Cambridge Forum on AI: Culture and Society (2026), **2**, e17, 1–13
doi:10.1017/cfc.2026.10029

Cambridge Forum

REFLECTION

ATLAS: Harnessing retrieval-augmented generation (RAG)

James Smithies[1], Glen Berman[1], Karaitiana Taiuru[2], Martin Spychal[3] and John Moore[4]

[1]The Australian National University, HASS Digital Research Hub, Canberra, Australia; [2]Taiuru & Associates Ltd, Wellington, New Zealand; [3]The History of Parliament Project, London, UK and [4]The National Archives, Kew, London, UK
Corresponding author: James Smithies; Email: james.smithies@pm.me

(Received 28 February 2025 UTC; revised 26 March 2026 UTC; accepted 5 May 2026 UTC)

Abstract
This paper describes a prototype test harness that has been developed to evaluate the use of large language model retrieval-augmented generation systems for humanities and social science research, with an emphasis on collections held in the galleries, libraries, archives and museum sector. We propose the development of new modes of research software engineering oriented towards the values and principles of those communities, and rigorous modes of software testing grounded in industry best practices.

Keywords: artificial intelligence; GLAM; HASS; large language models (LLMs); research software test engineering

1. Introduction

The cultural heritage datasets maintained by GLAM (galleries, libraries, archives and museums) institutions and studied by humanities, arts and social sciences (HASS) scholars are logical application domains for large language models (LLMs). Indeed, given the facility LLMs have demonstrated at summarising large volumes of textual data, the increasing sophistication of retrieval-augmented generation (RAG), which provides an LLM with a defined set of documents to "ground" their output with expert knowledge (Callaghan and Vieira 2025), it is unsurprising that LLM and RAG technologies are being integrated into GLAM institutions' digital services (Finn and Khosrowi 2026). Little work has been undertaken to assess the alignment of research tools powered by LLMs with the standards and values of HASS scholars and GLAM institutions (Bode and Goodlad 2023; Raley and Rhee 2023), however. This is particularly the case with products enabled by RAG. As with commercial applications, RAG systems have significant potential for research and GLAM workflows, extending from summarisation of corpora to the provision of expert chat interfaces, but their incorporation into HASS research and GLAM institutions is complex. From a practical viewpoint, RAG systems demand significant computational resources and introduce additional layers of complexity into digital research infrastructure (RI). More substantively, however, the incorporation of RAG systems into research workflows and methods alters them, introducing new epistemologies, new variables, new potential sources of bias and new potentials for failures in reproducibility and replicability. Digital humanities (DH), which sits at the boundary of humanities inquiry and computational practice and has used cultural heritage datasets intensively for decades, is well placed to address these substantive risks.

For DH scholars, however, critical engagement with LLM technologies presents difficulties. Commercial LLM services are black-boxed, foreclosing possibilities of critical analysis of their components. LLMs, even when released as open-source models (Solaiman 2023), produce

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike licence (http://creativecommons.org/licenses/by-nc-sa/4.0), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the same Creative Commons licence is used to distribute the re-used or adapted article and the original article is properly cited. The written permission of Cambridge University Press or the rights holder(s) must be obtained prior to any commercial use.

https://doi.org/10.1017/cfc.2026.10029 Published online by Cambridge University Press

inscrutable model outputs and are nondeterministic. Evaluation frameworks designed in industry or Science, Technology, Engineering, Mathematics (STEM) settings do not reflect DH's understanding of computational technologies as inherently sociotechnical. This paper uses critical technical practice (Agre 1997) to offer a path forward for DH scholars. In explicating a process of building and experimenting with LLM and RAG technologies, the paper demonstrates one approach DH scholars may apply to critical engagements with LLM technologies. In particular, in this paper, we report on the development of ATLAS (Analysis and Testing of Language Models for Archival Systems), which is a prototype evaluation harness developed initially as an exercise in critical technical practice, but now open-sourced as part of a broader effort to support robust and responsible use of RAG systems in research involving GLAM datasets. While from the perspective of state-of-the-art RAG systems, ATLAS is not novel (i.e. it uses industry standard RAG techniques), for DH scholars, the development and configuration of ATLAS as an open-source and highly adaptable harness for evaluating LLM and RAG components represent a substantive technical contribution. Conceptually, the paper's demonstration of critical technical practice through the use of AI coding tools and open-source AI frameworks provides DH scholars and the broader GLAM community with a high-level approach to experimentation with LLM technologies.

The paper proceeds as follows. We begin by situating our approach within the current landscape of LLM evaluation, arguing that existing frameworks fail to account for the priorities of DH and GLAM institutions. We then outline critical technical practice as a framework for DH engagement with LLM technologies. This lays the foundation for our exposition of ATLAS, in which we detail the configurable components of its RAG pipeline and the evaluation infrastructure that supports transparent, reproducible experimentation. We conclude by reflecting on the implications of ATLAS's development for DH scholars and the broader GLAM sector.

2. Situating the methodology

ATLAS is a product of the AI as Infrastructure (AIINFRA) project, which is implementing methodological baselines and tools to support the evaluation of RAG systems in a way that opens them up for empirical as well as epistemological analysis by HASS researchers. This is pressing work being undertaken in a fast-moving technological and methodological environment that is not always well aligned to HASS and GLAM priorities. Existing mainstream approaches to the evaluation of LLMs largely focus on leaderboard tables, by which the performance of different LLMs is tracked against benchmark tests. These are designed by industry and STEM disciplines to their standards, assess foundation models rather than RAG architectures (see below) and do not account for the semantics of historical source documents or HASS and GLAM sector values or scholarly requirements (Lee et al. 2023; Li et al. 2023; Lin et al. 2022). As the International Panel on the Information Environment (IPIE 2025) noted in early 2025, issues with labour laws, violations of human rights and the rights of Indigenous modes of knowledge and scholarship are never noted in such frameworks. Environmental impact is often considered but political and gender bias, and racism, are not adequately confronted (IPIE 2025, 11).

This set of problems speaks to a wider issue. A benchmark specifically designed to test historical knowledge has concluded that "while the overall performance of the LLMs on expert historical knowledge is better than random guessing, it falls short of comprehensive expert-level knowledge" (Hauser et al. 2024). Those tests were conducted by comparing the output of LLMs against the *Sheshat Global History Databank*, a broad-ranging human-curated collection of historical data. While these results are important, not only for demonstrating the limitations of current LLMs for HASS research but for establishing a baseline against which future versions can be tested, they are not reflective of the ways in which LLMs are likely to be deployed in GLAM or HASS settings. Like the commercial leaderboards they are derived from, they were not designed to test the performance of LLM

https://doi.org/10.1017/cfc.2026.10029 Published online by Cambridge University Press

RAG systems or to evaluate how LLMs and LLM RAG systems might be evaluated in terms of broad sociotechnical metrics. This more holistic mode of evaluation is important because it reflects modes of deployment likely to occur across the HASS and GLAM sectors, as AI becomes more tightly integrated into RI. DH scholars need to gain a sense of the baseline accuracy of LLMs, but there are many other issues to consider as well, ranging from basic issues of scientific best practice (transparency, explainability, reproducibility) to environmental impact, cultural safety and deployment cost.

STEM and computational linguistics research communities are aware of the importance of testing and have been developing tools and methods for many years, noting that "[w]ith the rapid development and widespread use of LLMs, the importance of evaluating them in practical applications and research has become crucial" (Chang et al. 2024, 39:31). The implications of this statement are more profound than might be supposed. The commercialisation of LLMs has *depended* on the interaction between evaluation, testing and subsequent technical development, initially through empirical (often automated) benchmarking but increasingly through human-in-the-loop evaluation that allows for "greater human feedback during the evaluation process" (Chang et al. 2024, 39:31). The ability of DH scholars to design and develop LLM research tools is heavily dependent upon their capacity to engage in wide-ranging software test engineering.

Yet engagement with software test engineering alone is insufficient. The opacity of AI technologies demands not just technical competence but critical reflection on the systems being built and evaluated. Testing has always been integral to software development, but the opacity of AI technology heightens the stakes. As is well known from landmark publications such as Emily M. Bender et al.'s (2021) "stochastic parrots" article "… an LM [language model] is a system for haphazardly stitching together sequences of linguistic forms it has observed in its vast training data, according to probabilistic information about how they combine, but without any reference to meaning …" (Bender et al. 2021, 617). This results in the well-known tendency towards "hallucination" of facts and citations (Xu et al. 2024), and the deep opacity of AI systems that lend them the appearance of classic "black box" technologies (impenetrable to meaningful interpretation) (Pasquale 2015). Opacity is increased again through the commercialisation of LLMs into products and apps that actively seek to hide complexity from end users, in the interests of protecting intellectual property (IP) rights and improving user experience. As David Berry has pointed out, it is important not to over-compensate for these issues and assume meaningful knowledge of AI products is impossible, but commercial products make robust evaluation of LLMs difficult (Berry 2023). Evaluation and testing can never provide full transparency, but they can ameliorate the opacity of LLM-based products to a significant degree and enable informed decisions regarding LLM use.

This is of huge importance in epistemological as well as political and sociotechnical terms. If DH scholars can harness the power of language models, gaining leverage over their tendency to reflect biases from their training data in their output, they will be better positioned to integrate them into research workflows, contribute to their ethical design, development and regulation and support the broader GLAM community with its engagement with LLM technologies. Our contribution to this space is to implement a prototype test harness in the spirit of "critical technical practice" presented by Philip Agre in 1998, introducing technical methods that might have some lasting value in the development of responsible AI for HASS and GLAM (Agre 1997). Our prototyping method is necessarily reliant on established engineering methods but is informed by a commitment to the values and principles of the HASS and GLAM communities, incorporating the spirit of responsible and slow AI (Huggett 2024; McLennan et al. 2020). If we, as DH scholars, can create models, workflows and prototypes that support thinking through the realities of working with and against AI technologies, we stand a much better chance of either developing production-grade versions of those technologies or (at least) understanding them well enough to build mutually beneficial relationships with commercial vendors.

https://doi.org/10.1017/cfc.2026.10029 Published online by Cambridge University Press

3. Architecture

Having outlined the case for critical technical engagement with LLM technologies, this section describes the design choices that have shaped ATLAS's development and details its architecture.[1] ATLAS is a prototype developed using the Python frameworks FastAPI and LangChain, the frontend framework Vue and the vector store Chroma DB (Figure 2). The key-value store Redis is used to enable feedback telemetry and manage the message queue in production. AWS Cognito is used for user authentication. Telemetry is enabled using the Open Telemetry schema for delivery to the Phoenix Arize cloud platform. Heavy use has been made of AI coding tools, including GitHub Copilot, Cursor, Windsurf and Claude Code. The sole developer (Smithies) is an amateur programmer, augmented by insights gained from two decades working with software engineering teams in the commercial, government and academic sectors. This process of coding with AI in a research context is outside the scope of this paper but is an important topic for further investigation, with significant implications for HASS research.

The ATLAS harness enables reproducible and transparent experiments with relatively large corpora and multiple models. The system uses a mixture of deterministic and probabilistic technologies and can be configured to use any number of LLMs, including commercial (OpenAI, Anthropic, Google) and more open models (Llama, Mistral, DeepSeek). The user interacts with a standard question and answer (Q&A) chatbot interface (Figure 3). Although the choice of technologies stemmed from a preference for Python-based tools, it was strongly influenced by the state of technology when development started in mid-2024. At that point in time, there were few low and no-code solutions, and desktop and commercial products had limited functionality and worse transparency. The LangChain framework was the leading option but was under rapid development and criticised by experienced programmers for being unnecessarily abstract (Safjan 2023). Rival frameworks were even less mature. At the time of writing, version 0.2.1 of the ATLAS prototype is about to be released and the tool has been used in multiple production-level contexts.

The hope is that ATLAS can continue to develop alongside rapid changes in the technical environment, but it is envisaged as an ephemeral prototype. The sustainability of using complex development tools, designed for enterprise use cases, is questionable, and it is unlikely that the project has the broad appeal needed to develop a technical community. The primary benefit of the chosen technology stack is its ability to provide a view "behind the veil" of LLM RAG systems, exposing the components that influence the system output (Figure 1) and modelling possible approaches to LLM RAG testing, rather than as a long-term solution suitable for widespread production use.

RAG system architectures are largely deterministic (traditional web applications), except for the foundation model (LLM) that provides the "brains" of the system, and include multiple configurable elements. The responses from the LLM are guided by system prompts that define the role the LLM should adopt in its answer. This constrains but does not override the different "personalities" of the different LLMs (Serapio-García et al. 2023). The primary system prompt for ATLAS is configurable and can make a significant difference to the output. At the time of writing, its default prompt is 66 lines long, divided into sections defining its role, corpus selection, citation guidelines, evidence handling, uncertainty handling and edge cases. Its primary "identity" is defined in the following way:

> You are an expert historical research assistant specializing in 1901 Hansard parliamentary records from Australia, New Zealand, and the United Kingdom. Your expertise is limited to these historical records and their context, but you can make relevant comparisons to contemporary issues when appropriate. Present your findings in a clear, authoritative manner …. (Smithies and Berman 2025b)

[1] ATLAS is an open-source project. The codebase can be found on GitHub at https://github.com/AI-as-Infrastructure/aiinfra-atlas/releases/tag/v0.2.0 (Smithies and Berman 2025b). Example implementations, example telemetry data and further documentation can be found on Zenodo at https://zenodo.org/communities/aiinfra/.

https://doi.org/10.1017/cfc.2026.10029 Published online by Cambridge University Press

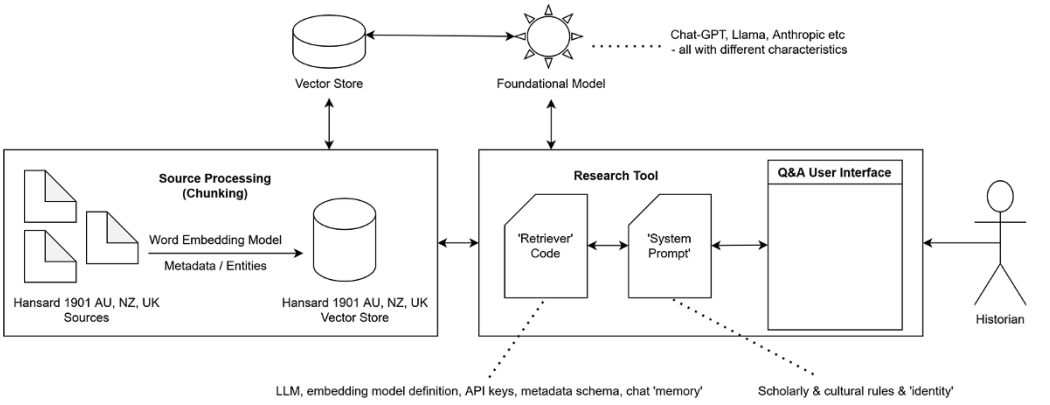


Figure 1. ATLAS large language model (LLM) retrieval-augmented generation (RAG) architecture.

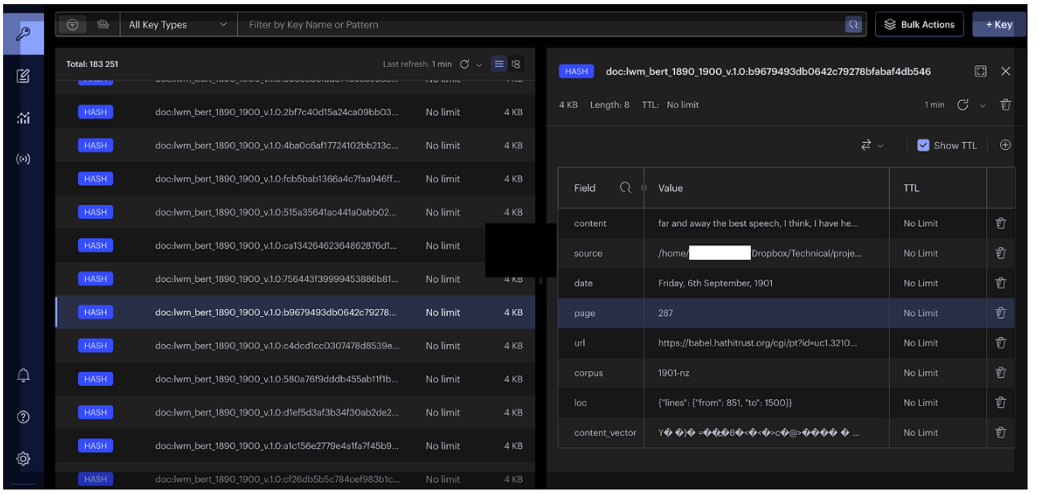


Figure 2. ATLAS vector store database.

Another system prompt is needed to guide multi-turn Q&A, instructing the LLM:

> Given a chat history and the latest user question which might reference context in the chat history, formulate a standalone question which can be understood without the chat history. Do NOT answer the question, just reformulate it if needed and otherwise return it as is.[2]

Using the system prompts, the LLM uses a "retriever" module, written using LangChain, to search across a vector database containing the transnational corpus of Hansard documents, stripped of their XML in the initial version to create a straightforward text corpus and divided into ~95,000 "chunks." Significantly, the open architecture of ATLAS allows for the implementation of multiple different retrievers, optimised for different vector stores and emerging forms of Model Context Protocol servers that can provide accessible programmatic access to data sources of various kinds.

One of the benefits of using open-source components, therefore, is the ability to create multiple vector store databases using the baseline Hansard corpus. As the primary data source for the system,

[2]This phrasing was taken from LangChain documentation. See https://python.langchain.com/v0.1/docs/use_cases/question_answering/chat_history/.

https://doi.org/10.1017/cfc.2026.10029 Published online by Cambridge University Press

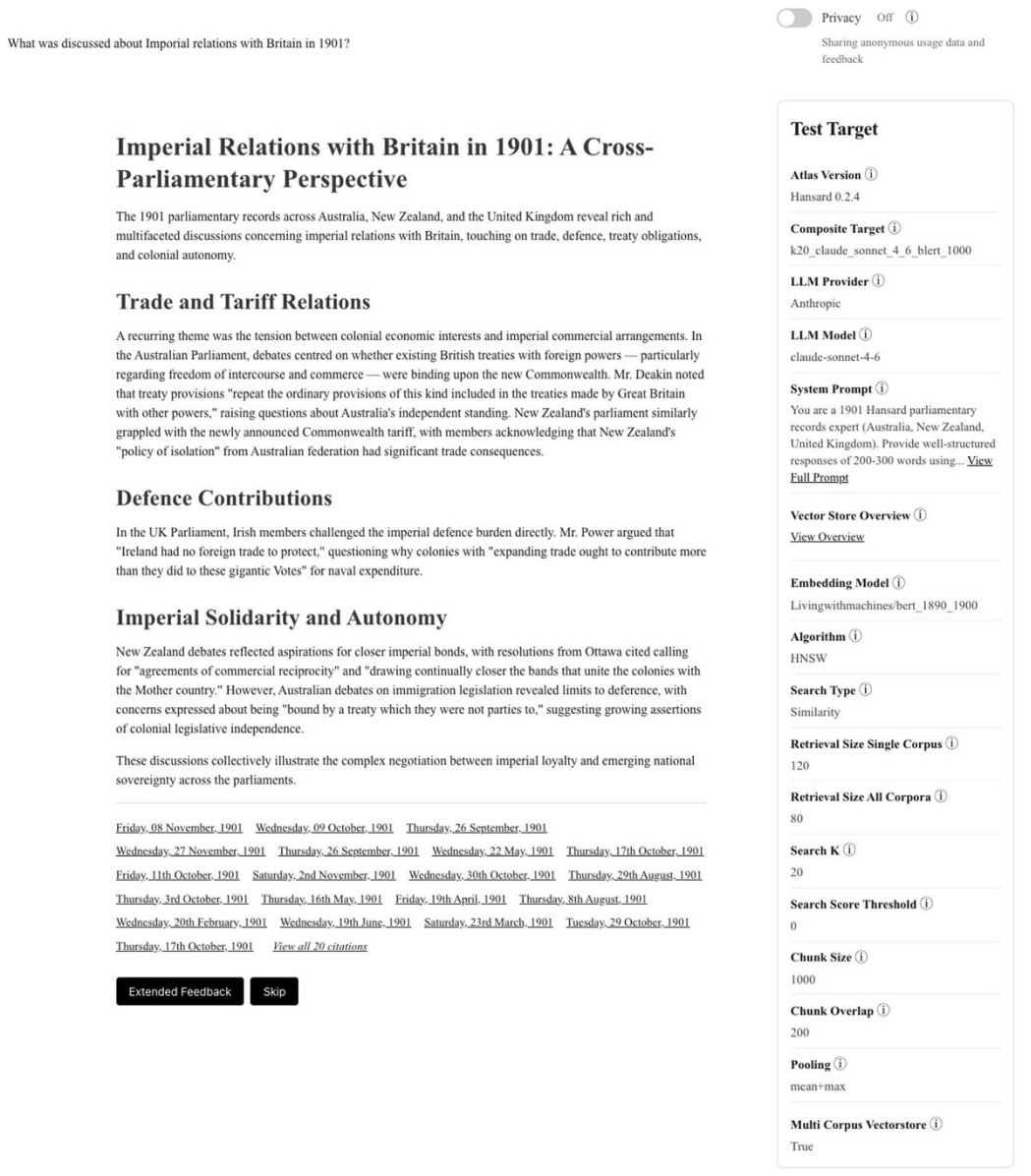


Figure 3. ATLAS user interface.

the vector store has a crucial effect on the eventual output, implying the need for testing different approaches to vector store design. This is especially the case with the transformation of the document chunks into word embeddings that are stored in the database, a process that is needed because of the limited "context window" of LLMs. The LLM will only retrieve a limited number of chunks that match the user question, grounding its answers in them. Different document "chunking" sizes can therefore make a significant difference to the answers given by the LLM, depending on whether the chosen chunks support robust inferencing. The baseline ATLAS vector store adopts the widely accepted approach of chunking the Hansard debates into tokens of 1000, with an overlap of 200, resulting in ~95,000 chunks with associated metadata. The number of chunks returned by the LLM

https://doi.org/10.1017/cfc.2026.10029 Published online by Cambridge University Press

(known as the "search-k" or "fetch-k" number) can also be set, with a range between 10 and 100 producing reasonable results depending on the LLM used. A "temperature" setting influences the shape of probability distribution through which an LLM generates a response, with a setting of 0 being the least random (or, most deterministic). Future work will assess whether larger or smaller chunks improve performance, and whether the ever-increasing context windows of LLMs will allow the analysis of HASS and GLAM-sized corpora without the use of RAG architectures.[3]

This does not seem likely in the short term, although the pace of development is rapid. It is important to understand the trade-off with using RAG systems for the analysis of corpora, regardless. The fetch-k setting drastically reduces the amount of information the LLM has available to analyse, reducing the information space and evidentiary context of the resulting answer, but also drastically reducing the computational, financial and environmental cost of analysis. In the case of the default transnational Hansard corpus used in ATLAS, costs are reduced from over $USD1000 *per query* (depending on the model used) to a matter of cents.[4] The technique for achieving this relies on a radical reduction in the information space available to the LLM to analyse. Rather than analysing all 11,967,262 words (70,581,735 tokens) in the corpus, which ChatGPT-4o would need to do in 579 batches to fit the data into its 128,000 token context window at the time of writing, only the document *chunks* defined by the fetch-K variable are analysed. This monetary reality corresponds to environmental impact in important ways, with the cost of each query reflecting the computational intensity of the related processing.

The chunks are chosen using the Hierarchical Navigable Small World algorithm that selects chunks based on their semantic similarity to the query (making the choice of word embedding significant), then re-ranked using a standard process. This reduces the 95,000 chunks to 500 and then 30 that are analysed by the LLM, depending on the test target configuration. As with all LLM RAG systems, the information space is so reduced that a degree of serendipity is introduced, requiring careful attention to the response, but the underlying architecture guarantees a relationship between query, response and citations. The precise sociotechnical and epistemological nature of this relationship is a question of some complexity, especially given the probabilistic nature of the LLM responsible for analysing the document chunks, but parallels could perhaps be found in the analogue world of browsing library shelves and card catalogues. The point is not to assume the epistemological relationship between researcher and tool is stable or known (and certainly not to treat LLM RAG systems as oracles), but to problematise such systems in terms of critical technical practice and evaluate their potential role in the discovery and research process on that basis.

Aside from gaining the ability to extract and attach metadata to document chunks (potentially including concepts, entities and sentiments alongside basic information about the Hansard session, country of origin and so on), the primary advantage of having complete control over the vector store is the ability to use any number of different word embedding models in the development process, ensuring the semantic relationships represented in the vector space of the database correspond to *historical*

[3] At the time of writing, even Google's NotebookLM has a context window (1.5–2 million tokens depending on the model), an order of magnitude too small to fit the Hansard corpus, even though the corpus only stores debates from a single year across the three countries. See Google.com. "Long Context | Gemini API." Google AI for Developers. Accessed 21 February 2025. https://ai.google.dev/gemini-api/docs/long-context.

[4] Cost estimation for full-corpus query vs. RAG retrieval using GPT-4o, 128k context. ~Tokens per chunk: 500 tokens; ~total chunks: 183,000; ~total tokens needed for context window: 183,000 × 500 = 91,500,000 tokens. A full-corpus query using GPT-4o (128k context window) requires processing 91,500,000 tokens as input, with an estimated output of 9,150,000 tokens (assuming a summary with citations is 10% of the input size). Based on OpenAI's December 2024 pricing: input cost: (91,500,000 ÷ 1,000,000, because OpenAI prices per 1 million tokens) × $2.50 (the cost of processing 1 million tokens) = $228.75; output cost: (91,500,000 ÷ 1,000,000) × $10.00 = $915.00; total cost per query: $1143.75. However, using a RAG approach, only a subset of relevant chunks is retrieved rather than processing the entire corpus. For instance, if a fetch-K value retrieves 20 document chunks of 500 tokens each (10,000 tokens per query), the input cost drops significantly: RAG-based input cost: (10,000 ÷ 1,000,000) × $2.50 = $0.025; output cost (assuming 10% of input): (1000 ÷ 1,000,000) × $10.00 = $0.01; total cost per query (RAG): ~$0.035 per query.

https://doi.org/10.1017/cfc.2026.10029 Published online by Cambridge University Press

rather than contemporary semantics (McGillivray 2020). Most commercial LLM RAG products limit the word embeddings that can be used (if they even allow the user to see which embeddings have been used to transform the documents they have uploaded), as well as significantly limiting the number of documents that can be uploaded to the vector store. The primary ATLAS vector store uses word embeddings developed by the *Living With Machines* project. The embeddings are based on the well-regarded Bidirectional Encoder Representations from Transformers (BERT) model, fine-tuned using one billion tokens of content from the British Library's collection of books from 1890 to 1901 (Hosseini et al. 2021). It is worth noting the foresight of that project in producing the embeddings and making them available for wider use. Future work could test our assumption that a vector store developed using the *Living With Machines* BERT word embeddings will perform better than vector stores developed using word embeddings optimised for contemporary content. While it is tempting to assume historical semantics will have a demonstrable effect on the quality of the output, the possibility remains that a sufficiently powerful LLM might be able to draw on its own background knowledge and the context of the corpus to choose appropriate chunks and infer appropriate meaning.

Pushing back against the opacity of many commercial AI products, ATLAS provides detailed technical information about the "calibration" of each system target (model version, word embedding, vector store characteristics, system prompt), as well as the source documents used by the foundation

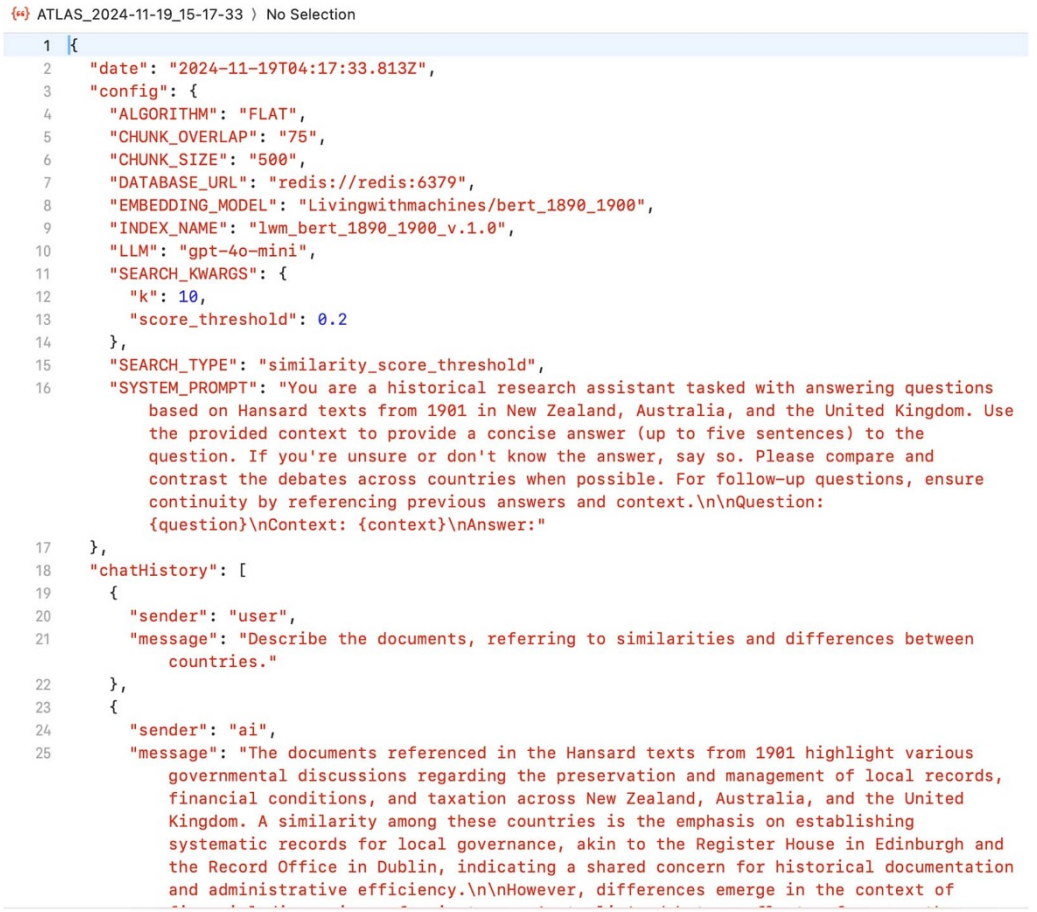


Figure 4. ATLAS .json export.

https://doi.org/10.1017/cfc.2026.10029 Published online by Cambridge University Press

model to generate its response. The first version of the user interface only includes a simple multi-turn Q&A interface, with answers including citations to the document chunks used in the analysis. A privacy setting allows users to perform searches without their queries being captured. Data that is captured is fully anonymised, including the IP of the computer used for testing or any information about the user beyond their questions and feedback ratings.

To facilitate reproducibility, the user interface (UI) also includes export functionality that produces a .json file with the details of the system target and a record of the Q&A and citations (Figure 4). This is a rudimentary nod towards reproducibility but has been augmented with rich telemetry data that captures not only the input and output data but steps in the RAG chain, such as guardrail assessment (which can be configured to detect hate speech or language related to Indigenous culture or other sensitive topics), the retrieval and re-ranking process, and citation generation. By recording token counts for each query and response pair, a sense can be gained of the financial but also environmental cost of the system, and by sending test target metadata, it becomes possible to compare different configurations of the system.

In keeping with our focus on human responses to AI, ATLAS also includes functionality to capture human feedback on the LLM RAG output. We view this as crucial to the evaluation process, allowing subject domain and general users to enter feedback that is added to the telemetry sent to Phoenix Arize (Figure 5). This has been augmented with an inter-rating feature that allows feedback on Q&A sessions to be submitted by multiple users, improving the quality of the dataset. Although the data collected by ATLAS is anonymous, it should allow the project to develop a rich database of test configurations and associated user feedback for later analysis and possible model fine-tuning and LLM-as-Judge testing. AI "observability" frameworks and services (often also offering automated testing services), such as Phoenix Arize, are important for HASS and GLAM communities. Such tools have undeniable connections to the age of "surveillance capitalism" (Zuboff 2019) but reflect a convergence between traditional notions of transparency and reproducibility in HASS research and the technology industry's need for tools and methods capable of derisking probabilistic systems. Such

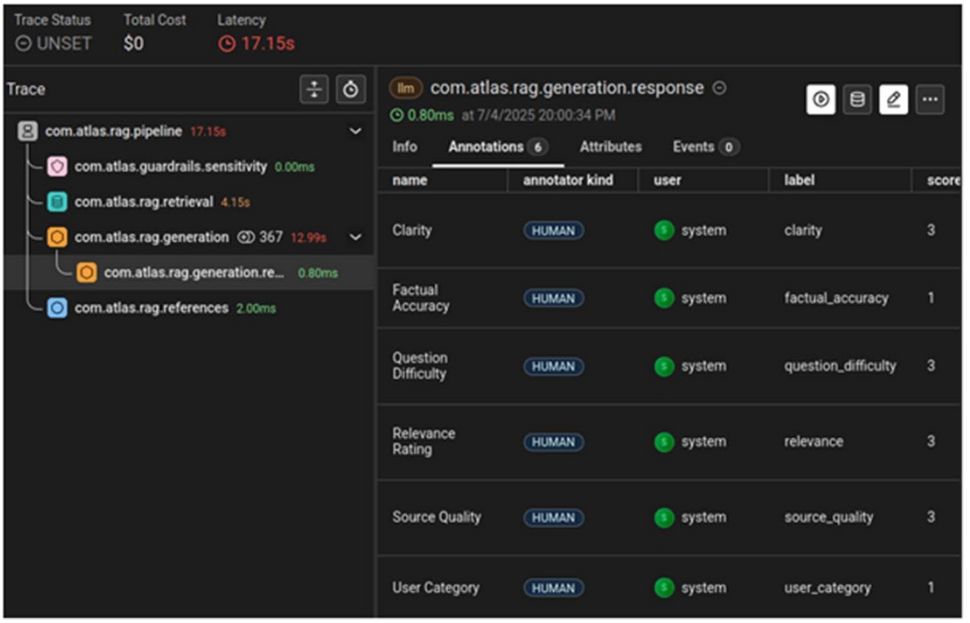


Figure 5. ATLAS user feedback in the Phoenix Arize observability platform.

https://doi.org/10.1017/cfc.2026.10029 Published online by Cambridge University Press

tools meet the needs of commercial auditing and compliance requirements and provide software engineers and data scientists with large amounts of detailed test data. They also provide HASS and GLAM sector practitioners with a level of rigour appropriate to their research and public sector activities.

Ultimately, the intention is to use the telemetry and feedback data for automated testing, whereby a second LLM is trained to assess the quality of responses and asked to grade the output of a range of different system targets. Despite this potential, it is important to note that testing with ATLAS requires substantial attention to human ethics and test planning, to ensure any subsequent operationalisation of the system is properly contextualised (Berman and Smithies 2026). This crucial supporting material is under development at the time of writing. The test plan, used to guide focus groups and online interaction with ATLAS, will be made available along with the source code in due course. The plan uses principles for testing LLM systems informed by HASTRICT principles, referring to the need for testing to be H: Human-focused; A: Accountable; S: System-aware; T: Transparent; R: Reproducible; I: Iterative; C: Contextually Informed; and T: Task based (Smithies and Berman 2025a). These principles pair the technical ATLAS system with controlled human processes to generate actionable research data and will be the ultimate test of the harness.

4. Conclusion

Developing ATLAS as an exercise in critical technical practice has yielded a functional prototype. ATLAS's development reveals the extent of configurability choices (embeddings, chunking, retrieval strategy, system prompts) necessitated by the application of RAG to GLAM datasets and highlights the significant epistemological questions raised by these choices, alongside the radical information-space reduction these choices enable. The choice of word embedding model informs which semantic relationships are encoded in the vector store; the chunking strategy determines what units of evidence the LLM can access; the system prompt shapes the persona and interpretive frame applied to that evidence; and the fetch-k setting determines how much of the corpus is brought to bear on any given query. Each of these choices carries epistemological weight, yet in most commercial deployments, they are made by engineers with no domain expertise in the materials being analysed and are invisible to the researcher or GLAM professional using the resulting tool. The value of ATLAS as a prototype thus lies less in its technical novelty than in the way it makes these choices explicit, configurable and available for critical examination.

ATLAS models a constructive path for DH engagement with LLM technologies through critical technical practice, highlighting that the barriers to critical technical engagement with LLMs are lower than may be implied by AI technologists. ATLAS was developed by a single amateur programmer using AI coding tools, without a dedicated engineering team. While commercial software products have their role, ATLAS demonstrates the potential for DH scholars to build their own prototypes and illustrates the benefits of doing so. The challenges of using emerging and rapidly developing open-source tools to develop experimental LLM applications expose DH scholars to the inevitable design trade-offs and compromises involved in the integration of LLM technologies in research workflows. This exposure can generate a form of embodied understanding that neither user-level interaction with commercial products nor scholarly critique of such products can provide. It is this experiential, practice-based knowledge that Agre's (1997) critical technical practice surfaces.

The development of ATLAS also reveals an unexpected convergence between DH and industry priorities that creates favourable conditions for this kind of work. For DH scholars, transparency and reproducibility are preconditions for critical technical practice; for LLM technologists, traceability and observability are commercial necessities driven by the need to audit and derisk probabilistic systems. Industry is building tools to address these concerns, notably observability platforms such as Phoenix Arize, open telemetry standards and detailed logging of RAG chain processes. These can be repurposed by DH scholars to enable highly transparent and somewhat reproducible experimental LLM systems. This convergence has limits. Data provenance and cultural stewardship, for instance,

https://doi.org/10.1017/cfc.2026.10029 Published online by Cambridge University Press

are foundational concerns for DH and GLAM communities that the technology sector has little incentive to prioritise. ATLAS demonstrates, however, that responsible, research-standard data practices can be realised through the development of open-source tools, rather than through dependence on commercial product licenses whose priorities may not align with those of the research communities they serve.

Nonetheless, at the time of writing, the possibility remains that our experiment in critical technical practice will fail to achieve its objectives. Limitations abound, from the sole developer model to the pace of technological change, the inscrutability of models, the architectural complexity of LLM RAG systems and the need to upgrade to more performant approaches, and problems moving from prototype to full production. Conceived in the heat of the hype around LLMs in June 2024 and developed using high-risk rapid development processes, the cascading levels of technical and sociotechnical complexity, paired with the need to wrap testing in robust human ethics and testing processes, pale in comparison to the enterprise-scale resources and experience commercial teams can bring to bear. Ultimately, although the project remains on a positive trajectory at the time of writing, it could well be that our raison d'être is overtaken by increasingly powerful general-purpose models, the feedback data we aim to collect does not support meaningful analysis or model fine-tuning or the sheer systemic complexity of LLM RAG architecture overwhelms sociotechnical interpretation. Agre's (1997) exhortation to engage in critical technical practice sustains us, a reminder that we must never give up our effort to harness technology for the social good.

Acknowledgements. We would like to thank Barbara McGillivray for her comments on a draft of this article, and the wider AIINFRA team for their support and guidance.

Funding statement. The AI as Infrastructure project is funded by the Australian National University Futures Scheme (J.S.) 2024–2027.

Competing interests. The authors declare no competing interests.

References

Agre, P. E. (1997) 2014. "Toward a Critical Technical Practice: Lessons Learned in Trying to Reform AI." In *Social Science, Technical Systems, and Cooperative Work*: *Beyond the Great Divide,* edited by G. C. Bowker, L. Gasser, L. Star, and B. Turner, 131–57. Computers, Cognition, and Work. New York: Psychology Press.

Bender, E. M., T. Gebru, A. McMillan-Major, and S. Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–23. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922.

Berman, G., and J. Smithies. 2025. "Responsible AI Principles and the GLAM Sector: An Initial Review." *Zenodo*. https://doi.org/10.5281/zenodo.17776154.

Berman, G., and J. Smithies. 2026. "Can AI be Contextualized? Remediating General Purpose AI for Context Specific Use." In *Operationalizing Responsible AI: Opportunities and Challenges*, edited by D. Quercia, and M. Constantinides. Cambridge: Cambridge University Press. Forthcoming.

Berry, D. M. 2023. "The Explainability Turn." *Digital Humanities Quarterly* 17, 2. https://doi.org/10.63744/mcdt3b3nknku.

Bode, K., and L. M. E. Goodlad. 2023. "Data Worlds: An Introduction." *Critical AI* 1 (1–2): 1–2. https://doi.org/10.1215/2834703X-10734026.

Callaghan, S., and M. Vieira. 2025. "Prototyping a RAG System for Digital Humanities: Exploring AI/ML with Indigenous Data." *King's Digital Lab Blog*, January 16. https://kdl.kcl.ac.uk/blog/ireal-rag/.

Chang, Y., X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, W. Ye, Y. Zhang, Y. Chang, P. S. Yu, Q. Yang, and X. Xie 2024. "A Survey on Evaluation of Large Language Models." *ACM Transactions on Intelligent Systems and Technology* 15 (3): 39:1–45. https://doi.org/10.1145/3641289.

Finn, F., and D. Khosrowi. 2026. "AI Assistants in the Archive and the Lure of 'Instant History'." *Cambridge Forum on AI: Culture and Society* 2: e6. https://doi.org/10.1017/cfc.2025.10012.

Hauser, J., D. Kondor, J. Reddish, M. Benam, E. Cioni, F. Villa, J. S. Bennett, D. Hoyer, P. Francois, P. Turchin, and R. M. del Rio-Chanona. 2024. "Large Language Models' Expert-Level Global History Knowledge Benchmark (HiST-LLM)." In *The Thirty-Eighth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. https://openreview.net/forum?id=xlKeMuyoZ5.

Hosseini, K., K. Beelen, and G. Colavizza. 2021. "Neural Language Models for Nineteenth-Century English." *Journal of Open Humanities Data* 7: 22. https://doi.org/10.5334/johd.48.

Huggett, J. 2024. "Slow AI." *Introspective Digital Archaeology*, November 27. https://introspectivedigitalarchaeology.com/2024/11/27/slow-ai/.

International Panel on the Information Environment (IPIE). 2025. "Towards a Global AI Auditing Framework: Assessment and Recommendations." International Panel on the Information Environment. https://www.ipie.info/research/sr2024-3.

Lee, T., M. Yasunaga, C. Meng, Y. Mai, J. S. Park, A. Gupta, and Y. Zhang 2023. "Holistic Evaluation of Text-to-Image Models." *Advances in Neural Information Processing Systems* 36 (NeurIPS 2023): 69981–70011.

Li, J., X. Cheng, X. Zhao, J.-Y. Nie, and J.-R. Wen. 2023. "HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models." In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, edited by H. Bouamor, J. Pino, and K. Bali, 6449–64. Singapore: Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.397.

Lin, S., J. Hilton, and O. Evans. 2022. "TruthfulQA: Measuring How Models Mimic Human Falsehoods." *arXiv*. https://doi.org/10.48550/arXiv.2109.07958.

McGillivray, B. 2020. "Computational Methods for Semantic Analysis of Historical Texts." In *Routledge International Handbook of Research Methods in Digital Humanities*, edited by K. Schuster, and S. Dunn, 261–74. London: Routledge.

McLennan, S., A. Fiske, L. A. Celi, R. Müller, J. Harder, K. Ritt, S. Haddadin, and A. Buyx. 2020. "An Embedded Ethics Approach for AI Development." *Nature Machine Intelligence* 2 (9): 488–90. https://doi.org/10.1038/s42256-020-0214-1.

Pasquale, F. 2015. *The Black Box Society: The Secret Algorithms that Control Money and Information*. Cambridge: Harvard University Press.

Raley, R., and J. Rhee. 2023. "Critical AI: A Field in Formation." *American Literature* 95 (2): 185–204. https://doi.org/10.1215/00029831-10575021.

Safjan, K. 2023. "Problems with Langchain and How to Minimize Their Impact." *Krystian Safjan's Blog*, September 1. https://safjan.com/problems-with-Langchain-and-how-to-minimize-their-impact/.

Serapio-García, G., M. Safdari, C. Crepy, L. Sun, S. Fitz, P. Romero, M. Abdulhai, A. Faust, and M. Matarić. 2023. "Personality Traits in Large Language Models." *arXiv*. https://doi.org/10.48550/arXiv.2307.00184.

Smithies, J., and G. Berman. 2025a. "The HASTRICT Principles: A Brief Overview." *Zenodo*. https://doi.org/10.5281/zenodo.17374271.

Smithies, J., and G. Berman. 2025b. "AI-as-Infrastructure/aiinfra-atlas: V0.2.0 (V0.2.0)." *Zenodo*. https://doi.org/10.5281/zenodo.17204370.

Solaiman, I. 2023. "The Gradient of Generative AI Release: Methods and Considerations." In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 111–22. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3593013.3593981.

Xu, Z., S. Jain, and M. Kankanhalli. 2024. "Hallucination is Inevitable: An Innate Limitation of Large Language Models." *arXiv*. https://doi.org/10.48550/arXiv.2401.11817.

Zuboff, S. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York: Public Affairs.

James Smithies is a professor of digital humanities and director of the HASS Digital Research Hub at the Australian National University. He was previously a professor of digital humanities and the founding director of King's Digital Lab at King's College London. His research is grounded in the history of technology and ideas, and he brings a strong applied aspect to his work. His latest monograph is *Digital Modernity: Why We Need to Think Historically About the Digital Age* (2026).

Glen Berman is a researcher at the HASS Digital Research Hub at the Australian National University and a postdoctoral research fellow at the School of Historical and Philosophical Studies at the University of Melbourne. His research focuses on the implications of artificial intelligence technologies for the science system and the production of scientific knowledge.

Karaitiana Taiuru is a preeminent Māori expert in the intersection of *tikanga Māori* (cultural values), *mātauranga Māori* (traditional knowledge) and emerging digital technologies. He specialises in areas such as AI ethics, Māori Data Sovereignty, digital governance and intellectual property rights. From cultural appropriation in algorithm design to the risks and opportunities AI presents for all communities, Dr Taiuru provides useful and practicable insights and advice into the ethical, governance and rights-based issues that can't be ignored.

Martin Spychal is a historian of modern Britain, specialising in nineteenth-century politics, society and culture. He is the author of *Mapping the State: English Boundaries and the 1832 Reform Act* (University of London Press, 2024). A senior research fellow at the History of Parliament, he works on the House of Commons 1832–1945 project and is also Digital Humanities Lead for the Trust.

https://doi.org/10.1017/cfc.2026.10029 Published online by Cambridge University Press

John Moore is a computer scientist and former member of the Systems Research Group at the University of Cambridge, and Head of Emerging Technologies Research at the National Archives. His specialities include artificial intelligence (AI), data infrastructure, digital humanities, environmental sustainability and International Image Interoperability Framework (IIIF). John served as a senior lecturer in computing for several years and is a senior fellow of the Higher Education Academy. His research interests encompass applied computing challenges within digital humanities, focusing on providing access to digital objects at scale through standards like IIIF and leveraging emerging technologies such as AI.

Cite this article: Smithies, J., Berman, G., Taiuru, K., Spychal, M. and Moore, J. 2026. ATLAS: Harnessing retrieval-augmented generation (RAG). *Cambridge Forum on AI: Culture and Society*, **2**, e17, 1–13. https://doi.org/10.1017/cfc.2026.10029

https://doi.org/10.1017/cfc.2026.10029 Published online by Cambridge University Press