

Developing deep learning algorithms to generate radiation oncology clinical summaries

Ningyuan Li^{1,*}, Odette Rios-Ibacache¹, Amal Zouaq², John Kildea¹

¹ Medical Physics Unit, McGill University

² Department of Computer Engineering and Software Engineering, Polytechnique Montréal



Centre universitaire de santé McGill
Institut de recherche

McGill University
Health Centre
Research Institute

Introduction & Objective

Clinical notes in oncology are generally **free-text** documents that often repeat clinical information copied from elsewhere in the patient's medical record, **creating inefficiency and risk of error** [1,2]. Modern AI tools can potentially alleviate these problems [3], but they have not yet been applied systematically to oncology records.

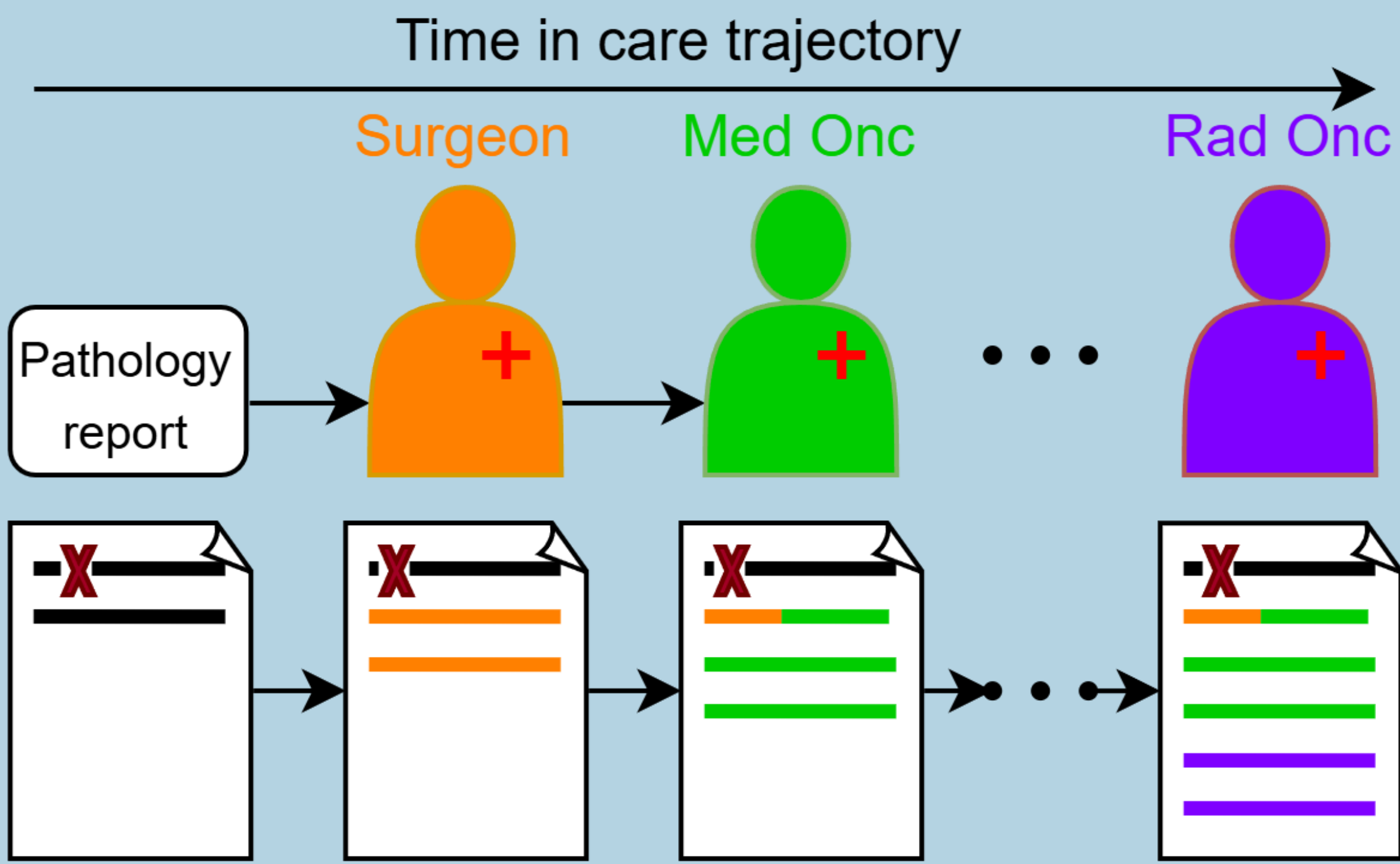


Figure 1: **Example of a typical clinical note-taking pipeline for a patient record in oncology.** Clinicians often copy-forward existing information that they deem important to include from past clinical notes and reports, then add new information after. This can cause past errors to propagate in a patient's record.

Therefore, our primary research goal is to develop and evaluate deep learning (DL) algorithms to generate audience-tailored clinical summaries in oncology.

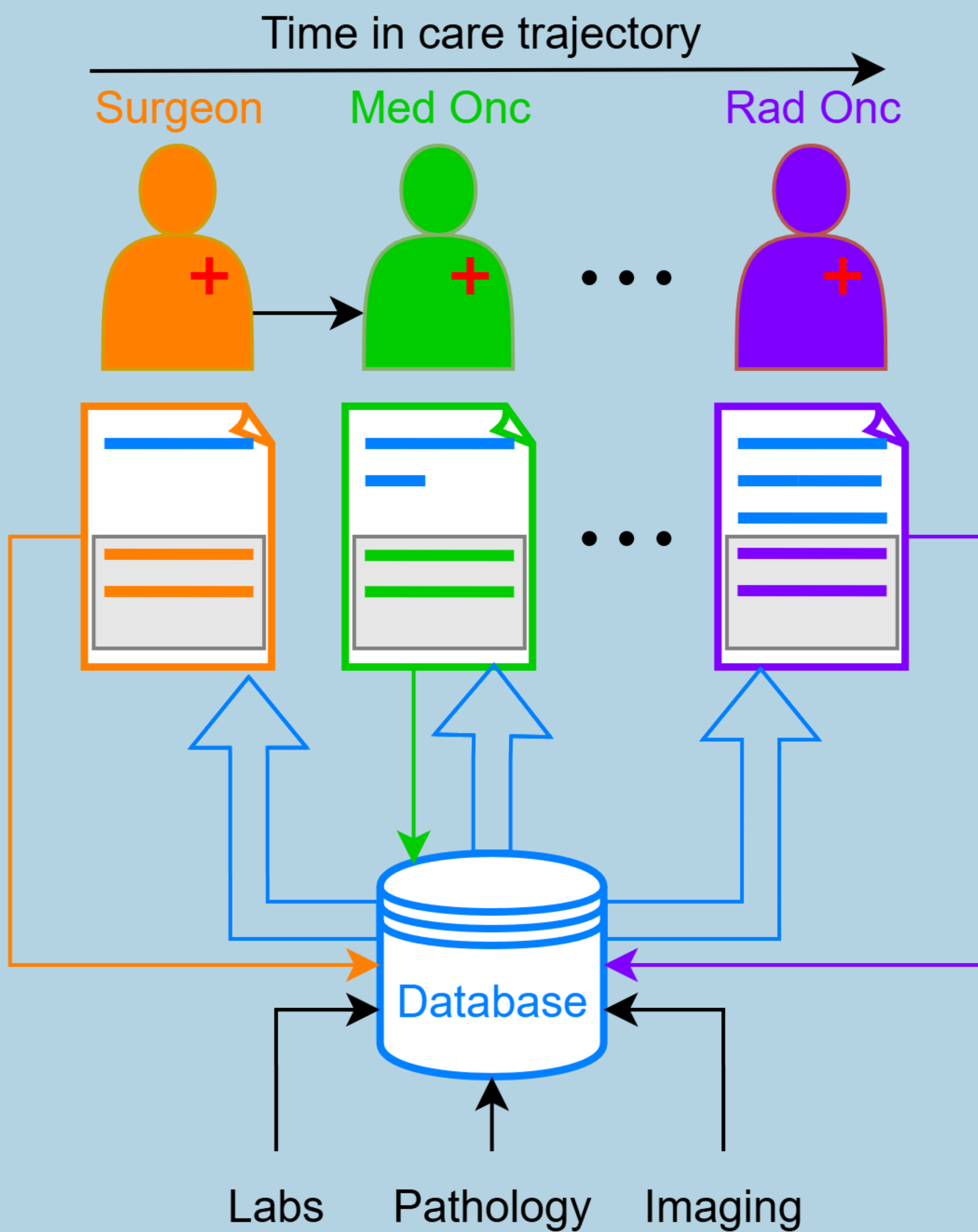


Figure 2: **Proposed clinical note-taking pipeline using our algorithm.** Instead of relying on the error-prone copy-and-pasting procedure, we use a continuously-updated database populated with results from tests and past consultations as the source of truth. Our algorithm generates the existing content referencing the database, whilst leaving room for the clinicians to fill in new information.

Methodology

Our algorithms are being trained and validated using existing oncology information system (OIS) data (structured data) and clinical notes (unstructured data) available at the McGill University Health Centre (MUHC). Specifically, we are:

- (1) building and testing software that, for clinical notes of each origin (e.g., from radiation oncologists, nurses, or nutritionists), identifies data already stored in the OIS ("redundant information") and stores them in a structured data format called mCODE (minimal Common Oncology Data Elements);
- (2) analysing the extracted results from part (1) to infer content expected in the notes by different specialist types;
- (3) reviewing and corroborating the findings with clinicians and patient partners; and
- (4) developing generative deep learning algorithms to generate semi-populated radiation oncology clinical notes after consultations such that only the specialist-provided information needs to be added.

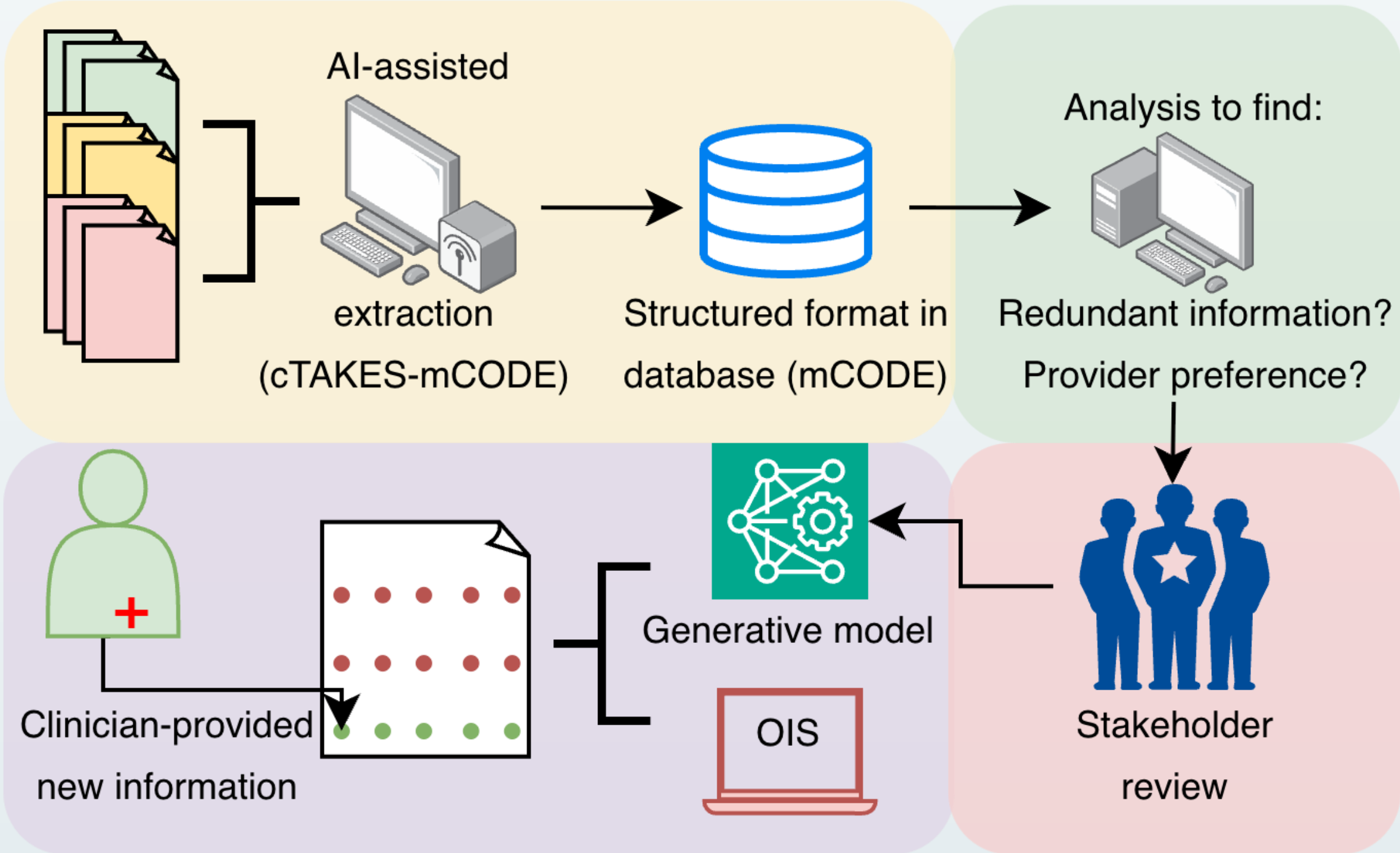


Figure 3: **Overall project pipeline.**

Current Results

Preliminary data set analysis

- We started with an analysis of our internal data set of 588 patients to investigate the content types present in our database.
- This visualisation helps us determine the document types that need to be processed by our software.

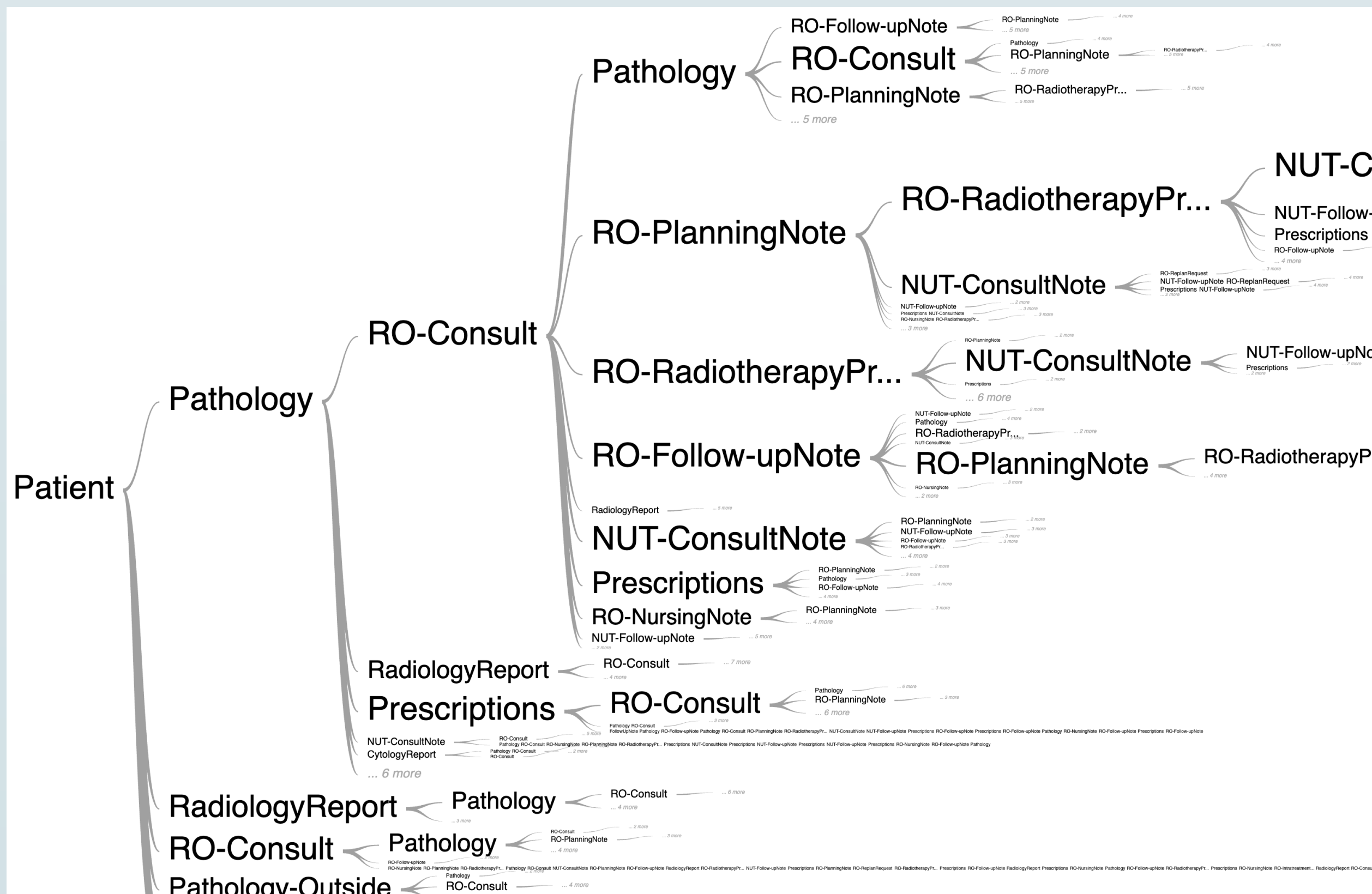


Figure 4: **Chronology of clinical documentation for our internal data set.** The documentation progression of patients in our study was visualised chronologically using a Google WordTree.

Information extraction from clinical notes

- We initially selected mCODEGPT [4], a **large language model (LLM)**-based tool to process unstructured oncology notes into mCODE, because it has been published and benchmarked.
- It was computationally too slow with local LLMs for practical use within the hospital environment.
- We, therefore, shifted from a fully LLM-based pipeline to a **hybrid design** in which an LLM serves only as **optional augmentation**.
- Our new software, **cTAKES-mCODE**, uses Apache cTAKES [5] to process notes and extracts mCODE fields from cTAKES-generated XMI output using a primarily **deterministic** pipeline, with selected fields **optionally** enhanced by lightweight local LLM models.
- The tool is still under active development, but early results on synthetic data are encouraging, and evaluation on hospital hardware and data is now beginning.

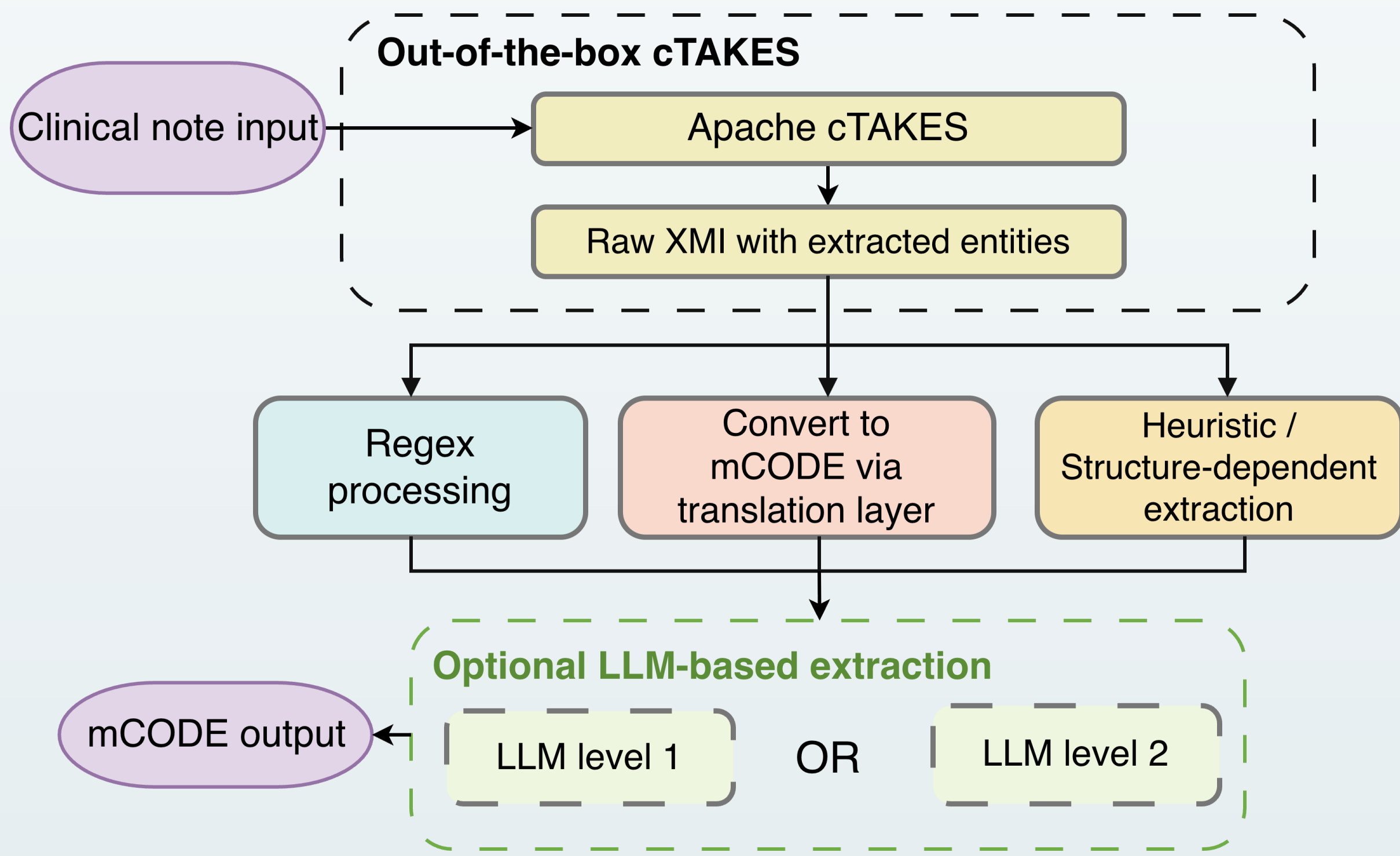


Figure 5: **cTAKES-mCODE design pipeline.** The package is designed in a modular fashion, which enhances customisability and transparency. The LLM-based extractor can be configured or entirely skipped based on hardware availability.

Next Steps & Conclusion

- We are developing an analysis pipeline combining the **extraction results** and **available metadata** for the notes to infer the expected content for different specialists. The results will be **reviewed and validated by key stakeholders**.
- The model will **autofill** audience-specific **existing** information while leaving room for clinicians to add **new developments**.
- In conclusion, by using AI to write summaries for different audiences, we aim to make clinical notes in radiation **oncology clearer, more concise, and more accurate**.

Acknowledgements

Mes travaux bénéficient d'un octroi du



Correspondence email:
ningyuan.li@mail.mcgill.ca

References

- [1] Kroth PJ, et al. The electronic elephant in the room: Physicians and the electronic health record. JAMIA Open. 2018;1(1):49–56.
- [2] Furlow B. Information overload and unsustainable workloads in the era of electronic health records. Lancet Respir Med. 2020;8(3):243–244.
- [3] Zhang J, et al. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. arXiv. 2020;1912.08777 [cs.CL].
- [4] Zhang K, Huang T, Malin BA, et al. Introducing mCODEGPT as a zero-shot information extraction from clinical free-text data tool for cancer research. Commun Med. 2025;5:422.
- [5] Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, component evaluation and applications. J Am Med Inform Assoc. 2010;17(5):507–513.