MACHINE LEARNING FOR HEALTH

LLMs and Multimodality: The New Frontier of Clinical Intelligence

From statistical fluency to safe, auditable, human-in-the-loop care

TECHNICAL HANDOUT · SESSION 10

Professor Flávio Luiz Seixas
Graduate Program in Computing

Academic teaching material based exclusively on the PDFs available in the project

Table of Contents

Introduction

Language models expand the possibilities for generating, transforming, and querying text, but fluency is not evidence of correctness. In health care, every use must separate support for organizing information from autonomous decision-making and must make input data, outputs, and human review verifiable. [1]

The updated references support a deeper discussion of tokenization, autoregressive generation, retrieval augmentation, hallucination, ambient documentation, explainability, and the safety of clinical LLMs. [1]

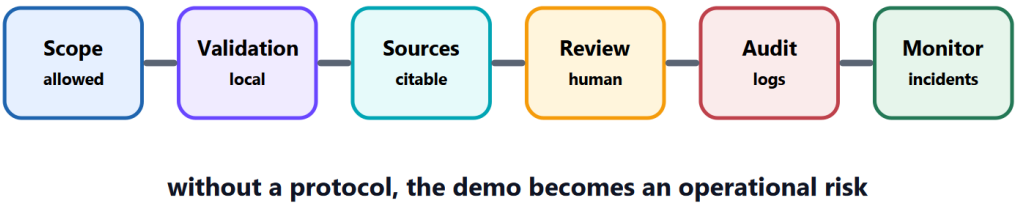

Figure 1 — Conceptual safety protocol for the clinical use of LLMs. Source: original illustration based on [1]–[5].

1. What an LLM processes and produces

Text is converted into tokens associated with embeddings before passing through Transformer blocks. In generative models, output is constructed one token at a time: each selected token is appended to the context and feeds the next pass. This mechanism explains why linguistic coherence is not equivalent to automatically consulting a clinical knowledge base [6].

Encoder-only architectures such as BERT are used especially to represent and classify text; generative architectures use decoding to produce sequences. The distinction helps determine whether to extract a variable, retrieve a document, or draft an open-ended answer [6].

KEY CONCEPT
Text is converted into tokens associated with embeddings before passing through Transformer blocks.

2. Contextual models and generation

Transformers produce context-dependent representations through attention mechanisms. BERT is used mainly to represent and understand text; generative models estimate continuations and can produce sequences. Pretraining followed by adaptation makes it possible to reuse representations in biomedical tasks [1].

In a clinical system, the input may include notes, a question, and retrieved context; the output may be a summary, code, answer, or draft. Evaluation must separately examine faithfulness to the record, omissions, unsupported additions, and suitability for the user [1].

3. Summarization, coding, and dialogue

Summarization condenses documents but must preserve facts, uncertainty, and temporality. Coding maps text to terminologies; errors can affect research, billing, and decisions. Dialogue systems must manage goals, turns, confirmation, and clarification rather than merely respond to isolated statements [2].

In ambient documentation, one possible workflow captures audio, transcribes it and identifies speakers, extracts clinical elements, proposes a structured note, and gives it to the professional for editing and signature. The output remains a draft: clinical sign-off is an explicit part of the process [7].

4. Multimodality and retrieval

Clinical data include text, images, signals, and measurements. Multimodal models seek to combine representations from these sources. The potential gain depends on temporal alignment, availability, and the meaning of each modality; systematic missingness may introduce bias [3].

Information retrieval adds a stage that locates relevant documents or passages before the response. It can provide verifiable material and reduce exclusive dependence on model parameters, but it does not guarantee that retrieval is correct or that generation adheres to the content [4].

In retrieval-augmented generation (RAG), the query and documents are represented for search, results may be reranked, and selected passages enter the generator's context. The chain should preserve the query, collection, version, passages, and response to support auditing; indexing or retrieval failures can still contaminate the synthesis [6] [8].

5. Hallucination, explanation, and evaluation

Hallucination is the production of plausible but false or unverifiable content. In health care, an invented laboratory value, an incorrect drug interaction, or a nonexistent reference can change care. Minimum safeguards include grounding in sources, clinical validation, and a human in the loop [7].

A useful explanation must be understandable and faithful to the system's behavior. RAG can expose external evidence but does not prove that the conclusion is correct; narrative explanations may sound convincing without reflecting the model's mechanism. Evaluation should combine faithfulness, robustness, comprehensibility, and clinical utility [8].

The testing protocol should separate retrieval, generation, and effects on work. In addition to language metrics, it should measure omitted or added facts, agreement with the source, stability, confidence calibration, subgroup performance, review time, human corrections, and incidents [7] [8].

6. Human-in-the-loop safety

A well-written output may contain information absent from the source. Protocols should require links to evidence, communication of uncertainty, verification by an authorized professional, a record of changes, and the option to reject the output. The greater the impact of an action, the more independent its verification should be [5].

Privacy requires control over the content sent, destination, retention, and access. LLMs also expand the attack surface through prompt manipulation, poisoning, membership inference, and leakage of identifiable information; controls must cover training, adaptation, inference, and integration [5] [9].

Session summary

What an LLM processes and produces: text is converted into tokens associated with embeddings before passing through transformer blocks.
Contextual models and generation: transformers produce context-dependent representations through attention mechanisms.
Summarization, coding, and dialogue: summarization condenses documents but must preserve facts, uncertainty, and temporality.
Multimodality and retrieval: clinical data include text, images, signals, and measurements.
Hallucination, explanation, and evaluation: hallucination is the production of plausible but false or unverifiable content.
Human-in-the-loop safety: a well-written output may contain information absent from the source.

References

[1] Xu, H.; Roberts, K. Natural Language Processing. In: Intelligent Systems in Medicine and Health. Springer, 2022. External link

Local file: pdfs/3-Intelligent Systems/VII-Natural Language Processing.pdf · PDF pages used: 8, 11, 14, 16.

[2] Bickmore, T.; Wallace, B. Intelligent Agents and Dialog Systems. In: Intelligent Systems in Medicine and Health. Springer, 2022. External link

Local file: pdfs/3-Intelligent Systems/IX-Intelligent Agents and Dialog Systems.pdf · PDF pages used: 1, 2, 6, 18.

[3] Sim, I.; Sirota, M. Data and Computation: A Contemporary Landscape. In: Intelligent Systems in Medicine and Health. Springer, 2022. External link

Local file: pdfs/3-Intelligent Systems/III-Data and Computation- A Contemporary Landscape.pdf · PDF pages used: 3, 7, 10, 14.

[4] Hersh, W. Information Retrieval. In: Biomedical Informatics. Springer, 2021. External link

Local file: pdfs/2-Biomedical Informatics/XXIII-Information Retrieval.pdf · PDF pages used: 3, 8, 13, 24.

[5] Goodman, K. W.; Miller, R. A. Ethics in Biomedical and Health Informatics: Users, Standards, and Outcomes. In: Biomedical Informatics. Springer, 2021. External link

Local file: pdfs/2-Biomedical Informatics/XII-Ethics in Biomedical and Health Informatics- Users, Standards, and Outcomes.pdf · PDF pages used: 8, 9, 10, 24.

[6] Alammar, J.; Grootendorst, M. Hands-On Large Language Models: Language Understanding and Generation. O'Reilly Media, 2024.

Local file: pdfs/1-Machine Learning/Hands-On Large Language Models_ Language Understanding and -- Jay Alammar, Maarten Grootendorst -- 1, 2024 -- O'Reilly Media, Incorporated -- isbn13 9781098150969 -- 44bf8ae4fd6eb0c873190856f9ec6b35 -- Anna’s Arch.pdf · PDF pages used: 40, 65, 66, 128, 248, 252, 397.

[7] Shaw, L.; Mahajan, S.; Upreti, K. (eds.). Natural Language Processing for Healthcare: The Rise of Intelligent Assistants. Academic Press, 2026.

Local file: pdfs/1-Machine Learning/Natural Language Processing for Healthcare.pdf · PDF pages used: 30, 32, 33, 65, 157, 159.

[8] Zamanifar, A.; Taherkordi, A.; Farhadi, A. (eds.). Explainable Large Language Models in Healthcare Applications. Springer, 2026. External link

Local file: pdfs/1-Machine Learning/Explainable Large Language Models in Healthcare Appliadeh Zamanifar & Amir Taherkordi & Amirfarhad Farhadi.pdf · PDF pages used: 84, 143, 162, 207, 283, 294, 311.

[9] Khalifa, N. E. M.; Taha, M. H. N. (eds.). Next-Gen Healthcare: AI-Powered Medical Innovations. Springer, 2026. External link

Local file: pdfs/1-Machine Learning/Next-Gen Healthcare_ AI-Powered Medical Innovations -- Nour Eldeen M_ Khalifa, Mohamed Hamed N_ Taha -- 2026 -- 0743a517976fb4e01964bc1deb4997db -- Anna’s Archive.pdf · PDF pages used: 180, 186, 187, 204.