Pattern Discovery in Clinical Workflows for Anemia Diagnosis

Project in MAS, realized by Elisa Castagnari.

Background:

The Project is an extension for the ethics project dealing: Explainability is a critical requirement in **clinical decision-making**, where the **reasoning** must be **clear and transparent.**

Usually diagnostic guidelines specify the **sequential steps needed** to reach an accurate diagnosis.

The project will be further improved in collaboration with Inria center of research, that will provide access to chatgpt-4.

Brief summary of previous work:

- Prompt Engineering of two open sources LLMs (Mistral7B v0.3 and LLaMA3)[8][12].
- Using of advanced techniques: setting of Persona, 1-shot, Decision Tree, Chain Of Thougth, Plain and Sequential prompt strategy [13][14].
- Deep Analysis of the results, thanks to the explainability obtained by the use of LLMs [2].
- Example of results obtained compared to the one obtained with DRL on the same dataset made on paper [9]:

LLM	Accuracy	F1-Score	ROC-AUC	Time
LLaMA	72.5	72.56	84.29	65m 47.6s
Mistral	29.29	26.27	59.96	406m 47.0s

Plain

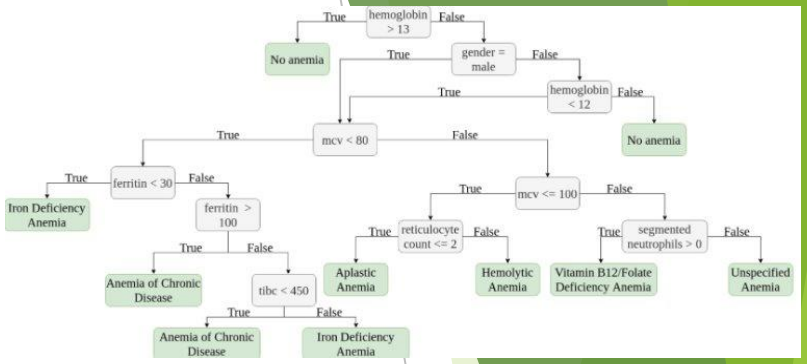
LLM	Accuracy	F1-Score	ROC-AUC	Avg. Length	Time
LLaMA	61.4	62.98	78.62	6.10	262m 1.0s
Mistral	24.9	20.88	56.89	3.13	106m 9.7s
DQN	97.5	97.50	98.60	4.82	1.6s

Sequential

MAS in the previous project:

- **LLMs as Agents:** The LLMs act as "agents," working in tandem with a physician function that retrieves patient-specific data from a database.
- **Physician Function:** This physician function acts as a placeholder for real physician input but lacks decision-making capabilities and is not an autonomous agent.
- **Collaboration of Expert Agents:** (like doctors with different specialties or models trained for specific diagnoses) could be explored but not for anemia.
- **Diagnostic Process:** The conversation between LLMs and the physician function aims to derive a diagnosis based on patient-specific data retrieved from the database.
- **Potential for Human Interaction:** Real physicians could be incorporated to interact with the LLMs and collaborate.
- **Independent LLMs:** The two LLMs (LLaMA and Mistral) operate independently, facilitating a comparative analysis of their diagnostic outputs.

Limitations in the previous work



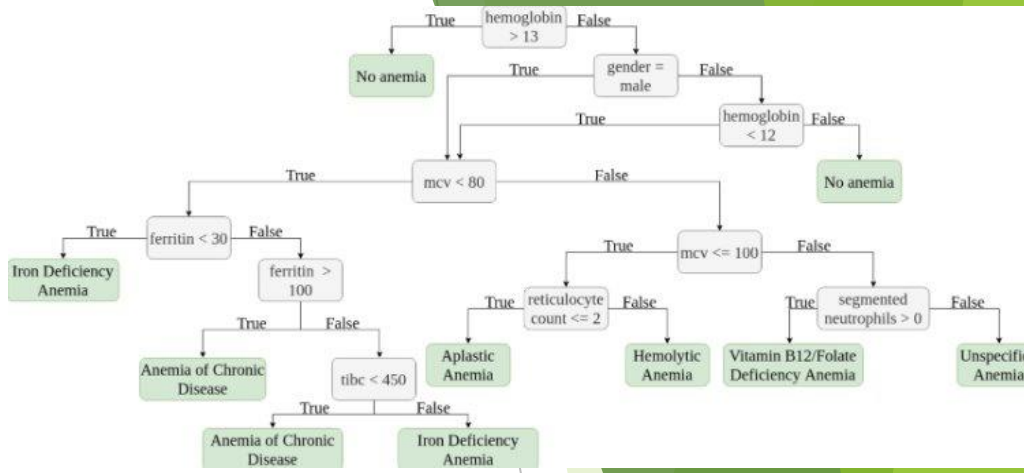
- Deviations from the prescribed decision tree rules do not necessarily indicate errors. These deviations could represent valid alternative diagnostic pathways. Involving a **physician to assess whether these deviations are correct** and if the LLMs can identify **more efficient or accurate diagnostic pathways** would be an important next step.

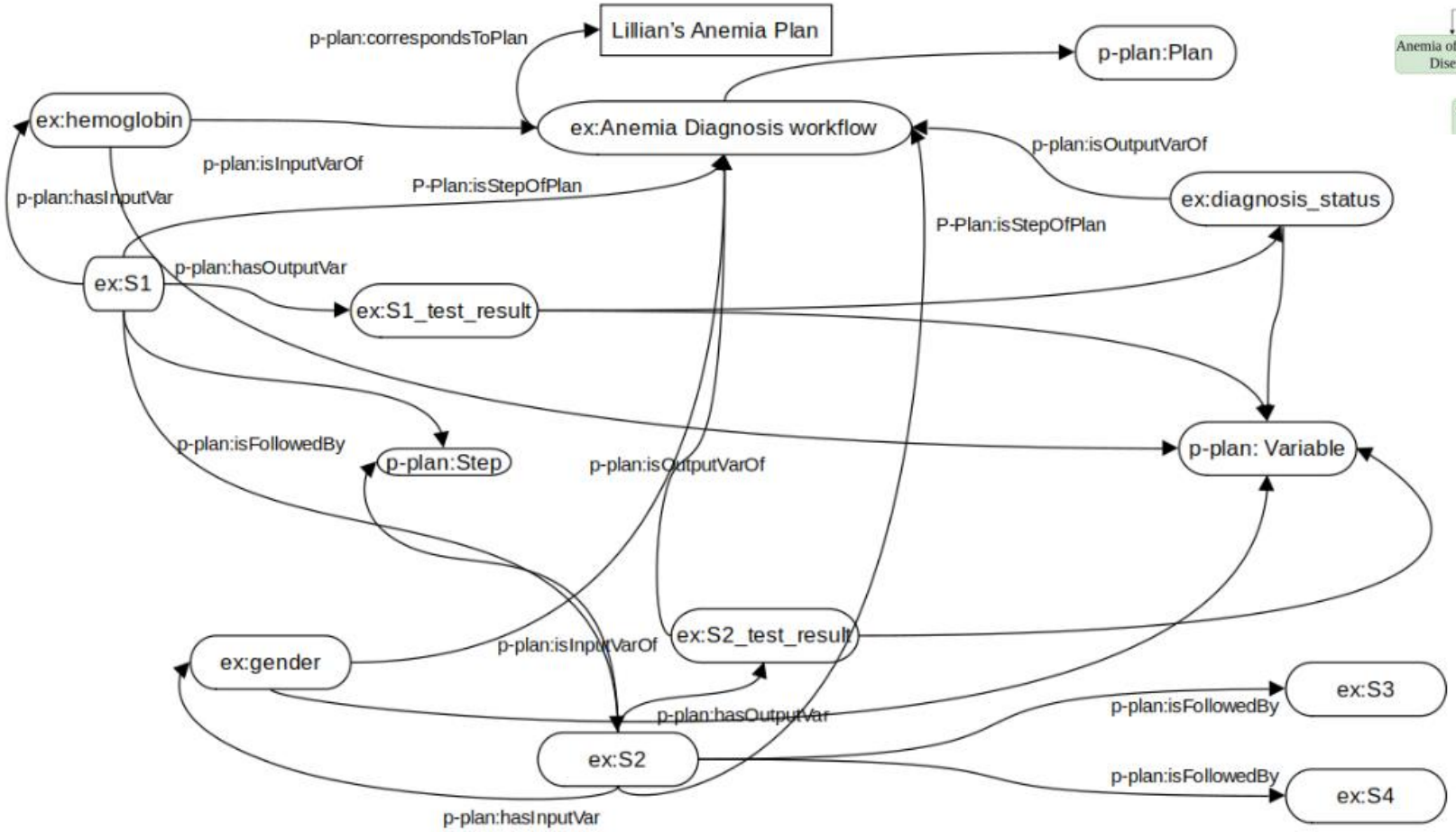
To simplify the work of the physician we need to identify **interesting patterns or the most common pathways** generated by the LLMs for further analysis by physicians. Additionally, using **P-Plan**, it would be valuable to measure the costs and time associated with specific diagnostic pathways, helping to determine the most efficient and cost-effective options. However, in searching for such data, I found that anemia diagnosis often relies on standard blood tests, making it less suitable for proving this point. Lupus, for example, could serve as a better case study.

P-plan

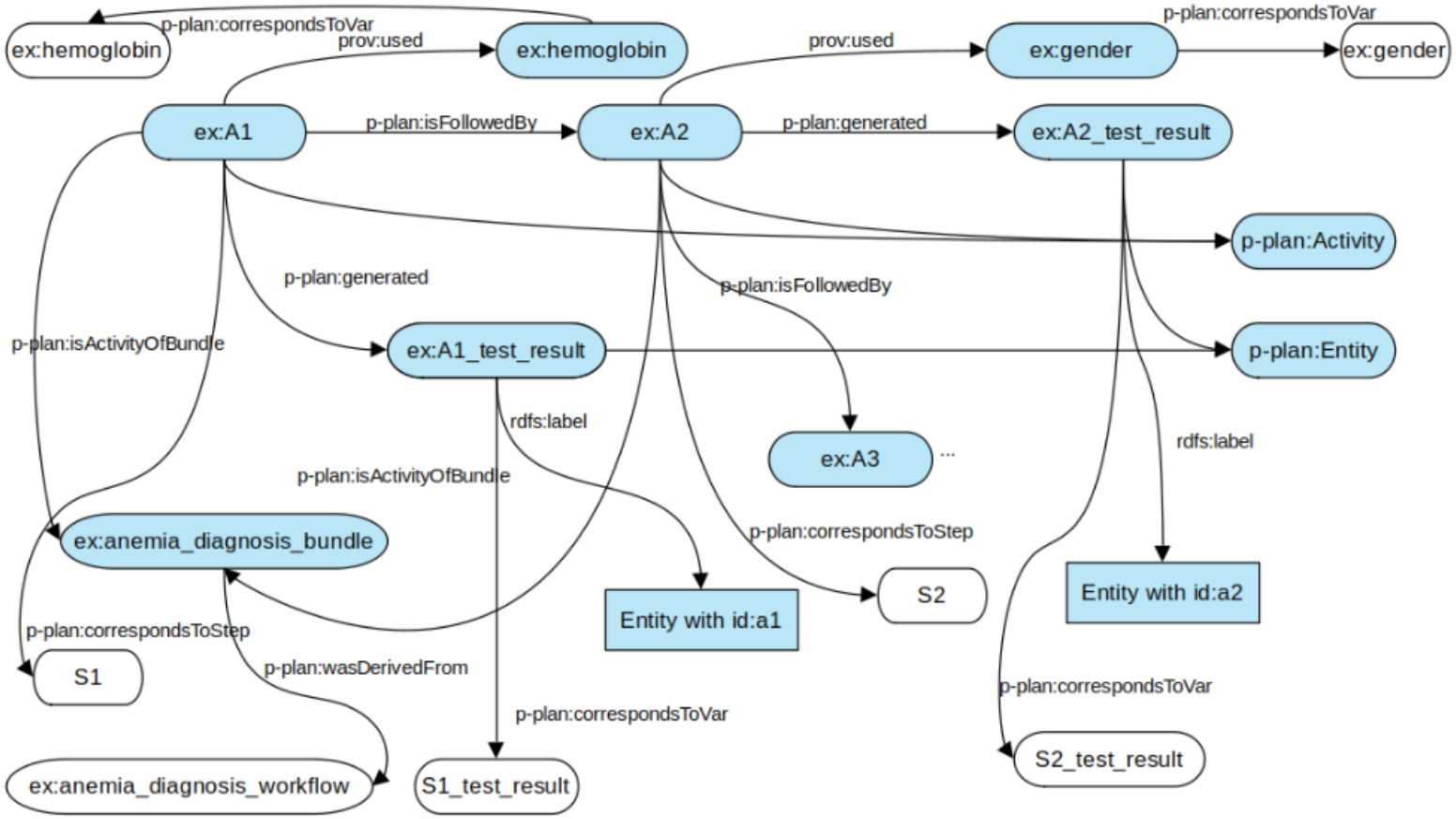
- P-plan [10] is built on another system called PROV-O [7], an ontology using the OWL2 Web Ontology Language (OWL2). PROV-O provides a set of classes, properties, and restrictions that can be used to represent and interchange provenance information generated in different systems and contexts.
- The objective was to have a detailed record (provenance) of the pathway, showing all the steps and lab tests used, and explaining the original plan of how the experiment was supposed to proceed. P-Plan fills the gap left by PROV-O by defining a method to describe and connect plans to the provenance.
- P-Plan serves as a crucial tool for enhancing the representation and tracking of clinical guidelines, which can be viewed as workflows outlining the processes involved in making medical decisions.

Original Plan DT





Specific Pathway for a Patient



Workflow Comparison Overview

- **Focus:** Structured comparison of diagnostic workflows using dataflow DAG (directed acyclic graph).
- Following the content of the paper [11].
- **Key Components:**
  - Global inputs and outputs (variables).
  - Data processing modules with attributes:
    - Label
    - Input/Output
    - Type of operation (e.g., lab test)
    - Datalinks
- **Assumptions:**
  - Modules are deterministic and functionally equivalent.
  - Two levels of similarity: individual modules and entire workflows.

Four Main Tasks for Workflow Comparison

1. **Module Similarity**: Compare each pair of modules.
2. **Module Mapping**: Use module similarity to establish mappings between modules.
3. **Topological Comparison**: Compare workflows based on established mappings.
4. **Normalization**: Normalize results based on workflow size.

Task 1 – Module Comparison

- Modules organized as:
  - `{activity, corresponding step, used, followed by}`
- **Example**:
  - m1 = {‘activity’: ‘A gender’,’corresponding step’: ’S2’, ‘used’: ‘gender’, ’followed by’: ’A hemoglobin’}
  - m1' = {’activity’: ’A hemoglobin’, ’corresponding step’: ’S1’, ’used’: ’hemoglobin’, ’followed by’: ’A rbc’}

(m1, m2, . . . , mi, . . . , mk) and (m′1, m′2, . . . , m′i, . . . , m′l) are two workflows for the same patient.

- **Comparison Method**: Levenshtein edit distance
  - Smallest number of edits to transform one string into another.
  - Insertion, deletion, and substitution.

Distance Assumptions

These are the basic assumptions made about this distance:

- It is equal to zero if and only if the two strings are equal: $dist(a, b) = 0$ iff $a = b$.
- It is symmetric: $dist(a, b) = dist(b, a)$.
- The value is **at most** the length of the longer string: $dist(a, b) \leq \max(|a|, |b|)$.
- The value is **at least** the size difference of the strings: $||a| - |b|| \leq dist(a, b)$.
- Triangle inequality: the distance between two strings is no greater than the sum of their distances from another string: $dist(a, b) \leq dist(a, c) + dist(b, c)$.

Also one could assign a weight to each attribute. Some could be checked using exact string matching.

Task 2 Mapping of Modules:

Pairwise Module Comparison: Establish similarity between module pairs.

Mapping Strategy:

- Maximum weight matching (mw).
- Maximum weight non-crossing matching (mwnc) if module order is given.

Task 3:

Represent workflows as directed acyclic graphs (DAGs).

- Gwf = (Vwf, Ewf)

Use additive similarity score for mapped modules.

Formula:

- Non-normalized similarity score (nnsimMS).

$$nnsim_{MS} = \sum sim(m, m') | (m, m') \in mw(V_{wf1}, V_{wf2})$$

Task 4: Normalization

Goal: Maximize global match between workflows.

Normalization Example:

- Jaccard Index for workflow comparison.

- Formula

$$sim_{MS} = \frac{nnsim_{MS}}{|V_{wf1}| + |V_{wf2}| - nnsim_{MS}}$$

$$nnsim_{MS} = |mw(V_{wf1}, V_{wf2})| = |V_{wf1}| = |V_{wf2}|, and\ sim_{MS} = 1.$$

Implementation Overview

Implementation Overview

- **Two Pathways:**
  - Same patient was considered, sequential prompting.

- **Tools Used:**
  - P-Plan.
  - Turtle file to capture workflow.

Mapping and Hungarian Algorithm

- **Initial Approach**: Greedy selection of mapped modules.
- **Improvement**: Hungarian algorithm for optimal mapping.
  - Minimizes total cost in bipartite graph.

Conclusion and Recommendation

- **Findings**: Detailed mappings offer insights into the diagnostic process.
- **Complexity**: Approach too complex for anemia but useful for diseases like Lupus.
- **Suggested Simplification**: Use Sankey diagrams for simpler workflow comparisons based on lab tests and pathway length.

Why a Lupus implementation could be interesting:

- Complex Diagnostic Process: Lupus (systemic lupus erythematosus, SLE) has a **complex** and multifaceted diagnostic process. The symptoms can vary widely among patients, and there is no single test that definitively diagnoses lupus. This complexity allows for the exploration of multiple diagnostic pathways based on various symptoms, laboratory tests, and clinical findings.

- Diverse Diagnostic Tests: Diagnosing lupus typically involves a combination of laboratory tests, including [1]:

  1. Antinuclear antibody (ANA) test
  2. Anti-double-stranded DNA (anti-dsDNA) test
  3. Anti-Smith (anti-Sm) antibody test
  4. Complement levels (C3, C4)
  5. Urinalysis for kidney involvement

These tests can provide opportunities to measure the time taken and costs associated.

Why a Lupus implementation could be interesting:

- Multi-disciplinary Approach: Lupus often requires **collaboration among different specialists** (e.g., rheumatologists, nephrologists, and hematologists) for an accurate diagnosis. This multi- disciplinary aspect that incorporates various expert opinions and knowledge, resembles a multi-agent system approach.
- Variability in Patient Presentation: Patients with lupus can present with a range of symptoms, including fatigue, joint pain, skin rashes, and organ involvement [3]. This variability can lead to different diagnostic pathways based on individual patient profiles, allowing for the study of how different pathways are formed and which ones are **more efficient or cost-effective**.
- Increased Need for Decision Support: Given the complexity of lupus diagnosis and the potential for misdiagnosis or delayed diagnosis, decision support systems can be particularly beneficial.
- Potential for Real-World Data: Lupus is a prevalent condition with a **substantial amount of clinical data available from electronic health records (EHRs)[5].** This data can be leveraged to train models and validate findings, enhancing the generalizability of the results.
- Cost Implications: Understanding the costs associated with diagnosing lupus can be crucial for healthcare systems. This condition often **requires long-term management and follow-up,** so identifying the most cost-effective pathways can have significant implications for patient care and healthcare budgeting [4].

Sankey Diagram for Anemia

- Generated Pathways LlaMA:

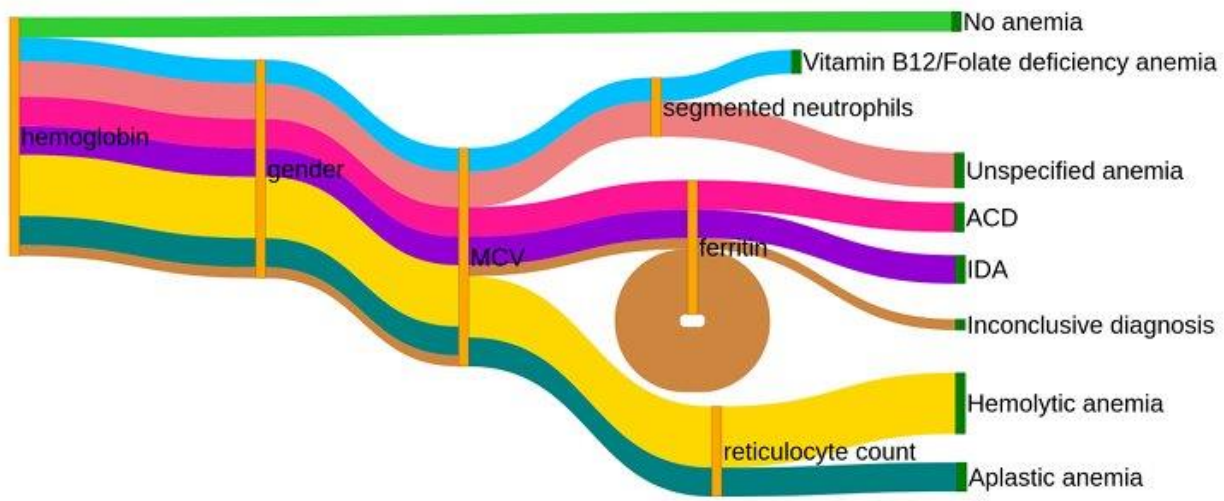
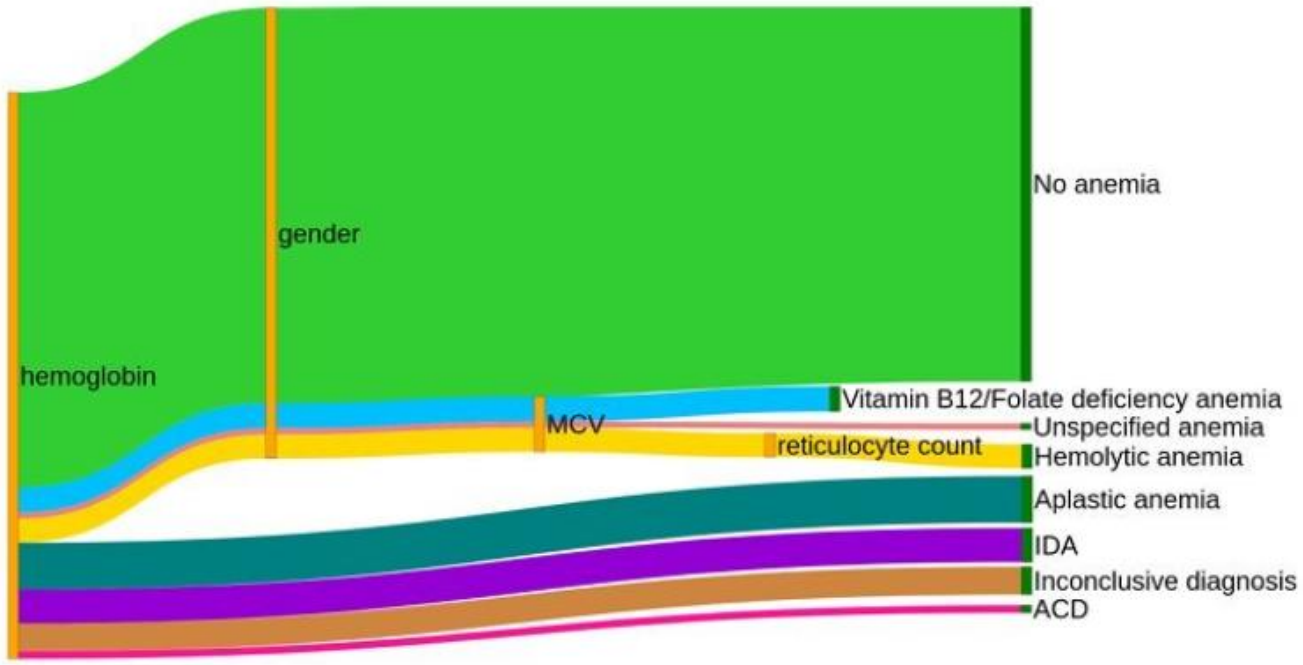


Figure 3: The commonest pathway to each anemia class for *LLaMA*.

Sankey Diagram for Anemia

- Generated Pathways Mistral:



Conclusions and Future Work:

- Structured Representation of Diagnostic Process: The integration of P-Plan facilitates a **structured framework**, allowing for clear organization of diagnostic elements and transparent tracking of decision-making steps, improving clinical assessment quality and reliability.
- Future Work Expansion to Lupus: The study could expand to complex diagnoses like **lupus**, which requires more innovative strategies, as a simple decision tree may not capture its diagnostic complexities.
- Innovative Prompting Strategies: Future work would explore alternatives like retrieval-augmented generation (RAG) and prompt tuning to handle lupus diagnosis efficiently.
- Multidisciplinary Diagnostic Approach: Lupus diagnosis involves multiple specialists, making it a valuable case for studying how large language models (LLMs) can support complex, multidisciplinary diagnostic pathways.
- Costs and Time Analysis for Pathways: A critical objective will be to analyze **costs and time across different diagnostic pathways**, particularly for lupus, aiming to identify the most efficient and cost-effective approaches.
- Physician Expertise Integration: Collaboration with physicians will enhance the analysis by validating pathways generated by LLMs and ensuring that they align with real-world clinical guidelines and practices.

References

- [1] M. Aringer, K. Costenbader, D. Daikh, R. Brinks, M. Mosca, R. Ramsey-Goldman, J. S. Smolen, D. Wofsy, D. T. Boumpas, D. L. Kamen, and et al. 2019 european league against rheumatism/american college of rheumatology classification criteria for systemic lupus erythematosus. Arthritis & Rheumatology, 71(9):1400–1412, 2019.
- [2] Julia Adler-Milstein, Jonathan H. Chen, and Gurpreet Dhaliwal. Next-Generation Artificial Intelligence for Diagnosis: From Predicting Diagnostic Labels to "Wayfinding". JAMA, 326(24):2467–2468, 12 2021.
- [3] A. Fayyaz, A. Igoe, B. T. Kurien, D. Danda, J. A. James, H. A. Stafford, and R. H. Scofield. Haematological manifestations of lupus. Lupus Science & Medicine, 2(1):e000078, 2015.
- [4] N. Galanopoulos, A. Christoforidou, and Z. Bezirgiannidou. Lupus thrombocytopenia: pathogenesis and therapeutic implications. Mediterranean Journal of Rheumatology, 28(1):20–26, 2017.
- [5] Lin Lawrence Guo, Jason Fries, Ethan Steinberg, Scott Lanyon Fleming, Keith Morse, Catherine Aftandilian, Jose Posada, Nigam Shah, and Lillian Sung. A multi-center study on the adaptability of a shared foundation model for electronic health records. NPJ Digital Medicine, 7(1):171, 2024.
- [6] RDF Working Group. Rdf, 2014. Publication date: 2014-02-25 (with a previous version published at: 2004-02-10).
- [7] Timothy Lebo, Satya Sahoo, and Deborah McGuinness. Prov-o: The prov ontology. http://www.w3.org/TR/2013/REC-prov-o-20130430/, 2013. W3C Recommendation 30 April 2013.

References

- [8] Mistral AI. Accessed: 2024-08-06.
- [9] Lillian Muyama, Antoine Neuraz, and Adrien Coulet. Deep reinforcement learning for personalized diagnostic decision pathways using electronic health records: A comparative study on anemia and systemic lupus erythematosus. arXiv preprint arXiv:2404.05913,2024.
- [10] The p-plan ontology. http://vocab.linkeddata.es/p-plan/version/13032014/. Release 12 March 2014.
- [11] Starlinger, Sarah Cohen-Boulakia, Sanjeev Khanna, Susan B. Davidson, and Ulf Leser. Effective and efficient similarity search in scientific workflow repositories. Humboldt-Universitat zu Berlin, Institut f¨ur Informatik, 2019. Universite Paris-Sud, Laboratoire de Recherche en Informatique, CNRS UMR 8623, INRIA, LIRMM, France; University of Pennsylvania, Department of Computer and Information Science, 3330 Walnut Street Philadelphia, PA 19104-6389, US.
- [12] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothee Lacroix, Baptiste Roziere, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971, 2023.
- [13] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35:24824–24837, 2022.
- [14] Jamil Zaghir, Marco Naguib, Mina Bjelogrlic, Aurelie Neveol, Xavier Tannier, and Christian Lovis. Prompt engineering paradigms for medical applications: scoping review and recommendations for better practices. arXiv preprint arXiv:2405.01249, 2024.