Human Factors in Design, Engineering, and Computing, Vol. 238, 2026, 2238–2248
https://doi.org/10.54941/ahfe1008450

A Distributed Local LLM and Centrally Governed RAG Architecture for Privacy-Preserving Cancer Registry Support

Mitsuhiro Takasaki[1], Kazumi Sasaki[2], Narue Nakabayashi[3], Megumi Hori[3], and Tomohiro Matsuda[3]

[1]Faculty of Science and Engineering, Saga University, 1 Honjo, Saga, Saga 840-8502, Japan
[2]Cancer Center, Saga University Hospital, 5-1-1 Nabeshima, Saga-shi, Saga 849-8501, Japan
[3]Center for Cancer Registries, National Cancer Center Institute for Cancer Control, 5-1-1 Tsukiji, Chuo-ku, Tokyo 104-0045, Japan

ABSTRACT

Hospital-based cancer registration requires registrars to interpret detailed and frequently updated coding rules, disease-specific manuals, staging classifications, and historical question-and-answer records. Retrieval-augmented generation (RAG) can provide natural-language access to these resources, but cloud-based implementations raise concerns about transmitting clinical information outside the hospital. Moreover, retrieval of the correct document does not guarantee correct application of a registration rule. This paper presents a prototype architecture for deploying a local large language model and vector database within each hospital, while centrally governing and distributing approved reference documents, metadata, retrieval policies, and selected deterministic validation rules. The prototype uses FastAPI, Ollama, ChromaDB, multilingual embeddings, and locally executable language models. Questions are routed to prioritized document groups including the standard registration form, ICD-O materials, staging and site-specific resources, and historical Q&A records. For closed, explicitly enumerated, or high-risk rules, generative processing is supplemented by deterministic safeguards. A formative evaluation involving prostate-specific antigen illustrates the limitation of conventional RAG: although PSA is clinically associated with prostate cancer, it is not one of the four combinations permitted as a site-specific tumor marker for diagnostic-basis coding in the Japanese cancer registry. The case illustrates that semantic plausibility and registry correctness are not necessarily equivalent. The proposed functional allocation assigns natural-language interpretation, retrieval, synthesis, and explanation to the LLM; exact membership checking to deterministic components; and final confirmation to the registrar. This hybrid design is intended to improve privacy, traceability, maintainability, and calibrated human reliance.

Keywords: Cancer registry, Retrieval-augmented generation, Local large language model, Privacy-preserving AI, Human-in-the-loop, Deterministic guardrail

INTRODUCTION

Hospital-based cancer registries support cancer surveillance, quality improvement, clinical research, and health policy. Registry personnel

Received August 8, 2026; Revised September 27, 2026; Accepted October 6, 2026; Available online December 1, 2026

© 2026 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

transform information from pathology reports, imaging reports, operative records, and other clinical documents into standardized data elements. The work requires medical knowledge and exact interpretation of registration rules, ICD-O codes, staging classifications, site-specific instructions, and nationally defined procedures. Because the necessary knowledge is distributed across heterogeneous and versioned documents, registrars frequently search several sources before reaching a decision.

Large language models (LLMs) offer a natural-language interface to these materials. Retrieval-augmented generation (RAG) can ground responses in external documents and update the knowledge base without retraining the model (Lewis et al., 2020). However, two problems remain. First, questions may contain sensitive clinical details that hospitals cannot transmit to external cloud services. Second, retrieval of a relevant passage does not ensure that the model will apply a restrictive registration rule correctly. Retrieval quality, evidence faithfulness, and answer relevance are distinct dimensions of performance (Es et al., 2023; Saad-Falcon et al., 2023).

This paper presents a distributed architecture that combines local computation with central knowledge governance. In the proposed architecture, each participating hospital operates its own LLM, embedding model, vector database, and user interface. A designated central governance body would maintain approved source documents, metadata, retrieval configurations, terminology mappings, deterministic safeguards, and regression tests. Patient-level questions and generated responses remain within the hospital. The primary contribution is a human-centered functional allocation among generative AI, deterministic validation, and registrar review.

METHOD AND DESIGN REQUIREMENTS

This study used a design-and-formative-evaluation method comprising three stages: defining requirements from cancer-registry workflow and AI-governance considerations, implementing a local RAG prototype, and examining a closed-rule case in which clinical plausibility could conflict with registry correctness. The prototype was designed around five requirements. First, privacy-sensitive questions, retrieved passages, and generated answers should remain within the institution. Here, privacy preservation refers to avoiding routine transmission of patient-level questions, retrieved passages, and generated responses to external model services; it does not replace institutional controls for access, endpoint security, audit, backup, and data retention. Second, the knowledge base should distinguish authoritative manuals from historical Q&A and retain version, page, site, role, and priority metadata. Third, users should be able to inspect supporting evidence and recognize when information is insufficient. Fourth, common rules should be governed centrally to avoid incompatible local interpretations. Fifth, the system should support appropriate abstention and retain final registrar oversight when information is insufficient or uncertain (NIST, 2023; World Health Organization, 2021, 2024). These requirements separate protection of case-specific information from governance of shared registration knowledge. Table 1 summarizes the requirements and corresponding mechanisms.

Table 1: Principal design requirements and corresponding system mechanisms.

Requirement	System Mechanism
Local confidentiality	Local LLM inference, embeddings, vector search, logs, and institutional documents.
Authoritative grounding	Metadata-aware retrieval with document-group and priority controls.
Traceability	Separate display of generated answers, source file, page or row, and document identifier.
Maintainability	Centrally reviewed packages containing documents, metadata, routing policies, guardrails, and regression tests.
Human oversight	Explicit evidence display, abstention when evidence is insufficient, and final registrar confirmation.

DISTRIBUTED SYSTEM ARCHITECTURE

Figure 1 shows the proposed separation between local computation and central governance. Within each hospital, a browser interface communicates with a FastAPI application. The application performs question classification, local embedding, vector retrieval, prompt construction, and inference through a locally running model. ChromaDB stores persistent vectors for official references and historical Q&A in separate collections. The central governance function distributes approved non-patient resources but does not receive clinical questions or answers.

The prototype was implemented with FastAPI, Ollama, ChromaDB, a multilingual BGE-M3 embedding model (Chen et al., 2024), and a browser-based interface. Locally executable model options evaluated in the prototype include models from the Qwen, Gemma, Phi, and gpt-oss families. The interface allows the user to select a model, submit a natural-language question, review the generated response, and inspect retrieved sources. The architecture is hardware-independent; a compact workstation with sufficient unified or accelerator memory is one practical deployment option.

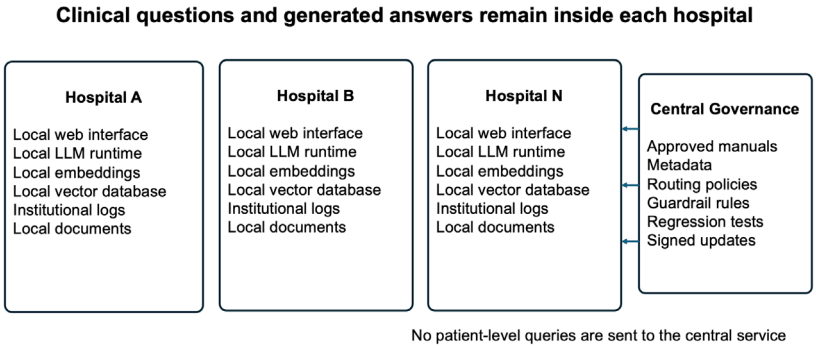


Figure 1: Distributed local inference and centrally governed knowledge architecture. Patient-level questions remain within each hospital, while approved documents, metadata, routing policies, and guardrails are distributed centrally.

Table 2: Prototype components and their roles.

Component	Role
Browser interface	Question entry, model selection, response presentation, and evidence inspection.
FastAPI application	Question routing, retrieval orchestration, prompt construction, and response formatting.
Ollama	Local execution of selectable LLMs and embedding models.
ChromaDB	Persistent local vector collections for formal references and historical Q&A.
Central update package	Approved manuals, metadata, terminology, routing rules, deterministic safeguards, and tests.

RETRIEVAL AND RESPONSE GENERATION

Reference documents are converted into structured records containing extracted text and metadata such as source filename, page, worksheet, row, document group, document role, cancer site, and priority. Formal documents and historical Q&A are indexed separately. The prototype distinguishes the standard registration form, ICD-O topography materials, ICD-O morphology materials, staging documents, site-specific resources, and historical consultation records.

A lightweight classifier routes the question before vector search. Terms related to primary site or C-codes prioritize topography documents; morphology and pathology terms prioritize morphology documents; TNM, stage, and extent-of-disease terms prioritize staging documents; and anatomical terms activate site-specific collections. The standard registration form is searched for core registration items. Targeted and broad results are merged, deduplicated, and ranked using semantic similarity together with document authority and metadata.

The generation prompt instructs the model to use only supplied references, prioritize formal manuals over historical Q&A, avoid unrelated cancer-site materials, and abstain when evidence is insufficient. The answer format contains a conclusion, rationale, cautions, and references. Figure 2 illustrates the prototype interface with the question, selected local model, generated answer, and retrieved references in a single view.

HYBRID GENERATIVE AND DETERMINISTIC SUPPORT

RAG accuracy depends on document quality, extraction, chunking, embeddings, retrieval, ranking, context selection, prompt adherence, rule interpretation, and answer presentation. A correct passage may be retrieved but misapplied. This is especially likely when a manual defines a closed list using expressions such as “only in the following cases.” Semantic similarity encourages related concepts, whereas a closed rule requires exact membership checking.

The prototype therefore does not treat every question as an unrestricted generation task. Generative AI handles natural-language interpretation, terminology variation, retrieval, document synthesis, explanation, and identification of missing information. Deterministic processing is reserved for explicitly enumerated combinations, prohibited values, mutually exclusive options, numeric thresholds, and other high-risk decisions that can be represented as stable structured rules. The registrar remains responsible for confirming the clinical facts and accepting or rejecting the recommendation.

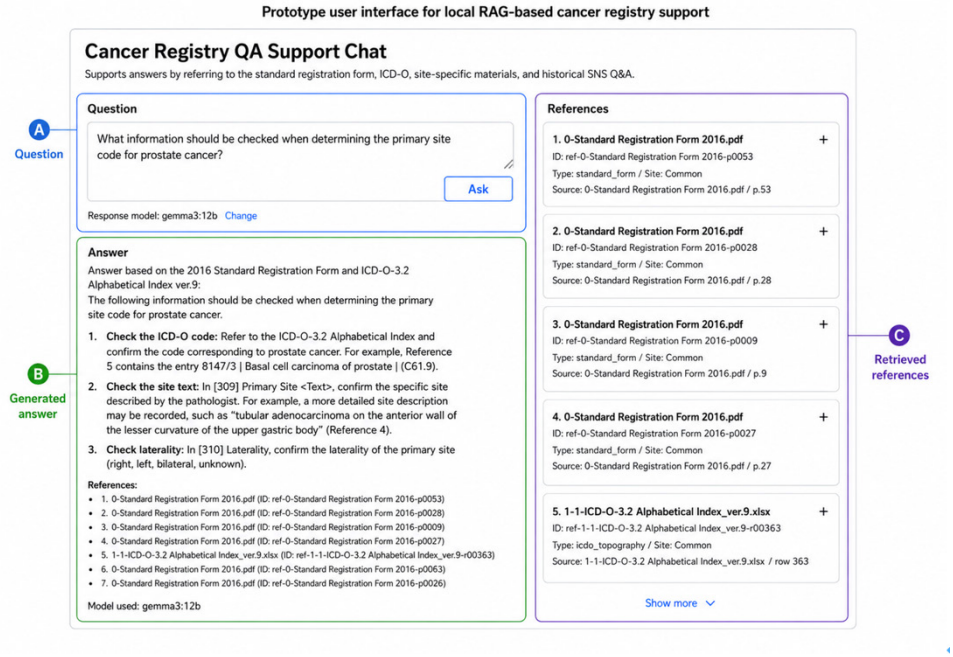


Figure 2: Prototype user interface for local RAG-based cancer registry support. The interface separates the user question, generated response, and retrieved evidence to support traceability and human verification.

Table 3: Functional allocation in the hybrid decision-support design.

Function	Preferred Mechanism	Reason
Interpret an incomplete natural-language question	LLM	Handles paraphrases, spelling variation, and ambiguous wording.
Retrieve and synthesize several manuals	RAG + LLM	Supports contextual explanation with traceable evidence.
Check membership in an exhaustive permitted list	Deterministic validation	Avoids semantic over-generalization and inconsistent generation.
Identify missing clinical facts	LLM with constrained prompt	Can explain what additional information is required.
Make the final registration decision	Human registrar	Retains professional responsibility and contextual judgment.

FORMATIVE EVALUATION: PSA AND DIAGNOSTIC BASIS

The Japanese Standard Registration Form specifies that diagnostic basis 4, site-specific tumor marker, may be used only for four combinations: elevated AFP in hepatocellular carcinoma with morphology code 8170/3; elevated HCG in choriocarcinoma with morphology code 9100/3; elevated VMA in neuroblastoma with morphology code 9500/3; and elevated immunoglobulin in Waldenström macroglobulinemia with morphology code 9761/3 (National Cancer Center Japan, 2016). The wording defines an exhaustive list rather than representative examples.

The PSA case was selected for formative evaluation because it creates a controlled contrast between a clinically plausible association and an exhaustive registry rule. A general-purpose LLM may infer that PSA is a site-specific tumor marker and recommend diagnostic basis 4, although PSA is not included in the permitted list. Conventional RAG may also fail if the decisive page is not retrieved, the limiting statement is separated from the list during chunking, unrelated prostate materials outrank the rule, or the model combines the retrieved rule with general medical knowledge.

In the formative evaluation, the prototype restricted retrieval to the authoritative standard-form page, suppressed historical Q&A, checked the case against the permitted combinations, and generated a constrained explanation. Figure 3 shows the resulting interface, in which the system presents the negative determination, the four permitted conditions, the authoritative source, and the absence of historical Q&A. This result illustrates the intended rule-in-the-loop behavior: deterministic validation constrains the generative explanation, while the registrar retains final control.

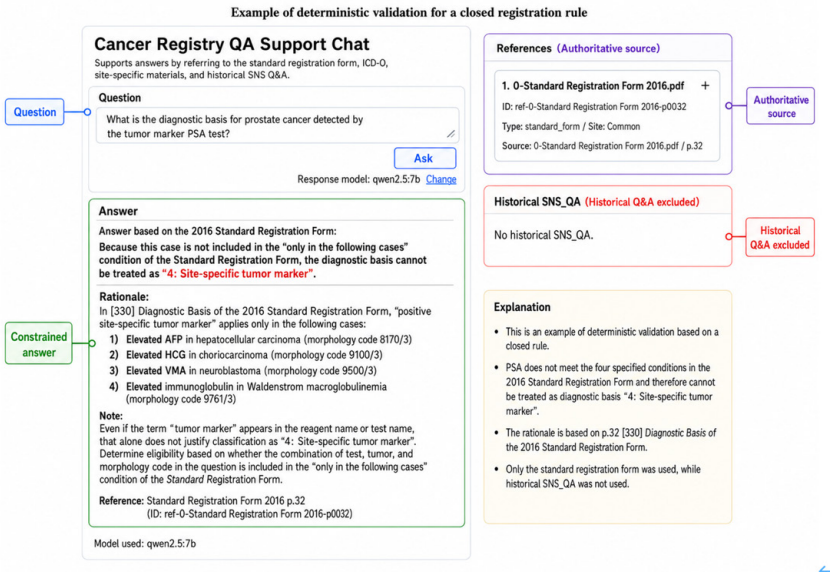


Figure 3: Deterministic validation for a closed registration rule. Diagnostic basis 4 is rejected for the PSA-related prostate-cancer case because PSA is not among the four permitted tumor-marker combinations; retrieval is restricted to the authoritative source, and historical Q&A is excluded.

FACTORS AFFECTING ANSWER ACCURACY

Source quality and version control are foundational. Superseded, incomplete, or duplicated documents cannot be corrected by a larger model. Every indexed resource should have an explicit version, effective period, authority level, and document role. When historical versions are retained, retrieval should use the case date to identify the applicable rule.

Document extraction and chunk construction determine whether logical units remain intact. Cancer registration manuals contain tables, footnotes, codes, and multi-column layouts. Rule-aware chunks should preserve the heading, definition, limiting words, complete enumerated list, relevant footnotes, and associated code. In the PSA case, separating "only in the following cases" from the four combinations would remove the essential logical constraint.

Retrieval should combine dense semantic search with metadata filters and exact lexical matching. Alphanumeric codes such as 8170/3 and abbreviations such as PSA may be more reliably located through hybrid retrieval than through embeddings alone. Ranking should also consider document authority, version applicability, cancer-site match, and whether the passage contains a formal definition rather than a secondary discussion.

Prompt constraints can reduce unsupported generation but cannot guarantee compliance. Excessive context may dilute attention to the decisive passage. Primary rules should appear first, closed lists should be marked explicitly, and every conclusion should cite evidence. Low-temperature decoding may improve consistency, but deterministic decoding does not correct an invalid reasoning path.

Human factors influence operational accuracy. The interface should make source authority, document version, evidence location, missing information, guardrail activation, and uncertainty visible. The objective is calibrated reliance: users should neither accept fluent answers without checking evidence nor reject useful assistance merely because it is generated by AI (NIST, 2023; World Health Organization, 2021).

CENTRAL GOVERNANCE AND LOCAL DEPLOYMENT

A centrally governed release package should contain approved source files, checksums, extraction outputs, chunk definitions, metadata, terminology dictionaries, embedding and indexing configuration, query-routing policies, deterministic rules, regression questions, expected answers, version information, and release notes. Packages should be verifiable, and each hospital should be able to activate an update locally with rollback capability.

Institution-specific workflow manuals may be indexed in a separate local namespace. Local documents can explain institutional processes but should not silently override national rules. When central and local documents differ, the interface should display their authority and provenance. A proposed update cycle includes document verification, change detection, re-extraction, guardrail revision, regression testing, central review, signed distribution, local verification, and controlled activation.

PRACTICAL DEPLOYMENT FEASIBILITY FOR MID-SIZED HOSPITALS

A practical advantage of the proposed architecture is that it does not require each hospital to operate a data-center-class GPU server. A compact workstation can host the language model, embedding model, vector database, and web application as a single institution-controlled appliance. The current Mac Studio platform provides high-bandwidth unified memory, local solid-state storage, and built-in 10-Gigabit Ethernet in a 19.7 × 19.7 × 9.5 cm enclosure (Apple, 2026), making workstation-class deployment practical where server-room space or specialist staffing is limited.

For a mid-sized hospital, the expected workload differs substantially from that of a public cloud service, with a limited number of registrars using interactive question answering rather than high-volume model training. Depending on response-time and concurrency requirements, such a workload may be supported by a quantized model with 7–14 billion parameters and a moderate-sized local knowledge base. Unified memory allows model weights and inference data to share one memory pool, avoiding separate GPU-memory constraints. Deployment can scale from a single workstation to additional nodes or a larger local server without changing the centrally governed document package.

Economic feasibility should be evaluated through total cost of ownership rather than purchase price alone. A compact workstation can reduce the need for a dedicated GPU chassis, specialized cooling, complex driver management, and recurring per-query cloud charges while keeping sensitive prompts within the hospital. Costs remain for backup, access control, maintenance, update validation, and local technical support; therefore, the proposal is not that Mac Studio is universally optimal, but that workstation-class local inference may make privacy-preserving deployment financially and operationally practical for mid-sized hospitals. A formal comparison with cloud APIs and GPU servers is planned as part of the multi-institutional evaluation.

Table 4: Illustrative deployment profile for a mid-sized hospital.

Deployment Element	Illustrative Configuration	Operational Rationale
Local inference host	One workstation-class system, such as Mac Studio, with sufficient unified memory for a quantized 7–14B model	Avoids a custom GPU server while keeping prompts, retrieval, and generation on premises
Knowledge storage	Local SSD for approved reference packages, vector indexes, logs, and institution-specific documents	Provides predictable response time and permits local backup and retention control
Network integration	Hospital LAN connection; external access not required for routine inference	Supports internal browser access without transmitting case questions to an external AI service

(Continued)

Table 4: Continued.

Deployment Element	Illustrative Configuration	Operational Rationale
Users and workload	Small registrar team with interactive, low-concurrency question answering	Matches the workload of many mid-sized hospitals rather than cloud-scale throughput
Growth path	Additional workstation or migration to a larger local server	Allows redundancy and concurrency to expand without changing the governance model
Cost controls	Capital purchase, scheduled updates, and local support; no mandatory per-query model fee	Improves budget predictability while retaining institutional control of sensitive data

FUTURE EVALUATION

The present study provides a prototype implementation and formative evaluation of the proposed architecture; systematic comparative evaluation is required before operational use. An expert-annotated test set should cover registration eligibility, diagnostic basis, topography, morphology, TNM, extent of disease, site-specific rules, missing-information cases, outdated-rule traps, and misleading clinical assumptions. Questions should include paraphrases, abbreviations, and adversarial wording.

Four configurations should be compared: local LLM without retrieval; conventional vector-search RAG; metadata-routed RAG with strict prompting; and hybrid RAG with deterministic safeguards. Retrieval measures should include authoritative-document recall, decisive-passage recall, ranking, and version correctness. Answer measures should include decision accuracy, false-positive and false-negative rates, evidence faithfulness, citation accuracy, appropriate abstention, and repeated-run consistency.

A human-factors study with cancer registrars should evaluate task completion time, correction rate, workload, confidence calibration, source-inspection behavior, usability, trust, and inappropriate reliance. Operational testing should assess local data flow, access control, update verification, response latency, memory consumption, concurrent use, and recovery from failed updates.

DISCUSSION

The proposed architecture addresses two opposing centralization problems. Clinical data should not be centralized merely to obtain AI assistance, while registration rules should not be independently reinterpreted at every hospital. Distributing computation and centralizing governance offers a practical compromise. The PSA case illustrates how registry-operational correctness can differ from medical-semantic plausibility. In administrative and coding tasks, a medically reasonable answer can still be formally wrong.

The hybrid design is not a return to a fully rule-based expert system. Encoding all registration knowledge as rules would be expensive, brittle, and

poorly suited to incomplete or explanatory questions. Instead, deterministic safeguards are narrow controls for stable, explicit, and consequential rules. The generative component remains responsible for the conversational interface and most document interpretation. This allocation uses each component where its strengths are greatest.

Limitations include the absence of a large expert-annotated performance study, keyword-based routing in the current prototype, variable PDF extraction quality, and the narrow scope of the initial guardrail. A production marker rule must validate the complete tuple of tumor, morphology code, marker, and result. Local deployment also requires institutional access control, audit policy, workstation security, and user training. Central governance requires an accountable organization to review changes and resolve conflicting interpretations.

CONCLUSION

This paper presented a distributed local LLM and centrally governed RAG architecture for privacy-preserving cancer registry support. Clinical questions, retrieval, and generation remain within each hospital, while approved reference knowledge, metadata, routing policies, and deterministic safeguards are centrally maintained and distributed. The prototype combines question routing, metadata-aware retrieval, evidence display, constrained generation, and selective rule-in-the-loop validation.

The PSA case illustrates a situation in which retrieval alone may be insufficient: an LLM may prefer a familiar clinical association even when a closed registry rule excludes it. The proposed functional allocation assigns natural-language interpretation and explanation to the LLM, exact closed-rule validation to deterministic components, and final responsibility to the registrar. Future work will include expert benchmark construction, model and retrieval comparison, usability testing, formalization of additional high-risk rules, and multi-institutional operational evaluation.

ACKNOWLEDGMENT

The authors would like to thank the cancer registration professionals who provided practical insights into the interpretation of registration rules and the development of system requirements. This work was supported in part by the National Cancer Center Research and Development Fund under Project No. 2025-A-19, “Development of Privacy-Preserving Data Sharing Methods and an International Collaborative Database.”

REFERENCES

Apple (2026) Mac Studio - Technical Specifications. Apple Japan. Available at: https://www.apple.com/jp/mac-studio/specs/ (Accessed: 26 July 2026).

Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D. and Liu, Z. (2024) ‘BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings through Self-Knowledge Distillation’, arXiv preprint, arXiv:2402.03216.

Es, S., James, J., Espinosa-Anke, L. and Schockaert, S. (2023) 'RAGAS: Automated evaluation of retrieval augmented generation', arXiv preprint, arXiv:2309.15217.
Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W-t., Rocktäschel, T., Riedel, S. and Kiela, D. (2020) 'Retrieval-augmented generation for knowledge-intensive NLP tasks', Advances in Neural Information Processing Systems, 33, pp. 9459–9474.
National Cancer Center Japan (2016) Standard Registration Form for Hospital-Based Cancer Registries, 2016 Edition [in Japanese]. Tokyo: National Cancer Center Japan.
National Institute of Standards and Technology (2023) Artificial Intelligence Risk Management Framework (AI RMF 1.0). Gaithersburg, MD: NIST.
Saad-Falcon, J., Khattab, O., Potts, C. and Zaharia, M. (2023) 'ARES: An automated evaluation framework for retrieval-augmented generation systems', arXiv preprint, arXiv:2311.09476.
World Health Organization (2021) Ethics and Governance of Artificial Intelligence for Health: WHO Guidance. Geneva: World Health Organization.
World Health Organization (2024) Ethics and Governance of Artificial Intelligence for Health: Guidance on Large Multi-Modal Models. Geneva: World Health Organization.