



Developing Large Language Models in Healthcare

Day 4: The medical LLM development and implementation cycle

Ruurd Kuiper – Julius Center, UMC Utrecht
Based on work by Yu-Wen Chen, Anne de Hond, Tuur Leeuwenberg, Karel Moons and others

Day 4: The medical LLM life-cycle

Theme: Understanding the comprehensive process of developing and deploying LLMs specifically for medical applications, culminating in a detailed project plan.

Learning Objectives:

1. Map out the end-to-end lifecycle of an AI model in healthcare.
2. Identify critical considerations and challenges unique to medical AI development.
3. Develop a structured plan for setting up, training, and implementing a foundation model.

Day 4: The medical LLM life-cycle

Planning

1. Finish lecture and practical from day 3

Break

2. Lecture on the medical LLM life-cycle

3. Discuss assignment

4. Create groups

Lunch

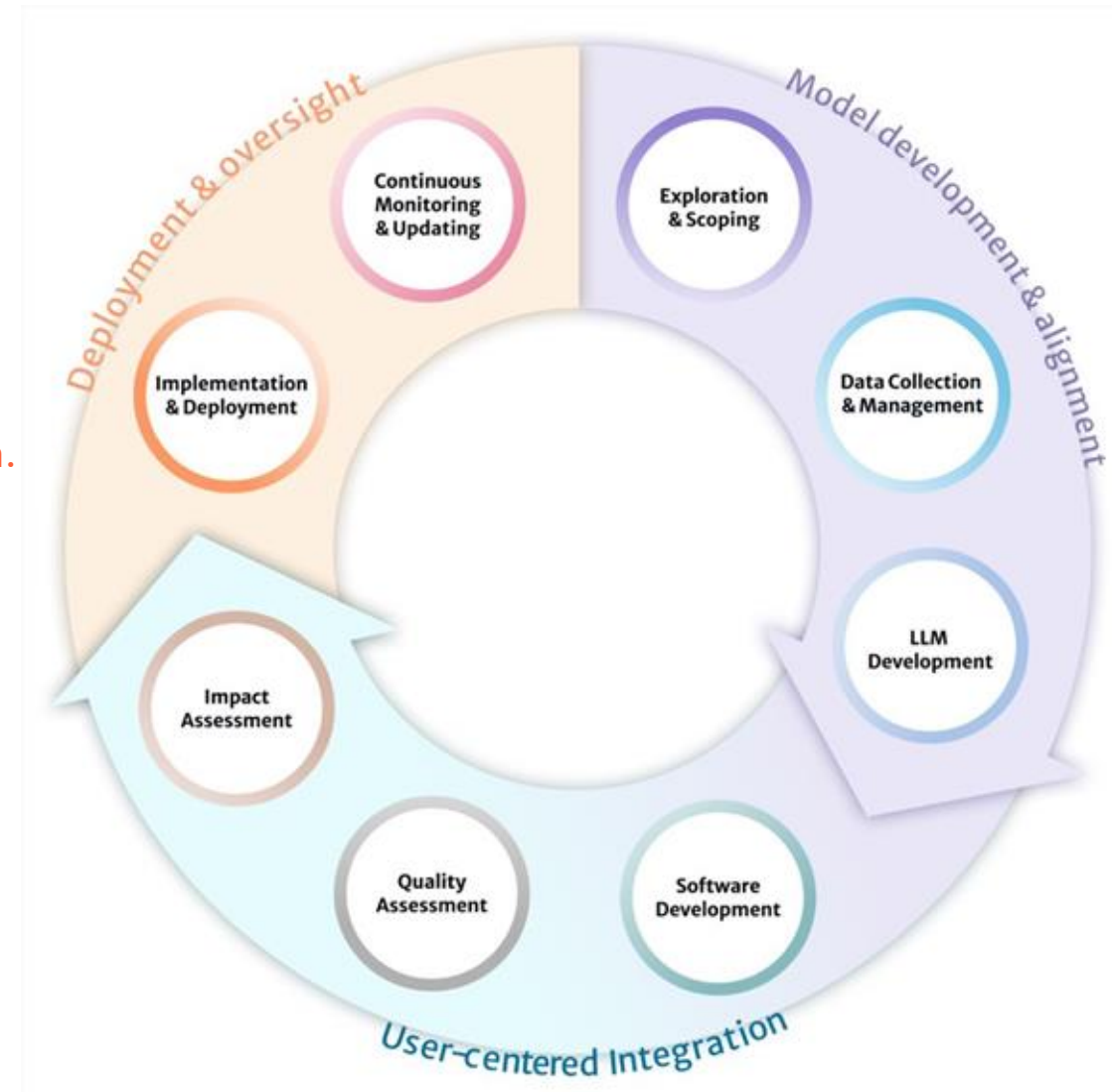
5. Groupwork

PRESENTATIONS ON FRIDAY AFTERNOON 2PM

Healthcare AI Lifecycle Overview

Overview

1. **Exploration and scoping:**
Define the problem, purpose, and LLM applicability.
2. **Data collection and management:**
Curate reliable, well-structured and representative data.
3. **LLM development:**
Train or fine-tune LLMs and design prompts for context-specific use.
4. **Software development:**
Integrate the LLM in a secure, and user-centered software application.
5. **Quality assessment:**
Evaluate and ensure LLM output meets quality standards.
6. **Impact assessment:**
Assess clinical, operational, economic, and environmental impact.
7. **Implementation and deployment:**
Deploy and integrate LLMs with training and ongoing support.
8. **Continuous monitoring and updating:**
Monitor performance, update, and decommission responsibly.



Healthcare AI Lifecycle Overview

Final Group Assignment

- 1. Form groups of four students:**
Try to have a mixed group with different expertise and problem-solving skills.
Think about skills in machine-learning, python, powerpoint, who has an interesting idea to work on etc.
- 2. Pick a problem you want to work on:**
Preferably a problem that one of you has encountered and described, but if no-one has good LLM applicability, pick something that interests you.
- 3. Focus first on the initial step of the AI lifecycle:**
Spend the most time on researching and ideating:
What solutions already exist? What datasets and models are available? Would an LLM be applicable?
- 4. Once you have a good idea, work through the whole healthcare AI lifecycle up to stage 6:**
Discuss and write a small paragraph for each step of the lifecycle up to stage 6.
- 5. Based on your plan, try to implement stage 1, 2 and 3:**
Choose a dataset
Choose and train/fine-tune/prompt-engineer a model
Validate your output using at least 3 metrics
- 6. Prepare a presentation for Friday afternoon**
Contents of the presentation are detailed on next slide!

Healthcare AI Lifecycle Overview

Assignment

PRESENTATION CONTENTS (rough guide):

1. Opening (names, project title) slide
2. Introduction (problem statement/description/background) slide
3. ± 3 slides on the planning of the healthcare lifecycle up to stage 6
4. Dataset slide
5. Model choice slide
6. Model training/prompting/finetuning slide
7. Validation (metrics) slide
8. Discussion and conclusion

Exploration and scoping

Define the problem, purpose, and LLM applicability

Item	Description
Health(care) opportunity and objective	• Collaborate with stakeholders to examine the current healthcare workflow and identify opportunities where LLMs could make meaningful contributions. • Prioritize these opportunities and define the objective and added value of using an LLM-powered application over non-LLM applications.
Intended purpose and context	• Define the intended purpose and context, including service type (e.g., administrative task or patient-centered, such as a decision-making application), user (e.g., patients, clinicians, administrators), healthcare setting, and required level of human oversight.
Risk assessment and acceptance criteria	• Conduct a preliminary risk assessment with stakeholders to identify privacy risks, misuse potential, and safety measures. • Establish risk acceptance criteria to guide decisions on risk mitigation or early termination of the LLM-powered application.
Social, ethical, and legal considerations	• Assess the social and ethical considerations (e.g., increased carbon footprint, changes in the physician-patient relationship) and compliance with local regulations.
Impact criteria	• Establish criteria aligned with stakeholder requirements for the operational, clinical, and patient-centered impacts (e.g., patient engagement).
Market analysis and implementation planning	• Conduct a market scan to identify potentially suitable LLMs for the intended purpose. • Plan feasible implementation strategies (e.g., training from scratch or fine-tuning a pre-trained model) based on available technical, human, and organizational resources.



Data collection and management

Curate reliable, well-structured and representative data.



Item	Description
Data format identification	Identify required data formats based on training, fine-tuning, or evaluation needs (e.g., medical QA pairs for fine-tuning, instructions with examples for prompting, or annotations for evaluation).
Data management plan	Develop a data management plan that covers the entire data lifecycle, including quantitative and qualitative data descriptions, governance measures, and preprocessing details.
Data collection, curation, and preprocessing	- Collect and curate distinct datasets for different purposes (e.g., training, fine-tuning, prompting, evaluation) to prevent data leakage. - Ensure that all collected data is legally compliant, relevant, justified, and obtained from trusted sources. - Mitigate privacy risks and minimize traceability using appropriate methodologies (e.g., anonymizing, pseudonymizing, or synthetic data)
Data characterization	Describe key data characteristics (e.g., metadata, distributions), and evaluate data quality regarding completeness, representativeness, and fitness for the intended purpose.
Fairness consideration	Consider the fairness aspects that the data does not disadvantage groups of people based on characteristics such as age, sex or gender, race or ethnicity, or socioeconomic status.

LLM development

Train or fine-tune LLMs and design prompts for context-specific use

Item	Description
Model selection	From the list of potentially viable pre-trained LLMs, select the appropriate models considering their capabilities, limitations, risks, and alignment with the intended purpose
Model training and fine-tuning	- Train and/or fine-tune the model using tailored, task-specific development methods such as instruction tuning and preference optimization. - Ensure alignment with human values and medical ethics, for instance, preventing harmful language.
Prompt engineering	- Refine prompts iteratively based on predefined criteria (e.g., accuracy, hallucination) and user feedback to improve response reliability and appropriateness. - Stress-test the application under real-world scenarios, such as harmful inputs and dynamic user behaviors. - Incorporate context-enriched tools such as Retrieval-Augmented Generation (RAG), if feasible, to support fact-based responses.
Version control and logging	- Implement model version control and logging to track key metadata, including architecture, version, and last date of training, fine-tuning, or updating. - Document the data sources for prompt development, prompt engineering techniques, and version history of prompts.



Software development

Integrate the LLM in a secure, and user-centered software application

Item	Description
Development standards and reproducibility	- Consider using relevant standards (e.g., Medical Device Software Lifecycle Processes IEC 62304:2006 + A1:2015) for best software development practices. - Employ version control (e.g., Git), publish code with anonymized or synthetic data, and document key configurations (e.g., environment settings) to enhance reproducibility.
System robustness and security	- Establish fallback procedures (e.g., routing to human experts) to ensure system robustness against unexpected errors. - Test the application regularly for vulnerabilities and implement security safeguards, such as input/output validation and strict access control. - Ensure that sensitive internal data is protected and not transmitted to external model providers (if relevant).
Logging	- Establish internal logging mechanisms to track usage activities (e.g., prompts, responses, latency) for monitoring patterns and detecting malicious use. - Ensure that logged data complies with local regulations, such as anonymization, secure storage, and strict access controls.
User-centered design and transparency	- Involve end-users in designing user interfaces, considering their literacy and digital skills, and build in accessible ways for them to provide feedback. - Clearly label AI-generated content in the interface to support transparency.



Quality assessment

Evaluate and ensure LLM output meets quality standards



Item	Description
Evaluation design	• Ensure stakeholder involvement throughout all evaluation components. • Select suitable evaluation methods (e.g., human, automatic, LLM-as-a-judge), considering the resource feasibility, benefits, and limitations. • Design a systematic evaluation and/or annotation scheme, including evaluation frequency, instructions, the required number of human references, generated content, and evaluators (if relevant). • Involve multiple, diverse, and representative evaluators for outputs of LLM-powered applications (if relevant).
Reference standard	• Determine whether a reference standard is needed and, if so, select one that aligns with the application's intended purpose (e.g., human-written clinical notes, annotations, or conventional diagnostic standards). • Ensure that the reference standard is task-specific, locally relevant, and developed in collaboration with stakeholders. • Describe the potential limitations or deficiencies of the reference standards.
Quality metrics	• Consider using established quality instruments (e.g., Provider Documentation Summarization Quality Instrument) and select metrics (e.g., factuality, coverage, fluency) aligned with the intended purpose. • Avoid relying solely on accuracy or lexical overlap and ensure quality metrics measure what they are intended to measure.
Fairness and unintended output	• Conduct a fairness analysis by evaluating metric outcomes across key attributes (e.g., gender, ethnicity, age) to identify potential disparities, often referred to as algorithmic bias. • Estimate the consequences of unintended output such as hallucination and harmful language. • Determine necessary mitigation measures based on risk acceptance criteria.

Impact assessment

Assess clinical, operational, economic, and environmental impact



Item	Description
Human oversight	• Test the LLM-powered application with human oversight in the targeted workflow before the final application deployment. • Involve real users (e.g., healthcare providers, patients) and collect user interactions to uncover any usability issues.
Healthcare impact	• Assess the impact on individual patient health outcomes, such as improvement in quality of life. • Evaluate the impact on patient experience, such as satisfaction, perceived quality of care and accessibility. • Examine cost-effectiveness of care, such as workflow efficiency and time saved. • Assess the impact on healthcare provider and staff well-being such as reduced burnout and improved work satisfaction.
Societal and environmental impact	• Address, where possible, the environmental impact resulting from deploying and operating the LLM application, such as energy consumption and the sustainability of the provider. • Evaluate societal impact, such as public acceptance and equitable access to LLM-guided services across demographic groups.
Ethical, legal, and operational evaluation	• Ensure the integrated workflow complies with ethical considerations and applicable legal requirements. • Assess operational risks that arise from the real-world use of LLM-powered applications in the workflow (e.g., automation bias, workflow misfit) and determine necessary mitigation measures.

Implementation and deployment

Deploy and integrate LLMs with training and ongoing support



Item	Description
Implementation plan	- Develop a comprehensive implementation plan that outlines the integration into the existing IT infrastructure and workflow (e.g., considering a pilot deployment). - Involve end-users, the development and management team, regulatory and ethics experts to ensure feasible and purpose-driven integration.
Secure deployment and data transmission	Implement security protocols for deployment and data transmission that are appropriate for the local organization (e.g., authentication, authorization, encryption).
Training	Provide training for all end-users, including activities covering the intended purpose, proper usage, impact on workflows, and ethical considerations.

Continuous monitoring and updating

Monitor performance, update, and decommission responsibly



Item	Description
Monitoring and incident response	• Monitor system operations in real-time and analyze logs to detect errors, unusual activities, and unexpected outputs. • Establish an incident response plan to ensure the timely handling of alerts and potential security issues.
Governance and audit	• Assign clear responsibilities for oversight, maintenance, and long-term accountability. • Set up audit processes with stakeholder involvement, specifying who will conduct them (e.g., internal or third-party), what they will assess (e.g., data security, fairness, legal compliance), and how often they will take place.
Updating	• Periodically review data pipelines, deployment settings, and user access to ensure security and compliance with legal and clinical standards. • Update (or decommission) models, systems, and procedures based on findings from reviews, audits, or best practices.

Healthcare AI Lifecycle Overview

Assignment

- 1. Form groups of four students:**
Try to have a mixed group with different expertise and problem-solving skills.
Think about skills in machine-learning, python, powerpoint, who has an interesting idea to work on etc.
- 2. Pick a problem you want to work on:**
Preferably a problem that one of you has encountered and described, but if no-one has good LLM applicability, pick something that interests you.
- 3. Focus first on the initial step of the AI lifecycle:**
Spend the most time on researching and ideating:
What solutions already exist? What datasets and models are available? Would an LLM be applicable?
- 4. Once you have a good idea, work through the whole healthcare AI lifecycle up to stage 6:**
Discuss and write a small paragraph for each step of the lifecycle up to stage 6.
- 5. Based on your plan, try to implement stage 1, 2 and 3:**
Choose a dataset
Choose and train/fine-tune/prompt-engineer a model
Validate your output using at least 3 metrics
- 6. Prepare a presentation for Friday afternoon**
Contents of the presentation are detailed on next slide!

Healthcare AI Lifecycle Overview

Assignment

PRESENTATION CONTENTS (rough guide):

1. Opening (names, project title) slide
2. Introduction (problem statement/description/background) slide
3. ± 3 slides on the planning of the healthcare lifecycle up to stage 6
4. Dataset slide
5. Model choice slide
6. Model training/prompting/finetuning slide
7. Validation (metrics) slide
8. Discussion and conclusion