anticipates the document intelligence

capabilities required for our downstream question-answering over the BCUL archive.

The fifth paper [5] is the foundational RAG work by Lewis et al., which introduces the dual-encoder dense retriever and BART-based generator trained end-to-end with retrieved documents as latent variables. Reading it last—after encountering the domain-specific challenges in Articles 1–4—inverts the conventional order to pedagogical effect: when the framework's baseline assumptions (clean text, monolingual content, static knowledge bases) are encountered *after* four papers that systematically violate them, each assumption registers not as a design choice but as a falsified prior, revealing the gap between the framework's encyclopedic English-Wikipedia origin and the multilingual, OCR-corrupted historical archive that motivated this entire selection with a clarity that chronological reading cannot achieve. Across all five papers, a single analytic thread runs: the assumptions embedded in each framework are most severely tested by the document types that most urgently require it.

II. Summary and Motivation of Selected Publications

A. Article 1: Tran et al. (2024)

Full reference: Tran, T.T., González-Gallardo, C.-E., & Doucet, A. [1]. *Retrieval Augmented Generation for Historical Newspapers.* In Proceedings of the 24th ACM/IEEE Joint Conference on Digital Libraries (JCDL 2024), Article 58, 5 pp. ACM, 2024.

1) Summary: Tran et al. present one of the first RAG-based systems designed explicitly for question-answering over historical newspaper archives. The pipeline is evaluated on the **NewsEye** dataset—a multilingual corpus of historical newspapers in English, French, German, and Finnish—and consists of four modular components: a query router; a dense retrieval module using multilingual E5 embeddings indexed in Chroma; a hybrid reranking module combining a Cohere multilingual cross-encoder with a custom NER-based scoring component derived from a historical multilingual BERT model (hmBERT); and a LLaMA 3 generator. A further methodological contribution is the proposal of *ground-truth-free evaluation metrics*—faithfulness and relevancy scored by an LLM judge—for settings where annotated QA pairs are unavailable. Results show that dense retrieval substantially outperforms BM25 on OCR-corrupted historical text, and that cross-encoder reranking provides approximately 0.1 NDCG@10 improvement; NER-enhanced reranking shows mixed results depending on OCR and NER quality.

2) Strengths: This paper's most significant contribution is its direct engagement with the full pipeline of challenges present in archived historical newspaper corpora: OCR noise, multilingualism, temporal language drift, and the absence of annotated QA ground truth. The finding that dense multilingual embeddings substantially outperform BM25 on OCR-corrupted historical text has direct implications for any archival retrieval system: vocabulary mismatch between corrupted OCR output and clean query terms makes keyword-based retrieval fragile in exactly this setting. The hybrid reranking design—combining cross-encoder relevance with entity-level NER signals from hmBERT—is well-motivated given the entity-density of historical newspaper content, where person names, place names, and organisation mentions are primary relevance signals. The ground-truth-free evaluation metrics address a genuine methodological gap for under-annotated archival collections, providing a path to principled evaluation that does not require expensive expert annotation. JCDL is the premier ACM/IEEE venue for digital libraries research, ensuring the paper's evaluation by domain experts in digital preservation and information retrieval.

3) Weaknesses: At five pages, the paper is a short-paper contribution and many architectural details—the query router's classification logic, the exact weighting function for the hybrid reranking score, chunking parameters—are insufficiently specified for replication. The NER reranking results are mixed across languages and underanalysed: the paper does not determine when entity-based reranking helps versus hurts as a function of OCR quality, language, or entity density, leaving the component's contribution empirically underspecified. The ground-truth-free evaluation relies on an LLM judge whose own biases—position bias, verbosity bias, and English-centric evaluation quality for non-English archives—are not addressed [3]. More fundamentally, the central methodological claim of the paper hinges on the reliability of faithfulness and relevancy scores computed by an LLM judge operating over multilingual, OCR-degraded

input; this reliability is never established, leaving the evaluation framework without a validity foundation precisely in the domain conditions for which it was proposed. The Cohere reranker is a commercial closed-source API, raising sustainability and privacy concerns for national library deployments; no open-weight reranker alternative is evaluated. Finally, no ablation of the generator is presented: it is unclear how much of the observed performance is attributable to better retrieval and reranking versus the inherent generative capabilities of LLaMA 3.

B. Article 2: Guan et al. (2024)

Full reference: Guan, S., Xu, C., Lin, M., & Greene, D. [2]. *Effective Synthetic Data and Test-Time Adaptation for OCR Correction.* In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024), pp. 15412–15425. ACL, 2024.

1) Summary: This paper addresses the central data scarcity challenge in post-OCR correction: obtaining aligned (noisy OCR, ground truth) training pairs at scale is expensive. The authors propose two complementary contributions. First, they develop a principled synthetic data generation framework that extracts character-level OCR error distributions from the ICDAR 2017/2019 datasets and injects these at varying Error Levels (ELs) into clean text corpora, using weak supervision (retaining all aligned pairs without manual filtering). An Optimal Alignment Curve (OAC) is fitted over the resulting data to identify a CER threshold ($\approx$20.1%) separating "effective" from "ineffective" synthetic training examples. Second, they propose Self-Correct-Noise Test-Time Adaptation (SCN-TTA): a seven-step pipeline that uses BookNLP to extract proper nouns from the test text, generates synthetic noisy versions with those nouns preserved, and fine-tunes the model on this text-specific data at inference time. Evaluated on the RETAS benchmark of 19th-century English novels, ByT5 with multi-noise synthetic data and SCN-TTA achieves a 68.67% CER reduction without any manually annotated training data.

2) Strengths: The principled approach to synthetic data generation—varying noise levels and empirically determining the effective threshold—is the paper's most significant methodological contribution. The OAC provides a rare, data-driven tool for calibrating synthetic corpora, with potential applicability beyond post-OCR tasks. The SCN-TTA mechanism is creative and well-motivated: proper nouns are a known weakness of generic correction models, and the test-time adaptation elegantly exploits the redundancy of proper nouns within a single long text to generate labelled examples on the fly. The result of achieving near-benchmark-level performance *without* any manual annotation is practically compelling for large-scale digitisation projects where labelling is prohibitively expensive.

3) Weaknesses: The evaluation is restricted to English 19th-century novels, a relatively narrow domain. More fundamentally, the Optimal Alignment Curve—the paper's flagship methodological contribution—is fitted on a single combination of model architectures (ByT5, Flan-T5, mBART) and a single pair of datasets (ICDAR 2017/2019 as the error source, RETAS as the target). The resulting 20.1% CER threshold is presented with a precision that implies generalisability, yet the paper provides no evidence that this threshold is stable across different OCR engines with distinct error distributions, different clean text corpora, or different language families. For archival collections digitised with heterogeneous scanning equipment over decades—as is typical of national library holdings—the OAC is essentially an unfalsified claim. The SCN-TTA mechanism compounds this concern: it relies on BookNLP for proper noun identification, which can fail on heavily corrupted OCR text, and assumes that proper nouns appear frequently enough within a single text for the adaptation to be meaningful—an assumption that fails for short newspaper articles or archival fragments. Furthermore, the term "weak supervision" is used loosely and diverges from established frameworks (e.g., data programming [6]), which could mislead practitioners. Constrained decoding approaches, which explicitly penalise outputs that deviate character-by-character from the OCR input and have been shown to outperform both prompting and standard fine-tuning on archival texts [7], are absent from the comparison entirely—a gap that leaves the practical significance of SCN-TTA inadequately situated within the field's current state of the art.

C. Article 3: Gu et al. (2026)

Full reference: Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu,

H., Wang, S., Zhang, K., Lin, Z., Zhang, B., Ni, L., Gao, W., Wang, Y., & Guo, J. [3]. *A Survey on LLM-as-a-Judge. The Innovation*, 101253. DOI: 10.1016/j.xinn.2025.101253, 2026.

1) Summary: This survey provides a comprehensive overview of the "LLM-as-a-judge" paradigm, in which large language models are employed as automated evaluators for complex generation tasks. The paper offers a formal definition and taxonomy classifying LLM-based judgements by input format (pointwise, pairwise, listwise), output format (binary, scored, ranked, open-ended), and by usage dimension: evaluation, selection, and verification. The survey systematically reviews strategies for enhancing reliability, including improving consistency, mitigating biases (position bias, verbosity bias, self-enhancement bias, authority bias), and adapting to diverse evaluation scenarios. A meta-evaluation benchmark is proposed for assessing the reliability of LLM judges themselves, addressing dimensions of consistency, robustness, sensitivity, and correlation with human judgement. Applications are surveyed across core NLP tasks (natural language generation evaluation, code generation, reasoning, summarisation) and high-stakes domains (healthcare, law, finance, education). The paper concludes with a research agenda emphasising trustworthy and hybrid human-AI evaluation frameworks.

2) Strengths: The survey's most significant contribution is its formal taxonomy of a paradigm that had previously been applied ad hoc under inconsistent terminology. The systematic classification by input format, output format, and usage dimension gives the field a shared vocabulary and enables principled comparison of disparate approaches. The treatment of evaluation biases is particularly rigorous: position bias, verbosity bias, self-enhancement bias, and authority bias are each documented with empirical evidence and paired with concrete mitigation strategies (position swapping, calibration, multi-judge ensembles), making the survey immediately actionable for practitioners designing evaluation pipelines. The meta-evaluation framework—evaluating the evaluators—addresses a genuine methodological gap: if LLM judges are used to assess system outputs, a principled method for determining those judges' own reliability is essential for the field to make meaningful progress. The breadth of application coverage—from NLG evaluation and code generation to healthcare and legal reasoning—demonstrates the cross-cutting relevance of the paradigm. The forward-looking research agenda, including hybrid human-AI evaluation and epistemically grounded assessment, articulates a substantive vision beyond cataloguing.

3) Weaknesses: The survey's most consequential methodological weakness is that its central contribution—the meta-evaluation benchmark for assessing the reliability of LLM judges—is proposed conceptually without empirical grounding: the paper does not demonstrate how well the benchmark discriminates between good and poor LLM judges under any specific task or domain condition, nor does it characterise the benchmark's own failure modes or scope of applicability. A framework for evaluating evaluators that cannot itself be evaluated is not yet a scientific instrument; it is a research agenda. This gap is directly operative for the pipeline developed in Tran et al. [1] and Guan et al. [2]: both rely on LLM-based quality assessment in exactly the conditions—OCR-corrupted, multilingual, under-annotated archival text—where the reliability of LLM judges is most uncertain and the meta-evaluation benchmark would be most needed. The survey's connection to RAG and post-OCR correction pipelines more broadly is indirect: it does not discuss how LLM-based evaluation applies to retrieval quality assessment or character-level OCR scoring, nor how archival metrics (NDCG@10, CER, WER) relate to the scoring frameworks surveyed. The overwhelming majority of cited studies rely on GPT-4 or proprietary equivalents, a dependency the survey acknowledges but does not resolve: if LLM-as-judge evaluations are tied to specific model versions that may be deprecated or silently updated, longitudinal comparisons become unreliable—precisely the reproducibility concern the framework is meant to address. Calibration and uncertainty quantification receive insufficient attention; in high-stakes domains, and equally in archival post-correction pipelines where errors propagate to downstream information extraction, the confidence of a quality judgement may be as important as the judgement itself. Cost-accuracy tradeoffs—employing a fine-tuned open-weight model as judge versus GPT-4, or single-judge versus multi-judge ensemble evaluation—are not systematically compared, an omission that is particularly consequential for resource-constrained digitisation projects at European national libraries. Finally, the survey's treat-

ment of multilingual and cross-cultural evaluation is brief: the biases of English-centric frontier LLMs when judging non-English content are noted but not empirically characterised, leaving open the practical reliability of LLM-based quality assessment for the multilingual archival collections—French, German, Italian—that constitute the Swiss newspaper corpus central to our research.

D. Article 4: Sun et al. (2025)

Full reference: Sun, L., He, L., Jia, S., He, Y., & You, C. [4]. *DocAgent: An Agentic Framework for Multi-Modal Long-Context Document Understanding.* In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025), pp. 17701–17716. ACL, 2025.

1) Summary: DocAgent addresses the challenge of understanding long, multi-modal documents (research papers, financial reports, guidebooks) by introducing a multi-agent framework inspired by human reading practices. The system comprises four components: a *Planner* that extracts a structured tree-formatted document outline and decomposes user queries; an *Actor* equipped with retrieval tools (search, `get_section_content`, `get_image`, `get_table_image`) to selectively access document content; a *Reviewer* that cross-checks actor responses using complementary evidence; and a *Memory* module that shares task-agnostic knowledge across multiple questions about the same document. Evaluated on MMLongBench-Doc and DocBench using GPT-4o, Claude 3.5 Sonnet, and Gemini 2.0 Flash, DocAgent surpasses all baselines by 9–12.4% on MMLongBench-Doc and reaches within 1.3% of human performance on DocBench, while accessing only ≈24% of document pages on average.

2) Strengths: The cognitively inspired architecture—outline-scanning followed by selective reading and verification—is well-motivated and elegantly solves the context window bottleneck that limits naive long-context approaches. The consistent improvements across three diverse base LLMs demonstrate that the framework's gains are model-agnostic rather than artefacts of a specific backbone. The latency analysis (10.6 s/query vs. 43–45 s for competing agent-based approaches) and the selective page access analysis (24% of pages) provide convincing evidence of efficiency. The reviewer agent's ability to cross-check text answers against tables and figures is particularly relevant for the multi-modal financial and scientific documents where errors are most consequential. The fine-grained analysis by evidence page count and position reveals that DocAgent is notably more robust than baselines as evidence is distributed across more pages—a key result for real-world archives.

3) Weaknesses: All experiments rely on commercial, closed-source LLMs (GPT-4o, Claude 3.5, Gemini 2.0); no open-weight alternatives are tested, raising cost and reproducibility concerns that are particularly acute for under-resourced archival institutions. The outline-based planner is the framework's most critical architectural assumption and its most significant limitation for archival research: it requires documents to expose a discernible hierarchical structure from which a tree-formatted outline can be extracted. This assumption is systematically violated by precisely the document types that most need intelligent understanding—historical newspapers with irregular multi-column layouts, manuscripts with marginal annotations, and archival dossiers that aggregate heterogeneous document types without consistent sectioning. Tran et al. [1] and Guan et al. [2] document that OCR errors and layout complexity compound across historical newspaper corpora; DocAgent's planner would receive a degraded, non-hierarchical input in exactly these cases, yet no robustness experiments under such conditions are reported. The evaluation further uses GPT-4o-based answer extraction to convert lengthy responses to concise answers while GPT-4o simultaneously serves as the primary backbone model; this circularity directionally inflates reported performance for the GPT-4o condition—where the judge and the system share the same generative priors—and may deflate performance for the Claude and Gemini conditions, meaning the benchmark comparisons across backbone models are not evaluated on a level playing field. The memory module is not ablated independently; it is always paired with the reviewer, making it impossible to isolate its individual contribution—and, consequently, impossible for practitioners to determine which component is responsible for the reported benchmark gains or to make principled decisions about which modules to retain when deploying the framework under latency or cost constraints.

E. Article 5: Lewis et al. (2020)

Full reference: Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. [5]. *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.* In Advances in Neural Information Processing Systems 33 (NeurIPS 2020), pp. 9459–9474. Curran Associates, Inc., 2020.

1) Summary: Lewis et al. introduce *Retrieval-Augmented Generation* (RAG), a framework that combines pre-trained *parametric memory* (a seq2seq language model) with *non-parametric memory* (a dense vector index over an external corpus). The retriever—a Dense Passage Retriever (DPR) based on a BERT dual-encoder—retrieves the top-k relevant passages for a given query using Maximum Inner Product Search over a pre-built FAISS index. The generator—BART-large—conditions its output on both the query and the concatenated retrieved passages. Retrieved documents are treated as latent variables and marginalised during joint end-to-end training, which is the key technical advance over prior pipeline-based retrieve-then-read architectures. Two formulations are proposed: *RAG-Sequence*, which uses the same retrieved passage to generate the entire output; and *RAG-Token*, which allows the generator to draw on different passages when generating each token, enabling multi-source synthesis. Evaluated on open-domain question answering (Natural Questions, TriviaQA), fact verification (FEVER), and knowledge-grounded generation tasks, RAG establishes new state-of-the-art results, outperforming both parametric-only (T5) and extractive retrieve-then-read pipelines (REALM, DPR-based reader).

2) Strengths: The foundational contribution of this paper is the joint training of the retriever's query encoder and the BART generator, with retrieved documents treated as latent variables marginalised during the forward pass. Crucially, the *document encoder is kept fixed*—the paper notes that periodic re-indexing of the 21M-passage Wikipedia corpus during training is computationally prohibitive—meaning only the query encoder and generator receive gradient updates. This engineering constraint leaves open the question of whether fully joint training would further improve retrieval-generator alignment. The non-parametric memory framing is nonetheless the paper's most durable conceptual contribution: unlike parametric memory (model weights), the retrieval index can be updated without any retraining—a decisive architectural advantage for applications where the knowledge base grows over time. The NeurIPS publication and 12,700+ citations confirm the paper's foundational status; virtually every subsequent RAG-based system, including those reviewed in Articles 1–4 of this exam, descends from this architectural framework.

3) Weaknesses: The RAG framework as introduced here rests on three assumptions that are violated in historical archival applications—assumptions whose real-world consequences become fully legible only after examining the domain-specific systems in Articles 1–4. First, it indexes English Wikipedia—a clean, well-structured, encyclopedic text source. The retrieval quality degradation caused by OCR-corrupted documents in the index is not examined; yet vocabulary mismatches between corrupted historical text and clean queries are precisely the conditions under which dense retrieval is expected to fail, as Tran et al. [1] subsequently document. Second, all experiments are monolingual (English); multilingual dense retrieval, required for any archive spanning French, German, Italian, or Finnish historical newspapers, is not addressed. Third, the knowledge base is a static Wikipedia snapshot; the engineering and retrieval quality implications of continuously growing archival collections are not discussed. Additionally, the training-time marginalisation over top-k documents requires k generator forward passes per training step—a computational cost that scales poorly with corpus size and is not systematically analysed.

III. Conclusion

The five papers reviewed in this document collectively illuminate the current state and key open challenges of LLM-based retrieval and document intelligence applied to historical archival corpora. Reading them in the order presented—from domain application back to architectural foundation—makes the research gap visible with full analytical force.

Tran et al. [1] establish the problem: applying RAG to multilingual, OCR-corrupted historical newspaper archives requires dense multilingual retrieval, hybrid NER-based reranking, and ground-truth-free evaluation metrics—but the re-

sulting pipeline remains partially evaluated, dependent on a commercial reranker, and unevaluated for cross-lingual corpora. Guan et al. [2] address the data scarcity that makes training RAG-ready corrected text expensive, offering a principled synthetic data pipeline that achieves strong CER reduction without manual annotation; their Optimal Alignment Curve, however, is calibrated on English 19th-century novels and provides no evidence of transferability to the multilingual, multi-century Swiss newspaper corpus. Gu et al. [3] provide the evaluation framework that underlies both preceding papers: the LLM-as-a-judge taxonomy and bias analysis is essential for understanding the reliability of the ground-truth-free metrics proposed by Tran et al., and the multilingual evaluation bias documented by Gu et al. is directly operative in a corpus spanning French, German, and Italian historical text judged by English-centric frontier models. Sun et al. [4] project the RAG paradigm forward into agentic document understanding, demonstrating that multi-agent frameworks with selective retrieval can approach human performance on long-context document benchmarks—yet their outline-based planner presupposes document structure that is by definition absent from historical newspapers with irregular multi-column layouts and OCR-degraded content.

Lewis et al. [5] provide the architectural foundation for all four preceding papers: the RAG framework's non-parametric memory enables knowledge updating without retraining—the decisive advantage for growing digitised archives—but its three baseline assumptions (clean text, monolingual content, static knowledge bases) define precisely the research gap that Articles 1–4 each, from their distinct vantage points, attempt to close. Reading Lewis et al. last reveals, retrospectively, that the entire literature reviewed here is an attempt to reconcile a foundational framework designed for encyclopedic English text with the reality of multilingual, historically degraded, structurally irregular archival corpora.

A structural theme running across all five papers is the cost and reproducibility burden imposed by closed-source, API-dependent LLMs: Lewis et al. index English Wikipedia with FAISS and train with BART; Tran et al. depend on the commercial Cohere reranker; Sun et al. use only GPT-4o, Claude, and Gemini; and Gu et al. document that LLM-as-judge evaluations are tied to proprietary model versions that may be deprecated. For digitisation-at-scale at national libraries and archival institutions—where processing millions of pages is the operative challenge—this cost constraint is not a minor omission but a central deployability problem. Our research on the BCUL Swiss newspaper archive directly confronts the intersection of all five limitation clusters simultaneously: multilingual OCR-corrupted text that violates Lewis et al.'s clean-text assumption; cross-lingual retrieval quality that is inadequately evaluated in Tran et al.'s pipeline; synthetic data calibration that cannot be imported from Guan et al.'s English-only OAC; LLM judge reliability that must be independently established for French, German, and Italian archival content per Gu et al.'s bias taxonomy; and agentic document understanding that cannot rely on Sun et al.'s hierarchical outline extraction over irregular newspaper layouts.

REFERENCES

[1] T. T. Tran, C.-E. González-Gallardo, and A. Doucet, "Retrieval augmented generation for historical newspapers," in *Proceedings of the 24th ACM/IEEE Joint Conference on Digital Libraries*, ser. JCDL '24. New York, NY, USA: Association for Computing Machinery, 2025. [Online]. Available: https://doi.org/10.1145/3677389.3702542

[2] S. Guan, C. Xu, M. Lin, and D. Greene, "Effective synthetic data and test-time adaptation for OCR correction," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 15 412–15 425. [Online]. Available: https://aclanthology.org/2024.emnlp-main.862/

[3] J. Gu, X. Jiang, Z. Shi, H. Tan, X. Zhai, C. Xu, W. Li, Y. Shen, S. Ma, H. Liu, S. Wang, K. Zhang, Z. Lin, B. Zhang, L. Ni, W. Gao, Y. Wang, and J. Guo, "A survey on llm-as-a-judge," *The Innovation*, p. 101253, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666675825004564

[4] L. Sun, L. He, S. Jia, Y. He, and C. You, "DocAgent: An agentic framework for multi-modal long-context document understanding," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng, Eds. Suzhou, China: Association for Computational Linguistics, Nov. 2025, pp. 17 701–17 716. [Online]. Available: https://aclanthology.org/2025.emnlp-main.893/

[5] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 9459–9474. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

[6] A. Ratner, C. D. Sa, S. Wu, D. Selsam, and C. Ré, "Data programming: creating large training sets, quickly," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, ser. NIPS'16. Red Hook, NY, USA: Curran Associates Inc., 2016, p. 3574–3582.

[7] I. Sastre, L. Etcheverry, G. Rey, G. Moncecchi, and A. Rosá, "Post-OCR Correction Using Large Language Models with Constrained Decoding," Jul. 2025. [Online]. Available: https://www.researchsquare.com/article/rs-6823036/v1