Unlock Clinical Data with AI: The Rise of LCLMs

Joe Xing, Ph.D

Yifan Zhang
Founding ML Engineer at C. Light Tech

April 16, 2026

Visiting Professor at Tsinghua University, AI and Data Science Advisor for The Functional Neurology Center and Cofounder of and Cofounder of C. Light Tech.

Clinical Language

Clinical Language

- Clinical language is not just standardized terminology. It is a way of representing how the body functions in real patients. Terms like “Transferrin,” “eGFR,” or “HDL cholesterol in mg/dL” do not simply label data; they reflect underlying biological processes and physiological states. In this sense, clinical language acts almost like embodied intelligence, linking measurements to what is physically happening in the body and enabling clinicians to interpret and act with precision.

AI Language

- Learning representations of reality that can be computed and acted on. It does not directly describe the body like clinical language. Instead, it encodes patterns, relationships, and latent states from data.
- Rather than explicitly modeling physiology, it approximates it through signals, embeddings, and learned features.

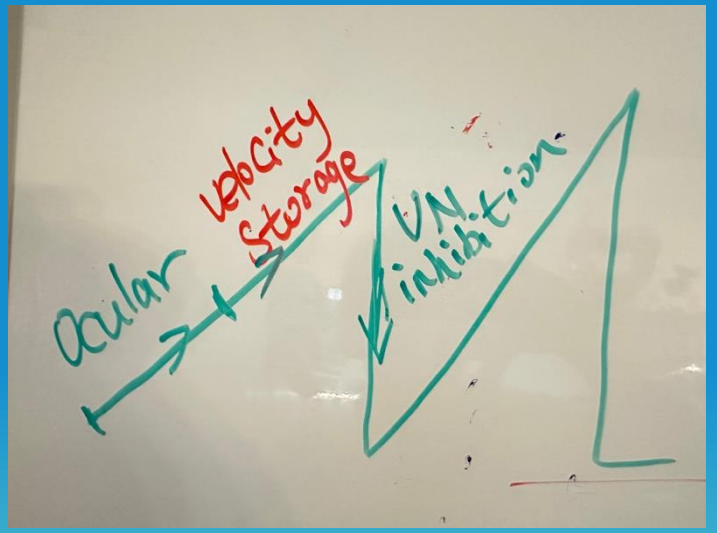
In this way, AI becomes a bridge between clinical reality and computation, translating real world biological processes into representations that machines can reason over and make

Physical Reality

Computational Representation

An Example: Optokinetic Nystagmus (OKN) Reflex

Joe's little sketch…



The ocular following response initiates tracking by generating a rapid, cortical driven eye movement in response to visual motion. This signal is then sustained and amplified by the velocity storage system, primarily involving the cerebellum, allowing continued tracking over time. As the response persists, the vestibular nuclei are modulated through cerebellar pathways to regulate and shape the output. Together, these systems create a coordinated progression from initial detection to sustained and controlled eye movement.

Adam Mani, et al., "A circuit suppressing retinal drive to the optokinetic system during fast motion," *Nature Communications*, 2023

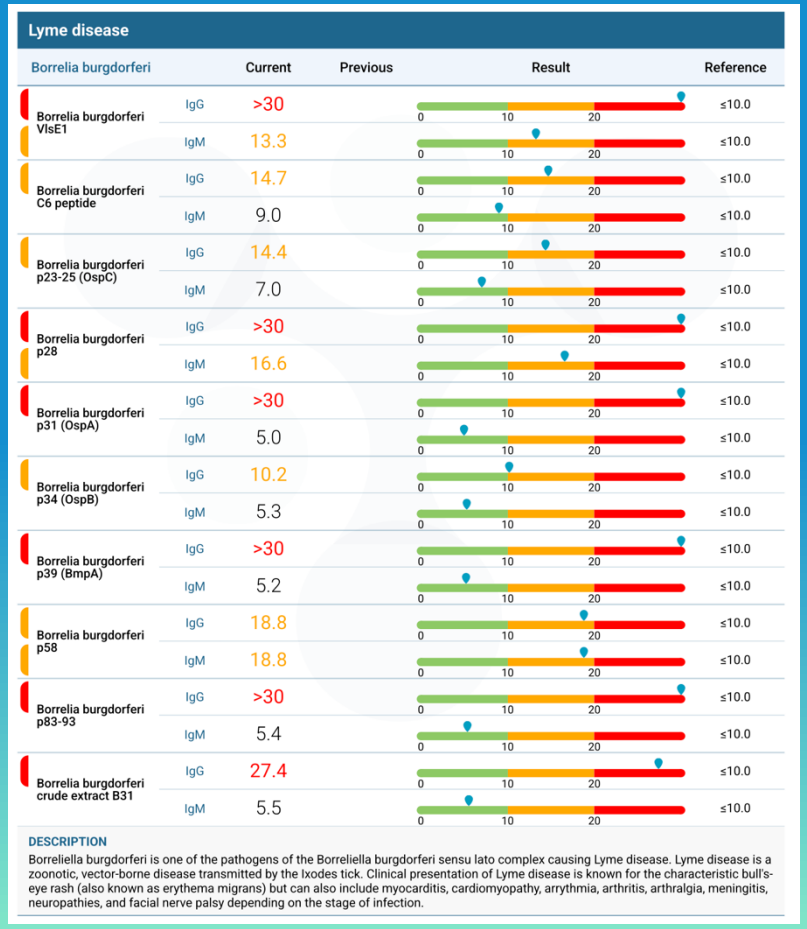
Paradigm Shift

Healthcare data from tables & labels →
→ Multimodal world representations

Medical AI from language understanding →
→ World understanding

Clinical care from episodic visit & reactive diagnosis →
→ Continuous monitoring & patient modeling and care

Lyme Disease Case Study



Lyme disease

Borrelia burgdorferi		Current	Previous	Result	Reference
Borrelia burgdorferi VlsE1	IgG	>30		0 10 20	≤10.0
	IgM	13.3		0 10 20	≤10.0
Borrelia burgdorferi C6 peptide	IgG	14.7		0 10 20	≤10.0
	IgM	9.0		0 10 20	≤10.0
Borrelia burgdorferi p23-25 (OspC)	IgG	14.4		0 10 20	≤10.0
	IgM	7.0		0 10 20	≤10.0
Borrelia burgdorferi p28	IgG	>30		0 10 20	≤10.0
	IgM	16.6		0 10 20	≤10.0
Borrelia burgdorferi p31 (OspA)	IgG	>30		0 10 20	≤10.0
	IgM	5.0		0 10 20	≤10.0
Borrelia burgdorferi p34 (OspB)	IgG	10.2		0 10 20	≤10.0
	IgM	5.3		0 10 20	≤10.0
Borrelia burgdorferi p39 (BmpA)	IgG	>30		0 10 20	≤10.0
	IgM	5.2		0 10 20	≤10.0
Borrelia burgdorferi p58	IgG	18.8		0 10 20	≤10.0
	IgM	18.8		0 10 20	≤10.0
Borrelia burgdorferi p83-93	IgG	>30		0 10 20	≤10.0
	IgM	5.4		0 10 20	≤10.0
Borrelia burgdorferi crude extract B31	IgG	27.4		0 10 20	≤10.0
	IgM	5.5		0 10 20	≤10.0

DESCRIPTION

Borreliella burgdorferi is one of the pathogens of the Borreliella burgdorferi sensu lato complex causing Lyme disease. Lyme disease is a zoonotic, vector-borne disease transmitted by the Ixodes tick. Clinical presentation of Lyme disease is known for the characteristic bull's-eye rash (also known as erythema migrans) but can also include myocarditis, cardiomyopathy, arrythmia, arthritis, arthralgia, meningitis, neuropathies, and facial nerve palsy depending on the stage of infection.

Lyme disease is caused by *Borrelia burgdorferi (tick bite)*, which spreads through the body and triggers a widespread inflammatory immune response, particularly affecting the nervous system. This neuroinflammation can disrupt normal brain and vestibular function, leading to symptoms like **headache, dizziness, and brain fog**. These symptoms are driven less by the bacteria itself and more by the immune system's effect on neural signaling and sensory processing.

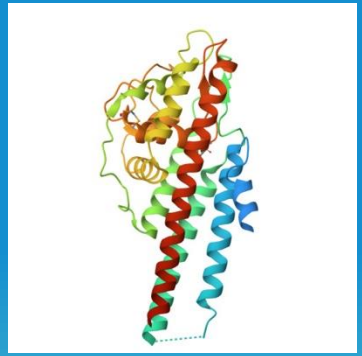
This cohort at our clinic consists of chronic brain health patients presenting with persistent neurological symptoms including headaches, dizziness, and cognitive dysfunction

Mahajan VK. Lyme Disease: An Overview. *Indian Dermatology Online Journal*. 2023; doi:10.4103/idoj.idoj_418_22

Biomarker

In medicine, a biomarker (biological marker) is any measurable indicator of a biological state or condition. HDL (High-Density Lipoprotein), LDL (Low-Density Lipoprotein), Total Cholesterol, Triglycerides, and the components of a CBC (Complete Blood Count), are all used by doctors to assess your health status.

What is VlsE?
Vmp (Variable Major Protein)-like sequence Expressed — a surface lipoprotein on Borrelia burgdorferi.

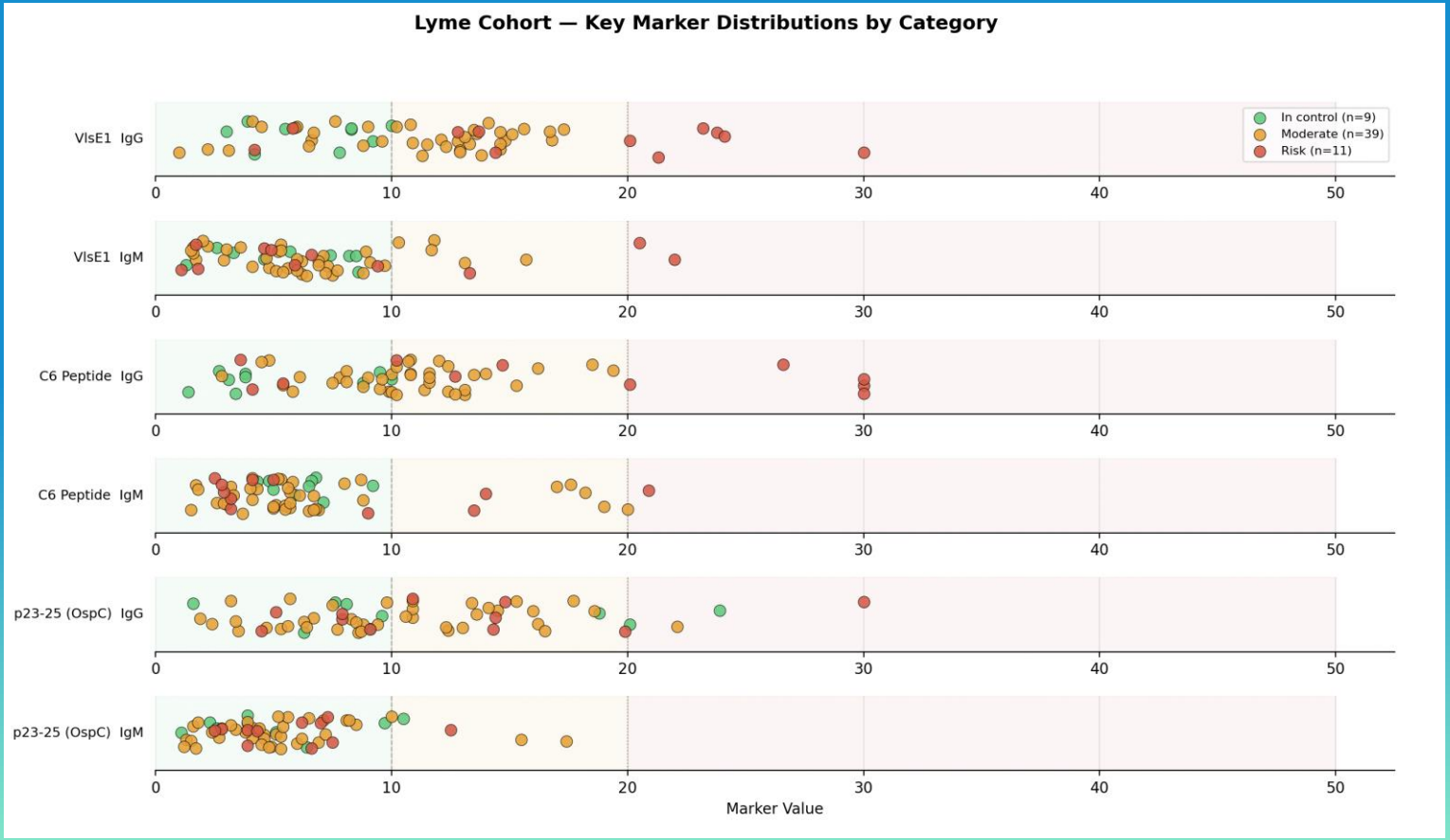


What is Immunoglobulin macroglobulin (IgM) / gamma (IgG) ?

Marker	Meaning
VlsE1 IgM / C6 IgM	Acute infection, recent tick bite, symptoms < 30 days, fast but not precise
VlsE1 IgG / C6 IgG	Established or chronic infection, symptoms > 30 days, slow but more effective and specific

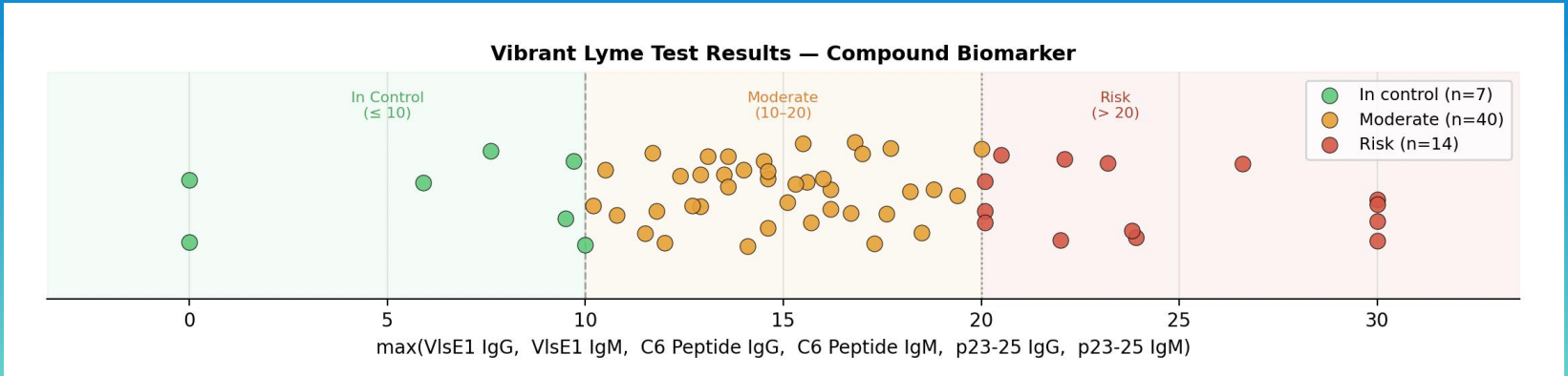
Wozinska et al. (2023) Diagnostics 13(23):3547. doi:10.3390/diagnostics13233547
Porwancher et al. (2011) Clin Vaccine Immunol 18(5):851–859. doi:10.1128/CVI.00409-10
Miraglia (2020) J Chiropr Med 19(3):201–202. doi:10.1016/j.jcm.2020.09.001
Bacon et al. (2003) J Infect Dis 187(8):1187–1199. doi:10.1086/374395
CDC (2024) Standard Two-Tiered Lyme Testing Interpretation. cdc.gov
CDC (2024) Modified Two-Tiered Lyme Testing Interpretation. cdc.gov

Individual Marker



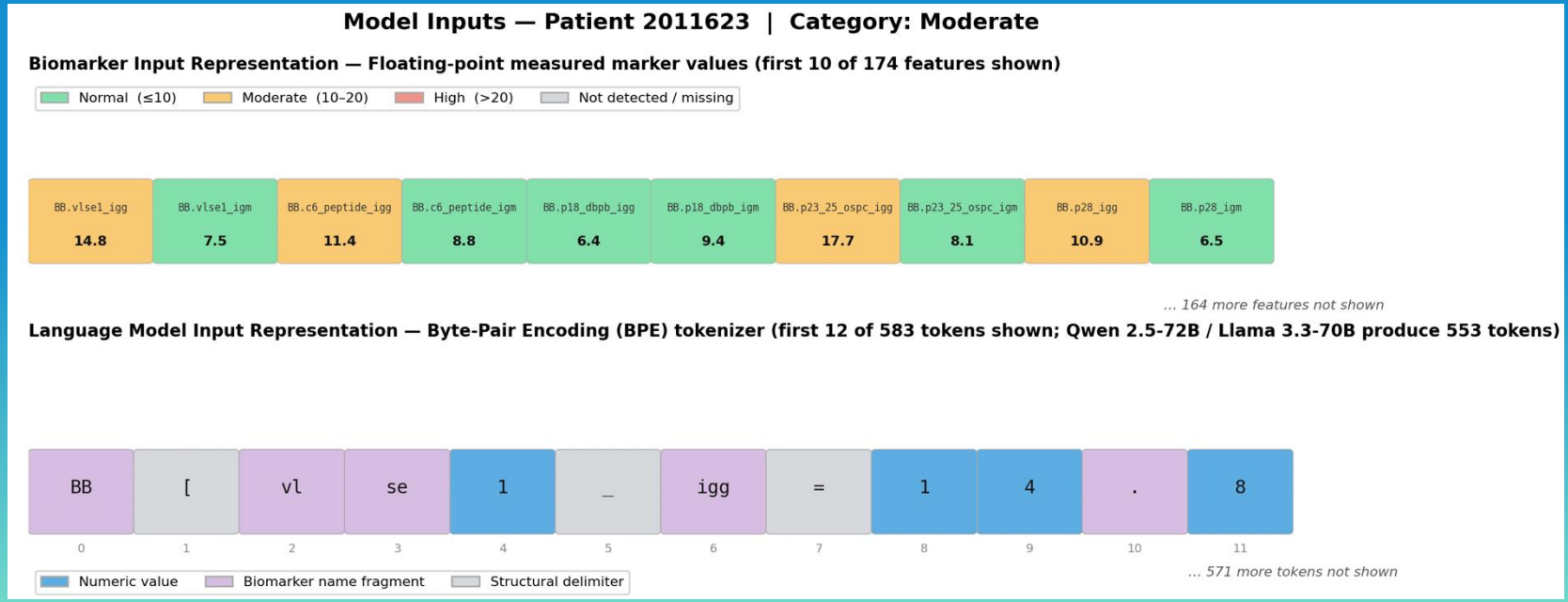
Compound Biomarkers

A simple composite marker is constructed by combining individual biomarkers such as VlsE1, C6, and p23–25, using the maximum value across these measures to represent overall signal strength. This provides a straightforward way to distinguish normal versus abnormal patterns and compare against future LLM outputs. A maximum score greater than 10 is considered abnormal, and values exceeding 20 are classified as high risk to reflect increasing severity.



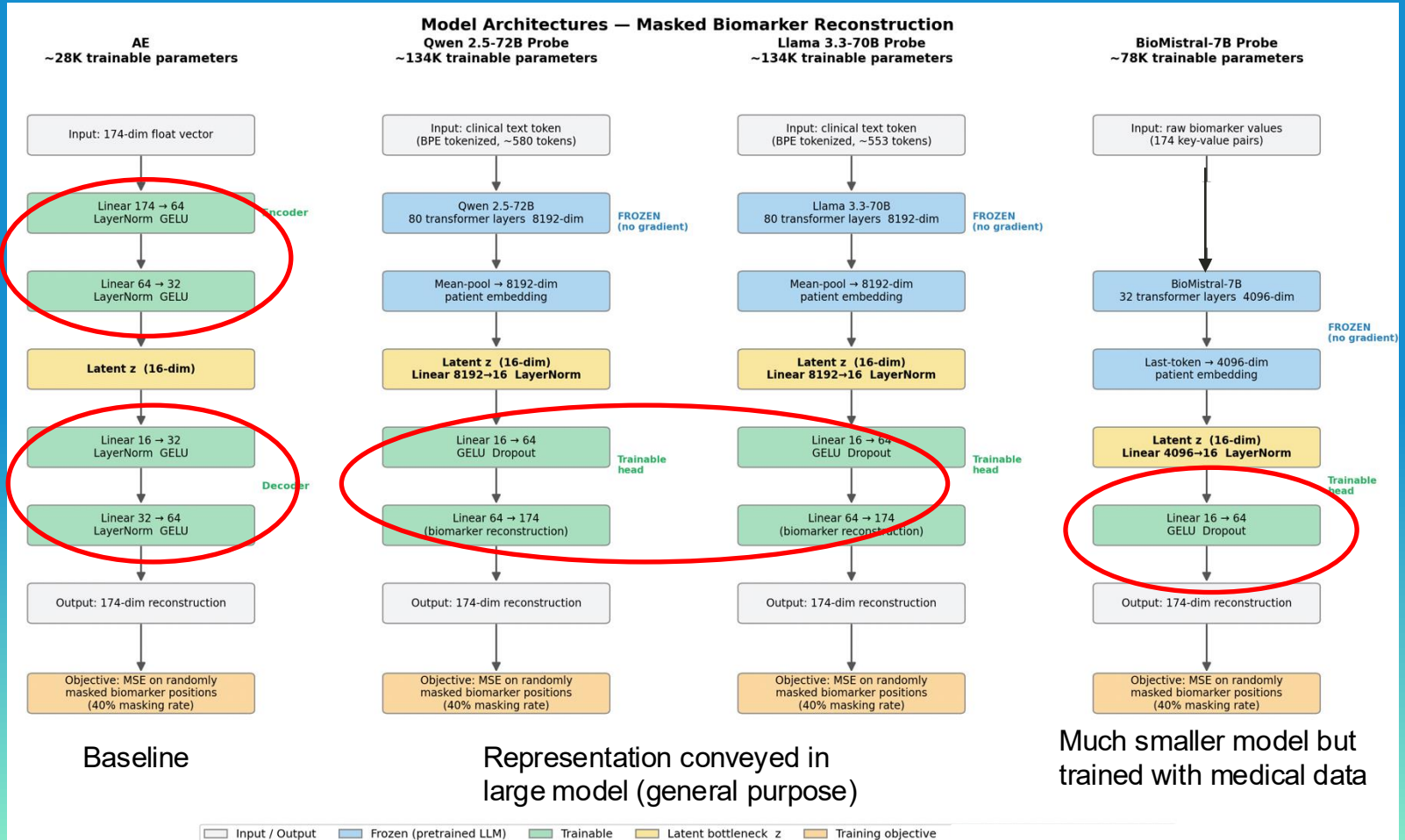
How To Build Tokens

We represent biomarker data as tokens using simple methods, either as floating-point vectors (e.g., for autoencoders) or via standard BPE on structured test results. This captures both numeric signals and text patterns in a model-friendly format, establishing a baseline before moving to more advanced tokenization.



Sennrich, R., Haddow, B., & Birch, A. (2016). Neural machine translation of rare words with subword units. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016), 1715–1725

Experimental Architectures



Baseline

Representation conveyed in large model (general purpose)

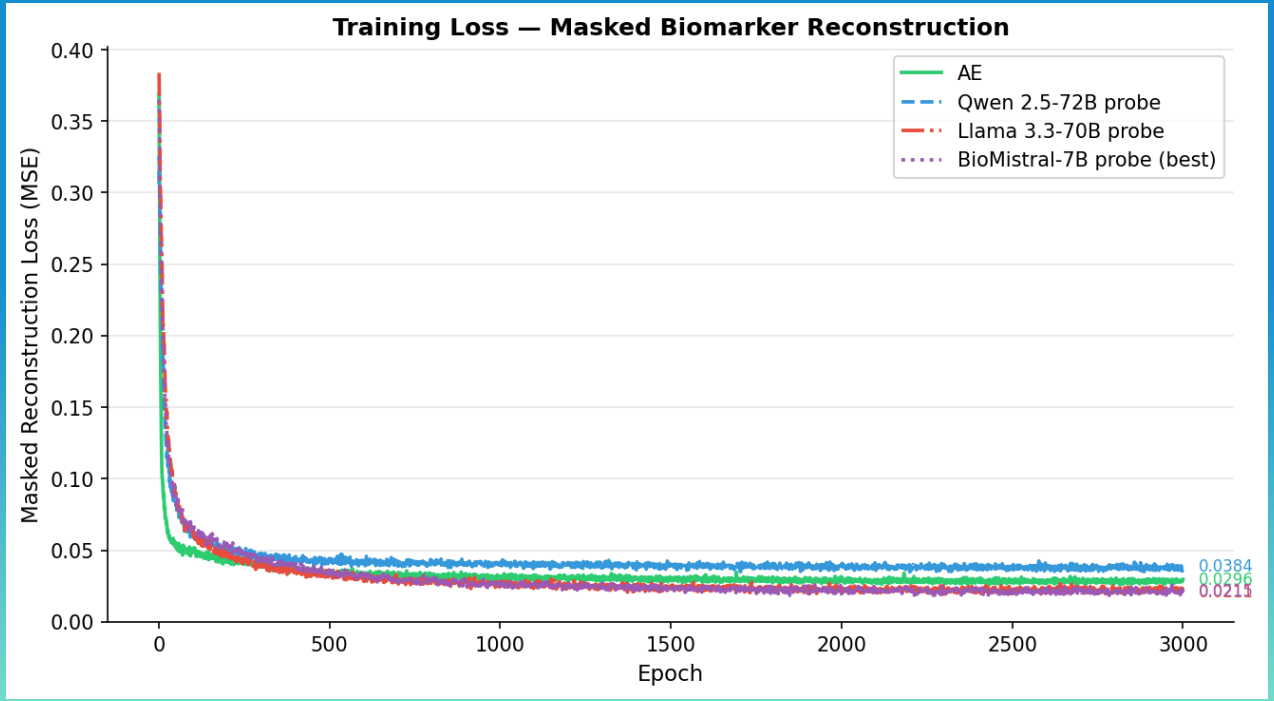
Much smaller model but trained with medical data

Labrak, Y., Bazoge, A., Morin, E., Gourraud, P.-A., Rouvier, M., & Dufour, R. (2024). *BioMistral: A collection of open-source pretrained large language models for medical domains*. Findings of the Association for Computational Linguistics: ACL 2024. https://arxiv.org/abs/2402.10373

A Probe to the Representations

A probe is a simple model trained on frozen representations (Qwen and Llama) to measure what information those representations contain.

This line of research uses small probe models on frozen LLM activations to test what information is already encoded. By learning simple decision boundaries, probes reveal whether specific signals are implicitly present without additional training.

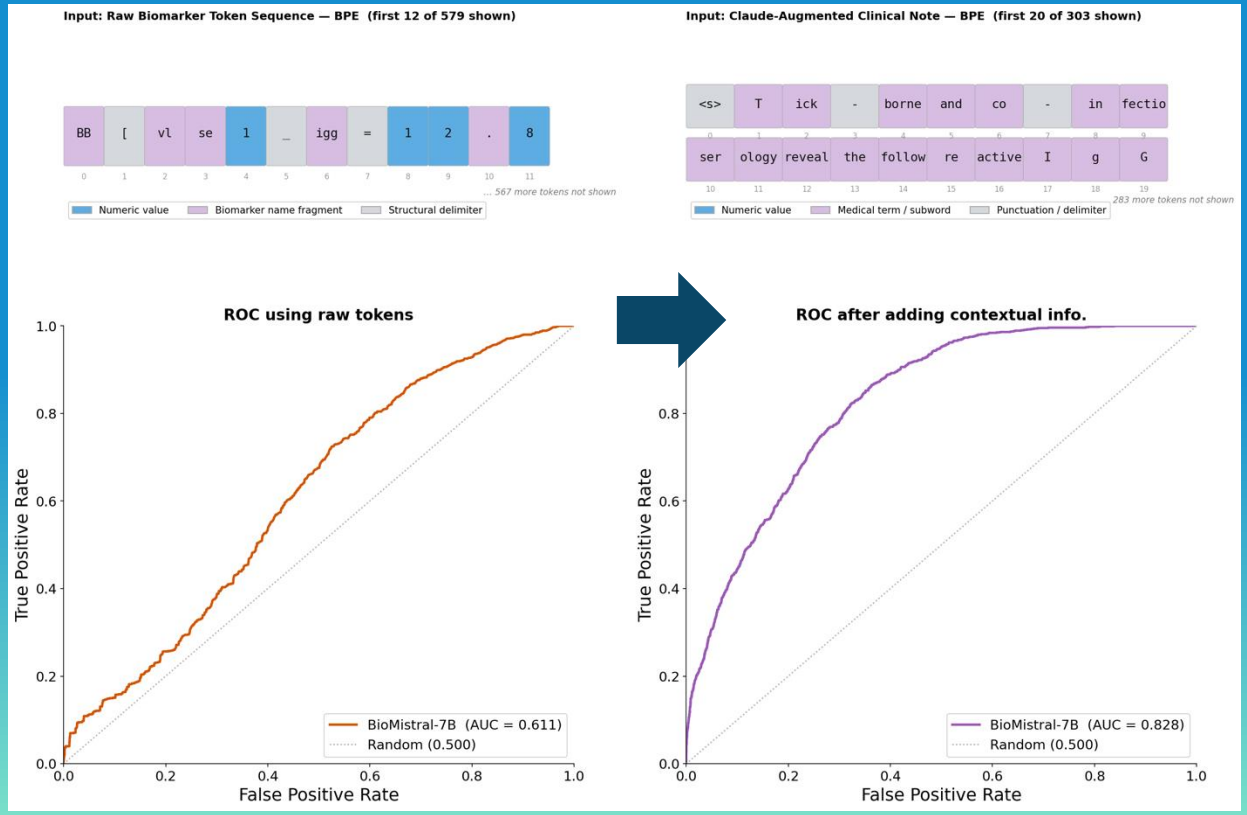


Iván Vicente Moreno Cencerrado et al. (2025). *No answer needed: Predicting LLM answer accuracy from question-only linear probes* (arXiv:2509.10625). arXiv. https://arxiv.org/abs/2509.10625

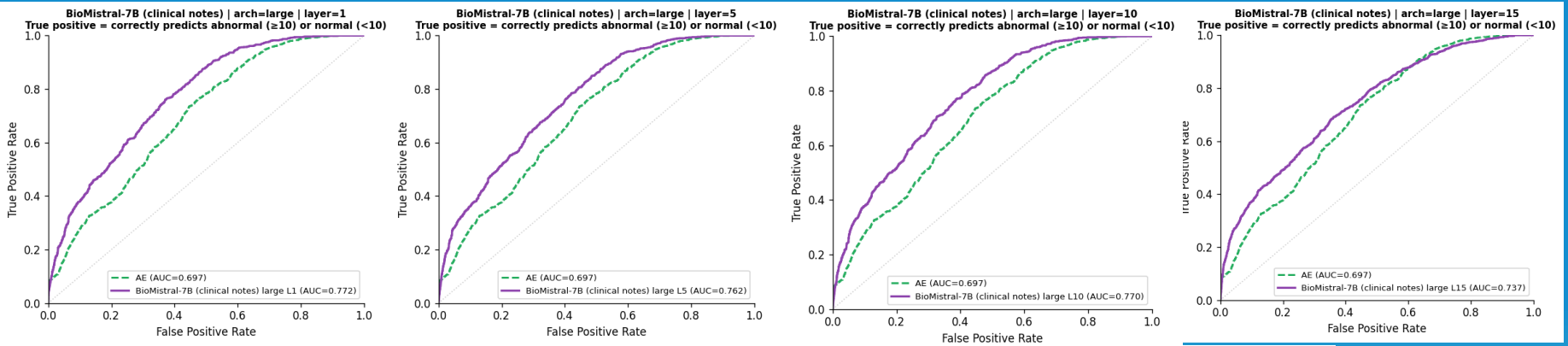
Power of Contextual Information – Ablation Study

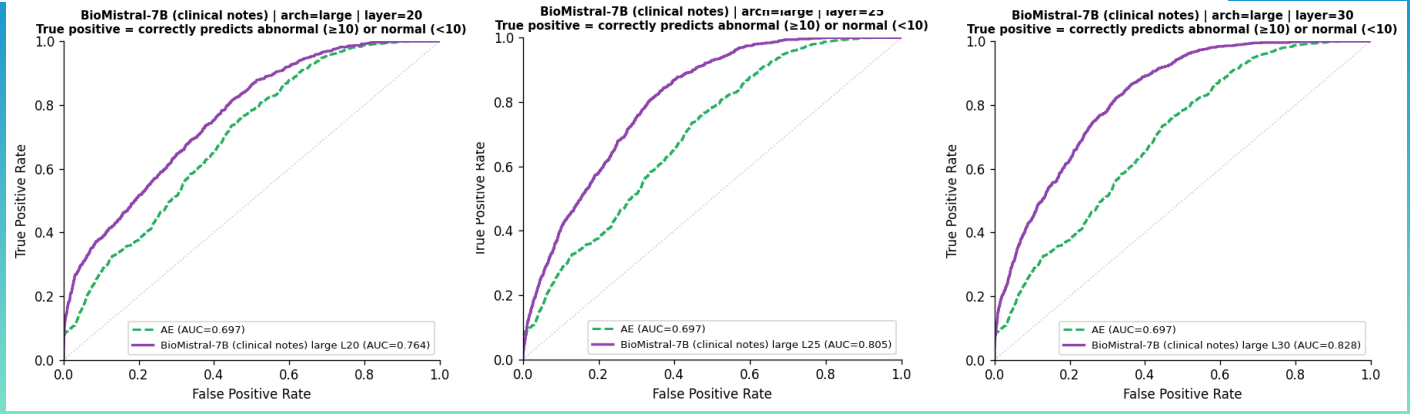
An interesting observation from our probing experiments is that smaller models like BioMistral 7B benefit significantly from added clinical context.

When biomarker tokens are augmented with brief explanatory notes, performance on masked biomarker prediction improves substantially compared to using raw numeric inputs alone. This suggests that smaller LLMs rely more heavily on contextual cues to effectively utilize underlying representations.



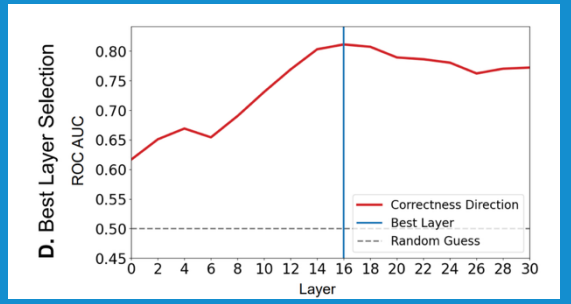
Identifying Informative Layers in BioMistral: Enhanced Biomarker Predictive Power in Deeper Representations



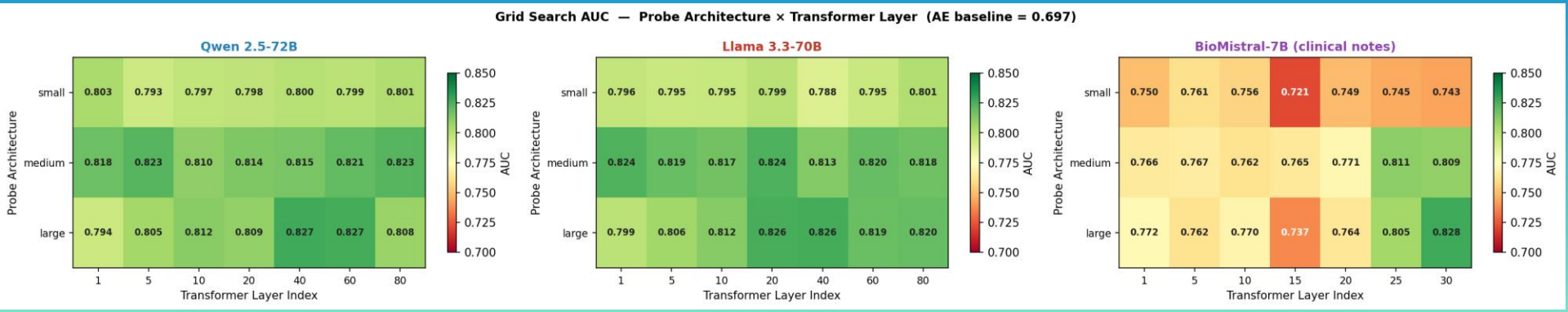


Systematic analysis of representations across layers and probe capacity

We control the size of the probe network by adjusting the input dimensionality of the representations. For smaller probes, the original 8192-dimensional activations are compressed using PCA (e.g., to 20 or 40 dimensions), significantly reducing the number of trainable parameters. In contrast, the large probe operates directly on the full 8192-dimensional space, resulting in a much higher-capacity model.



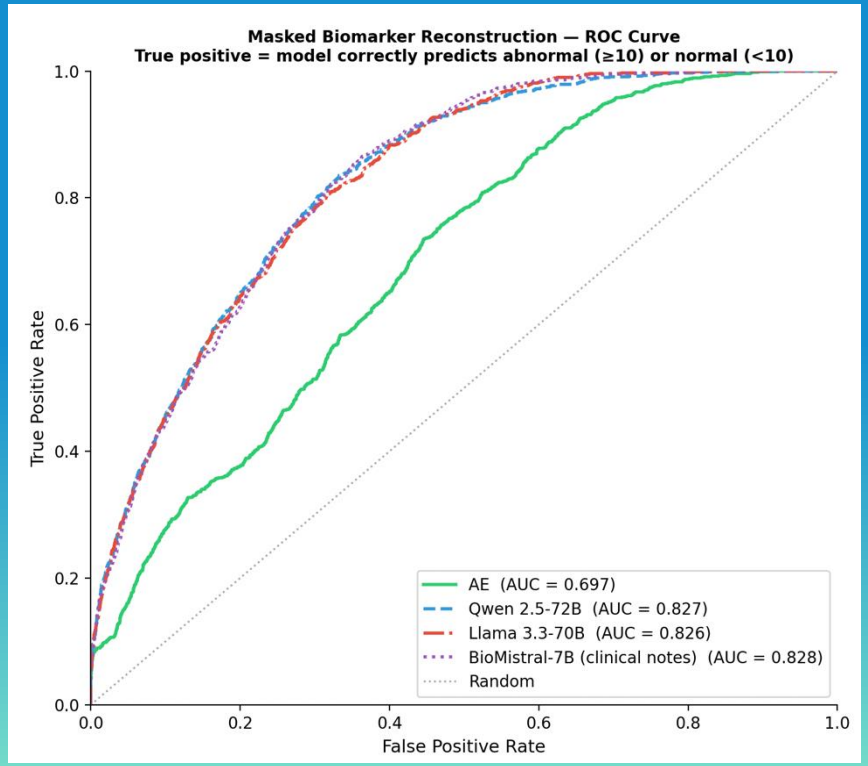
Iván Vicente Moreno Cencerrado et al. (2025). *No answer needed: Predicting LLM answer accuracy from question-only linear probes* (arXiv:2509.10625). arXiv. https://arxiv.org/abs/2509.10625



Results – Prediction of Abnormal and Normal Marker

We adopt a self-masked approach where the LLM, augmented with a trainable probe, predicts masked biomarker values as continuous outputs using RMSE as the training objective.

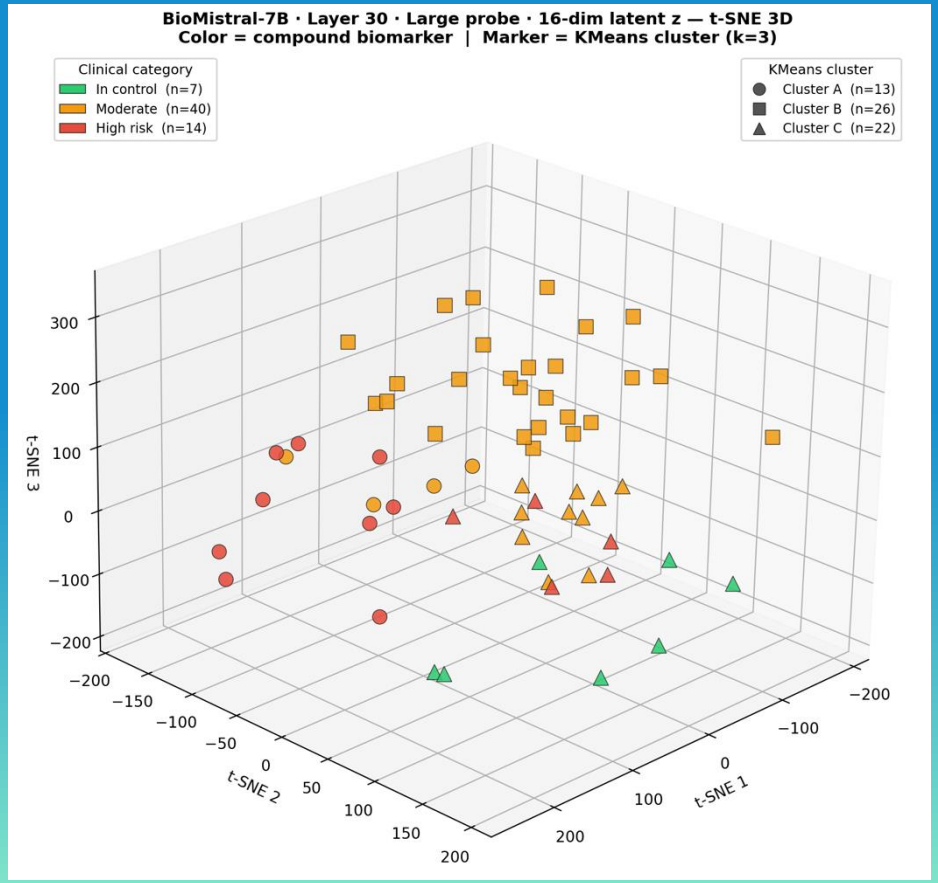
The model learns to reconstruct floating-point biomarker readouts from internal representations. Performance is then evaluated via AUC by classifying predictions into normal (<10) versus abnormal (>10) ranges.



Patient Representation

We derive patient representations from the trained LLM+probe system by extracting a compact latent embedding from the probe's bottleneck layer.

A learned 16-dimensional latent vector that captures the most task-relevant patient information. These embeddings are then visualized with t-SNE, providing an interpretable view of patient structure in a reduced-dimensional space.



Feature of Multi-biomarker Profile

PATIENT ID 198a322d-e0b6-4fec-9336-f7a9972e104d

GI-MAP

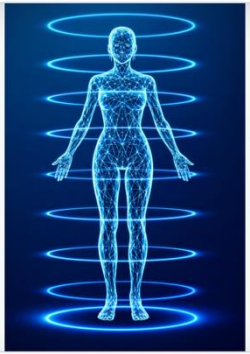
H. pylori: 577 org/g
Campylobacter: not detected
C. difficile Toxin A: not detected

Preview PDF

MYCOTOXIN

Ochratoxin A: 1.058 ppb
Aflatoxin Group (B1, B2, G1, G2): 0.358 ppb
Trichothecene Group (Macrocyclic): Roridin: 0.12 ppb

Preview PDF



PATIENT INTAKE

Sex: Female

Preview PDF

MARCONS

Positivity: Positive
Report ID: 14259126

Preview PDF

LABCORP BLOOD

glucose mg/dL: 78
bun mg/dL: 13
creatinine mg/dL: 0.97

Preview PDF

IMMUNE / LONG COVID

panel type: long_covid_immune
panel name: LYME MULTI-PEPTIDE ELISA ASSAY / IMMUNOSEROLOGY OF LYME

Preview PDF

Conclusions

Clinical tokens should not just encode numbers but meaningfully represent biological processes and physiological states to better reflect how the human body functions

By combining biomarker data with context, LLMs move beyond language understanding toward a deeper understanding of patients and real-world health conditions

In the era of large-scale AI, this unlocks a paradigm shift from reactive, episodic patient visits to continuous, intelligent monitoring that transforms care delivery

THANKS

contact@joexing.me

DO YOU HAVE ANY QUESTIONS?