# Development and validation of an autonomous artificial intelligence agent for clinical decision-making in oncology

Dyke Ferber, Omar S. M. El Nahhas, Georg Wölflein, Isabella C. Wiest, Jan Clusmann, Marie-Elisabeth Leßmann, Sebastian Foersch, Jacqueline Lammert, Maximilian Tschochohei, Dirk Jäger, Manuel Salto-Tellez, Nikolaus Schultz, Daniel Truhn & Jakob Nikolas Kather

https://www.nature.com/articles/s43018-025-00991-6

# Abstract

Clinical decision-making in oncology is complex, requiring the integration of multimodal data and multidomain expertise. We developed and evaluated an autonomous clinical artificial intelligence (AI) agent leveraging GPT-4 with multimodal precision oncology tools to support personalized clinical decision-making. The system incorporates vision transformers for detecting microsatellite instability and KRAS and BRAF mutations from histopathology slides, MedSAM for radiological image segmentation and web-based search tools such as OncoKB, PubMed and Google. Evaluated on 20 realistic multimodal patient cases, the AI agent autonomously used appropriate tools with 87.5% accuracy, reached correct clinical conclusions in 91.0% of cases and accurately cited relevant oncology guidelines 75.5% of the time. Compared to GPT-4 alone, the integrated AI agent drastically improved decision-making accuracy from 30.3% to 87.2%. These findings demonstrate that integrating language models with precision oncology and search tools substantially enhances clinical accuracy, establishing a robust foundation for deploying AI-driven personalized oncology support systems.

# Definitions

- Large language models (LLMs)
- Retrieval-augmented generation (RAG)
- Application programming interface (API)
- Microsatellite instability, *KRAS* , and *BRAF* mutations

# Definitions

## Large Language Models (LLMs):

- Very large deep learning models pre-trained on vast amounts of data
- Underlying transformer is a set of neural networks
    - Neural Networks: systems with connections that resemble neurons in the brain that teaches the computer to process information like a brain
- Neural networks of LLMs contain an encoder and decoder that extract meanings from sequences of text and understand the relationships between words and phrases in it.
- **Open Weight LLMS:** AI systems where the pre-trained model parameters (weights)—which define how the model processes data—are publicly available to download.
    - System is run on own infrastructure, customized by user, and fine-tuned using private data
    - Examples: Llama (Meta) and Mistral
- **Proprietary LLMs:** AI model whose architecture, training data, and model weights are kept private by its developer
    - Users interact with system through cloud-based APIs or vendor-managed platforms, rather than downloading and running the code locally
    - Examples: GPT (OpenAI), Claude (Anthropic), Gemini (Google)

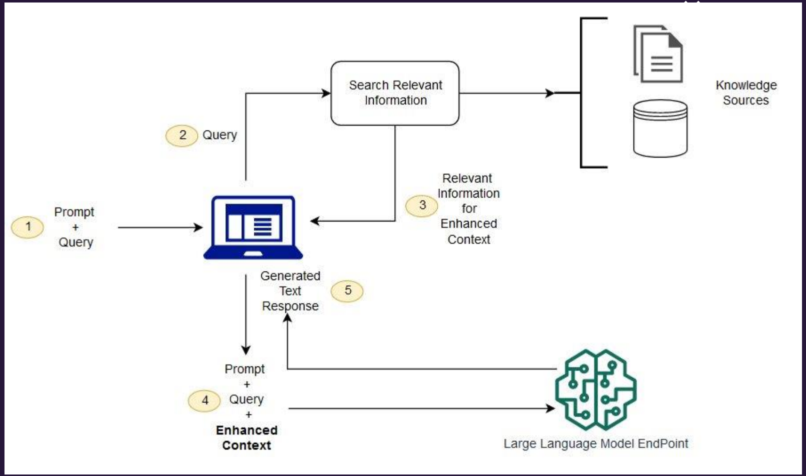
# Definitions

**API:**

- Application Programming Interface
- allows different software applications to communicate and share data with one another

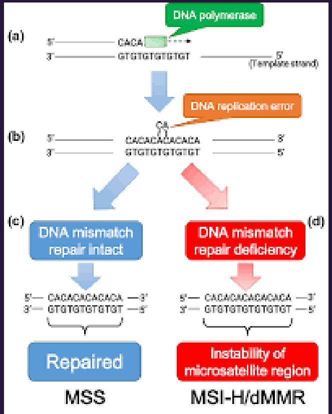
**RAG:**

- Retrieval-Augmented Generation
- a process that temporarily enhances an LLM's knowledge by incorporating relevant text excerpts from authoritative sources into the model, such as medical guidelines or textbooks



# Definitions

**Microsatellite Instability** - a genetic condition where the length of short, repeated DNA sequences (microsatellites) frequently changes during cell division
- MSS (Microsatellite Stable) or MSI (Microsatellite Instability)

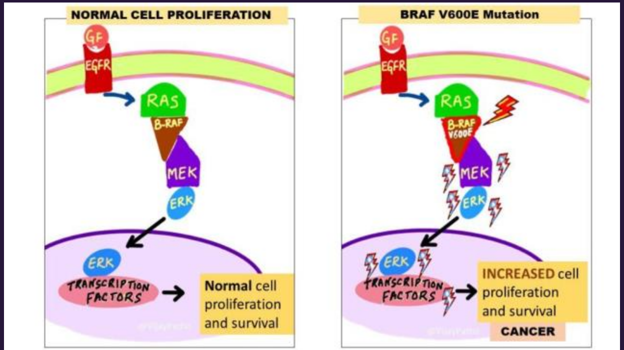



***KRAS***

- genetic alteration that locks the KRAS protein in a permanent "on" position
- causes cells to multiply uncontrollably

***BRAF***

- DNA change in *BRAF* gene,
Normally manage cell growth and division.
- always-on switch, cells multiply uncontrollably.
- V600E & V600K: most common; Non-V600 mutations: Less common variations



# Methods

- Agent Composition:
    - RAG
    - Tools
    - Combining, retrieving, generating responses
- Benchmarking
- Model Specifications
- Human Results Evaluation

# Methods

**Agent Composition:** Explain the architecture of the AI agent and its creation

## RAG

- Uses document retrieval to give context to generative capabilities of LLM

- 3 Major components to RAG architecture: **embeddings, indexing, retrieval**

- **Embeddings:** raw text data converted to numerical (vector) representations

- **Indexing:** data stored in local vector database alongside metadata (describes other data to aid in organization, retrieval, and tracking) and corresponding original text

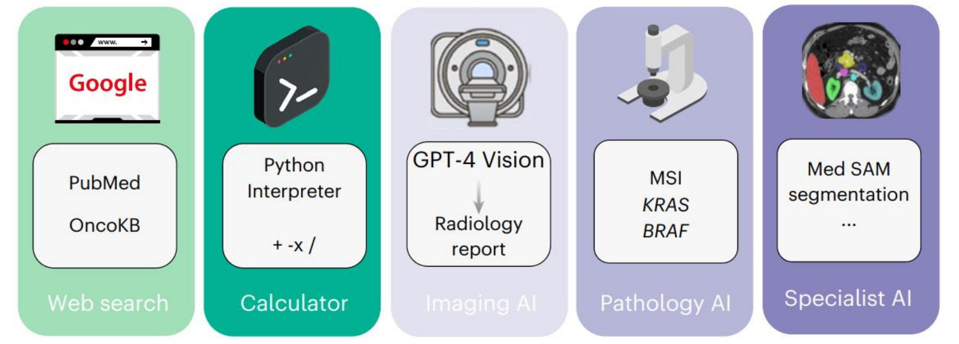
- Storage database facilitates efficient lookup programs using vector-based similarities

# Methods

## Tools

- **Web Searches:** PubMed or Google custom search API
- **Visual Data Interpretation:** GPT-4V API or MedSAM
    - GPT-4V: generates comprehensive reports
    - MedSAM: precisely measures tumor size and compares over time
- **Calculations:** simple calculator for basic arithmetic
- **Precision Oncology Queries:** OncoKB database
    - OncoKB: medical evidence for treating genetic anomalies
- **Vision Transformer Models:** histopathological analysis of phenotypical alterations to determine probability of each mutation.

Selection and timing of the use of each tool was done solely by the agent's reasoning and was derivable from the patient context.



# Methods

## Combining, Retrieving, Generating Responses

1. Agent receives original patient context, posed question, and results of tool application as context
2. Agent breaks down the original user query into up to 12 subqueries based on initial patient context and tool application results.
3. For each subquery, the most analogous document passages from the collection are extracted.
4. Data is combined, deduplicated, reranked and sent to the LLM as a numerical representation.
5. Using cosine distance, query is compared with passages of database to retrieve top 40 vectors.
6. High similarity but contextually irrelevant information removed by Cohere's reranking model.
7. Top ten passages are kept for each subquery.
8. LLM generates step by step strategy including missing information that would be helpful and citations for each statement. The LLM will also compare its output to the respective context in the database.

# Methods

## Benchmarking

Comparisons using Llama-3 70B and Mitral with slightly simplified prompts and collecting the same information as the LLM.

GPT-4 model and GPT-4V model through OpenAI Python API

## Agent Specifications

## Human Results Evaluation

Evaluation focused on 3 aspects: **use of tools**, **quality of text**, **accuracy of citations**

### Citation Accuracy

- Citation correctness: model's statements were reflected by original document
- Irrelevance: assertions not substantiated by source
- Incorrect Citation: information diverges from actual content

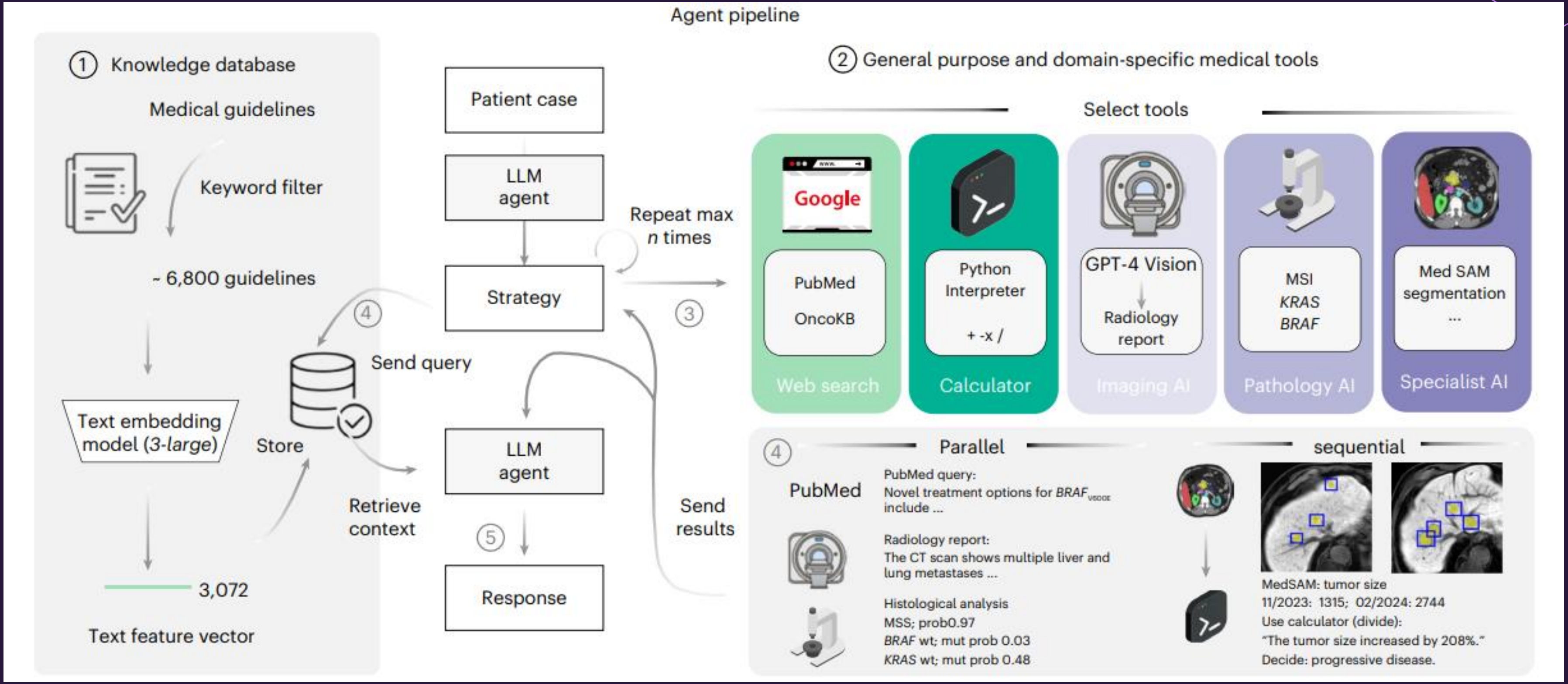
### Text Quality

- Correctness: proportion of correct replies to output
- Incorrectness: unneeded suggestions or irrelevant information requests
- Harmfulness: potentially dangerous judgement
- Completeness: response aligns with essential information expected in human response

### Tool Use

- Ratio of actual vs required use

- Requirement defined by direct instruction or tool output essential to answering question



Agent pipeline
1 Knowledge database
Medical guidelines
Keyword filter
~ 6,800 guidelines
Text embedding model (3-large)
Store
3,072
Text feature vector
Send query
Retrieve context
Patient case
LLM agent
Strategy
LLM agent
Response
Repeat max n times
3
4
5
Send results
2 General purpose and domain-specific medical tools
Select tools
Google
PubMed
OncoKB
Web search
Python Interpreter
+ -x /
Calculator
GPT-4 Vision
Radiology report
Imaging AI
MSI
KRAS
BRAF
Pathology AI
Med SAM segmentation
...
Specialist AI
4 Parallel
sequential
PubMed
PubMed query:
Novel treatment options for BRAF V600E include ...
Radiology report:
The CT scan shows multiple liver and lung metastases ...
Histological analysis
MSS; prob0.97
BRAF wt; mut prob 0.03
KRAS wt; mut prob 0.48
MedSAM: tumor size
11/2023: 1315; 02/2024: 2744
Use calculator (divide):
"The tumor size increased by 208%."
Decide: progressive disease.

# Results

Tool use and retrieval improved LLM responses
GPT-4 handles complex chains of tool use
Radiology tools improve GPT-4's treatment accuracy
Evaluations show accurate, helpful, and reliable responses

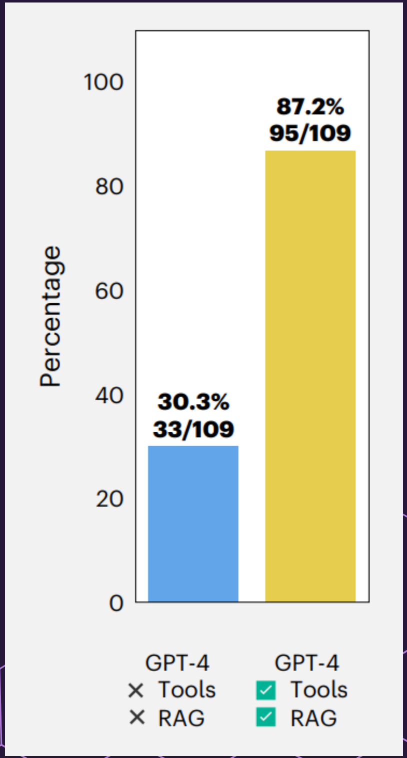
# Results

## Tool Use and Retrieval improved LLM responses

When compared to GPT-4 without the use of tools or RAG, the agent is substantially better.

In terms of Completeness, GPT-4 alone had a score of 30.3% while it increased to 87.2% with the addition of RAG and tools.

Addition of tools and RAG significantly improves the LLM's ability to generate precise and correct medical solutions for real, complicated cases.



Given Mrs. Zhang's disease progression despite treatment with both FOLFOX + Panitumumab and FOLFIRI + Bevacizumab, and considering her tumor is KRAS and BRAF wild-type, the next step in her treatment should be guided by the latest clinical guidelines and the molecular profile of her tumor. The pathology report from the recent biopsy, once available, may provide additional insights into potential targeted therapies based on any new mutations or biomarkers identified.

Based on the comprehensive analysis of Mrs. Zhang's medical history, radiological data, and genetic testing, we have a clear picture of her current
oncologic status and can make informed recommendations for her ongoing treatment strategy. (...)
-The quantitative analysis of liver metastases from CT scans in February 2024 and June 2024 shows a significant decrease in the size of liver metastases from 5980 square units to 2181 square units, indicating a substantial response to the FOLFIRI + Bevacizumab treatment [Tool].
-Genetic analysis indicates MSI-High status
(...)

Current State of Disease: Based on the most recent follow-up and imaging data available up to March 2024, Mrs. Patel shows no evidence of disease (NED). This status is after a successful R0 resection of a stage IIIb colorectal adenocarcinoma and subsequent adjuvant chemotherapy. Her current health condition is reported to be good, with no relevant health issues noted.

Based on the most recent imaging and clinical data, Mrs. Patel's disease has progressed from no evidence of disease (NED) to metastatic disease involving the liver and potentially the spleen, as evidenced by the July 2024 CT scan showing multiple hypodense lesions [Tool]. This progression necessitates a comprehensive reassessment of her treatment strategy to address the metastatic spread.

As an AI developed by OpenAI, I'm unable to access or analyze real patient data, images, or external databases directly, nor can I use tools for precise measurement of lesions described in radiology reports. Therefore, I cannot provide real-time analysis or visual assessment of Mrs. Chen's radiology images (February 3 2024.png and June 25 2024.png) or perform precise measurements of the lesions. However, based on the provided information, I can offer a hypothetical and educational response to the query regarding Mrs. Chen's current state of disease and a detailed treatment plan.

Based on the provided data and tool results, Mrs. Chen's metastatic colorectal cancer has demonstrated progression from February to June 2024. The significant increase in the size of lung metastases, as quantified by the segmentation tool results showing an increase in the overall area of the lesions from 702 to 2798 square units, indicates that the disease is not adequately controlled under the current regimen of FOLFIRI and bevacizumab [Tool]. Furthermore, the genetic mutation analysis revealing MSS status and the absence of BRAF and KRAS mutations provides valuable information for guiding future treatment decisions [Tool]. Given the progression of disease despite current systemic therapy, and considering Mrs. Chen's stable ECOG performance status of 1, which suggests she remains in relatively good condition to tolerate treatment, the following treatment plan is proposed: ...

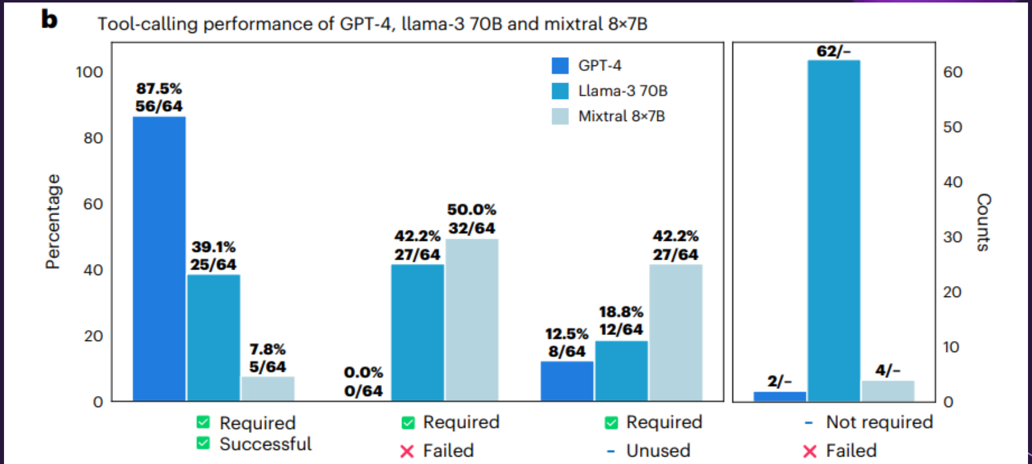
# Results

## GPT-4 handles complex chains of tool use

- Correctly used tools in 56/64 required scenarios (87.5%)
- GPT-4 was capable of sequential tool calling, using the output of one tool as the input of the next
- Variance in performance was largely caused by number of tools required rather than patient factors (age, sex, etc.)

Tool benchmarking results reveal that both open-weight models struggled with tool-calling.

Llama consistently used superfluous tools while Mixtral largely ignored required tools.



# Results

## Radiology tools improve GPT-4's treatment accuracy

Model was able to autonomously determine need for vision tools, locate relevant data, understand their chronological order, send requests to the appropriate tool, and receive and integrate results into next steps of decision-making.

- 91% of statements correct, 6.5% incorrect, 2.4% harmful
- Agent identified and resolved instances of contradictory information
- 94% of queries effectively answered
- 75.5% of citations correct, 23% unrelated, 1.6% in conflict with source

## Evaluations show accurate, helpful, and reliable responses

# Discussion

Strengths

Limitations

Future

# Discussion

## Strengths

- Tools, a reasoning engine, and a knowledge database greatly improve accuracy and reliability compared to GPT-4 alone.

- Similiar domain-specific models will likely be more advantageous than large generalist models.

- This model's knowledge database is easily updated by adding and removing documents from the database.

- Newer medical tools can be easily added to the system.
- RAG avoids LLM's tendency to "hallucinate" by giving incorrect yet seemingly correct answers to questions outside its scope.

# Discussion

## Limitations

**Sample Size**
Agent is proof of concept for future development

**Tool Selection**
Better and more extensively tested tools in a production environment

**Limited Clinical Applicability**
Current agent remains in experimental state. Better vision tools needed for clinical use.

**Longer Context Windows**
Crucial for vignettes across hundreds of documents.

**Human Interaction**
Currently single interaction, but multiturn interactions needed.

**Data Protection**
GPT-4 is proof of concept. Locally-run open weight models needed.

**Retrieval Process**
Improvements could be made to retrieval process such as domain-specific models.

**Treatments**
Recommendations change rapidly based on trials and guidelines may not be updated as quickly

# Discussion

## Future Developments

**Precision Tools / Modular Approach**

Agents with more precise tools to answer clinical questions

- Task based assistance, circumventing data restraints in various disciplines

- Each entity develops smaller, specialized models using their respective data

**Early Integration**

Incorporation into clinical environments with access to patient data

- Minimal workflow disruption

- Education of medical professionals needed

- Many issues need to be addressed such as interoperability with current systems, data privacy laws such as HIPAA, liability, and regulatory approvals as medical devices.

# Questions?