Healthcare information system engineering: AI technologies and open source approaches

Shen, Z.

Propositions

accompanying the dissertation

Healthcare Information System Engineering: AI Technologies and Open Source Approaches

1. Federated database architectures that preserve local data sovereignty are superior to centralized data warehouses for multinational clinical trials, as they balance accessibility with privacy and maintain system availability across borders.
   Chapter 2.

2. Lightweight API-based frameworks for clinical NLP systems offer superior cost-effectiveness and development speed compared to traditional monolithic NLP frameworks, without sacrificing performance.
   Chapter 3.

3. The API-based framework designed for traditional NLP services is ideally suited for today's Large Language Models—both rely on external service integration through APIs, proving that flexible architecture matters more than the specific AI technology.
   Chapter 3.

4. Rule-based NLP pipelines tailored to specific document structures can achieve high precision and recall in extracting adverse drug reactions, though their generalizability remains limited.
   Chapter 4.

5. Cloud-based big data frameworks enable biomedical literature mining at unprecedented scales and costs—processing over one million full-text articles for less than $2 demonstrates that computational resources are no longer the bottleneck.
   Chapter 5.

6. Systematic reviews of open source software repositories provide more timely and comprehensive insights than traditional literature reviews, as they capture the rapidly evolving landscape of clinical software development that publication cycles cannot match.
   Chapter 6.

7. The effective reuse of open source clinical software requires linking code repositories to scientific literature, as traditional software metrics (stars, forks, downloads) measure popularity but not clinical validity or effectiveness.
   Chapter 7.

8. Large Language Models have collapsed multi-stage clinical NLP pipelines into single API calls, but their lack of explainability and tendency to hallucinate make them unsuitable for high-stakes healthcare decisions without extensive validation.

9. The scarcity of publicly available clinical datasets in minor languages creates a systematic bias in both traditional NLP and modern AI research, perpetuating a linguistic divide that leaves low-resource language communities underserved by healthcare AI technologies.

10. Reproducibility in computational healthcare research depends less on publishing methods and more on making source code, data pipelines, and infrastructure configurations openly accessible.

11. The promise of open source software in healthcare is undermined by inadequate attention to security compliance, maintenance sustainability, and regulatory alignment—transparency alone does not guarantee trustworthiness.

12. The success of a PhD trajectory depends not only on technical contributions, but equally on learning when to transition from pursuit of perfection to acceptance of completion.

Zhengru Shen
Zoetermeer, October 28, 2025