AI and VistA

Use of SLMs and LLMs for Modernization and Sustainment

Augie Turano Ph.D.
Consultant
Adjunct Associate Professor
University of Pittsburgh
Health Informatics

Augie.Turano@gmail.com

VistA Community Meeting
February 20 2026

It's not what you look at that matters, it's what you see.
Henry David Thoreau

What I want to discuss…

- Modernizing VistA
- Enhancing Interoperability
- Having AI help with VistA code generation and maintenance
- TIU case study
- Using AI with Small and Large Language Models
- Optimizing VistA Data Search/Find

Code Development/Maintenance Acceleration

- “The role of the engineer is changing fundamentally. You will spend less time writing syntax and debugging, and more time commanding the computer to execute complex intent.” — Sam Altman (OpenAI)
- Shift from writing code to generating outcomes, write prompt, verify output, change prompt, verify again until goal is achieved.
- Jevons Paradox - making code easier and faster to produce often leads to more code being written overall, not less, because the lowered effort encourages greater demand for new features and rapid iteration.

AI Acceleration

- Ryan Dahl, creator of Node.js, posted a **viral take** arguing that AI tools have ended the era of manually writing code. He's not saying developers are obsolete, just that their role is shifting from writing syntax to higher-level work like system design, architecture and validating AI outputs.
- In November 2025, Anthropic released a study based on 100,000 real Claude conversations. Many were surprised by its conclusion: AI reduces task-solving time by 5x.

Anthropics Claude Building a C compiler

The Achievement

- Codebase: Composed ~100,000 lines of Rust in just 14 days.
- Human Role: Humans provided specifications and validation; no human wrote a single line of the compiler code.
- Validation: Successfully passes the vast majority of the GCC torture test suite.

Real-World Impact

Successfully compiled complex, industry-standard systems software:

- Linux Kernel & QEMU, PostgreSQL,Redis, FFmpeg, SQLite, more.
- Cost $20,000

The Bottom Line

Microsoft's Steve Sinofsky further qualified the claim that Claude did in two weeks the work that took human engineers 37 years

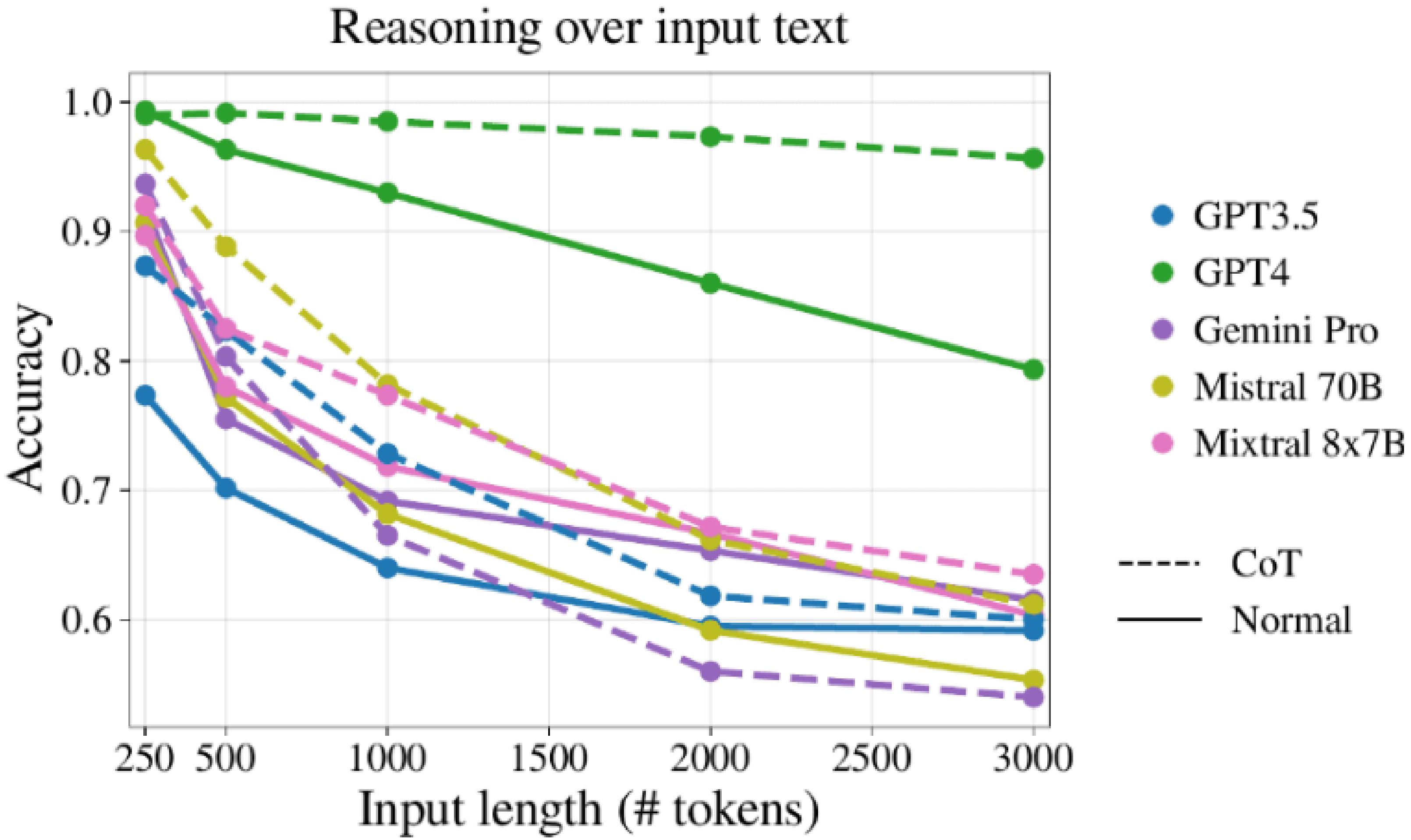
Sources: https://www.anthropic.com/engineering/building-c-compiler

Model Right-Sizing: A Framework for Selecting SLMs Over LLMs in Health Care Systems

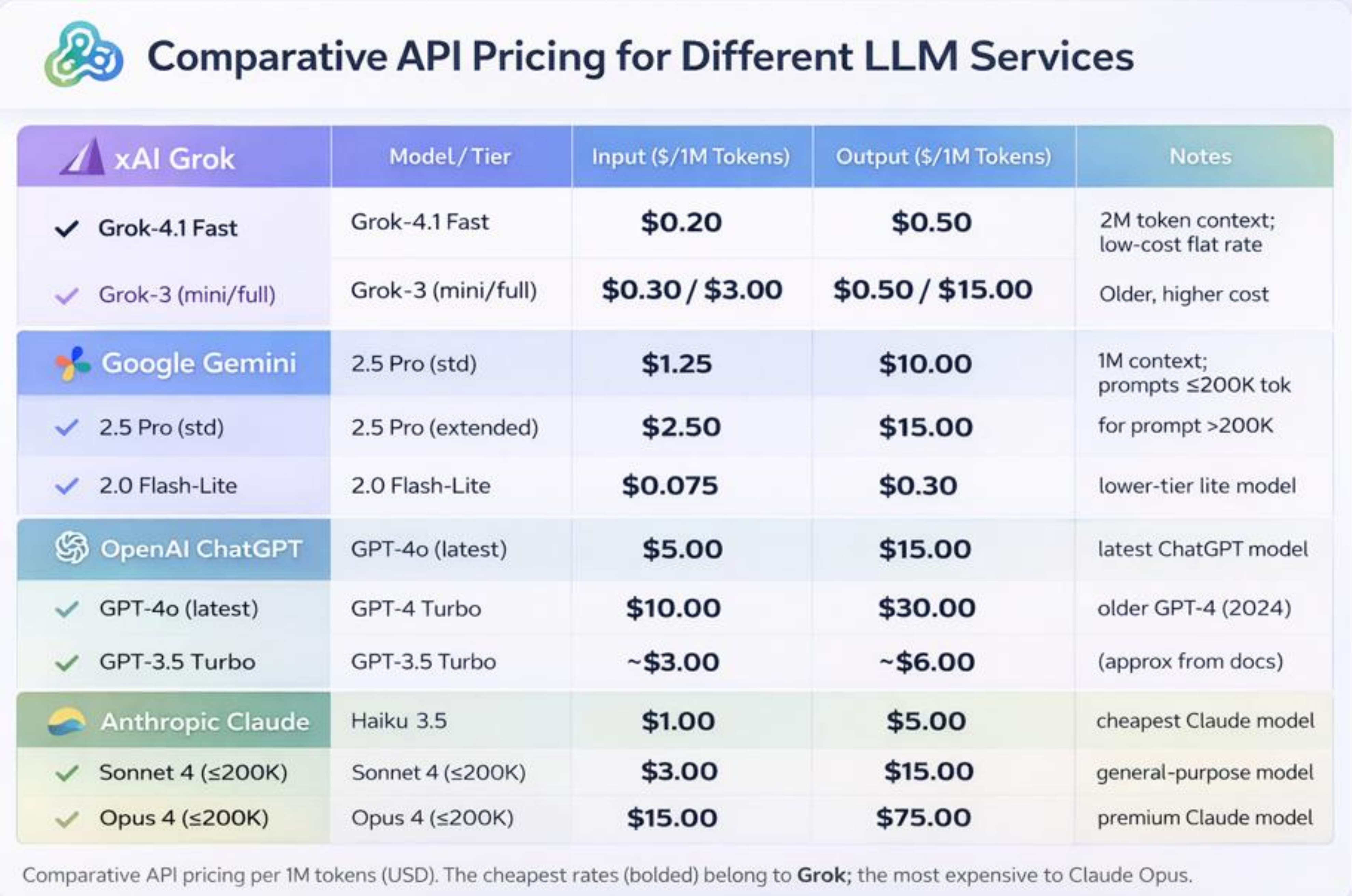
- **We don't need more powerful LLMs for many tasks.** Models need context about your work and specific training. Without this context, AI can't deliver what you need. (Few Shot examples, Fine tuning)
- **Context Windows** – Result worsen as context gets large
- **MOE** – Mixture of Experts small SLMs can be tuned for specific work.
    - Multi-SLM + RAG + Functionality Router
- **SLM's –** process data locally, no need for API keys, Cloud APIs and high costs.

SLM = Small Language Model
LLM = Large Language Model
Life on the edge – SLMs/LLMs on devices.

The Inverse Relationship Between Prompt Length and LLM Accuracy



https://gritdaily.com/impact-prompt-length-llm-performance/



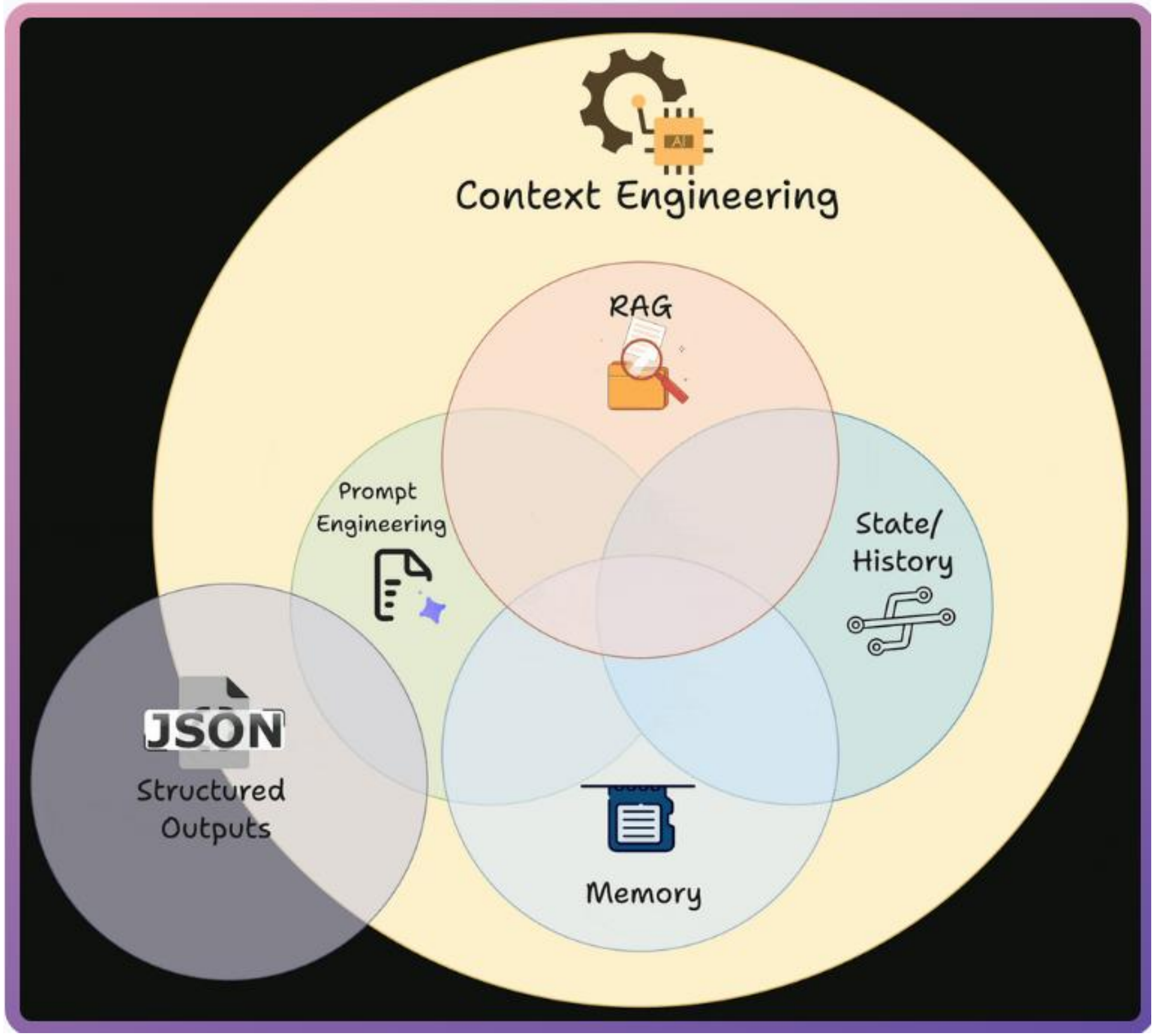
Comparative API Pricing for Different LLM Services

xAI Grok	Model/Tier	Input ($/1M Tokens)	Output ($/1M Tokens)	Notes
✓ Grok-4.1 Fast	Grok-4.1 Fast	**$0.20**	**$0.50**	2M token context; low-cost flat rate
✓ Grok-3 (mini/full)	Grok-3 (mini/full)	**$0.30 / $3.00**	**$0.50 / $15.00**	Older, higher cost
Google Gemini	2.5 Pro (std)	**$1.25**	**$10.00**	1M context; prompts ≤200K tok
✓ 2.5 Pro (std)	2.5 Pro (extended)	**$2.50**	**$15.00**	for prompt >200K
✓ 2.0 Flash-Lite	2.0 Flash-Lite	**$0.075**	**$0.30**	lower-tier lite model
OpenAI ChatGPT	GPT-4o (latest)	**$5.00**	**$15.00**	latest ChatGPT model
✓ GPT-4o (latest)	GPT-4 Turbo	**$10.00**	**$30.00**	older GPT-4 (2024)
✓ GPT-3.5 Turbo	GPT-3.5 Turbo	**~$3.00**	**~$6.00**	(approx from docs)
Anthropic Claude	Haiku 3.5	**$1.00**	**$5.00**	cheapest Claude model
✓ Sonnet 4 (≤200K)	Sonnet 4 (≤200K)	**$3.00**	**$15.00**	general-purpose model
✓ Opus 4 (≤200K)	Opus 4 (≤200K)	**$15.00**	**$75.00**	premium Claude model

Comparative API pricing per 1M tokens (USD). The cheapest rates (bolded) belong to **Grok**; the most expensive to Claude Opus.

https://intuitionlabs.ai/articles/ai-api-pricing-comparison-grok-gemini-openai-claude

An LLM’s Context Window Functions As Working Memory



https://www.dailydoseofds.com/llmops-crash-course-part-7/

Some Modernization Techniques of EHRs in General using AI (LLMs)

- EHRs have to many clicks, data hidden in nooks and crannies – NLP requests to show data can improve it
- Treatment summaries – AI/NLP
- Use AI to homogenize EHR data – convert terminologies, code sets, merge data from different sources.
- AI in coaching for treatments (rx’s, recommend tests etc)
- Deal with interop in much better ways – translate unstructured data to structured data
- Reduce administrative burden – automatic coding
- AI listening to Doctor/Patient dialog – summarizing – predicting course of action?

VistA Many Opportunities

- **Intelligence Layer :** Enabling natural language "Chat with EHR" by bridging VistA and Millennium data. Deploying LLMs for automated text summarization, rapid Q&A, and advanced NLP.
- **Data Retrieval/ Technical Access**: $GET Globals,VX130, File Man Queries, APIs, RPCs
- **Cloud** entirely? Cloud partially? All data in AWS.
- **Rich Media/ Storage**? FileMan, Objects, SQL, Key/Value stores, vectors, audio, video, graphics, Images?
- **VX130, VistA, VDIF, CDW, VIDA**? - VistA Aggregated data structures – map them, analyze the data
- **NLP/NER** – finding data in notes – there are a lot of them and they are important, Ambient Listening to capture dialog
- **Keyword Search and Vector Stores** – vectors necessary for embeddings for similarity Search, , Agentic AI; For keyword: BM25, TF/IDF.
- **Better Interop** – **:** Streamlining data exchange between the DoD, Oracle, and Community Care providers.
- **Differential Diagnosis** – Tackling high-complexity clinical reasoning,this is hard and needs big guns

Advanced Functionality For VistA as Oracle Migration Continues

In VA VistA runs on the **IRIS platform, IRIS/HealthShare** provides:

- **Object creation** from VA File Manager Files
- **ODBC/JDBC/REST/APIs** link to external sources
- **SQL capabilities** – this expands support staff and coding/report capabilities (Hook to PowerBI, Excel, Tableau,etc)
- **Python** Implementation
    - Facilitates use of OCR – interpret incoming faxes
    - PDF reading/data extraction
    - Data analysis – Pandas, Polars etc data Libraries
    - AI/ML, LLMs – enhance code, detect trends

Assume optimistically that 2 VAMCs are converted each month starting in April 2026.

- It would take more than 6 years to convert 170 VAMCs.
- Many VistA functions go beyond the EHR and there are areas not covered by Millennium
- **AI can assist with new code** and maintenance

Modest Goals for Increasing VistA Value

- Provide a **hybrid architecture** that is based on a creating VistA **objects**, **Microservices** using both VistA **global storage** and a standard reference model like SDA (software document architecture)or other models. (FHIR, Document)
- Provide VistA core object infrastructure in IRIS Objects, including the development platform, tools, utilities, distribution methods, standards and guidelines and use FileMan more sparingly and utilize **IRIS SQL, Python** like **SLMs**, **LLMs, NLP with vector stores**.
- Ensure a start-up set of core VistA objects for healthcare applications is in place. **Show value incrementally**.
- Use **LLMs, SLMs for VistA code description and creation**, also creating and decoding healthcare message in FHIR, HL7, CCDA etc. this would allow for better community care interop.
- **Expand the diversity of data ingestion** to include graphics, faxes, PDFs etc parse and turn that into usable structured data.

*LLM= Large Language Model, SLM=small language model.

Getting away from Menu Clicks

- Humans think in conversation, not menus
- Menus assume you already know where to look for data
- Chat becomes the primary interaction layer:
  - ➢Order 2 500 Mg tablets of Tylenol for patient John Doe.; What is his last H1AC? (if there is ambiguity, it can be handled with human interaction)
- Screens still matter and can be intertwined with NLP.
- Query/Chat – can be voice, ambient listening etc LLM/SLM can evaluate queries and direct to the appropriate source.

Increasing Outsourcing of VA Care

- On Dec 15, 2025, VA released a Request for Proposals (RFP) for new Community Care contracts that will govern the next decade of outsourced care.
- The VA is restructuring Community Care into **two massive national regions** with contracts worth **up to $1 trillion**, signaling a major long-term expansion of outsourced care. Spending on community care has risen sharply, reaching **nearly $30 billion in FY2023**, and VA plans for **$1 trillion in private-sector care purchases over the next decade**.
  - https://empowr.us/wp-content/uploads/2024/04/VA-Red-Team-Executive-Community-Care-Roundtable-Report-post.pdf
  - https://www.stripes.com/veterans/2026-01-22/democrats-va-community-care-proposal-20491262.html

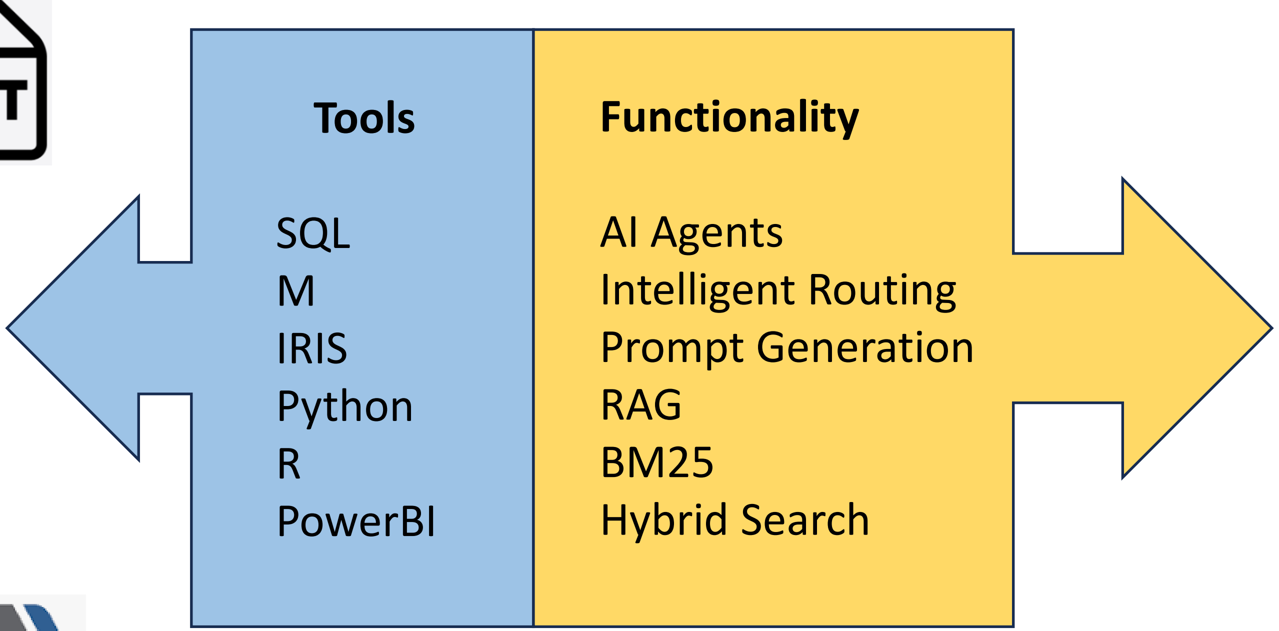
VA to reorganize community care contracts, reducing regions to 2

  - https://www.militarytimes.com/veterans/2025/12/15/va-to-reorganize-community-care-contracts-reducing-regions-to-2/

Aggregation and Amalgamation at Scale

TXT
PDF
SQL
VistA
ORACLE Health
NIH
CDC

Small SLMs or Large LLMs
Local or Frontier Labs

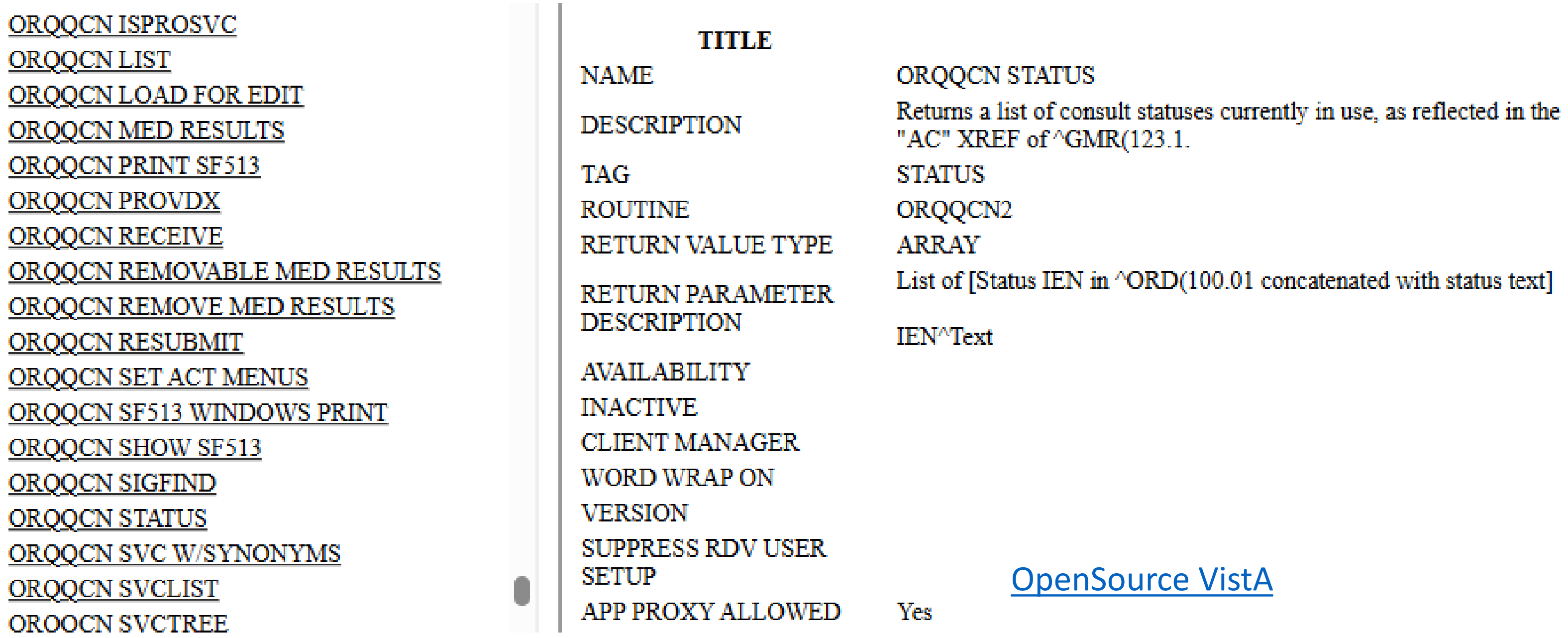


Tools	**Functionality**
SQL	AI Agents
M	Intelligent Routing
IRIS	Prompt Generation
Python	RAG
R	BM25
PowerBI	Hybrid Search

- **PHI data can stay local if necessary.**
- **Choosing the correct functionality based on input, then execute.**
- **Divide a large problem space into smaller, manageable ones**

Augment API catalog – Developer Kits

- RPCs – identified for key functions, aggregated and identified
- LLMs understand the RPCs and can help to rework or refactor



ORQQCN ISPROSVC
ORQQCN LIST
ORQQCN LOAD FOR EDIT
ORQQCN MED RESULTS
ORQQCN PRINT SF513
ORQQCN PROVDX
ORQQCN RECEIVE
ORQQCN REMOVABLE MED RESULTS
ORQQCN REMOVE MED RESULTS
ORQQCN RESUBMIT
ORQQCN SET ACT MENUS
ORQQCN SF513 WINDOWS PRINT
ORQQCN SHOW SF513
ORQQCN SIGFIND
ORQQCN STATUS
ORQQCN SVC W/SYNONYMS
ORQQCN SVCLIST
ORQQCN SVCTREE

TITLE	
NAME	ORQQCN STATUS
DESCRIPTION	Returns a list of consult statuses currently in use, as reflected in the "AC" XREF of ^GMR(123.1.
TAG	STATUS
ROUTINE	ORQQCN2
RETURN VALUE TYPE	ARRAY
RETURN PARAMETER DESCRIPTION	List of [Status IEN in ^ORD(100.01 concatenated with status text] IEN^Text
AVAILABILITY	
INACTIVE	
CLIENT MANAGER	
WORD WRAP ON	
VERSION	
SUPPRESS RDV USER SETUP	
APP PROXY ALLOWED	Yes

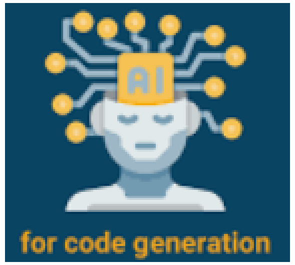
OpenSource VistA

Legacy Modernization is a mandatory prerequisite for AI.

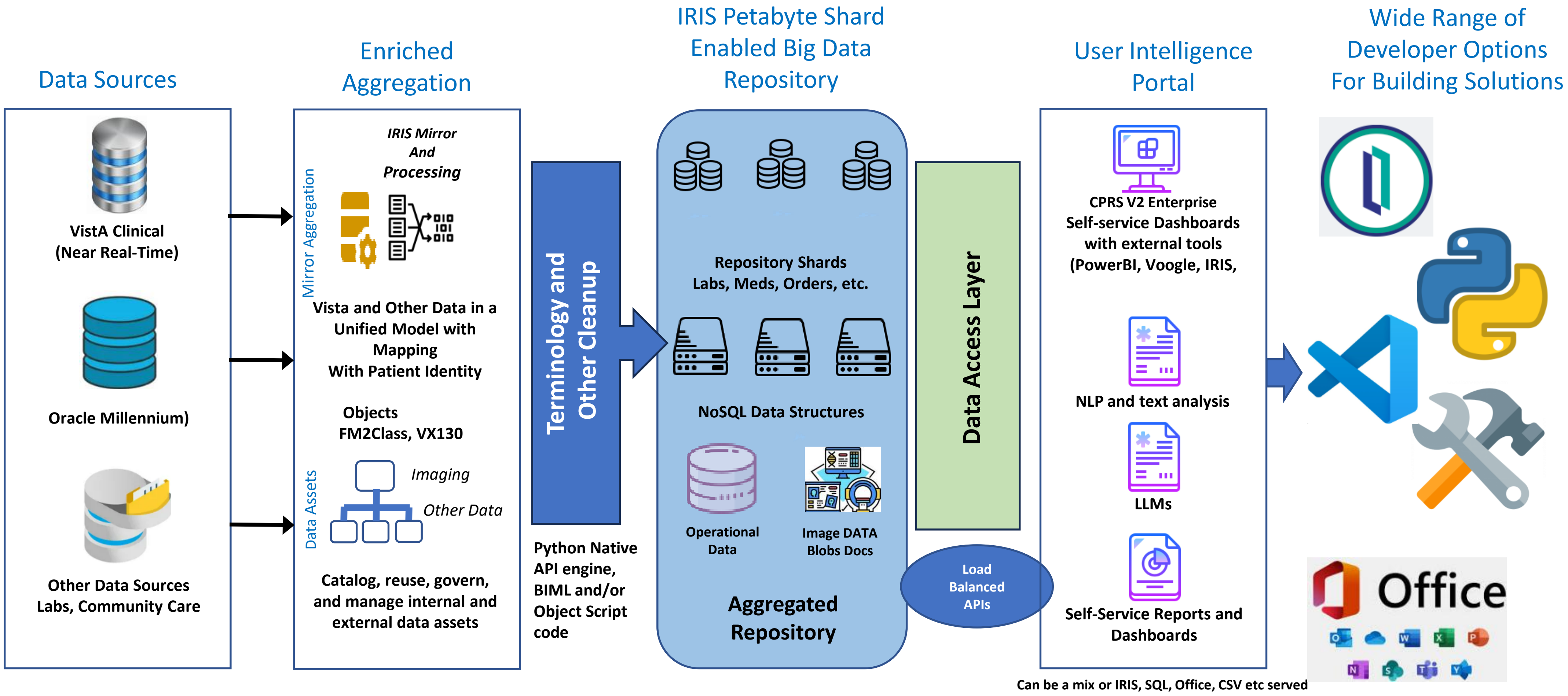
Grok, OpenAI, Claude, Gemini, Understand VistA

- The above LLMs were tested in code generation for VistA.
- Tests yielded the following:
  - FileMan utilities were used when prompt indicted to use them.
  - The LLM can generate M code or IRIS object script
  - All LLMs knew the location of the Globals, Grok and Claude were particularly good at relating VistA M routines to specific tasks like orders or labs.
  - AI understanding adds another dimension to VistA sustainability and enhancements.

Using AI for coding is like the jump from assembly to high level languages.



VistA Architecture Enhanced



Data Sources

VistA Clinical (Near Real-Time)

Oracle Millennium)

Other Data Sources Labs, Community Care

Enriched Aggregation

Mirror Aggregation

IRIS Mirror And Processing

Vista and Other Data in a Unified Model with Mapping With Patient Identity

Objects FM2Class, VX130

Data Assets

Imaging

Other Data

Catalog, reuse, govern, and manage internal and external data assets

Terminology and Other Cleanup

Python Native API engine, BIML and/or Object Script code

IRIS Petabyte Shard Enabled Big Data Repository

Repository Shards Labs, Meds, Orders, etc.

NoSQL Data Structures

Operational Data

Image DATA Blobs Docs

Aggregated Repository

Data Access Layer

Load Balanced APIs

User Intelligence Portal

CPRS V2 Enterprise Self-service Dashboards with external tools (PowerBI, Voogle, IRIS,

NLP and text analysis

LLMs

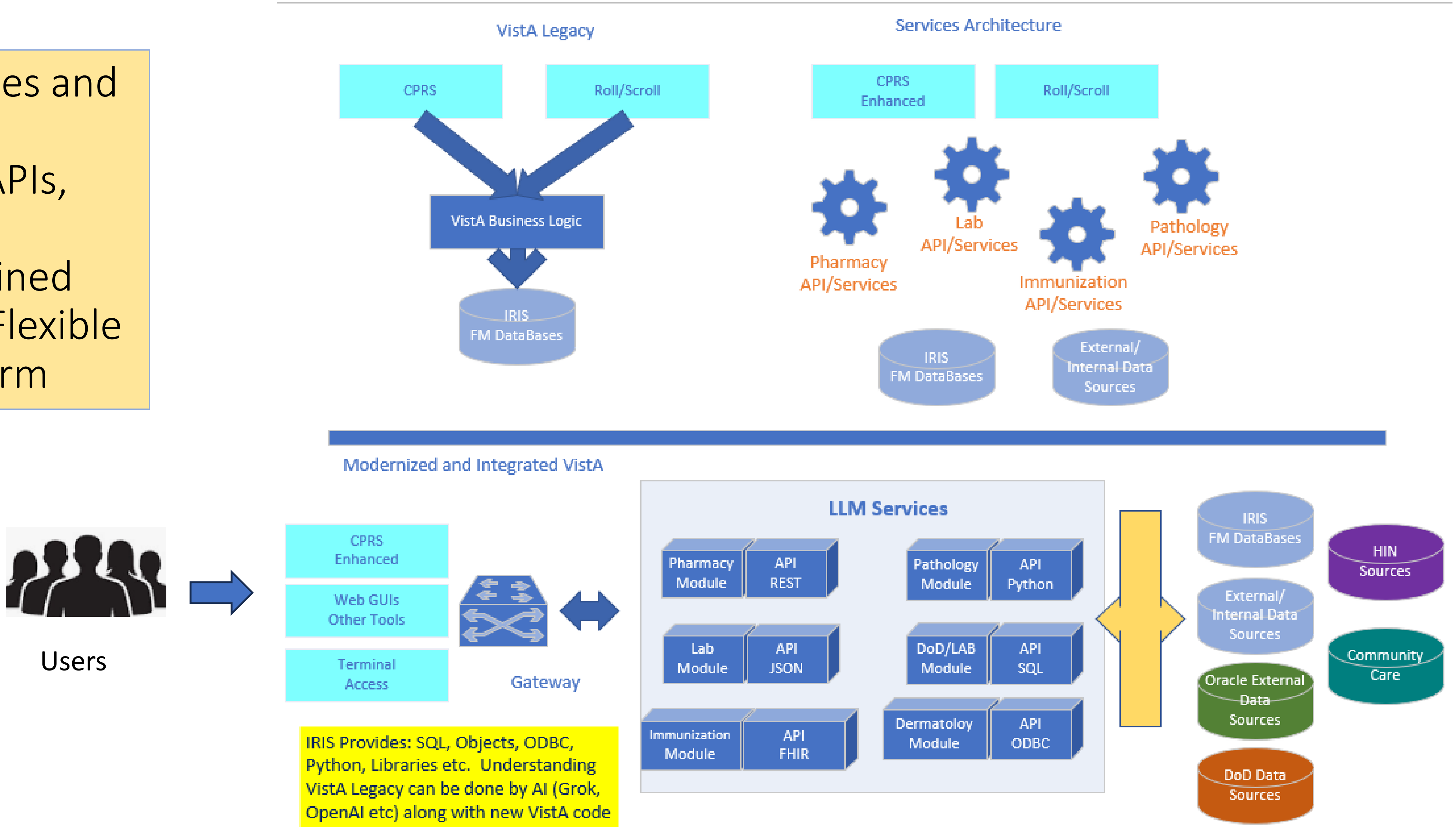
Self-Service Reports and Dashboards

Can be a mix or IRIS, SQL, Office, CSV etc served

Wide Range of Developer Options For Building Solutions

Modern Architecture Interop

Services and SLMs, LLM,APIs, DBs combined for a Flexible Platform



VistA Legacy

CPRS

Roll/Scroll

VistA Business Logic

IRIS FM DataBases

Services Architecture

CPRS Enhanced

Roll/Scroll

Pharmacy API/Services

Lab API/Services

Immunization API/Services

Pathology API/Services

IRIS FM DataBases

External/ Internal Data Sources

Modernized and Integrated VistA

Users

CPRS Enhanced

Web GUIs Other Tools

Terminal Access

Gateway

LLM Services

Pharmacy Module

API REST

Pathology Module

API Python

Lab Module

API JSON

DoD/LAB Module

API SQL

Immunization Module

API FHIR

Dermatoloy Module

API ODBC

IRIS FM DataBases

HIN Sources

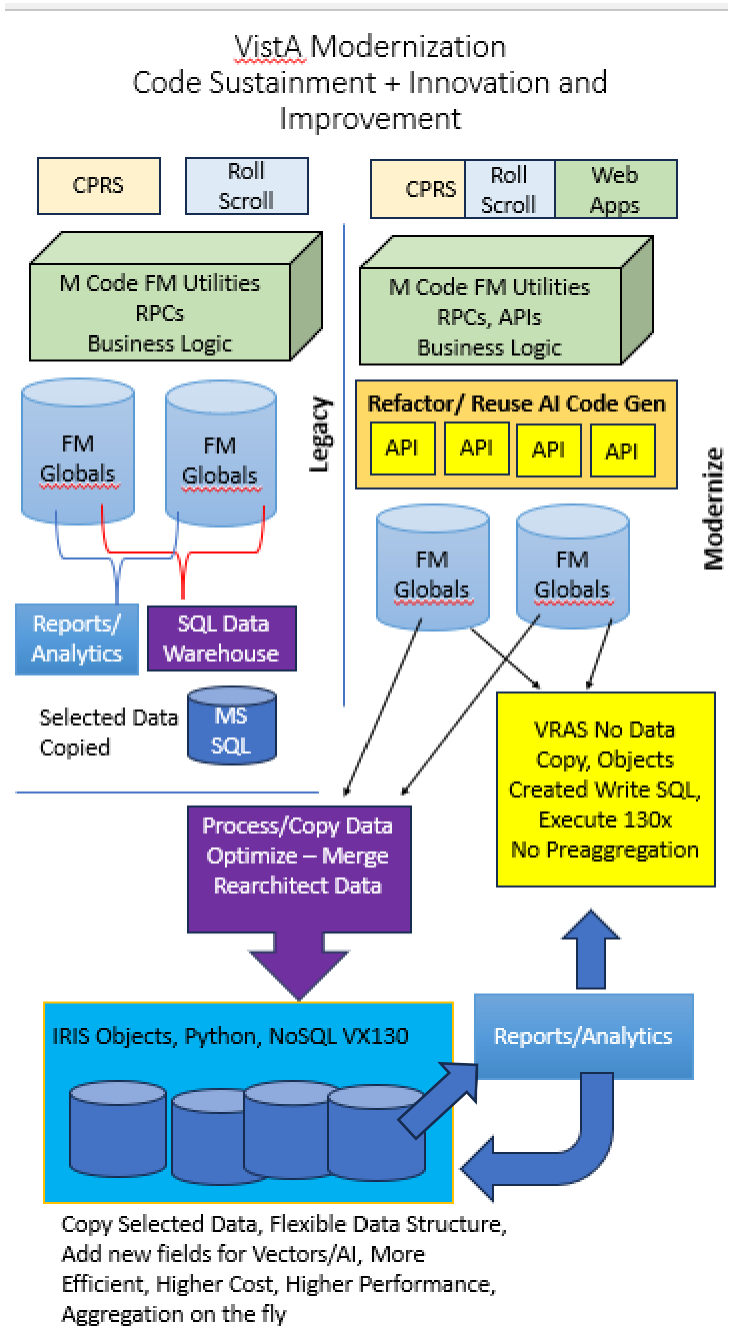
External/ Internal Data Sources

Community Care

Oracle External Data Sources

DoD Data Sources

IRIS Provides: SQL, Objects, ODBC, Python, Libraries etc. Understanding VistA Legacy can be done by AI (Grok, OpenAI etc) along with new VistA code generation.


VistA Modernization
Code Sustainment + Innovation and Improvement
CPRS
Roll Scroll
M Code FM Utilities
RPCs
Business Logic
FM Globals
FM Globals
Legacy
Reports/ Analytics
SQL Data Warehouse
Selected Data Copied
MS SQL
CPRS
Roll Scroll
Web Apps
M Code FM Utilities
RPCs, APIs
Business Logic
Refactor/ Reuse AI Code Gen
API
API
API
API
FM Globals
FM Globals
Modernize
VRAS No Data Copy, Objects Created Write SQL, Execute 130x No Preaggregation
Process/Copy Data Optimize – Merge Rearchitect Data
IRIS Objects, Python, NoSQL VX130
Reports/Analytics
Copy Selected Data, Flexible Data Structure, Add new fields for Vectors/AI, More Efficient, Higher Cost, Higher Performance, Aggregation on the fly

Refactoring VistA code to APIs for Analytics, Reporting and Interop.

Create Higher level APIs to simplify coding.

Real Time Data Access Across the Enterprise

Why not chat with the data with LLMs?

Prompt: lab result and put in HL7 format: patient: augie turano, lab result date: 4 25 2025, performed at UPMC hospital Montifiore; test is hemoglobin a1c from blood sample, result is 5.6, ordered by: dr. marcus welby

```
MSH|^~&|UPMC|UPMC||UPMC|202504251200||ORU^R01|20250425120001|P|2.5
PID|1||TURANO^AUGIE||TURANO^AUGIE||19700101|M|||123 MAIN
ST^^PITTSBURGH^PA^15213^USA||555-123-4567
PV1|1|O|MONTEFIORE^^^^UPMC||||WELBY^MARCUS^^^MD
OBR|1||20250425HBA1C|83036^Hemoglobin
A1c^CPT||202504250800|||N||||202504250800||WELBY^MARCUS^^^MD|||||202504251200|||F
OBX|1|NM|83036^Hemoglobin A1c^CPT||5.6|%|5.0-6.0|N|||F
```

Then asked to interpret each field:

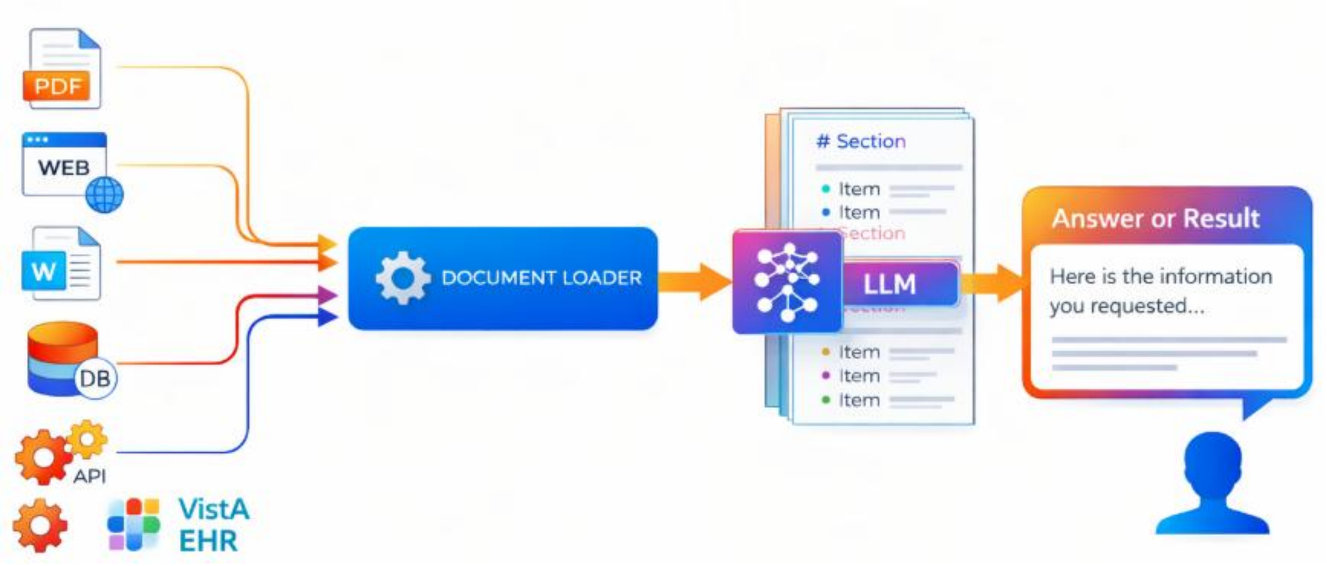
Can you describe each variable in the following HL7 message?

```
MSH|^~&|UPMC|UPMC||UPMC|202504251200||ORU^R01|20250425120001|P|2.5
PID|1||TURANO^AUGIE||TURANO^AUGIE||19700101|M|||123 MAIN
ST^^PITTSBURGH^PA^15213^USA||555-123-4567
PV1|1|O|MONTEFIORE^^^^UPMC||||WELBY^MARCUS^^^MD
OBR|1||20250425HBA1C|83036^Hemoglobin
A1c^CPT||202504250800|||N||||202504250800||WELBY^MARCUS^^^MD|||||202504251200|||F
OBX|1|NM|83036^Hemoglobin A1c^CPT||5.6|%|5.0-6.0|N|||F
```

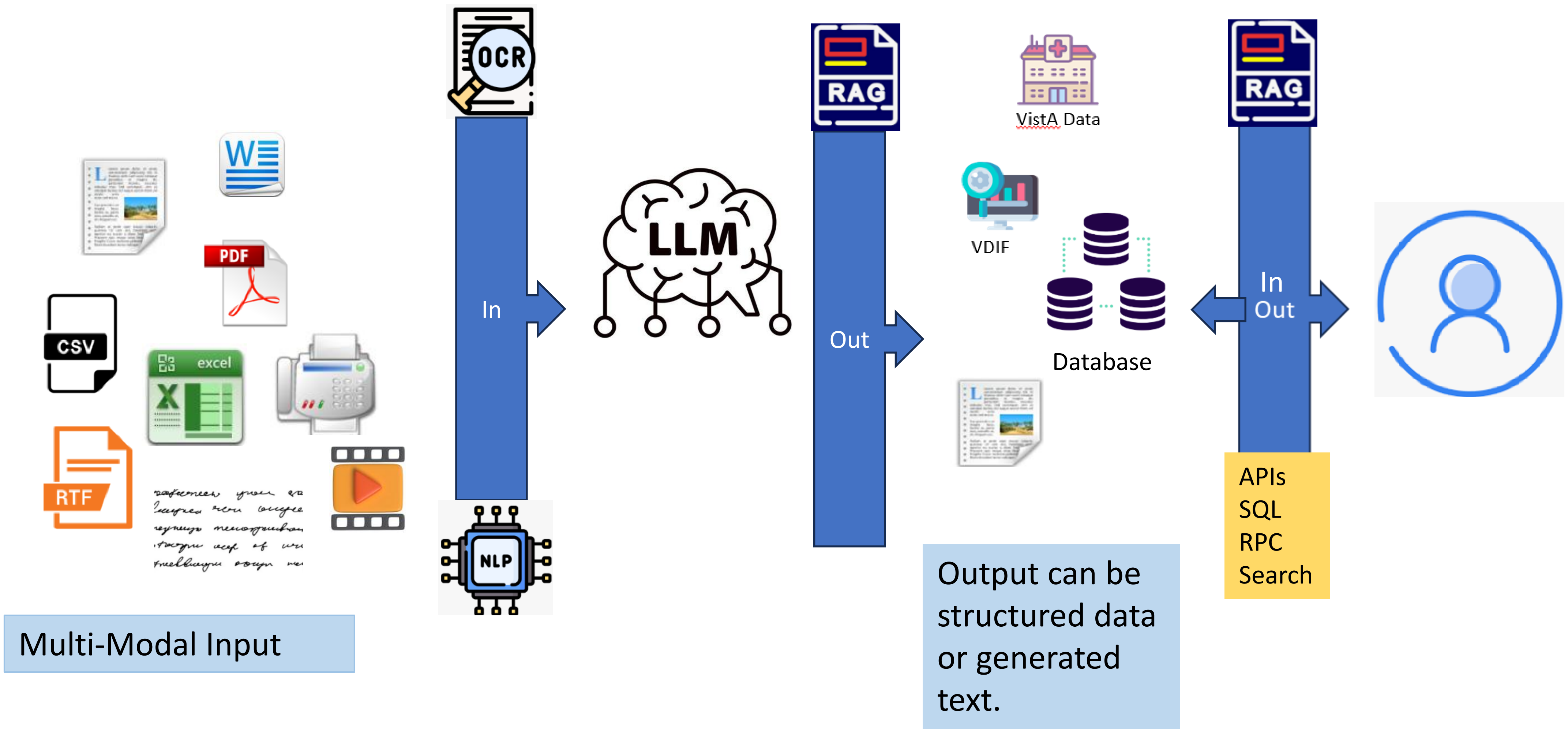
VistA data is hierarchical and idiosyncratic, but to interact with foreign systems standards like HL7 or FHIR are important. Epic/Oracle and many other healthcare providers have FHIR interfaces, LLMs can be very useful for assisting in coding for that data exchange.

Clinicians Need to Find Information

- Community care increasing at a rapid pace, which means data can be coming in a variety of formats. Easy to hook to TIU notes.
- NER (Named Entity Recognition)
- Search (Use AI router to decide best option)
  - Key Word
  - BM25,
  - Similarity Search (embeddings and vector search)
  - Structured



Community Care Consolidation to Usable Data



Multi-Modal Input

In

Out

VistA Data

VDIF

Database

In
Out

APIs
SQL
RPC
Search

Output can be structured data or generated text.

Chatting with VistA + (Millennium)

The clinician enters a question such as:

- Does this person have any allergies?
- What medications is this person taking?
- What surgeries has this person had?
- What does this person's latest A1C test show?
- Has this person had a colonoscopy? Were the results normal? Show references.

VA gets data back daily from the Oracle Millennium EHR

Deciding a Path (Picking a Route for Answer)

Given a clinical request a router decides:

- Is this a structured query? → SQL
  - Labs, vitals, meds, allergies
- Is this a keyword/similarity query? → BM25 + embeddings
  - Clinical notes
- Is this a reasoning question? → LLM
  - Explain why the creatinine is rising.
- Is this a domain-specific extraction task? → SLM
  - NER, ICD/SNOMED/CPT extraction, problem list extraction

Future?

- Imagine a EHR or CPRS that doesn’t just have a fixed menu but can rewrite a custom interface based on how one works best.
- **Current State:** Users adapt to the software.
- **Future State:** Software adapts to the user (in real-time).
- Can be a **massive opportunity** for developers.
  - Don’t build static tools. Build malleable tools that can adapt to each users requirements.
  - Always have a dialog command interface to either request or enter data. (LLM can screen and perform inserts/writes or reads in NRT.
  - Prompt engineering is a required skill for developers
- AI is not a tool you query. It’s an agent you direct.

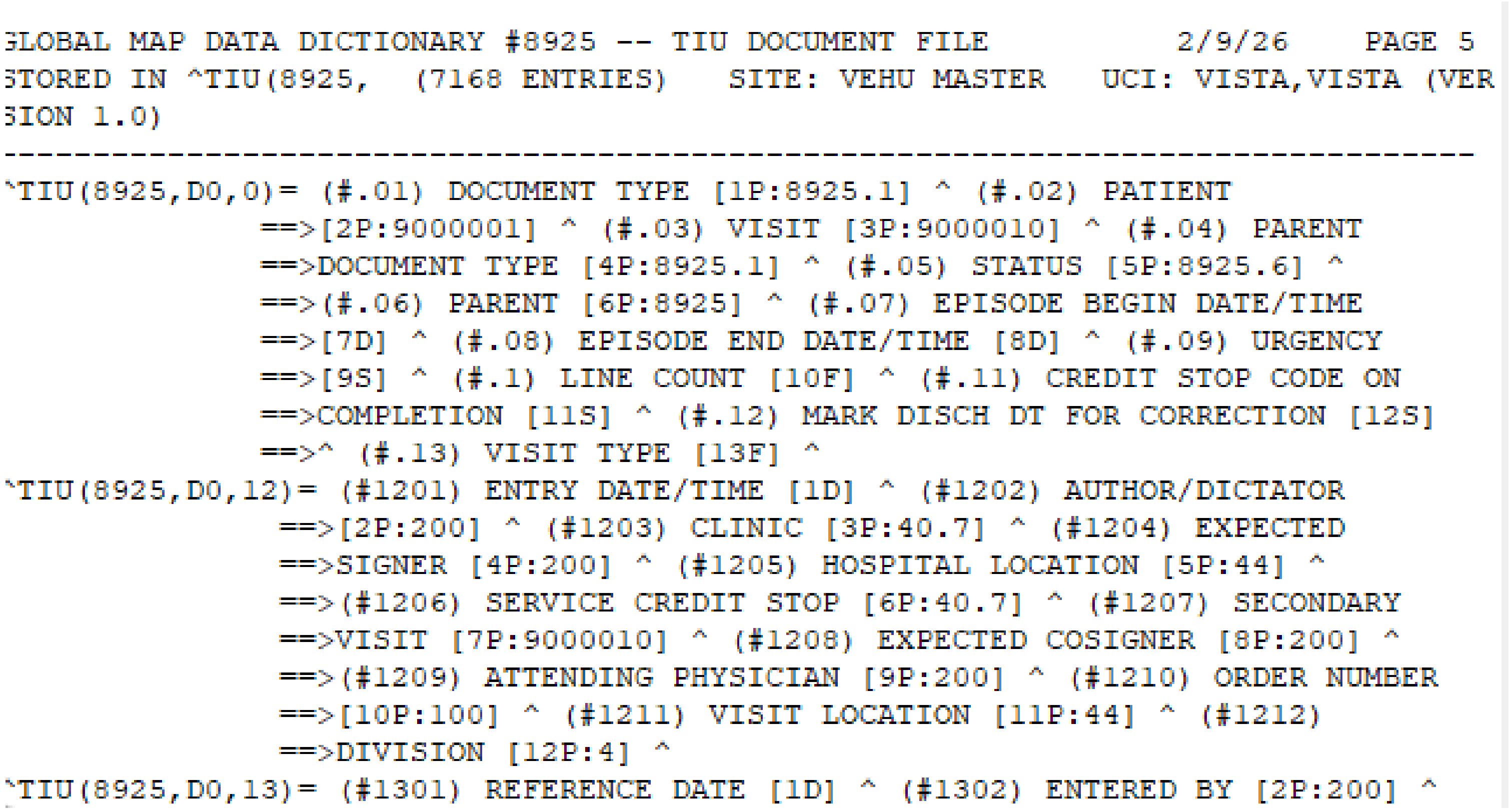
Conclusion

- Individuals who have the experience, expertise, and competence can effectively use LLM to enhance creativity and productivity
- Examine smaller, efficient models, benchmark them against operational constraints, and continuously refine them through iterative experimentation and error analysis.
- AI is different from traditional coding: Coding has bugs, you fix them (it's binary), in AI can't use classical tests (unit tests) to test against AI bugs, you need verification techniques.
- English-to-code is just another abstraction layer. Historically Assembly programmers became C/Basic/Pascal programmers became Python programmers. More software was built, there were not fewer programmers.
- Don't judge AI solely on free tiers, it's like comparing a flip phone with an Iphone 17.

Demos

1. Interop – HL7, FHIR
2. Text to Interop HL7 and FHIR.
3. VistA Code Explanation, Documentation
4. TIU Case Study
    - Interop, but data from many areas
    - Convert PDFs, PNG, Fax (TIFF), Word etc to text for addition to TIU. (then usable quickly by clinicians with Voogle or other search
    - TIUIN – take input above and put into TIU global
    - GROK – flawless generation of VistA TIU Note Display
5. LLM/SLM – generation of code mapped to objects, fall back to LLM or SLMs IRISRAG, NER

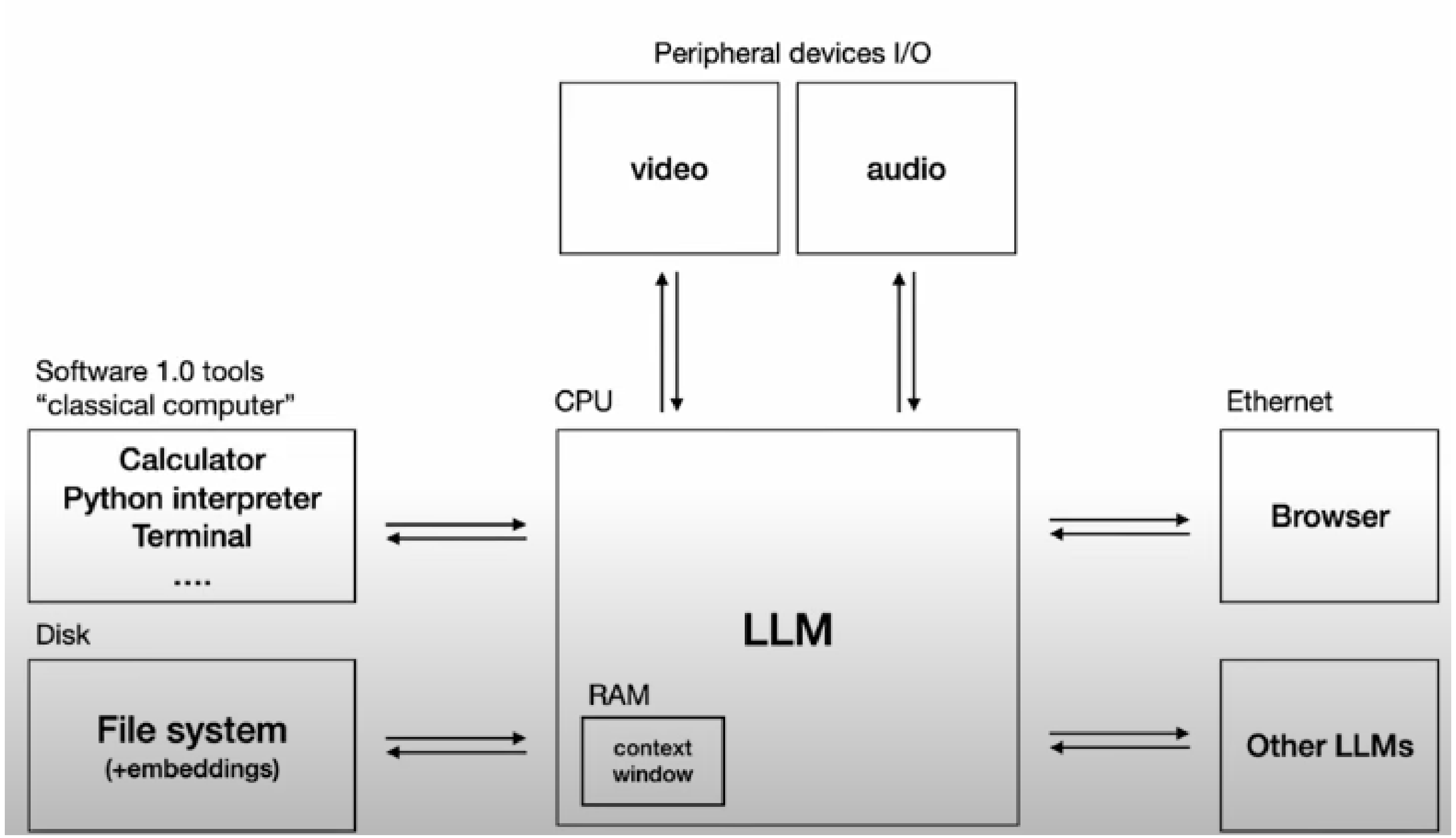
TIU Partial Schema (Clinical Notes)



```
GLOBAL MAP DATA DICTIONARY #8925 -- TIU DOCUMENT FILE          2/9/26    PAGE 5
STORED IN ^TIU(8925,  (7168 ENTRIES)   SITE: VEHU MASTER   UCI: VISTA,VISTA (VER
SION 1.0)
-------------------------------------------------------------------------------
^TIU(8925,D0,0)= (#.01) DOCUMENT TYPE [1P:8925.1] ^ (#.02) PATIENT
              ==>[2P:9000001] ^ (#.03) VISIT [3P:9000010] ^ (#.04) PARENT
              ==>DOCUMENT TYPE [4P:8925.1] ^ (#.05) STATUS [5P:8925.6] ^
              ==>(#.06) PARENT [6P:8925] ^ (#.07) EPISODE BEGIN DATE/TIME
              ==>[7D] ^ (#.08) EPISODE END DATE/TIME [8D] ^ (#.09) URGENCY
              ==>[9S] ^ (#.1) LINE COUNT [10F] ^ (#.11) CREDIT STOP CODE ON
              ==>COMPLETION [11S] ^ (#.12) MARK DISCH DT FOR CORRECTION [12S]
              ==>^ (#.13) VISIT TYPE [13F] ^
^TIU(8925,D0,12)= (#1201) ENTRY DATE/TIME [1D] ^ (#1202) AUTHOR/DICTATOR
               ==>[2P:200] ^ (#1203) CLINIC [3P:40.7] ^ (#1204) EXPECTED
               ==>SIGNER [4P:200] ^ (#1205) HOSPITAL LOCATION [5P:44] ^
               ==>(#1206) SERVICE CREDIT STOP [6P:40.7] ^ (#1207) SECONDARY
               ==>VISIT [7P:9000010] ^ (#1208) EXPECTED COSIGNER [8P:200] ^
               ==>(#1209) ATTENDING PHYSICIAN [9P:200] ^ (#1210) ORDER NUMBER
               ==>[10P:100] ^ (#1211) VISIT LOCATION [11P:44] ^ (#1212)
               ==>DIVISION [12P:4] ^
^TIU(8925,D0,13)= (#1301) REFERENCE DATE [1D] ^ (#1302) ENTERED BY [2P:200] ^
```

LLM OS

Programming With Prompts



Idea: Andrej Karpathy

Reference Paper: https://huggingface.co/blog/shivance/illustrated-llm-os
[Youtube: https://www.youtube.com/watch?v=LCEmiRjPEtQ&t=616s](https://www.youtube.com/watch?v=LCEmiRjPEtQ&t=616s)

Terms

- LLM/SLM – Large and Small Language Models
- Vector Retrieval – store words as number array, similarity search
- Tokens – words or pieces of words
- Chunking – how to lump tokens
- Prompt Engineering – creating commands and information to tell an LLM what to do or produce
- NER – Named Entity Resolution
- TIU – Text Integration Utility - Notes
- OCR – Optical Character Recognition
- Embeddings
- Neural Network
- Self attention – lets each word look at other words
- Transformers –type of Neural network designed to process sentences, code, time series data.
- MOE – Mixture of experts
- Temperature- the degree of randomness in a generated response.
- Hallucinations – LLM making things up.

SQL-Intended Queries

These should trigger the structured-data path.

- List all patients over age 65 with an A1c > 8.5 in the last year.
- How many admissions for heart failure occurred in January?
- Show me the most recent creatinine value for patient 12345.
- What is the average length of stay for pneumonia patients?
- Give me all lab results for patient 998877 between March 1 and March 15.
- Which patients are currently on both lisinopril and furosemide?
- Count the number of patients with a diagnosis of CKD stage 3.
- Show all prescriptions for metformin written in the last 90 days.

BM25, TF/TDF Keyword Search-Intended Queries

Short, code-like, or policy-like queries, rank documents

- ICD-10 code for “acute pancreatitis”
- Searches for “Diabetes”
- Look for terms that are negation. “Patient denies chest pain.”
- Synonyms: T2D = Type 2 Diabetes
- Long terms supported “Ventilator-associated pneumonia”
- Presentation of all terms per patient.

RAG-Intended Queries

These require retrieving and summarizing local notes, documents, or patient-specific text (could also execute a query for context).

- Summarize this patient’s last three cardiology notes. (risky)
- Why was patient 445566 admitted during their last hospitalization?
- Extract all mentions of medication changes from this note.
- What were the key findings in the discharge summary for patient 123456?
- Summarize the trend of this patient’s blood pressures over the last month.
- Identify all symptoms described in this triage note.
- What procedures has this patient had in the last five years?
- Find all notes mentioning “worsening shortness of breath”.

LLM-Reasoning-Intended Queries

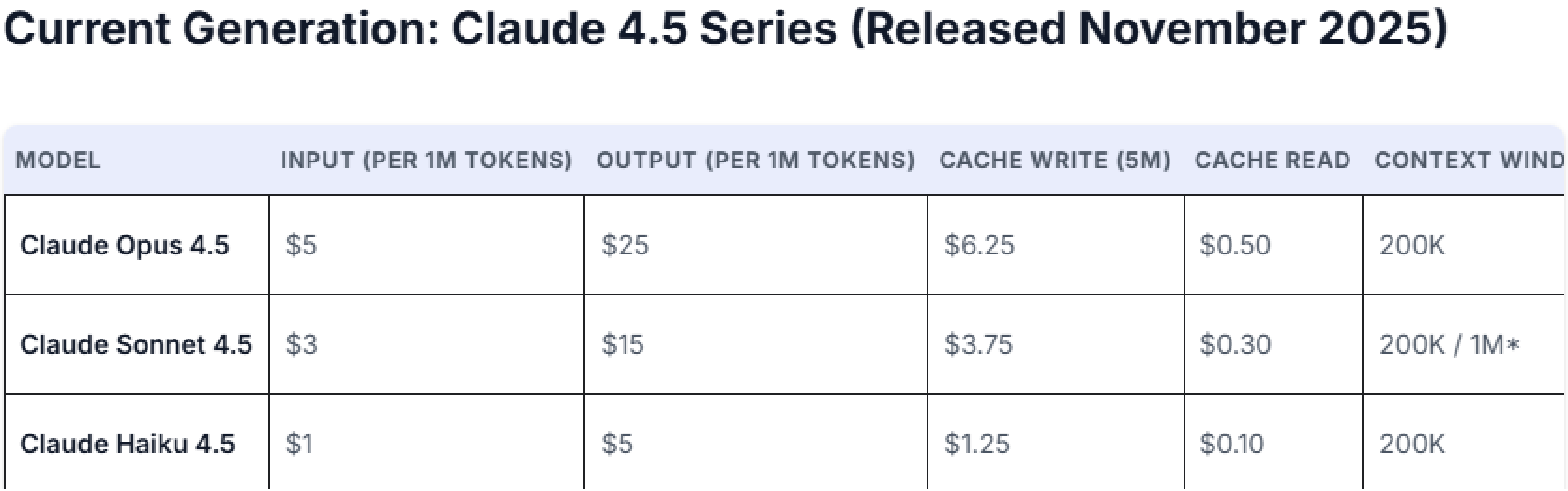
These should not hit SQL or RAG they are conceptual, explanatory, or hypothetical.

- Explain the difference between AKI and CKD to a medical student.
- How does metformin work?
- What are the risks of untreated hypertension?
- Explain why ACE inhibitors can increase creatinine.
- Describe the pathophysiology of diabetic ketoacidosis.
- What lifestyle changes help reduce cardiovascular risk?
- Explain the mechanism of action of beta blockers.

Interesting Quotes

- Jimmy Ba, co-founder of Elon Musk’s xAI, announced his departure. His farewell read less like a corporate goodbye and more like a warning flare: *“We are heading into an era of 100x productivity for anyone with the right tools. The recursive self-improvement loop will likely come online in the next 12 months. 2026 will be a crazy year, likely the busiest and most decisive in the future of our species.”*
- Anthropic CEO Dario Amodei has publicly stated that we are only one to two years away from AI autonomously building the next generation of AI.

Claude Costs



Current Generation: Claude 4.5 Series (Released November 2025)

MODEL	INPUT (PER 1M TOKENS)	OUTPUT (PER 1M TOKENS)	CACHE WRITE (5M)	CACHE READ	CONTEXT WIND
Claude Opus 4.5	$5	$25	$6.25	$0.50	200K
Claude Sonnet 4.5	$3	$15	$3.75	$0.30	200K / 1M*
Claude Haiku 4.5	$1	$5	$1.25	$0.10	200K

https://www.metacto.com/blogs/anthropic-api-pricing-a-full-breakdown-of-costs-and-integration