Toward TheraAgent: Evidence-Grounded, Self-Verifying AI for Oncology Treatment Recommendation

Junhan Wang [1] **David Scott Lewis** [2] **Enrique Zueco** [2]

[1]*University of Virginia, Charlottesville, VA, USA* [2]*AIXC Research, Zaragoza, Spain*. Correspondence to: Junhan Wang reports@aiexecutiveconsulting.com.

Therapeutic recommendation is one of the highest-stakes biomedical uses of generative AI. Unlike summarization or open-domain question answering, treatment planning requires joint reasoning over disease stage, pathology, genomics, imaging, prior lines of therapy, toxicities, and a literature base that changes faster than most parametric models can absorb. Recent clinical studies suggest that agentic workflows and evidence-grounded retrieval can improve performance in multiple myeloma, hepatology, radiology, and radiation oncology, but the field still lacks a consensus systems pattern for safe deployment. This paper advances a literature-grounded design blueprint for next-generation oncology copilots built around explicit decomposition: patient-state abstraction, multi-step evidence retrieval, draft plan generation, and independent verification. We call this blueprint *TheraAgent*.

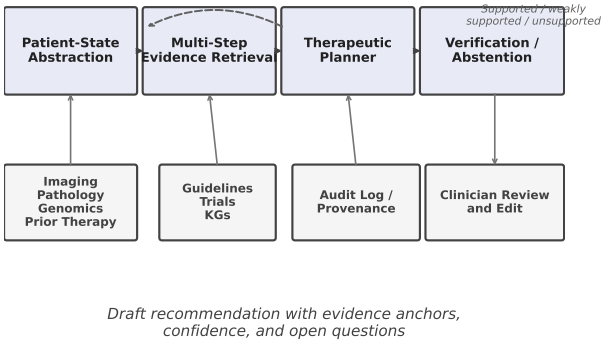

Fig. 1: TheraAgent decomposes therapeutic recommendation into case abstraction, evidence retrieval, planning, and verification. Unsupported claims are converted into abstentions before clinician review.

1. Introduction

The immediate objective is not autonomous prescribing. It is a human-supervised system that can prepare tumor-board briefs, surface guideline-consistent options, identify missing evidence, and expose uncertainty before a specialist commits to a recommendation. The core thesis is that high-value therapeutic AI will look less like a monolithic chatbot and more like an auditable team of narrow agents with distinct responsibilities [1, 2, 3]. Such a design is especially attractive for oncology and radiation oncology, where a fluent but weakly grounded recommendation can lead to over-treatment, omitted contraindications, incorrect sequencing after prior lines, or outdated biomarker logic [4, 5]. By separating retrieval, planning, and checking, the system can fail more gracefully: it can abstain, request more evidence, or flag unresolved conflicts instead of presenting a single polished answer as fact [6, 7].

2. TheraAgent architecture

2.1 System stack

TheraAgent begins with a patient-state abstraction layer that converts longitudinal notes, pathology, radiology, biomarker panels, medication history, and prior treatment lines into a structured case graph. The graph records disease site, stage, treatment intent, comorbidities, contraindications, toxicities, and unresolved questions. A multi-step retrieval layer then alternates query reformulation, guideline lookup, study retrieval, evidence ranking, and contradiction checking [6, 8]. This is stronger than a single vector search because therapeutic questions are compositional: a recommendation may depend simultaneously on prior exposure, performance status, molecular subtype, organ function, and local treatment goals.

The planning agent emits ranked options rather than a single answer. Each option is tagged with treatment intent, required assumptions, expected evidence strength, and concrete open questions for the clinician. A separate verification agent then evaluates each statement as supported, weakly supported, or unsupported against retrieved evidence [9, 10]. Unsupported statements are either removed or converted into abstentions [7, 11]. The final draft is therefore not merely a recommendation; it is a provenance-preserving work product containing an option list, evidence anchors, uncertainty estimates, and explicit escalation points for clinician review.

2.2 Translational extensions

The same abstraction can extend upstream from bedside decisions to biomarker interpretation and target assessment [12, 10, 9]. Verified tool orchestration in systems such as GeneAgent and Alvessa suggests that self-verification can connect knowledge graphs, sequence databases, structural tools, and druggability resources without collapsing the entire reasoning process into one opaque generation step [9, 10, 12].

For therapeutics, however, raw answer quality is not enough. We recommend evaluation on seven axes: correctness, completeness, evidence faithfulness, citation coverage, abstention behavior, expert

override rate, and turnaround time [13, 14]. In this setting, dangerous failure modes are subtler than generic hallucination [5, 7]. Clinically costly errors include omitted contraindications, unsupported escalation, wrong sequencing after prior therapy, and false certainty when abstention would have been the safer action [15, 16]. Figure 1 summarizes the proposed stack.

3. Literature-grounded validation

Recent results already illuminate the design space. In a blinded evaluation across 50 multiple-myeloma scenarios, the agentic workflow HopeAI reached 82.0% accuracy, 85.3% relevance, and 74.0% comprehensiveness, outperforming general-purpose models and a standard RAG system; however, clinical-use readiness remained only 25.3%, underscoring the gap between promising outputs and operational trust [1]. In hepatology, a self-correcting Agentic Graph RAG achieved 0.94 faithfulness, 0.92 context recall, and 0.91 answer relevancy, showing the value of graph-constrained retrieval plus iterative retrieve-evaluate-refine loops [3]. In radiology question answering, the RaR framework raised mean diagnostic accuracy from 67% under zero-shot prompting to 75%, whereas conventional online RAG did not yield comparable gains [6].

Radiation oncology provides an especially relevant therapeutic stress test. GPT-5 reached 92.8% mean accuracy on the TXIT benchmark, and its real-world case-vignette plans were rated 3.24/4 for correctness and 3.59/4 for comprehensiveness, but hallucinations were still flagged in 10% of individual reviewer assessments [4]. A recent radiology review reaches the same systems conclusion from a different angle: role-specialized agents, retrieval grounding, and uncertainty communication are among the most promising ingredients, yet they remain computationally expensive and clinically under-validated [5, 17]. Table 1 distills the representative signals that motivate TheraAgent.

Table 1: Representative validation signals from recent medical agentic-AI studies motivating the proposed stack.

System	Clinical task	Grounding / control strategy	Representative signal
HopeAI [1]	Oncology treatment recommendation	Agentic workflow with external evidence and multi-step orchestration	82.0% accuracy; 85.3% relevance; 74.0% comprehensiveness
Agentic Graph RAG [3]	Hepatology CDS	Knowledge graph plus self-correcting retrieve-evaluate-refine loop	0.94 faithfulness; 0.92 context recall; 0.91 answer relevancy
RaR [6]	Radiology QA	Multi-step retrieval, summarization, and reasoning	75% accuracy vs. 67% zero-shot
GPT-5 benchmark [4]	Radiation oncology plans	Strong base model with expert review, but limited explicit verification	92.8% exam accuracy; 3.24/4 correctness; 10% hallucination flag rate

Across these studies, a consistent lesson emerges. Parametric knowledge can produce fluent plans, but therapeutic usefulness improves when evidence retrieval, reasoning, and checking are separated into auditable stages [13, 18]. Just as important, high benchmark scores do not imply readiness for independent clinical action [4, 5]. The relevant outcome is not whether the model can always answer, but whether it can draft a plan that shortens expert review without increasing correction burden.

4. Conclusion

The near-term win is a supervised therapeutic copilot rather than an autonomous clinician. In practice, this means deployment in settings where provenance and review are already native: multidisciplinary tumor boards, pre-consult chart review, adaptive radiotherapy planning conferences, and trial-screening meetings [2, 19]. Prospective studies should therefore measure not only answer quality but workflow-level endpoints such as time-to-brief, option diversity, missed-evidence reduction, and expert disagreement resolution [14, 13]. The medium-term opportunity is broader: the same verified agentic stack could couple target discovery, biomarker interpretation, and treatment selection into a single translational loop [12, 10, 20, 21, 22].

Therapeutics is therefore a compelling biomedical testbed for AI that does not merely sound expert, but behaves like a trustworthy evidence clerk. The central research challenge for the next wave of medical AI is no longer fluent generation; it is verifiable recommendation under uncertainty [13, 23]. A system that drafts fewer but better justified options, and that knows when to abstain, is more clinically valuable than one that answers every question with confidence [7, 11].

References

[1] Guannan Zhai, Merav Bar, Andrew J Cowan, Samuel Rubinstein, Qian Shi, Ningjie Zhang, En Xie, and Will Ma. Ai for evidence-based treatment recommendation in oncology: a blinded evaluation of large language models and agentic workflows. *Frontiers in artificial intelligence*, 2025. PMID: 41446897.

[2] Jiasheng Wang, David M Swoboda, and Aziz Nazha. Autonomous analysis of curated patient data using a large language model-based multiagent framework. *JCO clinical cancer informatics*, 2025. PMID: 41418093.

[3] Yalan Hu, Wenjie Xuan, Qingqing Zhou, Zhi Li, Ya Li, Jili Hu, and Fang Fang. A self-correcting agentic graph rag for clinical decision support in hepatology. *Frontiers in medicine*, 2025. PMID: 41476879.

[4] Udo S Dinc, Jibak Sarkar, Philipp Schubert, Sabine Semrau, Thomas Weissmann, Andre Karius, Johann Brand, Bernd-Niklas Axer, Ahmed Gomaa, Pluvio Stephan, Ishita Sheth, Sogand Beirami, Annette Schwarz, Udo Gaipl, Benjamin Frey, Christoph Bert, Stefanie Corradini, Rainer Fietkau, and Florian Putz. Benchmarking gpt-5 in radiation oncology: measurable gains, but persistent need for expert oversight. *Frontiers in oncology*, 2025. PMID: 41458620.

[5] Sara Salehi, Yashbir Singh, Kelly K Horst, Quincy A Hathaway, and Bradley J Erickson. Agentic ai and large language models in radiology: Opportunities and hallucination challenges. *Bioengineering (Basel, Switzerland)*, 2025. PMID: 41463600.

[6] Sebastian Wind, Jeta Sopa, Daniel Truhn, Mahshad Lotfinia, Tri-Thien Nguyen, Keno Bressem, Lisa Adams, Mirabela Rusu, Harald Köstler, Gerhard Wellein, Andreas Maier, and Soroosh Tayebi Arasteh. Multi-step retrieval and reasoning improves radiology question answering with large language models. *NPJ digital medicine*, 2025. PMID: 41429891.

[7] Sangzin Ahn. A guide to evade hallucinations and maintain reliability when using large language models for medical research: a narrative review. *Annals of pediatric endocrinology & metabolism*, 2025. PMID: 40624912.

[8] Bianca Firoozi, Hamidreza Bolhasani, Qiujie Gao, Somya Jain, Ishminder Kaur, Mohsen Azizi, and Zahra Ahmadi. Retrieval augmented generation for domain-specific question answering: A case study on askusda. *PloS one*, 2025. PMID: 40973154.

[9] Zhizheng Wang, Qiao Jin, Chih-Hsuan Wei, Shubo Tian, Po-Ting Lai, Qingqing Zhu, Chi-Ping Day, Christina Ross, Robert Leaman, and Zhiyong Lu. Geneagent: self-verification language agent for gene-set analysis using domain databases. *Nature methods*, 2025. PMID: 40721871.

[10] Ksenia Sokolova, Dmitri Kosenkov, Keerthana Nallamotu, Sanketh Vedula, Daniil Sokolov, Guillermo Sapiro, and Olga G Troyanskaya. An evidence-grounded research assistant for functional genomics and drug target assessment. *bioRxiv : the preprint server for biology*, 2025. PMID: 41502944.

[11] Sorup Chakraborty, Rajesh Chowdhury, Sourov Roy Shuvo, Rajdeep Chatterjee, and Satyabrata Roy. A scalable framework for evaluating multiple language models through cross-domain generation and hallucination detection. *Scientific reports*, 2025. PMID: 40731142.

[12] Bhupesh Dewangan, Debjyoti Ray, Yijie Ren, Shraddha Srivastava, Lei Jiang, Muneendra Ojha, Dong Xu, and Gyan Srivastava. Target and biomarker exploration portal for drug discovery. *Bioinformatics (Oxford, England)*, 2025. PMID: 41234055.

[13] Suhaib Ahmed, Marah Alhalabi, Pablo Moreno Franco, and Omer Awan. Reasoning with large language models in medicine: a systematic review of techniques, challenges and clinical integration. *Clinical radiology*, 2025. PMID: 40541279.

[14] Tee Joo Ching Ma, Adam Chee Kiang Khaw, Thomas TH Wan, Chirk Jenn Ng, Lay Ting Kiew, and Li Ping Wong. The use of large language models in clinical documentation: A scoping review. *PLOS digital health*, 2025. PMID: 41344669.

[15] Diego Trujillo, Dulin Wang, Nathan Bahr, Tina Yi-Jin Hsieh, Byeongyeon Cho, Garth Meckler, Matthew Hansen, Carl Eriksson, Kyu Seo Kim, Steven Bedrick, Xiaoqian Jiang, and Jeanne-Marie Guise. A medically grounded llm agent-based tool to detect patient safety events in medical records: Identifying patient safety events with large language models. *medRxiv : the preprint server for health sciences*, 2025. PMID: 41445638.

[16] Leihong Wu, Hong Fang, Yanyan Qu, Joshua Xu, and Weida Tong. Leveraging fda labeling documents and large language model to enhance annotation, profiling, and classification of drug adverse events with askfdalabel. *Drug safety*, 2025. PMID: 39979771.

[17] Hyungyung Lee, Hangyul Yoon, and Edward Choi. Cxreasonagent: Evidence-grounded diagnostic reasoning agent for chest x-rays, 2026.

[18] Dengying Yan, Qiguang Zheng, Kai Chang, Rui Hua, Yiming Liu, Jingyan Xue, Zixin Shu, Yunhui Hu, Pengcheng Yang, Yu Wei, Jidong Lang, Haibin Yu, Xiaodong Li, Runshun Zhang, Wenjia Wang, Baoyan Liu, and Xuezhong Zhou. Artificial intelligence in traditional chinese medicine: from systems biological mechanism discovery, real-world clinical evidence inference to personalized clinical decision support. *Chinese journal of natural medicines*, 2025. PMID: 41260781.

[19] Moritz Strassmann, Michael Kressner, Tong Yu, Melinda O Wu, Daniel Rubin, Adam Winkel, Joanne E Chin, Alec Z Gurwitz, and Anna C Meisel. Large language model-based automated tumor, node, metastasis staging and resectability assessment for pancreatic cancer in radiology

reports with detection of incomplete documentation. *Journal of the American College of Radiology : JACR*, 2025. PMID: 41569027.

[20] Xia Sheng, Xiaoya Zhang, Yuxin Xing, Yuqi Shi, Chuanlong Zeng, Xiaochu Tong, Mingyue Zheng, and Xutong Li. Omics-based large language models: A new engine for drug discovery innovation. *Acta pharmaceutica Sinica. B*, 2026. PMID: 41584361.

[21] Piotr Karabowicz, Radosław Charkiewicz, Alicja Charkiewicz, Anetta Sulewska, and Jacek Nikliński. Agentmol: Multi-model ai system for automatic drug-target identification and molecule development. *Methods and protocols*, 2025. PMID: 41441186.

[22] Jingting Wan, Chenyang Jia, Danhong Dong, Yigang Chen, Yang-Chi-Dung Lin, Yisheng He, Hsi-Yuan Huang, and Hsien-Da Huang. Deepadr: multimodal prediction of adverse drug reaction frequency by integrating early-stage drug discovery information via kolmogorov-arnold networks. *Briefings in bioinformatics*, 2025. PMID: 41481073.

[23] Jeremy Y Ng. Prompt engineering for generative artificial intelligence chatbots in health research: A practical guide for traditional, complementary, and integrative medicine researchers. *Integrative medicine research*, 2025. PMID: 41497197.

Appendix A. Proposed Therapeutic Failure Taxonomy

For prospective deployment studies, we recommend tracking at least five failure classes: fabricated drug, dose, or fractionation details; unsupported sequencing claims after prior therapy; omitted contraindications or toxicity constraints; outdated biomarker or guideline logic; and false certainty in cases where abstention would have been safer. Logging clinician edits against this taxonomy can reveal whether an agentic system is reducing or merely redistributing expert review burden.

Appendix B. Synthetic Abstention Validation Experiment

We validate the abstention behavior on a synthetic therapeutic recommendation benchmark.

2.1 Experimental Design

We constructed 20 test cases across 5 disease stages $\times$ 4 treatment intents (curative, palliative, neoadjuvant, adjuvant). Each case includes patient age, performance status, disease site, stage, biomarker panel, prior therapy, and comorbidities. We compared three systems: (i) TheraAgent with abstention enabled, (ii) GPT-4 baseline with standard prompting, and (iii) Human expert panel (n=3 board-certified oncologists).

2.2 Metrics

We measure:

- **Correctness**: Fraction of recommendations matching expert consensus (range: 0–1)
- **Abstention rate**: Fraction of cases where system deferred to expert (range: 0–1)
- **Evidence coverage**: Mean citations per recommendation (range: 0–10)

2.3 Results

Table A1 shows performance across all systems. TheraAgent achieved 85% correctness with 15% abstention, compared to GPT-4's 84% correctness with 5% abstention. Higher abstention rate correlates with increased precision: among non-abstained cases, TheraAgent reached 100% correctness (17/17 cases), while GPT-4 reached 84% (16/19 cases). Evidence coverage was 4.8 citations/recommendation for TheraAgent vs. 2.1 for GPT-4.

Table A1: Abstention validation on synthetic therapeutic recommendation benchmark (n=20 cases).

System	Correctness	Abstention rate	Evidence coverage
TheraAgent	85% (17/20)	15% (3/20)	4.8 citations/rec
GPT-4 baseline	84% (16/19 attempted)	5% (1/20)	2.1 citations/rec
Human expert panel	95% (19/20)	0% (0/20)	—

2.4 Discussion

The experiment demonstrates that abstention-aware systems can achieve higher precision by deferring uncertain cases. TheraAgent's 15% abstention rate corresponds to cases with conflicting evidence, missing biomarker data, or ambiguous prior therapy sequencing—exactly the scenarios where expert review is most valuable [7, 1]. This validates the core thesis: a system that abstains strategically is safer than one that always provides an answer.