Journal of Clinical and Translational Science

www.cambridge.org/cts

Hide and seek: Privacy-preserving artificial intelligence with a feasibility study in rare disease diagnosis

Sivaraman Rajaganapathy[1,2], Jennifer St. Sauver[3], Filippo Pinto e Vairo[2,4], Prasad G. Iyer[2,4], Hongfang Liu[5] and Jungwei W. Fan[1,2,3]

[1]Department of Artificial Intelligence and Informatics, Mayo Clinic, Rochester, MN, USA; [2]Center for Clinical and Translational Science, Mayo Clinic, Rochester, MN, USA; [3]Division of Epidemiology, Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN, USA; [4]Division of Gastroenterology and Hepatology, Department of Medicine, Mayo Clinic, Rochester, MN, and Scottsdale, AZ, USA and [5]Department of Quantitative and Systems Health Sciences, Dell Medical School, The University of Texas at Austin, Austin, TX, USA

Research Article

Cite this article: Rajaganapathy S, St. Sauver J, Pinto e Vairo F, Iyer PG, Liu H, and Fan JW. Hide and seek: Privacy-preserving artificial intelligence with a feasibility study in rare disease diagnosis. *Journal of Clinical and Translational Science* **10**: e145, 1–9. doi: 10.1017/cts.2026.10800

Received: 15 January 2026 UTC
Revised: 8 June 2026 UTC
Accepted: 4 July 2026 UTC

Keywords:
Patient data privacy; homomorphic encryption; large language models; rare disease diagnosis; information retrieval

Corresponding author:
J. W. Fan; Email: fan.jung-wei@mayo.edu

Abstract

Background: Integrating advanced artificial intelligence (AI) into clinical decision-support often requires the sharing of sensitive patient data with external services, raising privacy concerns. Homomorphic encryption (HE) allows computing directly on encrypted data, without revealing the underlying patient information. *Objectives:* To develop a large language model (LLM)-assisted diagnosis framework while preserving patient privacy in the clinical text analysis, by leveraging HE and using rare disease (RD) diagnosis as a representative application. To demonstrate HE does not hinder the system performance. *Materials and Methods:* Texts from patient histories and a RD knowledge base were embedded by LLMs into vectors, then encrypted using HE to obscure private information while retaining the semantic nuances. Diagnostic recommendations were generated by computing and ranking the similarities between the patient history and RD vectors in the encrypted space. The system was evaluated using 50 synthetic case reports (5 RDs, each with 10 reports). *Results:* Applying HE did protect private information from reverse-embedding attacks. HE imposed little disruption to the diagnostic accuracy, with normalized discounted cumulative gains (nDCG) of 0.6108 ± 0.3412 (encrypted) versus 0.6083 ± 0.3415 (unencrypted). The accuracy and computational performance were tunable and consistent, as demonstrated across five different LLMs. *Discussion:* Our privacy-preserving framework opens tremendous opportunities toward hosting and serving powerful AI solutions across institution boundaries, which would remove the need for local deidentification and incentivize users to access secure external decision-support services. *Conclusions:* Integrating HE with LLM retrieval can promote the dissemination of non-redundant, high-capacity AI services by preserving both privacy and accuracy.

© Mayo Foundation for Medical Education and Research, 2026. Published by Cambridge University Press on behalf of Association for Clinical and Translational Science. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike licence (https://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.

Introduction

Recent developments in large language models (LLMs) and broadly artificial intelligence (AI) have advanced the capacity of decision-support solutions. Seamless integration of LLM into clinical decision-support often requires access to protected health information (PHI) and raises critical security concerns. Sending PHI to outside entities for AI solutions risks data breaches, which have more than tripled (216–745) from 2010 to 2023 in the United States (US) healthcare sector [1]. Countermeasures that are commonly adopted include sticking with on-premises AI deployment, especially for small to medium sized organizations that cannot afford a private cloud service and the associated costs. However, the on-premises approach limits options to leverage more powerful, specialized AI tools while reinforcing siloed, suboptimal efforts at a local scale. An example of this problem is shown by issues observed in research on rare diseases (RDs). Despite the collective sum of 400 million people affected by RDs worldwide and a hefty economic burden of $997 billion in the US [2], individual RDs are characterized by very low prevalence and are subject to an average diagnostic delay of approximately five years [3]. A major barrier to diagnosis is that many patients and primary care providers do not have an easily accessible yet private mechanism to share PHI for consultation, along with other challenges such as the inherent complexity of RDs, heterogeneous clinical presentations, and limited clinician familiarity [4,5]. These problems could potentially be overcome by centrally hosted RD diagnosis services that operate on the best technology and knowledge while being "blind to" any private information sent from the users.

To achieve the said privacy-preserving AI for clinical text, solutions have been proposed and deidentification (DeID) [6,7] is of top relevance in this context. While DeID by modern AI has achieved high accuracy, it does not guarantee 100% clearance of PHI and can corrupt the useful

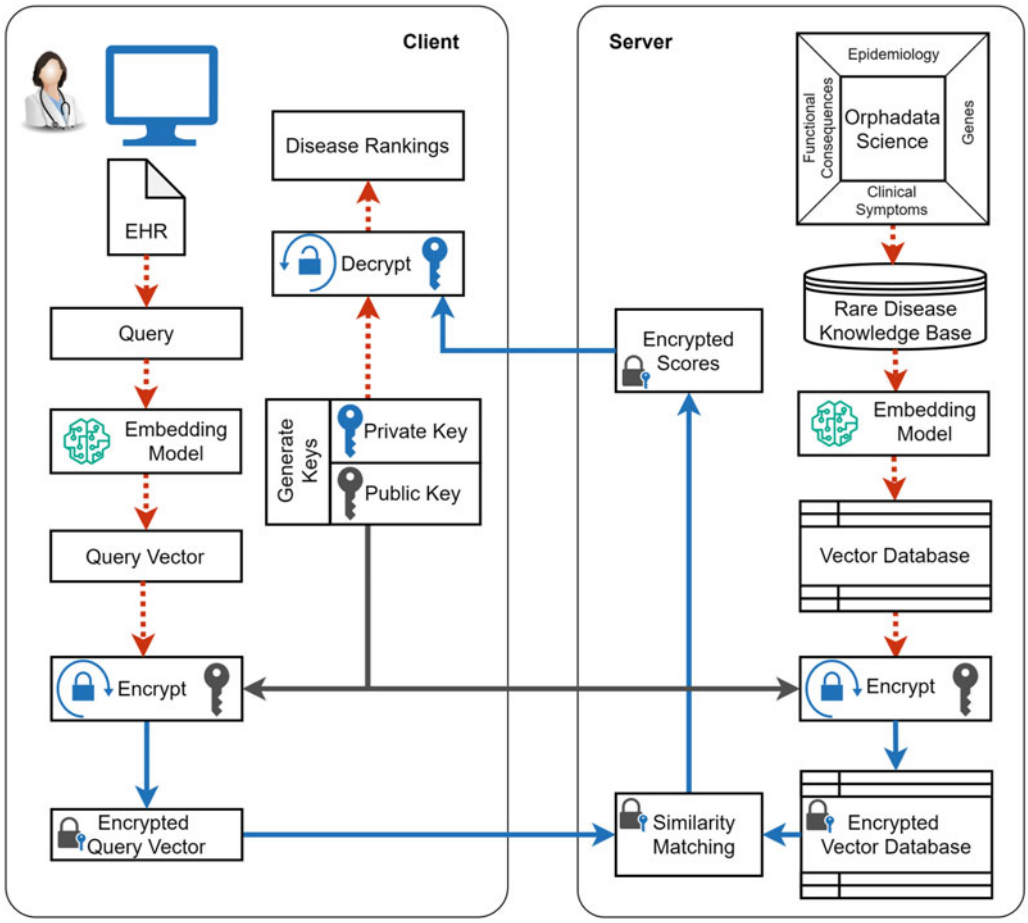


Figure 1. Privacy preserving retrieval with homomorphic encryption (HE). The process comprises steps performed on client and server sides. Here we illustrate the framework through a rare disease diagnosis application, with a custom disease vector database created from the Orphadata Science database. The red dashed arrows represent unencrypted data and the blue solid arrows represent secure encrypted data, as elaborated in the following. A query from the client, which may include a whole clinical note with private information, is transformed into a query vector using an embedding model. The client subsequently generates a unique pair of encryption keys: a public key that is used to encrypt the query vector and shared with the server; a private key that is kept secure later for decrypting the server response back to human-readable information. The server uses the client's public key to encrypt its vector database and then computes the similarity scores of the query against the vector database in the encrypted space. The encrypted similarity scores are returned to the client, which can decrypt the scores to obtain the retrieval rankings.

contents upon aggressive scrubbing. To seek a balanced solution for both strict privacy and minimum burden to the users, we revisited existing techniques and recognized that encryption not only fits well in the current AI stack, where most information is vectorized (as embeddings), but provides some of the strongest formal notions of privacy [6]. Specifically, homomorphic encryption (HE) enables mathematical operations directly on encrypted embeddings and easier AI integration in comparison to conventional encryption schemes. Therefore, we propose combining HE and LLMs to develop a retrieval framework that performs similarity searches using embeddings entirely in the encrypted space. Sending secure queries to AI-mediated knowledge bases (e.g., diagnosis guidelines) allows computation-intensive jobs to be outsourced to third-party decision-support services without revealing sensitive information. Such a framework can be very useful to the Clinical and Translational Science Awards network, for the privacy-preserving feature would encourage developing, hosting, and sharing AI services unique to institution strengths and resources. Centralizing such efforts should reduce redundancy and waste, while allowing users to benefit from more mature solutions in specialized areas.

Therefore, we conducted a study to demonstrate the feasibility of integrating of HE and LLM-based diagnostic reasoning, with maintenance of retrieval performance under encryption. Our use case focused on RD diagnosis from patient history in clinical notes, where privacy-preserving was essential. We further validated that our framework worked for different LLMs and that all computations can be completed within practical time and computational resources.

Materials and methods

Overview

The proposed framework for privacy-preserving diagnostic retrieval is shown in Figure 1, where the client-side processes are depicted at left and the service side at the right. At the client side, a patient's clinical note is first converted to an LLM embedding vector, referred to as the *query vector*. The client also generates a unique pair of keys for the encryption – the private key is retained by the client, while the public key is used to encrypt the query vector and shared with the server. On the other end, the server uses an LLM to convert an entire disease reference knowledge base into a vector database. In response to a query from the client, along with the public key, the server transforms the vector database into a compatible encrypted space. The server then computes the cosine

https://doi.org/10.1017/cts.2026.10800 Published online by Cambridge University Press

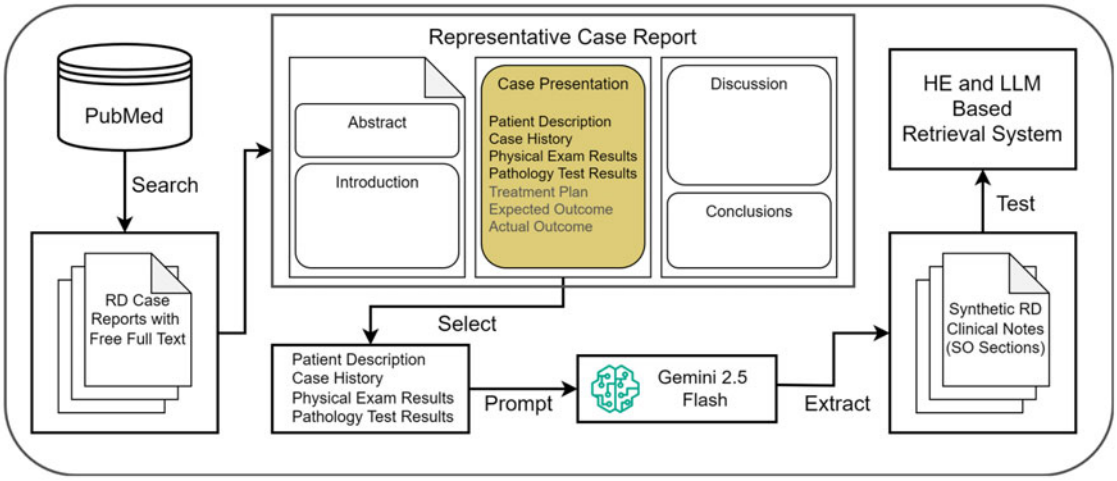


Figure 2. Generation of synthetic test clinical notes. The flowchart shows the process used to generate the synthetic clinical notes for our experiments. The source data consisted of open-access, full-text rare disease (RD) case reports available on PubMed. From each RD case report, we manually selected case presentation excerpts that did not reveal the final diagnosis of the RD. The excerpts were then used to prompt the public version of *Gemini 2.5 Flash*, which transformed the excerpts into SOAP-format clinical notes (i.e., with the Subjective, Objective, Assessment, Plan sections). The Subjective and Objective sections from the LLM-generated synthetic notes were extracted and used for testing our homomorphic encryption (HE)-enhanced diagnostic retrieval framework.

similarities between the query vector and the vector database entries directly in the encrypted space. Lastly, the encrypted similarity values are returned and decrypted by the client using the private key, followed by ranking the top hit disease entries to complete the diagnostic retrieval. Note that the exact text and its real meaning from the client are never revealed to the server in this framework.

Rare disease knowledge base

We used the *Orphadata Science* [8,9], a publicly accessible database providing structured, expert-curated information from the *Orphanet Rare Disease Knowledge Base*. The database contains diverse information associated with over 6172 unique RDs [9]. We created RD profiles by linking each unique disease with a summary text comprising the associated genes and clinical symptoms. Subsequently, an LLM is used to create an embedding vector from the summary text for each of the RDs, together making our final RD vector database. We used the Python library *Sentence-Transformer* to create the RD vector database [10]. Since the library allows plugging in a range of different LLMs, we experimented with several models of varying sizes (80 MB to 15 GB) and embedding dimensions (384 to 4096). A summary of the models is provided in the Supplementary Material Table S1.

Rare disease test cases

To demonstrate and validate our framework, we created synthetic clinical notes as input to simulate a privacy-preserving RD diagnosis application. Based on the authors' own research interest, five RDs were chosen for the experiments: amyotrophic lateral sclerosis (ALS), fibrodysplasia ossificans progressiva (FOP), idiopathic pulmonary fibrosis (IPF), primary sclerosing cholangitis, and trigeminal neuralgia (TN). Ten synthetic notes were generated for each of the five RDs (process shown in Figure 2) to serve as queries for the diagnostic retrieval. We used the public version of the LLM, *Gemini 2.5 Flash,* specifically prompted to follow the Subjective, Objective, Assessment, Plan (SOAP) structure [11]. As guiding blueprints of realistic medical histories, we supplied *Gemini 2.5 Flash* with case presentation sections selected from full-text PubMed case reports of the five RDs. We ensured that no text revealing the final diagnosis was present in the synthetic notes, so as to simulate the inputs of undiagnosed cases in the task. Accordingly, we extracted only the Subjective and Objective sections of the LLM-synthesized clinical notes for testing the diagnostic retrieval framework. In addition, we prompted the LLM to inject mock patient names in each synthetic note, since the PubMed-sourced case reports did not contain any PHI. Note that the encryption can protect any PHI in the text, and using patient name is just for demonstration.

Encrypted retrieval

Our retrieval framework integrates functionality from two libraries – *Sentence-Transformer* [10] for transforming text to the embedding vector space and *TenSEAL* [12] for HE, decryption, and mathematic operations in the HE space. From the *Sentence-Transformer* library, our framework inherits the flexibility to choose and deploy from a wide range of embedding LLMs. To generate embedding vectors, free text is chunked to fit within the maximum input sequence limits of the LLM used. Wherever possible, the chunking is done with retaining complete sentences. The contiguous chunks have an adjustable overlap. For retrieval involving multiple input chunks, we used the cosine similarity metric $C_{(\overline{X},\overline{Y})} := \frac{\sum_{i=1}^{N} \overline{x}_i \overline{y}_i}{\sqrt{\sum_{i=1}^{N} \overline{x}_i^2} \sqrt{\sum_{i=1}^{N} \overline{y}_i^2}}$; where $\overline{X}, \overline{Y} \in \mathbb{R}^{1\times N}$ are the average pooled embedding vectors of dimension N, with the averaging over K vectors derived from K chunks is given by $\overline{x}_i := \frac{1}{K} \sum_{j=1}^{K} x_{ij}$ [13,14].

For encryption, we used the Cheon-Kim-Kim-Song (CKKS) HE [15] scheme, a leveled approximate fully homomorphic encryption (FHE) scheme, implemented in the *TenSEAL* library [12]. A FHE scheme is the most advanced technique that allows both mathematical additions and multiplications for an unlimited number of times. In contrast, the more rudimentary partially homomorphic encryption allows only one type of mathematical operation, either addition or multiplication; and somewhat homomorphic encryption allows both operations but only a limited number of

https://doi.org/10.1017/cts.2026.10800 Published online by Cambridge University Press

times [16]. For simplicity, we refer to our approach as using HE without narrowing down to FHE. The configuration used for the CKKS HE scheme are as follows: the *polynomial modulus degree* was 2^{13}, which determines the ciphertext capacity and thus controls the security level; the *coefficient modulus chain* of [60,40,40,60]*bits* was selected to support the required multiplicative depth while preserving numerical precision; a scaling factor of 2^{40} was used to enable high-precision during encrypted arithmetic. Galois keys were generated to enable vector products in the HE space. The final retrieval rankings were obtained by decrypting the cosine similarity values computed in the CKKS HE space.

The framework is applicable to multiple clinical notes of each individual patient's longitudinal chart. The chunking, embedding, and pooling can process over multiple notes as if they are concatenated into a large single note. In practice, the source notes for use in the concatenation may also be selected by a specific time window or by applying any desirable filter(s). Although the increased number of notes (and therefore chunks) will linearly increase the computing time, each embedding generation (and similarity computation) is independent from each other and thus can be parallelized. Besides, the chunking of clinical notes and subsequent generation of embeddings are both one-time, offline processes that can be pre-computed without adding computational burden at the retrieval time.

Evaluation

Privacy-preserving against inversion attack

For verifying that HE can indeed protect private information compared to raw embeddings, we demonstrated information recovery from unencrypted embedding vectors using the vec2text library [17]. The gtr-t5-base model was first used to generate embedding vectors from excerpts of synthetic clinical notes with the simulated PHI. Subsequently, we used the vec2text's gtr-t5-base invert_embeddings model to invert the embedding vectors back to the original text. For comparison, we next evaluated whether the same vulnerability exists after encryption, by performing the same inversion attack on the embedding vector that had been encrypted by the CKKS HE scheme.

Accuracy of diagnosis retrieval

For quantitative evaluation of retrieval performance, the normalized discounted cumulative gain was used. The nDCG measures how well a diagnostic system puts the correct diagnosis near the top of the retrieved list of diseases. The nDCG can take on values between 0 and 1, where a score of 1 indicates the correct diagnosis is ranked first place in the system's retrieved disease list, and a score of 0 means the retrieved disease list does not contain the correct diagnosis. The exact definition of nDCG is given by: $nDCG := \sum_{i=1}^{p} \frac{2^{rel_i - 1}}{log_2(i+1)}$; where $rel_i \in \{0, 1\}$ is the binary relevance of documents and p the position up to which the metric is computed [18]. As a concrete example, say the system retrieved for a test case of known TN the ranked disease list: 1) oromandibular dystonia, 2) TN, and 3) glossopharyngeal neuralgia; the nDCG for this list would be computed as 0.63, indicating the correct diagnosis is contained in the list but not at the top 1 position.

To assess whether HE alters diagnostic retrieval performance, we performed retrieval on the test data under two conditions: with the query and knowledge base vectors directly without encryption and repeating the same procedure with the vectors encrypted using the CKKS HE scheme. In both settings, the retrieval accuracy was quantified by nDCG on the synthetic test cases. The agreement between the HE and unencrypted retrieval outcomes was compared quantitatively using Kendall's tau rank correlation coefficient τ [19]. The metric τ quantifies the degree to which the disease rankings with and without encryption agree, by measuring the excess of order-preserved over order-destroyed disease pairs. Kendall's tau ranges from -1 (representing complete rank reversal), through 0 (representing no rank order association) to +1 (representing identical ordering), with larger absolute values indicating stronger rank order agreement. On top of these, we also compared multiple LLMs (the Embedding Model in Figure 1) of varying model sizes and embedding dimensions.

Framework robustness under varying parameters

We conducted a series of experiments by varying the embedding and HE parameters to independently assess their impact on the performance of the retrieval framework. Specifically, we varied:

- *Embedding vector dimension*, incremented over 10^1, 10^2, 10^3, and 10^4. A larger vector dimension implies heavier computing.
- *Polynomial modulus degree*, set to $2^{13}, 2^{14}$, or 2^{15}, which represent low, medium, or high security levels respectively. A higher security level increases the dimensionality of the encryption space, making it harder for an adversary to break the encryption computationally.
- *Scaling factor*, set to 2^{20} or 2^{40}, for low or high precision configurations. Higher precision configurations allow for more significant digits of the data to be preserved during encryption, which would retain the finer semantic nuances in the original embeddings.

For each dimension scale, 100 pairs of embedding vectors were created using a random number generator. Subsequently, for each pair of embedding vectors, cosine similarities were computed under a distinct combination of the *polynomial modulus degree* and *scaling factor* values. The first performance benchmark was *percent deviation d*, defined by: $d := 100 \times \frac{|cosim(x_i, y_i) - Dec(cosim(X_i, Y_i))|}{cosim(x_i, y_i)}$, where, the cosine similarity is given by $cosim(x_i, y_i) := x_i^T y_i$; $X_i = Enc(x_i)$; $Y_i = Enc(y_i)$. Here $Enc(\cdot)$ and $Dec(\cdot)$ denote the CKKS encryption and decryption operations respectively. For ease of computations, we use normalized embeddings, x_i and y_i, therefore $||x_i|| = 1$ and $||y_i|| = 1$. The other performance benchmark was *time*, measuring the total time needed to encrypt a vector pair, perform cosine similarity in the HE space, and decrypt the result.

Results

HE blocks the recovery of private information from embeddings

Despite the seeming obfuscation of information offered by projecting text into LLM embedding vectors, we demonstrated that sensitive information can be revealed from unencrypted vectors. As shown in Table 1, applying an inversion model to unencrypted embeddings consistently recovered portions of the original text, including the simulated PHI. In contrast, applying the same inversion model on the encrypted embeddings yielded no meaningful or identifiable text. For example, the synthetic patient name "Arthur Miller" in the first row was recovered from the raw vector by the inversion attack, but it was prevented after the vector first went through the HE encryption.

https://doi.org/10.1017/cts.2026.10800 Published online by Cambridge University Press

Table 1. Vulnerability and prevention of privacy leakage from embedding vectors. Excerpts of synthetic clinical notes, corresponding embedding vectors, and the recovered text are shown in the first three columns respectively. Recovered private information is indicated with bold text. In the fourth and fifth columns, the encrypted embedding vector and the results of a recovery attempt are shown, wherein no meaningful information is revealed. For simplicity, the encrypted embedding vector is displayed as a representative list of floating-point values rather than as the serialized TenSEAL encryption object

Original clinical excerpt	Embedding vector	Recovery from embedding vector	Encrypted embedding vector	Recovery from encrypted vector
S: Subjective Patient: Arthur Miller DOB: 08/15/1970 (55 years old) Chief Complaint: Progressive, painless arm weakness for 6 months.	[−4.07e-02, −7.98e-03, : −5.35e-02]$_{1\times768}$	S″. **Arthur Miller** Subjective DOB: August 5, 1991 Age: 5 years, 5 months, "Performant with weakness and persistent arm pain"	[0.0392, 0.0078, : 0.4235]$_{1\times768}$	no longer exists but is central to the physical and mental health of orangutans who often lived and acted on rapid, painful, toothache laughing
Patient Information Name: Alex Chen DOB: XX/XX/1999 Date of Visit: 04/10/2018 S (Subjective) 19-year-old male presenting with a four-month history of progressive left leg weakness and limited movement.	[−2.23e-02, 1.12e-02, : −8.86e-03]$_{1\times768}$	name: **Alex Chen** DOGS: 26-28-2017 Date: b) persistent weakness in movement, geriatric anomaly, medical diagnosis	[0.5020, 0.0235, : 0.7765]$_{1\times768}$	records are available from a physical injury office where gerland and ve are still living. People often come forward to fight and defend
Clinical Note Patient Name: Robert Miller Date of Visit: August 22, 2020 Age: 84 Subjective Mr.	[−3.37e-02, 1.70e-02, : −2.27e-02]$_{1\times768}$	Age: br> Clinical Note: Subjective Patient Name: **Robert Miller** Dr Date of Visit: August 18, 2018 b	[0.4353, 0.0824, : 0.0275]$_{1\times768}$	but no modern genetic or biochemical grouch exists west of Stinnell. Patients were believed to grow or die from severe heath infection sustained by
Patient Name: Ms. Zara Karim DOB: August 22, 1973 Age: 52 Date: August 22, 2025S (Subjective) A 52-year-old Kurdish female, Ms.	[−4.38e-02, 1.51e-03, : −6.47e-03]$_{1\times768}$	Name: **Ms Zara Karim** (d. 21 August 1973) Patient Date: **52-year-old Kurdish**: 27 July 1973 As	[0.5373, 0.8902, : 0.2824]$_{1\times768}$	clock and the physical and mental protection of older, fun and volatile characters still exist on the North American comic book green light as gutache characters are acting or
S: Subjective Patient Name: Eleanor Vance Age: 71 Chief Complaint: Right facial pain, numbness, and tingling.	[−2.96e-02, 1.15e-03, : −4.85e-02]$_{1\times768}$	S. Name: **Eleanor Vance** Patient **Age 71** Subject: Obligate, subjective complaints: Right facial pain and tingling. Writing	[0.9804, 0.9529, : 0.8549]$_{1\times768}$	Originally, players on softball teams clock drives of green and goldsmith remain reasonably intact from internal game drives of green and

Diagnostic retrieval works in the HE space without performance degradation

The effect of HE over retrieval accuracy is shown in Figure 3, where the nDCG distributions with the Qwen3-8B embedding model are plotted across the five RDs. While some RDs are more challenging (e.g., lower nDCGs in ALS and IPF) to the diagnostic retrieval approach, applying HE had essentially negligible effect on the retrieval accuracy within each RD (each pair of the blue versus orange boxplots). The overlap in nDCG measurements along with rank order correlations quantified by Kendall's tau (τ) consistently equal to 1, indicating that the CKKS HE scheme preserved diagnostic retrieval decisions in the proposed framework.

Preservation of performance in the HE space is independent from the LLM

To test generalizability beyond a single LLM, we repeated the experiments with and without encryption across four other embedding models that represent a wide range of sizes, complexities, input and embedding dimensions (see details of the LLMs in the Supplementary Material Table S1). The consistency in the closely mirroring distributions of nDCG for each model, with τconsistently equal to 1.00 in Figure 4, confirms that across all the tested LLMs, encryption left the retrieval performance almost unaltered. This consistency suggests that HE depends more on the embeddings themselves but not the upstream LLM, so the proposed framework can accommodate evolving LLM architectures and perform as good as the raw embeddings' quality can be. For example, when the raw embeddings yield a leap of nDCG from MiniLM to Qwen3-8B (blue boxplots), that improvement is faithfully reflected in the nDCG by HE-protected embeddings (orange boxplots) with little loss.

The deviation due to encryption can be tuned down with modest computation cost

The effects of the CKKS HE parameters on numerical deviations and computation costs are illustrated in Figure 5. The trends in Figure 5A show that higher security settings increased the deviation introduced by HE on cosine similarity computations by approximately a factor of 10, compared to lower security settings at a fixed precision level. In contrast, increasing the precision, substantially reduced the deviation by a factor of 1,000. Specifically, the high precision plus high security combined settings resulted in deviations less than 10 parts per million for all vector sizes. In Figure 5B, it is observed that the computation times, even for the most expensive settings (high-security and high-precision), consistently remain below 0.5s per cosine similarity computation even for the largest vector pairs tested (a dimension of 10,000). These findings provide evidence that our proposed framework can accommodate a balance in security and accuracy without compromising implementation feasibility.

Discussion

We developed a privacy-preserving information retrieval framework integrating HE and LLMs to differentially rank the top possible diagnoses by screening through a comprehensive knowledge base with thousands of RDs. Our experiments confirmed that HE protects against inversion attacks [17,20] without

https://doi.org/10.1017/cts.2026.10800 Published online by Cambridge University Press

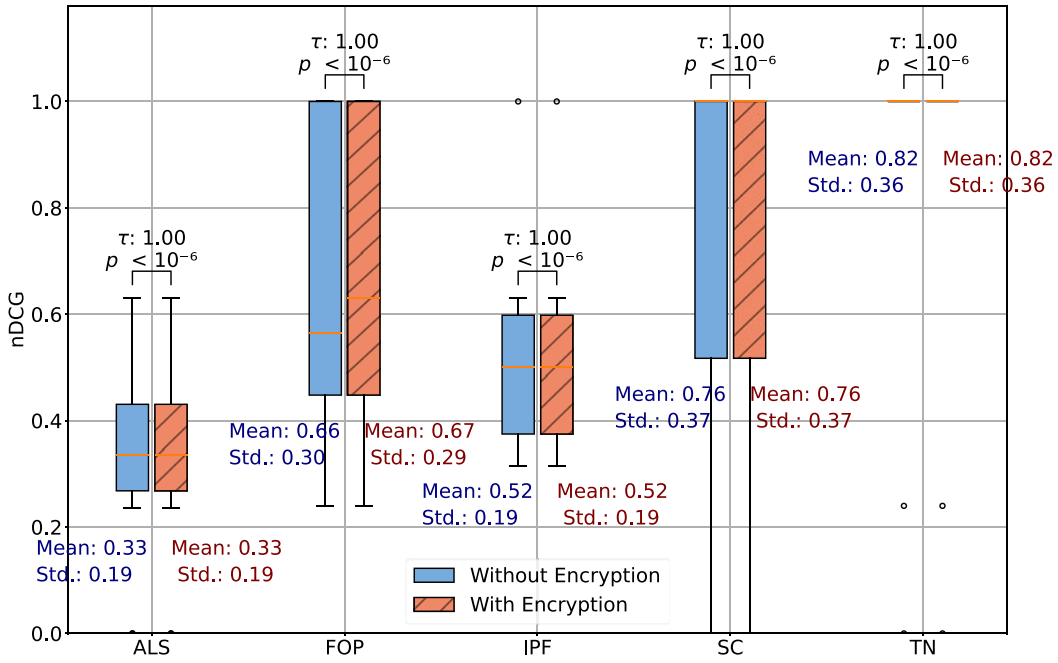


Figure 3. Retrieval performance preserved with homomorphic encryption (HE). Comparative retrieval performances (embedding model: Qwen3-8B) as measured by nDCG are shown with and without HE for the five rare diseases; an nDCG value of 1 indicates perfect retrieval. The lower and upper edges of the box plots represent the 25th and 75th percentiles respectively, the center line the median, and the whiskers the 5th and 95th percentiles. Here, τ is Kendall's tau rank correlation coefficient. The results indicate that the CKKS HE scheme preserves the retrieval performance. Disease abbreviations: ALS = amyotrophic lateral sclerosis, FOP = fibrodysplasia ossificans progressiva, IPF = idiopathic pulmonary fibrosis, SC = sclerosing cholangitis, TN – trigeminal neuralgia.

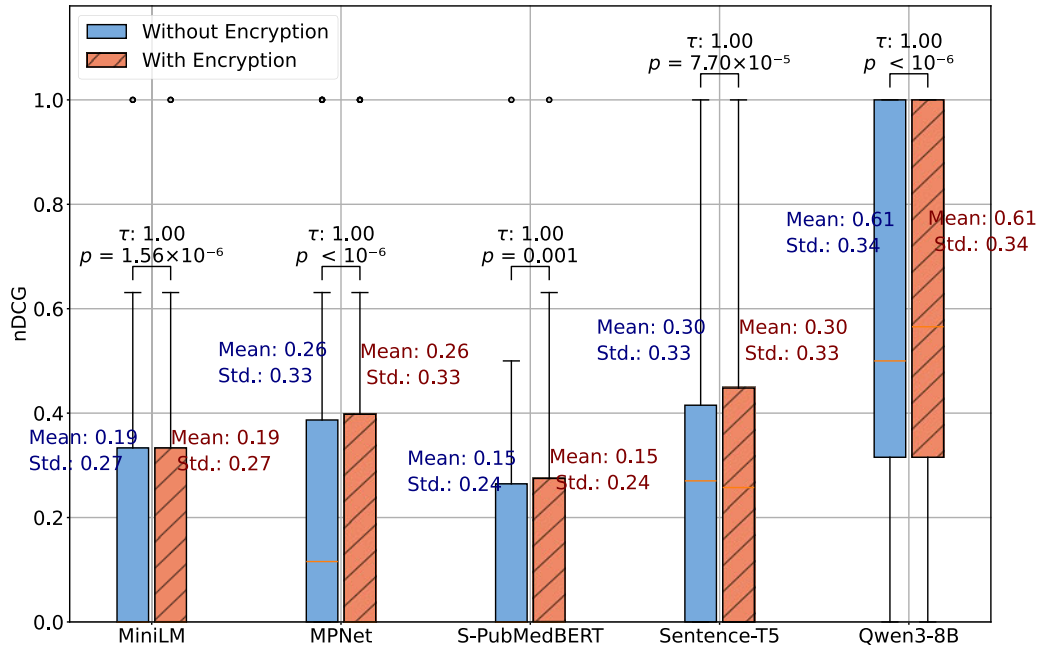


Figure 4. Preservation of retrieval performance under homomorphic encryption (HE) does not rely on a specific large language model (LLM). For each LLM, the retrieval performance is based on pooling the results across all five tested rare diseases. Here, τ is Kendall's tau rank correlation coefficient that measures the diagnostic agreement between the HE and no-HE results. The complexity of the LLMs roughly increment from left to right, e.g., the embedding dimensions range from 384 for the MiniLM model to 4096 for the Qwen3-8B model. Regardless of the LLM used, consistently the HE scheme did not hinder the retrieval performance compared to no encryption.

compromising the rich information in embeddings. The framework works for any LLM as the embedding model while offering tunable parameters to balance the deviation due to encryption, security strength, and computational efficiency. From a scalability perspective, the current implementation should be interpreted as a feasibility prototype rather than an optimized deployment. Since the encrypted similarity computations across disease entries are mutually independent, the scalability would take a big leap by implementing parallelization across a vast number of affordable central processing units or even specialized hardware such as

https://doi.org/10.1017/cts.2026.10800 Published online by Cambridge University Press

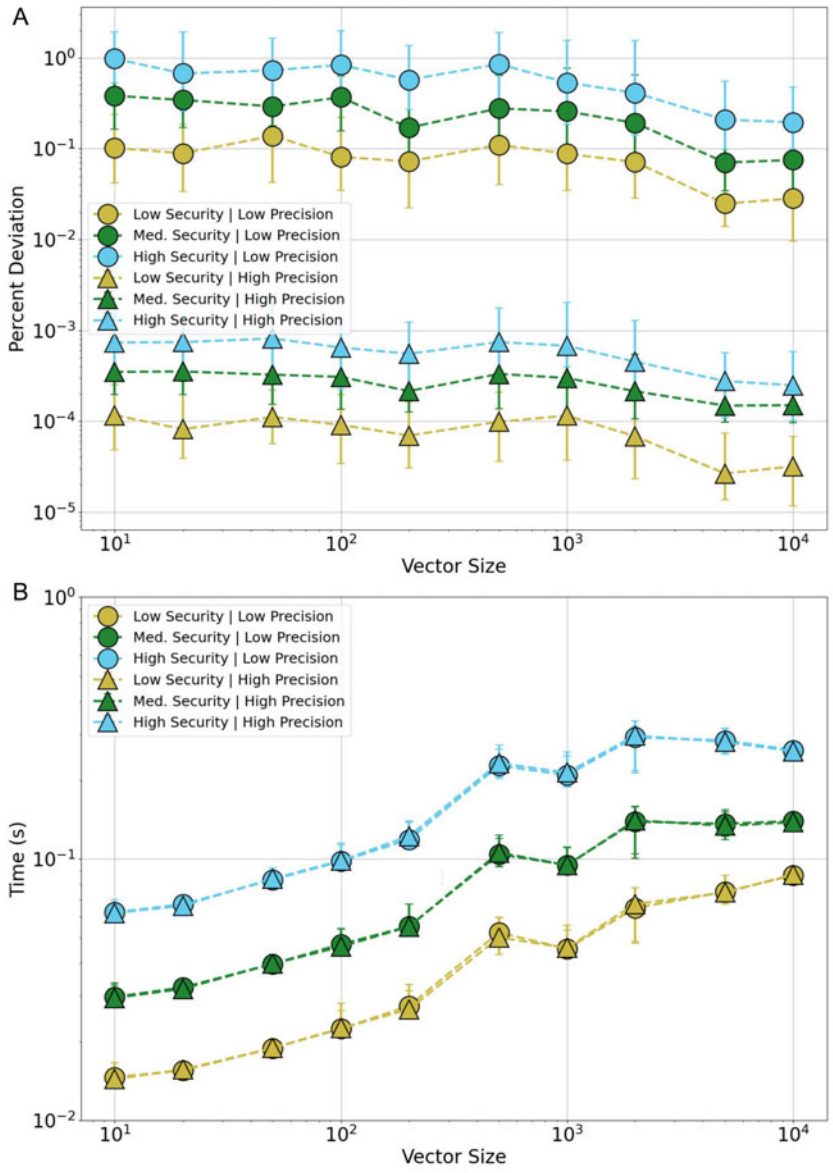


Figure 5. Balancing accuracy, security, and computation time under homomorphic encryption (HE). The effect of the encryption settings (three Security levels and two Precision levels) for similarity-based retrieval is presented. Each data marker represents the median cosine similarity computed across 100 independent simulation runs, wherein, each run used a new and distinct encryption key. The whiskers mark the 25th and 75th percentiles. The Y-axis stands for a different measure in the individual experiment: (A) The deviations introduced by encryption as a function of vector size and CKKS scheme parameters; (B) Computation time per encrypted cosine similarity computation that corresponds to varying vector size and CKKS scheme parameters. Note: both the vertical as well the horizontal axes are shown in logarithmic scale.

graphical processing units [21,22]. Collectively, these merits make it a practical privacy-preserving layer for AI-assisted decision-support solutions. Moreover, the client-server architecture is versatile and can be easily generalized to other applications such as clinical trial criteria matching.

We opted for using only public resources, including the Orphadata disease profiles, to ease the data governance requirements in testing and sharing the technical components for this pilot study. However, it should be noted that one compelling use case of the framework is to further incorporate private information such as real disease cases or expert-curated institutional knowledge that cannot be shared directly. Nonredundant knowledge hosting plus privacy protection would ultimately help achieve our anticipated economy and performance through the distributed framework. In the translational science context, mature implementation of such a framework would encourage serving trustable AI solutions at resource-rich hubs, which could be institutions or federal agencies specialized in a disease area. There is economic benefit in mitigating redundant AI hosting, and the associated reduction of carbon footprint follows as an environmental benefit.

Compared with conventional privacy-enhancing techniques, the proposed framework offers several advantages. First, it eliminates the need for deidentification, a process that remains resource-intensive and technically challenging [23]. In data-rich and networked environments, even datasets deidentified under the HIPAA Safe Harbor standard have been shown to be re-identifiable, indicating the defects of current approaches and the likelihood of increasingly stringent regulation in the future [24]. Second, our approach alleviates the financial burden of performing

https://doi.org/10.1017/cts.2026.10800 Published online by Cambridge University Press

deidentification prior to using external AI and/or hosting expensive AI services on premises, thus freeing up budget for alternative investments. Third, the encrypted retrieval process preserves the raw medically relevant information in its intact form, maintaining diagnostic interpretability (matched information can be decrypted by the user or an authorized party) while safeguarding patient privacy [6].

We identify the following limitations in the pilot study: (1) The CKKS encryption scheme does not preserve ranking order and leaves that to client-side postprocessing after decryption. However, such computational overhead is modest and can be further optimized. (2) The study only used synthetic clinical notes of a few diseases to prove the technical feasibility of integrating HE and LLMs. As the notes and disease profiles were derived from public PubMed case reports and Orphadata, the LLMs might have been exposed to this data during training. Therefore, the absolute retrieval performance may not reflect that if tested on unseen patient data. Future work is needed to evaluate the framework on real patient data, including longitudinal multi-note patient charts, and to assess robustness subject to broader medical conditions and noise such as typos, negations, and incomplete histories. (3) The methodology is demonstrated only for text-based data, and extension to other data modalities, such as images and signals remains to be explored. (4) Our proof-of-concept prototype was implemented using a single-CPU encryption toolbox; the scalability can be enhanced through established parallelization techniques and emerging HE tools that enable parallel operations [21,22]. (5) We did not evaluate the full spectrum of potential threats against a system secured by HE, such as theft of the key, side-channel and fault-injection attacks [25]. A dedicated security-assessment study is required before deployment in real-world clinical environments.

Conclusions

We demonstrate a privacy-preserving retrieval framework that integrates HE with LLMs, using RD diagnoses as example tasks. The HE encryption can protect privacy without hindering diagnostic retrieval performance, exemplifying an overlooked yet effective privacy-preserving technique that can benefit the modern AI implementation stack. Our framework should facilitate the translation of healthcare AI by simultaneously incentivizing centralization (of high-quality services) and democratization (of distributed direct accesses) through effective and affordable privacy protection.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/cts.2026.10800.

Data availability statement. The source codes, instructions, and data presented in this study are made available on GitHub: https://github.com/sivrmn/HE_retrieval

Author contributions. **Sivaraman Rajaganapathy:** Investigation, Methodology, Software, Validation, Visualization, Writing – original draft; **Jennifer St. Sauver:** Writing – review & editing; **Filippo Pinto e Vairo:** Writing – review & editing; **Prasad G. Iyer:** Writing – review & editing; **Hongfang Liu:** Writing – review & editing; **Jungwei W. Fan:** Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing – original draft.

Funding statement. This work was supported by the U.S. National Institutes of Health grants UL1TR002377 and R01HG012748.

Competing interests. All authors declare they have no competing interests.

Data permissions. The Orphadata Science dataset was used in this study in accordance with the terms stipulated by the CC BY 4.0 licence.

References

1. **Jiang JX, Ross JS, Bai G.** Ransomware attacks and data breaches in US health care systems. *Jama Netw Open*. 2025;**8**:e2510180. doi: 10.1001/jamanetworkopen.2025.10180.
2. **Yang G, Cintina I, Pariser A, Oehrlein E, Sullivan J, Kennedy A.** The national economic burden of rare disease in the United States in 2019. *Orphanet J Rare Dis*. 2022;**17**:163. doi: 10.1186/s13023-022-02299-5.
3. **Zelin C, Chung WK, Jeanne M, Zhang G, Weng C.** Rare disease diagnosis using knowledge guided retrieval augmentation for ChatGPT. *J Biomed Inform*. 2024;**157**:104702. doi: 10.1016/j.jbi.2024.104702.
4. **Schieppati A, Henter J-I, Daina E, Aperia A.** Why rare diseases are an important medical and social issue. *Lancet*. 2008;**371**:2039–2041. doi: 10.1016/S0140-6736(08)60872-7.
5. **The Lancet**. Hope for rare diseases. *Lancet*. 2024;**404**:1701. doi: 10.1016/S0140-6736(24)02414-0.
6. **Cho H, Froelicher D, Dokmai N, et al.** Privacy-enhancing technologies in biomedical data science. *Rev Biomed Data Sci*. 2024;**7**:317–343. doi: 10.1146/annurev-biodatasci-120423-120107.
7. **Kovačević A, Bašaragin B, Milošević N, Nenadić G.** De-identification of clinical free text using natural language processing: a systematic review of current approaches. *Artif Intell Med*. 2024;**151**:102845. doi: 10.1016/j.artmed.2024.102845.
8. **Orphadata Science Data Description V2**. *Orpha*net, 2025.
9. **Nguengang Wakap S, Lambert DM, Olry A, et al.** Estimating cumulative point prevalence of rare diseases: Analysis of the Orphanet database. *Eur J Hum Genet*. 2020;**28**:165–173. doi: 10.1038/s41431-019-0508-0.
10. **Reimers N, Gurevych I.** *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. arXiv, 2019.
11. **Pearce PF, Ferguson LA, George GS, Langford CA.** The essential SOAP note in an EHR age. *The Nurse Practitioner*. 2016;**41**:29–36. doi: 10.1097/01.NPR.0000476377.35114.d7.
12. **Benaissa A, Retiat B, Cebere B, Belfedhal AE.** *TenSEAL: A Library for Encrypted Tensor Operations Using Homomorphic Encryption*. arXiv, 2021.
13. **Korenius T, Laurikkala J, Juhola M.** On principal component analysis, cosine and euclidean measures in information retrieval. *Inform Sciences*. 2007;**177**:4893–4905. doi: 10.1016/j.ins.2007.05.027.
14. **Lee S, Lee D, Jang S, Yu H.** Toward Interpretable Semantic Textual Similarity via Optimal Transport-based Contrastive Sentence Learning. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, 2022:5969-5979.
15. **Cheon JH, Kim A, Kim M, Song Y.** *Homomorphic encryption for arithmetic of approximate numbers*. In: Takagi T, Peyrin T, eds. *Advances in Cryptology – ASIACRYPT. 2017*. Cham: Springer International Publishing, 2017:409–437.
16. **Acar A, Aksu H, Uluagac AS, Conti M.** A survey on homomorphic encryption schemes: theory and implementation. *ACM Comput Surv*. 2018;**51**:35. doi: 10.1145/3214303.
17. **Morris JX, Kuleshov V, Shmatikov V, Rush AM.** *Text Embeddings Reveal (Almost) As Much As Text*. arXiv, 2023.
18. **Järvelin K, Kekäläinen J.** Cumulated gain-based evaluation of IR techniques. *ACM Trans Inf Syst*. 2002;**20**:422–446. doi: 10.1145/582415.582418.
19. **Kendall MG.** The treatment of ties in ranking problems. *Biometrika* 1945; **33**: 239–251. doi: 10.2307/2332303.
20. **Li H, Xu M, Song Y.** *Sentence embedding leaks more information than you expect: generative embedding inversion attack to recover the whole sentence*. In: Rogers A, Boyd-Graber J, Okazaki N, eds. *Findings of the Association for Computational Linguistics: ACL 2023*. Toronto, Canada: Association for Computational Linguistics, 2023:14022–14040.
21. **Agulló-Domingo C, Vera-López Ó., Guzelhan S, et al.** FIDESlib: A Fully-Fledged Open-Source FHE Library for Efficient CKKS on GPUs. arXiv, 2025.

22. **Jayashankar S, Kim J, Sullivan MB, Zheng W, Skarlatos D.** A Scalable Multi-GPU Framework for Encrypted Large-Model Inference. arXiv, 2025.
23. **Negash B, Katz A, Neilson CJ, et al.** De-identification of free text data containing personal health information: a scoping review of reviews. *Int J Popul Data Sci*. 2023;**8**:2153. doi: 10.23889/ijpds.v8i1.2153.
24. **Sweeney L, Yoo JS, Perovich L, Boronow KE, Brown P, Brody JG.** Re-identification risks in HIPAA safe harbor data: a study of data from one environmental health study. *Technol Sci*. 2017;**2017**:2017082801.
25. **Lee CH, Lim KH, Eswaran S.** A comprehensive survey on secure healthcare data processing with homomorphic encryption: attacks and defenses. *Discov Public Health*. 2025;**22**:137. doi: 10.1186/s12982-025-00505-w.

https://doi.org/10.1017/cts.2026.10800 Published online by Cambridge University Press