AI Agents for construction of Longitudinal Patient Profiles from Sparse EHR Data

Gerasimos Grammenos

PhD Candidate

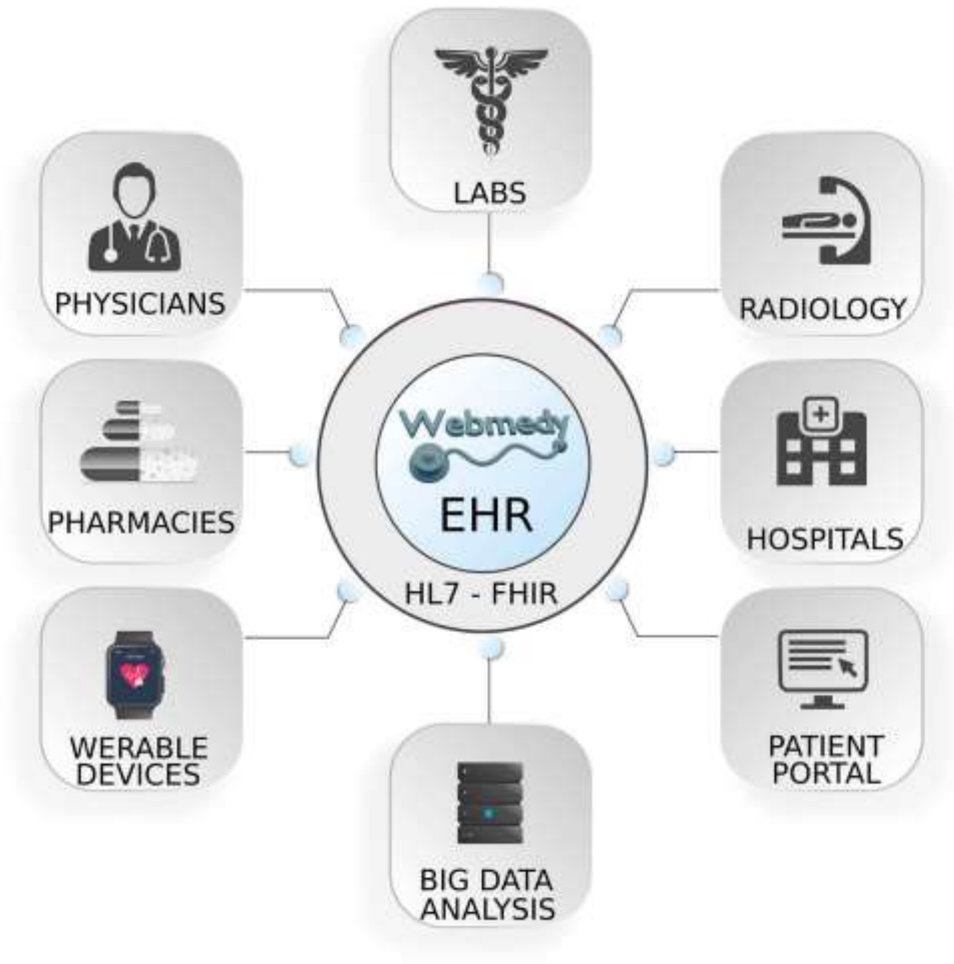
Bioinformatics and Human Electrophysiology Laboratory

Supervisor

Assoc. Professor Themis Exarchos

What are Electronic Health Records

1. Digital records that capture patient care across visits, departments, and often in different clinical facilities
2. Heterogeneous data like diagnoses, medication history, lab results, procedures, and clinician notes.
3. They often include multimodal data (imaging, ultrasound etc.)



Problems of analyzing EHRs

1. Real-world hospital scale health databases store data across many separate tables which makes relationships and trajectory reconstruction difficult
2. Records are often collected under no specific protocol which makes systematic analysis challenging
3. Irregular visit timing creates fragmented patient timelines which makes temporal modeling difficult at population scale
4. Clinical relationships are often hidden due to sparsity, heterogeneity, and the tabular structure of EHR data

Clinical need of EHR synthesis

1. Databases like JH’s PMC hosts patient data scattered across different tables
2. Patient data are sparse often difficult to construct a trajectory
3. Fetching them can be time-consuming and difficult for non-technical users
4. Stored data do not provide context regarding patient’s profile

Large Language Models on EHR Reasoning

Digital Twin- GPT (Makarov et al 2025)

- An LLM based digital twin model for clinical trajectory forecasting and reasoning from longitudinal EHR data that operates directly on raw records and is robust to missingness and noise
- Outperforms state of the art machine learning models across cancer, ICU, and Alzheimer's datasets

EHR Agent (Shi et al 2024)

- An LLM based agent for complex clinical information retrieval from multi table EHR databases using natural language, which plans sequences of actions, generates and executes database queries, and refines its code through execution feedback
- Achieves up to 29.6% higher task success rate compared to the strongest baseline on real world EHR datasets

1. Makarov, N., Bordukova, M., Quengdaeng, P., Garger, D., Rodriguez-Esteban, R., Schmich, F., & Menden, M. P. (2025). Large language models forecast patient health trajectories enabling digital twins. *npj Digital Medicine*, *8*(1), 588.
2. Shi, W., Xu, R., Zhuang, Y., Yu, Y., Zhang, J., Wu, H., ... & Wang, M. D. (2024, November). Ehragent: Code empowers large language models for few-shot complex tabular reasoning on electronic health records. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 22315-22339).

Limitations on existing approaches

1. Most industry software is built for general data management and is not optimized for disease-specific datasets such as Alzheimer's research data.
2. Existing AI tools rely on generic reasoning frameworks that may not fully capture the structure and relationships in longitudinal clinical data.
3. Commercial AI systems often rely on predefined query templates and limit full user control over complex analytical queries.
4. Many industry solutions prioritize operational workflows rather than the flexibility required for experimental research.

Beyond state of the art

We propose an LLM-based multi agentic pipeline that leverages graph-powered reasoning to reconstruct holistic patient profiles from real-world Alzheimer's data sourced from the EHR database.

1. Agentic LLMs have shown promising capabilities in EHR reasoning, yet their application on real-world EHR Alzheimer’s data remains unexplored
2. Existing SOTA methods lack a unified framework for temporal trajectory reconstruction under sparse, irregular real-world visit data.

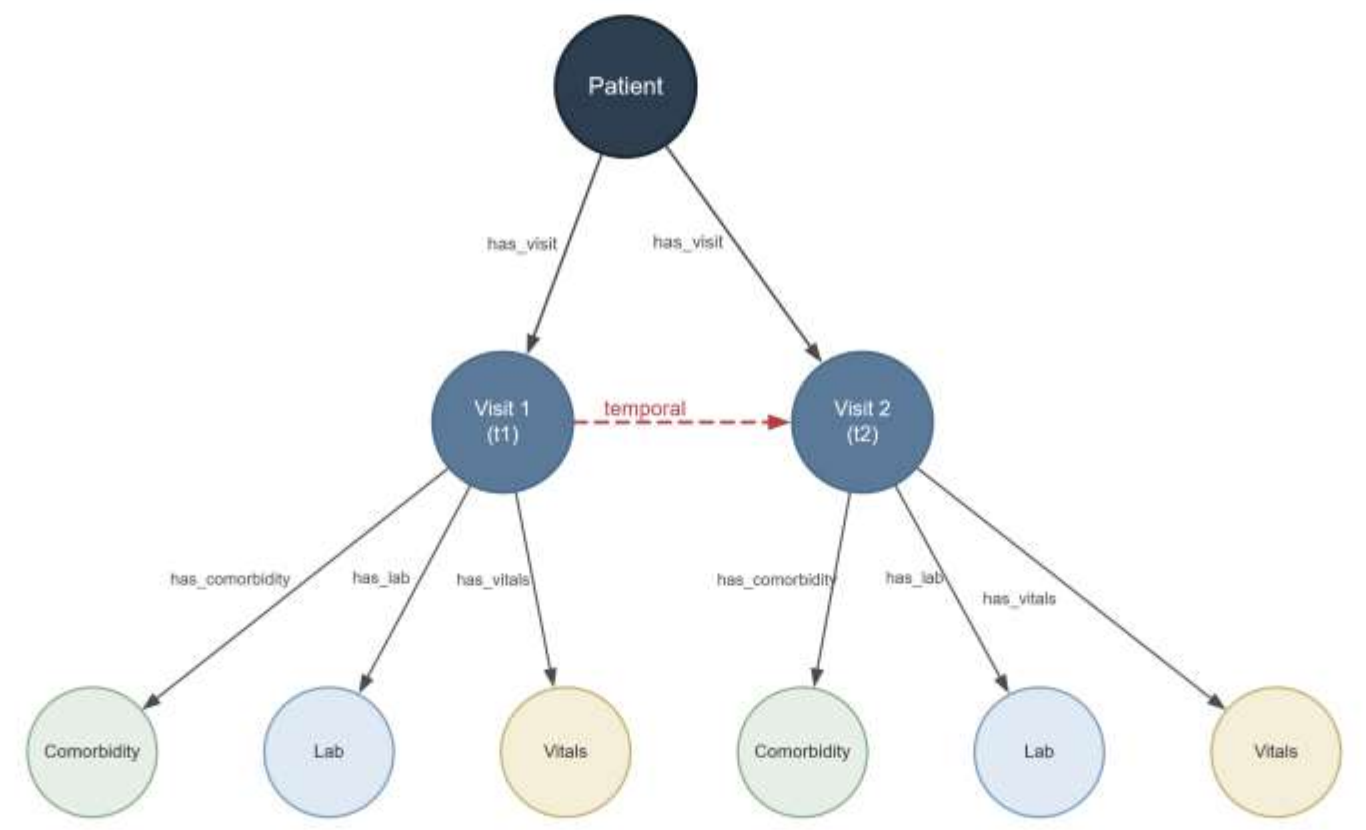
Beyond state of the art

3. Utilize Knowledge Graphs for exploratory analysis of patient trajectory and potentially help to uncover hidden connections across disconnected records of Dementia patients (Tan et al 2025)
4. To the best of our knowledge, no prior work has explored the use of LLMs for real-world EHR-based patient trajectory construction and summarization in Alzheimer's disease and related dementias.

Why explore Knowledge Graphs?

1. Structured Reasoning
   - Knowledge graphs represent clinical entities and relationships explicitly, enabling structured reasoning over relational patient data.
2. Temporal Coherence
   - Temporal edges encode chronological relationships and preserve longitudinal patient trajectories across sparse and irregular visits.
3. Hallucination Mitigation
   - Constraining reasoning to verified graph relations grounds outputs in patient data and reduces hallucination risk.

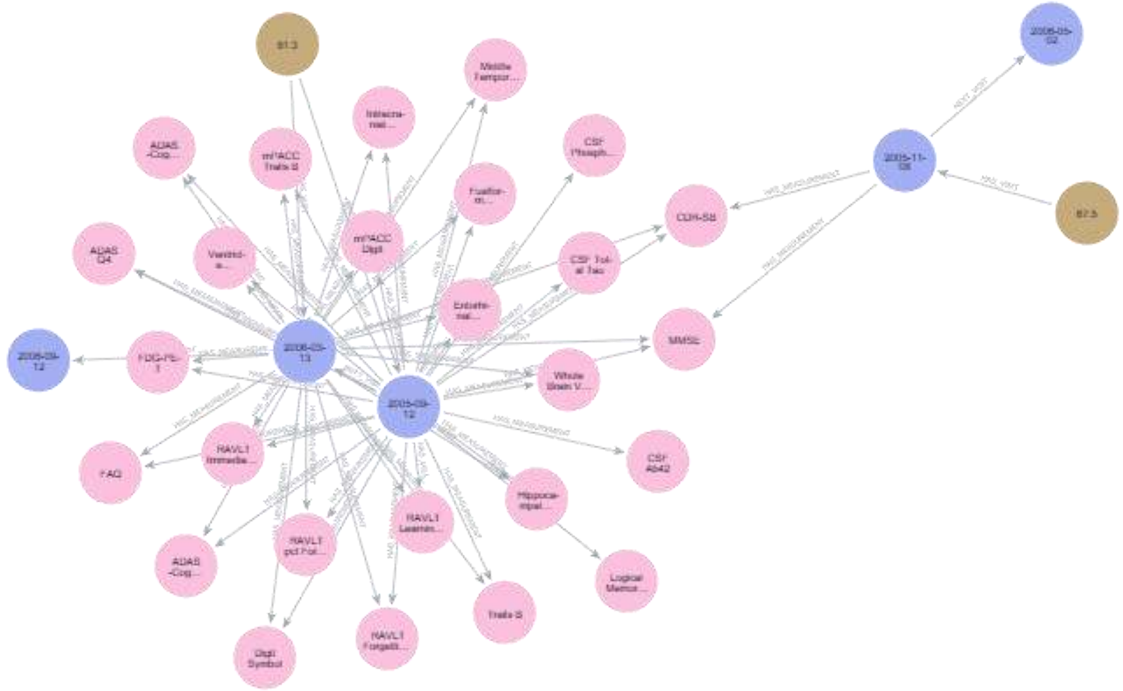
Why explore Knowledge Graphs?



The case of ADNI

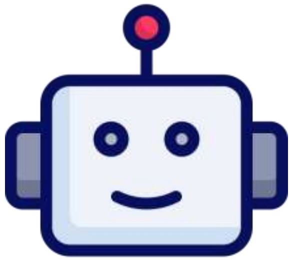
Example using Alzheimer's Neuroimaging Initiative Data

1. Graph Nodes
    1. Patients
    2. Visits
    3. Measurements
2. Edges
    1. HAS_MEASUREMENT
    2. HAS_VISIT
    3. TRANSITION
    4. NEXT_VISIT
    5. DIAGNOSED_AS
    6. BASELINE_DIAGNOSIS
3. Nodes are connected with edges based on the measurement availability
4. Each edge encodes as weight the table feature corresponding value

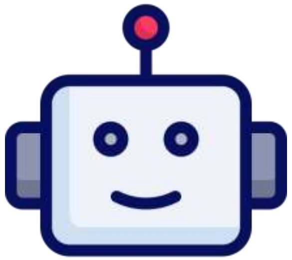


System Architecture

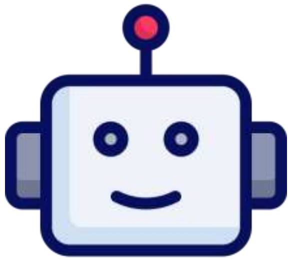
1. The system consist of 4 AI Agents powered by the Claude family models
2. We assigned a specific role on each agent within the pipeline to simulate the fetch and experimentation process performed by humans
3. We utilized Neo4J Graph database to host the graph generated by the conventional ADNI files



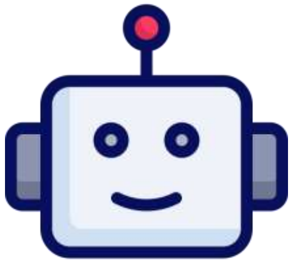
Orchestrator Agent



Database Administrator Agent



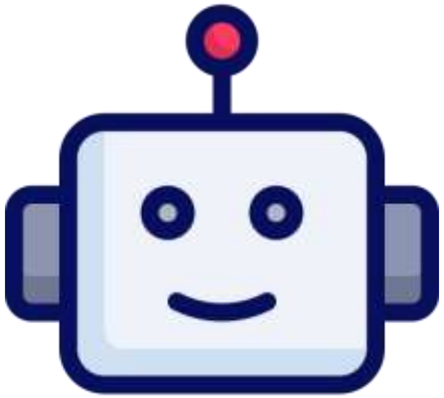
Statistician Agent



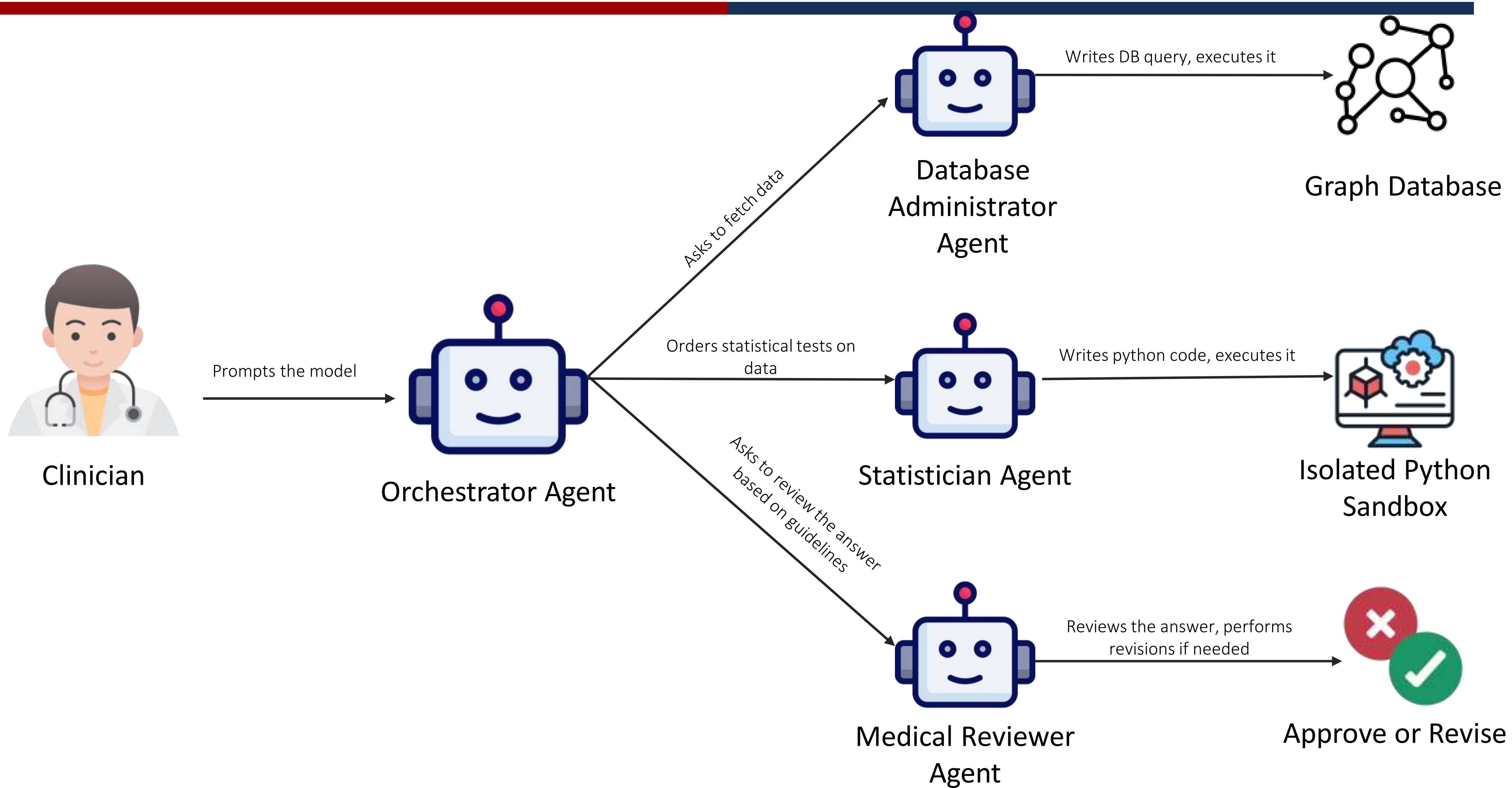
Medical Reviewer Agent

System Architecture

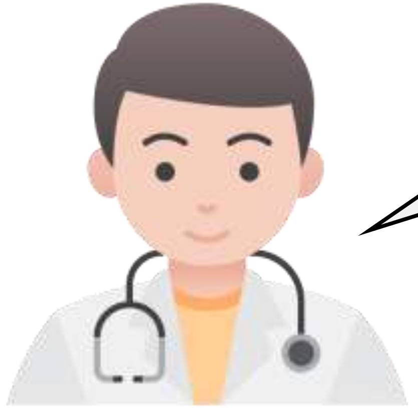
1. Orchestrator Agent
   - Generates a workflow plan based on user's specific request and returns the result once generated
2. Database Administrator agent
   - Receives instructions from the orchestrator agents and queries the graph database
3. Statistician Agent
   - Writes and executes python code on isolated sandbox if statistical tests are needed to verify observed claims
4. Medical Reviewer agent
   - Instructed to convert the output into pure informative narrative and redact any diagnostic suggestions or assumptions generated by the orchestrator agent



System Architecture



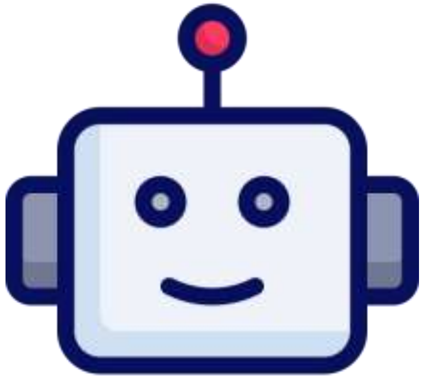
Proof of concept on ADNI dataset



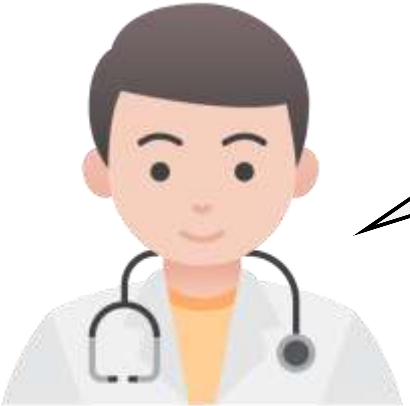
Show me patient's 005_S_0222 trajectory summary

Clinician

AI Agent



Proof of concept on ADNI dataset



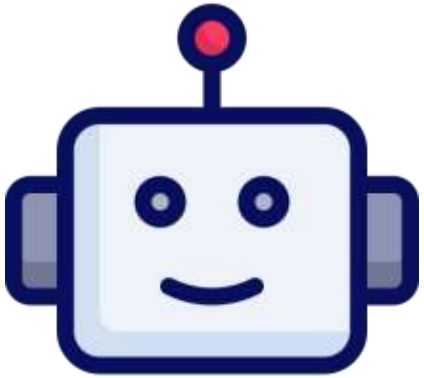
Clinician

Show me patient's 005_S_0222 trajectory summary

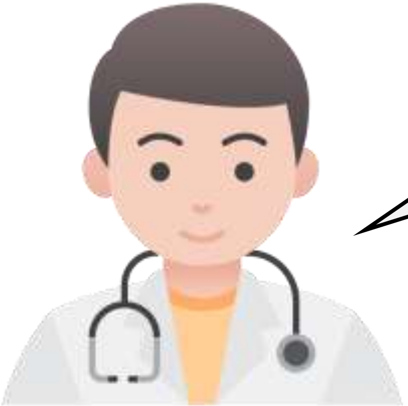
This 85.9-year-old male participant in ADNI1 showed the following longitudinal patterns over approximately 2.5 years of follow-up:

- Research diagnostic classification changed from MCI to Dementia at the 12-month timepoint
- CDR Sum of Boxes scores increased from 3.0 to 7.0
- Logical Memory Delayed recall scores declined from 5 to 1
- Global cognitive measures (MMSE, ADAS-Cog) remained relatively stable
- CSF Aβ42 levels decreased by approximately 12%
- FDG-PET values declined by approximately 6%
- Multiple brain regions showed volume decreases with corresponding ventricular expansion
- Participant had zero APOE ε4 alleles

AI Agent



Proof of concept on ADNI dataset



Clinician

Show me more details

CSF Biomarker Data

Amyloid Beta 42: Decreased from 620.3 pg/mL at baseline to 535.4 pg/mL at 12 months and 548.3 pg/mL at 24 months (approximately 12% decrease from baseline)

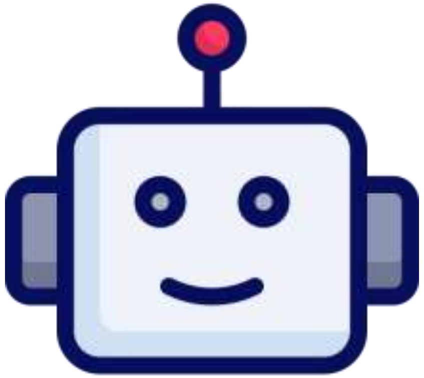
Phospho-Tau: Remained stable across timepoints (31.9→32.0→32.6 pg/mL)

Total Tau: Showed slight increase from 319.7 pg/mL at baseline to 343.2 pg/mL at 24 months (approximately 7% increase)

Neuroimaging and Volumetric Data

Metabolic Imaging:
- FDG-PET values declined from 1.108 at baseline to 1.044 at 24 months (approximately 6% decrease) [.....].

AI Agent



Application in Real World Data

1. We are currently working on applying the proposed methodology on John's Hopkins PMC database
2. Johns Hopkins PMC can be represented as a Knowledge Graph due to their unique real-world characteristics.
3. PMC database contains patient records spanning more than 12 years of observations collected under diverse clinical conditions.

JOHNS HOPKINS
MEDICINE
SCHOOL OF MEDICINE

BiH.E. Lab

About PMC Database

1. 700.000+ unique patients across all Johns Hopkins Medical facilities
2. More than 98M visit records since 2014
3. 153.902 unique patients in the dementia cohort
   - 139,477 marked as Control
   - 14.425 marked as Dementia at any given timepoint
4. More than 4M (timepoints) since 2014 in the dementia cohort
5. More than 90% of timepoints contain missing values
6. Irregular time interval between each visits
7. Laboratory measurements were conducted across multiple facilities using disparate equipment and varying analytical thresholds

JOHNS HOPKINS
MEDICINE
SCHOOL OF MEDICINE

Preliminary Results

Raw EHR data

```
================================================================================
=============
Patient ID: g7uz8gt0ih

Report Period: March 2016
Observation Dates: 2016-03-23, 2016-03-24
================================================================================
=============

## comorbidities
- HYPERTENSIVE DISEASES: 1
- METABOLIC DISORDERS: 1
- MOOD (AFFECTIVE) DISORDERS: 1

## epic
- BLOOD PRESSURE: 99.33
- PULSE: 96.00
- PULSE OXIMETRY: 96.00
- R BMI: 34.50
- RESPIRATIONS: 14.00
- TEMPERATURE: 97.80
- WEIGHT/SCALE: 3417.60

## labs
- ALBUMIN: 0.39
- ALKPHOS: 0.57
- ALT: 0.71
- ANIONGAP: 0.67
- AST: 0.55
- BILITOT: 0.42
- BUN: 0.73
- CALCIUM: 0.62
- CHOL: 0.95
- CL: 0.92
- CO2: 0.20
- CREATININE: 0.71
- GLU: 0.82
- K: 0.69
- LDLCALC: 0.67
- NA: 0.69
- PROT: 0.14
- TRIG: 0.49
```

```
Patient Medical Report
Start date of report: 2016-03-23
End date of report: 2016-03-24

Findings :
- **Comorbidities (as recorded):**
  - Hypertensive diseases: 1 (present)
  - Metabolic disorders: 1 (present)
  - Mood (affective) disorders: 1 (present)
- **Vital signs / anthropometrics (same values on both dates):**
  - Blood pressure (MAP): 99.33
  - Pulse: 96.00
  - Pulse oximetry: 96.00
  - Respirations: 14.00
  - Temperature: 97.80 °F
  - R BMI: 34.50 (obese range)
  - Weight/scale: 3417.60 oz
- **Laboratory results (single timepoint, all within the normalized healthy range 0–1 inclusive):**
  - Albumin: 0.39 – within healthy range
  - Alkaline phosphatase: 0.57 – within healthy range
  - ALT: 0.71 – within healthy range
  - Anion gap: 0.67 – within healthy range
  - AST: 0.55 – within healthy range
  - Total bilirubin: 0.42 – within healthy range
  - BUN: 0.73 – within healthy range
  - Calcium: 0.62 – within healthy range
  - Cholesterol: 0.95 – within healthy range
  - Chloride: 0.92 – within healthy range
  - CO₂: 0.20 – within healthy range
  - Creatinine: 0.71 – within healthy range
  - Glucose: 0.82 – within healthy range
  - Potassium: 0.69 – within healthy range
  - LDL-calc: 0.67 – within healthy range
  - Sodium: 0.69 – within healthy range
  - Protein: 0.14 – within healthy range
  - Triglycerides: 0.49 – within healthy range

Impression :
All laboratory parameters fall within the predefined healthy interval (0–1 inclusive), suggesting no overt biochemical abnormalities at the time of measurement. The patient's BMI of 34.50 places them in the obese category, which aligns with the presence of metabolic disorder comorbidity. The recorded mean arterial pressure of 99.33 mmHg is at the upper edge of typical MAP values, and together with a pulse of 96 bpm, may reflect the documented hypertensive disease, though the MAP itself is not overtly elevated. Pulse oximetry at 96 % is marginally lower than the usual ≥98 % threshold for healthy adults, indicating a mildly reduced oxygen saturation. Temperature of 97.8 °F is slightly below the conventional lower limit of normal (≈98.6 °F), but without trend data this single value cannot be interpreted definitively. Respiratory rate is within normal limits. The presence of mood (affective) disorder is noted but cannot be further characterized from the available data. No longitudinal trends are observable because each variable was captured only once (or identically) across the two consecutive dates; therefore, stability cannot be confirmed nor change inferred.

Overall Assessment:
At the end of the reporting window (2016-03-24), the patient exhibits a generally normal laboratory profile, an obese BMI, and marginally elevated MAP with a high-normal pulse, consistent with known hypertensive and metabolic comorbidities. Mildly low pulse-oximetry and temperature are noted but lack trend context. Mood disorder is present but not further detailed. No acute laboratory or vital sign abnormalities are evident from the available data.
```

AI agent powed synthesized profile

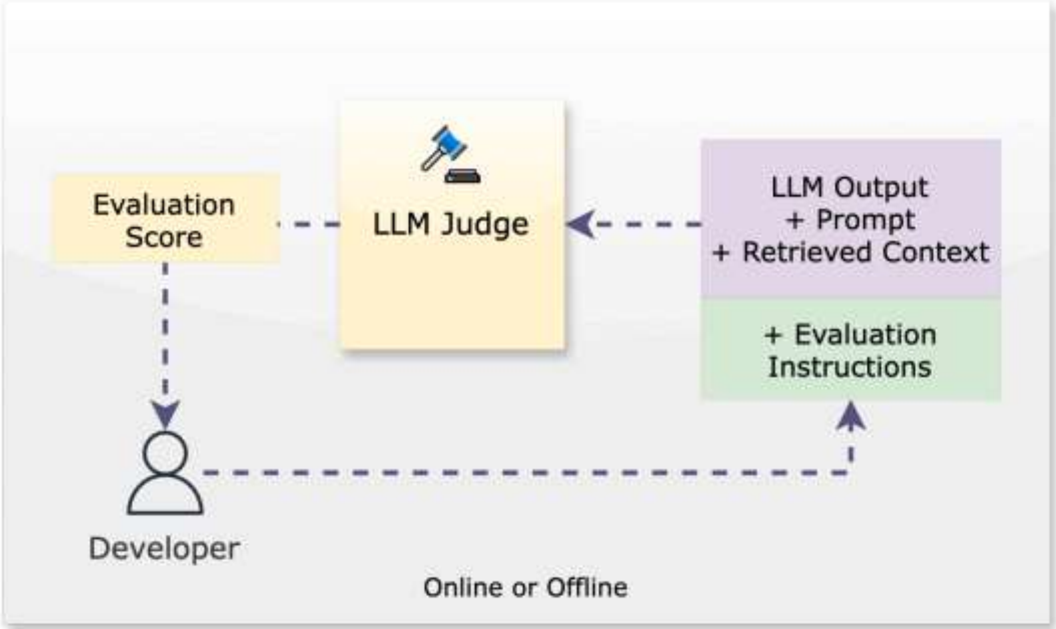
Validation

We are actively working towards designing a Qualitative and Quantitive Validation pipelines based on recent SOTA validation methods for agentic systems

- **Qualitative Assessment**
  - Human expert review of generated clinical narratives
  - Sample-based evaluation of medical accuracy, coherence, and interpretability of the generated narrative.
  - Inter-observer variability evaluation using a standardized protocol
    - Ask a group of experts to evaluate a set of generated narratives using a given rating scale and measure the degree of disagreement to assess their clinical usefulness.

Validation

- **Quantitative Assessment**
  - LLM as a Judge evaluation (Zheng et al 2023 & Gu et.al 2024)
    - A "smarter" LLM is instructed to evaluate the output of another LLM based on given instructions and returns the evaluation score



1. Zheng, L., Chiang, W. L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., ... & Stoica, I. (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, *36*, 46595-46623.
2. Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., ... & Guo, J. (2024). A survey on llm-as-a-judge. *The Innovation*.

Thank you