Upgrading Search Applications in the Era of LLMs: A Demonstration with Practical Lessons

Shuang Yu[1] , **Nirandika Wanigasekara**[1] , **Jeff Tan**[1] , **Kent Fitch**[2] and **Jaydeep Sen**[3]
[1]IBM Australia
[2]National Library of Australia
[3]IBM Research
{shuang.yu, Nirandika.Wanigasekara}@ibm.com, jeffetan@au1.ibm.com, kfitch@nla.gov.au,
jaydesen@in.ibm.com

Abstract

While traditional search systems have mostly been satisfactorily relying on lexical based sparse retrievers such as BM25, recent research advances in neural models, the current day large language models (LLMs) hold good promise for practical search applications as well. In this work, we discuss a collaboration between IBM and National Library of Australia to upgrade an existing search application (referred to as NLA) over terabytes of Australian Web Archive data and serving thousands of daily users. We posit and demonstrate both empirically and through qualitative user studies that LLMs and neural models can indeed provide good gains, when combined effectively with traditional search. We believe this demonstration will show the unique challenges associated with real world practical deployments and also offer valuable insights into how to effectively upgrade legacy search applications in the era of LLMs.

1 Introduction and Related Work

Classical retrievers are known to use sparse methods [Robertson and Zaragoza, 2009] with bag-of-words based lexical matching for retrieval. Popular approaches include BM25 [Stanford:BM25, 2009] used by Solr [Solr, 2010] for relevance scoring for accurate retrieval. Such methods have been seen to be performing well for a long time over many domains with low overhead of index management. Recent advances in language modelling research [Devlin *et al.*, 2019; Liu *et al.*, 2019] have ushered the retrieval community into a new age including use of LLMs [Zhao *et al.*, 2023; Touvron *et al.*, 2023] and neural models [sentence transformers, 2022; Karpukhin *et al.*, 2020; Khattab and Zaharia, 2020], for producing state-of-the-art retrievers. However, the key challenge for adopting neural retrievers is the need for training data across different domains and a relatively expensive index management and search mechanism over embedding vectors, often needing a dedicated engine known as a vector database [VectorDBs, 2010].

In this demonstration, IBM in collaboration with National Library of Australia, take up the task of of upgrading an existing search application (referred to as NLA) for the Australian Web Archive, which comprises more than 100 TB of data, 2 billion files indexed in the last year alone and supports thousands of users per month. Considering usage patterns from the users and noticeable drawbacks of NLA, we arrive at a well defined set of desired features for NLA as detailed in Section 2. We believe these features are not just useful for our application but are applicable in general to any search application to help users navigate web archives in an easy way. To measure the helpfulness of these features, we performed a qualitative user study which we detail in Section 4.

We propose to use modern advancements with LLMs [Touvron *et al.*, 2023] and neural retrievers [Khattab and Zaharia, 2020] to upgrade NLA which uses Solr [Solr, 2010], a sparse retriever. In this process, we also attempt to address the long standing question of sparse vs dense retrievers in practical applications. We detail our design choices and experiments in Section 3. We posit that, sparse retrievers indeed have their own advantages in terms of lightweight infrastructure and cost efficiency, and need not be given up completely. However, it is essential to design the search application in such a way that can leverage the advancements of neural models effectively to improve over sparse retriever results, thus combining the best of both worlds. We back up our position both empirically and through qualitative user studies.

We believe our demonstration will help the community learn about the practical challenges of a heavily used search application and will also provide relevant insights on designing and/or upgrading existing search applications with the help of modern day neural models.

2 NLA: Application Details

First we discuss the existing search application NLA which forms the backbone of our demonstration and our proposed upgrade. The National Library of Australia currently performs twice-yearly bulk harvests of the Australian web domain (.au), three-monthly bulk harvests of Australian Government websites, and in conjunction with state libraries, thousands of curated, targeted harvests including harvests of social media sites, many daily, starting from 1996. In the last year alone, over 160 TB in 2 billion files was collected. Harvested contents are stored in WARC (web archive) files [WARC, 2024], basic metadata is indexed in CDX format [CDX, 2024] and extracted text indexed using Solr. Searching the archive is done through the National Library of Australia's Trove[1] web site. The web archive search interface hosts about 40,000 sessions each month, and is popular with government personnel, journalists, researchers, and university students.

2.1 Challenges

Huge Volume of Data: The search corpus contains thousands of large data harvests performed since 1996, which are indexed via CDX and Solr.
Streaming Data: The index contains information crawled and stored over a span of 28 years, and hence has duplicate crawled responses for unchanged URLs.
User Base and Needs: There are on average 25,000 users and 40,000 search sessions each month, including government personnel, journalists, university students. While users can find related documents from keyword queries, they face difficulties with queries which are looking for some precise information or answers to specific questions.

2.2 Desired Feature Upgrades

Accurate Document Retrieval for Search Queries: NLA needs to be upgraded to support search queries for accurate retrieval while retaining keyword search capability.
Support for Answering/Summarizing the results: NLA should also provide a textual summary of retrieved results to help users understand the results better.
Assistance via follow-up Recommendations: NLA should guide users to search for related information by providing follow-up queries.

3 System

Figure 1 shows the two variants of system upgrades we attempted over NLA. We describe their key architectural features as follows.

3.1 Reranked-NLA

In this variant, we benefit from the lightweight indexing of Solr search by reusing it as a 1st stage retriever but thoughtfully introduce LLMs and neural models to improve NLA. The key upgrades over NLA are enumerated as follows.

LLM for Keyword Extraction. Instead of using the raw user query with Solr, we use the LLM Llama2 [Touvron *et al.*, 2023] with prompts to extract important keywords from the user query. This ensures that the documents retrieved by Solr are focused towards the extracted entities from the user query and thus Solr acts as an effective metadata filter over the huge search corpus.

Neural Model as Reranker. We use ColBERT [Santhanam *et al.*, 2022] to rerank the documents retrieved by Solr in the 1st stage. ColBERT is a state-of-the-art neural retriever architecture, which relies on contextual token embedding and aggregates similarity scores between each query token and document token to evaluate the similarity of a document given a query. Using ColBERT as a reranker over Solr output bridges the semantic gap between the search intent of the user query and target documents, by reranking the appropriate documents higher. Also note that we do not need a separate indexer for ColBERT, which we use only as a reranker over retrieved documents. So Reranked-NLA still remains quite lightweight using only inverted index from Solr.

LLM as Summarizer and Follow-up Query Generator. We use Llama2 again to summarize the retrieved results back to the user and also to generate follow-up query suggestions. Note that we only provide task specific prompts to Llama2 without any finetuning. We experimented with different prompts for the LLMs to identify the optimal choice of prompts. Using Llama2 out-of-the-box without incurring any cost to fine-tune it makes Reranked-NLA cost effective too.

3.2 Neural-NLA

In designing Neural-NLA, we adopt an end-to-end neural retrieval architecture. We use all-MiniLM-L6-v2 [sentence transformers, 2022] to create embedding vectors for documents and use Milvus [Milvus, 2023] as a dedicated Vector DB to handle the huge scale of document embedding vectors needed in NLA. Given a user query, Neural-NLA uses all-MiniLM-L6-v2 to compute the query embedding vector and retrieves the top 50 similar documents searched by Milvus over the document index. The retrieved documents are then used for generating summaries and follow-up questions using Llama2, similar to Reranked-NLA. It is important to note that Neural-NLA uses all-MiniLM-L6-v2 out of the box, therefore it is prone to suffer from domain shift and vocabulary mismatch. This is specifically important where NLA is dealing with a corpus which has lot of domain specific entities.

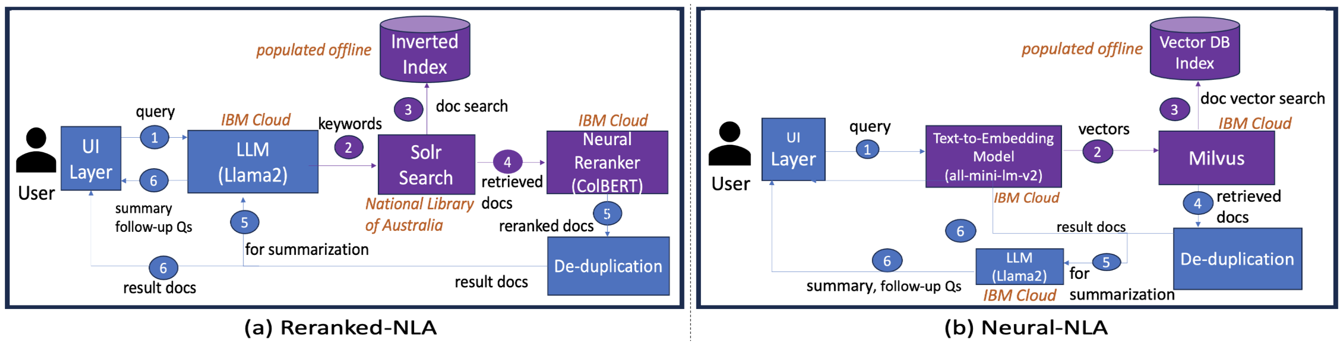
3.3 De-duplication of Results

A common step used for both Reranked-NLA and Neural-NLA is to de-duplicate the retrieved documents to remove documents with the same content indexed multiple times during periodic crawls over time. For Reranked-NLA, comparing ColBERT scores up to 4 decimal points is used to deduplicate reranked documents. For Neural-NLA the score returned from Milvus is used for de-duplication. We review the related results in the next section.

4 Results and Key Takeaways

We evaluate the system designs we proposed in Section 3 both empirically and qualitatively via user study metrics. For empirical evaluation we use popular information retrieval metrics such as precision@10, recall@10, Mean Reciprocal Rank (MRR) [Rank, 2008] and more robust metric Normalized Discounted Cumulative Gain (NDCG) [Wang *et al.*, 2013]. While recall is the basic metric for retrieval accuracy, MRR and NDCG are more practical to measure the accuracy of the deployed retriever, as they consider the rank of the correct document in the retrieved list. We evaluated on a pilot set of 49 queries provided by the National Library of Australia team where NLA was performing poorly and needed enhancement.

[1] https://trove.nla.gov.au/



(a) Reranked-NLA

(b) Neural-NLA

Figure 1: Comparison of two system architectures. Similarities are marked in Blue, Key differences are colored in Purple

4.1 Empirical Results

As we see in Table 1, Reranked-NLA improves over NLA for all the metrics. Better precision value shows the results from Reranked-NLA is more relevant for the user query than NLA and Neural-NLA. Although there is some gap in recall numbers between Reranked-NLA and Neural-NLA, we emphasize that MRR and NDCG are more realistic metrics for user-facing deployed systems, as they consider the rank of the retrieved documents. Reranked-NLA comes out as a better choice than Neural-NLA with comparable NDCG numbers but hugely improved MRR numbers. A better MRR value ensures the correct document is ranked as high as possible to help the user find it quickly.

Metric	NLA	Reranked-NLA	Neural-NLA
Precision@10	0.27	0.32	0.24
Recall@10	0.33	0.35	0.57
MRR	0.38	0.54	0.35
NDCG@10	0.43	0.86	0.88

Table 1: Evaluation Results

4.2 Qualitative Metrics: User Studies

We also performed user studies to qualitatively compare both approaches. We engaged 14+ people to assess how Reranked-NLA and Neural-NLA performed against the Baseline NLA on the helpfulness of (1) result summaries and (2) follow-up queries, where the user rated the systems in likert scale [likert, 2020] i.e. 0:not helpful, 5:very helpful.

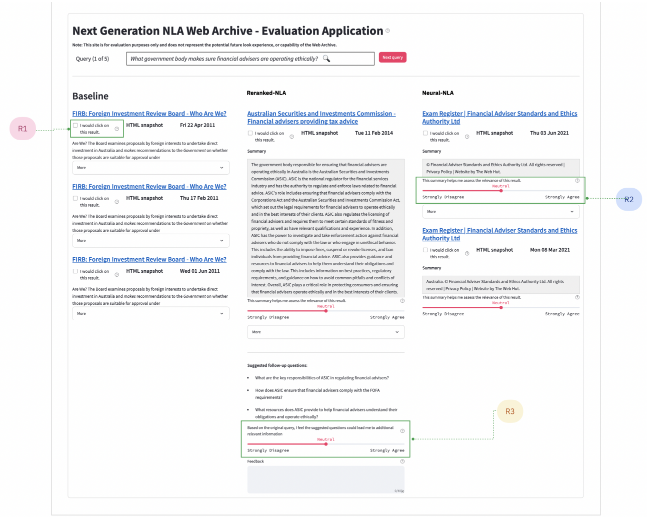
Result summaries	Reranked-NLA	Neural-NLA
Summary Quality	3.5	2.4
Follow-up questions Relevancy	3.4	3.1

Table 2: User Study Results

Table 2 summarizes the results of the user study. Users rate Reranked-NLA better than Neural-NLA for both the tasks of generating summaries from the results and generating follow-up questions. It is important to note here, both of these tasks have an implicit dependency on the accuracy of the retrieved documents. Therefore, better user study scores not only ensures the quality of the associated user-facing tasks such as summarization or follow-up query generation, but also confirms better retrieval results with Reranked-NLA.

5 Demonstration

The demonstration will include a comparative display of the baseline NLA along with our two proposed variations Reranked-NLA and Neural-NLA as shown in Figure 2. The screen would not only include questions and their responses from 3 systems but also enable users to provide feedback for the system responses. Thus an end user would be able to mimic the journey of the evaluators who actually evaluated the systems by their comparative performance.



Next Generation NLA Web Archive - Evaluation Application

Note: This site is for evaluation purposes only and does not represent the potential future look experience, or capability of the Web Archive.

Query (1 of 5) *What government body makes sure financial advisers are operating ethically?*

Baseline

FIRB: Foreign Investment Review Board - Who Are We?

□ I would click on this result. HTML snapshot Fri 22 Apr 2011

Are We? The Board examines proposals by foreign interests to undertake direct investment in Australia and *makes* recommendations to the *Government* on whether those proposals are suitable for approval under

FIRB: Foreign Investment Review Board - Who Are We?

□ I would click on this result. HTML snapshot Thu 17 Feb 2011

Are We? The Board examines proposals by foreign interests to undertake direct investment in Australia and *makes* recommendations to the *Government* on whether those proposals are suitable for approval under

FIRB: Foreign Investment Review Board - Who Are We?

□ I would click on this result. HTML snapshot Wed 01 Jun 2011

Are We? The Board examines proposals by foreign interests to undertake direct investment in Australia and *makes* recommendations to the *Government* on whether those proposals are suitable for approval under

Reranked-NLA

Australian Securities and Investments Commission - Financial advisers providing tax advice

□ I would click on this result. HTML snapshot Tue 11 Feb 2014

Summary

The government body responsible for ensuring that financial advisers are operating ethically in Australia is the Australian Securities and Investments Commission (ASIC). ASIC is the national regulator for the financial services industry and has the authority to regulate and enforce laws related to financial advice. ASIC's role includes ensuring that financial advisers comply with the Corporations Act and the Australian Securities and Investments Commission Act, which set out the legal requirements for financial advisers to operate ethically and in the best interests of their clients. ASIC also regulates the licensing of financial advisers and requires them to meet certain standards of fitness and propriety, as well as have relevant qualifications and experience. In addition, ASIC has the power to investigate and take enforcement action against financial advisers who do not comply with the law or who engage in unethical behavior. This includes the ability to impose fines, suspend or revoke licenses, and ban individuals from providing financial advice. ASIC also provides guidance and resources to financial advisers to help them understand their obligations and comply with the law. This includes information on best practices, regulatory requirements, and guidance on how to avoid common pitfalls and conflicts of interest. Overall, ASIC plays a critical role in protecting consumers and ensuring that financial advisers operate ethically and in the best interests of their clients.

This summary helps me assess the relevance of this result.

Neutral

Strongly Disagree Strongly Agree

Suggested follow-up questions:

- What are the key responsibilities of ASIC in regulating financial advisers?
- How does ASIC ensure that financial advisers comply with the FOFA requirements?
- What resources does ASIC provide to help financial advisers understand their obligations and operate ethically?

Based on the original query, I feel the suggested questions could lead me to additional relevant information

Neutral

Strongly Disagree Strongly Agree

0/400

Neural-NLA

Exam Register | Financial Adviser Standards and Ethics Authority Ltd

□ I would click on this result. HTML snapshot Thu 03 Jun 2021

Summary

© Financial Adviser Standards and Ethics Authority Ltd. All rights reserved | Privacy Policy | Website by The Web Hut.

This summary helps me assess the relevance of this result.

Neutral

Strongly Disagree Strongly Agree

Exam Register | Financial Adviser Standards and Ethics Authority Ltd

□ I would click on this result. HTML snapshot Mon 08 Mar 2021

Summary

Australia. © Financial Adviser Standards and Ethics Authority Ltd. All rights reserved | Privacy Policy | Website by The Web Hut.

This summary helps me assess the relevance of this result.

Neutral

Strongly Disagree Strongly Agree

Figure 2: Evaluation Screen. R1 - Check box to assess the relevancy of the result , R2- Likert scale for helpfulness of result summaries , R3- Likert score for helpfulness of follow-up questions.

We will also have a dedicated screen for the best performing system i.e. Reranked-NLA, where an end user can interact with the system to ask questions and get responses back. The user would also be able to see the generated summaries and might take up follow-up query suggestions from Reranked-NLA. Like the comparative screen, the dedicated screen will also have user feedback buttons so that the user can meaningfully engage with the system and provide real-time feedback.

Acknowledgments

The authors acknowledge the foresight of National Library of Australia and their Pandora partners for building and maintaining the archive since 1996.

Thanks to the National Library of Australia staff involved in the evaluation and special thanks to Alex Osborne for technical assistance and advice. The authors acknowledge the support of David Wong (National Library of Australia CIO) and Terence Ingram (National Library of Australia Project Manager) in facilitating this work.

Thanks to Drew Johnson, Claire Dunworth, Nick Merrett, Thanh Nguyen and Alice Tai from IBM Australia for facilitating the implementations.

References

[CDX, 2024] CDX. Cdxformat. https://archive.org/web/researcher/cdx_file_format.php, 2024.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[Karpukhin *et al.*, 2020] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. Dense passage retrieval for open-domain question answering, 2020.

[Khattab and Zaharia, 2020] Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 39–48, New York, NY, USA, 2020. Association for Computing Machinery.

[likert, 2020] likert. Likert scale. https://en.wikipedia.org/wiki/Likert_scale, 2020.

[Liu *et al.*, 2019] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692, 2019.

[Milvus, 2023] Milvus. Milvus vector db. https://milvus.io/, 2023.

[Rank, 2008] Wiki:Mean Reciprocal Rank. Mean reciprocal rank. https://en.wikipedia.org/wiki/Mean_reciprocal_rank, 2008.

[Robertson and Zaragoza, 2009] Stephen E. Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389, 2009.

[Santhanam *et al.*, 2022] Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3715–3734, Seattle, United States, July 2022. Association for Computational Linguistics.

[sentence transformers, 2022] sentence transformers. all-mini-lm-l6-v2. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2/tree/main, 2022.

[Solr, 2010] Solr. Solr search engine. https://solr.apache.org/, 2010.

[Stanford:BM25, 2009] Stanford:BM25. Bm25 retrieval. https://nlp.stanford.edu/IR-book/html/htmledition/okapi-bm25-a-non-binary-model-1.html, 2009.

[Touvron *et al.*, 2023] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv e-prints*, page arXiv:2307.09288, July 2023.

[VectorDBs, 2010] Summary: VectorDBs. Vector databases. https://www.elastic.co/what-is/vector-database, 2010.

[Wang *et al.*, 2013] Yining Wang, Liwei Wang, Yuanzhi Li, Di He, Tie-Yan Liu, and Wei Chen. A theoretical analysis of ndcg type ranking measures. *Journal of Machine Learning Research*, 30, 04 2013.

[WARC, 2024] WARC. Warc format. https://www.loc.gov/preservation/digital/formats/fdd/fdd000236.shtml, 2024.

[Zhao *et al.*, 2023] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Z. Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jianyun Nie, and Ji rong Wen. A survey of large language models. *ArXiv*, abs/2303.18223, 2023.