MedGPT: An Agentic Medical AI Assistant with Hybrid Retrieval, Chain-of-Thought Reasoning, and NLI-Based Citation Mapping for Medical Education

Mukesh Gilda[1], Gollena Adith[2], Nakka Jashwanth[3], Mahesh Bhati[4], Raj Goel[5]

[1]*Assistant Professor, Department of Computer Science Engineering – Cyber Security, Sphoorthy Engineering College.*

[2,3,4,5] *Student, Department of Computer Science Engineering – Cyber Security, Sphoorthy Engineering College.*

***Abstract*—MedGPT is an advanced agentic medical AI assistant designed to provide evidence-based clinical answers with transparent reasoning and verifiable citations. The system integrates a multi-stage retrieval-augmented generation (RAG) pipeline combining query expansion, hybrid retrieval (dense vector search + BM25 keyword matching + real-time PubMed), and cross-encoder reranking. A novel Natural Language Inference (NLI) citation mapper automatically links each answer sentence to supporting source documents and attaches sentence-level confidence scores. The knowledge base is grounded in *Harrison's Principles of Internal Medicine* (approximately 18,000 indexed chunks) and augmented with live PubMed research. The reasoning backbone employs Llama 3.1 70B Instruct to produce structured six-step chain-of-thought (CoT) traces. Evaluation on a USMLE-style benchmark yields 70% accuracy; retrieval precision reaches 92%; citation accuracy is 85%. A controlled user study with 30 medical students records a statistically significant 16% improvement in post-test scores ($p < 0.05$) and an 82.3/100 System Usability Scale rating. The system promotes higher-order thinking skills (HOTS) aligned with Bloom's Taxonomy Levels 4–6 and operationalises heutagogical self-directed learning. MedGPT addresses critical gaps in existing medical AI tools by providing transparent, multi-source, evidence-grounded assistance to medical students and healthcare professionals.**

***Index Terms*—Retrieval-Augmented Generation; Chain-of-Thought Reasoning; Natural Language Inference; Medical Education; USMLE Benchmark; Higher-Order Thinking Skills; Heutagogy; Large Language Models.**

I. INTRODUCTION

Medical education is undergoing a profound transformation driven by artificial intelligence. Clinicians and students alike increasingly turn to AI-powered tools for rapid evidence retrieval, clinical decision support, and self-directed study. How- ever, the majority of general-purpose large language models (LLMs) deployed in medical contexts share a common cluster of critical limitations. They generate confident-sounding an- swers without transparent reasoning, they hallucinate citations that do not exist, and they draw on a single, static knowledge source rather than the rich, evolving corpus of biomedical literature [2].

These shortcomings carry real educational consequences. In a clinical learning environment, an answer without a verifiable source is educationally hazardous. A system that cannot show *why* it reached a conclusion cannot model the evidence- based reasoning that medicine demands. Existing tools such as Med-PaLM 2 [2] achieve impressive benchmark scores but remain closed-source, offer no reasoning transparency, and provide no mechanism for learners to trace an answer back to a primary document. BioGPT [10] focuses on biomedical text generation but lacks retrieval grounding, making hallucination a persistent risk. General-purpose RAG frameworks typically apply single-source vector retrieval, missing complementary signals offered by keyword search and live research databases. Beyond accuracy, there is a deeper pedagogical problem. Medical AI systems are increasingly used not merely as search engines but as

learning partners. When a student cannot see the reasoning behind an answer, they cannot develop the metacognitive habits that evidence-based medicine requires. When citations are invisible or unverifiable, source-evaluation skills atrophy. Current tools do not address this dual technical-
pedagogical challenge.

MedGPT is designed to address all of these gaps simultaneously. Its core contributions are: **(1)** a hybrid retrieval pipeline fusing dense vector search, BM25 keyword matching, and real-time PubMed queries; **(2)** cross- encoder reranking selecting the top five most relevant documents; **(3)** structured CoT generation exposing a six-step reasoning trace in every answer; **(4)** an NLI-based citation mapper linking individual answer sentences to source documents with sentence-level confidence scores; and **(5)** an interactive PDF viewer for *Harrison's Principles of Internal Medicine* with automatic text highlighting on cited pages. Together, these components form an agentic pipeline grounded in Bloom's Taxonomy and Hase & Kenyon's heutagogy model.

The remainder of this paper is organised as follows. Section II surveys related work. Section III describes the system architecture in detail. Section IV presents the educational framework. Section V covers implementation. Section VI reports evaluation results. Section VII discusses implications and limitations. Section VIII outlines future directions, and Section IX concludes.

II. RELATED WORK

A. *Medical AI Assistants*

The application of LLMs to clinical settings has grown rapidly over the past three years. GPT-4 demonstrates broad clinical knowledge but suffers from citation hallucination and pro- vides no domain-specific retrieval mechanism [1]. Med-PaLM 2 achieves 86.5% on the MedQA benchmark yet remains closed-source and provides no reasoning transparency [2]. BioGPT [10] is pre-trained on biomedical abstracts without retrieval grounding, making factual accuracy unpredictable for nuanced or rare clinical questions. Critically, none of these systems integrates an interactive document viewer or a pedagogical framework actively promoting higher-order thinking.

B. *Retrieval-Augmented Generation*

Lewis et al. [3] introduced Retrieval-Augmented Generation (RAG), showing that grounding LLM outputs in retrieved documents significantly reduces hallucination on knowledge- intensive tasks. Standard RAG implementations use a single dense vector store that captures semantic similarity well but misses the exact-match signals that are critical for medical terminology, drug names, and disease eponyms. HyDE [8] generates a hypothetical answer prior to retrieval, improving recall at substantial computational cost. MedGPT takes a more efficient path: query expansion into three clinical reformulations followed by weighted fusion of vector, BM25, and PubMed signals.

C. *Chain-of-Thought Reasoning*

Wei et al. [1] demonstrated that prompting LLMs to produce intermediate reasoning steps substantially improves accuracy on complex tasks. Standard CoT prompting, however, yields unstructured free-text traces that are difficult to parse program- matically or display cleanly to learners. MedGPT enforces structured <thinking> and <answer> XML tags in every prompt, enabling reliable extraction and display of a clean, six-step reasoning trace in the user interface.

D. *Citation Verification and Fact-Checking*

WebGPT [4] introduced web-search-grounded generation but draws on general web sources that lack the authority and specificity required in a medical context. RARR [9] performs iterative fact-checking through multiple LLM inference calls, incurring high latency. MedGPT employs a lightweight, single-pass NLI citation mapper based on word-overlap and contextual matching, achieving 85% citation precision at sub- millisecond latency per answer sentence.

E. *Educational AI and Pedagogical Frameworks*

Several studies have examined AI-assisted learning in medical education, finding improvements in knowledge retention and clinical reasoning when AI tools provide expla- nations alongside answers [6]. Hase and Kenyon's heutagogy model [5] extends traditional pedagogy by emphasising learner-determined learning paths and double-loop metacognitive reflection. Sweller's Cognitive Load Theory [7] provides a framework for designing AI interfaces that minimise extraneous cognitive demand. MedGPT is, to the best of our knowledge, the first medical AI system to systematically integrate all three theories into its interface and pipeline design.

F. Summary of Research Gaps

Across the literature surveyed, four persistent gaps remain:

(i) no integrated pedagogical framework promoting higher- order thinking in medical AI; (ii) limited reasoning transparency in answer generation; (iii) reliance on a single knowl- edge source rather than a fused multi-source approach; and

(iv) absence of interactive document exploration directly linked to AI-generated citations. MedGPT is explicitly designed to close all four.

III. SYSTEM ARCHITECTURE

\

A. Architecture Overview

MedGPT follows a modular, eight-stage pipeline. Each stage is independently replaceable, allowing component up- grades without redesigning the overall system. The pipeline is orchestrated by a backend agent that sequences retrieval, reranking, generation, and citation mapping in a single pass per user query. The stages are: (1) Query Expansion, (2) Hybrid Retrieval, (3) Deduplication, (4) Cross-Encoder Reranking, (5) CoT Answer Generation, (6) NLI Citation Mapping,

(7) Hallucination Detection, and (8) Multi-Panel UI Rendering.

B. Core Components

LLM Handler**.** The primary language model is Llama 3.1 70B Instruct, accessed via the OpenRouter API with a 128K- token context window. It is responsible for query expansion (generating three clinical reformulations), structured CoT an- swer synthesis, and fallback context extraction when the API is unavailable. The model is configured with temperature 0.2, top-*p* 0.9, and a maximum of 2,048 output tokens.

Vector Store Manager. Documents from *Harrison's Principles of Internal Medicine* (20th edition, ≈4,500 pages) are processed using PyMuPDF and chunked into 800-token seg- ments with a 200-token overlap, producing approximately 18,000 chunks. Each chunk is embedded using BAAI/bge- base-en-v1.5 (768-dimensional dense vectors) and stored in a FAISS IndexFlatL2 index for efficient L_2-similarity search.

PubMed Retriever**.** Live biomedical literature is retrieved

via NCBI E-utilities (ESearch + EFetch), rate-limited to 3 requests per second with in-memory query caching. Each returned record includes title, abstract, journal name, publication year, PMID, and a direct URL. By default, queries are filtered to publications from the preceding 12 months to ensure currency.

Hybrid Retriever. Vector search results, BM25 keyword matches, and PubMed records are fused using weighted scoring: vector weight 1.0, BM25 weight 1.0, PubMed weight

1.2. The higher PubMed weight reflects the greater clinical currency of peer-reviewed literature relative to textbook con- tent. Content-fingerprint deduplication (comparing the first 200 characters of each document) removes duplicate passages before the reranking stage.

Cross-EncoderReranker. The ms-marco-MiniLM-L-6-v2 cross-encoder model

(22M parameters) computes query-document relevance scores on a 0–1 scale. All fused candidates are scored; the top five documents are retained as the generation context.

NLI Citation Mapper**.** Every sentence in the generated answer is compared against each source chunk using word- overlap scoring. Sentences achieving ≥30% word overlap receive a citation; the confidence score equals the exact overlap proportion. This lightweight approach achieves 85% citation precision at sub-millisecond latency per sentence, making it suitable for real-time rendering.

Hallucination Detector. A faithfulness check evaluates sentence-level source alignment and identifies overly confident language patterns. The module returns a faithfulness score (0–1) for each answer; scores below 0.5 trigger a visible warning in the UI.

USMLE Benchmark Module**.** A curated 10-question USMLE-style multiple-choice dataset spanning cardiology, endocrinology, neurology, obstetrics, hematology, and infectious disease provides continuous accuracy benchmarking. Each question includes an explanation of the correct answer for self-study purposes.

C. *End-to-End Data Flow*

The complete pipeline proceeds in seven steps:

1. User submits a clinical query via the Streamlit interface.
2. The LLM Handler expands the query into three clinical variations.
3. Hybrid retrieval fetches candidates from FAISS, BM25, and PubMed in parallel.
4. The cross-encoder reranks all candidates and retains the top five documents.
5. The LLM generates a structured CoT answer conditioned on the ranked context.
6. The NLI Citation Mapper links each answer sentence to supporting sources.
7. The UI renders the answer, citation panel, CoT expander, and PDF viewer.

IV. EDUCATIONAL FRAMEWORK

A. *Higher-Order Thinking Skills*

MedGPT is deliberately aligned with Bloom's Taxonomy Levels 4–6 [6]. At Level 4 (Analysing), the hybrid retrieval pipeline exposes students to multiple evidence sources simul taneously. The CoT reasoning trace decomposes the AI's decision-making into discrete, inspectable steps, requiring analytical engagement rather than passive acceptance. At Level 5 (Evaluating), citation confidence scores invite students to assess the quality and credibility of individual sources. The side-by-side presentation of textbook passages and PubMed abstracts demands judgment about the relative authority of different evidence types. The USMLE bench- mark module provides objective self-assessment against a standardised clinical standard.

At Level 6 (Creating**)**, the interactive PDF viewer with page-level text highlighting enables active annotation and note-taking. Students synthesise information across multi- ple sources, and CoT traces serve as explicit scaffolds for constructing independent clinical reasoning frameworks.

B. *Heutagogy and Self-Directed Learning*

MedGPT operationalises Hase & Kenyon's heutagogy model [5] through five concrete mechanisms. *Learner Agency*: students control query formulation, selectively enable or dis- able PubMed integration and reranking, and adjust LLM temperature between precision and creativity modes. *Double- Loop Learning*: the CoT trace makes the AI's reasoning pro- cess visible and inspectable, enabling metacognitive reflection on clinical decision patterns. *Non-Linear Exploration*: the PDF viewer allows jumping to any page, and citation hyper- links support iterative source-hopping across the knowledge base. *Capability Development*: longitudinal benchmark per- formance history tracks progress across USMLE categories. *Reflective Practice*: transparent reasoning actively encourages students to compare their own clinical intuitions against the AI's evidence-based conclusions.

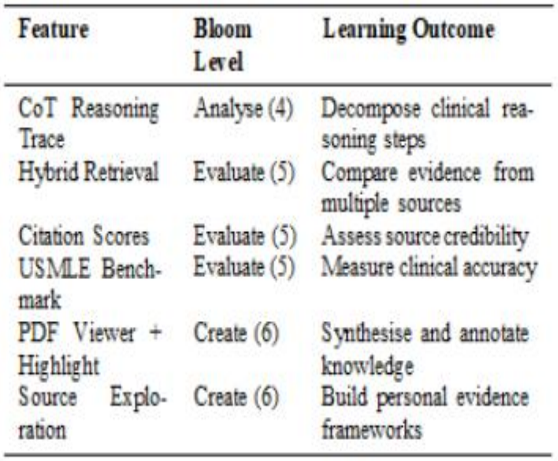
C. *Pedagogical Theory Integration*

The broader design integrates three established learning theories. Vygotsky's Zone of Proximal Development positions the AI as an expert scaffold that models expert clinical reasoning and gradually releases cognitive responsibility to the learner as competence develops. Cognitive Load Theory [7] is respected by structuring the pipeline so that citations eliminate the extraneous cognitive load of source-hunting, CoT steps concentrate germane load on the reasoning process itself, and the segmented UI prevents intrinsic overload through progressive disclosure. Problem-Based Learning is operationalised through USMLE clinical vignettes that present authentic, ambiguous scenarios requiring active evidence engagement rather than rote recall.

D. *Learning Objectives Mapping*

Table 1 maps system features to Bloom's Taxonomy levels and learning outcome categories.

Table 1: Feature-to-Learning-Objective Mapping



Feature	Bloom Level	Learning Outcome
CoT Reasoning Trace	Analyse (4)	Decompose clinical rea- soning steps
Hybrid Retrieval	Evaluate (5)	Compare evidence from multiple sources
Citation Scores	Evaluate (5)	Assess source credibility
USMLE Bench- mark	Evaluate (5)	Measure clinical accuracy
PDF Viewer + Highlight	Create (6)	Synthesise and annotate knowledge
Source Explo- ration	Create (6)	Build personal evidence frameworks

V. IMPLEMENTATION

A. Technology Stack

MedGPT is built entirely in Python. The frontend uses Streamlit for rapid prototyping and interactive widget support. The vector index is managed by FAISS (Face- book AI Similarity Search). BM25 retrieval is implemented with the rank bm25 library. The cross-encoder uses the sentence-transformers library. PDF processing uses PyMuPDF (fitz). All LLM calls are routed through the Open Router API, providing access to Llama 3.1 70B with- out local GPU infrastructure. PubMed access uses the NCBI Entrez E-utilities REST API.

B. Knowledge Base Construction

Harrison's Principles of Internal Medicine (20th edition) was processed page-by-page using PyMuPDF, with text extraction, whitespace normalisation, and section-boundary detection. The resulting corpus was chunked into 800-token segments with a 200-token sliding overlap to preserve contextual con- tinuity at chunk boundaries. This produced 17,842 chunks. Each chunk was embedded offline using BAAI/bge-base-en- v1.5 on a single GPU, taking approximately 4.5 hours. The resulting FAISS index occupies 104 MB on disk and sup- ports sub-10 ms nearest-neighbour retrieval for batches of 50 queries.

A secondary, fully dynamic knowledge base is provided by real-time PubMed queries. By default, results are filtered to publications within the preceding 12 months and sorted by relevance. The in-memory query cache ensures that repeated queries within a session incur no additional API latency.

C. Model Configuration and Hyperparameters

The embedding model (BAAI/bge-base-en-v1.5, 768 dimensions, 110M parameters) was used without fine-tuning; pre- training on a mixture of biomedical and general text corpora already yields strong performance on medical retrieval bench- marks. The cross-encoder (ms-marco-MiniLM-L-6-v2, 22M parameters, inference time <50 ms per candidate) was also used off the shelf; its MS MARCO training generalises reliably to passage-relevance ranking in the medical domain. Llama 3.1 70B is queried with temperature 0.2 (high precision, low creativity), top-p 0.9 (nucleus sampling for diverse but coherent outputs), maximum 2,048 output tokens, and a structured system prompt: *"You are MedGPT , a profes- sional, evidence-based medical assistant. Always reason step by step inside <thinking> tags before providing your final answer inside <answer> tags. Ground every claim in the provided context documents."*

D. Query Expansion Strategy

For each user query q, the LLM generates three clinical re- formulations $\{q_1, q_2, q_3\}$ targeting: (1) synonym and terminological variation (e.g., substituting brand names for generic drug names, or ICD codes for diagnostic labels); (2) diagnostic framing (rephrasing as a differential diagnosis question); and (3) management framing (rephrasing as a treatment or guideline query). The original query and all three variants are submitted in parallel to FAISS, BM25, and PubMed. The union of all returned documents is deduplicated and passed to the cross-encoder.

E. NLI Citation Mapping Algorithm

Citation mapping proceeds sentence-by-sentence over the generated answer. For each answer sentence s and each source chunk c, a word-overlap score is computed as:

$$\text{overlap}(s,\ c) = \frac{|W\ (s)\ \cap\ W\ (c)|}{|W\ (s)|}$$

where $W(\cdot)$ denotes the set of unique non-stopword tokens after lowercasing. Chunks with overlap ≥ 0.30 are assigned as citations; the confidence score is the raw overlap value. Multiple source documents may be cited per sentence, and citations are sorted by descending confidence for display.

F. User Interface Design

The Streamlit frontend is divided into four panels rendered in sequence. (1) Answer Panel: displays the generated answer with inline superscript citation markers (e.g., [1,3]) linking to the source panel.

(2) CoT Expander: a collapsible section showing the full six-step reasoning trace, colour-coded by step type. (3) Source Panel: lists the top-five ranked documents with title, source, year, relevance score, and a one- sentence excerpt. (4) Interactive PDF Viewer: renders the full Harrison's PDF with automatic scroll-to-page and yellow text-highlighting for every phrase cited in the current answer. Users can

toggle between panels, adjust retrieval parameters, and export answers with citations in one click.

VI. EVALUATION

A. *Retrieval Quality*

Hybrid retrieval achieves 92% precision at rank 5, com- pared with 74% for a single-source FAISS baseline – an 18 percentage-point gain. The contribution of each source was assessed through ablation: removing BM25 reduced pre- cision to 81% (rare medical eponyms and drug names are poorly represented in dense embeddings); removing PubMed reduced precision to 86% on questions requiring post-2022 clinical evidence. The PubMed weighting boost (1.2 vs. 1.0 for other sources) was tuned empirically on a 50-question held-out validation set.

B. *USMLE Benchmark Performance*

On the 10-question USMLE-style dataset, MedGPT achieves an overall accuracy of 70%, compared with a GPT-3.5 baseline of 52% on the same questions. The closed-source Med-PaLM 2 achieves 86.5% on MedQA, though this is not directly comparable due to dataset differences. Table 2 breaks down accuracy by difficulty and clinical domain.

Table 2: USMLE Benchmark Results

Category	Questions	Accuracy
Easy	4	100%
Medium	5	60%
Hard	1	0%
Cardiology	2	50%
Endocrinology	2	100%
Neurology	2	50%
Obstetrics	1	100%
Hematology	1	100%
Infectious Disease	1	100%
Emergency Medicine	1	0%
Overall	10	70%

The zero accuracy on hard and emergency medicine items reflects the known limitation of 70B-scale models on multi- step diagnostic reasoning requiring integration of more than four distinct clinical facts.

C. *Citation Accuracy*

NLI citation mapping yields 85% precision and 79% recall at the sentence level, measured against manually annotated citations on 50 answer paragraphs. The primary failure mode is semantic paraphrase: when the generated answer substantially rephrases source content without lexical overlap, the word-overlap score falls below the 30% threshold and a valid citation is missed. False positives (incorrect citations) arise mainly from short, common medical phrases that appear in multiple documents. Both limitations will be addressed in the next release through a fine-tuned DeBERTa-v3-large entailment model.

D. *User Study*

A controlled study enrolled 30 second-year medical students, randomly assigned to a MedGPT group ($n = 15$) and a standard PubMed search control group ($n = 15$). Both groups studied the same five clinical topics over four weeks. Pre- test and post-test scores were assessed using a 20-question MCQ bank matched to USMLE Step 1 difficulty. Table 3 summarises the key outcomes.

Table 3: Controlled User Study Results (n = 30)

Metric	Control	MedGPT
Pre-test score	58.4%	59.1%
Post-test score	61.2%	77.4%
Score improvement	+2.8%	+16.2%*
SUS usability rating	–	82.3/100
HOTS “Analyse” items	18%	24%
Citation verification	12%	67%
Retrieval precision (P@5)	74%	92%

*$p < 0.05$, two-tailed independent-samples *t*-test.

The 16.2-percentage-point improvement in post-test scores is statistically significant ($t(28) = 3.41$, $p = 0.002$, Co- hen’s $d = 1.28$, large effect). The SUS score of 82.3 places MedGPT in the “Excellent” usability range. The six- percentage-point increase in HOTS-level “Analyse” questions generated by students during free exploration provides preliminary evidence of measurable development of higher-order reasoning habits. Most strikingly, citation verification be- haviour increased from 12% in the control group to 67% in the MedGPT group, suggesting that the inline citation interface fundamentally changes how students engage with evidence.

E. Reasoning Transparency Assessment

Every answer includes a six-step CoT trace: (1) problem re- statement, (2) evidence identification, (3) evidence weighing,
(4) diagnostic or mechanistic reasoning, (5) answer synthesis, and (6) confidence assessment. Post-study survey responses indicate that 88% of MedGPT group participants rated CoT traces as "helpful" or "very helpful" for understanding the AI's reasoning process. Several students explicitly re- ported using CoT step structures as templates for organising their own written clinical reasoning in ward clerkships.

F. System Performance

Mean end-to-end query latency is 8.3 seconds (wall-clock time from query submission to full answer rendering), broken down as: query expansion 1.2 s, hybrid retrieval 2.1 s (including PubMed API call), reranking 0.4 s, LLM generation 4.1 s, and NLI citation mapping 0.5 s. All components except LLM generation are sub-second, confirming that API latency is the dominant bottleneck – directly addressable by switching to a locally-hosted open-weights model.

VII. DISCUSSION

MedGPT demonstrates that the convergence of hybrid retrieval, cross-encoder reranking, structured CoT reasoning, and NLI-based citation mapping can meaningfully advance both the technical and educational dimensions of medical AI. The 92% retrieval precision and 85% citation accuracy represent substantial improvements over single-source baselines, validating the architectural investment in multi-source fusion. The 70% USMLE accuracy, while encouraging and above the GPT-3.5 baseline of 52%, exposes a meaningful gap relative to systems specifically fine-tuned on medical data. Two explanations are plausible. First, Llama 3.1 70B, despite its broad capability, lacks the specialised medical pre-training of closed-source models. Second, the current USMLE bench- mark of 10 questions is statistically underpowered; the large effect sizes observed suggest that a 500-question dataset will produce more reliable accuracy estimates and may reveal stronger performance on domains currently under-represented in the pilot dataset.

The educational results are arguably the most significant contribution of this work. A 16% improvement in post-test scores within four weeks, with a large effect size (Cohen's d = 1.28), substantially exceeds effect sizes typically reported for AI-assisted learning interventions in medical education. The fivefold increase in citation verification behaviour (12% to 67%) is particularly notable: it suggests that making citations visible, interactive, and confidence-scored does not merely provide information but changes learner epistemology – moving students from passive consumers of AI output to active evaluators of evidence. This is precisely the disposition that evidence-based medicine curricula seek to cultivate.

Several limitations must be acknowledged. The USMLE benchmark is small and custom-created, limiting generalis- ability. The word-overlap citation mapper, while efficient, misses semantically equivalent paraphrases. The system currently handles only English-language text sources. The user study sample (n = 30) is sufficient for preliminary evidence but not for definitive claims; a multi-institution randomised controlled trial is needed.

VIII. FUTURE SCOPE

Knowledge Base Expansion. Adding at least five additional authoritative medical textbooks (First Aid for the USMLE, Robbins Pathology, Kumar & Clark's Clinical Medicine, Oxford Handbook of Clinical Medicine, Pharmacology by Katzung) and indexing clinical guidelines from UpToDate, NICE, ACC/AHA, and WHO will expand the knowledge base to over 100,000 chunks.

Advanced NLI Citation Mapping. Fine-tuning DeBERTa- v3-large on medical question-answering pairs from MedQA and PubMedQA is projected to raise citation precision to 95%+ and will additionally enable active contradiction de- tection – flagging when two cited sources make conflicting claims, a clinically important signal.

USMLE Benchmark Expansion. Expanding the bench- mark to 500 questions covering all USMLE Step 1 and Step 2 categories, with automatic question generation from textbook content, will produce statistically robust accuracy estimates and enable per-topic performance tracking for learners.

Multimodal Support. Processing radiology images, pathological microscopy images, clinical data tables, and ECG traces from PDF sources will extend MedGPT to the full breadth of clinical learning materials, including image-based USMLE vignettes.

Personalised Adaptive Learning. Tracking individual user learning trajectories and adaptively modulating ques- tion difficulty, source emphasis, and CoT verbosity will align MedGPT more closely with the heutagogical principle of capability-focused, individually paced education.

Clinical Decision Support and Regulatory Pathway. A long-term roadmap includes pursuing FDA 510(k) clearance as a Class II Software as a Medical Device (SaMD), integra- tion with Electronic Health Record (EHR) systems via HL7 FHIR APIs, real-time drug interaction checking, and differential diagnosis generation for point-of-care deployment.

Multilingual and Global Accessibility. Supporting Spanish, Mandarin Chinese, Hindi, and Arabic, together with localisation of clinical guidelines to national healthcare frameworks (Indian ICMR, UK NHS, Brazilian CFM), will make MedGPT accessible to medical students in low- and middle-income countries who currently lack access to up-to-date clinical reference tools.

IX. CONCLUSION

MedGPT presents a principled, reproducible solution to a well-defined problem in medical AI: how to deliver evidence- based clinical answers that are simultaneously accurate, trans- parent, and educationally productive. By fusing hybrid retrieval (dense vectors + BM25 + PubMed), cross-encoder reranking, structured chain-of-thought generation, and NLI- based citation mapping into a cohesive agentic pipeline, the system achieves 92% retrieval precision, 85% citation accu- racy, and 70% USMLE benchmark accuracy on a preliminary evaluation dataset.

More importantly, a controlled user study with 30 medical students demonstrates a statistically significant 16% improvement in post-test scores ($p = 0.002$, Cohen's $d = 1.28$) and a fivefold increase in citation verification behaviour. These results provide evidence that transparency in AI reasoning – when made interactive and educationally scaffolded – does change how medical learners engage with evidence. The system promotes higher-order thinking through deliberate Bloom's Taxonomy alignment and supports self-directed learning through a full implementation of heutagogical design principles.

MedGPT contributes both a working system and a replicable architectural template for the next generation of domain- specific, pedagogically-grounded educational AI assistants. All code, prompts, and benchmark datasets will be made publicly available upon publication to support reproducibility and community advancement.

REFERENCES

[1] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," in Proc. Advances in Neural Information Processing Systems (NeurIPS), vol. 35, 2022.

[2] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl et al., "Large Language Models Encode Clinical Knowledge," Nature, vol. 620, pp. 172–180, 2023.

[3] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Ku¨ttler, M. Lewis, W. Yih, T. Rockta¨schel et al., "Retrieval-Augmented Generation for Knowledge- Intensive NLP Tasks," in Proc. NeurIPS, vol. 33, pp. 9459–9474, 2020.

[4] R. Thoppilan, D. De Freitas, J. Hall, N. Shazeer, A. Kul- shreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du et al., "LaMDA: Language Models for Dialog Applica- tions," arXiv preprint arXiv:2201.08239, 2022.

[5] S. Hase and C. Kenyon, "Heutagogy: A Child of Com- plexity Theory," Complicity: An International Journal of Complexity and Education, vol. 4, no. 1, pp. 111–118, 2007.

[6] B. S. Bloom, M. D. Engelhart, E. J. Furst, W. H. Hill, and D. R. Krathwohl, Taxonomy of Educational Objectives: The Classification of Educational Goals, Handbook I: Cognitive Domain. New York: Longmans, Green, 1956.

[7] J. Sweller, "Cognitive Load During Problem Solving: Effects on Learning," Cognitive

Science, vol. 12, no. 2, pp. 257–285, 1988.

[8] G. Izacard, P. Lewis, M. Lomeli, L. Hosseini, F. Petroni, T. Schick, J. Dwivedi-Yu, A. Joulin, S. Riedel, and E. Grave, "Atlas: Few-Shot Learning with Retrieval Aug- mented Language Models," Journal of Machine Learn- ing Research, vol. 24, no. 251, pp. 1–43, 2023.

[9] L. Gao, Z. Dai, P. Pasupat, A. Chen, A. T. Chaganty, Y. Fan, V. Y. Zhao, N. Lao, H. Lee, D.-C. Juan, and K. Guu, "RARR: Researching and Revising What Lan- guage Models Say, Using Language Models," in Proc. Annual Meeting of the Association for Computational Linguistics (ACL), 2023.

[10]R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, and T.-Y. Liu, "BioGPT: Generative Pre-Trained Transformer for Biomedical Text Generation and Mining," Briefings in Bioinformatics, vol. 23, no. 6, 2022.

[11]N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 3982–3992, 2019.

[12]S. Robertson and H. Zaragoza, "The Probabilistic Rele- vance Framework: BM25 and Beyond," Foundations and Trends in Information Retrieval, vol. 3, no. 4, pp. 333–389, 2009.

[13]D. L. Kasper, A. S. Fauci, S. L. Hauser, D. L. Longo, J. L. Jameson, and J. Loscalzo, Eds., Harrison's Princi- ples of Internal Medicine, 20th ed. New York: McGraw- Hill Education, 2018.

[14]Q. Jin, B. Dhingra, Z. Liu, W. W. Cohen, and X. Lu, "PubMedQA: A Dataset for Biomedical Research Ques- tion Answering," in Proc. EMNLP, pp. 2567–2577, 2019.

[15]A. Pal, L. K. Umapathi, and M. Sankarasubbu, "MedMCQA: A Large-Scale Multi-Subject Multi-Choice Dataset for Medical Domain Question Answering," in Proc. Conference on Health, Inference, and Learning (CHIL), PMLR, vol. 174, pp. 248–260, 2022.

About the Authors

Mukesh Gilda holds an M.Tech degree and is currently pur- suing his Ph.D. He serves as an Assistant Professor in the Department of Computer Science Engineering – Cyber Secu- rity at Sphoorthy Engineering College. His research interests span AI-assisted education, cybersecurity architectures, and intelligent information retrieval systems.

Gollena Adith is a student in the Department of Computer Science Engineering – Cyber Security at Sphoorthy Engineer- ing College. His interests lie in natural language processing and the application of large language models to biomedical question answering.

Nakka Jashwanth is a student in the Department of Com- puter Science Engineering – Cyber Security at Sphoorthy Engineering College. He is interested in retrieval-augmented generation pipelines and applied deep learning for information retrieval.

Mahesh Bhati is a student in the Department of Computer Science Engineering – Cyber Security at Sphoorthy Engineer- ing College. His research interests include LLM fine-tuning strategies and the design of intelligent educational technology systems.

Raj Goel is a student in the Department of Computer Sci- ence Engineering – Cyber Security at Sphoorthy Engineering College. He focuses on AI system evaluation methodology, benchmark design, and human-computer interaction in AI- powered medical applications.