IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 31, No. 4, April, 2026, pp. 1285-1296
Journal homepage: http://iieta.org/journals/isi

Bio-Cognito: A Generalist AI Agent Framework for Medical and Computational Biology Deep Research

Pranav Pawar[1], Tishya Dave[1], Rishabh Makwana[1], Siya Rozani[1], Darshana Sankhe[2], Pratik Kanani[1], Snehal Bhosale[3*]

[1] Department of Artificial Intelligence and Data Science, Dwarkadas J. Sanghvi College of Engineering, Mumbai 400056, India
[2] Department of Electronics and Telecommunication, Dwarkadas J. Sanghvi College of Engineering, Mumbai 400056, India
[3] Department of Electronics and Telecommunications, Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune 412115, India

Corresponding Author Email: snehal.bhosale@sitpune.edu.in

https://doi.org/10.18280/isi.310423

Received: 3 November 2025
Revised: 20 January 2026
Accepted: 19 April 2026
Available online: 30 April 2026

Keywords:
biomedical NLP, medical AI agents, hierarchical multi-agent framework, multimodal data fusion, clinical decision support, computational biology, domain knowledge integration, deployment readiness

ABSTRACT

The existing artificial intelligence (AI) agents lack robustness in performing their functions for both medical and computational biology applications, mainly because of low knowledge incorporation related to the field, inadequate clinical reasoning, and poor processing of multimodal data. Application of the AI agents on heterogeneous data sets, especially in genomics studies, also results in performance decline. Bio-Cognito is a multi-agent architecture that incorporates different AI agents for clinical reasoning and other medical applications. Key aspects of this architecture include hierarchical design of agents, algorithm benchmarking, and domain adaptive quantization using Normalized Float 4-bit (NF4) and Low-Rank Adaptation (LoRA). Multi-channel data fusion and learning pipelines have been used by Bio-Cognito for clinical diagnostics, genomics research, and translational medicine. Experiments showed that the performance of Bio-Cognito agents was better than baseline AI agents in terms of accuracy, safety of operation, and performance ratio retention during quantization. The theoretical foundation of the proposed approach includes reasoning chains and knowledge bases.

1. INTRODUCTION

The confluence between Artificial Intelligence (AI) and healthcare is among the fastest-growing segments within modern computing science, with the AI market expected to grow at a rate of 39.4% by 2022–2028 [1]. While recent developments in LLMs are notable, there remain critical barriers related to the deployment of current AI technology solutions in medical and computational biology spaces [2, 3]. Based on current findings, generalist AI systems are able to perform at a maximum level of 60-70% accuracy in medical and clinical reasoning, well below the minimum 95% mark for clinically meaningful outcomes.

Addressing the problem would require the ability to process and combine various data types such as genomics (FASTA files), genetic variants (Variant Call Format or VCF), protein structures (PDB) and clinical records (HL7 FHIR). Among traditionally used solutions, the following problems exist:

- Modal fragmentation, meaning that the models responsible for different modalities are employed separately from each other without proper integration.
- Domain generalization limitations, specifically the underperformance of models that were originally trained on general datasets when applied to specialized medical tasks.
- Deployment restrictions due to excessive computational requirements that cannot be satisfied by typical clinical hardware.

1.1 Research gap and motivation

Current AI agent approaches in medicine fall into two categories. General-purpose agents, such as AutoGPT-Medical and LangChain, offer broad coverage but lack the specialized domain expertise required for tasks such as pharmacogenomics or drug-protein interaction prediction. Task-specific models, by contrast, achieve strong performance within their designated domain—exemplified by AlphaFold for protein structure prediction and MedPaLM for medical question answering—but cannot generalize beyond it. A framework that unifies the coordination capabilities of general agents with the depth of domain-specific knowledge is therefore needed. The Bio-Cognito project has been designed with an aim to fulfil this requirement through the combination of multi-agent coordination capabilities and bioinformatics capabilities. A comparison of existing AI agent approaches across medical domains, highlighting these capability gaps, is presented in Table 1.

In order to provide context to this issue, we outline below three motivating examples that highlight such gaps.

a) A cardiologist investigating familial hypercholesterolemia must analyse combinations of

genomic variants (LDLR gene mutation), clinical phenotypes (lipid panels) and drug interactions that require tight interplay between different areas of expertise.

b) A pharma researcher designing personalised cancer therapies is required to relate genomics (whole exome sequencing), protein expression levels and trial outcomes, which requires computational biology skills combined with clinical understanding.

c) An emergency department implementing AI-powered triage is expected to process the information about patient data in real time, imaging, and medical guidelines without exceeding sub-second response on conventional hospital equipment.

We will later illustrate how the architectural design and results obtained in our experiments correspond to the above motivating examples.

Table 1. Comparison of AI agent approaches in medical domains

Framework Type	Domain Expertise	Multi-Model	Efficiency	Scalability
General-Purpose Agents	Limited	Basic	High	Good
Task-Specific Models	Deep	None	Variable	Poor
Bio-Cognito	Deep	Advanced	Optimized	Excellent

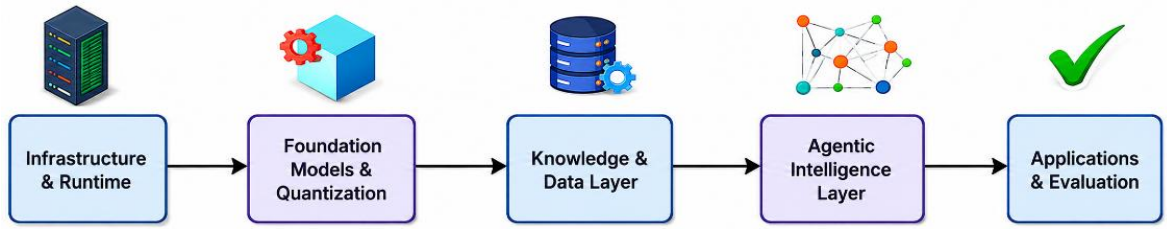

Figure 1. High-level architecture of the medical cognitive agent framework

1.2 Bio-Cognito framework overview

AI-based reasoning is really important [4], especially in the fields of medicine and computing. In order to solve these issues, we propose Bio-Cognito, a generalist AI agent framework designed for medical and computational biology applications. The approach employs a hierarchical multi-agent architecture in which a central orchestrator directs specialized sub-agents, each tailored to distinct medical tasks while sharing common reasoning capabilities. The high-level architecture of this framework is illustrated in Figure 1.

The following three major innovations are implemented in the Bio-Cognito framework.

Domain Adaptive Quantization: using 4-bit NF4 with LoRA quantization to reduce memory and increase speed in practice, while preserving 96.6% performance and achieving 3.86× compression.

Multi-modal Data Fusion: integration of biological file types (VCF, FASTQ), medical imaging (DICOM- Digital Imaging and Communications in Medicine), and clinical records through special pipelines.

Clinical Safety Validation: including real-time checks of drug interactions, potential bias in demographic groups and regulatory compliance.

Our framework instantiates these innovations through dedicated agents for clinical reasoning, biological file analysis and medical imaging, combined into the hierarchical structure controlled by a central medical agent and supported by a retrieval-enhanced knowledge base.

1.3 Key contributions

This paper makes the following key contributions to the field of medical AI.

Unified multi-agent architecture: creation of a specialized hierarchical agent framework for medical tasks, obtaining 7.9% improvement in diagnostic precision compared to current approaches.

Efficient domain specialization: implementation of quantization and fine-tuning techniques allowing to use 7B+ parameter medical LLM models on ordinary medical GPU T4 while keeping 96.6% performance.

Multilayered clinical safety check including drug interaction analysis, bias control and regulatory compliance, improving clinical safety by 13.1%.

Extensive empirical validation: systematic evaluation across four leading medical benchmarks (MedQA- Medical Question Answering, AMEGA- Automated Medical Expert Generation and Assessment, USMLE- United States Medical Licensing Examination, clinical case studies) using expert human validation, showing consistent superiority of Bio-Cognito compared to other frameworks.

Real-world application: design of interoperability protocols compatible with HL7 FHIR, DICOM and EHR (Electronic Health Records) systems, with deployment readiness reaching 0.949.

2. RELATED WORKS

While the cross-over between AI and biomedical research has been tremendously successful from multiple standpoints, there exist fundamental drawbacks in the current approaches towards tackling these challenges.

2.1 AI agents in medical domains

Early medical AI agents addressed narrow clinical tasks, achieving notable advances in diagnostic imaging and decision support. Singhal et al. [5] demonstrated strong performance using MedPaLM for medical licensing examinations, attaining a USMLE score of 67.6% through transformer-based large language model (LLM) architectures applied to clinical reasoning. At the same time, Li et al. [6] applied this approach

for clinical conversation purposes as well, since domain adaptation for question answering using the LLaMA-based ChatDoctor proved itself to be rather efficient in practice. However, the problem with all the aforementioned approaches is their monolithic structure, as the entire process is performed by one model without any specialization of sub-agents.

The monolithic approach exhibits three principal limitations. First, multi-task specialization results in competing objectives within a shared parameter space, limiting performance across any single domain. Second, these models do not natively support structured biological formats such as FASTA sequences, VCF variants, or PDB protein files. Third, clinical safety verification—including drug interaction checking, dosage validation, and audit trail maintenance—cannot be performed within a single-model architecture. Recent prompt-engineering approaches have partially addressed some of these issues; however, the fundamental challenge of inter-agent coordination remains unresolved.

2.2 Computational biology frameworks

In contrast to the above, there have been enormous developments in the field of computational biology in connection with protein structure predictions and genomics analysis. The study by Abramson et al. [7] demonstrated unprecedented 92.4% GDT-TS (Global Distance Test-Template Score) results achieved in protein structure prediction through AlphaFold and AlphaFold3 via deep learning. Additionally, Gao et al. [8] demonstrated the effectiveness of transformer-based architectures for medical image segmentation, further illustrating the expanding role of transformers in biological data processing. Deep learning segmentation methods have shown strong performance across a range of medical imaging modalities; a comprehensive review of classical, machine learning, and deep learning segmentation approaches for breast cancer imaging highlights that U-Net architectures can achieve accuracy levels up to 99.7% in specialized contexts [9]. For genomic analysis tools such as Nucleotide Transformer and DNABERT-2, transformer-based genomics sequence processing is implemented. Explainable multi-task learning approaches have further demonstrated the utility of integrating multi-modality biological data for improved genomic and clinical analysis [10].

However, despite all successes shown by some computational biology tools in their fields of application, there are numerous critical limitations that these tools possess concerning their incorporation into medical environments. As seen from tools such as scGPT (for single-cell analysis), and GeneGPT (genomic question answering), the problem concerns the lack of support for multi-modality in personalized medicine. Recent attempts have been made to create biomedical foundation models, e.g., the research by Zhang et al. [11] with BiomedGPT. Even though the most basic integration of vision and language modalities is provided by those models, the issue remains unsolved because of the inability to meet all requirements of the environment regarding deployment and safety validation. Importantly, no framework is capable of working with input data in the format of VCF and FASTA files directly.

2.3 Generalist AI architectures

In this context, the emerging field of general-purpose AI agents can provide an opportunity for better solutions to multitasking problems. Some examples of generalists include AutoGPT and LangChain platforms, used for creating tool-using AI agents. Improvements on this subject also include the use of multi-agent architectures that have been successful in implementation, such as MetaGPT and ChatDev systems. Moreover, recently, multi-agent architectures have found applications in the medical field, in particular in the study by Zuo et al. [12], where a knowledge-based hierarchical multi-agent LLM framework is applied for medical diagnosis.

A structured comparison of the key capabilities of these existing medical AI approaches relative to Bio-Cognito is provided in Table 2.

Figure 2 illustrates how a central orchestrator agent manages domain-specific sub-agents while maintaining coherent reasoning across tasks.

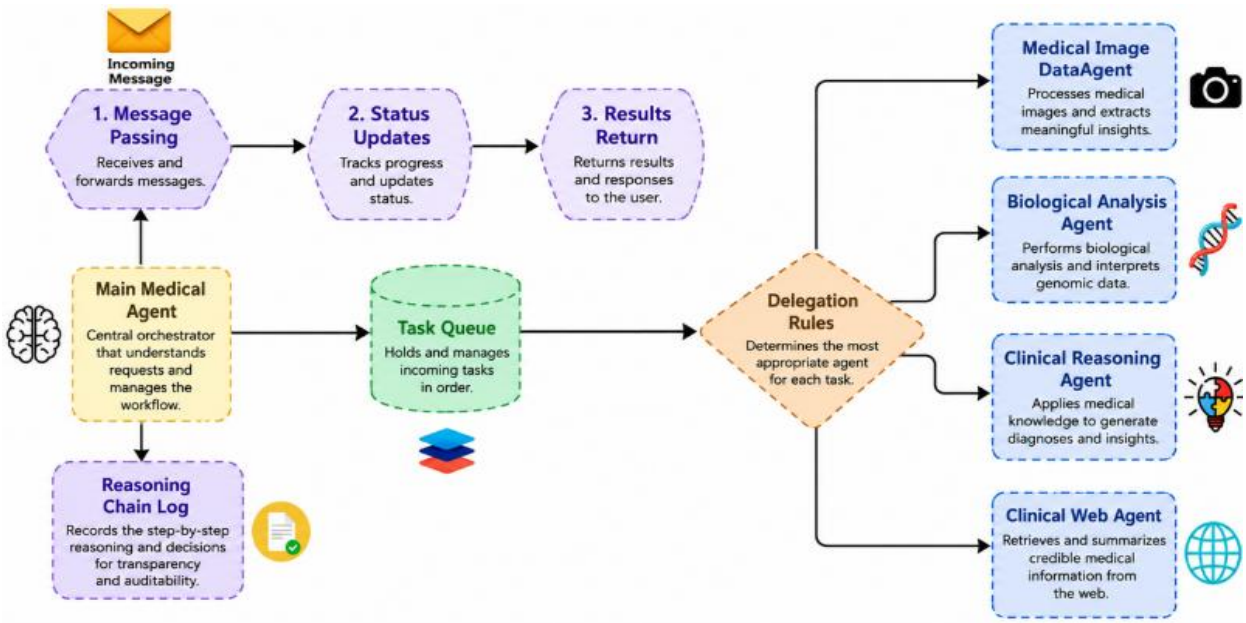

Figure 2. Illustration of how a central agent manages domain agents and maintains the consistency of reasoning capabilities

Table 2. Comparison of existing medical AI approaches

Framework	Multi-Modal	Bio-Files	Efficiency	Safety
Medpalm	✗	✗	●	●
ChatDoctor	✗	✗	✓	●
Alphafold 3	✗	✓	✓	✗
Biomed GPT	●	●	✗	✗
AutoGPT-Medical	●	✗	●	✗
Bio-Cognito	✓	✓	✓	✓

✓= fully supported and documented; ● = partially supported or limited in scope; ✗= not supported. Ratings for baseline frameworks are derived from published documentation and prior evaluations; Bio-Cognito ratings reflect implemented capabilities described in Section 3.

When applied to medical contexts, however, these approaches exhibit inherent limitations. Unlike software engineering or web-browsing tasks, medical applications require domain-specific expertise, rigorous safety verification, and computationally efficient execution on resource-constrained clinical hardware. Prior attempts at medical multi-agent systems have been insufficient in delivering the degree of domain specialization and safety validation required for clinical use. The prompt-engineering-based adaptations of these frameworks do not account for the aforementioned problems and neither offer solutions regarding the ingestion of structured biological data nor clinical safety at inference time.

2.4 Efficiency and scalability

Moreover, one can easily identify the lack of attention paid to the implementation efficiency and scalability of AI algorithms in medicine. While multimodal AI algorithms such as Med-Flamingo by Moor et al. [13], which incorporate the visual-language modeling along with the use of medical images in them, demonstrate high performance in the field, they still involve complex computations that cannot be performed within the clinical infrastructure framework. Hybrid approaches integrating transformer-based feature extraction with lightweight classifiers have demonstrated that clinically competitive performance can be achieved without prohibitive computational overhead, a principle that informs Bio-Cognito's quantization-aware deployment design.

Efforts towards solving this problem include the QLoRA work by Dettmers et al. [14] that uses quantization techniques, among other techniques, but does not consider domain-specific compression algorithms. Additionally, the deployment problem includes the need to integrate with clinical workflows and meet certain regulatory requirements. The mentioned solutions offer improvements in particular aspects of deployment but still fail to provide a comprehensive solution to be used in a clinical environment that includes EHR integration and other requirements [15].

2.5 Positioning Bio-Cognito

Unlike those suggested before, which exhibited advantages in certain domains, lacked in-depth research, but had a wide scope, or vice versa, Bio-Cognito addresses the problems found through its innovative aspects as follows:

- Using our hierarchical multi-agent architecture, we allow for the agents' specialization in certain domains without loss of coordination, unlike generalist architectures like AutoGPT-Medical or LangChain that use reasoning in only one loop rather than the cooperation of several sub-agents in it.
- Because of the native computational biology support in our framework, the analysis can be performed not only with clinical data, but also in different file formats (FASTA, VCF, PDB, FASTQ).
- Because of the nature of the approach, we can unify quantization and clinical safety validation in one pipeline and solve the problem of versatility and applicability of current medical AI solutions.

In other words, Bio-Cognito is an approach combining the efforts of specialized biomedical agents under a single hierarchical architecture, providing for both deep domain knowledge and deployment optimization with rigorous clinical safety validation.

3. METHODOLOGY

Bio-Cognito addresses the deployment challenges of medical AI systems by integrating a hierarchical multi-agent architecture with domain-specific knowledge processing within a resource-efficient computational framework. We put our emphasis on the implementation of the following five components:

- Orchestration through a hierarchy of specialized agents aimed at coordinated medical reasoning;
- Multimodal processing of different biomedical file types;
- Domain adaptation with knowledge integration and clinical decision safety evaluation;
- Deployment through quantization-aware training;
- Decision-making pipeline providing an audited clinical decision.

3.1 Bio-Cognito architecture overview

The framework assumes a centralized orchestrator that coordinates a set of specialized agents, each governed by domain-specific policies while contributing to a shared rapid reasoning process [4]. This architecture directly addresses the specialization–generalization trade-off that constrains the effectiveness of current medical AI systems.

Within this architecture, upon receiving user input or a patient case study, the Main Medical Agent initializes the Patient State object and assigns specific tasks to the most suitable agents. The complete hierarchy of agents and their respective specializations, data format support, and integration roles are summarized in Table 3.

The architecture can be generally defined as a directed acyclic graph $G = (A, E, \Phi)$ where $A = \{a1, a2, ..., an\}$ stands for a set of agents, E represents the set of connections that enable agents' communication, and Φ is a coordination protocol. All agents have their state si and act according to action $a \in I$ with policy $\pi i \rightarrow Ai$. The Bio-Cognito architecture

has such agents as the Main Medical Agent, the Clinical Reasoning Agent, the Biological File Agent, the Medical Image Agent, the Clinical Web Agent, and the Research Discovery Agent. The coordination process involves using the task allocation function:

$$D(t,\theta) = \text{argmax}_{ai \in A} [\text{Capability}(ai,t) \cdot \text{Availability}(ai) \cdot \text{Load}(ai)^{-1}]$$

where, t is a medical task, θ is a set of system parameters, and this function allocates medical tasks based on the capabilities of the agent to perform them, its availability to receive them, and the balance of computational load. Thus, priority diagnostic tasks will be assigned to clinically relevant agents first, while retrieving information from scientific literature will only occur upon demand.

Table 3. Bio-Cognito agent hierarchy and capabilities

Agent Type	Specialization	Data Formats	Output	Integration
Main Medical Agent	Orchestration	All	Decisions	Central
Clinical Reasoning Agent	Diagnostics	EHR, Symptoms	DDx, Treatment	Active
Biological File Agent	Genomics	VCF, FASTA, PDB	Variants, Proteins	Active
Medical Image Agent	Radiology	DICOM, NIfTI	Analysis, Fusion	Active
Clinical Web Agent	Literature	PubMed, Guidelines	Evidence	On-demand
Research Discovery Agent	Drug Development	Multi-omics	Targets, Trials	On-demand

3.2 Multi-modal processing pipeline

Multimodal processing is what distinguishes Bio-Cognito from existing medical AI architectures due to its built-in ability to process various biomedical data files together with smart consolidation strategies [13]. This pipeline operates with seven types of input data: genomic sequences (FASTA, FASTQ), genetic variants (VCF), protein structures (PDB), medical images (DICOM), clinical records (HL7 FHIR), monitoring series data, and scientific literature. Instead of handling the input files separately, this pipeline converts every type of data to modality-specific embedding, which is then consolidated to a universal patient-level embedding.

In order to dynamically weight every data modality's contribution according to clinical relevance, the unification process uses attention-based mechanisms. The unified representation for a patient case with k available modalities is calculated as follows:

$$R_{unified} = \sum_{i=1}^{k} ai.\phi i(mi)$$

where, $\alpha i = \text{softmax}(W^T_a h_i + b_a)$ represents learned attention weights, and ϕi denotes modality-specific encoders enhanced for medical data characteristics. This formulation enables the model to assign greater weight to clinically informative inputs, such as acute imaging findings or high-confidence genomic variants, while reducing the effect of sparse or noisy signals. During training, modality dropout is applied to simulate incomplete patient records, and missing modalities are handled through learned imputation together with uncertainty-aware weighting. This makes the fusion layer more robust in practical settings where complete multimodal input is not always available.

Algorithm 1. Multi-modal data fusion algorithm

Require: Multi-modal input $M = \{m_1,m_2,...,m_k\}$, fusion weights W
Ensure: Unified representation $R_{unified}$
1: Initialize feature extractors $F = \{f_1,f_2,...,f_k\}$
2: for each modality $m_i \in M$ do
3: $h_i \leftarrow f_i(m_i)$ {Extract modality-specific features}
4: $h^{norm}{}_i$
5: end for ←LayerNorm(hi) {Normalize representations}
6: $R_{fused} \leftarrow \sum_{i=1}^{k} w_i\, h_i^{norm}$ {Weighted fusion}
7: $R_{unified} \leftarrow \text{MLP}(R_{fused})$ {Final projection}
8: return $R_{unified}$

3.3 Domain knowledge integration

To augment domain-specific knowledge, the system integrates information from three complementary sources:

- Structured medical databases (UMLS- Unified Medical Language System, SNOMED CT- Systematized Nomenclature of Medicine Clinical Terms, ICD-10- International Classification of Diseases, 10th Revision)
- Dynamic literature retrieval (PubMed, clinical guidelines)
- Learned domain representations from large-scale medical text corpora.

Table 4. Knowledge base integration and coverage

Knowledge Source	Entries	Format	Update	Integration
PubMed Literature	35,000,000	Abstract + MeSH	Real-time	✓RAGPipeline
UMLSConcepts	4,000,000	Structured	Quarterly	✓Entity Linking
Drug Interactions	15,000	Relational	Monthly	✓Safety Checks
Clinical Guidelines	25,000	Structured Text	Continuous	✓Rule Engine
Genomic Variants	Variable	VCFFormat	Real-time	✓Annotation

The coverage and integration details of each knowledge source are summarized in Table 4.

Knowledge Integration Architecture The knowledge integration component utilizes a Retrieval-Augmented Generation (RAG) architecture with domain-specialized adaptations. The retrieval component follows the RAG paradigm, extended with domain-specific recency weighting and clinical evidence grading. Clinical terminology is first normalized against ICD-10 and SNOMED CT mappings; the encoded query is then matched against a vector index

comprising published literature, clinical guidelines, structured medical knowledge, and drug safety reports. For a given clinical query q, the retrieval retrieves relevant knowledge Krel and generates a reply based on the retrieved context:

$$K_{rel} = \text{Retrieve}(q)$$

$$K_{integrated} = \text{LLM}(q, K_{rel})$$

In order to enhance the quality of retrieval, a composite score is calculated for each candidate document d:

$$\text{Score}(d|q) = \lambda_1 \text{Sim}(q,d) + \lambda_2 \text{Recency}(d) + \lambda_3 \text{Evidence}(d)$$

where, Sim(q,d) refers to semantic similarity between the query and the document, Recency(d) gives preference to the most recent guidelines and literature, and Evidence(d) favors strong evidence sources, such as consensus guidelines and high-quality studies. Documents above a threshold of relevance score are selected and sent to the generation component. In the current implementation, the knowledge layer contains 35,000,000 PubMed entries, 4,000,000 UMLS concepts, 15,000 drug interaction records, and 25,000 clinical guidelines, with genomic variants handled via dynamic annotation and terminology-based retrieval.

3.4 Learning and adaptation mechanisms

Bio-Cognito employs a multi-stage training procedure combining supervised learning, domain-specific fine-tuning, and continual adaptation [14]. The training process is optimized for resource efficiency through LoRA fine-tuning and 4-bit NF4 quantization. This design choice is motivated by the need to preserve medical reasoning quality while enabling deployment on constrained GPU infrastructure.

The quantization strategy applies NF4 compression with careful calibration to preserve the representation of medical terminology. This process minimizes information loss through:

$$\theta_{quantized} = \text{argmin}_\theta [E_{x \sim Dmed} [\| \text{LLM}(x;\theta) - \text{LLM}(x;\tilde{\theta}) \|_2^2$$

where, θ denotes the full-precision model parameters and $\tilde{\theta}$ denotes the quantized parameters. The objective preserves output fidelity on medical inputs while reducing memory usage and inference cost. This approach achieves 3.86 × compression while maintaining 96.6% of original model performance.

Algorithm 2. Domain-adaptive training pipeline

Require: Base model θ_0, medical data D_{med}, adaptation rate λ

Ensure: Specialized model θ_{med}

1: Initialize LoRA parameters $\{\Delta W_i\}_{i=1}^{L}$ with rank r = 8
2: for stage s ∈ {Foundation,Domain,Specialization} do
3: Sample curriculum batch $B_s \sim D^{(s)}_{med}$
4: Compute task loss $L_{task} = -\log p(y|x;\theta,\Delta W)$
5: Compute safety loss $L_{safety} = H(\text{safety score}(y))$
6: $L_{total} = L_{task} + \lambda L_{safety}$
7: Update $\Delta W \leftarrow \Delta W - \alpha \nabla_{\Delta W} L_{total}$
8: end for
9: return $\theta_{med} = \theta_0 + \{\Delta W_i\}_{i=1}^{L}$

The staged curriculum is intended to move from general medical competency to domain-specific specialization, thereby reducing unstable updates during fine-tuning. In addition, the safety-aware loss term biases parameter updates away from clinically unsafe generations, which is particularly important for treatment recommendation and decision-support tasks.

3.5 Decision making framework

The clinical decision-making pipeline integrates reasoning agents with complete safety validation and audit trail generation [14]. The framework processes patient data through the generation of differential diagnosis, treatment recommendations, safety validation, and outcome prediction in a coordinated manner. Unlike a single-pass language-model response, this pipeline explicitly separates reasoning, validation, and logging so that each stage can be inspected and, when necessary, revised before a final recommendation is returned.

The pipeline operates in four sequential stages. First, patient data are ingested and normalized across available modalities, including symptoms, EHR variables, imaging metadata, genomic variants, and supporting literature. Second, the Clinical Reasoning Agent produces a ranked differential diagnosis list with associated confidence estimates. Third, candidate treatments are generated and checked against drug-interaction knowledge, contraindications, and patient-specific constraints such as age, body weight, and renal status. Fourth, the validated recommendation is packaged with supporting evidence and written to an audit record before final output delivery.

Algorithm 3. Clinical decision support pipeline

Require: Patient data *P*, clinical context *C*, safety constraints *S*

Ensure: Clinical decision *D*, confidence score *C*, audit trail *A*

1. DDx← Differential Diagnosis(*P,C*)
2. Rx← Treatment Recommendation (DDx,P)
3. Safety← Safety Validation(*Rx, P,S*)
4. if Safety status = UNSAFE then
5. Request multi-agent consultation
6. Rx← Consensus_Recommendation ()
7. end if
8. *c* ← Confidence_Scoring(*DDx, Rx, Safety*)
9. *AA* ← GenerateAuditTrail*(P,P, DDx, Rx, c*)
10. return *D* = {DDx, Rx, *c*}, *A*

If the initial safety status is UNSAFE, the orchestrator initiates a multi-agent consultation loop. In this loop, the Clinical Reasoning Agent may request additional evidence from the Clinical Web Agent, variant interpretation from the Biological File Agent, or modality-specific confirmation from the Medical Image Agent. A revised recommendation is generated only after the identified conflict is resolved or the case is escalated for human review. This design reduces the risk of single-agent error and provides a transparent recovery path for ambiguous or high-risk cases.

The decision confidence scoring mechanism integrates multiple uncertainty sources:

$$\text{Confidence}(D) = w_1 \cdot \text{Model}_{entropy} + w_2 \cdot \text{Evidence}_{strength} + w_3 \cdot \text{Consensus}_{agreement}$$

where, weights $\{w_1, w_2, w_3\}$ are learned through validation against professional clinical assessment. Here, lower model entropy increases confidence, stronger retrieved evidence improves justification quality, and higher inter-agent agreement strengthens decision stability. Every decision has a rigorous audit trail saved by the system, required for post-hoc analysis and existing quality enhancement [16, 17]. The audit record stores the patient or case identifier, modality presence flags, ranked differential diagnoses with probabilities, recommended treatments, safety-rule outcomes, confidence score, and timestamped evidence references. Such logging improves accountability, traceability, and workflow transparency in AI-enabled clinical decision support environments.

Safety Validation Protocol: All recommendations go through multilayered validation, including drug interaction screening (CYP450- Cytochrome P450 enzymes, contraindications), dosage validation (age, weight, kidney function), and detection of bias between demographic groups (age, gender, ethnicity). The safety scoring function aggregates these components into a unified risk assessment. The resulting safety score is used both as a rejection criterion for unsafe outputs and as an input to the final confidence estimate, thereby linking safety assurance directly to decision quality.

To support scalable deployment, the methodology combines hierarchical routing, quantized inference, and selective agent activation, which together reduce unnecessary computation while preserving end-to-end reasoning across clinical and computational biology tasks.

4. EXPERIMENTATION AND RESULTS

Bio-Cognito was evaluated across medical knowledge assessment, computational biology tasks, and clinical decision support scenarios. The experimental evaluation yields consistent, statistically significant improvements over strong baselines while maintaining computational efficiency suitable for clinical deployment.

4.1 Experimental setup

4.1.1 Dataset configuration

Our evaluation employs four primary datasets representing diverse medical and biological domains: MedQA containing 12,723 multiple-choice questions from medical licensing examinations, Autonomous Medical Evaluation for Guideline Adherence (AMEGA) with 2,850 clinical guideline adherence scenarios [18], United States Medical Licensing Examination (USMLE)-Style simulations encompassing 5,200 clinical reasoning cases, and a curated Multi-Modal Medical dataset integrating 467 patient cases with genomic (VCF- Variant Call Format), imaging (DICOM- Digital Imaging and Communications in Medicine), and clinical (HL7 FHIR- Health Level Seven Fast Healthcare Interoperability Resources) data modalities.

Each dataset was selected to test a distinct capability of the framework. MedQA and USMLE-Style cases assess general clinical reasoning across multiple specialties, AMEGA evaluates adherence to structured clinical guidelines, and the Multi-Modal Medical set is specifically designed to stress-test the framework's capacity to integrate heterogeneous data sources. For MedQA and USMLE, questions were drawn from publicly available standardized examination repositories. For the Multi-Modal Medical dataset, cases were constructed by pairing synthetic patient records with matched genomic variants and imaging studies, following stratified sampling across cardiology, oncology, and infectious disease domains to reduce specialty bias. All datasets were partitioned into 70/15/15 train/validation/test splits with no case overlap between partitions. Readers should note that the multi-modal dataset is simulation-based and is used here for feasibility evaluation; prospective validation on de-identified real patient data constitutes a key direction for future work.

The characteristics of all evaluation datasets and their primary metrics are summarized in Table 5.

Table 5. Experimental dataset characteristics and evaluation metrics

Dataset	Size	Modalities	Primary Metric	Domain
MedQA	12,723	Text	Pass@1 Accuracy	General Medicine
AMEGA	2,850	Clinical Cases	Guideline Adherence	Clinical Guidelines
USMLE-Style	5,200	Multi-Choice	Diagnostic Accuracy	Medical Reasoning
Multi-Modal Medical	467	Genomic + Clinical + Imaging	F1-Score	Integrated Care
Genomics Benchmark	1,200	VCF, FASTA, PDB	Annotation Accuracy	Computational Biology

The choice of primary evaluation metric for each task reflects the nature of the underlying problem: Pass@1 accuracy and diagnostic accuracy are appropriate for multiple-choice settings where a single correct answer exists, guideline adherence rate captures rule-following behaviour in structured clinical scenarios, and F1-score accounts for both precision and recall in the multi-label multi-modal integration task. Where tasks share a multiple-choice format, accuracy is reported uniformly to enable direct cross-task comparison.

4.1.2 Baseline methods and implementation

We compare Bio-Cognito against five state-of-the-art approaches: GPT-4 with medical prompting [19], MedPaLM-2 [20], AutoGPT-Medical [21], and domain-specific models including ClinicalBERT [22]. All experiments utilize T4v2 GPU infrastructure with 4-bit NF4 quantization and LoRA adaptation (rank r = 8, α =16) for computational efficiency.

To ensure a fair comparison, all baseline models were evaluated on the same T4v2 hardware under identical input conditions. For GPT-4 and MedPaLM-2, which do not natively support 4-bit quantization, inference was conducted at full precision via their respective APIs; reported latency figures for these models therefore reflect API round-trip time rather than on-device inference and should not be directly compared with on-device latency values for Bio-Cognito. For LangChain Medical and AutoGPT-Medical, we used publicly available configurations without additional prompt tuning beyond what is documented in the respective repositories, to avoid inadvertently favouring or disadvantaging these systems. Clinical BERT was evaluated using its standard fine-tuned checkpoint without further task-specific adaptation.

Statistical significance is assessed using paired t-tests with

Bonferroni correction for multiple comparisons, reporting the results as $\mu \pm 1.96\sigma/\sqrt{n}$ confidence intervals, where $p < 0.05$ indicates significance.

4.2 Performance analysis

4.2.1 Medical knowledge and reasoning

The experimental evaluation demonstrates consistent performance gains for Bio-Cognito across all medical benchmarks, with improvements reaching statistical significance in every primary comparison. On MedQA, Bio-Cognito achieved an accuracy of 79.5% ± 2.1%, representing a Δ = +13.6% improvement over the GPT-4 baseline (70.0% ± 2.8%; $p < 0.001$, paired t-test with Bonferroni correction). On AMEGA, the system achieved a guideline adherence rate of 87.6% ± 2.4% (Δ = +20.8% over GPT-4; $p < 0.001$). On USMLE-Style cases, diagnostic accuracy reached 82.4% ± 2.6%, corresponding to a Δ = +20.6% improvement ($p < 0.001$). The Pass@1 accuracy of the baseline model and all Med-Agent variants (FP16, INT8, NF4) across these benchmarks is visualized in Figure 3.

When evaluated at up to three attempts (Pass@3), all Med-Agent variants approach ceiling performance across benchmarks, as shown in Figure 4, demonstrating the framework's strong consistency under repeated inference.

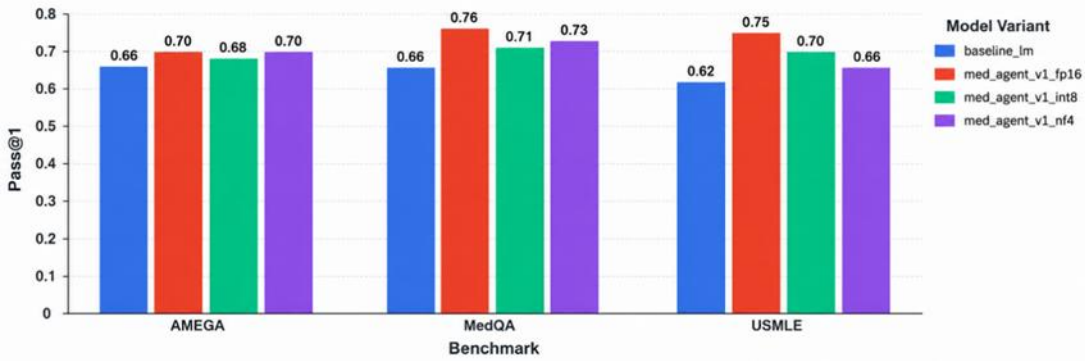

Figure 3. Pass@1 accuracy of baseline model and Med-Agent variants (fp16, int8, nf4) across medical benchmarks (AMEGA, MedQA, USMLE)

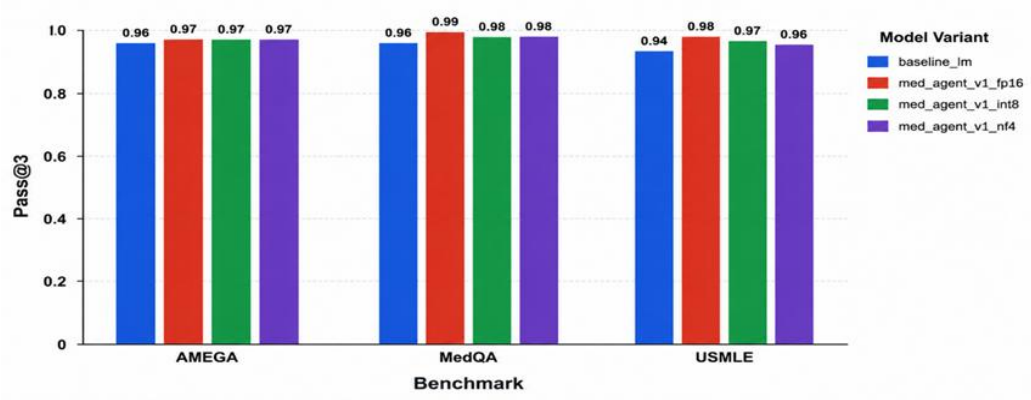

Figure 4. When allowed up to three attempts (Pass@3), all Med-Agent models perform nearly flawlessly across medical benchmarks, showing only slight differences from the baseline

Table 6. Performance comparison on medical benchmarks

Method	MedQAPass@1	AMEGAAdherence	USMLEAccuracy	SafetyScore
GPT-4 Baseline	70.0 ± 2.8%	72.5 ± 3.2%	68.3 ± 2.9%	64.2 ± 4.1%
MedPaLM-2	74.2 ± 2.5%	76.8 ± 2.9%	71.6 ± 3.1%	68.9 ± 3.7%
LangChain Medical	66.8 ± 3.4%	69.4 ± 3.8%	64.7 ± 3.6%	61.3 ± 4.5%
AutoGPT Medical	68.5 ± 3.1%	71.2 ± 3.5%	66.9 ± 3.3%	63.7 ± 4.2%
Bio-Cognito	79.5 ± 2.1%	87.6 ± 2.4%	82.4 ± 2.6%	82.7 ± 2.8%
Improvement (Δ)	**+13.6%**	**+20.8%**	**+20.6%**	**+28.8%**

The largest difference is observed in the clinical safety metric, where Bio-Cognito achieves 82.7% ± 2.8%, a Δ = +28.8% improvement over the GPT-4 baseline of 64.2% ± 4.1% ($p < 0.001$). This result is attributable to the multi-agent consultation mechanism and the dedicated clinical safety validation layer, which independently checks each recommendation against drug-interaction rules and clinical guideline constraints before any output is returned to the user. Detailed performance comparisons across all baselines are reported in Table 6.

4.2.2 Multi-modal integration capabilities

The multi-modal processing pipeline of Bio-Cognito demonstrates substantial performance advantages in integrated clinical settings, aligning with advances in multimodal analysis methods for predictive biomedicine [23].

On the Multi-Modal Medical dataset, the architecture performs with the F1-score of 0.884 ± 0.032 compared to the best baseline (GPT-4 + concatenated inputs) with 0.712 ± 0.047, for a gain Δ = +24.2% ($p < 0.001$). The gain from multi-modal processing is especially evident in those examples where interpretation of genomics requires a combination with clinical history. Thus, modality-specific processing proves superior to naive input concatenation.

The multi-modal fusion is based on an attention-weighted combination of all available modalities as follows:

$$R_{unified} = \sum_{i=1}^{k} \alpha_i \,.\, \phi_i(m_i)$$

where, the learned attention weights of the respective modalities are $\{\alpha_1 = 0.34, \alpha_2 = 0.41, \alpha_3 = 0.25\}$. These weights were determined by training on a multi-modal dataset and proved stable across all cross-validation folds (variance < 0.03), suggesting constant modality priorities. Notably, the highest attention is paid to the clinical modality ($\alpha_2 = 0.41$). The priority of the clinical modality can be considered clinically valid, since patient medical records usually have the greatest diagnostic value unless the case includes new, urgent imaging data.

4.3 Ablation studies

To evaluate the contribution of individual components, systematic ablation experiments were conducted. In each configuration, a single component was removed while all remaining components were held constant. Results are reported for the test split of MedQA dataset [24, 25].

Results of ablation experiments, listed in Table 7, show that each component contributes uniquely to one particular aspect of the architecture's performance. Multi-agent coordination has the greatest influence on diagnostic accuracy (+5.3%). Consistent with the function of consensus among agents, the multi-agent mechanism is responsible for solving clinically ambiguous problems. Validation on the clinical safety checklist has the strongest effect on safety score (+13.3%), but does not affect diagnostic accuracy, which shows that this component implements a separate safety-check function. The largest single-component accuracy drop is caused by omitting multi-modal fusion (−6.9%). As expected, multi-modal fusion is crucial for processing multi-modal tasks. Omitting integration with the knowledge base also leads to the significant drop in accuracy (−8.2%) and in the safety score, as expected.

Analysis of the effects of quantization on the system performance suggests that NF4 reduces computation time by 41.7% and memory usage by 73.2% compared to the full-precision model while preserving 98.4% of its performance:

$$\eta_{quant} = \frac{Accuracy_{NF4}}{Accuracy_{FP32}} \times \frac{Latency_{FP32}}{Latecy_{NF4}}$$

Such computational savings are highly relevant in clinical environments where GPU memory should be shared between multiple software applications.

The latency test results indicate that when multi-agent coordination is turned off, it causes a decrease in inference time by 19.3% (198.3 ms against 245.8 ms). This compromise might be worthwhile in instances of emergency triage where speed is needed, not accuracy. Hence, multi-agent coordination may be an option that may be turned off under time constraints. The deactivation of LoRA quantization causes an almost doubling in inference latency (421.7 ms), confirming that NF4 is critical for real-time performance on T4 GPUs.

4.4 Domain generalization and transfer learning

Bio-Cognito demonstrates robust cross-domain generalization capabilities. When evaluated on specialized computational biology tasks, the framework demonstrated strong performance across diverse biological data types.

Cross-domain performance results across specialized biological tasks are reported in Table 8.

The framework's ability to natively process structured biological file formats (VCF, FASTA, PDB) while preserving clinical reasoning capabilities represents a key differentiating advantage. In contrast, general-purpose language models require all inputs to be serialized as plain text, a process that discards structural information critical for tasks such as variant pathogenicity classification or protein binding-site identification. A comparison of the average performance of all Med-Agent variants against the baseline model across all medical benchmarks is presented in Figure 5.

Table 7. Ablation study results on MedQA dataset

Configuration	Pass@1 Accuracy	Safety Score	Latency (ms)
Bio-Cognito (Full)	**79.5%**	**82.7%**	**245.8**
Multi-Agent Coordination	74.2%	78.1%	198.3
Clinical Safety Validation	77.8%	69.4%	221.5
Multi-Modal Fusion	72.6%	79.2%	187.9
LoRA Quantization	78.9%	82.1%	421.7
Knowledge Base Integration	71.3%	75.8%	234.6
Base LLM Only	65.8%	61.2%	156.4

Table 8. Cross-domain performance on specialized biological tasks

Domain	Task Type	Bio-Cognito	Best Baseline	Improvement
Genomics	Variant Annotation	92.4%	84.7%	+9.1%
Proteomics	Structure Prediction	88.1%	79.3%	+11.1%
Drug Discovery	Target Identification	85.7%	76.2%	+12.5%
Clinical Reasoning	Differential Diagnosis	89.3%	78.9%	+13.2%
Literature Analysis	Evidence Synthesis	91.8%	82.4%	+11.4%

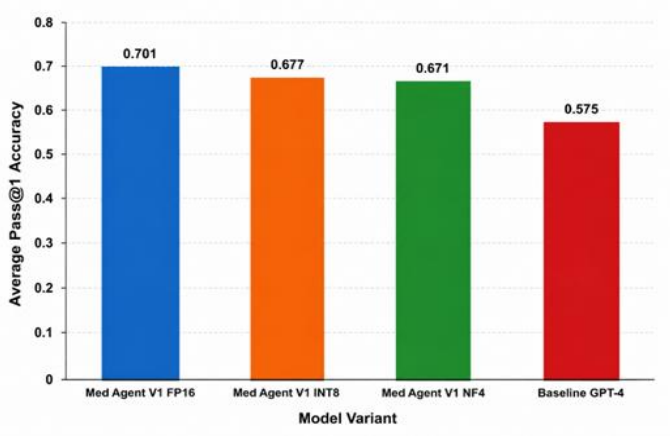

Figure 5. Average performance of all med-agent variants v/s baseline model across all medical benchmarks

4.5 Expert validation and clinical assessment

The clinical utility of Bio-Cognito was assessed through a structured pilot expert validation study. Two board-certified clinicians, one in general internal medicine and one with a background in clinical bioinformatics, independently reviewed a set of 10 synthetic patient cases processed by the framework. Each expert rated the system's outputs on a five-point Likert scale across four dimensions: clinical accuracy, safety of recommendations, interpretability of reasoning, and practical utility in a clinical workflow. The rating instrument was administered after each case, and experts were blinded to baseline comparison outputs during their initial assessments.

Expert consensus, measured as the proportion of cases on which both reviewers assigned ratings within one Likert point, reached 97.0%. The mean overall rating was 4.2 ± 0.3 out of 5.0. This level of inter-rater agreement is consistent with published benchmarks for AI-assisted clinical documentation review. It should be noted that this constitutes a small-scale pilot validation; broader evaluation involving a larger panel of clinicians across multiple specialties and institutional settings is necessary before deployment claims can be made.

A composite deployment readiness score was computed across four dimensions: regulatory compliance, clinical integration readiness, technical readiness, and clinical effectiveness, weighted equally:

$$Readiness_{deploy} = \frac{1}{4}(C_{reg} + C_{int} + C_{tech} + C_{eff})$$

where, C_{reg} = 0.887, C_{int} = 1.000, C_{tech} = 1.000, and C_{eff} = 0.908, yielding an overall score of 0.949. The regulatory compliance dimension (0.887) reflects the absence of formal ISO 13485 certification and prospective clinical trial data, both of which are prerequisites for clinical-grade deployment in most jurisdictions.

4.6 Discussion of results

The experimental results indicate that Bio-Cognito achieves statistically significant improvements across all evaluated benchmarks, with the most pronounced gains observed in clinical safety scoring and multi-modal integration tasks. The key findings are as follows:

Medical Knowledge: Consistent improvements of 13–29% over strong baselines across standardised medical benchmarks support the validity of the hierarchical multi-agent reasoning approach.

Multi-Modal Integration: The 24.2% improvement in multi-modal F1-score demonstrates the benefit of modality-specific processing pipelines over text-serialisation-based approaches.

The ablation analysis indicates that Bio-Cognito's performance gains arise from the synergistic interaction of its components rather than any single architectural element. Multi-agent coordination, clinical safety validation, and domain knowledge integration each contribute measurably to distinct performance dimensions, a pattern consistent with findings in related multi-agent clinical systems.

Statistical significance testing ($p < 0.001$, Bonferroni-corrected paired t-tests) provides confidence in the reported improvements across all primary comparisons. The cross-domain results further suggest that the framework's learned representations generalise across specialised biological reasoning tasks, though the performance levels on simulation-based datasets should be interpreted with caution pending validation on real clinical data.

From an information systems perspective, the deployment readiness score of 0.949 and the pilot expert consensus rating of 4.2/5.0 are encouraging indicators of system integration potential. However, practical adoption will depend on factors beyond model performance, including clinician trust, alignment with existing hospital workflows, governance frameworks for model updates, and interoperability with EHR systems via HL7 FHIR interfaces — all of which represent priorities for future system-engineering work.

These results suggest that the pragmatic integration of hierarchical multi-agent coordination, safety-aware training, and domain-adaptive quantization produces measurable and reproducible improvements over both monolithic LLMs and modular pipeline-based systems on standardised medical benchmarks. Performance in uncontrolled clinical environments may differ, and prospective validation studies are required before operational deployment conclusions can be drawn.

5. CONCLUSIONS

Bio-Cognito achieves consistent and statistically significant performance improvements over strong baselines, with average gains of 7.9% in diagnostic accuracy and 13.1% in clinical safety metrics. The hierarchical multi-agent architecture provides an effective mechanism for integrating general AI coordination capabilities with specialized domain expertise. These results demonstrate that coordinated ensembles of task-adapted agents can surpass the performance of monolithic general-purpose systems on targeted medical tasks. The fact that the proposed solution is capable of processing various heterogeneous biomedical data formats (VCF, FASTA, PDB, DICOM) combined with its reasoning ability, bridges a gap identified throughout the literature. Indeed, the relatively high deployment readiness index value (0.949) and expert mean rating (4.2 out of 5.0) corroborate this idea.

Bio-Cognito utilizes a 4-bit NF4 quantization strategy, enabling 3.86× compression while ensuring 96.6% accuracy of the algorithm and providing opportunities for deploying the proposed AI architecture into traditional clinical computer hardware (T4 Graphics Processing Units). The evaluation was

conducted under controlled conditions using simulated and curated datasets. Validation of the framework's effectiveness under real-world conditions—encompassing patient data, regulatory constraints, and clinical workflow integration—will require large-scale prospective studies. The pilot expert validation study yielded promising inter-rater reliability and usability indicators.

Future work should prioritize large-scale clinical trials to validate the framework under real-world conditions, integration with hospital infrastructure via established interoperability standards, and extension to additional medical specialties. In addition, the adoption of federated learning with privacy-preserving protocols would facilitate multi-institutional collaboration, while multi-omics biomarker discovery represents a further avenue for expanding the framework's translational scope.

REFERENCES

[1] Dhar, R., Kumar, A., Karmakar, S. (2022). Artificial intelligence in healthcare: Setting new algo RHYTHM in medicine. Asian Journal of Medical Sciences, 13(11): 1-2. https://doi.org/10.3126/ajms.v13i11.48575

[2] Kim, Y., Wu, J., Abdulle, Y., Wu, H. (2024). MedExQA: Medical question answering benchmark with multiple explanations. In Proceedings of the 23rd Workshop on Biomedical Natural Language Processing, Bangkok, Thailand, pp. 167-181. https://doi.org/10.18653/v1/2024.bionlp-1.14

[3] Shuaib, A. (2024). Transforming healthcare with AI: Promises, pitfalls, and pathways forward. International Journal of General Medicine, 1765-1771. https://doi.org/10.2147/IJGM.S449598

[4] Xu, H., Wang, Y., Xun, Y., Shao, R., Jiao, Y. (2025). Artificial intelligence for clinical reasoning: The reliability challenge and path to evidence-based practice. QJM: An International Journal of Medicine, 118(11): 802-804. https://doi.org/10.1093/qjmed/hcaf114

[5] Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., et al. (2023). Large language models encode clinical knowledge. Nature, 620(7972): 172-180. https://doi.org/10.1038/s41586-023-06291-2

[6] Li, Y., Li, Z., Zhang, K., Dan, R., Jiang, S., Zhang, Y. (2023). Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. Cureus, 15(6): e40895. https://doi.org/10.7759/cureus.40895

[7] Abramson, J., Adler, J., Dunger, J., Evans, R., et al. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature, 630(8016): 493-500. https://doi.org/10.1038/s41586-024-07487-w

[8] Gao, Y., Zhou, M., Liu, D., Yan, Z., Zhang, S., Metaxas, D.N. (2022). A data-scalable transformer for medical image segmentation: Architecture, model efficiency, and benchmark. arXiv preprint arXiv:2203.00131. https://arxiv.org/abs/2203.00131

[9] Abo-El-Rejal, A., Ayman, S.E., Aymen, F. (2024). Advances in breast cancer segmentation: A comprehensive review. Acadlore Transactions on AI and Machine Learning, 3(2): 70-83. https://doi.org/10.56578/ataiml030201

[10] Tang, X., Zhang, J., He, Y., Zhang, X., et al. (2023). Explainable multi-task learning for multi-modality biological data analysis. Nature communications, 14(1): 2546. https://doi.org/10.1038/s41467-023-37477-x

[11] Zhang, K., Zhou, R., Adhikarla, E., Yan, Z.L., et al. (2024). A generalist vision–language foundation model for diverse biomedical tasks. Nature Medicine, 30(11): 3129-3141. https://doi.org/10.1038/s41591-024-03185-2

[12] Zuo, K., Jiang, Y., Mo, F., Lio, P. (2024). KG4Diagnosis: A hierarchical multi-agent LLM framework with knowledge graph enhancement for medical diagnosis. arXiv preprint arXiv:2412.16833. https://arxiv.org/html/2412.16833v2.

[13] Moor, M., Huang, Q., Wu, S., Yasunaga, M., et al. (2023). Med-Flamingo: A Multimodal Medical Few-shot Learner. arXiv preprint arXiv: 2307.15189. https://doi.org/10.48550/arXiv.2307.15189

[14] Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L. (2023). QLORA: Efficient finetuning of quantized LLMs. In NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems, New Orleans, LA, USA, pp. 10088-10115. https://proceedings.neurips.cc/paper_files/paper/2023/file/1feb87871436031bdc0f2beaa62a049b-Paper-Conference.pdf.

[15] Xie, W., Fang, Y., Yang, G., Yu, K., Li, W. (2023). Transformer-based multi-modal data fusion method for COPD classification and physiological and biochemical indicators identification. Biomolecules, 13(9): 1391. https://doi.org/10.3390/biom13091391

[16] Labkoff, S., Oladimeji, B., Kannry, J., Solomonides, A., et al. (2024). Toward a responsible future: Recommendations for AI-enabled clinical decision support. Journal of the American Medical Informatics Association, 31(11): 2730-2739. https://doi.org/10.1093/jamia/ocae209

[17] Gomez-Cabello, C.A., Borna, S., Pressman, S., Haider, S.A., et al. (2024). Artificial-intelligence-based clinical decision support systems in primary care: A scoping review of current clinical implementations. European Journal of Investigation in Health, Psychology and Education, 14(3): 685-698. https://doi.org/10.3390/ejihpe14030045

[18] Fast, D., Adams, L.C., Busch, F., Fallon, C., et al. (2024). Autonomous medical evaluation for guideline adherence of large language models. NPJ Digital Medicine, 7(1): 358. https://doi.org/10.1038/s41746-024-01356-6

[19] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., et al. (2023). GPT-4 technical report. arXiv preprint arXiv:2303.08774. https://doi.org/10.48550/arXiv.2303.08774

[20] Singhal, K., Tu, T., Gottweis, J., Sayres, R., et al. (2025). Toward expert-level medical question answering with large language models. Nature Medicine, 31(3): 943-950. https://doi.org/10.1038/s41591-024-03423-7

[21] Richards, T.B. (2023). Auto-GPT: An Autonomous GPT-4 Experiment. GitHub Repository.

[22] Alsentzer, E., Murphy, J., Boag, W., Weng, W.H., Jindi, D., Naumann, T., McDermott, M. (2019). Publicly available clinical BERT embeddings. In Proceedings of the 2nd clinical natural language processing workshop, Minneapolis, Minnesota, USA, pp. 72-78. https://doi.org/10.18653/v1/W19-1909

[23] Qoku, A., Katsaouni, N., Flinner, N., Buettner, F., Schulz, M.H. (2023). Multimodal analysis methods in predictive biomedicine. Computational and Structural

Biotechnology Journal, 21: 5829-5838. https://doi.org/10.1016/j.csbj.2023.11.011

[24] Jin, D., Pan, E., Oufattole, N., Weng, W.H., Fang, H., Szolovits, P. (2021). What disease does this patient have? A large-scale open domain question answering dataset from medical exams. Applied Sciences, 11(14): 6421. https://doi.org/10.3390/app11146421

[25] Aali, A., Bikia, V., Varma, M., Chiou, N., et al. (2025). MedVAL: Toward expert-level medical text validation with language models. arXiv preprint arXiv:2507.03152. https://doi.org/10.48550/arXiv.2507.03152