CLIN-SUMM: Temporal Summarization of Longitudinal Clinical Notes

Valentina D'souza[1], Alaleh Azhir[1,2,3], Danielle F. Pace[1], Samuel Friedman[1], Arash Nargesi[1,3], Tristan Naumann[4], Christopher D. Anderson[1,3], Steven J. Atlas[2,5], Judy Hung[2], and Mahnaz Maddah[1]

[1]Broad Institute of MIT and Harvard
[2]Massachusetts General Hospital
[3]Massachusetts General Brigham
[4]Microsoft Research
[5]Harvard Medical School

Abstract

Electronic health records (EHRs) contain years of longitudinal clinical notes that capture evolving patient health, treatments, and outcomes. However, their value is difficult to realize in practice: notes are fragmented across encounters, highly redundant, and inconsistent, leaving temporal patterns underutilized for downstream modeling. We present **CLIN-SUMM** (Clinical Longitudinal Insight from Notes using Summarization), a framework that uses a large language model (LLM) to consolidate multi-visit EHR narratives into a concise, date-partitioned patient summary. **CLIN-SUMM** uses a two-prompt architecture to summarize both the first clinical note and subsequent incremental updates, a redundancy filter to skip near-identical notes, and a sliding-window approach to manage long patient histories. In a dementia prediction case study (N=1,500, 120,213 visit notes), **CLIN-SUMM** achieved up to 85% token reduction while preserving clinical fidelity in an initial clinician review (correctness: 4.84/5, completeness: 4.79/5). Fine-tuning Clinical-ModernBERT on these summaries yielded an Area Under the Receiver Operating Characteristic (AUROC) of 0.79 for classification up to 30 days before diagnosis and 0.75 for prediction three years prior to diagnosis. Analyses of prediction trajectories and saliency patterns revealed clinically meaningful temporal trends and risk markers. **CLIN-SUMM** can help bridge the gap between unstructured multi-visit narratives and predictive modeling by improving efficiency while surfacing rich temporal patterns and early risk signals, enhancing disease modeling, risk prediction, and other longitudinal reasoning tasks.

1 Introduction

Electronic health records (EHRs) contain years of clinical notes documenting a patient's evolving health, treatments, and responses to care. These notes capture details often missing from structured data, such as subtle symptoms, lifestyle factors, and clinician impressions. In clinical practice, providers routinely review EHRs spanning thousands of entries[1], synthesizing diagnoses, treatments, and lifestyle factors to inform care[2]. This longitudinal reasoning is crucial for chronic disease management and risk assessment, but remains challenging due to the volume, fragmentation, and inconsistency of clinical documentation, which is also highly redundant due to inconsistent copy-pasted content. As a result, the rich temporal patterns within clinical narratives remain underutilized for downstream modeling and decision support.

Large language models (LLMs) have advanced rapidly, now supporting longer context windows[3] and demonstrating strong performance across biomedical tasks, including clinical question answering on standardized exams[4], disease risk prediction[5], medication recommendations[6] and automated note generation[7]. They have also been used to summarize clinical notes, reducing cognitive burden for clinicians[8].

However, most LLM applications addressing biomedical text summarization focus on single-visit notes, failing to capture the rich longitudinal context captured across multi-year patient histories. For example, KeBioSum[9] introduced a knowledge-infusion framework integrating multiple pre-trained language models (BERT, RoBERTa, BioBERT, PubMedBERT) to summarize biomedical research articles. Other work[10] proposed a multi-head attention mechanism for extractive summarization of clinical notes, combining token, segment, and positional embeddings to identify salient sentences and visualize attention heatmaps. Similarly,[11] focused on extracting essential information, such as diagnoses and treatment plans, from oncology reports to ease clinicians' review of lengthy records. In general, generating concise, decision-ready summaries that surface evolving clinical context remains limited, as both extractive[10] and abstractive[12] approaches often fail to capture the most relevant updates.

In parallel, structured EHR sequence models (e.g., BEHRT[13], EHRMAMBA[14]) have been developed for longitudinal prediction, but rely primarily on coded data rather than unstructured text. While some approaches combine structured EHR data with clinical notes[15], they typically operate at the note- or visit-level, and do not fully address challenges like redundancy, temporal reasoning, and cross-note context integration.

To address these limitations, we introduce **CLIN-SUMM** (Clinical Longitudinal Insight from Notes using Summarization), a method for generating concise, patient-level summaries from

longitudinal clinical notes. **CLIN-SUMM** consolidates multi-visit narratives into a single, time-partitioned summary stored in a structured Parquet file, which for each patient includes a summary of the first clinical note followed by summaries of the incremental new updates described in subsequent notes. High-quality compression with preserved clinical fidelity benefits LLM-based pipelines, where token efficiency directly affects cost and feasibility. **CLIN-SUMM** supports downstream tasks such as disease classification, risk assessment, phenotype extraction, and clinical question answering.

Key challenges to be addressed include:

1. **Redundancy and inconsistency:** Clinical notes often reuse boilerplate text, vary in terminology, and provide incomplete updates, making it difficult for models to identify relevant changes. Over 50% of text is duplicated across encounters due to copy-paste and template note practices[16], making it difficult for traditional natural language processing (NLP) methods to extract accurate longitudinal signals[17].

2. **Limited temporal awareness:** Generating single-visit summaries with timestamps fails to capture how a patient's clinical state evolves across encounters. Key updates, such as new diagnoses, medication changes, or treatment responses, can be buried in repetitive text. Effective longitudinal summarization must identify novel updates while preserving temporal context, enabling better modeling of disease progression and evolving risk.

3. **Fragmented context:** Models confined to single visits fail to integrate observations across encounters, overlooking important temporal dependencies. While some methods combine multiple notes into a single summary[18], they are typically limited to episode-level summarization (e.g., a single hospitalization) and are not designed for longitudinal patient histories spanning months or years across multiple care settings.

Transforming fragmented and repetitive EHR narratives into coherent, longitudinally growing representations is essential for clinical decision support and downstream machine learning tasks. **CLIN-SUMM** addresses these challenges while improving efficiency by reducing redundant tokens that add little value. **CLIN-SUMM** introduces several methodological innovations:

1. **Two-prompt architecture:** Dedicated initial-summary LLM prompt and a change-focused update prompt to manage evolving patient histories.

2. **Redundancy control:** Jaccard similarity filter to avoid reprocessing near-identical notes.

3. **Sliding window mechanism:** Enables scalability to patients with long encounter histories without exceeding LLM context limits.

4. **Generalizable design:** **CLIN-SUMM** summaries can be applied to a wide range of phenotypes and downstream tasks.

This paper describes the architecture and methodology of **CLIN-SUMM**, and demonstrates its potential to produce accurate, interpretable, task-ready representations of patient trajectories from longitudinal clinical notes. In particular, we evaluated **CLIN-SUMM** by using the generated summaries to fine-tune a Clinical-ModernBERT[19] model for dementia prediction.

2 Methods

In this section, we introduce **CLIN-SUMM**, a framework designed to longitudinally summarize patient visit notes throughout their entire trajectory into a compressed, cohesive, date-partitioned patient-level summary. As illustrated in Figure 1, **CLIN-SUMM** integrates structured prompting and redundancy control to generate high-quality, token-efficient summaries suitable for downstream tasks such as disease classification, risk assessment, phenotype extraction, and clinical question answering. In addition, we describe the dataset and pilot experiments in dementia classification and prediction that evaluate **CLIN-SUMM** outputs and their utility for downstream disease prediction tasks.

2.1 Summarization Workflow

CLIN-SUMM uses two distinct prompts to iteratively summarize longitudinal clinical notes. The initial prompt, Figure 2(a), generates a comprehensive summary of the patient's inaugural entry note into the system, with sections encompassing "Chief Complaints", "History of Present Illness", "Past Medical History", "Medications & Allergies", "Vital signs or Lab/Imaging Findings", "Diagnosis", and "Treatment Plan and Follow-up". The second prompt, Figure 2(b), explicitly incorporates the previously generated summaries and expands the patient-level summary with new information from each subsequent note, focusing exclusively on summarizing only novel content not encompassed within all of the preceding note-level summaries, and appending a "Changes over time" section that summarizes key temporal updates across encounters, such as medication adjustments, new diagnoses, or changes in lab/imaging findings, to track the patient's evolving clinical state. Each summarized entry is timestamped using the original note's date and time, preserving the chronological order of events.

To ensure the growing summary fits within the LLM model's context window and to avoid rate limit concerns, **CLIN-SUMM** integrates a sliding window such that Prompt 2 processes the most recent note summaries to identify novel content, rather than processing the complete history of note summaries. The size of the sliding window parameter can be adjusted for LLMs with different context capacities.

To conserve tokens and Application Programming Interface (API) request charges, **CLIN-SUMM** implements a Jaccard similarity evaluation[20] between consecutive notes. If this similarity is high, the subsequent note is skipped to avoid redundant summarization of copy-pasted content.

CLIN-SUMM output is stored as a structured Parquet file, linking each patient's notes with the note date and time, brief note-level descriptions, and the generated summaries. This format

preserves longitudinal detail while enabling flexible aggregation of summaries by month, year, or clinical event.

2.2 Prompt Optimization and Parameter Selection

Prompts were refined following best practices[21]. For each component, several variants were evaluated. The chosen approach involved guiding the model to produce structured responses and defining the desired behavior when no new content was identified, improving performance over autonomous generation.

While the **CLIN-SUMM** framework is model-agnostic and compatible with any prompt-based LLM API, we used the GPT-4o model[22]. A sliding window of 50 notes was selected for practicality (Appendix A). The Jaccard similarity evaluation used an empirically chosen 95% threshold, which was set conservatively to retain subtle new information (Appendix A).

We also explored the LLMs temperature hyperparameter, which influences its conditional probability distributions during sampling and thus affects the likelihood of less probable token generation (Appendix A). Higher temperatures increase randomness and "creativity," while lower temperatures yield more deterministic outputs. Through an exploration of temperature values, it was empirically determined that 0.2 offered the optimal performance, aligning with the goal of factual summarization and minimizing hallucinations.

2.3 Study Cohort and Data

The EHRs data used for this study was from individuals receiving longitudinal care at Massachusetts General Brigham (MGB) primary care practices between 1976–2023. The dataset included both structured and unstructured data, including patient demographics, diagnosis and procedure codes, medications, health care encounters, and annual ambulatory visit notes. This study was approved by MGH IRB 2023P001077.

The dataset included ambulatory and visit notes from 11,194 patients. This includes a pilot cohort of 1,500 randomly sampled individuals aged 50–75 years, plus 9,694 additional individuals aged 45-85 years. The pilot cohort was used to perform an initial evaluation of the quality and utility of the generated summaries, after which experiments were scaled to the larger population. Subject selection is described in Section 2.5 below.

2.4 Summarization

We assessed the generated summaries from 1500 subjects for clinical validity and measured their compression efficiency.

2.5 Dementia Study

Our primary downstream evaluation was to develop a predictive model leveraging the rich textual data contained in patients' historical clinical notes and summarized in our **CLIN-SUMM** longitudinal dataset for dementia case identification and prognosis. The dementia model was designed for use as an early screening tool and encompassed two complementary tasks.

The first task is to develop a "classification" model, aimed at improving the accuracy of dementia case identification from notes as opposed to International Classification of Diseases (ICD) code assignments. The model was trained on **CLIN-SUMM** summaries up to 30 days before a formal dementia diagnosis, preventing direct mentions of the disease and ensuring that the model learns from preceding indicators rather than explicit diagnostic statements. This classification model is aimed at identifying dementia-consistent language patterns in clinical notes that may be undercoded or missing from structured data, thereby complementing ICD and SNOMED-based definitions by facilitating detection of individuals who may remain undiagnosed in coded records.

The second task is to develop a future risk "prediction" model, supporting opportunities for earlier proactive interventions through preventative measures, lifestyle changes, and therapeutic strategies and potentially transforming how chronic diseases are managed. The future prediction model was trained on data ending three years before diagnosis (or for controls, three years before the most recent available record), to capture subtle early signals that might signal disease onset well before symptoms emerge.

Subjects diagnosed with dementia were identified based on the presence of at least two relevant ICD or Systematized Nomenclature of Medicine (SNOMED) codes recorded on different dates within a 12-month period (Appendix B). An age-matched control group was selected from patients without any dementia-related codes, and the sampling maintained a balanced 50:50 case-to-control distribution for both the pilot cohort of 1,500 subjects and the entire cohort of 11,194 subjects. Importantly, these controls were not healthy. Many had multiple comorbidities and overlapping clinical risk factors, making the task of distinguishing dementia cases from controls more challenging. Detailed demographic characteristics of this study group are provided in Appendix B.

We fine-tuned the Clinical-ModernBERT[19] model, creating two distinct instances for the classification and future disease prediction tasks. The pilot cohort of 1,500 **CLIN-SUMM** patient-level summaries was stratified at the patient level into training (N=975, 65%), validation (N=375, 25%), and test (N=150, 10%) sets. Notes were preprocessed by removing punctuation, special characters, and stop words; the cleaned text was then tokenized and converted to tensors. The Clinical-ModernBERT model accepted sequences up to 8,192 tokens; for longer summaries, the most recent summaries were used. The Clinical-ModernBERT models were trained with a learning rate of 5×10^{-6} for up to 60 epochs, with a scheduler reducing the rate by 0.5 if validation loss plateaued for five epochs, and early stopping triggered after 15 stagnant epochs. The best-performing checkpoint was selected as the final model.

3 Results

We evaluated **CLIN-SUMM** for its ability to compress longitudinal clinical notes, maintain clinical fidelity, and support downstream dementia risk prediction. Each experiment used either the pilot test set (N=150) or an expanded test set that also included

the 9,694 additional subjects (N=9,844).

3.1 Summarization

The summarization experiments described below confirm that **CLIN-SUMM** produces concise, accurate longitudinal summaries suitable for downstream predictive modeling.

3.1.1 Compression

Summarization substantially compressed the clinical narratives in the pilot test dataset, reducing 140,587,252 words (182,763,428 tokens) to 20,585,169 words (26,760,721 tokens), yielding an overall compression rate of 85.3%.

While most summaries were shorter than their source notes, about 10% were longer. Manual review showed this typically occurred when notes had minimal text, in such cases, the summaries often included inferred clinical context based on prior encounters, and prior summaries' information, such as test orders, medication adjustments, or follow-up treatments.

To further characterize summarization behavior, we analyzed the distribution of compression levels across all notes (Figure 3). At the patient level, the mean compression rate was 79.2 ± 13.3%, indicating consistent summarization performance across patients. While the overall note-level compression was 85.3%, per-patient averages provide a less biased estimate by reducing the impact of patients with disproportionately many or lengthy notes. Additional results are shown in Appendix C.

3.1.2 Clinical Validation

A clinician manually validated 30 randomly sampled patient-level **CLIN-SUMM** summaries from the pilot test dataset, encompassing 698 notes. Two aspects were evaluated: completeness, assessing the level of information coverage from the notes and identifying any omissions; and correctness, which verifies if the summary fully aligns with the key points in the notes without factual errors or misinterpretations. A second clinician independently assessed 135 note-based summaries from five patients, corresponding to the same subset previously reviewed by the original clinician. Both clinicians rated each summary for correctness and completeness using a 5-point Likert scale (1 = poor, 5 = excellent), based on predefined questions. The validation questionnaire is included in Appendix D.

The evaluation yielded average scores of 4.79 for completeness and 4.84 for correctness. For correctness, the inter-rater agreement was 98.80% [95% CI 97.30 - 100], with Cohen's $\kappa = 0.77$; for completeness, the inter-rater agreement was 97.30% [95% CI 94.5 - 99.8], with Cohen's $\kappa = 0.81$, indicating substantial reliability between the two reviewers. Fewer than 4% of summaries exhibited hallucinations, mostly involving: (1) inferring diagnoses that were not explicitly documented based on presenting symptoms, (2) misinterpreting vitals or laboratory values due to inconsistent formatting, and (3) incorrectly attributing lack of treatment to unavailability (examples in Appendix D). Although this rate is low, we plan to conduct systematic error analyses and develop mitigation strategies to ensure the safe and reliable integration of **CLIN-SUMM** into clinical practice. While we plan a larger validation with multiple reviewers, these initial results indicate that **CLIN-SUMM** shows promise in capturing key information, reducing redundancy, and achieving substantial compression.

3.2 Dementia Classification and Prediction

We next evaluated how **CLIN-SUMM** summaries perform in our primary downstream demonstration of dementia risk classification and prediction.

3.2.1 Model performance

The predictive models' ability to use **CLIN-SUMM** patient-level summaries to identify dementia cases under the classification and future disease prediction scenarios was evaluated in the entire dementia cohort and summarized in Table 1. In the classification task, performance was strong (Area Under the Receiver Operating Characteristic (AUROC) 0.79, Area under the Precision-Recall Curve (AUPRC) 0.78, F1 score 0.73, precision 0.71, recall 0.76), reflecting high accuracy and a balanced trade-off between sensitivity and specificity. For future risk prediction, where the goal was to identify individuals who would be diagnosed with dementia at least three years later, performance was slightly lower but still robust (AUROC 0.75, AUPRC 0.73, F1 score 0.65, precision 0.71, recall 0.60). This drop was expected, as early indicators of future dementia diagnosis are more subtle, yet the model retained substantial discriminatory power. For context, a baseline logistic regression using only age and sex achieved an AUROC of 0.61 for three-year dementia prediction, underscoring the value of leveraging longitudinal clinical notes, which capture evolving symptoms, diagnoses, and treatments over time. Furthermore, the 3-year predictive performance (AUROC 0.75) compares favorably with multifactorial dementia risk models[23], which report AUROC values between 0.59 and 0.73 for 10-year prediction based on CAIDE[24] and Australian National University Alzheimer Disease Risk Index[25] scores.

For both the classification and prediction tasks, logistic baselines incorporating age, sex, and prevalent comorbidities (anemia, arthritis, depression, delirium, osteoporosis, dizziness, syncope, hyperlipidemia, dyspnea, Parkinson's disease, hearing and visual impairment, anorexia, anxiety, alcohol use, hypertension and falls, based on ICD codes) showed limited predictive ability, reflecting the insufficiency of structured risk factors alone (Table 1). Adding vitals (BMI, cholesterol and blood pressure) modestly improved AUROC but reduced the sample size due to variable missingness. Clinical-ModernBERT models trained on **CLIN-SUMM** summaries consistently outperformed all baselines, indicating that detailed longitudinal narratives contributes complementary predictive signals beyond structured covariates for dementia case identification and risk prediction.

While the classification model was trained on data up to 30 days before diagnosis and the prediction model was designed for risk prediction three years before diagnosis, we also evaluated the ability of the classification model to generalize as a prediction tool by applying it to test sets truncated at progressively earlier time points (6 months, 1, 2, 3, 4, and 5 years before diagnosis). We limited

the analysis to 5 years due to reduced data availability and smaller sample sizes beyond this point. Detailed evaluation metrics are reported in Table 1. Performance declined as the prediction window moved further from diagnosis (AUROC decreased from 0.79 at 30 days to 0.78 at 1 year, 0.77 at 3 years, and 0.73 at 5 years). The observed decline in performance with longer prediction windows is expected and likely multifactorial. Earlier records likely contain fewer explicit disease-related signals, with less frequent or subtler cognitive or neurological indicators, making dementia onset inherently harder to predict several years in advance. Additionally, summaries of earlier records may emphasize general medical history over cognitive or behavioral changes, which can lead to less informative embeddings. Despite the expected decline, the model maintained meaningful discriminatory power several years before diagnosis, suggesting potential use for risk stratification. Notably, the 30-day classification model achieved an AUROC of 0.77 on data truncated at 3 years, closely matching the future prediction model and demonstrating consistency between modeling strategies.

3.2.2 Longitudinal evaluation

In practice, **CLIN-SUMM** summaries can be appended in real-time as a patient's case evolves, and the accuracies of the dementia classification and prediction models are expected to change over time as a formal dementia diagnosis nears.

We assessed the performance of both models across different time windows to evaluate their ability to capture longitudinal patterns in dementia risk. Both models were applied to 3-year windows of data, starting from the first note and extending either to 30 days (classification model) or 3 years (prediction model) before diagnosis. Although the models were trained on the last subset of tokens that could fit within Clinical-ModernBERT's context window, we assessed whether they could generalize when applied to all available **CLIN-SUMM** summaries for each patient, such that together, these analyses used the entire available sequence of longitudinal summaries.

Heatmaps of classification and prediction scores (Figure 4) revealed clear temporal trends in dementia risk across 3-year windows (see Appendix E for window choice rationale). Most dementia cases had consistently high scores, confirming the models' ability to detect clinically relevant patterns even years before diagnosis, whereas most controls received low scores. Some high-risk predictions among controls appeared to correspond to age-related conditions (e.g., cervical spine issues, mobility limitations, hearing loss), suggesting partial overlap between general frailty language and dementia-related documentation. Interestingly, the first time window often produced the highest dementia risk scores, likely because initial summaries include comprehensive past medical histories, while later notes reflecting symptom improvement or lack of additional symptoms often corresponded to reduced scores over time. These temporal patterns suggest that the model's predictions are grounded in clinically meaningful cues that evolve over time.

We additionally assessed the longitudinal behavior of the classification and prediction model by applying it to 1-year cumulative windows, mirroring how risk estimation could be performed during annual clinical checkups. In this setup, each yearly window contains all **CLIN-SUMM** summaries of notes recorded from the beginning of a patient's history up until that year. The first window therefore reflects only the earliest available documentation, while each subsequent window adds the new year's notes to everything that came before. As a result, later predictions are made using a progressively richer clinical history, allowing us to evaluate whether the model's dementia risk estimates remain clinically meaningful when applied in a manner consistent with how new information is clinically accumulated. To characterize these temporal patterns, we plotted the mean predicted dementia risk across all patients for each cumulative window and fitted a linear trend line (Figure 5). The trajectories for dementia and control groups diverged sharply: dementia cases showed consistently higher mean scores that often remained elevated over time, while scores for controls generally stayed low or showed a decreasing trend.

3.2.3 Saliency

To further validate model performance and understand how the classification model identifies linguistic signals associated with dementia risk, we conducted a saliency analysis using Local Interpretable Model-agnostic Explanations (LIME)[26]. LIME perturbs segments of the **CLIN-SUMM** summaries and measures the impact on predicted risk, enabling us to identify words and phrases most influential in the model's decisions. At the global level, we aggregated feature importance scores across the pilot test set, identifying high-risk terms (e.g., "difficulties", "executive", "vision", "dysfunction") and protective terms (e.g., "screening", "monitoring", "exercise"), visualized in a word cloud (Figure 6). Encouragingly, the salient terms frequently identified (e.g., vision loss, memory difficulties, cognitive decline) correspond closely with symptoms consistent with early indicators of dementia. At the patient level, LIME highlighted specific sentences within **CLIN-SUMM** summaries that contributed most to predicted risk (Appendix F), providing examples of how these terms appear in real narratives. These results confirms that the model leverages clinically meaningful cues rather than spurious associations and demonstrate how interpretable outputs can help identify potential early indicators for dementia.

3.3 Additional Results

While GPT-4o was the LLM used for summarization in this study, scaling to larger cohorts using proprietary APIs remains cost-prohibitive. Open-source LLMs such as Qwen3[27] and GPT-OSS[28] may enable more scalable summarization. Appendix G discusses our preliminary results using Qwen3 experimenting with and optimizing open-source LLMs.

Additional pilot applications of **CLIN-SUMM** are presented in Appendix H.

4 Discussion

In this study, we developed **CLIN-SUMM**, a longitudinal summarization framework for condensing multi-visit EHR notes into

concise, date-partitioned summaries well-suited for downstream machine learning tasks. By reducing the total token count and emphasizing what is new in each note, **CLIN-SUMM** summaries improve computational efficiency and speeds processing for downstream tasks operating on longitudinal clinical notes. Clinical validation demonstrated high completeness and factual accuracy while achieving substantial token compression. Applied to dementia, **CLIN-SUMM** summaries supported robust classification performance (AUROC of 0.79) and meaningful future risk prediction up to 3 years before diagnosis (AUROC of 0.75), highlighting their value for both immediate and early screening applications.

By preserving temporal context while removing redundancy, **CLIN-SUMM** may help surface preclinical indicators of disease (e.g., dementia) that might otherwise be obscured in raw notes. Our findings indicate that summarization may serve as an intermediate representation for temporal reasoning tasks. While dementia prediction served as a case study, once computed **CLIN-SUMM** summaries are reusable and broadly applicable to other conditions requiring longitudinal reasoning. Even as longer LLM context windows reduce the need for compression, **CLIN-SUMM** shows how rich but underutilized longitudinal information can be unlocked for disease risk prediction.

Responsible use of large language models in healthcare requires attention to safety, transparency, and human oversight. Although **CLIN-SUMM** demonstrated low hallucination rates, even rare factual errors can have clinical implications. To mitigate this risk, we implement prompt-level safeguards that constrain outputs to verifiable information and incorporate clinician review during evaluation. Because documentation styles and EHR systems evolve over time, we also anticipate the potential for model drift and plan periodic auditing and prompt recalibration to maintain accuracy. Importantly, **CLIN-SUMM** is designed to support, not replace, clinical decision-making. Any prospective deployment should include clear disclosure of model provenance, ongoing monitoring for bias and fairness, and mechanisms for clinician feedback to ensure safe and equitable use.

Several directions remain for future work. Our experiments thus far aimed to validate the feasibility of **CLIN-SUMM** summaries as predictive inputs. Scaling beyond our initial dementia prediction application will require additional work evaluating open-source LLMs to improve cost-effectiveness, and further optimizing the redundancy filtering and windowing strategies for efficiency. Comparison to additional baselines (e.g., Clinical-ModernBERT on deduplicated notes or models such as BEHRT[13] and EHRMamba[14]) will be valuable. For dementia prediction,future work on bias correction and comorbidity adjustment is needed to improve generalizability across heterogeneous older populations. Finally, broader validation, systematic error analysis, and applying **CLIN-SUMM** across additional diseases, healthcare systems, and clinical decision support tools represent important next steps for assessing generalizability, robustness, fairness and real-world integration.

References

[1] Nathalie Huguet, Jorge Kaufmann, Jean O'Malley, Heather Angier, Megan Hoopes, Jennifer E. DeVoe, and Miguel Marino. Using electronic health records in longitudinal studies: Estimating patient attrition. *Medical Care*, 58(Suppl 6 Suppl 1):S46–S52, June 2020.

[2] Michael Wornow, Suhana Bedi, Miguel Angel Fuentes Hernandez, Ethan Steinberg, Jason Alan Fries, Christopher Ré, Sanmi Koyejo, and Nigam H. Shah. Context clues: Evaluating long context models for clinical prediction tasks on ehrs. *arXiv preprint arXiv:2412.16178*, December 2024.

[3] OpenAI. New models and developer products announced at DevDay. `https://openai.com/index/new-models-and-developer-products-announced-at-devday/`, 2023. [Online; accessed August 2025].

[4] Karan Singhal, Shekoofeh Azizi, Tao Tu, and et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, July 2023.

[5] Jonathan W. Cunningham, Pulkit Singh, Christopher Reeder, and et al. Natural language processing for adjudication of heart failure in the electronic health record. *Heart Failure (JACC Heart Failure)*, 11(7):852–854, 2023.

[6] Yoshitaka Inoue, Tianci Song, and Tianfan Fu. Drugagent: Explainable drug repurposing agent with large language model–based reasoning. *arXiv preprint arXiv:2408.13378*, August 2024.

[7] Tiffany I. Leung, Andrew J. Coristine, and Arriel Benis. Ai scribes in health care: Balancing transformative potential with responsible integration. *JMIR Medical Informatics*, 13:e80898, August 2025.

[8] Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, Nidhi Rohatgi, Poonam Hosamani, William Collins, Neera Ahuja, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, John Pauly, and Akshay S. Chaudhari. Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine*, 30:1134–1142, February 2024.

[9] Qianqian Xie, Jennifer Amy Bishop, Prayag Tiwari, and Sophia Ananiadou. Pre-trained language models with domain knowledge for biomedical extractive summarization. *Knowledge-Based Systems*, 252:109460, September 2022.

[10] Neel Kanwal and Giuseppe Rizzo. Attention-based clinical note summarization. *CoRR*, abs/2104.08942, 2021.

[11] C. Kalaiselvi, T. N. Vanitha, C. and R. Vinodavarshini. Nlp-powered oncology patient summary. In *Proceedings of the 7th International Conference on Electronics, Communication and Aerospace Technology (ICECA 2023)*, pages 253–258, 2023.

[12] Yuhao Zhang, Daisy Yi Ding, Tianpei Qian, Christopher D. Manning, and Curtis P. Langlotz. Learning to summarize radiology findings. *arXiv preprint arXiv:1809.04698*, September 2018.

[13] Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. BEHRT: Transformer for electronic health records. *Scientific Reports*, 10:7155, April 2020.

[14] Adibvafa Fallahpour, Mahshid Alinoori, Wenqian Ye, Xu Cao, Arash Afkanpour, and Amrit Krishnan. Ehrmamba: Towards generalizable and scalable foundation models for electronic health records. *arXiv preprint arXiv:2405.14567*, 2024.

[15] Weimin Lyu, Xinyu Dong, Rachel Wong, Songzhu Zheng, Kayley Abell-Hart, Fusheng Wang, and Chao Chen. A multimodal transformer: Fusing clinical notes with structured ehr data for interpretable in-hospital mortality prediction. In *Proceedings of the AMIA Annual Symposium*, pages 719–728, 2023.

[16] Jackson Steinkamp, Jacob J. Kantrowitz, and Subha Airan-Javia. Prevalence and sources of duplicate information in the electronic medical record. *JAMA Network Open*, 5(9):e2233348, September 2022.

[17] Elias Hossain, Rajib Rana, Niall Higgins, Jeffrey Soar, Prabal Datta Barua, Anthony R. Pisani, and Kathryn Turner. Natural language processing in electronic health records in relation to healthcare decision-making: A systematic review. *Computers in Biology and Medicine*, 155:106649, March 2023.

[18] Griffin Adams, Emily Alsentzer, Mert Ketenci, Jason Zucker, and Noémie Elhadad. What's in a summary? laying the groundwork for advances in hospital-course summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4794–4811. Association for Computational Linguistics, June 2021.

[19] Simon A. Lee, Anthony Wu, and Jeffrey N. Chiang. Clinical modernbert: An efficient and long context encoder for biomedical text. *arXiv preprint arXiv:2504.03964*, April 2025.

[20] Wikipedia contributors. Jaccard index — Wikipedia, The Free Encyclopedia. `https://en.wikipedia.org/wiki/Jaccard_index`. [Online; accessed October 2024].

[21] OpenAI. Best practices for prompt engineering with OpenAI API. `https://help.openai.com/en/articles/6654000-best-practices-for-prompt-engineering-with-openai-api`, 2023. [Online; accessed October 2024].

[22] Hurst, A., Lerer, A., Goucher, A. P., and et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, October 2024.

[23] Mika Kivimäki, Gill Livingston, Archana Singh-Manoux, Nina Mars, Joni V. Lindbohm, Jaana Pentti, Solja T. Nyberg, Matti Pirinen, Emma L. Anderson, Aroon D. Hingorani, Pyry N. Sipilä, and et al. Estimating dementia risk using multifactorial prediction models. *JAMA Network Open*, 6(6):e2318132, June 2023.

[24] Shireen Sindi, Elisabeth Calov, Jasmine Fokkens, Tiia Ngandu, Hilkka Soininen, Jaakko Tuomilehto, and Miia Kivipelto. The caide dementia risk score app: The development of an evidence-based mobile application to predict the risk of dementia. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 1(3):328–333, July 2015.

[25] Kaarin J. Anstey, Nicolas Cherbuin, Pushpani M. Herath, Chengxuan Qiu, Lewis H. Kuller, Oscar L. Lopez, Robert S. Wilson, and Laura Fratiglioni. A self-report risk index to predict occurrence of dementia in three independent cohorts of older adults: the anu-adri. *PLoS One*, 9(1):e86141, January 2014.

[26] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, August 2016.

[27] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, May 2025.

[28] OpenAI. gpt-oss-120b & gpt-oss-20b model card, 2025.

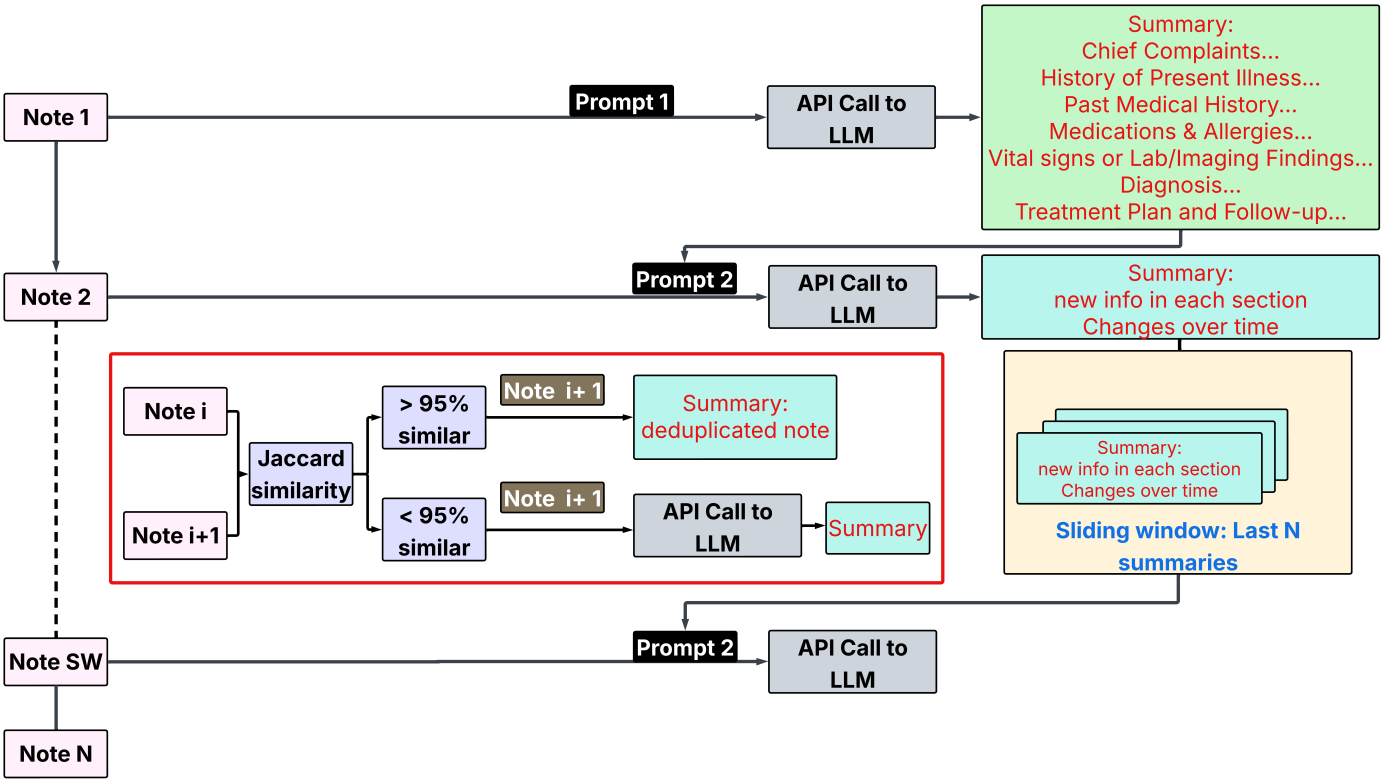

Figure 1: Overview of the **CLIN-SUMM** framework. Patient visit notes are processed using LLM-powered summarization prompts, redundancy filtering, and sliding-window management to generate a longitudinal, date-partitioned clinical summary.

(a) Prompt 1: Summarizing the first note

Prompt template for summarizing first note of a patient

```
Medical visit note (Timestamp: {note.name}):{note['text']}

Please generate a structured, comprehensive summary of the above medical note in an
abstractive manner, ensuring that the key details are captured and presented clearly and
redundant repetition is avoided

Organize the summary using the following sections:
**Chief Complaints**: [Summarize the primary concerns or symptoms reported by the patient]
**History of Present Illness**: [Provide a narrative of medical history for the symptoms
mentioned today, their onset, progression, associated factors, and any relevant context]
**Past Medical History**: [Summarize relevant past diagnoses, surgeries, or chronic
conditions]
"**Medications & Allergies**: [List current medications, dosages, and any known drug or food
allergies]
**Vital signs or Lab/Imaging Findings**: [Summarize vital sign measurements, lab results or
imaging results if any or respond 'nothing observed']
**Diagnosis**: [Summarize diagnoses made]
**Treatment Plan and Follow-up**: [Summarize the treatment plan including medications,
interventions, or lifestyle recommendations and outline if any follow-ups, referrals, or
additional tests were requested]

Ensure the summary is coherent, natural, and does not merely extract sentences from the text
but instead synthesizes the information into a well-structured report
```

(b) Prompt 2: Longitudinal update with changes

Prompt template for summarizing subsequent notes of a patient

```
Previous summary before (Timestamp: {current_note.name}):{previous_summary}
New visit note (Timestamp: {current_note.name}):{current_note['text']}

Identify any **new information** in the current visit that is not already included in the
past visit summaries
Update the relevant sections with new details while avoiding repetition
If a section has no new information, skip updating it

Format the updated summary with the following sections and definitions:
**Chief Complaints**:.................................
**History of Present Illness**: ......................................
.......................................
.......................................
**Diagnosis**: .............not mentioned before]".........
**Treatment Plan and Follow-up**: ......................................
**Changes Over Time**: [Outline changes observed over time if any or respond 'nothing
observed']
**Ensure the summary is coherent, natural, and does not merely extract sentences from the
text but instead synthesizes the information into a well-structured report…
```

Figure 2: Two-step summarization: (a) Initial note summary; (b) Incremental update with summarization of novel content and "changes over time" (optimized for GPT-4o).

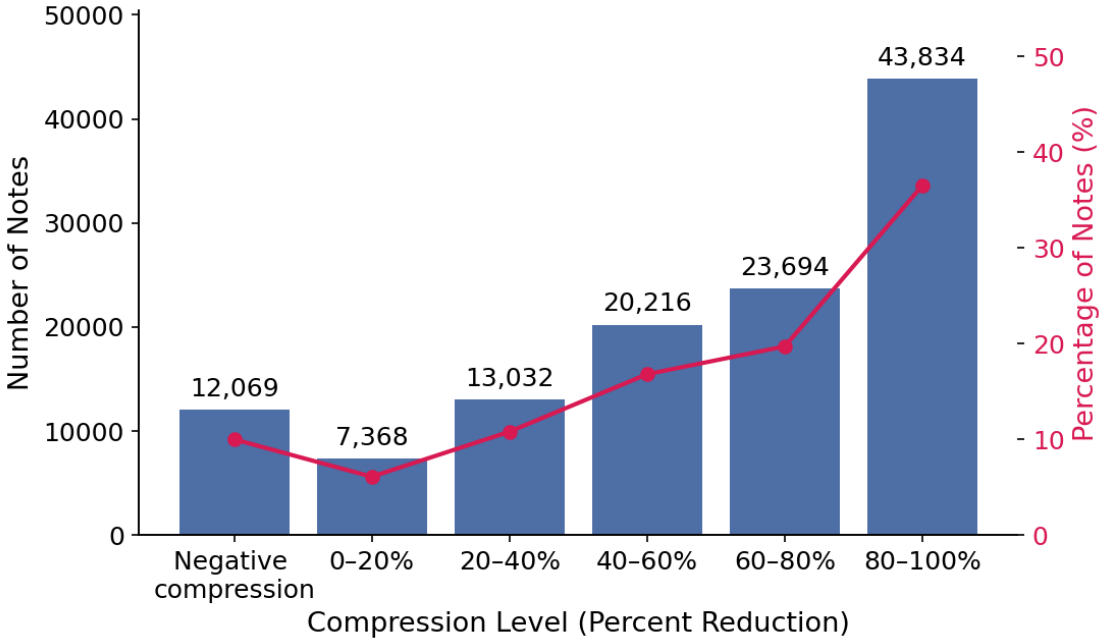

Figure 3: Distribution of summarization compression across 120,213 notes from 1,500 patients in the pilot cohort (percent reduction from note to summary).

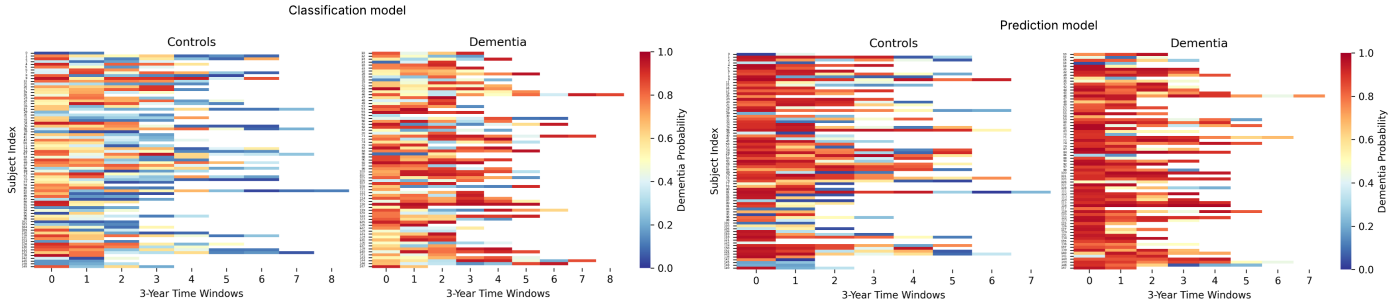

Figure 4: Temporal heatmaps of dementia classification and prediction probabilities across 3-year windows in the pilot test cohort. Each row represents a patient, and each column a longitudinal time window. Warmer colors indicate higher predicted risk. For dementia cases, the horizontal bars stop at 30 days before diagnosis for the classification model and 3 years before diagnosis for the prediction model. For control cases, the entire history is used.

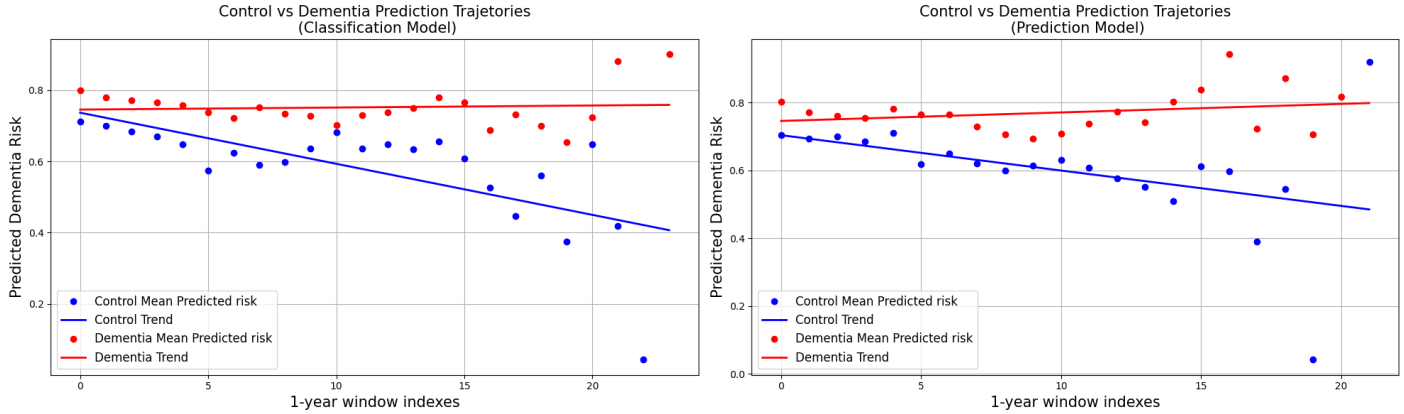

Figure 5: Longitudinal dementia risk trajectories generated by the classification (left) and prediction (right) models when applied to 1-year cumulative windows. Each point represents the mean predicted risk for dementia and control groups at a given yearly window, and trend lines illustrate how risk estimates evolve as the model incorporates progressively more patient history.

Figure 6: Word clouds generated using LIME in the pilot test set show the most influential terms contributing to dementia risk in the classification model. Left: high-risk terms; Right: protective terms.

Table 1: Performance of the Clinical-ModernBERT models using **CLIN-SUMM** summaries for (1) dementia classification, (2) future risk prediction, and (3) baseline comparisons including age–sex, age–sex–comorbidities, and age–sex–comorbidities–vitals. Time-horizon evaluation using the 30-day classification model applied to earlier cutoffs is reported separately below.

Metric	**Clinical ModernBERT**		**Age–Sex**	**Age–Sex–Comorbidities**		**Age–Sex–Comorbidities–Vitals**	
	Classification	**Prediction**		**Classification**	**Prediction**	**Classification**	**Prediction**
AUROC	0.79	0.75	0.61	0.68	0.61	0.75	0.70
AUPRC	0.78	0.73	0.57	0.70	0.63	0.73	0.65
F1 Score	0.73	0.65	0.61	0.59	0.52	0.55	0.35
Precision	0.71	0.71	0.61	0.68	0.63	0.72	0.63
Recall	0.76	0.60	0.62	0.52	0.44	0.44	0.24

Time Horizon Before Diagnosis (Classification Model Applied to Earlier Cutoffs)					
Metric	**30 days**	**6 months**	**1 year**	**3 years**	**5 years**
AUROC	0.79	0.78	0.78	0.77	0.73
AUPRC	0.78	0.77	0.76	0.74	0.69
F1 Score	0.73	0.73	0.73	0.72	0.70
Precision	0.71	0.70	0.69	0.66	0.59
Recall	0.76	0.76	0.76	0.81	0.86

Appendix

A Parameter Selection

A.1 Sliding window

Although all of our requests were well below the GPT-4o context limit (max $\approx$ 72K tokens even with 70 prior summaries), we encountered intermittent rate-limit and timeout errors during summarization. These issues stemmed not from token overflows but from API-level constraints, including transient throughput throttling, payload-size serialization overhead, and server-side latency caps on long requests. Increasing the window beyond 50 summaries consistently led to slower response times and occasional failed generations, despite remaining under the token limit. The window size of 50 was therefore chosen as the largest configuration that maintained stable API performance while preserving longitudinal context across encounters. In addition to technical stability, the fixed window size of 50 was chosen to ensure methodological consistency across patients. Because note volume varies widely by patient (some with fewer than 20 encounters, others exceeding 100), using a standardized context size avoids biasing the model toward patients with longer documentation histories. Finally, this choice also supports future scalability: as our pipeline can also be used using open-source or smaller-context models, maintaining a comparable 50-summary window facilitates seamless adaptation to models with tighter context limits while ensuring reproducibility across implementations.

A.2 Jaccard Similarity

To empirically select a similarity threshold for redundancy filtering, we computed entity-level recall (i.e., proportion of retained clinical entities) between consecutive notes as a function of Jaccard distance (Table A1). We observed that entity recall remained above 94% for note pairs with Jaccard distance ≤ 0.05 ($\approx$ 95% similarity), but dropped sharply beyond that range (Table A1). This inflection point indicates that notes exceeding this distance begin to introduce new clinical entities or contextual updates rather than templated repetition. Accordingly, we set a 0.05 distance threshold (95% similarity) to discard near-duplicate notes while preserving semantically distinct updates, minimizing information loss to < 6% on average.

A.3 Temperature

Lower-temperature decoding (< 0.3) reduces randomness in token sampling, leading to more deterministic and faithful outputs, an essential aspect when summarizing sensitive medical content. In exploratory runs, higher values (> 0.3) produced summaries that were longer but less stable, occasionally introducing mild factual drift (e.g., adding diagnosis from clinical conditions not explicitly mentioned in the source notes). Conversely, lower values (< 0.2) yielded overly compressed outputs that omitted secondary but clinically relevant information such as medication adjustments or lab trends. A temperature of 0.2 provided the best qualitative balance between completeness and factual precision, generating concise and stable summaries while maintaining high fidelity to the source notes. Although we did not conduct a full parameter sweep due to time constraints, this setting aligned with the goal of factual summarization and minimizing hallucinations in clinical text.

B Dementia study

B.1 ICD/snomed dementia codes

The codes used for dementia phenotyping are described in Table A2 and Table A3.

B.2 Demographic characteristics table

The characteristics of the dementia study cohort are described in Table A4.

C Compression

Figure A1 illustrates how the average compression rate evolves as additional notes are summarized for patients. On average, early notes show lower compression due to higher content density, whereas later notes exhibit greater compression as templated text and repeated information accumulate. Overall, **CLIN-SUMM** achieves consistently high compression rates while preserving clinical fidelity, demonstrating its scalability for processing large volumes of longitudinal notes.

D Clinical validation of summaries

D.1 Template questions

A clinician assessed **CLIN-SUMM** summaries using structured yes/no questions and a five-point scale (1 = poor, 5 = excellent) for correctness and completeness. The validation questionnaire included:

1. Does the summary include all required sections if related information was present in the note? (Completeness: 1–5)

2. Does the summary clearly identify the patient's main concerns or symptoms? (Completeness: 1–5, Correctness: 1–5)

3. Are listed medications, vitals, diagnoses, and treatment plans appropriately covered? (Completeness: 1–5, Correctness: 1–5)

4. Does the summary accurately reflect any improvement, worsening, or resolution of prior complaints? (Completeness: 1–5, Correctness: 1–5)

5. Is the summary free from hallucinations or invented content not supported by the source documentation? (Yes/No)

D.2 Detailed scores

Reviewer ratings (mean ± SD) and inter-rater agreement scores are described in Table A5: Overall, correctness and completeness scores were consistently high across all domains, with averages close to 5, and hallucinations were rare (4% of summaries).

An example of the LLM model's interpretation leading to possible hallucination is a patient whose original note mentions "increasing back pain, dysuria, discolored unit, advised patient to go to [urgent care]" and the summary mentions "Diagnosis: ... possible urinary tract infection or other renal condition". Though the model can possibly be correct, it is interpolating a diagnosis when none is mentioned in the original note. Another example of hallucination is when "HCM: colonoscopy was done in ..." in a primary care note, is mistakenly detected as a diagnosis of "hypertrophic cardiomyopathy" in the summary, instead of "healthcare maintenance". Lastly, in terms of completeness, CLIN-SUMM at times missed including that the patient was not adherent to a certain therapy, or pertinent negatives, such as "no AFIB detected on the ziopatch".

E Longitudinal evaluation

We selected a 3-year window size in the longitudinal evaluation experiments as a balance between maintaining sufficient temporal resolution and minimizing data truncation. As shown in Figure A2, the 95th percentile token count at 3-year windows (~4,663 tokens) remains well below the model's 8,192-token limit, resulting in a truncation rate of only 1%. In contrast, larger window sizes (≥ 6 years) exceed the token limit for a substantial fraction of patients, increasing truncation to ≥ 4–7%. Additionally, a 3-year window preserves an average of 4.5 windows per person, ensuring adequate longitudinal coverage while avoiding excessive fragmentation from smaller window sizes.

To quantitatively evaluate performance across time, we computed patient-level metrics using each patient's mean prediction score across all applicable 3-year windows as the basis of classification. These results, summarized in Appendix Table A6, complement the temporal patterns illustrated in Figure 4.

Figure A3 illustrates examples of longitudinal prediction scores alongside some highlighted text from two patients' entire histories of **CLIN-SUMM** summaries. In the control (left), the model assigns higher risk scores (red) in earlier windows due to symptoms such as pain and numbness, but shifts to lower predicted risk (blue) in later windows as the summaries document symptom improvement or resolution. In contrast, the dementia case (right) exhibits more fluctuation in predicted risk across windows reflecting variation in note content, for example when visits discuss unrelated conditions (e.g., thyroid disease or menopause) in greater detail, leading to changes in predicted scores.

F Saliency Analysis

Figure A4 illustrates an example of LIME-based interpretability applied at the patient level and providing additional context on how predictive signals manifest within real clinical narratives.

G Evaluation of open-source LLMs

We evaluated open-source Qwen3-generated **CLIN-SUMM** summaries for downstream dementia classification and prediction tasks to evaluate their accuracy relative to our GPT-4o-generated summaries. Qwen3 produces stable and faithful clinical summaries, with mean BERTScore recall of 0.91 (Table A7) between Qwen- and GPT-generated text. To achieve this level of alignment, we applied several prompt refinements and model-specific adjustments: (1) we disabled "think mode" and enforced stricter structural guardrails to ensure consistent sectioning; (2) we provided concrete examples of commonly missed clinical elements, such as specific laboratory names (e.g., TSH, HbA1c, CBC, BMP) and vital signs to improve coverage of quantitative clinical details; and (3) we increased top-p and top-k values to encourage the model to retain more clinical terms that open-source models tend to truncate under default decoding settings. When Clinical-ModernBERT dementia classification and prediction models were trained on these new Qwen3 summaries, the downstream results were comparable to GPT-4o performance, with AUROC = 0.79 vs. 0.81 for classification and 0.72 vs. 0.75 for prediction (Table A8). Other metrics, including AUPRC and F1, showed similarly small differences (< 5%). These findings demonstrate that, with prompt refinement and minor model-specific tuning, open-source alternatives such as Qwen3 can achieve near-parity with proprietary LLMs for generating useful longitudinal summaries of clinical notes at a fraction of the computational and financial cost. This substantially mitigates prior concerns about scalability, accessibility, and reproducibility. Moreover, since Qwen3 can be deployed entirely within a secure, institution-controlled environment, it eliminates data-transfer risks and enhances compliance with privacy regulations, an essential requirement for clinical translation.

The recent release of GPT-OSS may represent another alternative for cost-effective, high-quality summarization at scale.

H Additional applications of CLIN-SUMM summaries

To illustrate the versatility of **CLIN-SUMM** summaries, we conducted exploratory experiments beyond the main dementia classification and prediction analyses. These demonstrations highlight how longitudinal, date-partitioned summaries can support diverse machine learning and clinical research tasks.

H.1 Disease Progression Analysis

We performed an exploratory assessment of whether **CLIN-SUMM** summaries could support automated analysis of disease trajectories. We provided the GPT-4o LLM with patient-level **CLIN-SUMM** summaries and prompted it to extract a structured view of *Condition Evolution & Disease Progression*.

H.1.1 Example Output

For a representative patient, the model produced the following output:

1. **Initial Focus (2017–2018):** The early records highlight generalized anxiety disorder, sleep apnea, and gastroesophageal reflux disease (GERD). Symptoms included anxiety attacks, snoring, and sleep disturbances. During this time, medication adjustments (e.g., Sertraline to Venlafaxine) were frequent due to side effects like sexual dysfunction.

2. **Mid-course (2018–2019):** There's clear progression with the emergence of chronic cough, back pain, and vertebral compression fractures (due to coughing or osteopenia). This time frame also marks hospitalizations for pericardial effusion and early signs of cardiac disease (e.g., chest pain, syncope).

3. **Advanced Disease (2020–2023):** The patient experienced a STEMI (myocardial infarction) with stent placement and later AV block, necessitating a pacemaker. There's evidence of ischemic cardiomyopathy, and recurrent angina despite prior interventions. Progressive osteoporosis, recurring facial flushing, and chronic sinusitis persist.

4. **Recent Status (2022–2023):** The patient deals with continued back pain, increased medication burden, and new symptoms such as vision changes, fatigue, and possible autonomic symptoms (e.g., facial flushing with elevated tryptase levels suggesting mast cell activation or similar disorders).

This breakdown revealed evolving patterns in diagnoses, medications, and comorbidities, demonstrating the ability of summaries to reflect disease progression over time.

H.2 Track Symptom Trends

We also explored whether longitudinal summaries can reveal worsening, improved, or recurrent symptom trajectories. Again, we provided the GPT-4o LLM with patient-level **CLIN-SUMM** summaries and prompted it to extract a structured view of *Symptom Trends*.

H.2.1 Example Output

For the same representative patient, the model produced the following output:

1. **Worsening: Chronic back pain** has become more debilitating, despite physical therapy and injections. **Facial flushing and syncope** have increased in frequency and complexity. Chest discomfort episodes are recurring, even **post-STEMI and PCI**.

2. **Improving or Stabilizing: Anxiety symptoms** appear to be more controlled recently with escitalopram (Lexapro). **Sinus-related issues** are now part of a stable chronic management plan with allergy and ENT follow-ups.

3. **Recurring: GERD** and **postnasal drip, cough, and sleep apnea symptoms** have persisted throughout, with varying intensity.

Such trends, directly interpretable from the **CLIN-SUMM** parquet file, illustrate the potential for **CLIN-SUMM** to support trajectory analyses of symptoms relevant to clinical trial eligibility, clinical monitoring and diagnosis.

H.3 Other Applications

Additional pilot applications include:

- **Phenotyping:** Extraction of cardiovascular risk factors from summaries.
- **Clinical question answering:** Retrieval of key events from longitudinal trajectories.
- **Visualization:** Generation of patient-level timelines for chart review.

These exploratory tasks underscore the potential for **CLIN-SUMM** summaries to serve as intermediate representations for varied downstream analyses.

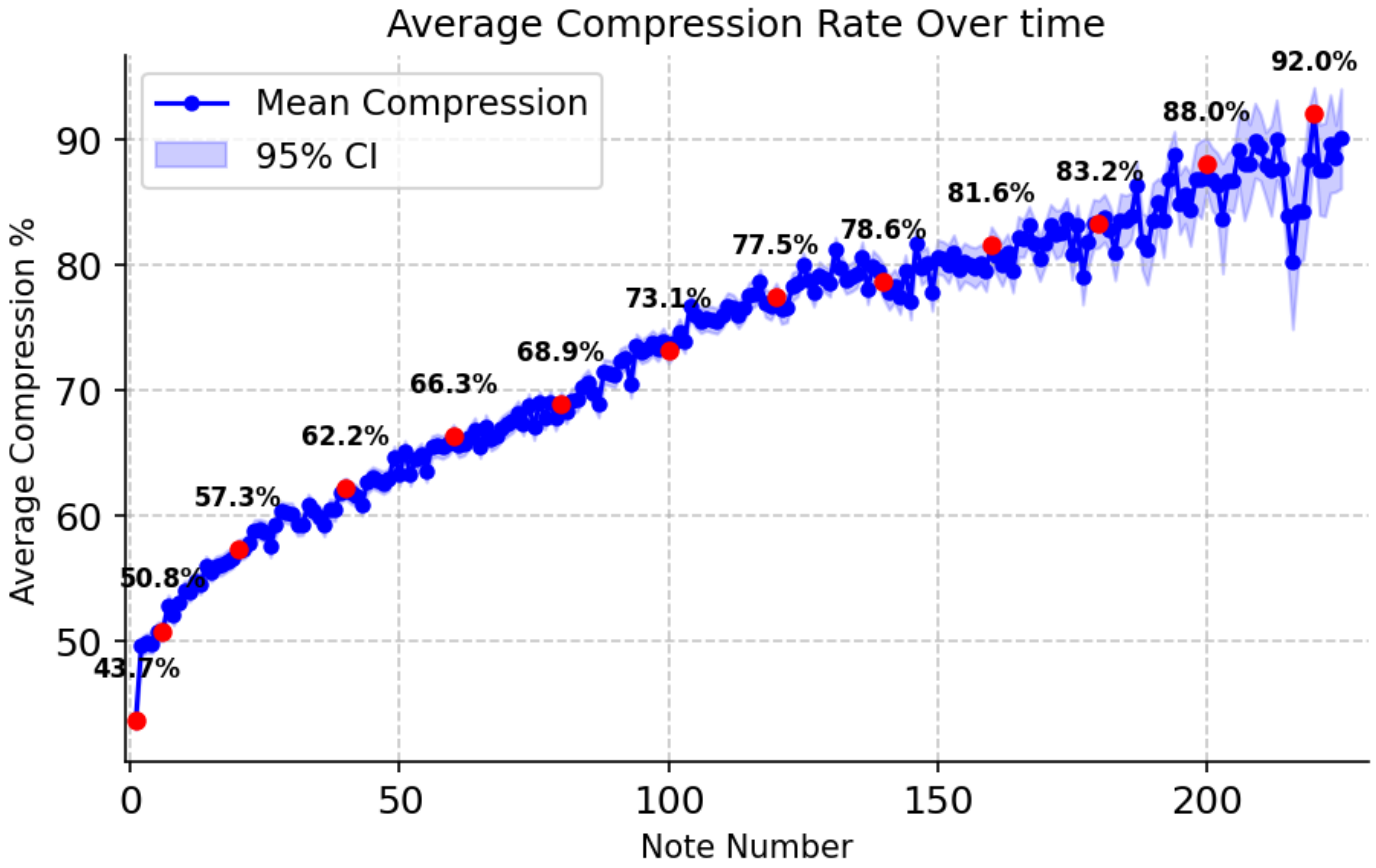

Figure A1: Average compression rate across clinical notes with 95% confidence intervals. The blue line represents the mean compression achieved by **CLIN-SUMM**, shaded regions indicate variability across patients, and red dots highlight key representative points (Larger confidence intervals on the right result from fewer patients with many notes) [In the pilot set of 1500]

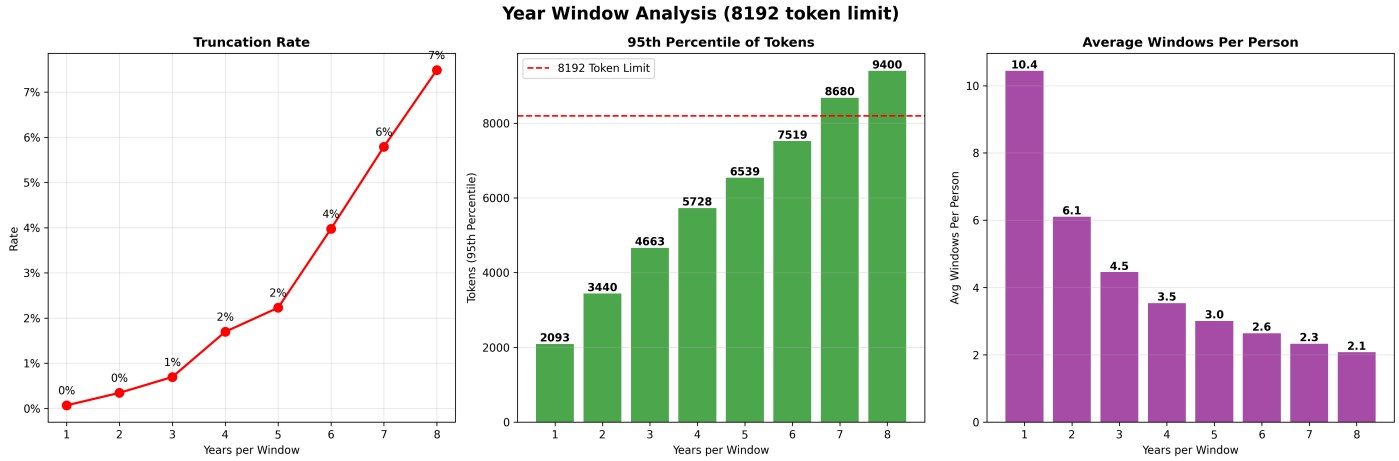

Figure A2: Analysis (In the pilot set of 1500) of different time window sizes for longitudinal modeling. Left: Truncation rate relative to the 8,192-token limit, showing low truncation ($\leq 1\%$) for window sizes up to 3 years but a sharp increase for larger windows. Middle: 95th percentile of token counts per window, demonstrating that 3-year windows (~4,663 tokens) remain comfortably below the model limit, whereas $\geq$ 6-year windows frequently exceed it. Right: Average number of windows per person, showing that 3-year windows preserve sufficient longitudinal coverage (~4.5 windows/person) while avoiding excessive fragmentation from shorter windows.

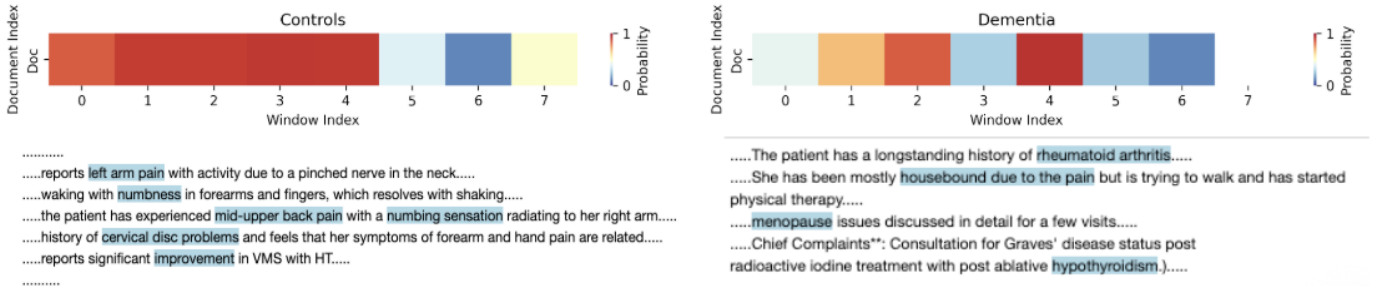

Figure A3: Example longitudinal scores from the dementia prediction model for a single control subject (left) and a single dementia case (right) from the pilot test cohort. The control experiences decreasing risk as symptoms improve, while the scores for the dementia case fluctuate due to detailed notes on other comorbidities.

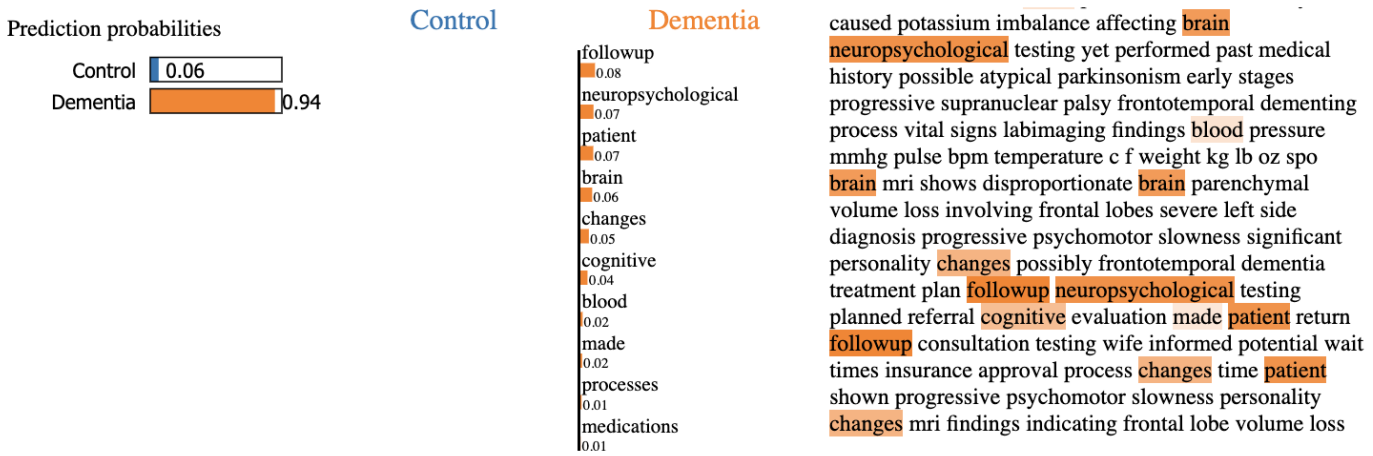

Figure A4: Example LIME-based interpretability for the classification model in a subject from the pilot test cohort. Left: Predicted dementia probability, Right: Key sentences from the patient's summarized notes, with highlights showing words contributing strongly to the model's prediction. Darker orange indicates stronger associations with higher dementia risk, while lighter shades indicate weaker contributions.

Table A1: Performance metrics by distance bin, including similarity, recall, and information loss, used to determine similarity threshold for Jaccard distance (In the pilot set of 1500).

Distance Bin	Mean Similarity (%)	Mean Entity Recall (%)	Info Lost (%)	# Pairs
(0.00 – 0.01]	99.97	99.18	0.82	1006
(0.01 – 0.02]	98.53	96.68	3.32	111
(0.02 – 0.03]	97.45	96.17	3.83	148
(0.03 – 0.04]	96.46	94.12	5.88	182
(0.04 – 0.05]	95.50	94.31	5.69	211
(0.05 – 0.06]	94.53	92.77	7.23	239
(0.06 – 0.07]	93.50	93.02	6.98	237
(0.07 – 0.08]	92.50	91.81	8.19	316
(0.08 – 0.09]	91.50	91.13	8.87	253
(0.09 – 0.10]	90.52	90.58	9.42	278

Table A2: ICD-9 and ICD-10 codes used for dementia phenotyping.

ICD Version	Codes
ICD-9	290.0; 290.10; 290.11; 290.12; 290.13; 290.20; 290.21; 290.3; 290.40; 290.41; 290.42; 290.43; 290.8; 290.9; 294.1; 294.10; 294.11; 294.20; 294.21; 331.0; 331.1; 331.11; 331.19; 331.2; 331.82
ICD-10	F01.x; F01.5; F01.50; F01.51; F02.x; F02.8; F02.80; F02.81; F03.x; F03.9; F03.90; F03.91; G23.1; G30.0; G30.1; G30.8; G30.9; G31.0; G31.01; G31.09; G31.1; G31.83

Table A3: SNOMED codes used for dementia phenotyping.

Name	Code	Name	Code
Dementia	52448006	Senile dementia with depression	191459006
Dementia associated with another disease	191519005	Senile dementia with depressive or paranoid features	191457008
Traumatic encephalopathy	230282000	Primary degenerative dementia of Alzheimer type, senile onset, uncomplicated	66108005
Cerebral degeneration presenting primarily with dementia	279982005	Multi-infarct dementia with delirium	10349009
Alzheimer's disease	26929004	Dementia of the Alzheimer type with behavioral disturbance	1581000119101
Vascular dementia	429998004	Mixed dementia	79341000119107
Dementia with behavioral disturbance	1591000119103	Multi-infarct dementia with depression	14070001
Vascular dementia without behavioral disturbance	16276361000119109	Senile dementia with delirium	191461002
Senile dementia	15662003	Presenile dementia with delusions	31081000119101
Presenile dementia	12348006	Senile dementia of the Lewy body type	312991009
Primary degenerative dementia of Alzheimer type, senile onset	416975007	Senile dementia with delusion	371024007
Uncomplicated senile dementia	191449005	Mild dementia	428051000124108
Frontotemporal dementia	230270009	Multi-infarct dementia with delusions	25772007
Primary degenerative dementia of Alzheimer type, presenile onset	416780008	Lewy body dementia with behavioral disturbance	135811000119107
Vascular dementia with behavioral disturbance	288631000119104	Presenile dementia with delirium	191452002
Diffuse Lewy body disease	80098002	Primary degenerative dementia of Alzheimer type, presenile onset, uncomplicated	6475002
Multi-infarct dementia	56267009	Dementia of frontal lobe type	278857002
Uncomplicated presenile dementia	191451009	Mixed cortical and subcortical vascular dementia	230287006
Multi-infarct dementia, uncomplicated	70936005	Psychoactive substance-induced organic dementia	111480006
Drug-induced dementia	191493005	Subcortical dementia	762707000
Dementia associated with alcoholism	281004	Subcortical vascular dementia	230286002
Presenile dementia with depression	191455000	Dementia due to disorder of central nervous system	724776007
Behavioral and psychological symptoms of dementia	10171000132106	Dementia due to Huntington chorea	442344002
Senile dementia with depression	191459006	Inhalant-induced persisting dementia	32875003

Table A4: Demographic characteristics for the entire dementia study cohort of 11,194 subjects. Values are reported as mean ± SD for continuous variables and *n* (%) for categorical variables.

Characteristic	Dementia (n=5,597)	Control (n=5,597)	All (N=11,194)
Age (years)	75.10 ± 7.42	75.06 ± 7.20	75.08 ± 7.31
Sex			
Male	2,457 (43.91%)	2,383 (42.58%)	4,840 (43.25%)
Female	3,138 (56.09%)	3,214 (57.42%)	6,352 (56.75%)
Race			
White	4,671 (83.49%)	4,544 (81.19%)	9,215 (82.34%)
Black	274 (4.90%)	266 (4.75%)	540 (4.82%)
Asian	146 (2.61%)	318 (5.68%)	464 (4.15%)
Other	504 (9.00%)	469 (8.38%)	973 (8.69%)
Average number of Notes			
Overall	272	167	219
Before diagnosis	182	–	170
Average Years of Data			
Overall	16	15	16
Before diagnosis	12	–	14

Table A5: Validation results for **CLIN-SUMM** summaries for a sample of 30 patients from the pilot set with over 700 notes. Scores reported on a 1–5 scale for correctness and completeness.

Validation Question	**Score**	**Inter- rater Agreement (%)** [95% CI]	**Cohen's κ**
Required sections included			
Completeness	4.87 ± 0.7	99.16% [97.52 – 100]	0.89
Patient's main concerns identified			
Correctness	4.84 ± 0.8	97.48% [94.66 – 100]	0.66
Completeness	4.82 ± 0.8	95.8% [92.19 – 99.4]	0.69
Coverage of medications, vitals, diagnoses, treatment plans			
Correctness	4.82 ± 0.8	99.03% [97.14 – 100]	0.66
Completeness	4.67 ± 0.8	96.12% [92.39 – 99.85]	0.85
Symptom improvement/ worsening captured			
Correctness	4.87 ± 0.7	100%	1.0
Completeness	4.80 ± 0.8	98.2% [95.72 – 100]	*(one rater gave same score for all)*
Free from hallucinations	96%	96.3% [93.11 – 99.48]	0.60

Table A6: Performance of the Clinical-ModernBERT model operating on 3-year windows of **CLIN-SUMM** summaries for dementia classification and future risk prediction in the pilot test cohort. Evaluation metrics were computed at the patient level, using each patient's mean prediction score across all 3-year windows as the basis for classification.

Metric	Classification	Future Prediction
AUROC	0.81	0.74
AUPRC	0.80	0.73
F1	0.75	0.67
Precision	0.70	0.54
Recall	0.82	0.93

Table A7: BERTScore comparison between Qwen3- and GPT-4o-generated **CLIN-SUMM** summaries in the pilot test cohort.

Metric	**Min**	**Max**	**Mean**	**Std. Dev.**
F1	0.83	0.95	0.90	0.017
Precision	0.82	0.95	0.90	0.018
Recall	0.83	0.95	0.91	0.017

Table A8: Performance of the Clinical-ModernBERT model using GPT-4o-generated versus Qwen3-generated **CLIN-SUMM** summaries for dementia classification and future risk prediction in the pilot test cohort.

Metric	**GPT-4o generated CLIN-SUMM summaries**		**Qwen3 generated CLIN-SUMM summaries**	
	Classification	**Prediction**	**Classification**	**Prediction**
AUROC	0.81	0.75	0.79	0.72
AUPRC	0.80	0.74	0.77	0.68
F1 Score	0.76	0.73	0.71	0.61
Precision	0.71	0.67	0.69	0.67
Recall	0.81	0.79	0.73	0.77