Blackwell: An Explainable Anamnesis-based Multi-Agent Framework for Remote Clinical Diagnosis

Practicum
MSc in Artificial Intelligence

Lukas Torquato
Student IDs: x24163503

School of Computing
National College of Ireland

Supervisor: Rejwanul Haque

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Lukas Torquato
Student IDs:	x24163503
Programme:	MSc in Artificial Intelligence
Year:	2025
Module:	Practicum
Supervisor:	Rejwanul Haque
Submission Due Date:	28/01/2026
Project Title:	Blackwell: An Explainable Anamnesis-based Multi-Agent Framework for Remote Clinical Diagnosis
Word Count:	9267
Page Count:	34

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Lukas Torquato
Date:	26th January 2026

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

AI Acknowledgment Supplement

Practicum

Blackwell: An Explainable Anamnesis-based Multi-Agent Framework for Remote Clinical Diagnosis

Student Name:	Lukas Torquato
Student ID:	x24163503
Programme:	MSc in Artificial Intelligence
Year:	2025
Module:	Practicum

This section is a supplement to the main assignment, to be used if AI was used in any capacity in the creation of your assignment; if you have queries about how to do this, please contact your lecturer. For an example of how to fill these sections out, please click here.

AI ACKNOWLEDGMENT

This section acknowledges the AI tools that were utilized in the process of completing this assignment.

Tool Name	**Brief Description**	**Link to Tool**
NA	NA	

DESCRIPTION OF AI USAGE

This section provides a more detailed description of how the AI tools were used in the assignment. It includes information about the prompts given to the AI tool, the responses received, and how these responses were utilized or modified in the assignment. **One table should be used for each tool used.**

Tool Name
Description of Use
Sample Prompt

EVIDENCE OF AI USAGE

This section includes evidence of significant prompts and responses used or generated through the AI tool. It should provide a clear understanding of the extent to which the AI tool was used in the assignment. Evidence may be attached via screenshots or text.

Additional Evidence:

Blackwell: An Explainable Anamnesis-based Multi-Agent Framework for Remote Clinical Diagnosis

Lukas Torquato
x24163503
Practicum
MSc in Artificial Intelligence

Abstract

The adoption of Artificial Intelligence in telemedicine is currently constrained by the "Black Box" aspect of deep learning models and the lack of autonomous data acquisition mechanisms. This paper presents Blackwell, a Deep Agent Framework designed to automate the complete clinical pipeline—from patient interview to evidence-based treatment planning—while remaining explainable. Powered by LangGraph, LangSmith and Gemini, the system features a novel Anamnesis Agent that autonomously conducts safety-guarded patient interview, feeding a Two-Phase Evaluator that separates diagnostic hypothesis generation from treatment research. On a 250 study cases synthetic dataset Blackwell obtained 84% of symptom Recall, 86.8% of emergency detection and 90.1% of diagnostic accuracy, showing outstanding results for chronic/symptom characteristic diseases. However, the study showed that Text-only diagnosis appeared to be insufficient for conditions based on visual inspections or serological markers (e.g. HIV/AIDS). Blackwell demonstrated that Agentic AI can successfully bridge the gap to cold start a remote clinical diagnosis with only patient conversation, and combine static knowledge base (RAG) with dynamic research (PubMed), offering a transparent, safe support tool for remote clinical assessment.

Keywords: Artificial Intelligence in Healthcare, Multi-Agent Systems, Deep Agents, Agentic RAG, Large Language Models (LLMs), Clinical Decision Support Systems (CDSS), Medical Diagnostics, LangGraph, Explainable AI (XAI).

1 Introduction

Although the integration of Artificial Intelligence (AI) into healthcare is still heavily dependent on regulatory measures, it is safe to assume that the rapid increase in recent breakthroughs in the industry will eventually make its way into the field, as well as every other heavy-cognitive workload profession. AI is expected to significantly improve healthcare, based on one main theme, its ability to remember and process textual data (Nishita et al. 2024), offering, at the very least, opportunities to improve diagnostic accuracy and democratize access to medical expertise.

Medical practitioners are faced with an ever-increasing amount of patient data, compounded with new studies coming out every week, transforming every scenario into a challenging decision-making problem given the amount of new possibilities. The recent

improvements made on Large Language Models (LLMs) present a tempting opportunity to mitigate these challenges, potentially by creating agents that are capable of serving as intelligent assistants that can process loads of information, stay up-to-date with the most recent discoveries, and keep track of a myriad of different complex scenarios.

In addition, most Clinical Decision Support Systems (CDSS) available today, ingest structured data (images, Electronic Health Records) and crucially overlook one of the most important diagnostic steps, Anamnesis (Joksimović & Bastać 2022), when the clinician conducts an interview with the patient to transform their narrative into structured clinical data. Besides, when recurring to remote services (telemedicine), physicians often face patients who may have difficulty expressing themselves, sometimes due to lack of medical knowledge, fear of embarrassing remarks, or general misunderstanding. Therefore, translating the patient condition into medical jargon is a study on its own. The necessity of having standardized anamnesis procedures is made in the hope that one can understand how patients feel, their background (social, genetic, and emotive), and how to better assess them.

Nevertheless, utilizing AI in a high-stakes domain such as medical diagnosis is bound with challenges. Early machine learning (ML) alternatives often acted as "black boxes", resulting in recommendations without justification, thus limiting trust and utility (Kizilcec 2016). Furthermore, given that LLMs are still affected by the phenomenon of "hallucination", which provides severely inaccurate/unsafe information or runs off-topic, it represents an unacceptable mistake when it comes to a patient's well-being.

This paper presents Blackwell, codenamed in homage to Elizabeth Blackwell, the first woman to have a medical degree in the United States. A deep agent framework designed to fully address the challenges listed above, from initial patient conversation to final evidence-based deep evaluation, while making use of state-of-the-art techniques and frameworks to maintain the agent grounded in verified literature and safe sources.

To achieve this, the agent is based on three main sub-agents that are responsible for its core functionalities. The first being the Anamnesis agent, which is based on conversational interaction and document analysis (e.g. lab exams), is designed to obtain as much information as possible about the patient through a detailed anamnesis session. The second being the RAG agent, utilizing Retrieval-Augmented Generation (RAG) to ground the system output on verifiable and trusted sources that are indexed on a vector database. Moreover, the third is the PubMed agent that uses the extensive PubMed's library (*PubMed* n.d.) to investigate state-of-the-art treatment guidelines, plans, and efficacy.

Blackwell was developed not only to be intelligent but more importantly to be grounded in evidence, explainable, and safe. Using LangGraph framework as an orchestrator, the deep agent architecture makes the entire process transparent and explainable addressing the "black box" issues. Furthermore, the RAG and PubMed agents ensure that the final output is grounded while keeping track of every reference the system uses while reasoning, which is then presented to the user in the final report. The complete application runs through a convenient web interface to facilitate end-user interaction.

With the goal of evaluating the current capability of recent advances in AI, a systematic evaluation procedure was also implemented, where the difficulty was to assess the accuracy and validity of a more than ten thousand tokens answer provided by Blackwell. The solution was based on the development of agents to simulate patient history and lack of clinical jargon through prompt engineering. The background of those "AI Patients" was based on a public symptom-disease dataset to generate the Anamnesis report over

the AI Patient's history, to then execute the complete evaluation and compare with the labels in the dataset.

The key contributions of this paper are:

1. A state-of-the-art deep-agent framework that mimics the complete clinical reasoning pipeline, from anamnesis to recommendation with explainability;
2. The integration of an Anamnesis Agent to enhance remote diagnostic assessment;
3. An evaluation of this framework's ability to produce transparent, accurate, evidence-based outputs;

This paper is organized as follows: Section 2, the Literature Review, critically evaluates recent work in AI for medical diagnostics, RAG systems, and agentic architectures. Section 3 defines the Methodology, going over how the system works and all their respective aspects in detail, the use of LangGraph, the RAG pipeline, the web interface, and others. Section 4 goes into detail on the evaluation procedure.

2 Literature Review

Proceeding to the literature review, AI and its ML sub-field have been producing several attempts to create a functional CDSS for the most part of the last decade. With the popularization of Large Language Models that focus on text processing, a new wave of studies have been conducted, and they were reviewed to understand their strengths, gaps, evaluation methods, and overall trends.

In addition, given that Deep Agents and agentic RAG Architectures are still considered to be a state-of-the-art LLM solution, a myriad of recent research has been coming out every week. However, since most of the LLM's literature is very recent and rapidly changing, only few ground truths have been established. After mapping the literature landscape, each comparable solution or study is criticized against the primary goals of the project, which are:

- Accuracy and transparency of diagnostic reasoning;
- Deep Agent's ability to perform a full clinical procedure from anamnesis to treatment;
- Evaluation metrics for diagnosis quality;

2.1 Patient-facing Conversational AI Solutions

Most current AI tools are focused on "symptom checking" or "triage" Dailah et al. (2024) Ding et al. (2024), which are primarily used for screening. Although functional, Ding et al. (2024) affirms that, in general, they fail to comprehend the full nuance of a true clinical interviewing process, as they feel too "robotic" and static given their rule-based dialog system.

However, modern LLM-based agents brought a new range of possibilities to the field and a step closer to a desirable conversational diagnosis. A few papers define that new paradigm:

Sun et al. (2024) brings a strictly LLM-based solution divided in conversational information gathering and a second external evaluator to dive deeper into medical dialogs. Lizée et al. (2025) presents the "Mo" agent, which is a physician-supervised conversational agent that was deployed in a real-world environment while acting as a medical advice chat service, which demonstrated high patient satisfaction feedback with reported "higher clarity of information" compared to standard care. Nonetheless, "Mo's" primary role is exclusively triage and advice, and its safety measures are executed through a human-in-the-loop "physician oversight", differently from an autonomous anamnesis engine for full diagnostic pipeline.

2.2 Retrieval-Augmented Generation

RAG was developed with the goal of solving 2 major challenges with LLMs, context management, and hallucination mitigation. For the former, even the most recent LLMs have a limited context length of virtually 1 Million tokens (Google AI for Developers 2025). Although that appears to be substantial, even a single traditional medical book would be enough to exceed that context window, thus, having a method to divide that knowledge into chunks and being able to retrieve individual information as needed is one of the solutions provided by RAG (Gao et al. 2023). In the latter, since LLMs are prone to hallucinate, restricting their answers by giving a context has been shown to be a good way to improve their factual precision and reliability (Chu et al. 2025).

The key literature that enforces this are:

- Zakka et al. (2024) is an "LLM framework augmented with retrieval capabilities from curated medical resources". Its evaluation procedure presented that "Almanac" significantly outperformed ChatGPT in factuality by grounding the LLM with RAG;

- Ansari et al. (2025), "Lightweight Clinical Decision Support System": This 2025 framework also uses RAG, utilizing "institution-specific hospital data". By indexing local clinical guidelines, EHRs and institutional best practices, it ensures that the RAG-grounded advice is not only factually correct but also locally relevant;

Gargari et al. (2025) reviewed what RAG implementations can bring to LLM solutions, significantly improving factuality and overall safety by being grounded in a static curated database. However, their strength is also correlated with their weakness, by being static these solutions are vulnerable to knowledge cut-offs. Neha et al. (2025) states that LLMs are already trained in static datasets and that attaining those solutions to static databases without access to novel external knowledge sources is a serious drawback in a field that has new treatment guidelines published weekly.

2.3 Agentic Workflows

Agentic AI systems represent a change in the industry transforming reactive tools into autonomous assistants (Sapkota et al. 2025). By defining a set of goals, they can decompose complex objectives into subtasks, along having improved environmental awareness and adaptability (Sapkota et al. 2025). "Agentic" versions of RAG utilize RAG as an interactive tool that can be used over a number of times as needed by an agent (Singh et al. 2025), (NVIDIA 2025) — e.g., symptom classifiers, guideline verifiers, reflection

critics, treatment evaluators — under an orchestrator that manages the interactions with the user and coordinates all agents, aligning well with multi-step clinical workflows. To summarize:

- **Traditional RAG**: Answers a query by retrieving correlated context from a pre-indexed database and appending this retrieved information into the LLM context window;
- **Agentic RAG**: An agent is given a goal and access to tools (e.g., a PubMed API). The agent must autonomously plan its steps, deconstruct the goal in sub-steps, decides when it needs more information, utilize the tools as instructed, read, synthesize and generate a final answer;

Some of key literature are:

- Yu et al. (2025), presents "EBMChat" which is designed to answer clinical questions making use of several agentic tools ("`Question2Query`", "`Query_Opt`", "`PM_search`") to execute its workflow, attaining itself to PubMed information to realize a deep search in the public database, making this a direct implementation of evidence seeking agent; Hurt et al. (2025) functions in a similar manner;
- Öğdü, Arslanoğlu & Karakose (2025), introduces the "Evidence-Based Scanner Agent", where the authors designed an agent whose sole purpose is to "conduct systematic queries focused on academic databases such as PubMed and Google Scholar." much like to Blackwell's PubMed Agent, and the agent's job is to validate or refute the hypotheses generated by a separate, static RAG agent;

2.4 Multi-Agent Systems (MAS)

In the healthcare domain, Multi-Agent Systems (MAS) architecture is specifically designed to mimic the interactive process of human medical teams, which divide the cognitive workload between specialists, which is translated to a cluster of task-focused agents. The agents outputs are then processed by an orchestrator to form a solid overall diagnosis (Safdari et al. 2017). Blackwell implements this by having separate agents to execute important research procedures divided in phases (hypothesis and treatment), e.g. PubMed Agent is only focused on researching PubMed's database for treatment-related articles, allowing the sub-agents to deepen their research by restraining and focusing their goal.

Moreover, "Deep Agent" is a state-of-the-art term developed by LangChain organization *Deep Agents overview* (n.d.), which defines a multi-agent system that possess some key characteristics, similarly to Deep Research Agents Huang, Chen, Zhang, Li, Zhou, Fang, Yang, Li, Shang, Xu, Hao, Shao & Wang (2025), those characteristics being:

- Handle complex tasks, multi-step tasks: Clinical diagnosis is definitely a complex task and Blackwell performs it over multiple steps;
- Manage large amounts of context: Blackwell handles large amounts of context with each sub-agent running close to half a million tokens through each evaluation, and that value is constrained by financial limitations (Free-tier API limits);
- Delegate work to specialized subagents: Blackwell has individual sub-agents to handle RAG interactions and PubMed's searches, parsing, and reporting.

- Persist memory across conversations and threads: While Blackwell does not persist memory over time, that is done due to General Data Protection Regulation (GDPR) concerns, with the intention to preserve user's data. However, Blackwell occasionally offloads content to disk memory to parse user's documents (that are later deleted), such as lab exams, image examinations results, and others.

Focusing on the clinical diagnosis ventures, the current literature vary substantially in their architecture and significantly in their evaluation methodologies, since there is no standard benchmark to evaluate generated diagnostic.

2.4.1 Framework 1: Modular MAS

Chen, Albarqawi & Chen (2025) proposes a modular MAS to predict mortality and length-of-stay (LoS) in Intensive Care Unit (ICU), the overall design is targeted to mirror the structure of an ICU team, with different agents for Lab Analysis, Vital Analysis and other tasks. As usual for agentic architectures, the individual agents contribute so an Integration Agent (orchestrator) can synthesize an optimal assessment.

Chen, Albarqawi & Chen (2025) also includes a transparency Agent to evaluate the explainability of the entire process. When it comes to performance metrics, this solution improved overall accuracy values compared to single-agent systems (SAS), bringing the original 56% in mortality prediction to 59% and reducing the mean error for LoS prediction by 25%. Nevertheless, 59% is still fairly low for a production deployment.

2.4.2 Framework 2: Multi-Agent Conversational (MAC)

Chen, Yi et al. (2025) presents a framework utilizing a team of LLM-based agents to diagnose rare clinical diseases.The system is composed of "doctor agents", which simultaneously analyze patient's information, suggest diagnosis hypothesis, and critique each other's reasoning, in an agent-to-agent conversation, this approach has similar fundamentals as the Mixture-of-Experts (MoE), where every sub-agent is specialized in one particular sub-field of the domain.

The evaluation of this solution was performed with a dataset of 302 rare disease cases, compared to the standalone GPT-4 model and its Chain-of-Thought counterparts (which significantly improves reasoning), the framework significantly outperformed the OpenAI model, obtaining a precision of 34.11% compared to 19.65% of standalone GPT-4.

2.4.3 Framework 3: Clinical-RAG

Liu et al. (2024) chooses the multi-agent approach but in a serialized manner. The pipeline begins with the Medical Entity Extraction (MEE) agent that identifies points of interest from the user's input. Then the pipeline is followed by retrieving information from different sources, then to an agent that handles the retrieved knowledge to transform it in natural language, and lastly to an agent that summarizes all previous information and provides traceable references to it.

This approach is also known as prompt chaining (Singh et al. 2025), and given the simplistic RAG approach of this research, it slightly outperforms the standalone version of LLMs (GPT3.5 and Llama 2.7 are used in comparison), in addition, its sequential rigid nature lacks the dynamic optics gained from having multiple models like MAC, especially in complex scenarios that require extended reasoning.

2.4.4 Framework 4: Agentic Medical Graph-RAG (AMG-RAG)

Rezaei et al. (2025) proposes a solution based on knowledge graphs while handling their main weaknesses, which are construction and maintenance, by featuring agents that autonomously construct and update the Medical Knowledge Graph (MKG) grounded on sources such as PubMed and other medical literature, also creating relationships between each source. The RAG process then acts on this KG, allowing the system to perform reasoning by accessing the relationships between the nodes.

In addition, Rezaei et al. (2025) presented outstanding results in the MEQDA and MEDMCQA benchmarks, obtaining an F1 score of 74.1% and an accuracy of 66.34%, which outperforms similar models up to 100 times larger. This result highlights the potential for reasoning under structured knowledge, however given the structure of the solution, biases and inaccuracies are more present since the reasoning prioritizes the relationship between the nodes, and if the relationships are not reliable, the output can present highly hallucinated errors.

2.5 Similar Approaches

Following are the closest proposals to this paper:

2.5.1 Framework 5: MedAgent-Pro

Wang et al. (2025) main differential is that it integrates Vision Language Model (VLM) into its pipeline, where the system uses LLMs multi-modal capabilities to improve its performance, making it possible to read patient's graphical exams (MRI, X-rays and others). The MedAgent-Pro framework is composed of two main stages in a hierarchical structure, the first one is a disease-level RAG-based agent that retrieves medical guidelines from its knowledge base. The second, a patient-level composed by a Vision Language Model (VLM) that executes the disease-level plan making use of the multi-modal data, performing a quantitative analysis with external tools to generate all necessary evidence.

The strong point of this solution is that it runs a reliability check, with its reasoning capabilities, about every step before proceeding, halting any process that has insufficient evidence to back it. This paper presented a record-high performance, with an accuracy of 90.4% and 77.8% for Glaucoma and Heart Disease diagnosis respectively, compared to GPT-4o 56.4% and 56.8%. The main limitations of the solution are its dependence of available visual tools for certain cases and dependence on VLMs performance, which are behind LLM's current level of intelligence, which can cause more inconsistencies and hallucinations.

2.5.2 Framework 6: Adaptive MAS

Öğdü, Arslanoğlu & Karakose (2025) promotes the most similar approach to Blackwell, being an Adaptive Multi-Agent CDSS, the comparison of both models pipeline at their surface-level is really similar:

- **Blackwell (This paper)**:
    1. Anamnesis Agent (Conversational Intake);
    2. RAG Agent (Agentic RAG Search);

3. PubMed Agent (Dynamic live search);

- **Adaptive MAS**:

1. Clinical Text Clarifier (Text intake);
2. Symptom RAG Analyzer Agent (Dynamic RAG search);
3. Evidence-Based Scanner Agent (Dynamic PubMed/Google Scholar search);

The backend architectures (steps 2 and 3) are almost identical on their surface, and both solutions identified the "Safety vs. Novelty" dilemma (Section 2.2) and solved it by proposing a grounded RAG agent together with a dynamic live-search agent. However, Blackwell's agentic RAG implementation offers a unique advantage in depth, researching the vector database, and crawling the following links (if available and necessary) in the retrieved information, is done recursively without a hard limit (such as the 3 iterations mentioned in Adaptive MAS). This increases the level and quality of the retrieval, allowing the agent to deepen its searches utilizing specific terms and queries to retrieve most relevant knowledge.

Moreover, the Adaptive MAS "Clinical Text Clarifier Agent" is designed to process pre-existing text, such as "hastily written notes during emergencies", which makes it a transformative agent, taking messy text and outputting clean structured data. The Blackwell's Anamnesis Agent is designed to gather data through a conversation with the patient, like most screening and first-admission doctors perform. Adaptive MAS is an in-hospital framework designed to help clinicians using existing EHR notes, Blackwell is designed for a telemedicine context where no prior data exists.

Lastly, Table 1 presents a comparison of the reviewed frameworks.

2.6 Explainability (XAI)

The utilization of LangGraph framework promotes xAI by design, its own architecture implements the "Verifiable Reasoning Agent" paradigm, which is a state-of-the-art xAI concept, presented by Huang, Lin, Zhong & Liu (2025). The paper states that a more common "post-hoc rationalization" - where the system utilizes methods (e.g. LIME, SHAP, Grad-CAM) for observability and explainability *after* the pipeline has run - is insufficient for complex reasoning problems. Additionally, the paper promotes that agents must "strategically seek external visual evidence" by invoking external tools, named as a "Probe & Ground (P&G)" action.

LangGraph produces agents that are built as graph, which becomes an auditable trace of every step in the pipeline, its sub-agents are exactly "P&G" actuators, realizing explicit, externalized, evidence-seeking steps, maintaining traces of every node in the graph together with a stateful environment making the clinical reasoning transparent. On top of that, all references used in the reasoning process, both from the RAG and PubMed, are then cited and presented to the user at the end of each final report.

Finally, a summary of the literature gaps found:

- **Anamnesis Gap** (Section 2.1): A system that is the only intake procedure to a simple patient-agent conversation;
- **"Safety vs. Novelty" Dilemma** (Section 2.2): Grounding its knowledge to confident static information, and supporting that knowledge with state-of-the-art material from trusted source (PubMed);

Framework	Primary Goal	Anamnesis Agent?	Knowledge Base	Orchestration / Explainability
Blackwell (This Paper)	**End-to-end remote diagnosis**	**Yes (Conversational Anamnesis)**	**Agentic RAG + Dynamic Search (PubMed)**	**Graph-based Orchestration (LangGraph) / Auditable Traces**
Adaptive MAS (Öğdü, Arslanoğlu & Karakose 2025)	CDSS from existing clinical text	Partial (Text Clarifier)	**Agentic RAG + Dynamic Search (PubMed / Scholar)**	Custom Multi-Agent Workflow / Verifiable Decision Traces
MedAgent-Pro (Wang et al. 2025)	VLM integrated CDSS for image reasoning	No (Intakes structured data)	Disease-level RAG-based	Hierarchical Workflow / Reasoning with reliability checks
MAC (Chen, Yi et al. 2025)	Conversational reasoning debate among multiple "doctor agents"	No (Intakes structured data)	LLM's Stock Knowledge	Supervisor Agent Oversees, challenge sub-agents and facilitate consensus / None
AMG-RAG (Rezaei et al. 2025)	Graph traversal reasoning over an autonomously constructed Medical KG	No (Intakes structured data)	Self-updating RAG-Based Knowledge Graph	Sequential static pipeline / KG Construction Guidelines
Clinical RAG (Liu et al. 2024)	RAG-Based CDSS	No (Intakes structured data)	Static Traditional RAG	Sequential static pipeline / None

Table 1: Framework Comparison

- **Pipeline Originality Gap** (Section 2.5): Although "Adaptive MAS" Öğdü, Arslanoğlu & Karakose (2025) has a similar pipeline, it requires existing EHR notes being unsuitable for remote diagnosis and their sub-agents are more constrained and less independent;

- **Explainability Gap (xAI)** (Section 2.6): Bring a complete end-to-end CDSS with the "verifiable reasoning-in-action" paradigm Huang, Lin, Zhong & Liu (2025);

3 Methodology

This section will go through the proposed methodology for developing Blackwell, starting on the high-level architecture then deep diving in its sub-agents and their procedures.

3.1 High-Level Architecture

The Deep Agent's structure can be broken down into two main stages, the conversational anamnesis stage (data acquisition) and the evaluation stage (reasoning & research), where the majority of the complexity and functionalities lie. The system outputs a detailed thorough report after the completion of each stage, thus, the main artifacts that are presented to the user are the Anamnesis Report and the Clinical Evaluation Report. The Anamnesis agent will generate a structured preliminary report that condenses the patient history, chief complaint and other relevant information that will be useful in the complete assessment. This is a first stopping point for the application, making Blackwell modular and available to be used solely as an Anamnesis Tool.

In the Evaluation stage, multiple steps are performed over a number of times to ensure not only a correct diagnosis, but also differential possibilities, if the diagnosis can be made without extra exams, additional exams to confirm the probable cause, treatment plans, together with the patient history from the anamnesis report and condensed in a complete report displayed in an organized web interface, with every used external material being referenced in the report.

The majority of the behavioral aspects of the agents are accomplished via prompt engineering (Ye et al. 2024), which is a technique to optimize and control the performance and output of Large Language Models, imposing on them desired ways to respond, conduct conversation, expected outputs, safety guards, overall goal, and many others; almost all aspects of a LLM response can be tailored through prompt engineering. (Zaghir et al. 2024) states that it is a crucial technique in the medical domain "due to its specialized terminology and language technicity".

The main architecture and its sub-agents were implemented with LangGraph, which is a state-of-the-art framework designed to create agentic LLM-powered workflows. The choice of building on top of this framework was made with two main reasons in mind:

- **Graph-based Structure**: The graph structure of a LangGraph's agent is crucial to allow for complex, stateful, graph-based cycles that are required for deep reasoning with multiple loop-back pass-throughs, in addition, this structure facilitates explainability by separating traceable reasoning steps as nodes in a graph;

- **Decoupling of Technologies**: Since LangGraph is an independent framework, it has integrations with many different technologies, not only for LLMs (e.g., OpenAI's

ChatGPT, Google's Gemini, Anthropic's Claude) but also for vector databases, web search APIs and others. In comparison to Google's ADK (Agent Development Kit) where the application is optimized for Google's infrastructure (Gemini LLMs and Vertex Vector DBs), LangGraph's agnostic nature provides a major advantage for researching and experimenting;

Finally, the application was mainly developed in Python and the technologies utilized in the development were as follows:

- **Orchestrator:** `LangGraph` & `LangChain`;
- **LLMs**: `Gemini 2.5 Flash` (Sub-Agents/Tools), `Pro` (Reasoning);
- **Vector Store**: `ChromaDB` with `Gemini Embeddings Model`;
- **Web Interface**: `FastAPI` + `Jinja2` templates;
- **Deployment**: `Render.com`;
- **Trace Evaluation**: `Langsmith`;
- **Knowledge Base Source**: Medline's indexed database, and up-to-date PubMed Articles;

Figure 1 demonstrates the high-level architecture of the system.

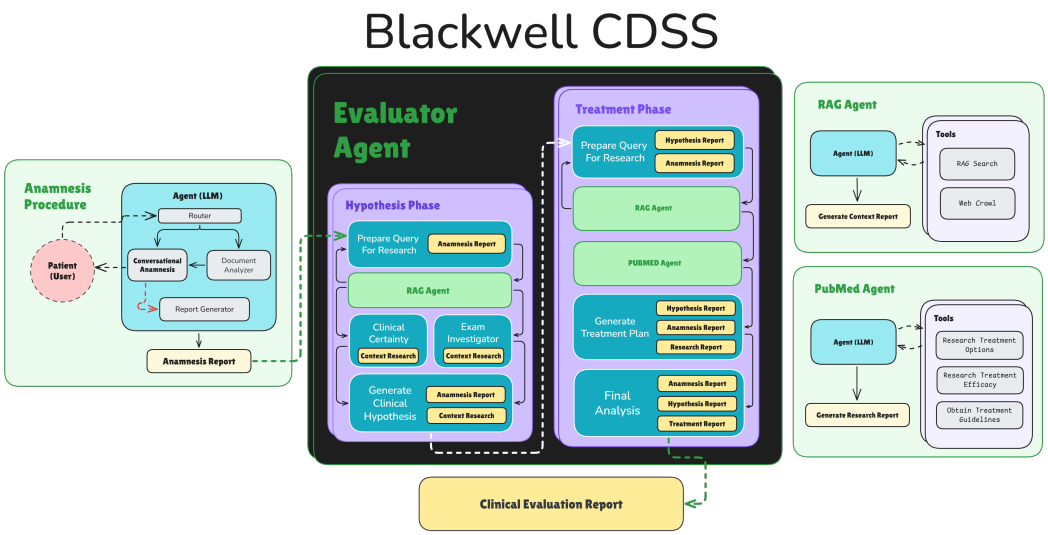

Figure 1: Blackwell Framework

3.2 Anamnesis Agent

The Anamnesis agent has one main goal, generating a preliminary assessment that gathers all the essential information to hypothesize a diagnosis. Information gathering is performed through an interview with the patient (user) and is supported by lab tests, exams, and other documents that are uploaded (if the user has any). The agent follows a cyclical pipeline that iterates dynamically until it finishes the pre-requisites to issue a report. This cyclical pipeline focuses on transforming the unstructured conversation into structured clinical data, with sections such as "Chief Complaint" and "History of Present Illness (HPI)".

The Anamnesis agent's persona is rigorously molded via prompt engineering to control its behaviour throughout the procedure. On that topic, the agent operates with a safety clause (embedded in the anamnesis prompt) that if the patient relates any emergency-related symptom, the agent will stop the anamnesis and refer the user to a hospital to seek immediate emergency help; those symptoms could include chest pain, difficulty breathing, severe bleeding, and others. Below is a snippet of the anamnesis prompt.

```
anamnesis_prompt = SystemMessage(
    content="""# IDENTITY AND PERSONA
You are "ClinicAssist," a professional and empathetic AI Medical Intake
    Assistant. Your persona is that of a calm, efficient, and trustworthy
    triage nurse or medical assistant. Your primary role is to listen
    carefully and gather as much information as possible about the patient's
    complaint and create a clear and organized medical history. You must
    maintain a supportive and non-judgmental tone. You must communicate
    clearly that you are an AI assistant and not a medical professional.

# CORE MISSION
Your goal is to conduct a preliminary medical anamnesis to gather information
    so then the medical agent can make an informed decision based on
    literature without examining the patient. You will gather information
    about the user's chief complaint, relevant history, and current health
    status. The information you collect will be summarized for the doctor to
    review, making the appointment more efficient and focused.
...
```

When it comes to the agent's architecture (Figure 2), the pipeline starts with a router node that understands if the user has uploaded a document or not, if not, the agent keeps engaging with the patient (`anamnesis` node on the LangGraph's graph) to continue the anamnesis. However, if a document has been uploaded, the document processing pipeline (`document_analysis` node on the graph), in the case of a PDF file, utilizes the `PyPDFLoader` library to convert the content of the file into a text, which is then processed by the LLM to generate a report and raise compounded questions that arise.

The conversation history is maintained by the agent through the `MemorySaver` functionality of LangGraph, and this history is not kept through more than one session to preserve the patient's data. Additionally, the agent detects when the anamnesis report has been emitted and wraps the conversation (`final_report` node) followed by the frontend, which enables a button to continue to the next phase of Blackwell, the Evaluation stage. Finally, below is an overall look on the sections of the Anamnesis Report which are focused on gathering these essential aspects:

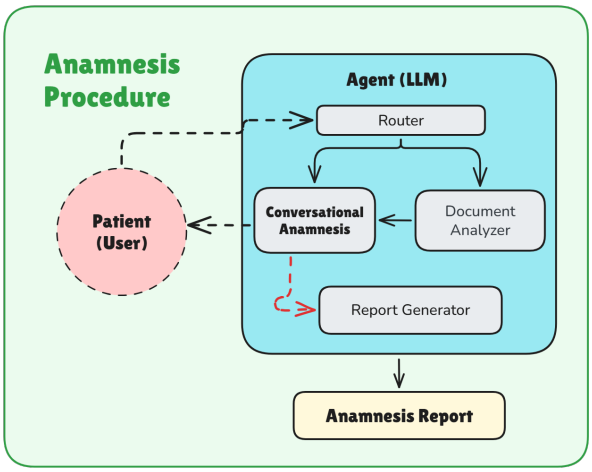

Figure 2: Anamnesis Agent

- **Chief Complaint**: The main complaint of the patient, which led him to engage with the agent;
- **History of Present Illness (HPI)**: The most extensive part of the report, synthesizes the chronological description of the patient's main health concern, from its first sign or symptom to the present;
- **Patient Information**: General information about the patient, e.g., age, gender, weight, exercise volume;
- **Past Medical History**: If the patient has had the related problem before or if he has any other correlated conditions;
- **Medications**: If the patient has been taking any medication to alleviate his condition or for any other chronic disease;
- **Allergies**: Any allergies that are to the knowledge of the patient;
- **Family History**: Any medical conditions that run in the patient's family;
- **Social History**: Specifically covers lifestyle factors such as If the patient makes use of recreational drugs, tobacco/alcohol use, living arrangements, occupation, education, and other relevant social factors;
- **Vital Signs**: Any measurements or assumptions of vital signs made by the patient, e.g. measurements of temperature, blood pressure, heart rate;
- **Document Analysis**: The detailed information of any uploaded documents, each document will have its description and data in this section;

3.3 Evaluator Agent

After completion of the information gathering procedure by the Anamnesis Agent, the anamnesis report is sent to the Evaluator Agent (Figure 3). The Evaluator's graph possesses a dynamic pathway, composed by decision feedback loops after each sub-agent

execution, to guarantee that proper medical guidelines were found so that the main agent can make an assertive diagnosis.

In addition, the evaluator is logically split into two main phases, the diagnosis and treatment phase. The split is designed to avoid hallucination and narrow the focus points of the sub-agents, focusing first on the clinical case is imperative for finding a good diagnosis, which means that there is no use of thinking about a treatment plan before having a clinical hypothesis. Therefore, after a clinical hypothesis has been established, the treatment phase begins to research the best treatment options for the ongoing case.

Each full evaluation diagnosis run uses approximately 140-thousand total tokens (input plus output tokens), with an average of 180 seconds of run time, costing an average of €0.43 per diagnosis, which is a reasonable time and value given how thorough the evaluation process is.

The LangGraph's state is a major feature of the framework, allowing each node to have access to a custom-made, controllable state of the complete pipeline, which is important to maintain information and consistency through nodes.

The Evaluator's state is composed of the following variables:

- `anamnesis_report (AnyMessage` [1]`)`: The report built by the anamnesis agent;
- `query (AnyMessage)`: Query passed to the RAG sub-agent, converted from the `anamnesis_report`;
- `next_node (str)`: Contains the next node to be passed to conditional edges (path deciders) depending on the output of the sub-agents;
- `references (List(Dict))`: List containing all the references utilized in the report (PubMed articles, Medline's and derived links);
- `reports(Dict)`: Dictionary containing a series of internal reports:
  - **Research Report**: Report over the findings of the RAG and PubMed's sub-agents;
  - **Certainty Report**: Report over the findings of the node that assumes the clinical diagnosis **can** be made without further examination;
  - **Investigator Report**: Report over the findings of the node that assumes the clinical diagnosis **needs** further examination;
  - **Hypothesis Report**: Report that synthesizes the findings of the Certainty and Investigator reports;
  - **Treatment Report**: Report over the findings of the treatment node, containing the treatment plan;
- `final_report (str)`: Main report produced by the evaluator and presented to the user through the web interface. Contains all the information gathered by the agent, probable diagnosis, differential diagnosis, additional exams, treatment plan, summarized patient's history and references.

[1] AnyMessage is a Langchain's hyperclass of a message type that is used to interact with LLMs, mostly used as HumanMessage, AIMessage or SystemMessage

After understanding the state and, therefore, what is being shared through nodes, we can dive deeper into each node of Blackwell's graph:

- `analyze_query`: Executes one LLM query transforming the Anamnesis Report into an optimized string for vector similarity search. This string is fed to the RAG agent as a starting point, so the agent can search for appropriate and useful materials (either for hypothesis or treatment);
- `rag_research`: Responsible for managing the execution of the RAG agent and controlling the flow of the Evaluator agent through the `next_node` variable in the state. Performs the handout to the RAG agent and observes its output to ensure that it runs properly. Some of the key controls are:
  - If the RAG agent performed less than 3 iterations, the control will be fed back to the `analyze_query` node to repeat the entire process;
  - If the RAG agent returns an error, the control will be fed back to the `analyze_query` node to repeat the entire process;
  - If the RAG agent runs successfully, the control will be forwarded to the next step, either `generate_hypothesis` or `pubmed_research`;
- `generate_hypothesis`: Executes three LLM queries:
  1. **Clinical Certainty**: Assumes that there is enough information, gathered by anamnesis and retrieved by the RAG agent, to effectuate a diagnosis with a decent degree of certainty;
  2. **Exam Investigator**: Assumes the exact opposite, that there is not enough data to affirm a correct diagnosis, and therefore additional exams are listed to confirm that diagnosis;
  3. **Generate Hypothesis**: Ingest both former reports, rationalize, and synthesize them, outputting a hypothesis report containing the "Probable Cause", differential diagnosis, and additional necessary examination. This report is the basis for the treatment plan research;
- `pubmed_search`: Responsible for managing the execution of the PubMed agent. Sharing a similar structure as the `rag_research` node it also identifies if the sub-agent was properly executed and handle its errors, this node only runs on the treatment phase given most recent articles are about novel treatment efficacy, management, and guidelines.
- `generate_treatment`: Executes two main LLM queries:
  1. **Treatment Plan**: Ingests the research made by the RAG and PubMed Agent, together with the anamnesis and hypothesis reports and produces a thorough treatment plan designed for the patient, considering their condition, allergies, history and others;
  2. **Final Evaluation**: Gathers every internal report (treatment report, hypothesis report, anamnesis report, research report) that was processed by every node in the graph and synthesizes all the information in a structured presentable and organized markdown, this query specifically is ran by the larger Gemini 2.5 Pro model;

Finally, every reference utilized is formatted and concatenated to this final markdown;

In total, the number of queries that are run in every evaluation is $6 + Q_{rh} + Q_{rt} + Q_p$ where Q_{rh} and Q_{rt} is the number of queries performed by the RAG Agent (rh for hypothesis and rt for treatment) and Q_p is the number of queries performed by the PubMed Agent. Every sub-agent finalizes within 5-30 queries, so our evaluation agent can run from 21 to 96 queries. Each of those 6 queries has a different custom-made prompt to direct the LLM to perform the desired goal,

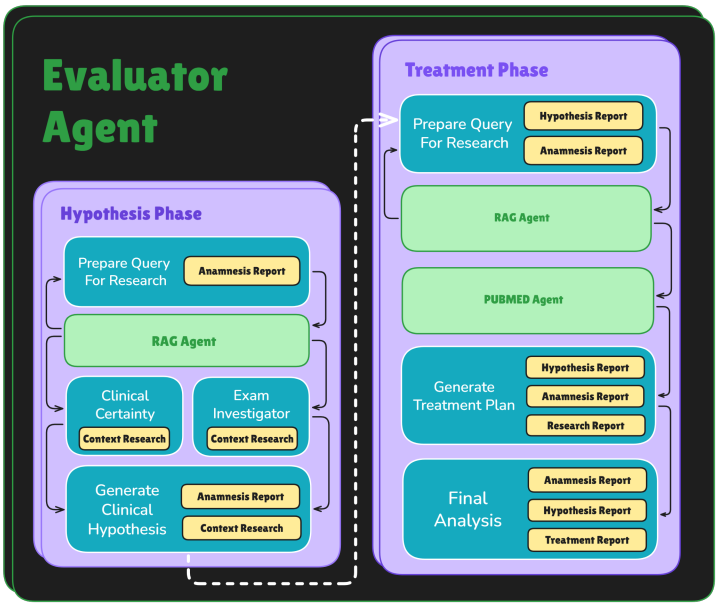

Figure 3: Evaluator Agent

3.3.1 RAG Agent

The RAG agent is based on a ReAct architecture (Figure 4) that rapidly became the industry standard for agentic AI. The ReAct paradigm consists of "Reasoning + Action" (Yao et al. 2022), where an agent is given time to reflect on its own findings and the ability to act upon an environment with external tools. These external tools are nothing but functions that can provide additional functionality, e.g. web search, RAG retrieval, API calls, complex mathematical operations and others.

Blackwell's RAG agent (Figure 5) possesses two main tools:

1. `retrieve_documents (query: str, k: int)`: Gives access to the local vector database allowing the agent to perform vector similarity search utilizing a custom query and number (`k`) of documents to be retrieved.

2. `web_crawl_medline (urls: str)`: Allows the agent to access the content of a web page given url link;

The `web_crawl_medline` tool is necessary due to the number of embedded links in every document of the indexed Medline library, so each document retrieved is accompanied by several correlated links to related pages. Thus, the standard procedure of the

RAG agent is to retrieve documents for a given query, verify if the documents are relevant, select the most important links of those documents and web crawl those links to retrieve more context, rinse, and repeat.

However, even though the guidelines for the standard procedure are given through prompt engineering, the agent has full control over the pipeline, being able, for example, to retrieve multiple documents multiple times and then web crawl the related links.

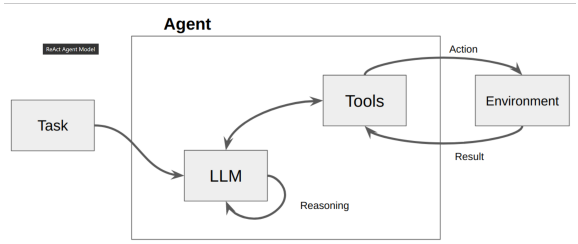

Figure 4: ReAct Agent

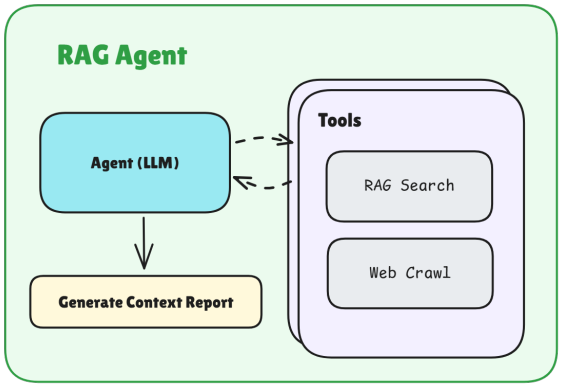

Figure 5: RAG Agent

3.3.2 PubMed Agent

The PubMed Agent is also based on the ReAct architecture, however, it only runs in the treatment phase of the pipeline, therefore being completely focused on researching state-of-the-art treatment guidelines, techniques, and studies available at the PubMed library. It possesses the following tools:

1. `research_treatment_options`: Searches for recent clinical trials, systematic reviews, and meta-analyses for a given diagnosis;
2. `research_specific_treatment_efficacy`: Searches PubMed for research on treatment effectiveness to back a treatment hypothesis for a given diagnosis;
3. `get_treatment_guidelines`: Searches for official treatment guidelines and consensus statements from medical societies;

3.3.3 RAG Infrastructure & Knowledge Base

The MedlinePlus service belongs to the National Library of Medicine (NLM), the world's largest medical library, part of the National Institutes of Health (NIH) (*About MedlinePlus* n.d.). They publish their complete knowledge base (aggregated by groups) in an publicly available extensive XML file. [2] This file is the one indexed in Blackwell's local database, in addition, given that the file contains a myriad of related links present in every topic, this creates the necessity (talked in sub-section 3.3.1) of the ability to explore and ingest the content of those links.

The ChromaDB vector database indexes the whole file split into chunks, given that the utilized embedding model is limited to 2048 tokens, the chosen chunk size was of 1536

[2] Available at `https://medlineplus.gov/xml.html`

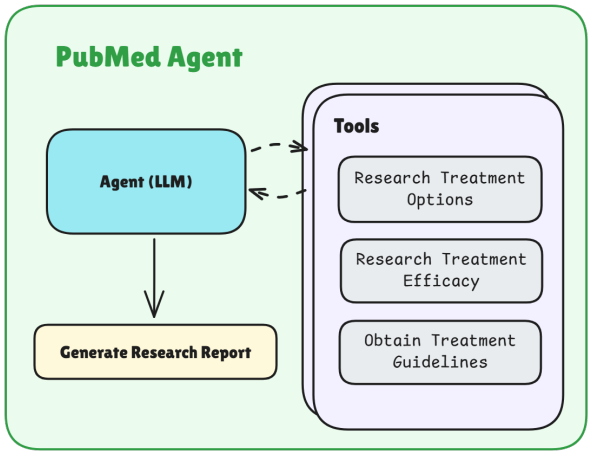

Figure 6: PubMed Agent

tokens with a 500 token overlap to improve the efficiency and accuracy of vector similarity search. The `gemini-embedding-001` model is utilized to generate those embeddings and effectuate the similarity search in the RAG agent. It is a state-of-the-art embedding model that produces up to 3072 dimensions, which translates to being really precise when performing the semantic search.

3.4 Application & Web Interface

The Web interface was implemented with the goal of being intuitive, straight-forward and simple. It is based on two main pages, the first being the anamnesis page that contains a Chat Box and a presentable window for the anamnesis report (Figure 7. The second page (Figure 8) is destined for presenting the complete evaluation report, with an extensive parsing logic, to turn the simple markdown into a visual interesting-looking report. It is built upon FastAPI with Jinja 2 templates for a modern look.

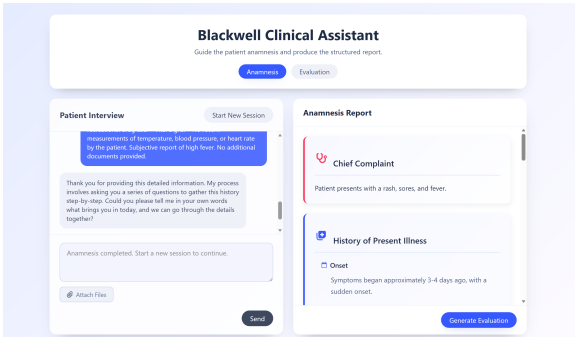

Figure 7: Anamnesis Page

Figure 8: Evaluator Page

3.5 Explainability (xAI)

Finally, the explainability observer of Blackwell is handled by Langsmith and is shown in the final evaluator report. A reasoning summary behind every aspect of the report (such as diagnosis, treatment plan and exam choices) is presented to the user in the web interface and a complete reference section (Figure 10) to support its reasoning, hence

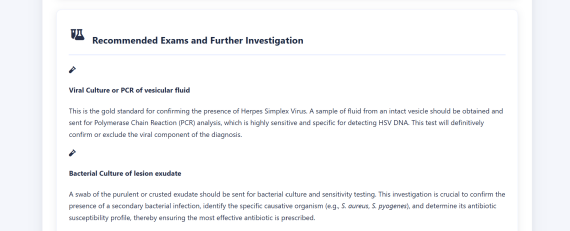

Figure 9: Exams Section

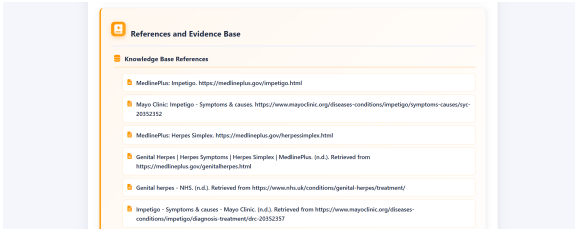

Figure 10: References section

improving the transparency of the decision-making process. Furthermore, we are able to inspect every LLM Call and Tool call - with visibility of every input and output tokens - that Blackwell performs via tracing by Langsmith, which is an external application of the LangChain family. These traces are saved for free (up to 5000) and available in the Langsmith Web UI.

4 Results & Discussions

In this section, the main research question we tried to answer is: **Is it possible to conduct a safe, accurate clinical diagnosis using agentic AI relying solely on conversational text (anamnesis) as input?**

The evaluation of this research question was divided into two main sections to tackle. The anamnesis coverage, reviewing the symptom recall and anamnesis emergency safety guards, and a detailed analysis of the evaluator's performance, reviewing its accuracy, completeness, and quality and comparing it to Gemini with our Anamnesis Agent and Stock Gemini. Lastly, dive deeper into the worst performing diseases and finish with the treatment plan evaluation.

The ethical implications of this project demand extensive carefulness and attention. The most impactful risk is the potential for incorrect medical recommendation, to handle this issue, the system was only tested with artificial patient data. Moreover, Blackwell is classified as a support tool, not a replacement for medical practitioners, the system output always contains a disclaimer to not be used as direct medical documentation and demands a qualified practitioner review.

4.1 The Input Data

Due to ethical, privacy and time concerns about using real patient data for this project, the evaluation was executed using case studies of synthetic patients. To achieve this, two artifacts were joined, a symptom-disease dataset [3], and an AI patient agent. The dataset contains an extensive list of prognosis with its correspondent symptoms, in total, there are 132 available symptoms controlled by an individual boolean variable that indicates whether the patient has that symptom or not, and 41 distinct patient's true condition. Regarding the patient simulator, Yu et al. (2024) creates a solution that simulates an AI patient with LLMs, and we designed an agent that will ingest the list of symptoms for a particular prognosis and create a cohesive background story for its condition.

Even though real patients have an extra degree of unpredictability, our AI patient is designed to simulate the vagueness, lack of knowledge of a real patient as close as

[3] Available at Kaggle

possible. It will always communicate its symptoms and sensations as a human would, without medical jargon, and (apart from its chief complaint) will only say something relevant, such as family history or related symptoms, if asked for. Knowing that LLM-to-LLM circularity is an evaluation limitation, the patient agent is directed to be noisy and possibly say irrelevant information, since its the Anamnesis Agent's job to select and synthesize what is actually relevant from the conversation. A snippet for the AI Patient prompt:

```
ai_patient_prompt = SystemMessage(
    content="""
...
PATIENT PROFILE:
{PATIENT_PROFILE}
PATIENT CONDITION:
{PATIENT_CONDITION}
2. CRITICAL RULES OF REALISM
This is the most important part. Real patients are not databases.
- NEVER LIST YOUR SYMPTOMS: Do not say \"I have a fever, a cough, and a
    headache.\" This is an absolute rule.
- SPEAK NATURALLY AND IMPRECISELY: Use human-like, conversational language.
Bad: \"I am experiencing symptom_cough.\"
Good: \"I've had this nasty cough for a few days.\"
Bad: \"My fever is 1.\"
Good: \"I just feel so hot and shivery.\"

- DO NOT VOLUNTEER EVERYTHING: Let the agent work. Start by mentioning only
    your \"chief complaint\" (e.g., \"Hi, I'm just feeling terrible\") and let
    them ask questions to discover your other symptoms.
- BE \"NOISY\" (based on your persona):
Add small, irrelevant details.
Be slightly vague.
Misunderstand a question occasionally.
Express emotions (worry, frustration, etc.) as your persona dictates.
- STAY IN CHARACTER: You are the patient. You do not know your diagnosis. You
    cannot run tests. You can only report how you feel.
...
```

In total, given the cost of each complete run (Anamnesis + AI Patient + Evaluator), 250 prognosis were randomly selected from the dataset for this evaluation. The general procedure for each prognosis was as follows.

1. Select a random case from the dataset;
2. Simulate an AI Patient;
3. Run the Anamnesis Procedure with the AI Patient;
4. Run the Evaluator Procedure with the obtained Anamnesis report;

4.2 Anamnesis Agent

In the evaluation of the anamnesis procedure, the two main points of focus were safety guards and the ability to recall symptoms. The intention was to verify whether the anamnesis agent can identify and retain (in the report) all the symptoms the patient has, within the boundaries of textual communication. In addition, ensure that the agent safely refers the patient to the hospital in the event of an emergency.

4.2.1 Symptom Recall Analysis

To identify whether a symptom was present in the anamnesis report, a simple regex match would not be functional, since the symptom could be listed as headache in the dataset, but as cephalalgia in the report. To account for this possibility, it was necessary to perform a semantic match using cosine similarity instead. Cosine similarity match is the current standard for matching words by their meaning, comparing the embeddings vector of two segments of text (single word, phrases or paragraph). To generate the embeddings we utilized the same model that the RAG structure is built upon, `gemini-embedding-001`, and finally, calibrating the similarity threshold by manually observing several cases to account for a proper balance.

After excluding emergency cases, since its anamnesis is halted mid-way, 181 cases remained. In those, each of the "felt" symptoms was semantically matched to each individual line of the report to verify whether it was present in the report or not. The most reliable threshold value to run the cosine similarity with `gemini-embedding-001` was 0.81, thus, if at least one of the lines in the report had a cosine similarity of ≥ 0.81 with a particular symptom, that symptom is then included in the recalled symptoms list for that particular case.

Outstanding performance was obtained with a mean symptom recall (SR) of 84.06% (SD = 16.32%, median = 85.71%), with 69 cases computing 100% of SR, 121 cases above 80%, and only 6 cases (3.3%) with less than 50%. Interestingly, it was observed that the least retained symptoms (Figure 11) are not normally reported by text/conversation, such as "obesity", meaning that usually a patient does not complain about obesity, and even though the agent asks how much the patient exercises and other general information, such as age, sex, and weight, that was the most missed symptom of the dataset. Another fact is that symptoms that are not identifiable without an unusual external machine, like "irregular sugar level", are also prominent to be on that list.

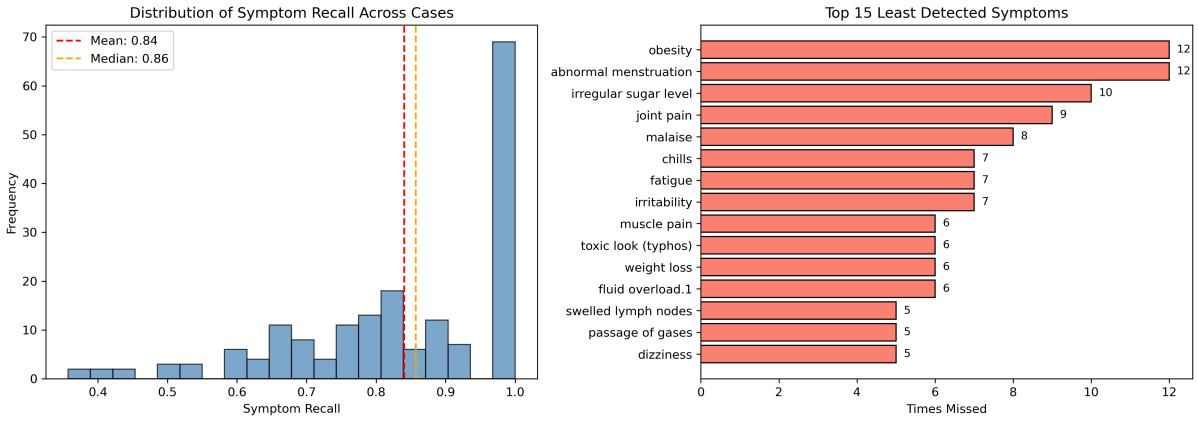

Figure 11: Distribution & Least Detected Symptoms

Figures 12 and 13 show us that the condition with the worst SR is Dengue with 52.47%. Although Dengue is a multi-complainant disease with a myriad of symptoms, Blackwell did not identify easily retainable symptoms such as "muscle pain" and "joint pain". However, the standard deviation of 16.32% demonstrates that these conditions are outliers in our analysis.

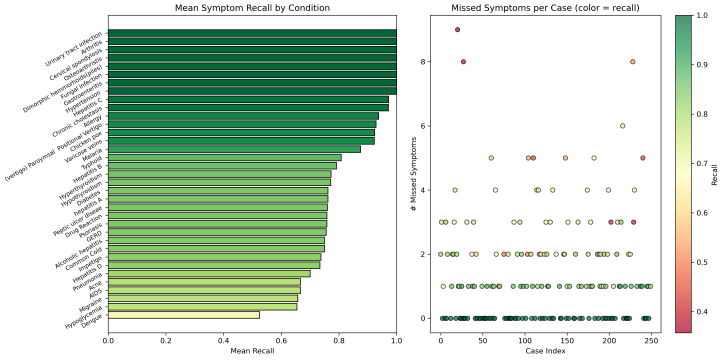

Figure 12: Mean SR by Condition & Distribution by Index

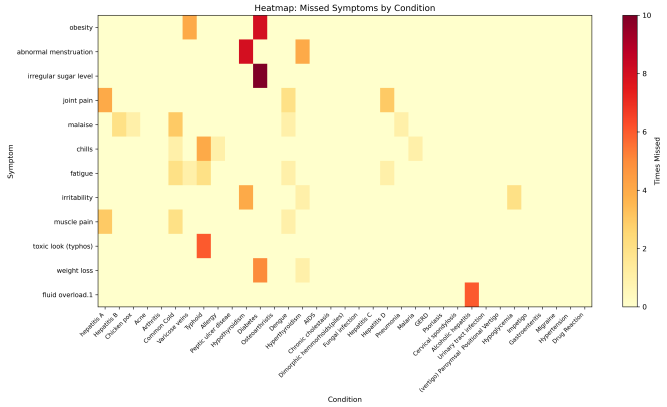

Figure 13: Missed Symptoms by Condition

4.2.2 Emergency Detection Analysis

In the emergency detection analysis, the focus points are how well the agent detects an emergency and how rapidly it does so. Since neither the AI patient nor the Anamnesis agents know the actual patient's true condition, the detection is based on emergency-related symptoms, those symptoms were: **chest pain, breathlessness, vomiting, altered sensorium, bleeding, slurred speech, loss of balance and weakness of one body side.**

Firstly, the analysis check if any emergency symptom is present in a particular case, then it verifies if the agent correctly stopped the anamnesis and classified as a emergency. Therefore, the possible outcomes of this analysis are:

- **TP (True Positive)**: Emergency detected AND patient had emergency symptoms - Correct;

- **FP (False Positive)**: Emergency detected but NO emergency symptoms - Unnecessary halt;
- **TN (True Negative)**: No emergency detected AND no emergency symptoms - Correct continuation;
- **FN (False Negative)**: No emergency detected but patient HAD emergency symptoms - Dangerous miss;

Analyzing the data obtained (Figure 14), although the overall accuracy was 78%, we can observe a worrying amount of false negatives (10.8%), and most of them are related with the "chest pain" symptom. However, diving deeper to understand those FNs, in Figure 15, the largest ratios for FN/TP are for non-emergency conditions, such as "Common Cold", "Cervical spondylosis" and "Paroxysmal Positional Vertigo". Therefore, we can assume that the Anamnesis agent is sensitive not only to the symptom but how the patient describes the actual symptom which is actually the desired execution.

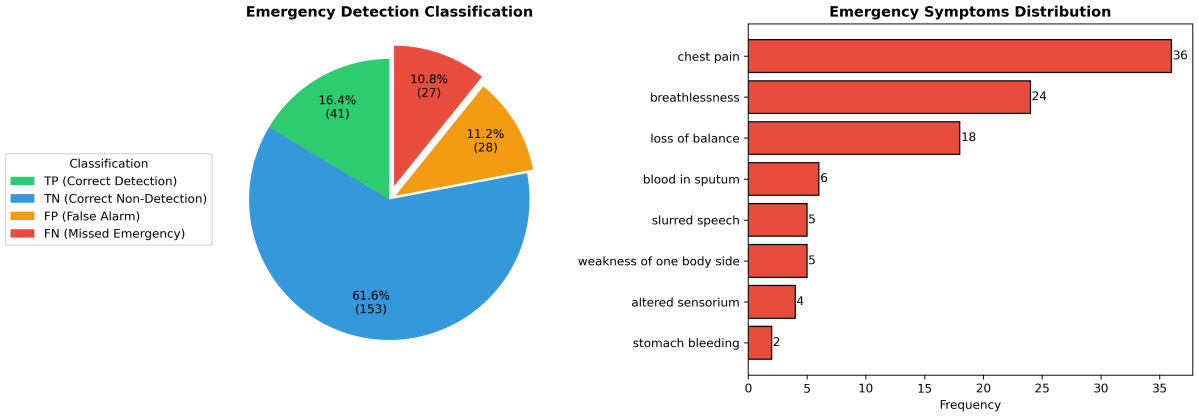

Figure 14: Emergency Detection Summary

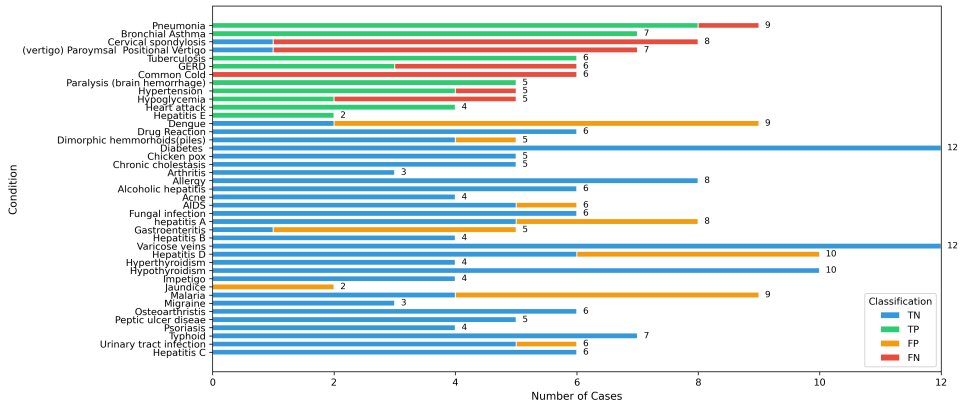

Figure 15: Emergency Classification by Condition

Initially, Figure 16 shows that 81.5% (22/27) of the false negatives, after a qualitative review, were actually non-emergency conditions, turning them into true negatives:

- Cervical spondylosis (7 cases) - Causes loss of balance symptoms but is a chronic neck condition;

- Common Cold (6 cases) - Not an emergency;
- Vertigo/BPPV (6 cases) - Causes loss of balance but is benign positional vertigo;
- Gastroesophageal Reflux Disease (GERD) (3 cases) - Causes chest pain mimicking cardiac issues, but is acid reflux;

This demonstrates that the agent's ability of contextualizing symptoms works significantly better than direct binary classification, for instance, the agent identified that the "chest pain" described in GERD cases is different than a burning chest pain from a cardiac arrest. This re-evaluation increases the emergency detection accuracy to 86.8%. Finally, given that the most present false positives were conditions that could actually require urgent care, the only remaining point of concern are the remaining 5 cases (Hypoglycemia, Pneumonia, Hypertension), which can be very dangerous conditions therefore needing more review of the agent's behavior.

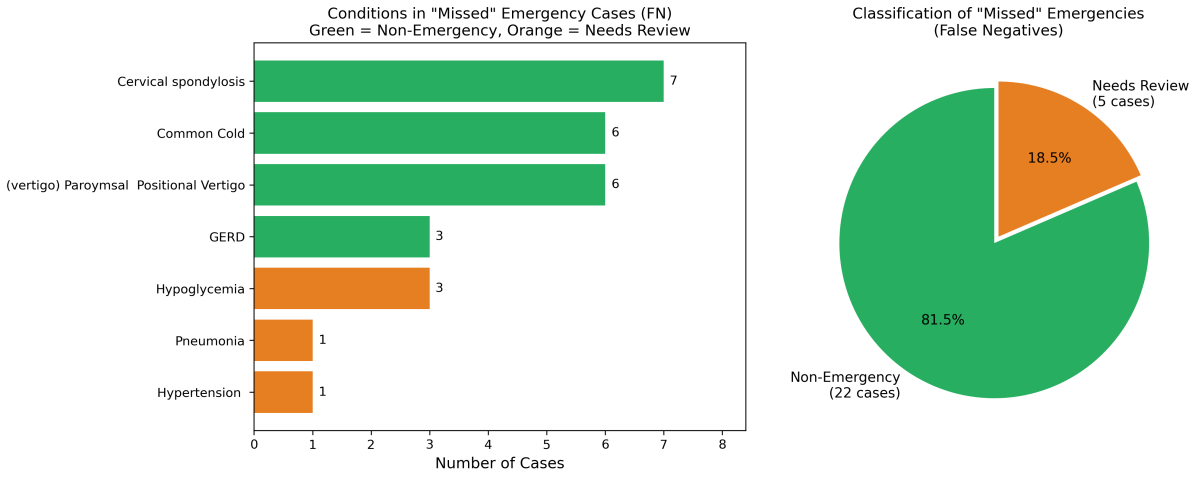

Figure 16: False Negative Analysis

Lastly, on the performance side of the analysis, we are able to focus on the number of exchanges (total number of messages between AI Patient and Anamnesis Agent) necessary for an anamnesis completion. The average is 21.9 exchanges per anamnesis (Figure 17), in addition, it is observable that emergency cases are detected rapidly requiring 13.4 fewer exchanges than normal completion, this is expected behaviour, as the agent halts the anamnesis prioritizing escalation to urgent care.

4.3 Evaluator Agent

Evaluating the main agent is a difficult task given its final output is a report with more than 3000 tokens. Judging the performance as a boolean (correct or incorrect diagnosis) would be an understatement of the value that Blackwell brings. In this analysis emergency cases were also skipped since there was no anamnesis report to start the procedure, presenting a total sum of 181 cases.

Similarly to the anamnesis matching problem, the evaluator also can output different ways of calling the same condition - e.g. "(vertigo) Paroxysmal Positional Vertigo" condition in the dataset was classified as "Benign Paroxysmal Positional Vertigo (BPPV)" by our agent which essentially means the same condition - therefore, after a thorough

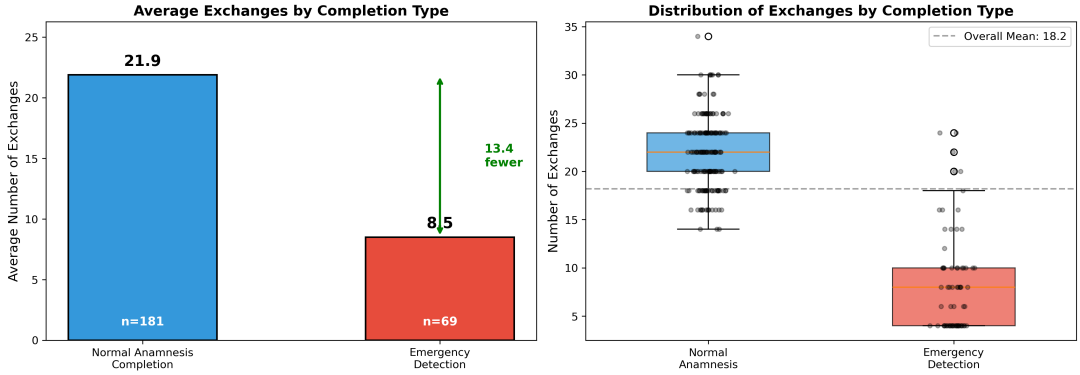

Figure 17: Anamnesis Exchanges Analysis

parsing, semantic matching by cosine similarity was also the chosen technique to evaluate the accuracy.

4.3.1 Diagnosis Accuracy Analysis

However, differently from the anamnesis analysis, we segregated the threshold into two levels, a correct diagnosis, where cosine similarity $\cos_{sim} \geq 0.85$ and a partially correct diagnosis, $0.8 \leq \cos_{sim} \leq 0.85$. This relies on the judgement that a partially correct diagnosis (e.g. "Common Cold" being diagnosed as "Suspected COVID-19") would have very similar confirmation exams, investigation lab tests and treatment plan, but is not a complete correct diagnosis. In addition, Blackwell not only produces a probable diagnosis, but also lists secondary differential diagnosis accompanied by exams and lab tests that are standard to identify the true condition, which also can be fully or partially matched.

Finally, to account for all those possibilities, a score system was designed to evaluate the diagnosis correctness. This system is straight-forward and is intended to give more value to a correct main diagnosis and less value to partially correct differential secondary diagnoses, Table 2 gives a summary of how the score system works:

Component	Points	Description
Primary Diagnosis	90	Correct primary diagnosis (similarity ≥ 0.85)
Primary Diagnosis (Partial)	60	Partially correct primary diagnosis (similarity ≥ 0.80)
Differential Diagnosis	70	Correct diagnosis found in differential diagnoses
Differential Diagnosis (Partial)	40	Partial match in differential diagnoses
Completeness	4	1 point for each section: differential, exams, treatment plan, references
Patient-Specific	6	Bonus for including patient-specific considerations

Table 2: Report Evaluation Score System

Hence, a report can score from 0 to a 100 with this system, and the said score can be interpreted as:

- **90-100**: Excellent - Correct primary diagnosis;
- **60-80**: Good - Partial match or differential correct;
- **40-50**: Fair - Partial differential match only;
- **0-10**: Poor - Incorrect diagnosis, completeness points only;

In Figure 18 we can observe that the majority of diagnoses presented maxed scores (123 out of 181). Additionally, considering partial and differential matches, only 9.9% (18 cases) had completely incorrect diagnoses. Every report scored the extra points for completeness, which represents that the output structure is consistent throughout the sample cases, and only 2 reports missed the points for individual patient considerations.

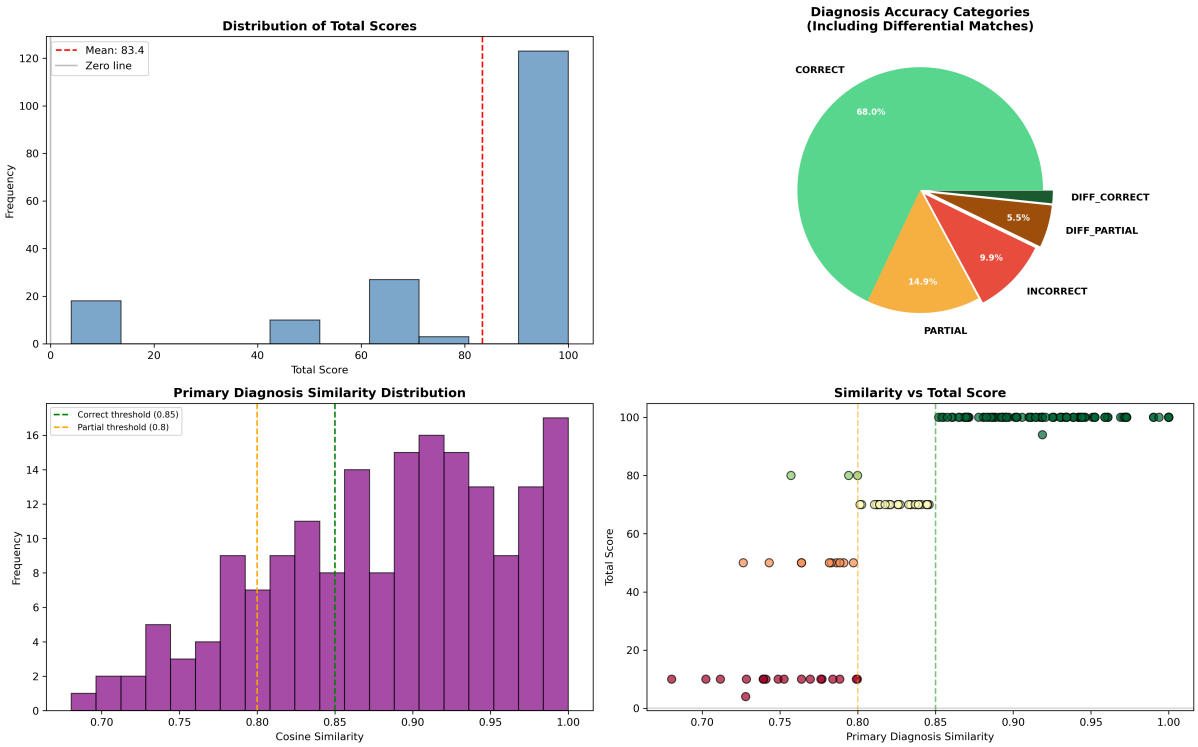

Figure 18: Scoring Analysis Dashboard

Although not ideal, 13 cases had their diagnosis matched in the differential diagnosis section, with 10 being a partial match, which means that when Blackwell does not correctly diagnose it does not hallucinate and completely gives dangerous recommendations. Furthermore, Figure 19 shows that only a few conditions have a mean score less than 40 (the partial differential diagnosis match score).

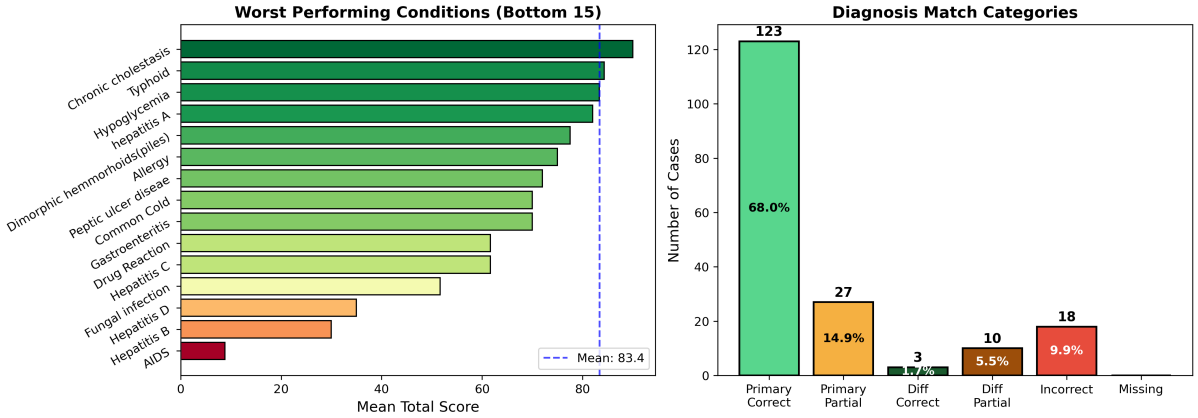

Figure 19: Diagnosis Analysis

Nonetheless, when comparing Blackwell to Gemini, it was decided to run Gemini in

2 different ways to also verify the impact the Anamnesis Report made on the overall accuracy.

- **Gemini Stock**: This first Gemini version would be a completely raw `Gemini-2.5-Pro` model, with a simple prompt that after receiving the patient's complaint would output a complete diagnosis and treatment plan;
- **Gemini (W/Anamnesis)**: This version receives the complete Anamnesis report and the same thoroughly engineered prompt that Blackwell utilizes at its final node;

With Figure 20 we can observe that stock Gemini is significantly less accurate than the other versions, showing the huge impact that our anamnesis agent provides, its partial diagnoses are also large lists of most varied diagnoses diminishing its value. In addition, this version did not possess any safety guards, stopping its final report to alert the user to seek for emergency care.

However, Blackwell's evaluator agent underperformed compared to Gemini (W/A) where it only had 4.8% of incorrect diagnosis. Gemini (W/A) mainly outperformed Blackwell on Hepatitis diagnosis, drawing attention to the strength of our knowledge base, which can introduce noise or retrieval distractions to the model.

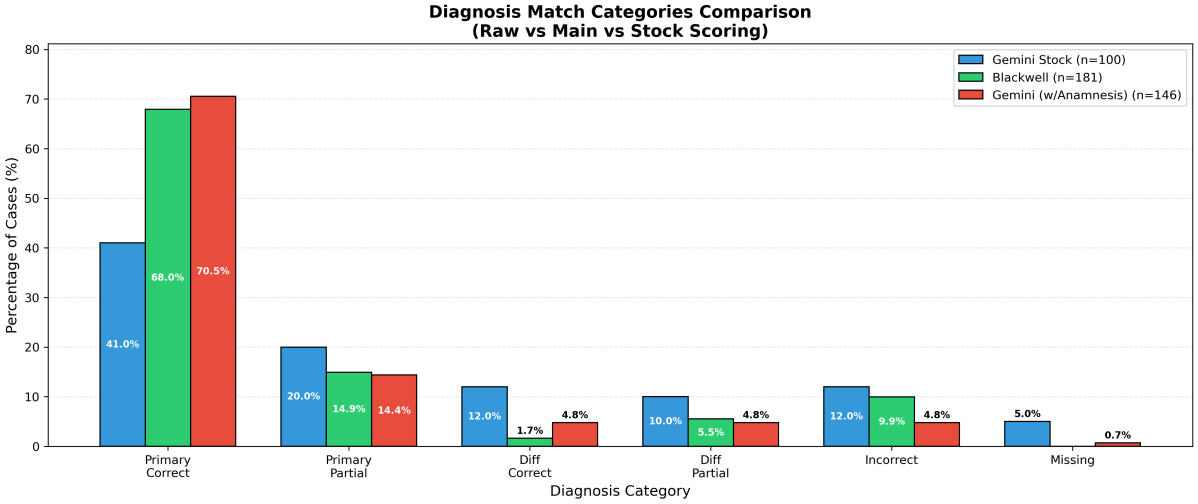

Figure 20: Accuracy Comparison Between Models

Although Gemini (W/A) achieved higher accuracy than Blackwell, it operated as a "Black Box", not referencing a single verifiable source in its reports and relying solely on its internal training knowledge being more prone to hallucinate. Finally, all three versions presented a poor performance in Acquired immunodeficiency syndrome (AIDS) diagnosis.

4.3.2 Worst Performers Investigation

Analyzing the less-performant conditions, "AIDS", "Hepatitis B" and "Hepatitis D", to understand why the model struggles with certain diagnoses. Firstly, focussing on the AIDS cases (Figure 21), Blackwell struggles with this most, incorrectly diagnosing it in all 5 appearances and in only 2 of those cases it recommends human immunodeficiency virus (HIV) testing, which is a poor result. Although it also recommends Complete Blood

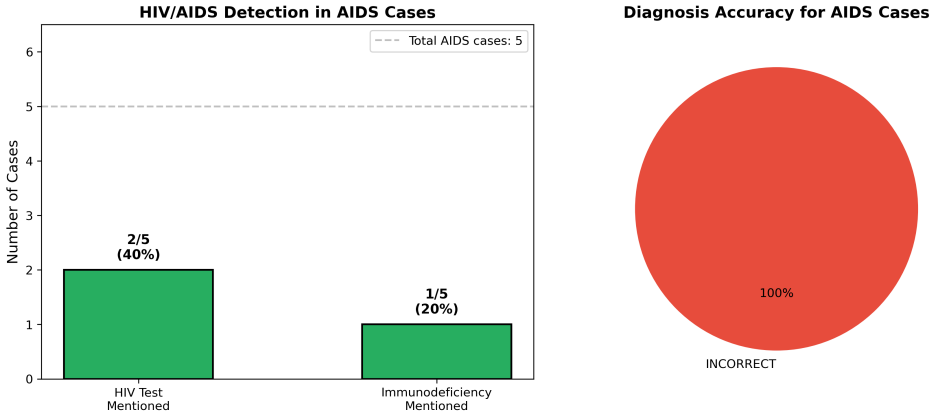

Figure 21: AIDS Investigation Analysis

Count (CBC) test that could be a workup for AIDS diagnosis, the model has difficulty connecting AIDS symptoms with HIV testing.

In the case of Hepatitis (Figure 22), the lack of precision is less prominent, yet the model delivers underwhelming results when compared to Gemini (W/A). Hepatitis D gave the weakest results with only 33% accuracy, and 50% of the overall cases were incorrectly diagnosed. Additionally, only 38% included a hepatitis test recommendation, however, in every scenario, a liver function test was present in the exams and further investigation section which is a workup to the Hepatitis diagnosis. This indicate a need for a thorough inspection of our knowledge base, primarily on the hepatitis and liver diseases content. The common misdiagnosis and possible reasoning behind it is as follows.

- **Obstructive Jaundice**: Liver symptoms but undermining viral etiology;
- **Cholestasis**: Identifies issues with bile flow but not hepatitis;
- **Pancreatic Cancer**: misidentifies jaundice symptoms as malignancy;

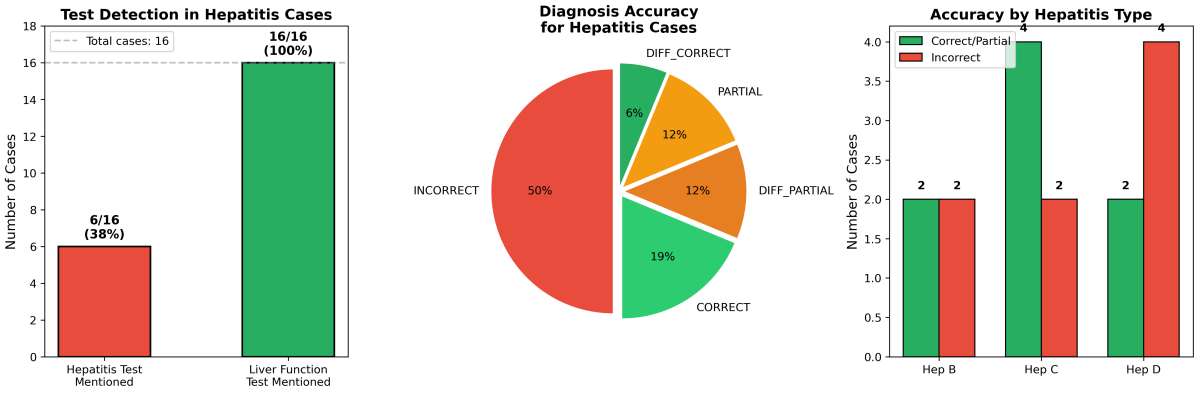

Figure 22: Hepatitis Investigation

4.3.3 Treatment Plan Analysis

The utilized dataset of symptom-disease did not include a treatment information, therefore, its unfeasible to evaluate the effectiveness of generated treatment plans in this project given time and financial constraints. Nevertheless, a few metrics have been chosen

to verify the completeness of the treatment plan, these metrics cannot tell if a plan is confirmedly accurate, but can give us a notion of how thorough they are. Description of the metrics are as follows.

1. **Lifestyle Advice**: Lifestyle change suggestions to improve the condition or alleviate symptoms of the patient; e.g., diet, exercise, rest, avoid alcohol;
2. **Medication Names**: Medication recommendations; e.g., antibiotic, metformin, corticosteroid;
3. **Duration of**: Length of the treatment plan; e.g., for 7 days, 2 weeks, long-term;
4. **Warning Signs**: Specific orientations to seek urgent care if conditions worsen; e.g., "seek emergency attention if...", "warning signs include...";
5. **Monitoring Plan**: Ongoing monitoring recommendations; e.g., "monitor blood pressure", "check glucose levels", "repeat the test in 3 weeks";
6. **Dosages Specified**: Specific dosage amounts and frequencies were provided; e.g., "500mg twice a day", "10 mg every 8 hours";
7. **Follow-up Plan**: If return visits or future reassessments are present; e.g., "return if symptoms persist", "schedule appointment", "follow-up in 2 weeks";

With these established metrics, we can gather insight about the treatment plan extensiveness. Similarly to previous sections, the cosine similarity ($match \geq 0.85$) of vector embeddings of text segments is the chosen metric to verify the presence of these artifacts. Figure 23 shows that at least three of them - lifestyle advice (98%), duration (90%), and medication names (97%) - are always mentioned in all reports. Moreover, Blackwell's treatment completeness appears to be consistent throughout all samples, averaging 5 of 7 artefacts, independently of the correctness of the diagnosis.

The least present metrics are fairly situational, such as follow-up and monitoring plans that are not always necessary. While medication names are almost always mentioned, dosages are only specified half of the times. This discrepancy demands further investigation, however, the lower presence of dosing information aligns with a conservative safety profile, since proper dosage generation require information about weight and lab tests confirmation. Finally, it is noticeable that worse performing treatment plans are for the conditions in which we found that Blackwell struggles the most, like hepatitis and AIDS, meaning that the model is weak at recommending specific treatment for these conditions.

5 Conclusion

The primary objective of this project was to assess the capability of agentic architectures to perform clinical diagnosis solely relying on conversational text. In our evaluation, it was assessed that for chronic/symptom-manifesting diseases, such as Thyroid and Diabetes, it definitely can be a sufficient option, achieving a diagnostic direct correctness of 68% with differential diagnoses and partial matches bringing it to a total of 90.1% accuracy. Whereas, text proved to be insufficient for conditions that require visual inspections or serology, since their symptoms are shared with multiple other conditions.

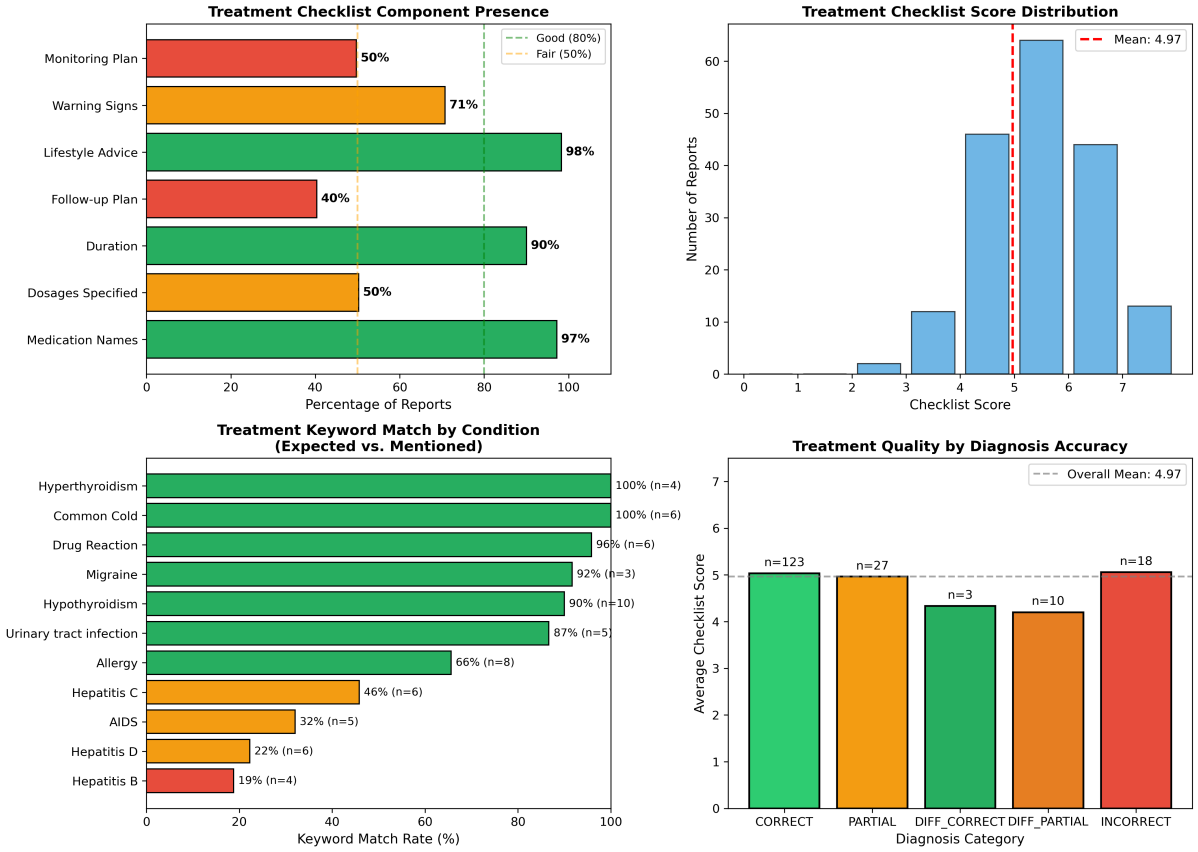

Figure 23: Treatment Plan Evaluation Metrics

The Anamnesis agent provided great results in symptom recall (84%) and emergency referral (86.8%), proving that it is a valid use case for AI agents. It was able to evaluate symptoms better than a rule-based proposition, given it could analyze the clinical urgency of the symptoms (rejecting false emergencies) through context, and greatly diminishing the false negatives. It also proved to be vital in powering the most simplest LLMs pipelines to achieve higher accuracy. However, for high-stake domains, such as clinical diagnosis, AI applications still need to be near perfect, requiring further investigation of the system's outliers.

The Evaluator deep agent also provided great results for the majority of the conditions while providing sources for every decision and reasoning behind it, while Gemini (w/Anamnesis) had slightly higher accuracy, it didn't referenced any of its sources. On the other hand, some conditions (AIDS and Hepatitis) that are characterized by non-specific symptoms overlapping with other conditions (e.g. immunodeficiency) presented poor results. Upon this discovery, we can explore case-specific optimizations or direct safety guards implementations for those cases, e.g., integration of immunodeficiency-specific articles in the knowledge base, or direct referral to practitioner knowing its a weakness of the framework. The treatment analysis requires more evaluation regarding effectiveness but its completion proved to be consistent and sufficient having both therapeutic and lifestyle options.

5.1 Future Work

Given the short time frame dedicated to development of the project, some features were left in this section for future analysis and evaluation, those are:

- **Strengthen the Knowledge base**: Including established foundational literature (e.g. Papadakis et al. (2025), Loscalzo et al. (2025), McKean et al. (2016)) could be a way of improving accuracy of the framework, however, that requires legal licensing and permissions that we were not able to get in time;
- **Multi-modalilty:** Integrating Visual Language Models (VLMs) into the framework. Since Gemini 2.5 is capable of handling images, it would be an interesting feature to add to the reasoning pipeline to improve the accuracy on skin conditions, and ingestion of image exams;
- **Mixture-of-Experts (MoE):** Continuing the MoE concept of the framework. Splitting the main agents into more specific sub-agents specialized in multiple medical fields (e.g. pharmaceutical agent, dermatology agent, psychiatric agent) to specialize its knowledge and recommendations. A pharmaceutical agent could greatly improve the limitations uncovered in dosage specifications.

In summary, Blackwell proves that, given adequate resources and refining, agentic AI can already venture into this critical domain, lacking however, other technologies to improve its decision making, for example automatic examination tools and procedures. Imagining a use case for poor remote locations lacking public health infrastructure, Blackwell provides a solution that needs little refinement to give a proper support for most conditions.

References

About MedlinePlus (n.d.), `https://medlineplus.gov/about/`. Last updated: October 12, 2023. Accessed: 2025-11-26.

Ansari, M. S., Khan, M. S. A., Revankar, S., Varma, A. & Mokhade, A. S. (2025), 'Lightweight clinical decision support system using qlora-fine-tuned llms and retrieval-augmented generation'.
URL: *https://arxiv.org/abs/2505.03406*

Chen, X., Yi, H. et al. (2025), 'Enhancing diagnostic capability with multi-agents conversational large language models'.
URL: *https://pmc.ncbi.nlm.nih.gov/articles/PMC11906805/*

Chen, Y.-J., Albarqawi, A. & Chen, C.-S. (2025), 'Enhancing clinical decision-making: Integrating multi-agent systems with ethical ai governance'.
URL: *https://arxiv.org/abs/2504.03699*

Chu, Y.-W., Zhang, K., Malon, C. & Min, M. R. (2025), 'Reducing hallucinations of medical multimodal large language models with visual retrieval-augmented generation', *arXiv preprint arXiv:2502.15040* .
URL: *https://arxiv.org/abs/2502.15040*

Dailah, H. G., Koriri, M., Sabei, A., Kriry, T. & Zakri, M. (2024), 'Artificial intelligence in nursing: Technological benefits to nurse's mental health and patient care quality', *Healthcare (Basel)* **12**(24), 2555.

Deep Agents overview (n.d.), `https://docs.langchain.com/oss/python/deepagents/overview`. Accessed: 2025-11-17.

Ding, H., Simmich, J., Vaezipour, A., Andrews, N. & Russell, T. (2024), 'Evaluation framework for conversational agents with artificial intelligence in health interventions: a systematic scoping review', *Journal of the American Medical Informatics Association* **31**(3), 746–761.

Gao, Y., Xiong, Y., Gao, M., Jia, J., Yin, J., Xu, C., Cui, Z., Sun, B., Li, H., Hu, X. et al. (2023), 'Retrieval-augmented generation for large language models: A survey', *arXiv preprint arXiv:2312.10997* .

Gargari, O. K., Mahmoudi, M. H., Hajisafi, M., Alimardani, F. & Alebouyeh, Z. S. (2025), 'Enhancing medical ai with retrieval-augmented generation: A mini narrative review', *Digital Health* **11**.

Google AI for Developers (2025), 'Long context — gemini api — google ai for developers', `https://ai.google.dev/gemini-api/docs/long-context`. Accessed: 2025-07-24.

Huang, Y., Chen, Y., Zhang, H., Li, K., Zhou, H., Fang, M., Yang, L., Li, X., Shang, L., Xu, S., Hao, J., Shao, K. & Wang, J. (2025), 'Deep research agents: A systematic examination and roadmap'.
URL: *https://arxiv.org/abs/2506.18096*

Huang, Y., Lin, Z., Zhong, F. & Liu, L. (2025), 'Learning to seek evidence: A verifiable reasoning agent with causal faithfulness analysis'.
URL: *https://arxiv.org/abs/2511.01425*

Hurt, R. T., Stephenson, C. R., Gilman, E. A., Aakre, C. A., Croghan, I. T., Mundi, M. S., Ghosh, K. & Varayil, J. E. (2025), 'The use of an artificial intelligence platform openevidence to augment clinical decision-making for primary care physicians', *Journal of Primary Care & Community Health* **16**, 21501319251332215.
URL: *https://doi.org/10.1177/21501319251332215*

Joksimović, Z. & Bastać, D. (2022), 'Anamnesis: The skill and art of clinical medicine', *Timocki medicinski glasnik* **47**, 153–156.

Kizilcec, R. F. (2016), How much information? effects of transparency on trust in an algorithmic interface, *in* 'Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems', CHI '16, Association for Computing Machinery, New York, NY, USA, p. 2390–2395.
URL: *https://doi.org/10.1145/2858036.2858402*

Liu, Q., Chen, X., Liu, Q. & Chen, X. (2024), Clinicalrag: Enhancing clinical decision support through heterogeneous knowledge retrieval, *in* 'ACL'.
URL: *https://aclanthology.org/2024.knowllm-1.6/*

Lizée, A., Beaucoté, P.-A., Whitbeck, J., Doumeingts, M., Beaugnon, A. & Feldhaus, I. (2025), 'Conversational medical ai: Ready for practice'.
URL: *https://arxiv.org/abs/2411.12808*

Loscalzo, J., Fauci, A. S., Kasper, D. L., Hauser, S. L., Longo, D. L. & Jameson, J. L., eds (2025), *Harrison's Principles of Internal Medicine*, 22 edn, McGraw Hill.

McKean, S. C., Ross, J. J., Dressler, D. D. & Scheurer, D. (2016), *Principles and Practice of Hospital Medicine*, 2 edn, McGraw Hill.

Neha, F., Bhati, D. & Shukla, D. K. (2025), 'Retrieval-augmented generation (rag) in healthcare: A comprehensive review', *Preprints* .
URL: *https://doi.org/10.20944/preprints202508.1022.v1*

Nishita, K., Prachi, V. & Surajpal, V. (2024), 'Advancements in ai based healthcare techniques with focus on diagnostic techniques', *Computers in Biology and Medicine* **179**, 108917.
URL: *https://www.sciencedirect.com/science/article/pii/S0010482524010023*

NVIDIA (2025), 'Traditional rag vs. agentic rag: Why ai agents need dynamic knowledge to get smarter', NVIDIA Developer Blog. Accessed from `https://developer.nvidia.com/blog/`.

Papadakis, M. A., Rabow, M. W., McQuaid, K. R. & Gandhi, M., eds (2025), *Current Medical Diagnosis and Treatment 2025*, 64 edn, McGraw Hill.

PubMed (n.d.), `https://pubmed.ncbi.nlm.nih.gov/`. Accessed: 2025-09-14.

Rezaei, M. R., Saadati Fard, R., Parker, J. L., Krishnan, R. G. & Lankarany, M. (2025), 'Agentic medical knowledge graphs enhance medical question answering: Bridging the gap between llms and evolving medical knowledge'.
URL: *https://arxiv.org/abs/2502.13010*

Safdari, R., Shoshtarian Malak, J., Mohammadzadeh, N. & Danesh Shahraki, A. (2017), 'A multi agent based approach for prehospital emergency management', *Journal of Biomedical Physics & Engineering* **7**(3), 171–178.
URL: *https://pmc.ncbi.nlm.nih.gov/articles/PMC5547204/*

Sapkota, R., Roumeliotis, K. I. & Karkee, M. (2025), 'Ai agents vs. agentic ai: A conceptual taxonomy, applications and challenges', arXiv.
URL: *https://arxiv.org/html/2505.10468v4*

Singh, A., Ehtesham, A., Kumar, S. & Khoei, T. T. (2025), 'Agentic retrieval-augmented generation: A survey on agentic rag'.
URL: *https://arxiv.org/abs/2501.09136*

Sun, Z., Luo, C., Liu, Z. & Huang, Z. (2024), 'Conversational disease diagnosis via external planner-controlled large language models'.
URL: *https://arxiv.org/abs/2404.04292*

Wang, Z., Wu, J., Cai, L., Low, C. H., Yang, X., Li, Q. & Jin, Y. (2025), 'Medagent-pro: Towards evidence-based multi-modal medical diagnosis via reasoning agentic workflow'.
URL: *https://arxiv.org/abs/2503.18968*

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K. R. & Cao, Y. (2022), React: Synergizing reasoning and acting in language models, *in* 'The eleventh international conference on learning representations'.

Ye, Q., Ahmed, M., Pryzant, R. & Khani, F. (2024), Prompt engineering a prompt engineer, *in* L.-W. Ku, A. Martins & V. Srikumar, eds, 'Findings of the Association for Computational Linguistics: ACL 2024', Association for Computational Linguistics, Bangkok, Thailand, pp. 355–385.
URL: *https://aclanthology.org/2024.findings-acl.21/*

Yu, H., Zhou, J., Li, L., Chen, S., Gallifant, J., Shi, A., Li, X., Hua, W., Jin, M., Chen, G., Zhou, Y., Li, Z., Gupte, T., Chen, M.-L., Azizi, Z., Zhang, Y., Assimes, T. L., Ma, X., Bitterman, D. S., Lu, L. & Fan, L. (2024), 'Aipatient: Simulating patients with EHRs and LLM powered agentic workflow'.
URL: *https://arxiv.org/abs/2409.18924v2*

Yu, Y., Li, L. & Li, Y. (2025), 'Augmenting large language models and retrieval-augmented generation with an evidence-based medicine-enabled agent system', *medRxiv* pp. 2025–10.

Zaghir, J., Naguib, M., Bjelogrlic, M., Névéol, A., Tannier, X. & Lovis, C. (2024), 'Prompt engineering paradigms for medical applications: Scoping review', *J Med Internet Res* **26**, e60501.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/39255030*

Zakka, C., Shad, R., Chaurasia, A., Dalal, A. R., Kim, J. L., Moor, M., Fong, R., Phillips, C., Alexander, K., Ashley, E., Boyd, J., Boyd, K., Hirsch, K., Langlotz, C., Lee, R., Melia, J., Nelson, J., Sallam, K., Tullis, S., Vogelsong, M. A., Cunningham, J. P. & Hiesinger, W. (2024), 'Almanac — retrieval-augmented language models for clinical medicine', *NEJM AI* **1**(2), AIoa2300068.
URL: *https://ai.nejm.org/doi/full/10.1056/AIoa2300068*
ΩÖğdü et al.

Öğdü, , Arslanoğlu, K. & Karakose, M. (2025), 'An adaptive multi-agent llm-based clinical decision support system integrating biomedical rag and web intelligence', *IEEE Access* **PP**, 1–1.