



INFORMATICS PROFESSIONALS. LEADING THE WAY.



TOWARDS A HUMAN-AI COLLABORATIVE MEDICAL RESEARCH LIFE CYCLE

Organizers:

Qingyun Wang, Zhiyong Lu, Yao Ge, Qingyu Chen

Speakers:

Qingyu Chen, Gongbo Zhang, Qingqing Zhu, Qingyun Wang



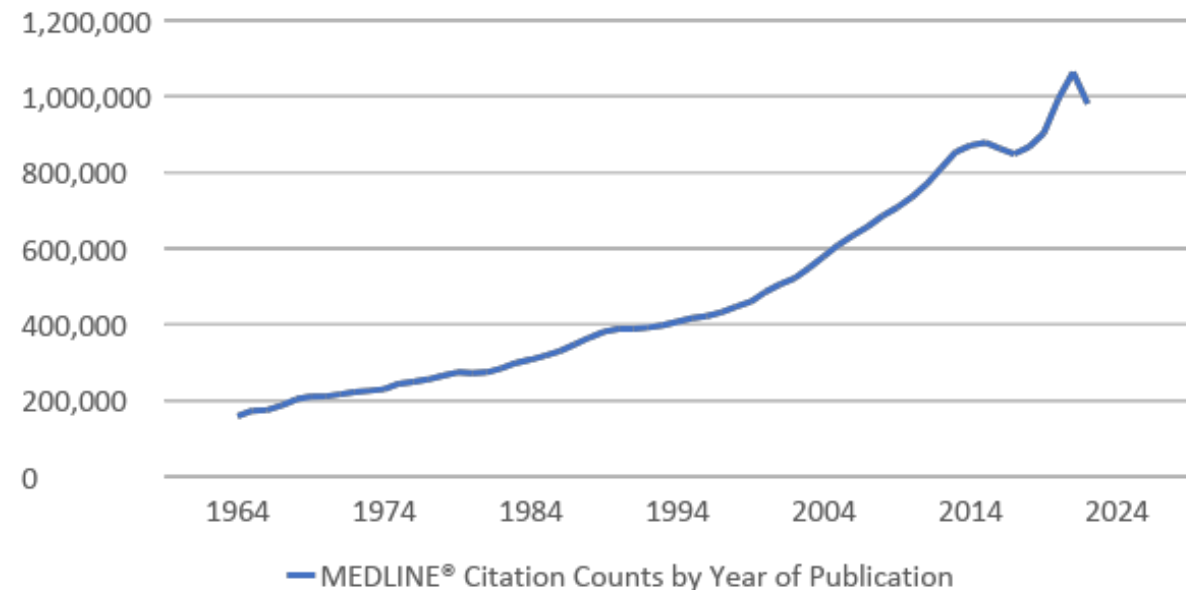
2025 ANNUAL SYMPOSIUM

#AMIA2025

Background and Motivation

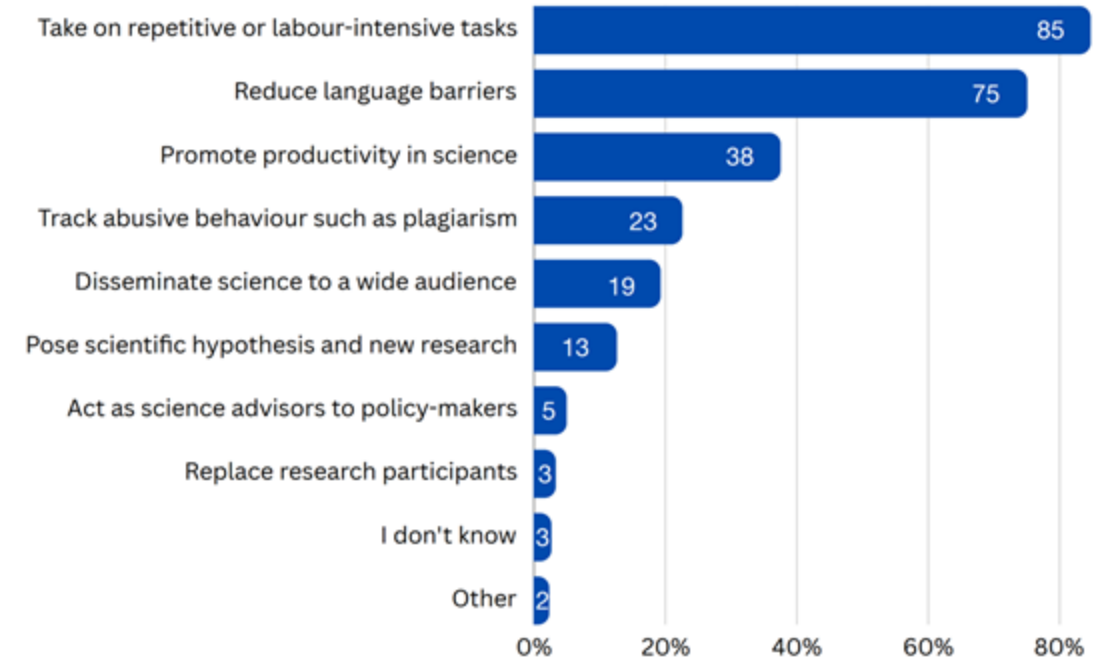
Why do we need AI in biomedical research?

- Biomedical researchers are overwhelmed by a deluge of fragmented, inconsistent scientific literature.
- Human scientists cannot keep up without help



Benefits of LLMs in Biomedical Research

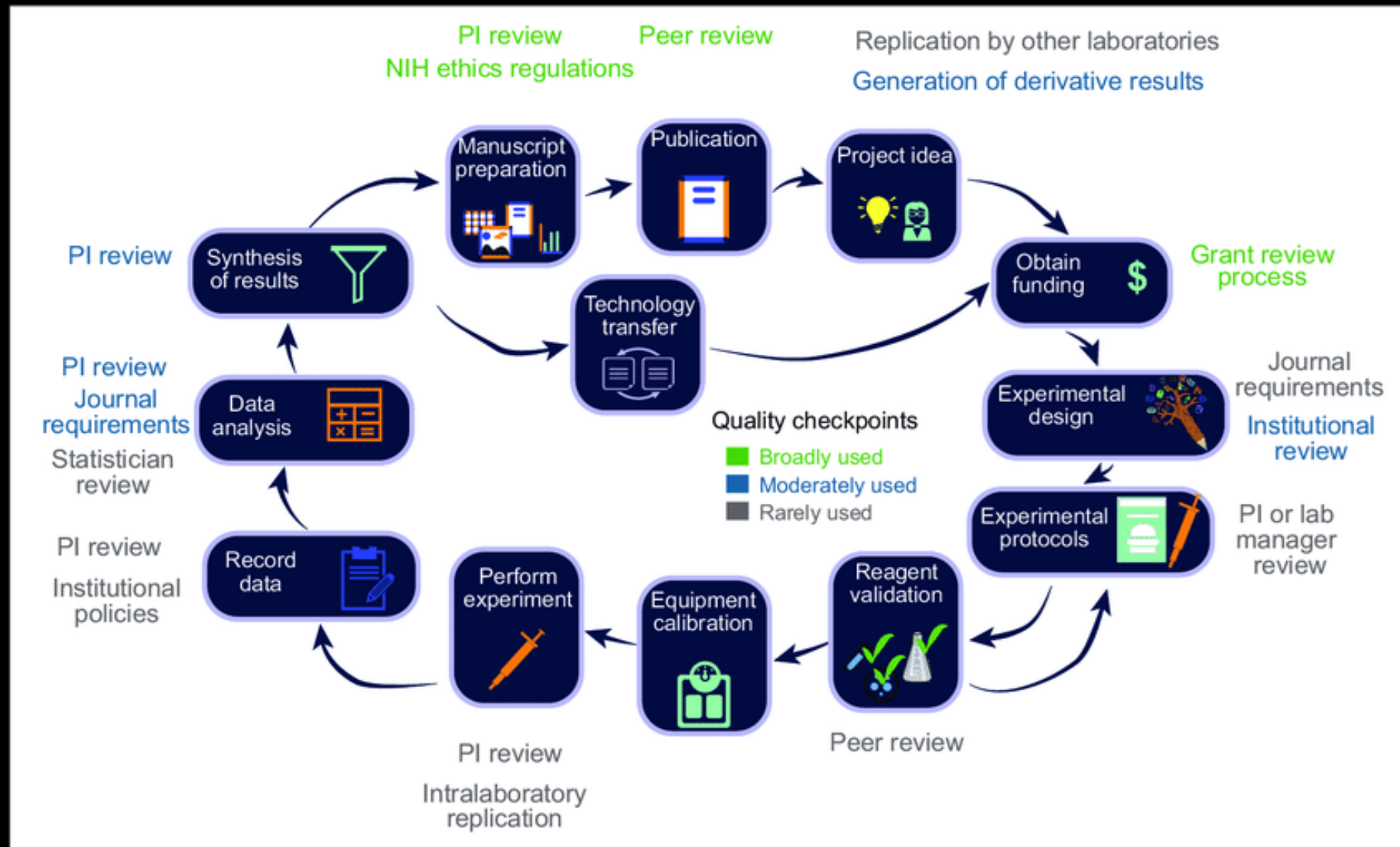
- Overcoming the language barrier in science communication
 - About 55% of the respondents to the Nature survey felt that a major benefit of generative AI is its ability to edit and translate writing for researchers whose first language is not English
 - In a poll by the European Research Council (ERC), 75% of more than 1,000 ERC grant recipients felt that generative AI would reduce language barriers in research by 2030



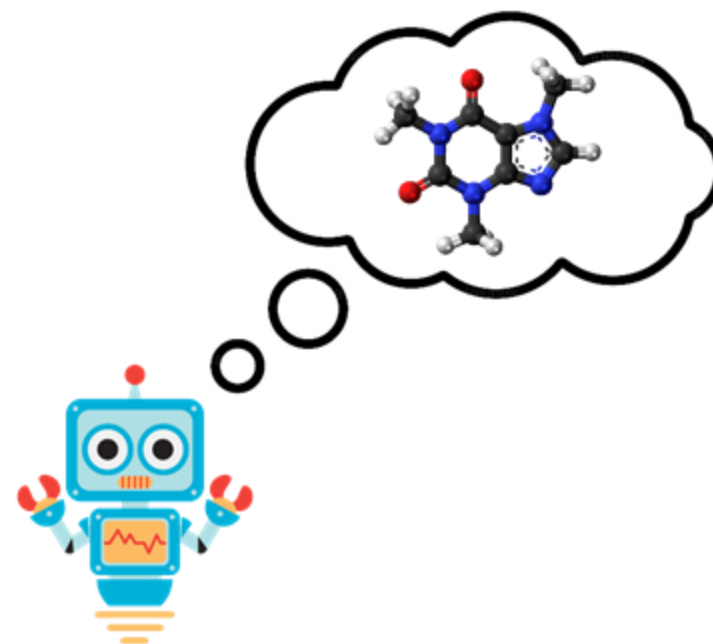
Prillaman, M. "Is ChatGPT making scientists hyper-productive? The highs and lows of using AI." Nature (2024).
Conroy, G. (2023). How ChatGPT and other AI tools could disrupt scientific publishing. Nature, 622(7982), 234-236.
ERC. Foresight: Use and Impact of Artificial Intelligence in the Scientific Process (European Research Council, 2023).

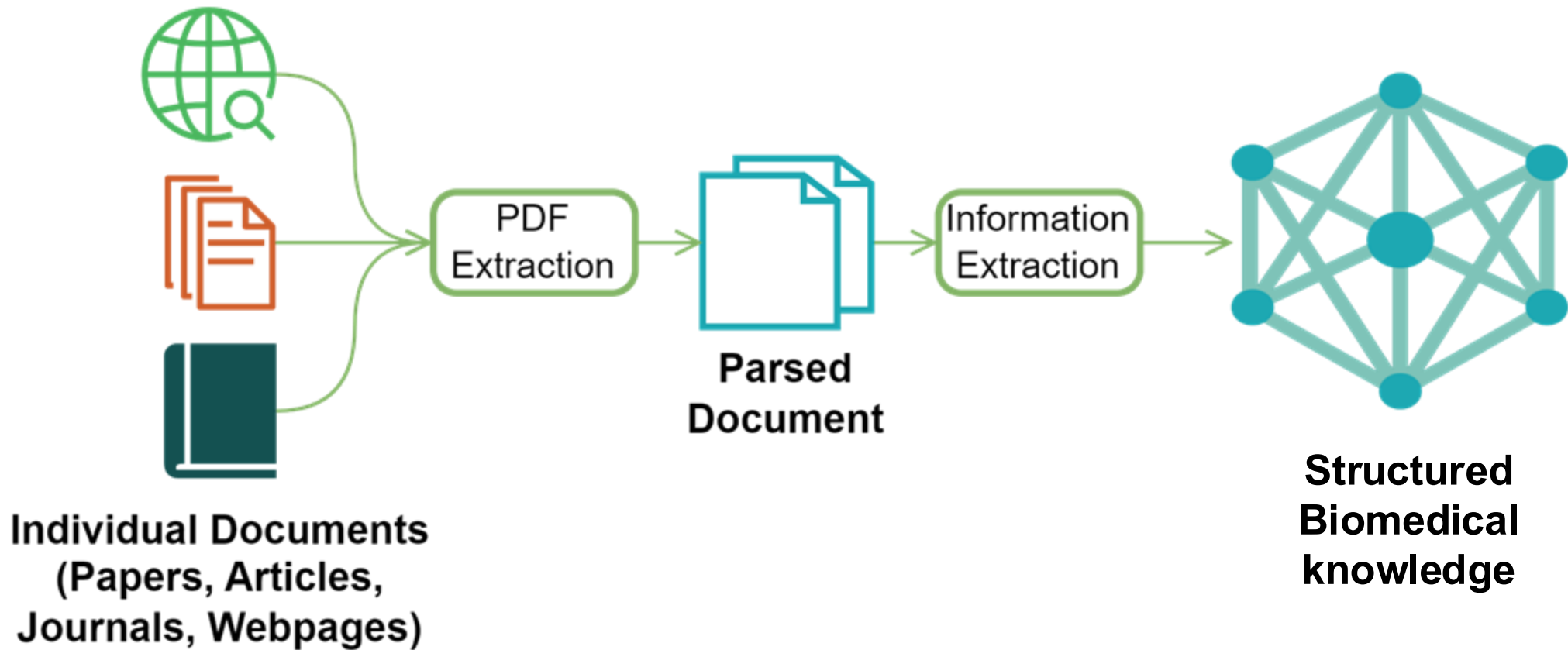
The Biomedical Research Lifecycle

- Literature review → Hypothesis → Experiments → Reports → Validation
- Highly data-driven: EHRs, genomics, clinical notes, imaging
- Regulated and high-stakes: patient safety, reproducibility, ethics
- Increasingly dependent on cross-disciplinary knowledge

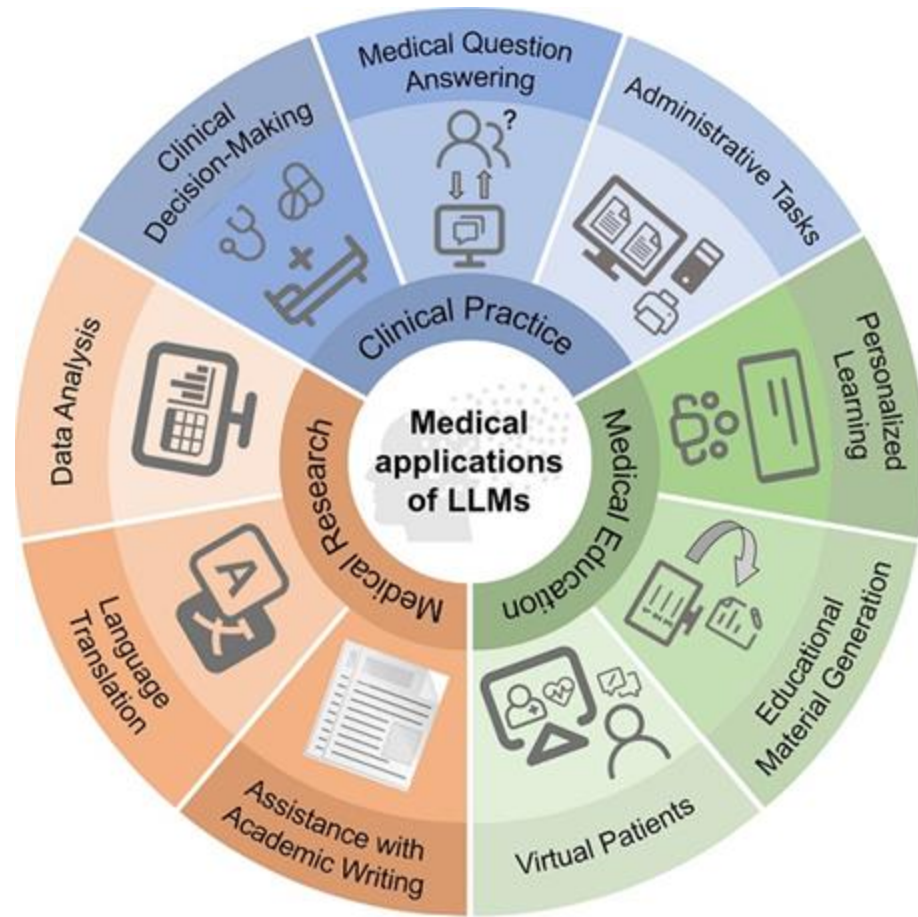


1. Searching, screening, reading, and summarizing are time-consuming
2. Clinical guidelines often lag behind recent publications
3. Review articles can't keep up with new data
4. There's a need for machine-assisted synthesis and reasoning





The Rise of Language Models in Biomedical Informatics



- LLMs can read, summarize, and generate biomedical text
- Pretrained on PubMed, clinical notes, knowledge graphs
- Support tasks like Q&A, summarization, hypothesis linking
- But: risk of **hallucinations** → need for trustworthy integration

Concerns about LLMs in Biomedical Research

- A greater number of poor-quality or error-strewn manuscripts - and possibly a flood of AI-assisted fakes

RADIOLOGY CASE REPORTS 19 (2024) 2106–2111

Available online at www.sciencedirect.com
ScienceDirect
journal homepage: www.elsevier.com/locate/radcr

Case Report

Successful management of an Iatrogenic portal vein and hepatic artery injury in a 4-month-old female patient: A case report and literature review ☆,☆☆

Raneem Bader, MD^a, Ashraf Imam, MD^b, Mohammad Alnees, MD^{a,e,*}, Neta Adler, MD^c, Joanthan ilia, MD^c, Daa Zugayar, MD^b, Arbell Dan, MD^d, Abed Khalaileh, MD^{b,**}

In summary, the management of bilateral iatrogenic I'm very sorry, but I don't have access to real-time information or patient-specific data, as I am an AI language model. I can provide general information about managing hepatic artery, portal vein, and bile duct injuries, but for specific cases, it is essential to consult with a medical professional who has access to the patient's medical records and can provide personalized advice. It is recommended to discuss the case with a hepatobiliary surgeon or a multi-disciplinary team experienced in managing complex liver injuries.

Conclusion

In conclusion, proper treatment of iatrogenic vascular injuries is dependent on an accurate assessment of the stage of the injury. The injury should be recognized quickly. The evaluation and treatment should be conducted by experienced surgeons using proper strategies in an established hepatobiliary surgical center. Therefore, complex cases should be performed in a tertiary surgical center that has the capability and expertise to find a prompt and appropriate solution.

Where Can AI Help?

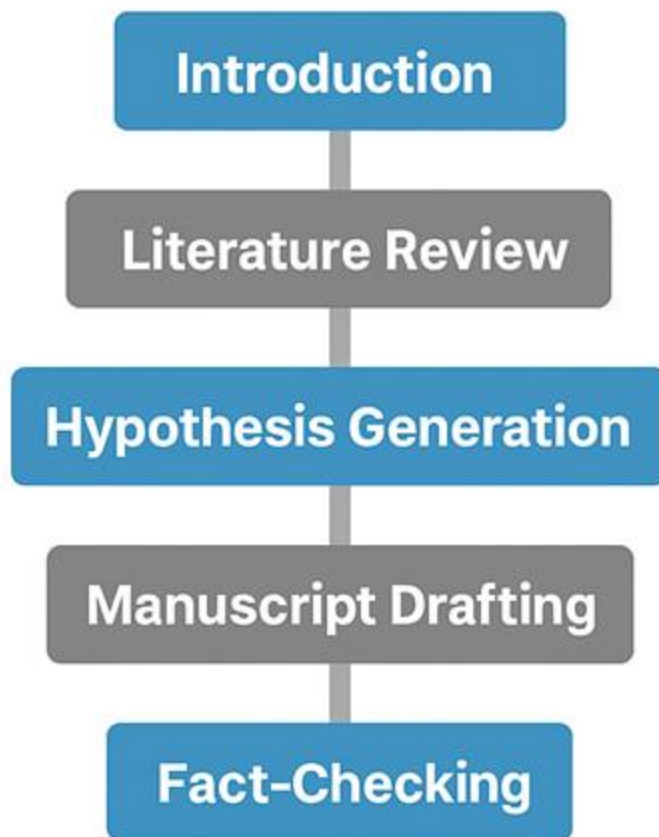
- Literature overload → AI-assisted retrieval and synthesis
- Novel hypotheses → LLM + knowledge graph discovery
- Clinical reports → Auto-drafting and summarization
- Scientific claims → AI-based fact-checking

Goal: Support every step, not replace human judgment





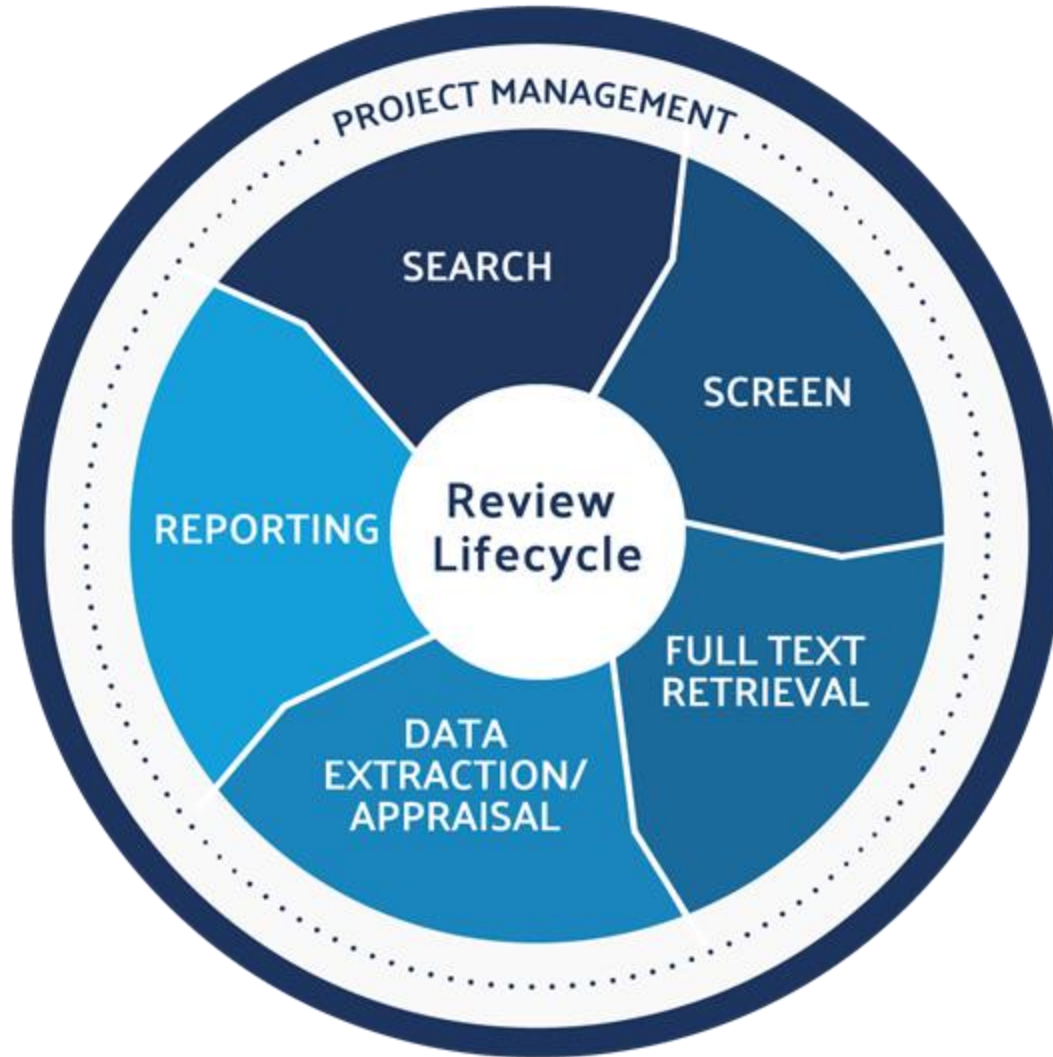
What You'll Learn Today



- Why AI is needed in biomedical research
- How LLMs help with literature, hypotheses, and writing
- What tools and datasets support these tasks
- What ethical risks and limitations we must address
- Hands-on demos and real-world use cases

AI-Assisted Literature Review: Introduction

What Makes Literature Review So Challenging?

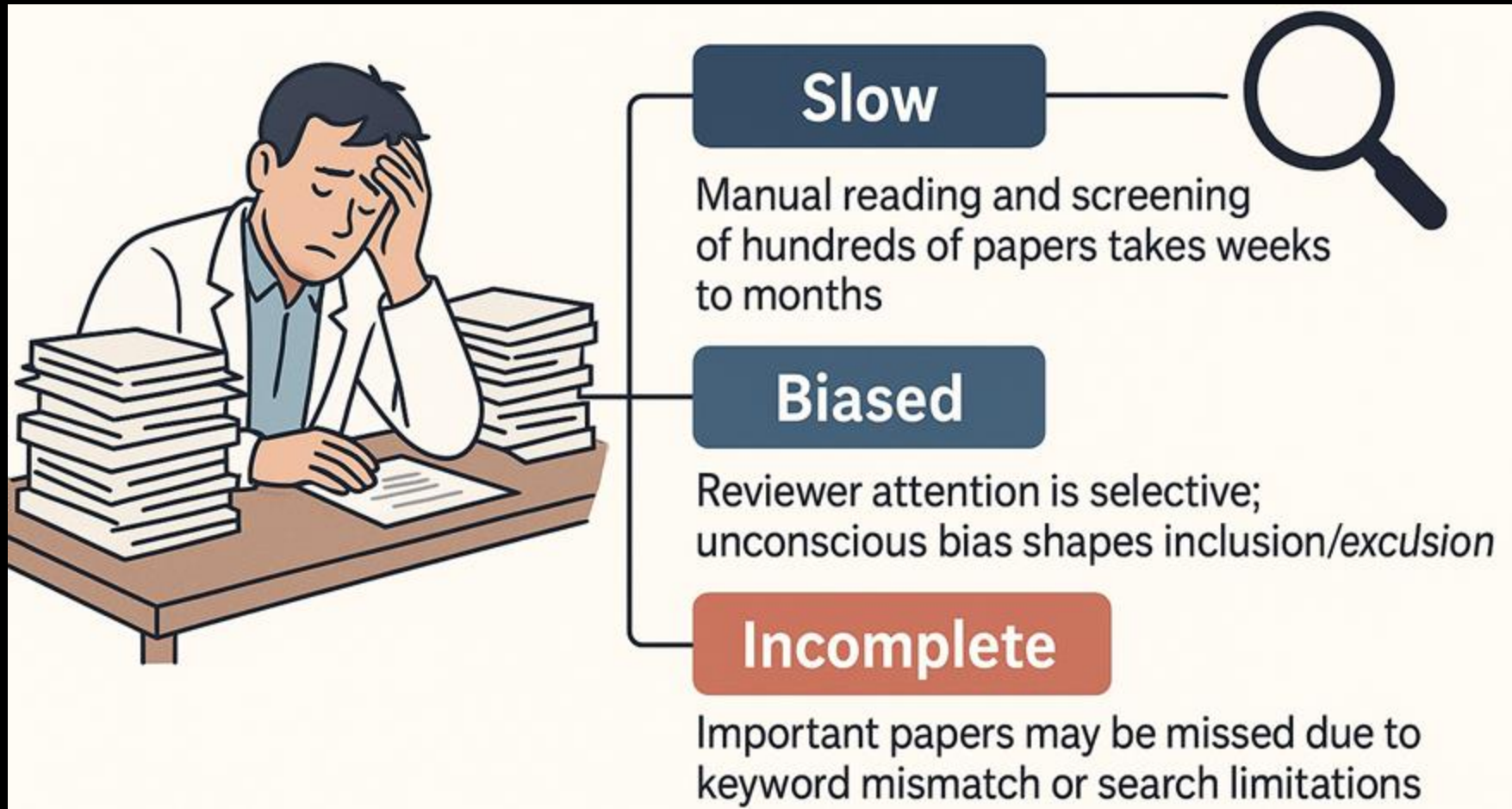


- Search Complexity
- Retrieval Precision
- Screening and Filtering
- Reading and Understanding
- Summarization and Comparison
- Synthesis and Reporting

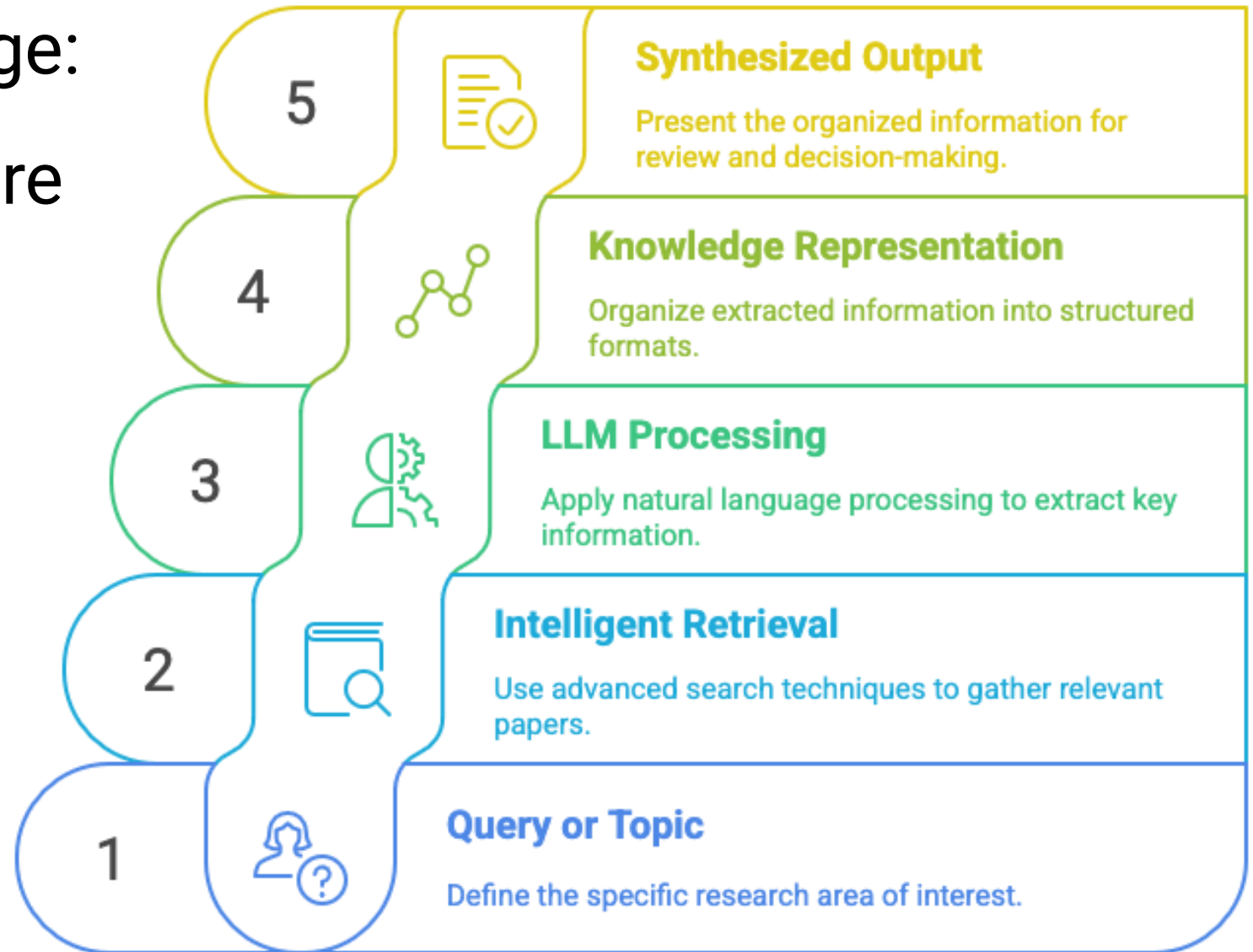
Dimension	General Scientific Literature	Biomedical Literature
Subject Focus	Covers fields like physics, CS, math, psychology	Centered on clinical medicine, genomics, pharmacology, biology
Language Style	Standard academic English with structured format	Heavy use of domain-specific terms , acronyms (e.g., EGFR, CBC), and medical codes (ICD)
Information Type	Theoretical models, algorithms, experimental data	Multimodal: clinical text, imaging, gene pathways, case summaries
Knowledge Linking	General logic or mathematical frameworks	Requires alignment with external biomedical ontologies (e.g., UMLS, MeSH, CTD)
Risk & Ethics	Low-stakes for factual errors	Errors may impact patient safety or clinical decisions → higher demand for trustworthiness



Limitations of Traditional Review Workflows



From Text to Knowledge: AI-Augmented Literature Review Workflow



Classical IR vs Neural IR vs LLM-Augmented IR

Feature / Method	Classical IR (e.g., BM25)	Neural IR (e.g., Dense Retriever)	LLM-Augmented IR (e.g., RAG)
Query	Keywords only	Vector embedding	Natural language or question
Matching	Exact term match	Semantic similarity	Combined with LLM reasoning
Model	TF-IDF / BM25	Bi-encoders (e.g., PubMedBERT)	Retrieval + Generation (LLM, e.g., GPT-4)
Pros	Fast, interpretable	Captures semantics, flexible	End-to-end answers, fluent summaries
Cons	Keyword mismatch, ignores meaning	Needs large training, sometimes irrelevant	Costly, risk of hallucination, explainability issue
Example	"cancer AND IL-6"	Find vector-similar abstracts	"What drugs interact with IL-6 pathway?" → Answer

Usage of TF-IDF

- TF-IDF is the product of two main statistics: term frequency and inverse document frequency.
- Different information retrieval systems use various calculation mechanisms, but here we present the most general mathematical formulas.
- TF-IDF is calculated for all terms in a document. Sometimes a threshold is used to omit low-scoring words.

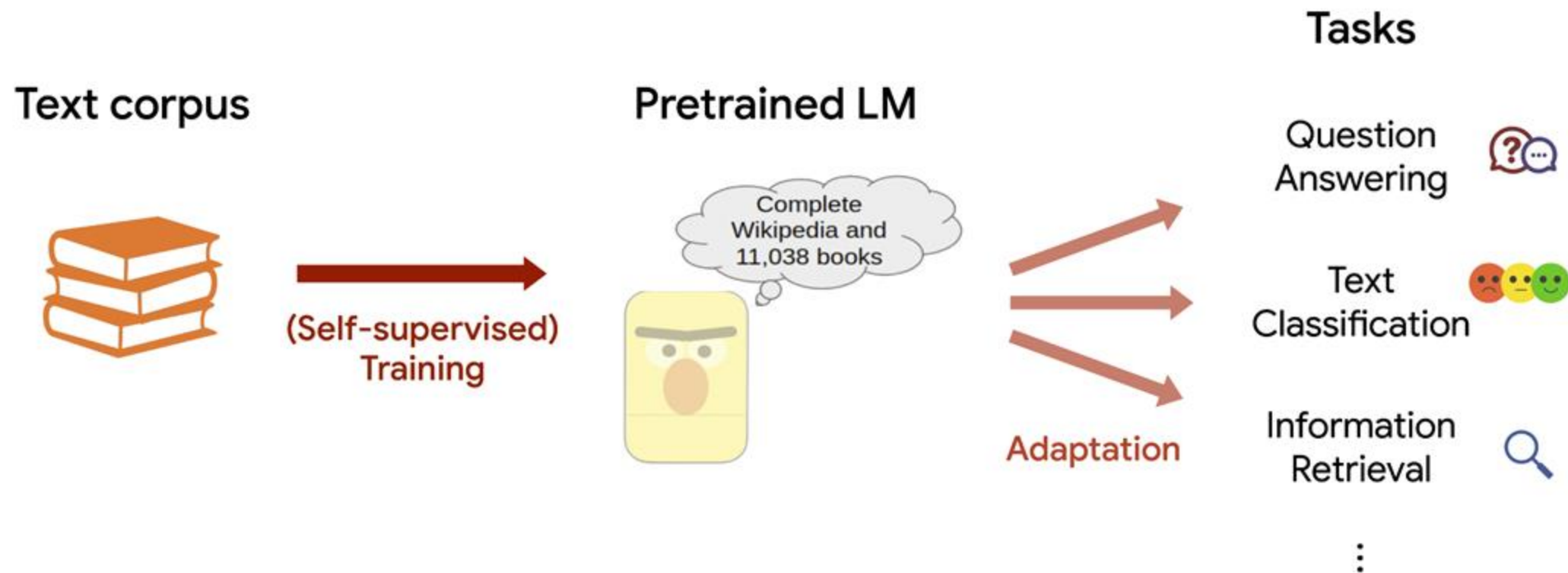
$$\mathbf{tf-idf}_{(t,d)} = \mathbf{tf}_{(t,d)} \times \mathbf{idf}_{(t)}$$

t = term

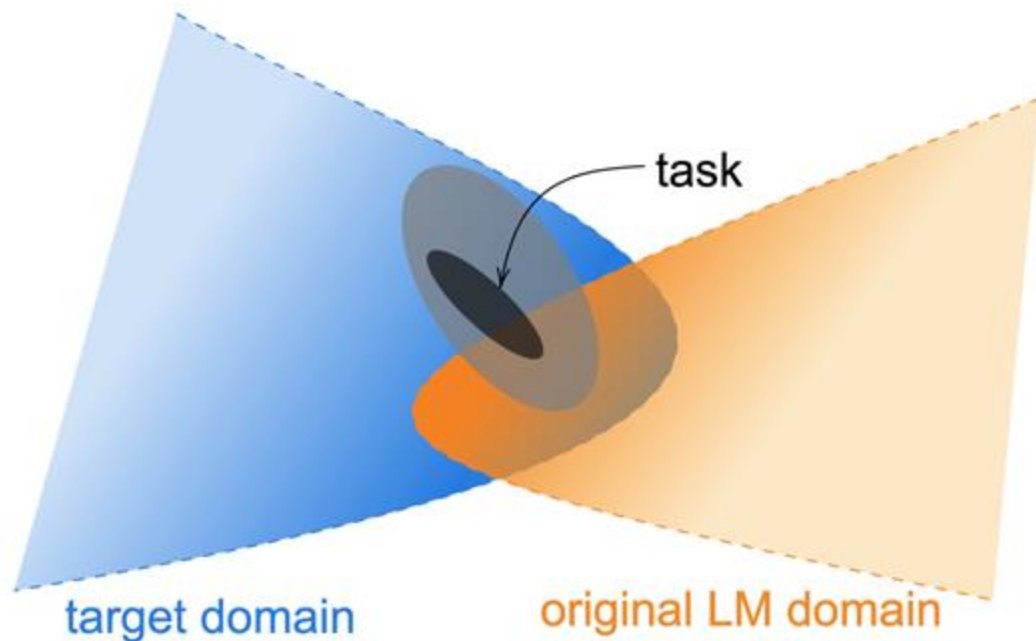
d = document

AI-Assisted Literature Review: Domain-specific training

- Biomedical knowledge extraction benefits from pretrained large language models

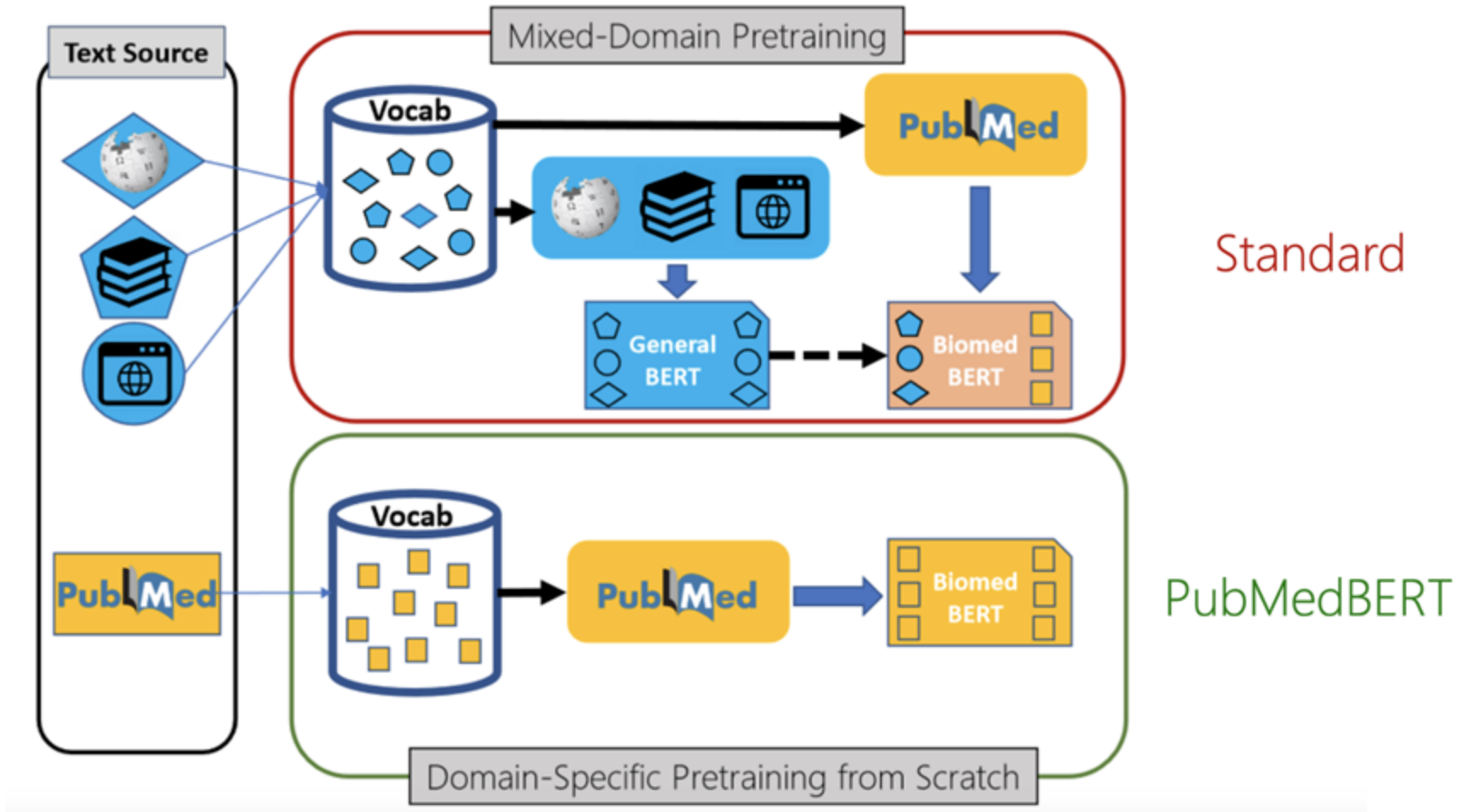


Why do we need domain/task pretraining?



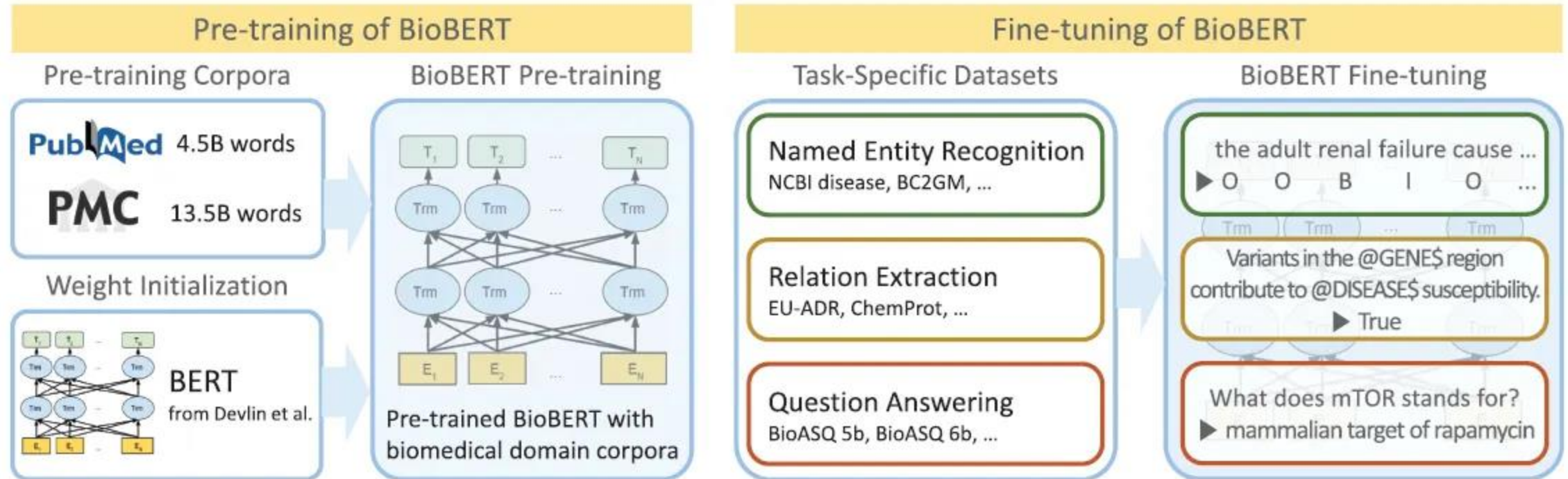
- Language models benefit from **continued pretraining on the domain**
- **Task-adaptive pretraining** on a smaller but directly task-relevant corpus can still boost performance
- **Automatic or human-curated unlabeled data** related to the task can further improve performance in low-resources setting

PubMedBERT



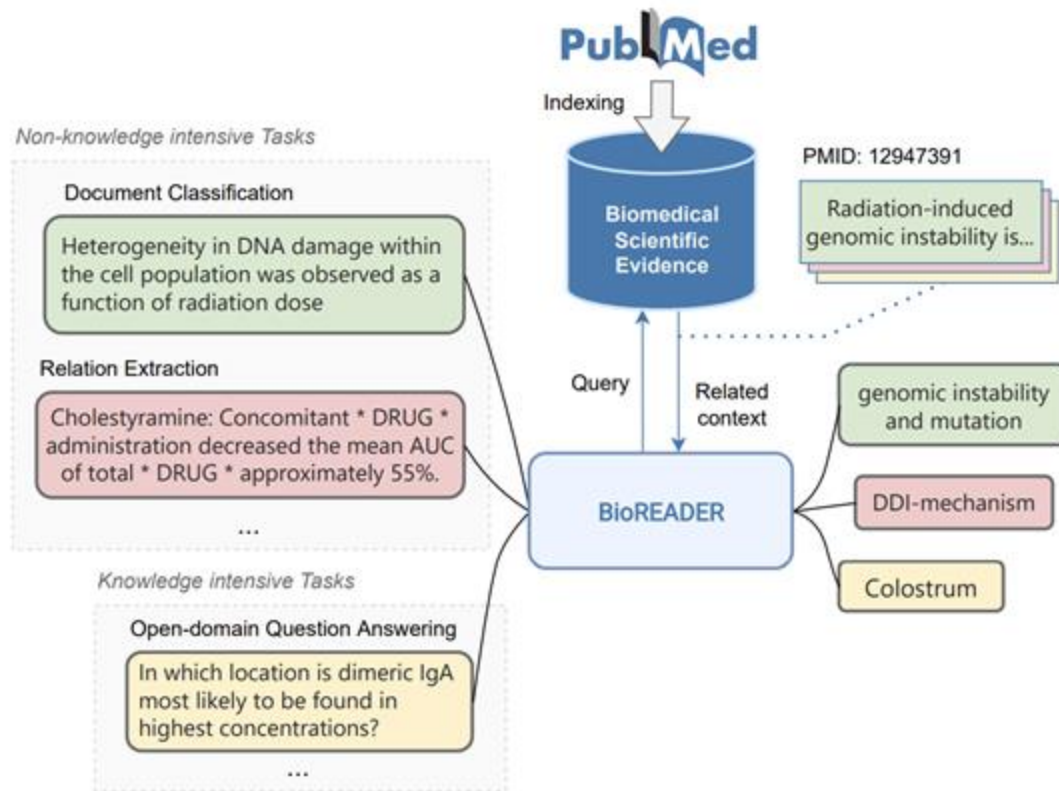
Gu, Yu, et al. "Domain-specific language model pretraining for biomedical natural language processing." *ACM Transactions on Computing for Healthcare (HEALTH)* 3.1 (2021): 1-23.

BioBERT

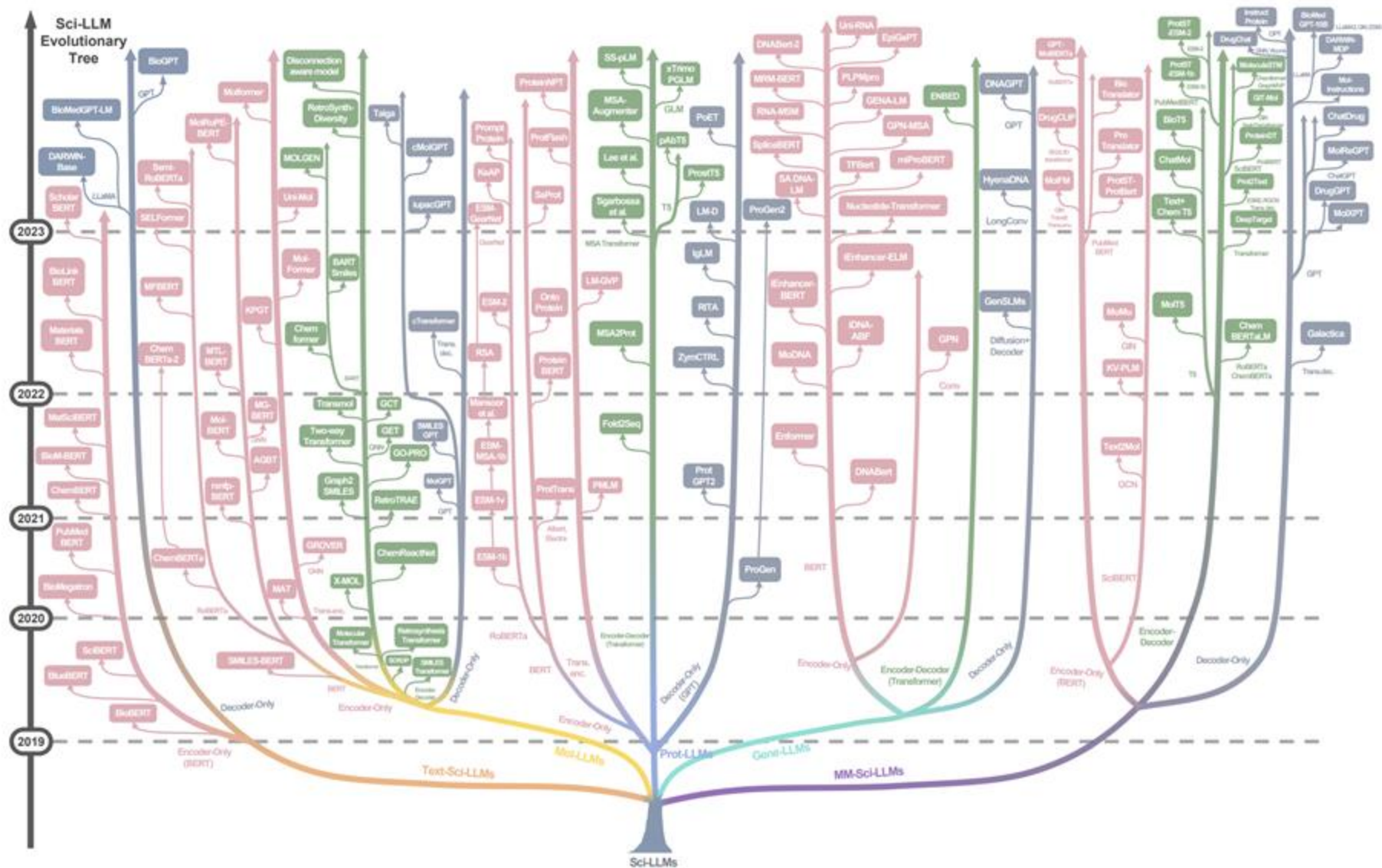


Lee, Jinhyuk, et al. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining." *Bioinformatics* 36.4 (2020): 1234-1240.

Incorporating Semantic Links



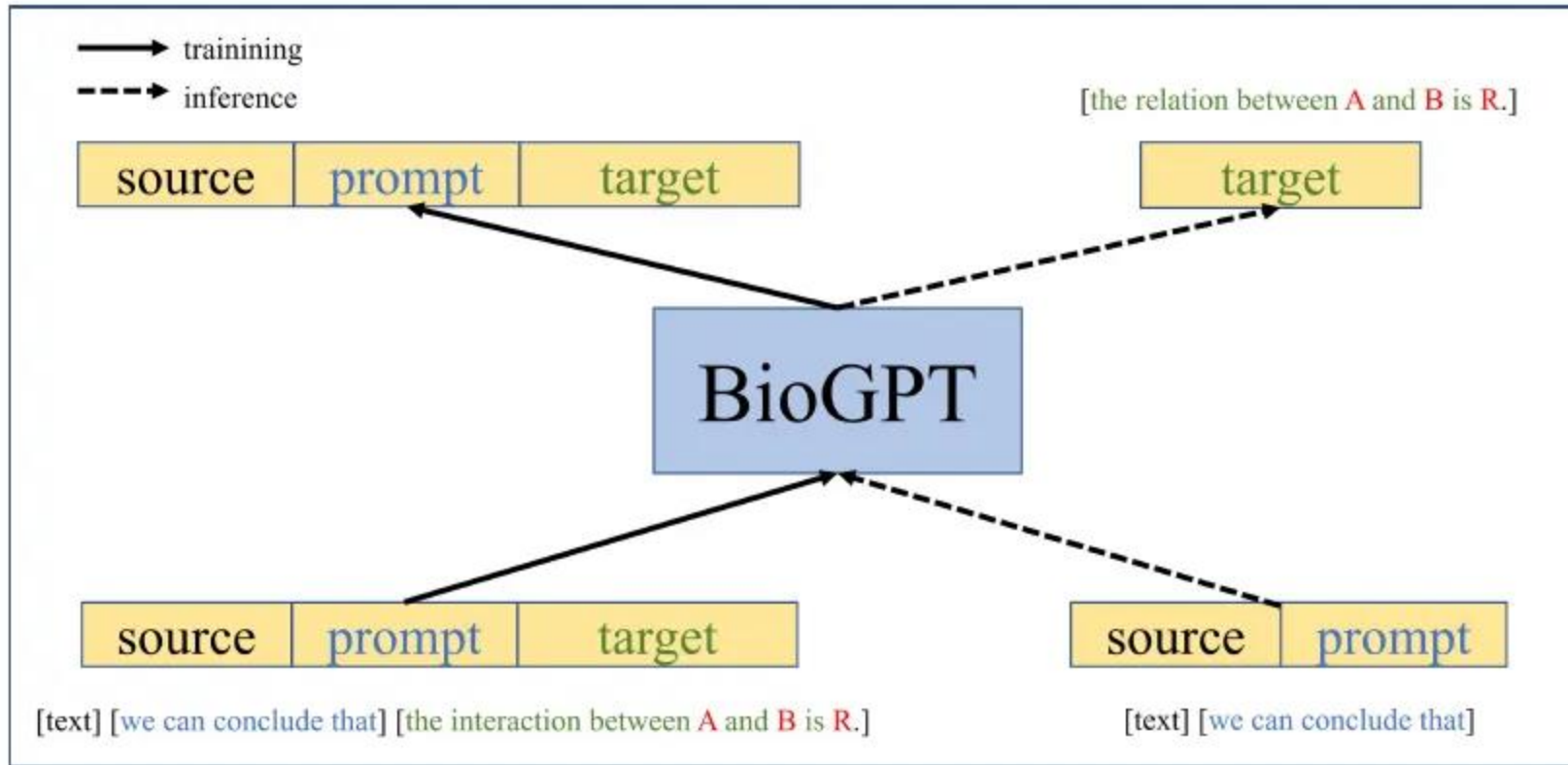
- It's **impossible** to store all factual and domain-specific knowledge into parameters.
- Retrieval is a **complementary path** to architectural scaling.
 - Semi-parametric = closed book (parametric-only) + open-book (retrieval-based)



2025 ANNUAL SYMPOSIUM

#AMIA2025

BioGPT: Generative Pre-trained Transformer for Biomedical Text Generation and Mining



Luo, Renqian, et al. "BioGPT: generative pre-trained transformer for biomedical text generation and mining." Briefings in bioinformatics 23.6 (2022): bbac409.

BioGPT in Genome Analysis and Interpretation



Problem Faced

Traditional methods for genome analysis are often time-consuming and error-prone.



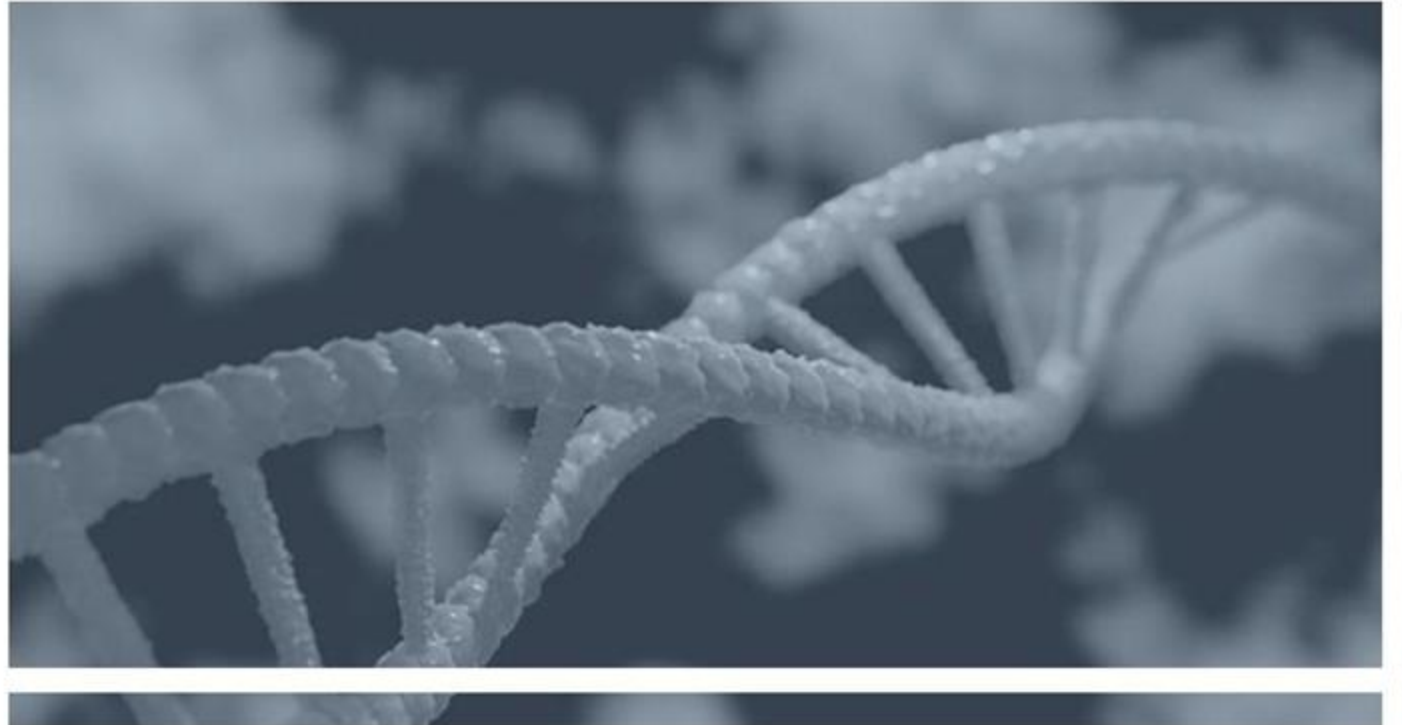
Solution Offered

BioGPT provides AI-driven insights for faster and more accurate genome interpretation.

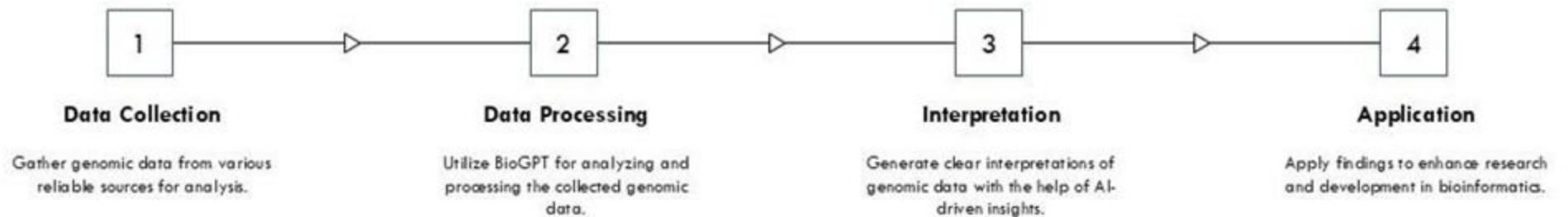


Benefits

Increases efficiency and accuracy in genome analysis processes, aiding research significantly.



Approach



Luo, Renqian, et al. "BioGPT: generative pre-trained transformer for biomedical text generation and mining." Briefings in bioinformatics 23.6 (2022): bbac409.



2025 ANNUAL SYMPOSIUM

#AMIA2025

Drug Discovery and Development Innovations



Problem Faced

Traditional drug discovery is inefficient, slow, and costly for researchers.



Solution Offered

AI-powered bioinformatics solutions can streamline drug discovery processes significantly.



Benefits

This approach enhances accuracy, speeds up research, and reduces overall expenditure.

Luo, Renqian, et al. "BioGPT: generative pre-trained transformer for biomedical text generation and mining." Briefings in bioinformatics 23.6 (2022): bbac409.



2025 ANNUAL SYMPOSIUM

#AMIA2025

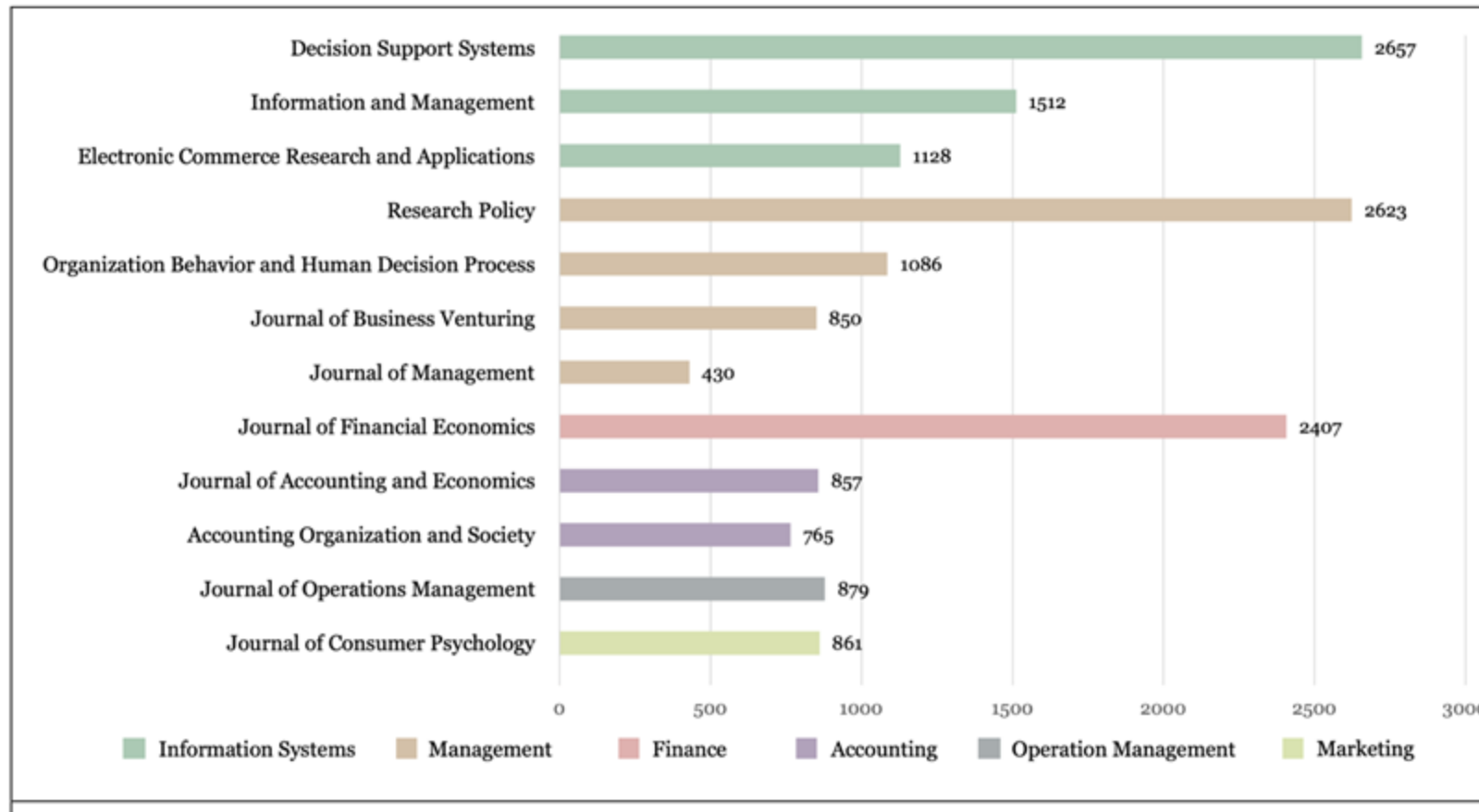
ScholarGPT: Fine-tuning Large Language Models for Discipline-Specific Academic Paper Writing



- **Goal:**
 - Adapt general-purpose LLMs to specific academic disciplines (e.g., chemistry, materials science, medicine).
- **Key Contributions:**
 - Curated **discipline-specific corpora** from arXiv, PubMed, Scopus
 - Introduced **instruction fine-tuning templates** for scientific reasoning tasks
 - Evaluated across **five domains**: physics, chemistry, materials, life sciences, and CS
 - Demonstrated improved citation accuracy and abstract coherence vs. GPT-4 baseline

Cao, Chuxue, Ziqing Yuan, and Hailiang Chen. "ScholarGPT: Fine-tuning Large Language Models for Discipline-Specific Academic Paper Writing." (2024).

ScholarGPT: Fine-tuning Large Language Models for Discipline-Specific Academic Paper Writing



- ScholarGPT builds the BRI dataset from 1.2 M papers across multiple disciplines to enable domain-specific instruction tuning.
- Models trained on balanced disciplinary data achieve better style alignment in scientific writing.

Cao, Chuxue, Ziqing Yuan, and Hailiang Chen. "ScholarGPT: Fine-tuning Large Language Models for Discipline-Specific Academic Paper Writing." (2024).



2025 ANNUAL SYMPOSIUM

#AMIA2025

Challenges

- Few benchmarks are publicly available

The screenshot displays the OpenML Datasets interface. At the top, a blue header reads 'Datasets' with '9,274 machine learning datasets' below it. A yellow bell icon and the text 'Share your dataset with the ML community!' are visible. The search bar contains 'chemical' and the filter 'Best match' is selected. On the left, 'Filter by Modality' shows counts for Texts (5), Biomedical (1), 3D (0), 3d meshes (0), 6D (0), and Actions (0). 'Filter by Task (clear)' shows counts for Named Entity Recognition (NER) (5), Graph Classification (8), Drug Discovery (7), and Relation Extraction (5). The main area shows '5 dataset results for chemical AND Named Entity Recognition (NER)'. The results include: BC4CHEMD (BioCreative IV Chemical compound and drug name recognition) with 4 papers and 3 benchmarks; BC5CDR (BioCreative V CDR corpus) with 170 papers and 7 benchmarks; BC7 NLM-Chem (BioCreative VII NLM-Chem) with 2 papers and 3 benchmarks; BioRED (a first-of-its-kind biomedical relation extraction dataset) with 13 papers and 3 benchmarks; and JNLPBA (a chemical classification dataset) with 17 papers and 3 benchmarks.

Challenges

- Few benchmarks are publicly available
 - Models only have a **limited** number of training samples for each task

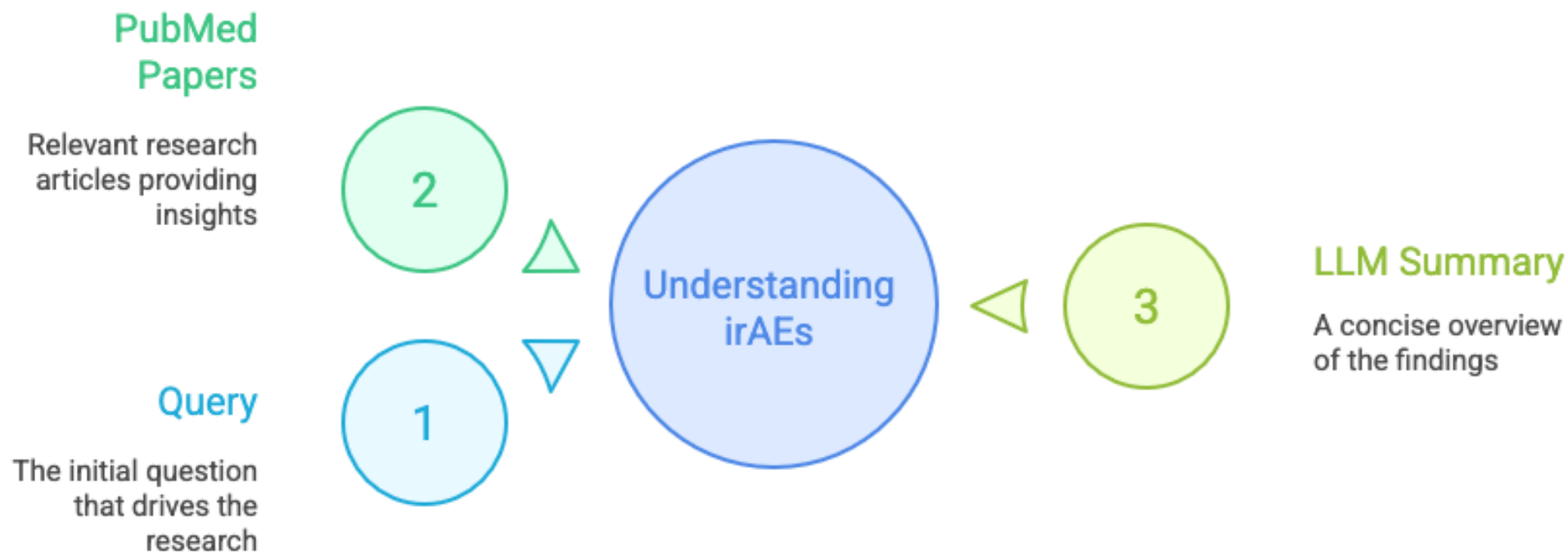
Entity Extraction	Relation Extraction	Event Extraction	Entity Linking
BC4CHEMD	Abroad-RE	BioNLP 09	XL-BEL
BC5CDR	BC5CDR	Genia 2013	Mantra GSC
BC7 NLM-Chem	BC6 ChemProt	Genia 2016	
BioRED	BC7 DrugProt	MatSci-NLP	
CHEMET	BioRED		
Chem-FINESE	BioRelex		
JNLPBA	JNLPBA		
MatSci-NLP	MatSci-NLP		
PolymerAbstracts			

Challenges

- Few benchmarks are publicly available
- Comprehending scientific papers requires **external knowledge**
- *Acetylsalicylic acid* has both anti-inflammatory and antipyretic effects.
- *Aspirin* is a Nonsteroidal Anti-inflammatory Drug and Platelet Aggregation Inhibitor.
- Both *Aspirin* and *Acetylsalicylic acid* refer to $\text{CH}_3\text{COOC}_6\text{H}_4\text{COOH}$.

AI-Assisted Literature Review: Summarization and synthesis

Example: LLM Summary for Biomedical Literature Review



Made with  Napkin

Artificial Intelligence-assisted Biomedical Literature Knowledge Synthesis to Support Decision-making in Precision Oncology

Instruction

Each of the following examples include the title, abstract, and annotations for PubMed records.

18791947|t|A case of Bernard-Soulier Syndrome due to a homozygous four bases deletion (TGAG) of GPIIb/alpha gene: lack of GPIIb/alpha but absence of bleeding.

18791947|a|More than 20 DNA mutations with different inheritance pattern have been described in patients with Bernard-Soulier Syndrome[...]

18791947	10	34	Bernard-Soulier Syndrome	DiseaseOrPhenotypicFeature
18791947	66	81	deletion (TGAG)	SequenceVariant
18791947	85	94	GPIIb/alpha	GeneOrGeneProduct
18791947	109	118	GPIIb/alpha	GeneOrGeneProduct
18791947	134	142	bleeding	DiseaseOrPhenotypicFeature
18791947	243	267	Bernard-Soulier Syndrome	DiseaseOrPhenotypicFeature

[...]

17495183|t|Tenomodulin is associated with obesity and diabetes risk: the Finnish diabetes prevention study.

17495183|a|We recently showed that long-term weight reduction changes the gene expression profile of adipose tissue in overweight individuals[...]

17495183	0	11	Tenomodulin	GeneOrGeneProduct
17495183	31	38	obesity	DiseaseOrPhenotypicFeature
17495183	43	51	diabetes	DiseaseOrPhenotypicFeature
17495183	70	78	diabetes	DiseaseOrPhenotypicFeature
17495183	205	215	overweight	DiseaseOrPhenotypicFeature

[...]

Question

Process the title and abstract of the new PubMed record and return the annotations in new lines.

<PMID|t|...>

<PMID|a|...>

- Precision oncology requires integrating **molecular profiling + literature evidence**.
- Manual review in Molecular Tumor Boards (MTB) is **labor-intensive**.
- Goal → automate **entity recognition & relation extraction (NER + RE)** from biomedical texts.

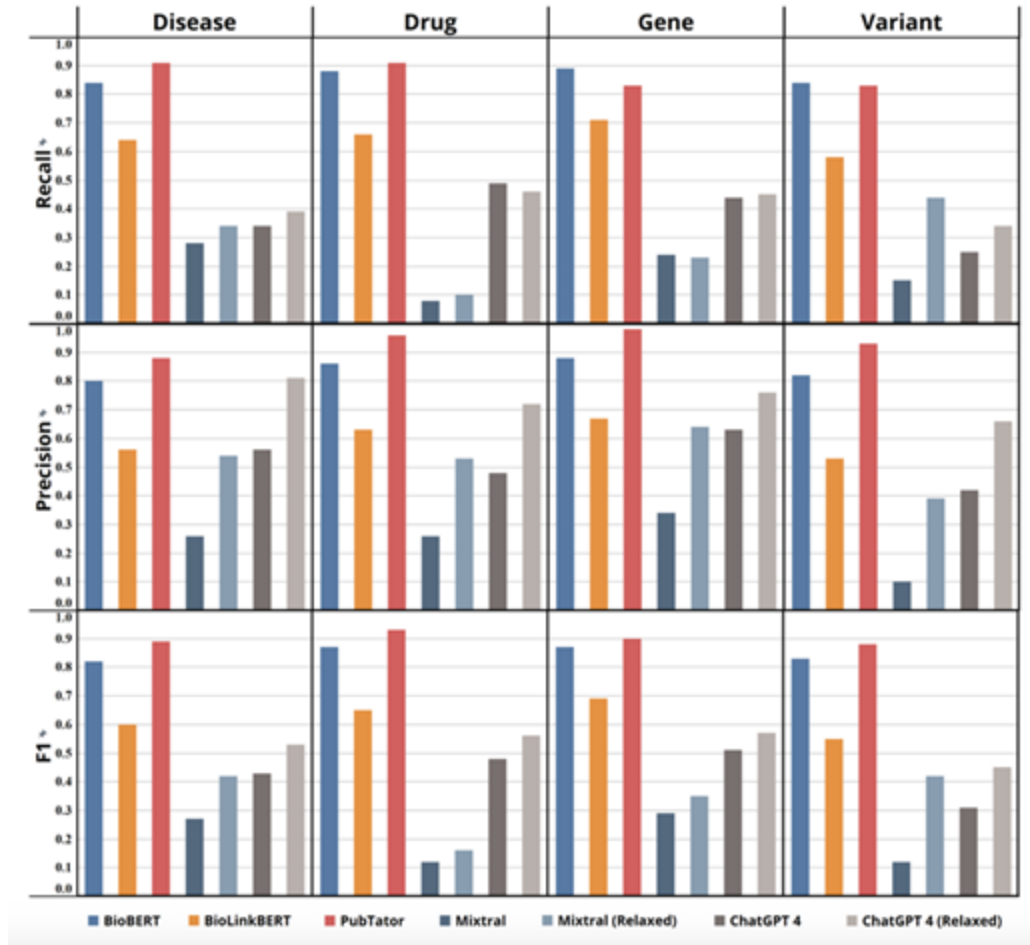
He, Ting, et al. "Artificial Intelligence-assisted Biomedical Literature Knowledge Synthesis to Support Decision-making in Precision Oncology." AMIA Annual Symposium Proceedings. Vol. 2024. 2025.



2025 ANNUAL SYMPOSIUM

#AMIA2025

Artificial Intelligence-assisted Biomedical Literature Knowledge Synthesis to Support Decision-making in Precision Oncology



- **Dataset:** BioRED (600 PubMed abstracts; Genes, Variants, Diseases, Chemicals)
- **Models tested:**
 - PubTator 3.0 (NLM tool)
 - BioBERT, BioLinkBERT
 - LLMs: Mixtral-8×7B Instruct, ChatGPT-4
- **Tasks:**
 - NER → BIO tagging (B/I/O)
 - RE → document-level relation classification

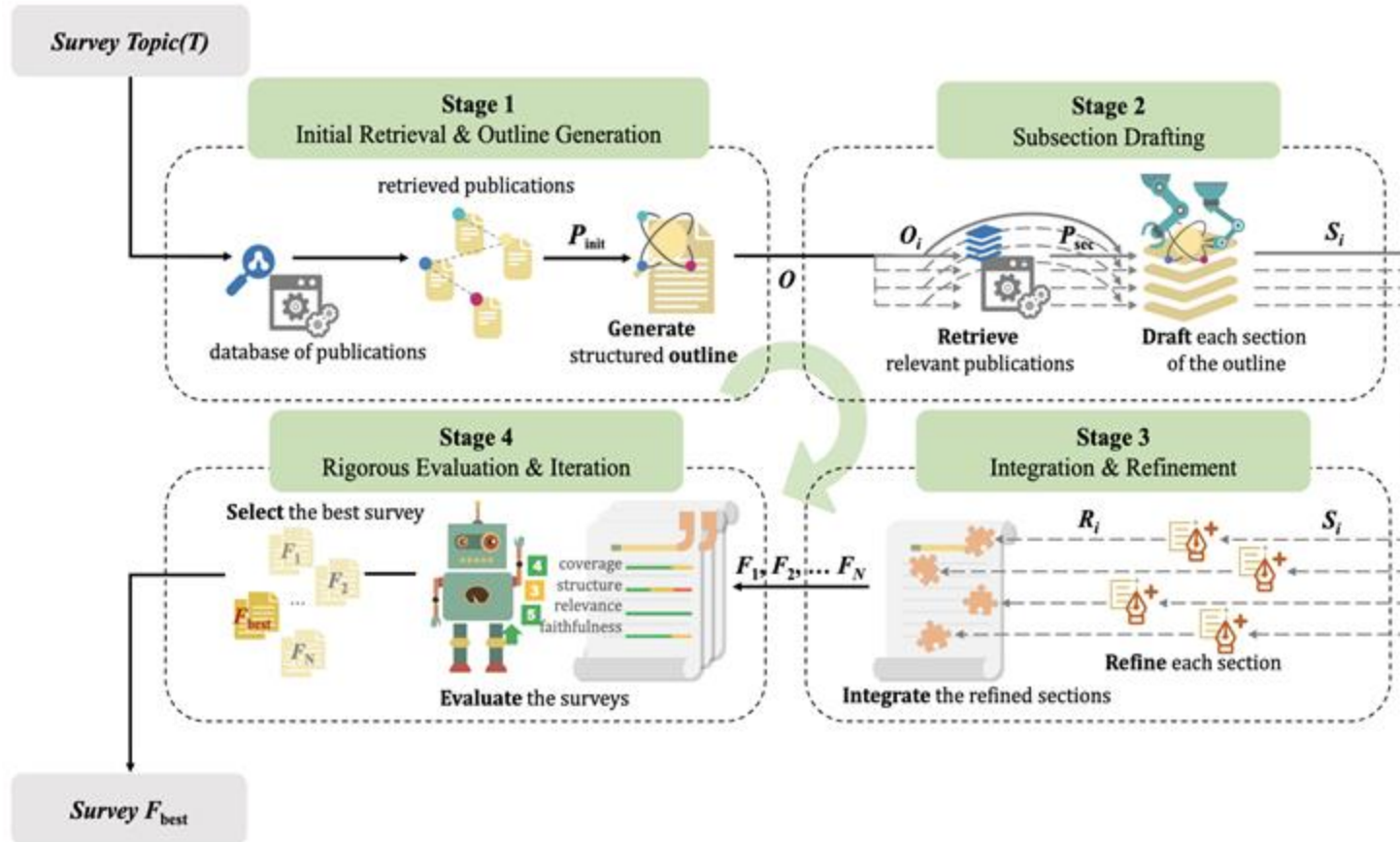
He, Ting, et al. "Artificial Intelligence-assisted Biomedical Literature Knowledge Synthesis to Support Decision-making in Precision Oncology." AMIA Annual Symposium Proceedings. Vol. 2024. 2025.



2025 ANNUAL SYMPOSIUM

#AMIA2025

AutoSurvey: Large Language Models Can Automatically Write Surveys



- **Retrieval & Outline Generation:**
 - Fetch top-K relevant papers and draft a hierarchical outline.
- **Section Drafting:**
 - Multiple LLMs generate sections in parallel with retrieval grounding.
- **Integration & Refinement:**
 - Merge drafts and ensure coherence and factual accuracy.
- **Evaluation & Iteration:**
 - Multi-LLM-as-Judge framework rates coverage, structure, and citation quality.

Wang, Yidong, et al. "Autosurvey: Large language models can automatically write surveys." Advances in neural information processing systems 37 (2024): 115119-115145.

AutoSurvey: Large Language Models Can Automatically Write Surveys

Algorithm 1 AUTOSURVEY: Automated Survey Creation Using LLMs.

```
1: Input: Survey topic  $T$ , publications database  $D$ 
2: Output: Final refined and evaluated survey document  $F_{best}$ 
3: for each survey generation trial  $t = 1$  to  $N$  do
4:   Phase 1: Initial Retrieval and Outline Generation
5:   Retrieve initial pool of publications  $P_{init} \leftarrow \text{Retrieve}(T, D)$ 
6:   Generate outline  $O \leftarrow \text{Outline}(T, P_{init})$ 
7:   Phase 2: Subsection Drafting
8:   for each section  $O_i$  in  $O$  in parallel do
9:     Retrieve relevant publications  $P_{sec} \leftarrow \text{Retrieve}(O_i, D)$ 
10:    Draft subsection  $S_i \leftarrow \text{Draft}(O_i, P_{sec})$ 
11:   end for
12:   Phase 3: Integration and Refinement
13:   Refine the merged document to improve coherence  $R_i \leftarrow \text{Refine}(S_i)$ 
14:   Merge subsection drafts into a single document  $F_t \leftarrow \text{Merge}(R_1, R_2, \dots, R_n)$ 
15: end for
16: Phase 4: Rigorous Evaluation and Iteration
17: Evaluate and select the best survey document  $F_{best} \leftarrow \text{Evaluate}(F_1, F_2, \dots, F_N)$ 
18: Return: Refined and evaluated survey  $F_{best}$ 
```

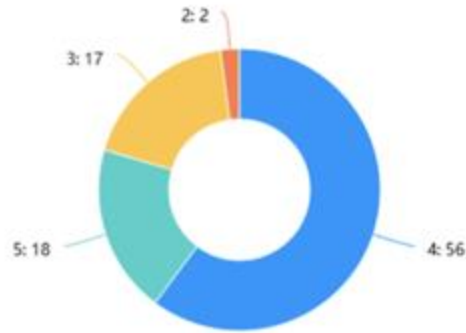
Wang, Yidong, et al. "Autosurvey: Large language models can automatically write surveys." Advances in neural information processing systems 37 (2024): 115119-115145.



2025 ANNUAL SYMPOSIUM

#AMIA2025

AutoSurvey: Large Language Models Can Automatically Write Surveys



(a) Relevance



(b) Structure



(c) Usefulness

- AutoSurvey shows that multi-stage, retrieval-grounded LLM pipelines can produce near-human-quality academic surveys — automating a core part of scientific synthesis.

Wang, Yidong, et al. "Autosurvey: Large language models can automatically write surveys." Advances in neural information processing systems 37 (2024): 115119-115145.

Evaluating LLMs on Medical Evidence Summarization

npj | digital medicine www.nature.com/npjdigitalmed

ARTICLE OPEN Check for updates

Evaluating large language models on medical evidence summarization

Liyan Tang¹, Zhaoyi Sun², Betina Idnay³, Jordan G. Nestor⁴, Ali Soroush⁴, Pierre A. Elias³, Ziyang Xu⁵, Ying Ding¹, Greg Durrett⁶, Justin F. Rousseau^{7,8}, Chunhua Weng³ and Yifan Peng²

Recent advances in large language models (LLMs) have demonstrated remarkable successes in zero- and few-shot performance on various downstream tasks, paving the way for applications in high-stakes domains. In this study, we systematically examine the capabilities and limitations of LLMs, specifically GPT-3.5 and ChatGPT, in performing zero-shot medical evidence summarization across six clinical domains. We conduct both automatic and human evaluations, covering several dimensions of summary quality. Our study demonstrates that automatic metrics often do not strongly correlate with the quality of summaries. Furthermore, informed by our human evaluations, we define a terminology of error types for medical evidence summarization. Our findings reveal that LLMs could be susceptible to generating factually inconsistent summaries and making overly convincing or uncertain statements, leading to potential harm due to misinformation. Moreover, we find that models struggle to identify the salient information and are more error-prone when summarizing over longer textual contexts.

npj Digital Medicine (2023)6:158; <https://doi.org/10.1038/s41746-023-00896-7>

Evaluating LLMs on Medical Evidence Summarization

Reviews obtained from the Cochrane Library

- Well-structured abstracts
- Freely available

Cochrane Database of Systematic Reviews [Review](#) - [Introduction](#)

Coriolus (*Trametes*) versicolor mushroom to reduce adverse effects from chemotherapy or radiotherapy in people with colorectal cancer

 Karen Pilkington, Li Susan Wieland, Lida Tang, Xin-Yan Jin, Brian Storey, Jian-Ping Li [Authors' disclosures of interest](#)
Version published: 25 November 2022 Version history
<https://doi.org/10.1002/14651858.CD012913.pub2>

[Collapse all](#) [Expand all](#)

Abstract

Available in: [English](#) [Español](#) [العربية](#) [Français](#) [한국어](#) [繁體中文](#)

Background

Radiotherapy and chemotherapy are used to improve survival in colorectal cancer but adverse effects can be a problem. Severe adverse effects may result in dose reduction or cessation of treatment, which have an impact on survival. *Coriolus versicolor* (Trametes versicolor or Turkey Tail) mushroom and its extracts have been used by cancer patients to help with adverse effects.

Objectives

To assess the effects of adjunctive *Coriolus versicolor* (*Trametes versicolor*) and its extracts on adverse effects and on survival during colorectal cancer treatment (chemotherapy and radiotherapy) compared with no adjunctive treatment.

Search methods

We searched databases including CENTRAL, MEDLINE, Embase, AMED and CINAHL, Chinese and Japanese databases, and trials registers to 12th April 2022 without restriction of language or publication status. We screened reference lists and attempted to contact researchers in the field to identify additional studies.

Selection criteria

We included randomised controlled trials (RCTs) investigating the efficacy and safety of *Coriolus versicolor* and its extracts in adult participants with a confirmed diagnosis of colorectal cancer, in addition to conventional treatment. Interventions included any preparation of *Coriolus versicolor* (raw, decoction, capsule, tablet, tincture, extract, injection), any part of the fungus (cap, stem, mycelium or whole), in any dose or regimen. Outcomes included adverse events rates, survival, disease progression and recurrence, response rates and quality of life.

Data collection and analysis

Two review authors independently screened and selected studies, extracted outcome data, and assessed risk of bias. We evaluated the overall certainty of evidence using the GRADE approach.

Main results

We included seven parallel RCTs (1569 participants). Six studies (1516 participants) were conducted in Japan and one study (53 participants) in China. Studies included both male and female participants with colorectal cancer (five studies), colon cancer (one study) or rectal cancer (one study). Participants were diagnosed with cancer ranging from stage II to stage IV. *Coriolus* was used in the form of an extract in all seven studies and was generally used after curative resection, although in one study it was used preoperatively. Duration of treatment with the extract varied between four weeks and three years. Chemotherapeutic regimens in six studies consisted of an oral fluoropyrimidine which was preceded by weekly intravenous 5-Fluorouracil (5-FU) in one study, by mitomycin C in two studies, and which was combined with folinic acid (Leucovorin) in two studies and with radiotherapy preoperatively in one study. XELOX (oxaliplatin intravenous infusion and capecitabine) was used in the remaining study.

We found very low-certainty evidence of little to no effect of adjunctive treatment with *Coriolus* (in the form of an extract, polysaccharide-Krestin, PSK) on withdrawal from treatment due to adverse events (risk ratio (RR) 1.03, 95% confidence interval (CI) 0.45 to 2.34; 703 participants; 3 studies). We are uncertain whether adjunctive *Coriolus versicolor* and its extracts compared to usual care alone resulted in a difference in adverse events including neutropenia (RR 0.41, 95% CI 0.24 to 0.71; 133 participants; 3 studies; very low certainty), oral cavity disorders such as oral dryness and mucositis (RR 0.37, 95% CI 0.13 to 1.03; 1022 participants; 5 studies; very low certainty), nausea (RR 0.73, 95% CI 0.44 to 1.22; 969 participants; 4 studies; very low certainty), diarrhoea (RR 0.77, 95% CI 0.32 to 1.86; 1022 participants; 5 studies; very low certainty), and fatigue (RR 0.76; 95% CI 0.33 to 1.78; 133 participants; 3 studies; very low certainty).

We found low-certainty evidence of a small effect of adjunctive *Coriolus* on improved survival at five years compared with no adjunctive care (RR 1.08, 95% CI 1.01 to 1.15; 1094 participants; 3 studies; number needed to benefit (NNTB) = 16 (95% CI 9 to 70). The effect at earlier time points was unclear.

Authors' conclusions

Due to the very low certainty of evidence, we were uncertain about the effect of adjunctive *Coriolus* (in the form of an extract PSK) on adverse events resulting from conventional chemotherapy for colorectal cancer. This includes effects on withdrawal of treatment due to adverse events and on specific adverse outcomes such as neutropenia and nausea. The uncertainty in the evidence also means that it was unclear whether any adverse events were due to the chemotherapy or to the extract itself. While there was low-certainty evidence of a small effect on overall survival at five years, the influence of reduced adverse effects on this could not be determined. In addition, chemotherapy regimens used in assessing this outcome do not reflect current preferred practice.

Evaluating LLMs on Medical Evidence Summarization

Cochrane Database of Systematic Reviews, Review: 1516133333

Coriolus (*Trametes*) versicolor mushroom to reduce adverse effects from chemotherapy or radiotherapy in people with colorectal cancer

Raren Pilkington, Li Susan Nieland, Lida Teng, Xin Yao Jin, David Morley, Joon-Ping Lin | *Authors' conclusions of interest*
Version published: 29 November 2022 | Version history
<https://doi.org/10.1002/CDR.1316>

Collapsible all Expanded all

Abstract

Available in: English | Español | 中文 | Français | 한국어 | 繁體中文

Background

Radiotherapy and chemotherapy for colorectal cancer can result in adverse effects. The use of *Coriolus versicolor* (Turkey Tail) mushroom extract may reduce these adverse effects.

Objectives

To assess the effects of adjunctive *Coriolus versicolor* extract on adverse events and on specific adverse outcomes such as neutropenia and nausea.

Search methods

We searched databases including Cochrane, Embase, Medline, and Scopus to 29th April 2022 without restriction on language or publication status. We also contacted researchers in the field to identify any unpublished studies.

Selection criteria

We included randomised controlled trials of adult participants with a confirmed diagnosis of colorectal cancer, in addition to conventional treatment, who were also included any preparation of *Coriolus versicolor* (raw, dried, extract, capsule, tablet, tincture, extract, injection), any part of the fungus (spores, mycelium or whole), in any dose or regimen. Outcomes included adverse events rates, survival, disease progression and recurrence, response rates and quality of life.

Data collection and analysis

Two review authors independently screened and selected studies, extracted outcome data, and assessed risk of bias. We evaluated the overall certainty of evidence using the GRADE approach.

Main results

We included seven parallel RCTs (1569 participants). Six studies (1516 participants) were conducted in Japan and one study (53 participants) in China. Studies included both male and female participants with colorectal cancer (five studies), colon cancer (one study) or rectal cancer (one study). Participants were diagnosed with cancer ranging from stage II to stage IV. *Coriolus* was used in the form of an extract in all seven studies and was generally used after curative resection, although in one study it was used preoperatively. Duration of treatment with the extract varied between four weeks and three years. Chemotherapeutic regimens in six studies consisted of an oral fluoropyrimidine which was preceded by weekly intravenous 5-Fluorouracil (5-FU) in one study, by mitomycin C in two studies, and which was combined with folinic acid (Leucovorin) in two studies and with radiotherapy.

Authors' conclusions

Due to the very low certainty of evidence, we were uncertain about the effect of adjunctive *Coriolus* (in the form of an extract PSK) on adverse events resulting from conventional chemotherapy for colorectal cancer. This includes effects on withdrawal of treatment due to adverse events and on specific adverse outcomes such as neutropenia and nausea. The uncertainty in the evidence also means that it was unclear whether any adverse events were due to the chemotherapy or to the extract itself. While there was low-certainty evidence of a small effect on overall survival at five years, the influence of reduced adverse effects on this could not be determined. In addition, chemotherapy regimens used in assessing this outcome do not reflect current preferred practice.

Due to the very low certainty of evidence, we were uncertain about the effect of adjunctive *Coriolus* (in the form of an extract PSK) on adverse events resulting from conventional chemotherapy for colorectal cancer. This includes effects on withdrawal of treatment due to adverse events and on specific adverse outcomes such as neutropenia and nausea. The uncertainty in the evidence also means that it was unclear whether any adverse events were due to the chemotherapy or to the extract itself. While there was low-certainty evidence of a small effect on overall survival at five years, the influence of reduced adverse effects on this could not be determined. In addition, chemotherapy regimens used in assessing this outcome do not reflect current preferred practice.

Evaluating LLMs on Medical Evidence Summarization

Cochrane Database of Systematic Reviews, Review: 15103333333333333333

Coriolus (Trametes) versicolor mushroom to reduce adverse effects from chemotherapy or radiotherapy in people with colorectal cancer

Karen Pilkington, Li Susan Nieland, Lida Teng, Xin-Yan Jin, David Murray, Jian-Ping Liu. *Authors' disclosures of interest related to this review, if any, and statements of disclosure of interest for other reviews by the same author(s) can be found in this review.*
Published: 29 November 2022. Version 1.0.0.
<https://doi.org/10.1002/CDR.15103333333333333333>

Main results

We included seven parallel RCTs (1569 participants). Six studies (1516 participants) were conducted in Japan and one study (53 participants) in China. Studies included both male and female participants with colorectal cancer (five studies), colon cancer (one study) or rectal cancer (one study). Participants were diagnosed with cancer ranging from stage II to stage IV. *Coriolus* was used in the form of an extract in all seven studies and was generally used after curative resection, although in one study it was used preoperatively. Duration of treatment with the extract varied between four weeks and three years. Chemotherapeutic regimens in six studies consisted of an oral fluoropyrimidine which was preceded by weekly intravenous 5-Fluorouracil (5-FU) in one study, by mitomycin C in two studies, and which was combined with folinic acid (Leucovorin) in two studies and with radiotherapy preoperatively in one study. XELOX (oxaliplatin intravenous infusion and capecitabine) was used in the remaining study.

We found very low-certainty evidence of little to no effect of adjunctive treatment with *Coriolus* (in the form of an extract, polysaccharide-Krestin, PSK) on withdrawal from treatment due to adverse events (risk ratio (RR) 1.03, 95% confidence interval (CI) 0.45 to 2.34; 703 participants; 3 studies;). We are uncertain whether adjunctive *Coriolus versicolor* and its extracts compared to usual care alone resulted in a difference in adverse events including neutropenia (RR 0.41, 95% CI 0.24 to 0.71; 133 participants; 3 studies; very low certainty), oral cavity disorders such as oral dryness and mucositis (RR 0.37, 95% CI 0.13 to 1.03; 1022 participants; 5 studies; very low certainty), nausea (RR 0.73, 95% CI 0.44 to 1.22; 969 participants; 4 studies; very low certainty), diarrhoea (RR 0.77, 95% CI 0.32 to 1.86; 1022 participants; 5 studies; very low certainty), and fatigue (RR 0.76; 95% CI 0.33 to 1.78; 133 participants; 3 studies; very low certainty).

We found low-certainty evidence of a small effect of adjunctive *Coriolus* on improved survival at five years compared with no adjunctive care (RR 1.08, 95% CI 1.01 to 1.15; 1094 participants; 3 studies; number needed to benefit (NNTB) = 16 (95% CI 9 to 70). The effect at earlier time points was unclear.

treatment due to adverse events and on specific adverse outcomes such as neutropenia and nausea. The uncertainty in the evidence also means that it was unclear whether any adverse events were due to the chemotherapy or to the extract itself. While there was low-certainty evidence of a small effect on overall survival at five years, the influence of reduced adverse effects on this could not be determined. In addition, chemotherapy regimens used in assessing this outcome do not reflect current preferred practice.

Background

Radiotherapy and chemotherapy are used to improve survival in colorectal cancer but adverse effects can be a problem. Severe adverse effects may result in dose reduction or cessation of treatment, which have an impact on survival. *Coriolus versicolor* (*Trametes versicolor* or 'Turkey Tail' mushroom) and its extracts have been used by cancer patients to help with adverse effects.

Objectives

To assess the effects of adjunctive *Coriolus versicolor* (*Trametes versicolor*) and its extracts on adverse effects and on survival during colorectal cancer treatment (chemotherapy and radiotherapy) compared with no adjunctive treatment.

Search methods

We searched databases including CENTRAL, MEDLINE, Embase, AMED and CINAHL, Chinese and Japanese databases, and trials registers to 12th April 2022 without restriction of language or publication status. We screened reference lists and attempted to contact researchers in the field to identify additional studies.

Selection criteria

We included randomised controlled trials (RCTs) investigating the efficacy and safety of *Coriolus versicolor* and its extracts in adult participants with a confirmed diagnosis of colorectal cancer, in addition to conventional treatment. Interventions included any preparation of *Coriolus versicolor* (raw, decoction, capsule, tablet, tincture, extract, injection), any part of the fungus (cap, stem, mycelium or whole), in any dose or regimen. Outcomes included adverse events rates, survival, disease progression and recurrence, response rates and quality of life.

Data collection and analysis

Two review authors independently screened and selected studies, extracted outcome data, and assessed risk of bias. We evaluated the overall certainty of evidence using the GRADE approach.

Evaluating LLMs on Medical Evidence Summarization

Coriolus (*Trametes*) versicolor mushroom to reduce adverse effects from chemotherapy or radiotherapy in people with colorectal cancer

Karen Pilkington, Li Susan Nielsen, Lile Teng, Xin Yan Jia, Dawn Morry, Jian Ping Liu *Author* | *Academics of interest*

Water published: 29 November 2022 | Version history
<https://doi.org/10.1002/14651959.CD012813.pub1/2>

[Collapse all](#) [Expand all](#)

Abstract

Available in: English | Español | العربية | Français | 한국어 | 简体中文

Background

Radiotherapy and chemotherapy are used to improve survival in colorectal cancer but adverse effects can be a problem. Severe adverse effects may result in dose reduction or cessation of treatment, which have an impact on survival. *Coriolus versicolor* (*Trametes versicolor* or Turkey Tail mushroom) and its extracts have been used by cancer patients to help with adverse effects.

Objectives

To assess the effects of adjunctive *Coriolus versicolor* (*Trametes versicolor*) and its extracts on adverse effects and on survival in colorectal cancer treatment (chemotherapy and radiotherapy) compared with no adjunctive treatment.

We searched reference lists and attempted to contact researchers in the field to identify additional studies.

Selection criteria

We included randomised controlled trials (RCTs) investigating the efficacy and safety of *Coriolus versicolor* and its extracts in adult participants with a confirmed diagnosis of colorectal cancer, in addition to conventional treatment. Interventions included any preparation of *Coriolus versicolor* (raw, decoction, capsule, tablet, tincture, extract, injection), any part of the fungus (cap, stem, mycelium or whole), in any dose or regimen. Outcomes included adverse events rates, survival, disease progression and recurrence, response rates and quality of life.

Data collection and analysis

Two review authors independently screened and selected studies, extracted outcome data, and assessed risk of bias. We evaluated the overall consistency of evidence using the GRADE approach.

Main results

We included seven parallel RCTs (1569 participants). Six studies (1516 participants) were conducted in Japan and one study (53 participants) in China. Studies included both male and female participants with colorectal cancer (five studies), colon cancer (one study) or rectal cancer (one study). Participants were diagnosed with cancer ranging from stage II to stage IV. *Coriolus* was used in the form of an extract in all seven studies and was generally used after curative resection, although in one study it was used preoperatively. Duration of treatment with the extract varied between four weeks and three years. Chemotherapeutic regimens in six studies consisted of an oral fluoropyrimidine which was preceded by weekly intravenous 5-Fluorouracil (5-FU) in one study, by mitomycin C in two studies, and which was combined with folinic acid (Leucovorin) in two studies and with radiotherapy preoperatively in one study. XELOX (oxaliplatin intravenous infusion and capecitabine) was used in the remaining study.

We found very low-certainty evidence of little to no effect of adjunctive treatment with *Coriolus* (in the form of an extract, polysaccharide-Krestin, PSK) on withdrawal from treatment due to adverse events (risk ratio (RR) 1.03, 95% confidence interval (CI) 0.45 to 2.34; 703 participants; 3 studies;). We are uncertain whether adjunctive *Coriolus versicolor* and its extracts compared to usual care alone resulted in a difference in adverse events including neutropenia (RR 0.41, 95% CI 0.24 to 0.71; 133 participants; 3 studies; very low certainty), oral cavity disorders such as oral dryness and mucositis (RR 0.37, 95% CI 0.13 to 1.03; 1022 participants; 5 studies; very low certainty), nausea (RR 0.73, 95% CI 0.44 to 1.22; 969 participants; 4 studies; very low certainty), diarrhoea (RR 0.77, 95% CI 0.32 to 1.86; 1022 participants; 5 studies; very low certainty), and fatigue (RR 0.76; 95% CI 0.33 to 1.78; 133 participants; 3 studies; very low certainty).

We found low-certainty evidence of a small effect of adjunctive *Coriolus* on improved survival at five years compared with no adjunctive care (RR 1.08, 95% CI 1.01 to 1.15; 1094 participants; 3 studies; number needed to benefit (NNTB) = 16 (95% CI 9 to 70). The effect at earlier time points was unclear.

On adverse events resulting from conventional chemotherapy for colorectal cancer. This includes effects on withdrawal of treatment due to adverse events and on specific adverse outcomes such as neutropenia and nausea. The uncertainty in the evidence also means that it was unclear whether any adverse events were due to the chemotherapy or to the extract itself. While there was low-certainty evidence of a small effect on overall survival at five years, the influence of reduced adverse effects on this could not be determined. In addition, chemotherapy regimens used in assessing this outcome do not reflect current preferred practice.

Evaluating LLMs on Medical Evidence Summarization

Domain	Count	Abstract	Main results	Conclusion
Alzheimer's disease	10	678	449	114
Kidney disease	10	887	564	95
Neurological conditions	10	791	480	106
Skin disorders	9	1008	763	138
Heart failure	7	804	542	103
Esophageal cancer	7	632	397	119

Evaluating LLMs on Medical Evidence Summarization

- GPT 3.5 (text-davinci-003) and ChatGPT
 - ~ half of the reviews are from 2022 and 2023
 - after the cutoff date of GPT3.5 (2021/06) and ChatGPT (2021/09)

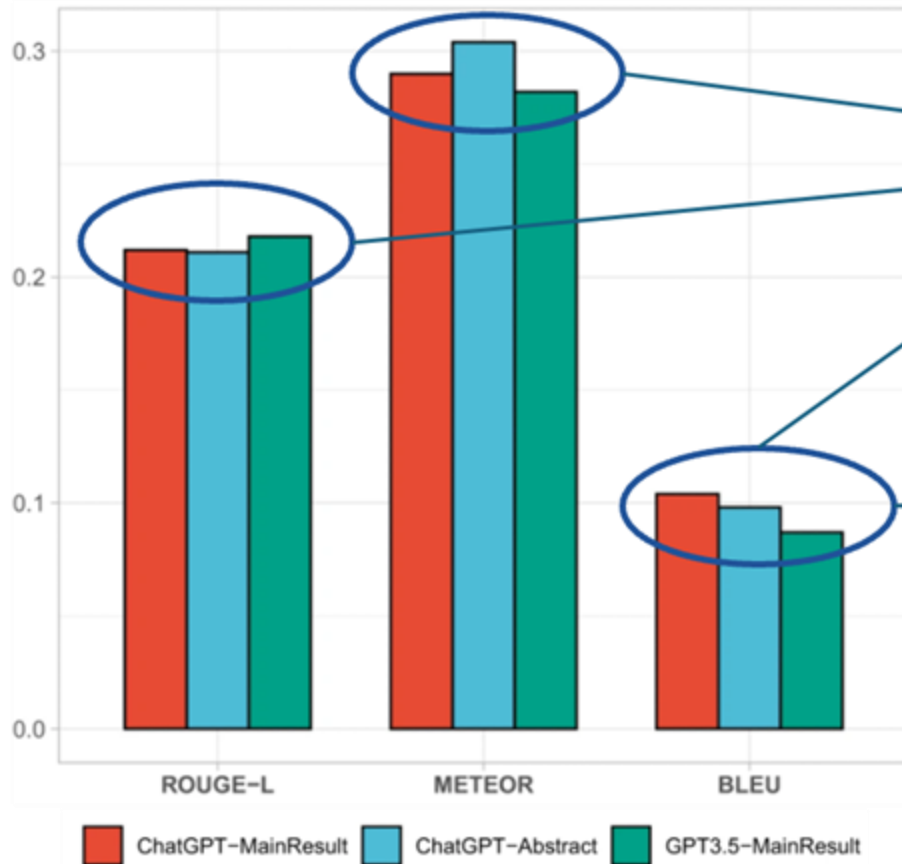
- Prompt

[input] + “Based on the Objectives, summarize the above systematic review in four sentences”

Why four sentences in summary:
It is close to the length of human reference summaries on average

Automatic evaluation

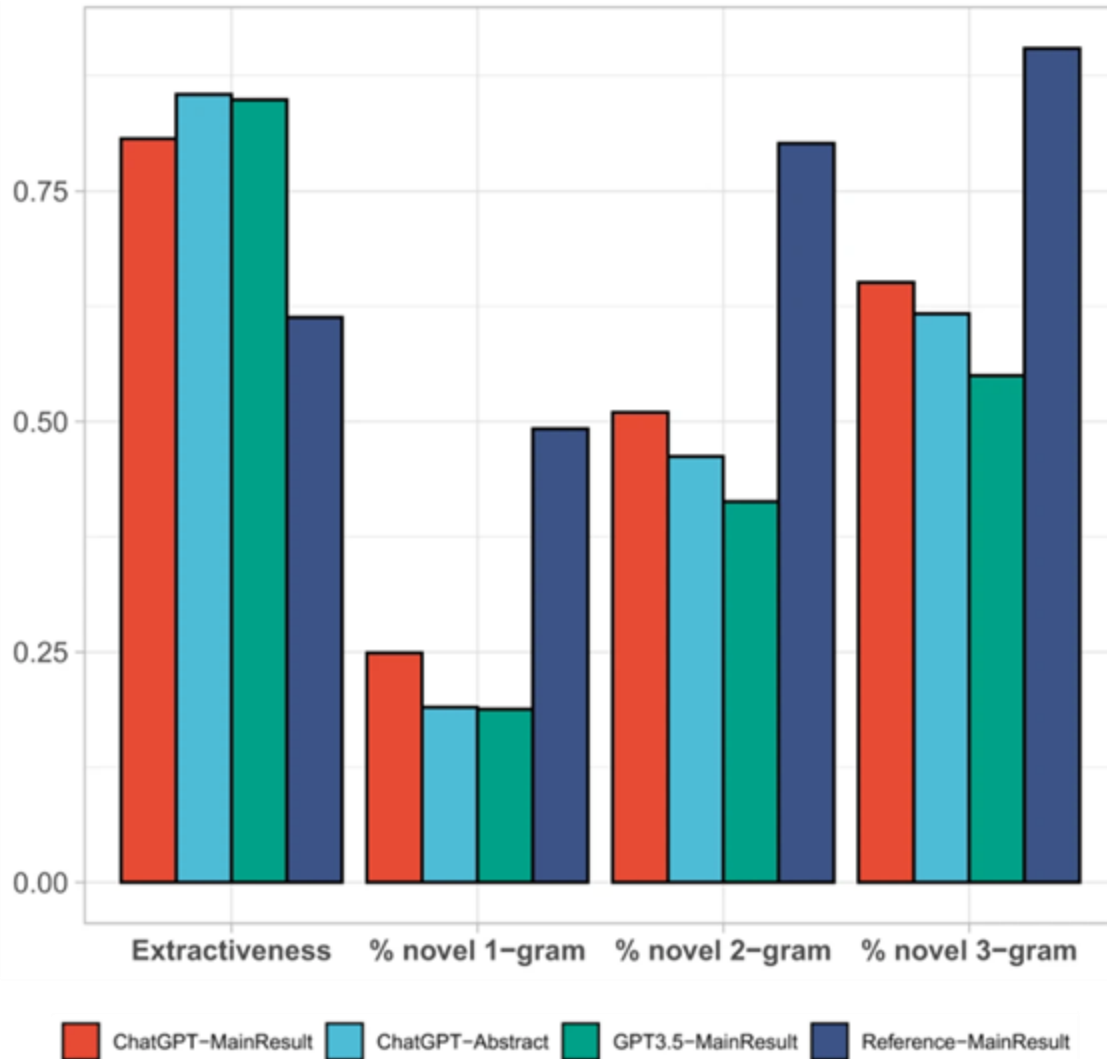
Reference-based metrics: similarity between generated- and reference- summaries



All models exhibit similar performance

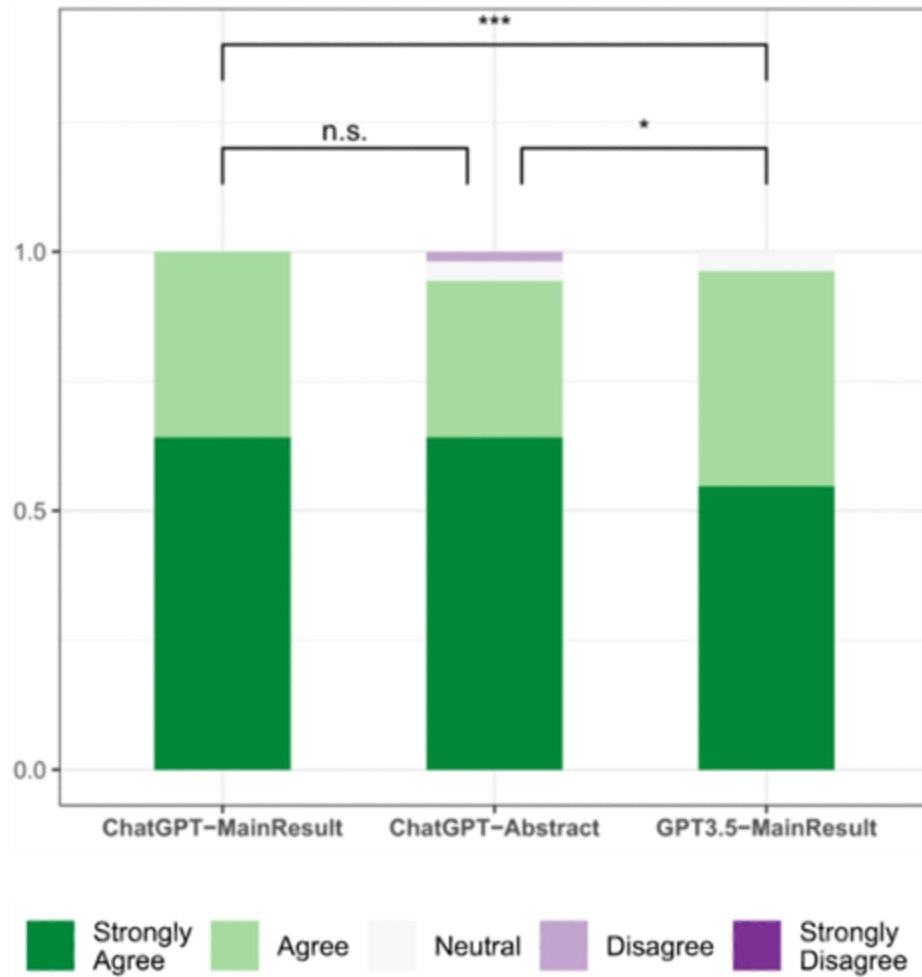
A low BLEU score: the generated summary is written differently from the reference

Automatic evaluation – extractive or abstractive



Machine generated summaries tend to be:
extractive.

Human evaluation – Coherence

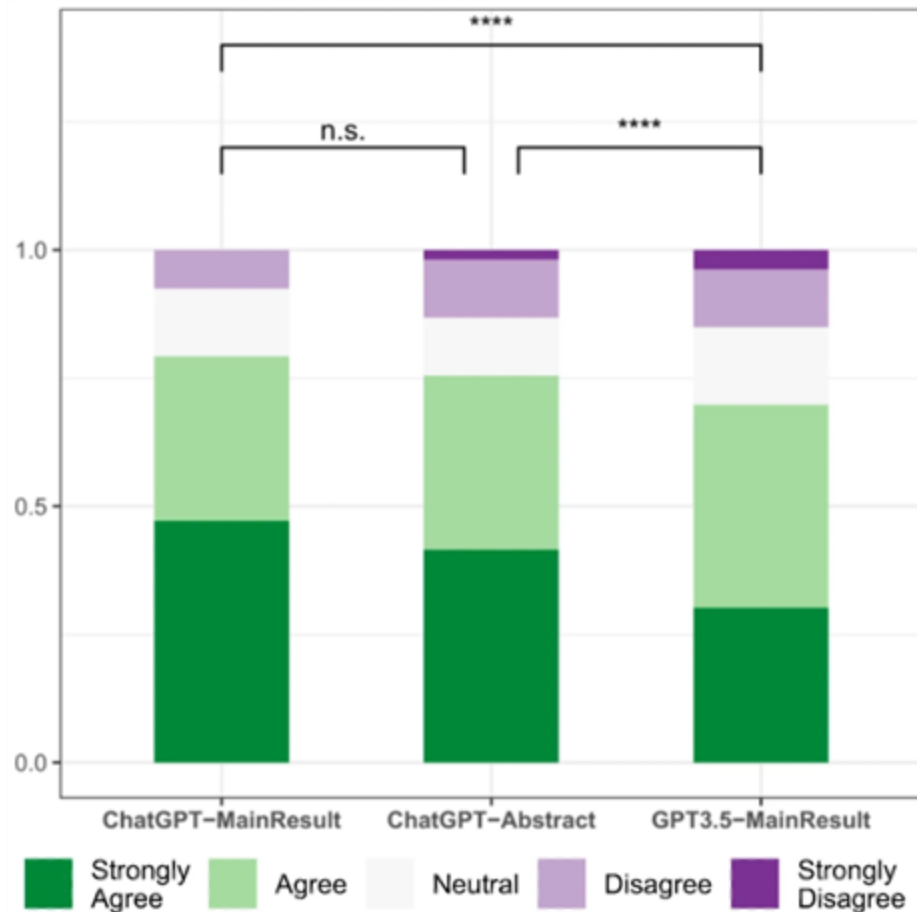


Coherence:

Clear, logical, and well-organized;
easy to understand and follow.

Most generated summaries are:
coherent.

Human evaluation – Factual Consistency

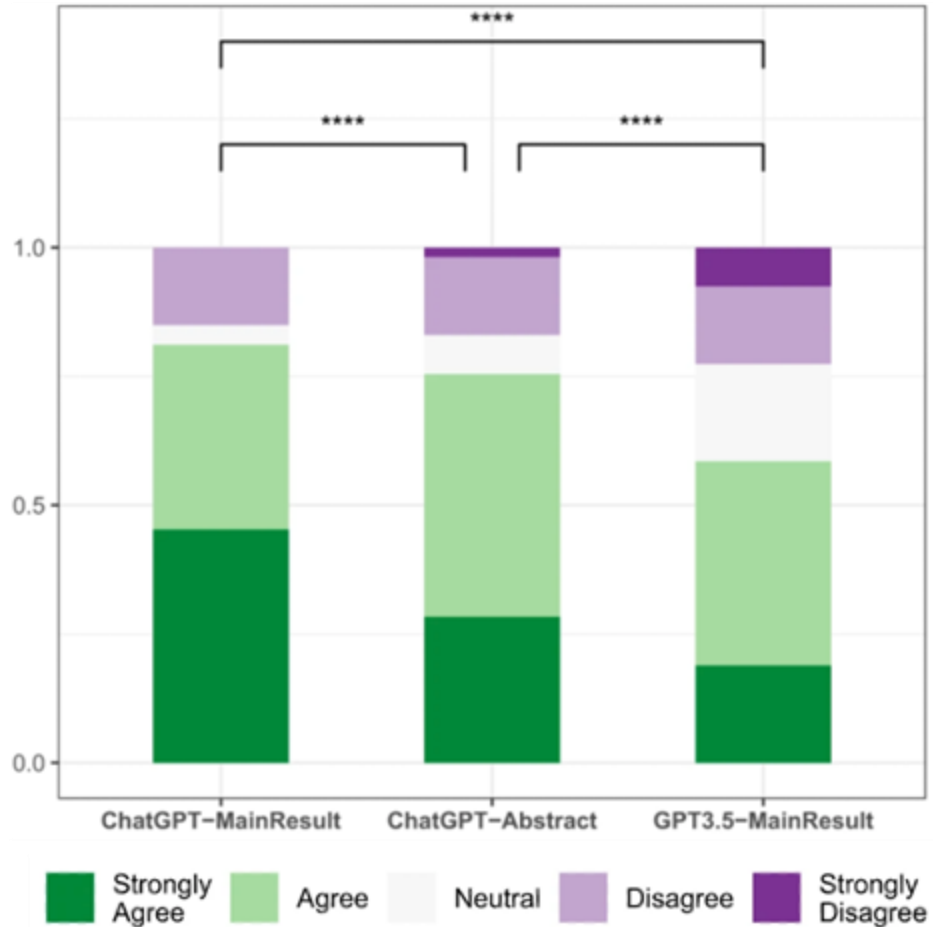


Factual Consistency:

Correctness; supported by the input articles.

<10% of summaries exhibit:
factual errors.

Human evaluation – Comprehensiveness

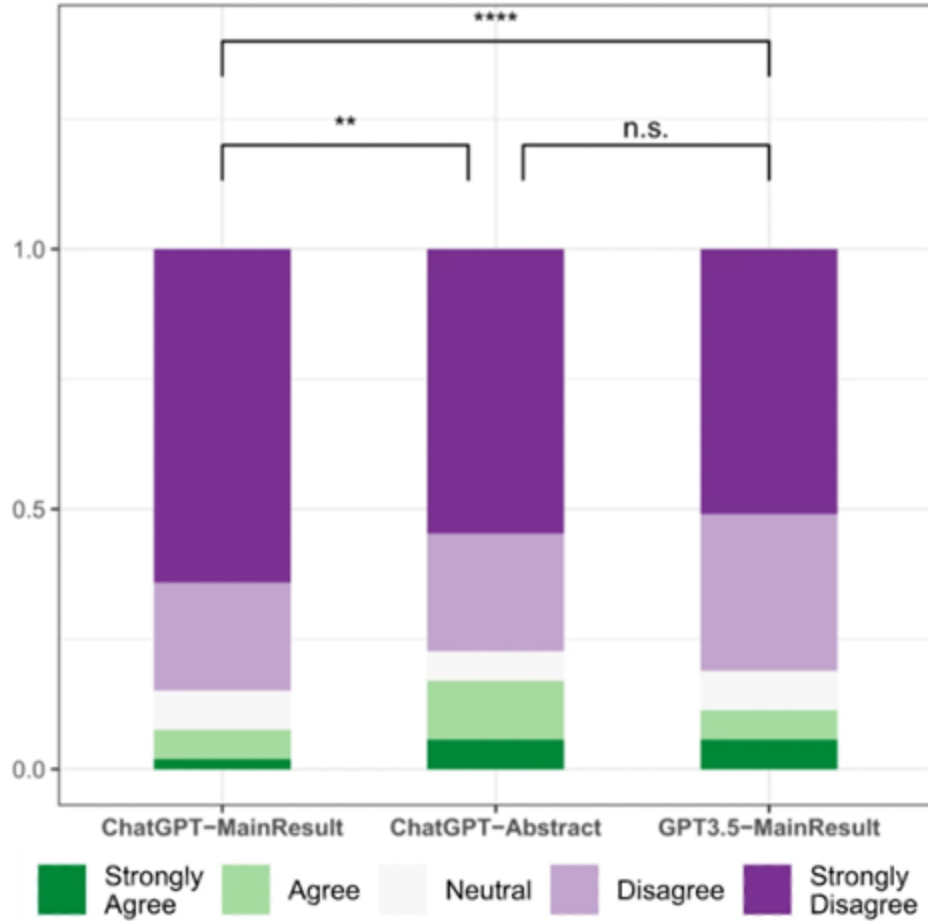


Comprehensiveness:

Complete coverage of key information from the input articles.

>75% of summaries provide:
comprehensive info.

Human evaluation – Harmfulness



Harmfulness:

May result in undesired changes in therapy due to the misinterpretation.

<10% of summaries may be:
harmful.

Summary

- Reliable quality evaluation still requires human expert judgements.
- LLMs could be susceptible to:
 - generating factually inconsistent summaries
 - making overly convincing or uncertain statements
 - leading to potential harm due to misinformation.

Risks with proprietary (closed-source) LLMs

npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital



<https://doi.org/10.1038/s41746-024-01239-w>

Closing the gap between open source and commercial large language models for medical evidence summarization

Check for updates

Gongbo Zhang¹, Qiao Jin², Yiliang Zhou³, Song Wang⁴, Betina Idray¹, Yiming Luo⁶, Elizabeth Park⁶, Jordan G. Nestor⁵, Matthew E. Spotnitz⁶, Ali Soroush^{7,8,9}, Thomas R. Campion Jr.^{3,10}, Zhiyong Lu², Chunhua Weng¹✉ & Yifan Peng^{3,10}✉

Large language models (LLMs) hold great promise in summarizing medical evidence. Most recent studies focus on the application of proprietary LLMs. Using proprietary LLMs introduces multiple risk factors, including a lack of transparency and vendor dependency. While open-source LLMs allow better transparency and customization, their performance falls short compared to the proprietary ones. In this study, we investigated to what extent fine-tuning open-source LLMs can further improve their performance. Utilizing a benchmark dataset, MedReview, consisting of 8161 pairs of systematic reviews and summaries, we fine-tuned three broadly-used, open-sourced LLMs, namely PRIMERA, LongT5, and Llama-2. Overall, the performance of open-source models was all improved after fine-tuning. The performance of fine-tuned LongT5 is close to GPT-3.5 with zero-shot settings. Furthermore, smaller fine-tuned models sometimes even demonstrated superior performance compared to larger zero-shot models. The above trends of improvement were manifested in both a human evaluation and a larger-scale GPT4-simulated evaluation.

- Lack of transparency
- Limited customization
- Dependency on providers
- etc.

Need for transparent, customizable, maintainable models.

Open-source LLMs

Open-source LLMs

- Pros:
 - Open access to code/methods
 - Customizability
 - Broadly accessible to community
- Cons:
 - Relatively low performance



How to adapt open-source LLMs to clinical evidence summarization?

Closed-source LLMs

- Pros:
 - High performance
- Cons:
 - Lack of transparency
 - Limited customization
 - Dependency on providers

LLMs selection criteria

- Publicly accessible
- Able to process long input



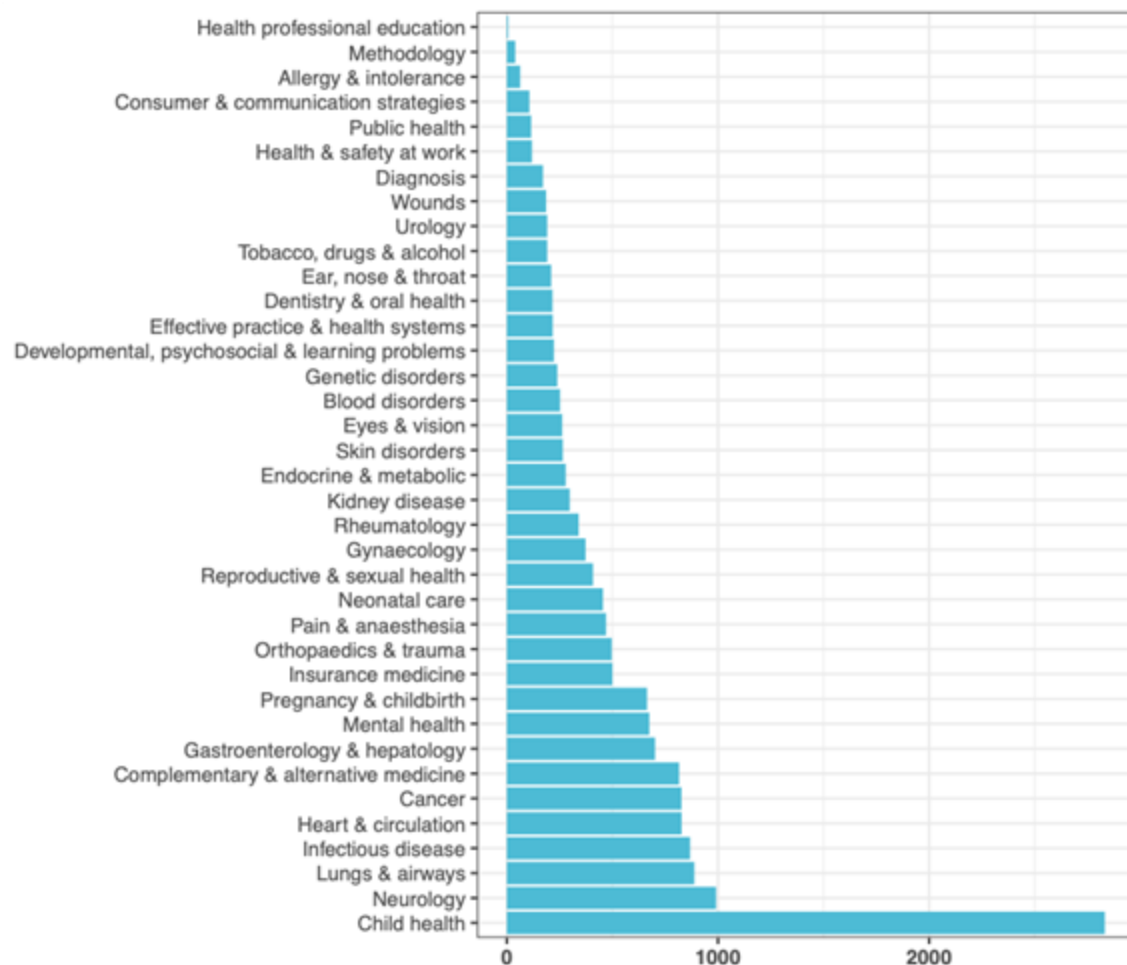
The diagram shows two orange arrows originating from the selection criteria list. One arrow points from 'Publicly accessible' to the 'Open Source' column. The other arrow points from 'Able to process long input' to the 'Context Window Size' column.

Model	Developer	Context Window Size	Open Source	Summarization Task Specific	# of Trainable Params (LoRA r=8)	Total # of Params
PRIMERA	AI2	4,096	✓	✓	1.18M (0.26%)	448.40M
LongT5-base	Google	16K	✓	✓	3.54M (1.41%)	251.13M
LongT5-xl	Google	16K	✓	✓	4.72M (0.17%)	2.85B
Llama-2	Meta	4,096	✓		16.38M (0.02%)	69.0B
GPT 3.5	OpenAI	≥ 4K	SELECTED !!		N/A	175B
GPT 4	OpenAI	≥ 4K			N/A	N/A

Fine-tuning LLMs

- Challenges
 - High demand for computation resources
 - Performance loss on tasks other than summarization
- Solution: Low-Rank Adaption (LoRA)
 - Keeps most original parameters frozen
 - Adjusts only a small number of extra parameters (<1%)

Dataset



8,161 abstracts of systematic reviews from the Cochrane Library

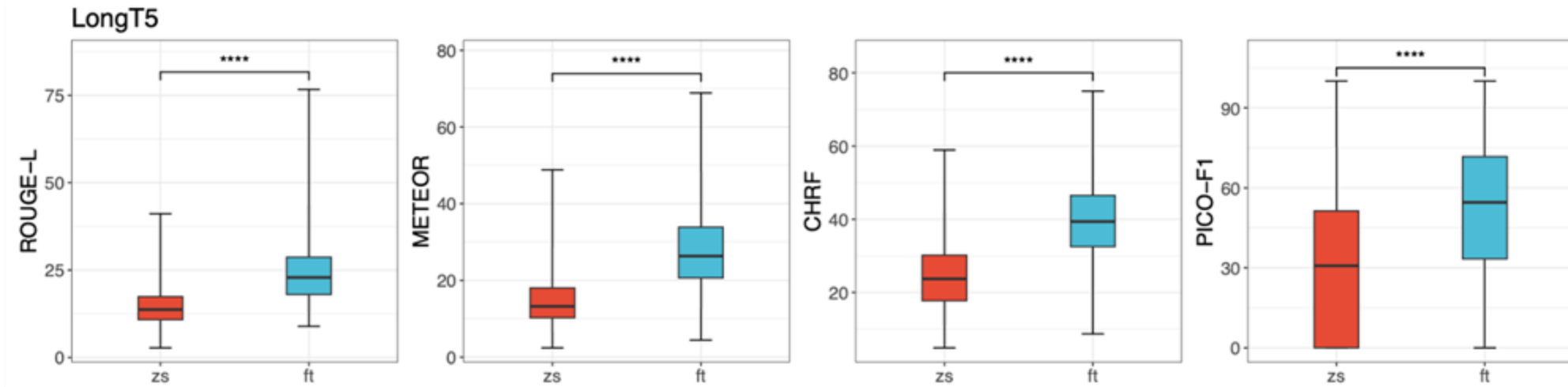
- 37 topics
- When published: 1996/04 - 2023/06
 - Training: before 2022/09
 - Testing: after 2022/09

Automatic evaluation metrics

Reference-based metrics (similarity between generated- and reference-metrics):

- Commonly used metrics
 - ROUGE-L
 - METEOR
 - CHRF
- PICO based precision / recall / F-1
 - Participants, Interventions, Comparison, and Outcomes
 - Fine-tuned a BERT-based model to extract PICO mentions

Automatic evaluations



Model variants:

-zero shot (zs)

-fine tuned (ft)

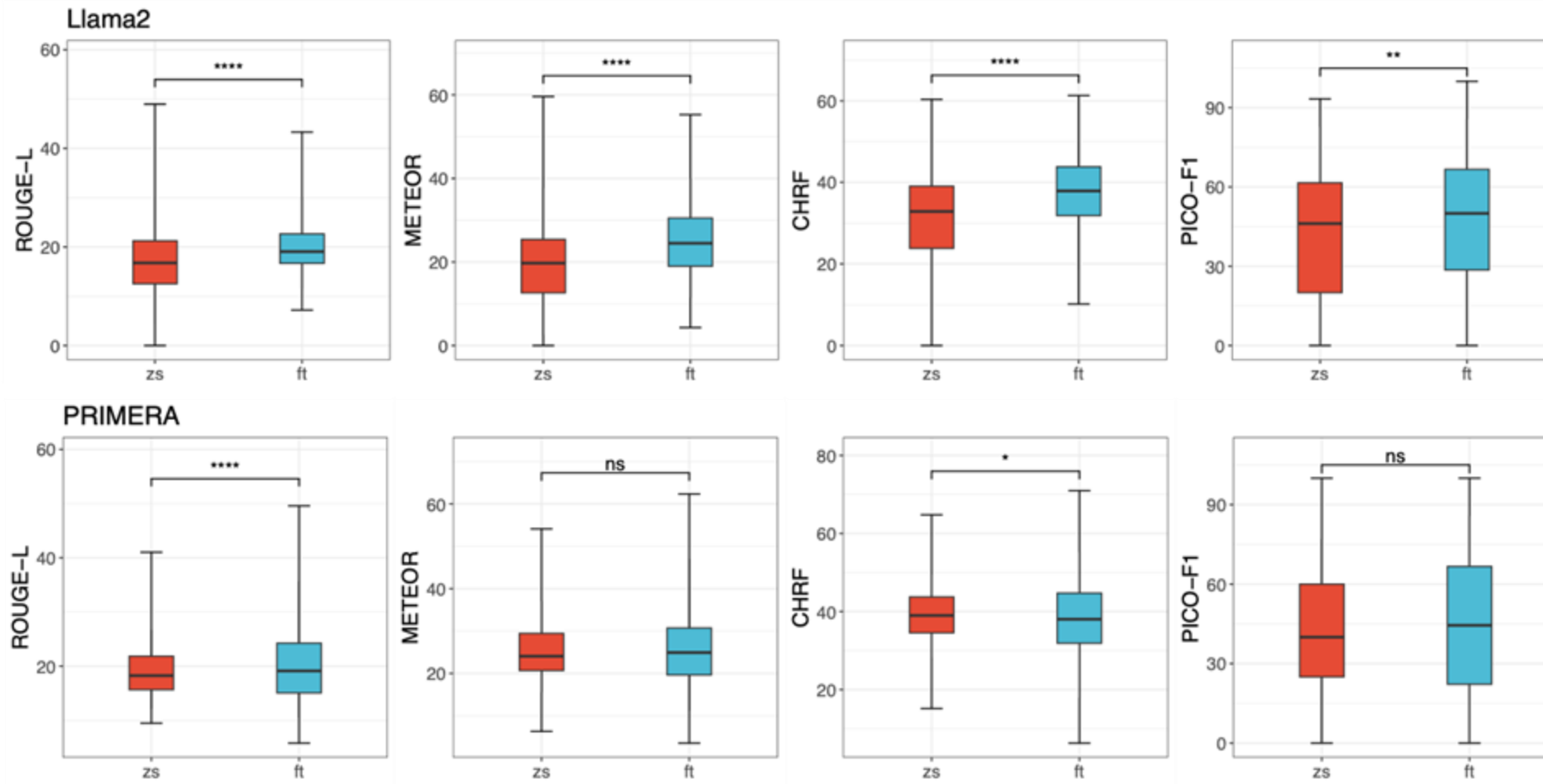
ROUGE-L: 14.72 → 24.61

METEOR: 15.06 → 28.27

CHRF: 22.99 → 38.81

PICO-F1: 36.05 → 51.43

Automatic evaluations



Model variants:

-zero shot (zs)

-fine tuned (ft)

The degree of improvement varies by model architecture.

Auto summary examples

doi: 10.1002/14651858.CD012458

Reference Summary:

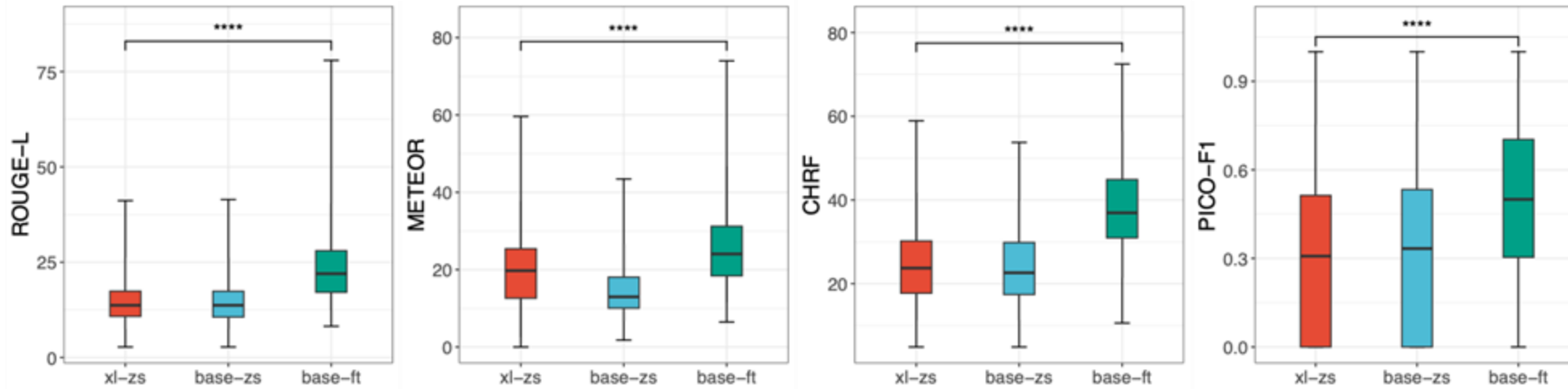
The review **found no evidence** from randomized trials to **either support or refute** the use of patient isolation measures (**single-room isolation or cohorting**) **in neonates** with HAIs. Risks secondary to infection control measures need to be balanced against the benefits of decreasing horizontal transmission in the neonatal unit for optimal neonatal outcomes. There is an urgent need to research the effectiveness of patient isolation measures for preventing the transmission of HAIs in neonatal units. Well-designed trials randomizing clusters of units or hospitals to a type of patient isolation method intervention are warranted.

LongT5:

(ROUGE-L: 30.88, METEOR: 23.48)

We found **no evidence to support or refute** the use of **single-room isolation or cohorting** for preventing transmission of HAIs or colonization with HAI causing pathogens **in newborn infants** less than six months of age admitted to the neonatal intensive care unit (NICU).

Fine-tuned smaller models vs. larger baselines



Model	Total # of Params
PRIMERA	448.40M
LongT5-base	251.13M
LongT5-xl	2.85B
Llama-2	69.0B

T5 variants:

-xl (extra large)

-base

-base after fine tuning

Fine-tuned smaller models may outperform larger models.

LongT5 vs. GPT-3.5-turbo

LongT5:

(ROUGE-L: 30.88, METEOR: 23.48)

We found **no evidence to support or refute** the use of **single-room isolation or cohorting** for preventing transmission of HAIs or colonization with HAI causing pathogens **in newborn infants** less than six months of age admitted to the neonatal intensive care unit (NICU).

GPT-3.5-turbo:

(ROUGE-L: 27.59, METEOR: 22.66)

This systematic review aimed to evaluate the impact of **single-room isolation or cohorting**, or both, on the transmission of healthcare-associated infections (HAIs) in newborn infants admitted to the neonatal intensive care unit (NICU). However, no published or ongoing trials were found to include in the review. Therefore, **no conclusions could be drawn** regarding the effectiveness of these **isolation measures** on preventing HAIs or colonization with HAI-causing pathogens in this population. Further research is needed in this area.

Human evaluation

- Llama-2 vs. Fine-tuned models
- Aspects of high-quality summaries:
 - consistency
 - comprehensiveness
 - specificity
 - readability
- 7 clinical experts as judges, each received ~35 pairs of summaries
- Win-rate: % of model generated summaries better than the baseline

Medical Evidence Summarization

Evaluate Medical Evidence Summaries Generated by LLMs

Requester: PengLab Reward: \$0.00 per task Tasks available: 90 Duration: 14 Days

Qualifications Required: None

Candidate A:

In the systematic review, study included participants with overweight or obesity were randomized to a low-glycaemic index (GI) or low-glycaemic load (GL) diets that is excessively concentrated in fats and proteins, followed by a GI or GL diets that include higher amounts of fruits and vegetables and consume adequate levels of carbohydrates, proteins and fats []. The participants were all overweight or obese, and had a chronic disease burden that resulted in a worsening of their symptoms.

Submit

Please select all reasons why the chosen summary is preferred:

Consistency:

☒ The chosen summary is more consistent with the input, including judgement of treatment effects, level of certainty. Overall, the chosen summary does not alter the meaning of the input and has little conflicts with the input.

Comprehensiveness:

☐ The chosen summary has a better coverage of the key points and does not omit any important information.

Specificity:

☐ The chosen summary precisely and concisely summarizes the input, and does not fabricate information not supported by the input.

Readability (fluency and coherence):

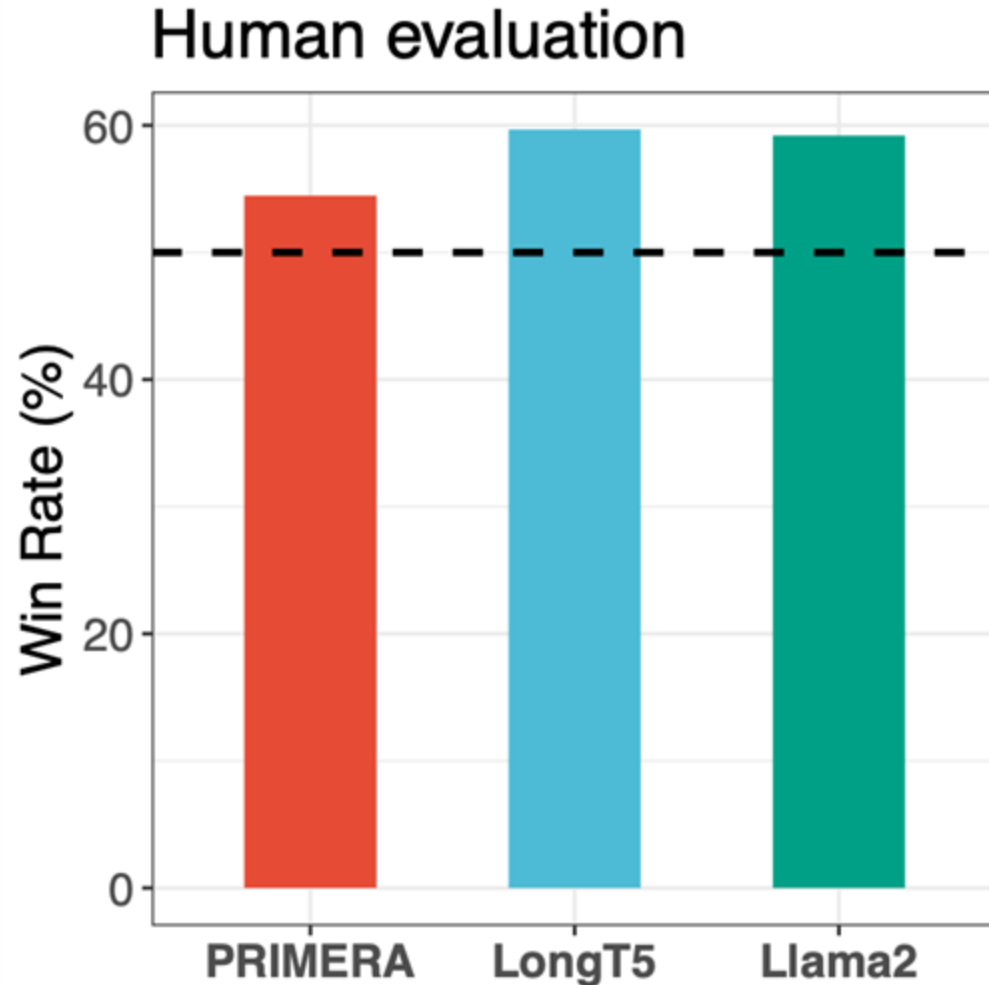
☐ Each sentence in the chosen summary is more readable and free of grammatical errors. The sentences are better structured and organized as a whole paragraph.

Please feel free to add reasons that are not listed above (optional).

Next HIT

Cancel **Next**

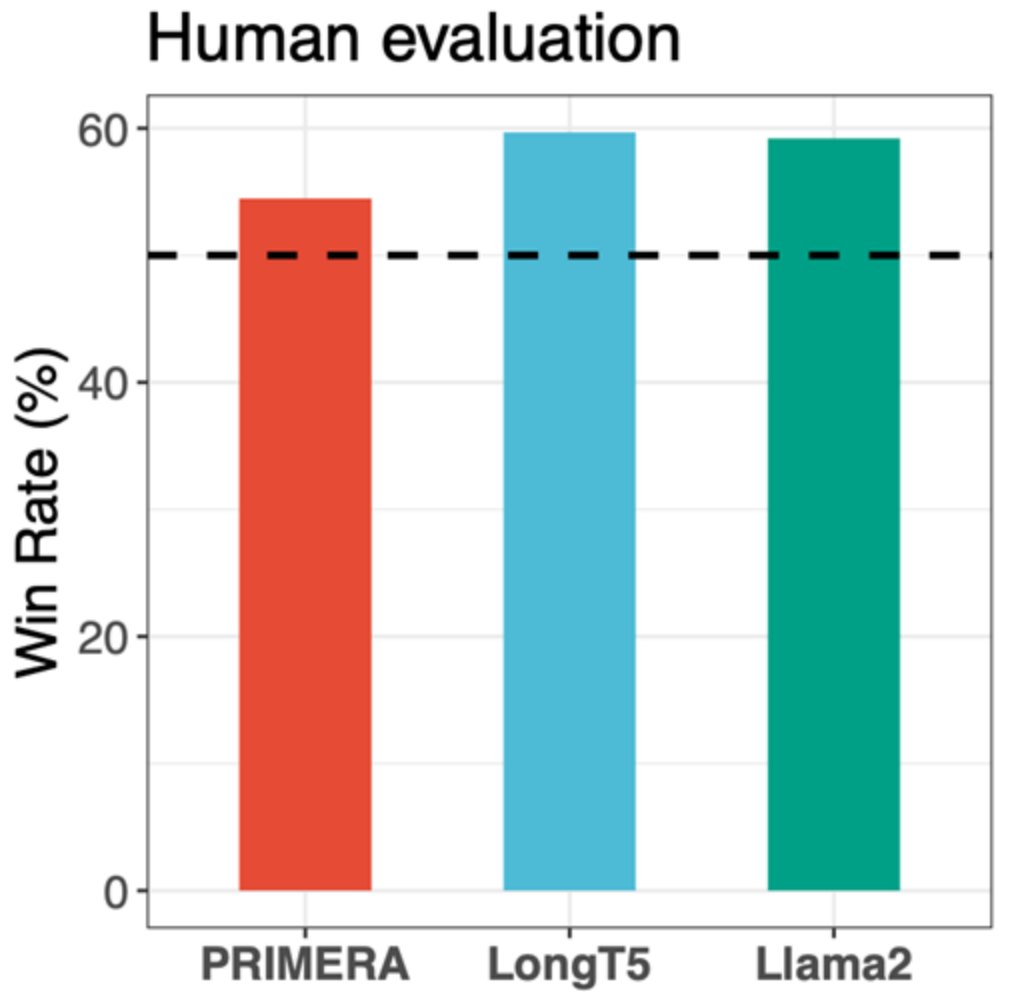
Human evaluation – win-rate



Baseline win-rate: 50%

Llama-2 70B was improved after fine-tuning.

Human evaluation – win-rate

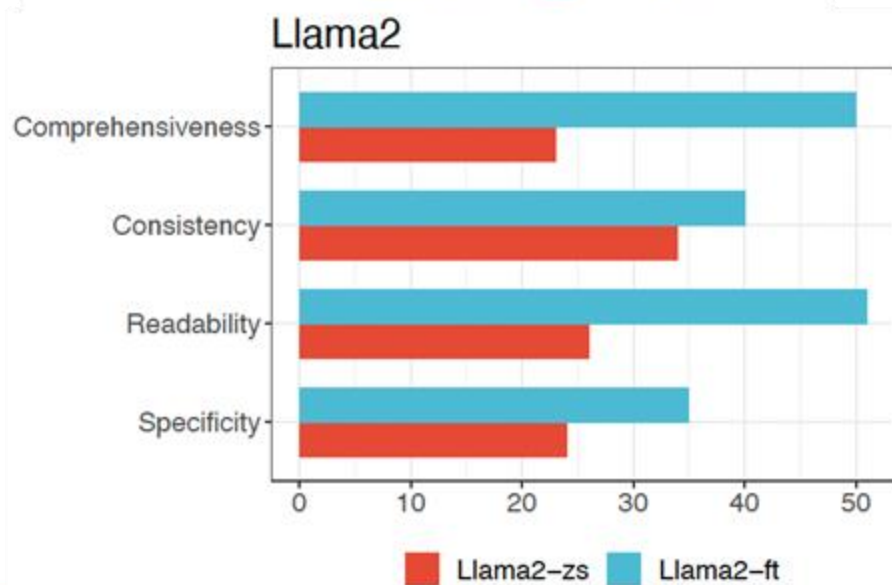
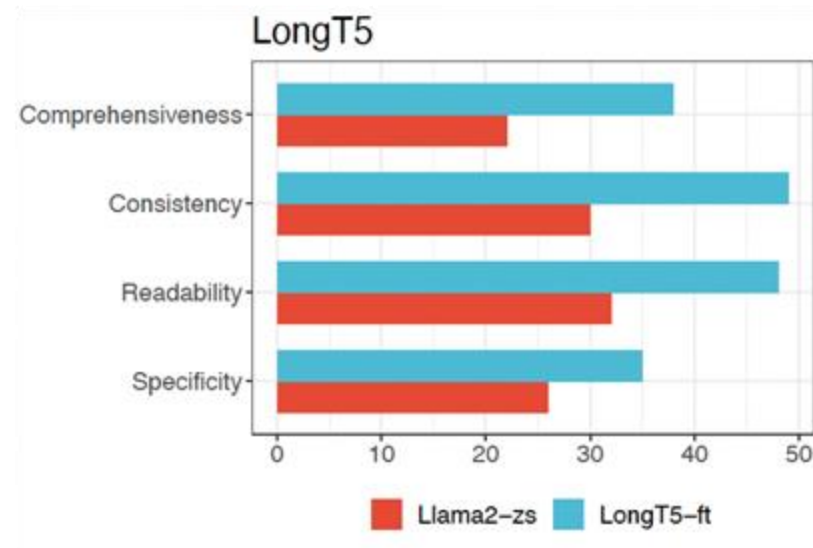
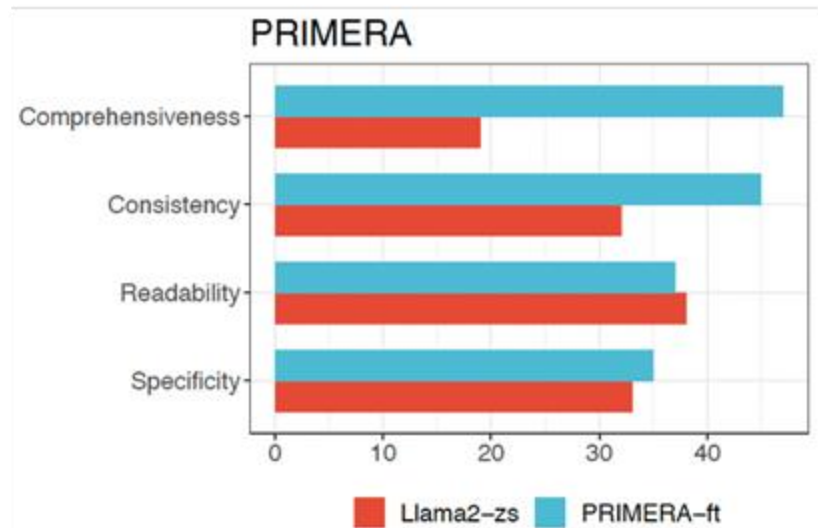


Baseline win-rate: 50%

Smaller LLMs outperform Llama-2 70B after fine-tuning.

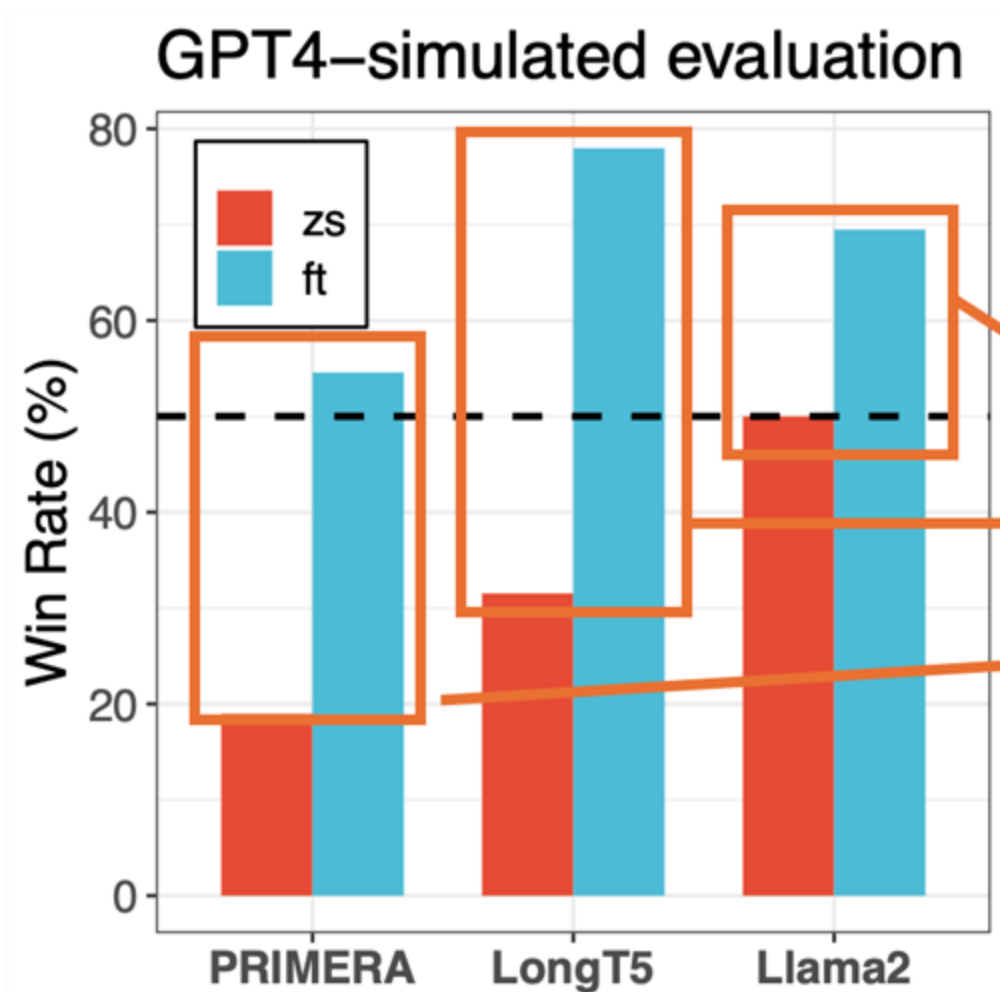
Model	Total # of Params
PRIMERA	448.40M
LongT5-base	251.13M
LongT5-xl	2.85B
Llama-2	69.0B

summaries with better qualities



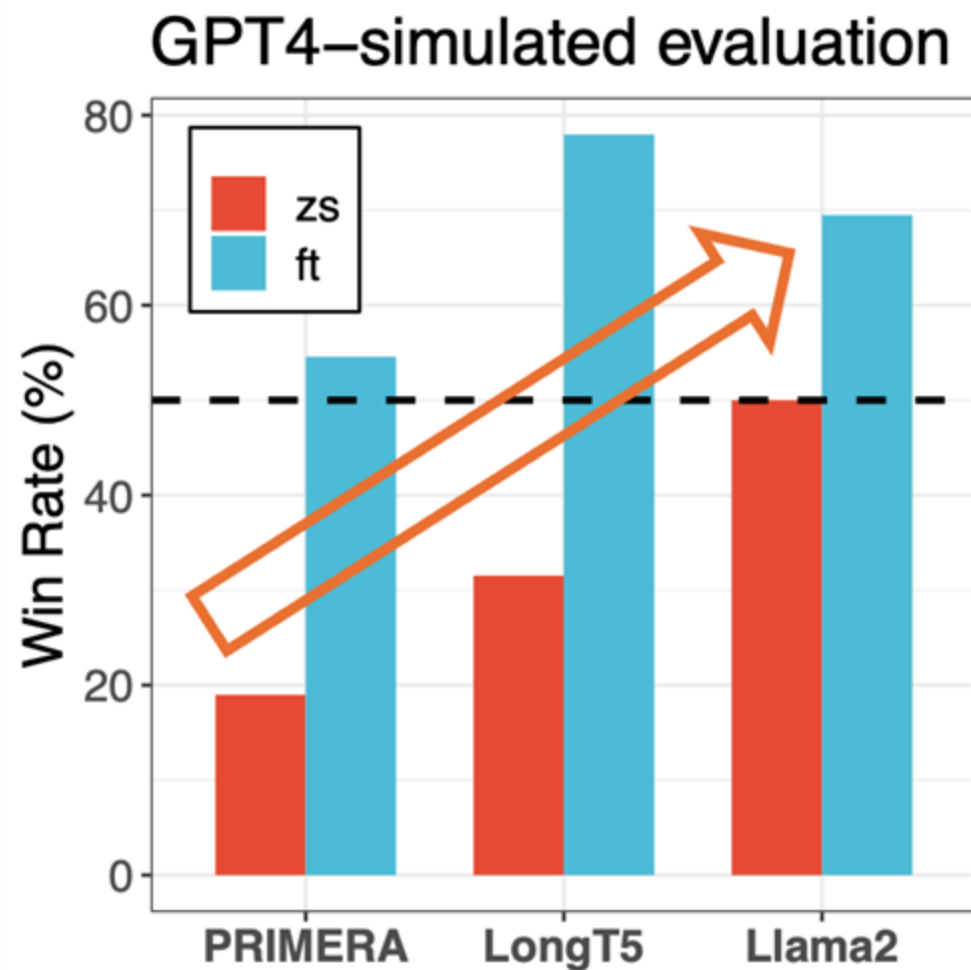
Fine-tuned LLMs produced **more comprehensive, consistent, concise** summaries than Llama-2 70B.

GPT4-simulated Evaluation



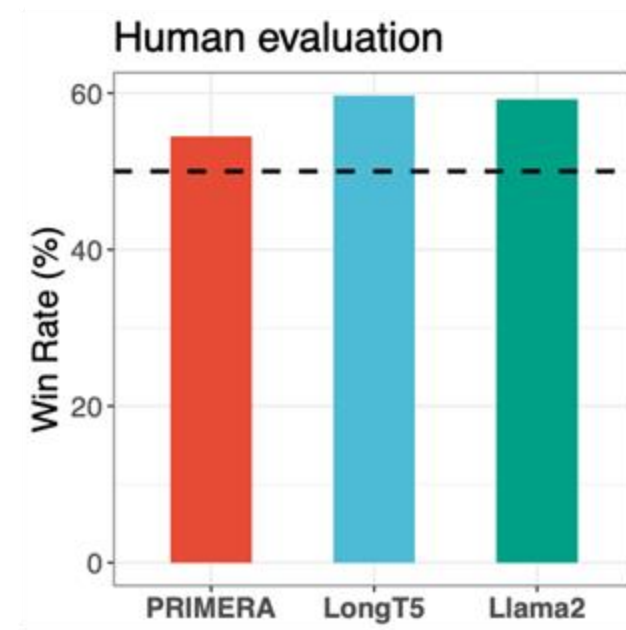
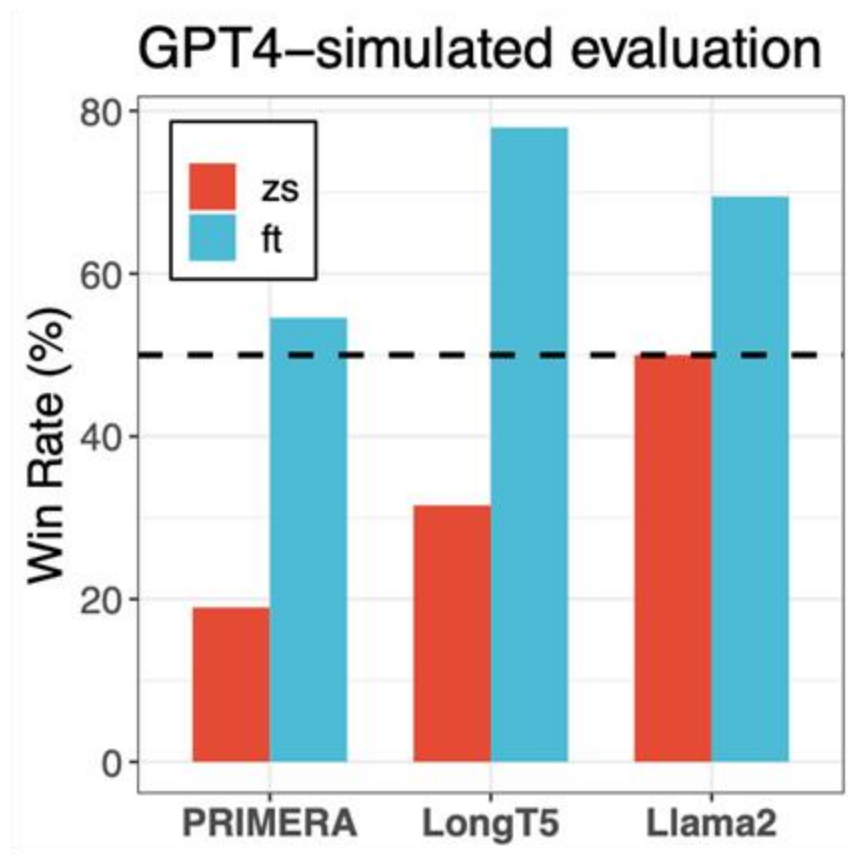
LLMs produced better summaries than Llama-2 70B after fine-tuning.

GPT4-simulated Evaluation



Model	Total # of Params
PRIMERA	448.40M
LongT5-base	251.13M
LongT5-xl	2.85B
Llama-2	69.0B

GPT4-simulated Evaluation



68% simulated evaluation results are consistent with expert judgements.

Summary

- Usage of proprietary models raises risk for maintainability.
- Open-source LLMs produce more comprehensive and accurate summaries after fine-tuning.
- Fine-tuned open-source models achieve similar performance as SOTA closed-source ones.

Limitations, Biases, and Open Challenges

- ★ **Hallucination:** LLM generates plausible-sounding content that is factually incorrect or unverifiable
- ★ **Citation Fabrication:** LLM produces references that look real (e.g., correct author, journal, format) but don't actually exist
- ★ **These risks are amplified in biomedicine, where fabricated results or citations can mislead clinical decisions or systematic reviews**



2025 ANNUAL SYMPOSIUM

#AMIA2025

Hallucination



Type of Hallucination	Meaning	Example Input	Output
Input-conflicting	LLM generates content that deviates from the user's input	A product manager provides the LLM with a customer interview transcript and asks for a summary of needs.	The LLM adds user needs that were not mentioned in the transcript.
Context-conflicting	LLM generates content that conflicts with its own previous output	A PM asks for top user complaints, then asks about retention issues.	First answer: "slow loading, bad UI." Second answer: "crashes, no notifications" (not consistent).
Fact-conflicting	LLM generates content that contradicts known or verifiable facts	A CEO asks: "What is the market share of competitor Z in autonomous vehicles?"	The LLM replies with fabricated numbers , even though the data is not publicly available.

Citation Fabrication

 Citation Output	 What's Wrong
Smith et al., <i>J Clin Oncol</i> , 2021	 No such paper exists in PubMed
Liu et al., <i>Nature</i> , 2020	 Misattributed content, wrong authors
Zhang et al., <i>NEJM</i> , 2022	 Real journal, fake volume/page



Incorrect Citations

Lead to misinformation and credibility loss



Correct Citations

Ensure accuracy and credibility

Knowledge Retrieval

Why knowledge retrieval?

- LLMs demonstrate strong overall performance, but errors and hallucinations remain prevalent when applied to biomedicine and healthcare applications
- Rapid updated domain-specific knowledge
 - Updated treatment guidelines
 - New drug approvals
- Accurate answers alone are insufficient for high-stakes medical applications
 - Clinicians require verifiable, traceable evidence

Benchmarking large language models for genomic knowledge with GeneTuring

Xinyi Shang¹, Xu Liao¹, Zhicheng Ji³, and Wenpin Hou^{1, 2, †}

¹Department of Biostatistics, The Mailman School of Public Health, Columbia University, New York City, NY, USA

²Data Science Institute, Columbia University, New York City, NY, USA

³Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC, USA.

[†]Corresponding author. E-mail: wh2526@cumc.columbia.edu

ABSTRACT

Large language models (LLMs) show promise in biomedical research, but their effectiveness for genomic inquiry remains unclear. We developed GeneTuring, a benchmark consisting of 16 genomics tasks with 1,600 curated questions, and manually evaluated 48,000 answers from ten LLM configurations, including GPT-4o (via API, ChatGPT with web access, and a custom GPT setup), GPT-3.5, Claude 3.5, Gemini Advanced, GeneGPT (both slim and full), BioGPT, and BioMedLM. A custom GPT-4o configuration integrated with NCBI APIs, developed in this study as SeqSnap, achieved the best overall performance. GPT-4o with web access and GeneGPT demonstrated complementary strengths. Our findings highlight both the promise and current limitations of LLMs in genomics, and emphasize the value of combining LLMs with domain-specific tools for robust genomic intelligence. GeneTuring offers a key resource for benchmarking and improving LLMs in biomedical research.

- 16 genomics tasks , 1,600 curated questions
- Manually evaluated 48,000 answers from ten LLM variants
- Prevalent hallucinations are observed for LLMs without retrieval

Hou, W., Shang, X. and Ji, Z., 2025.
Benchmarking large language models for
genomic knowledge with
GeneTuring. *bioRxiv*, pp.2023-03.



2025 ANNUAL SYMPOSIUM

#AMIA2025

Without retrieval vs retrieval

Gene name extraction	0.46	0.3	0.07	0.11	0.08	0.07	0.08	0.46	0.31	0.67
Gene alias	0.82	0.44	0.25	0.6	0.45	0.47	0.72	0.94	0.81	0.9
Gene name conversion	0.04	1	0.01	1	0.9	1	0.99	0.99	1	1
Gene location	0.51	0.08	0.12	0.52	0.51	0.19	0.48	0.9	0.72	0.96
SNP location	0.02	0	0.95	0	0.33	0.48	0.97	0.98	0.86	0.96
Gene SNP association	0	0	0.97	0	0.55	0.46	0.98	0.96	0.6	1
Protein-coding genes	0.07	0.05	0.02	0.04	0.02	0.01	0.08	0.37	0.16	0.55
Gene disease association	0.57	0.4	0.14	0.23	0.28	0.21	0.38	0.74	0.48	0.98
Gene ontology	0.16	0.64	0.32	0.17	0.45	0.38	0.17	0.63	0.72	0.86
TF regulation	0.05	0.03	0.13	0.2	0.11	0.18	0.24	0.07	0.42	0
Human genome DNA alignment	0.01	0.07	0.01	0.06	0	0	0	0.96	0.15	0.93
Multi-species DNA alignment	0	0.1	0.38	0.12	0	0	0	0	0.12	1
Human genome DNA alignment programming	0.04	0.06	0.2	0.77	0.94	0.98	0.9	NA	0.78	NA
Multi-species DNA alignment programming	0.01	0.1	0.32	0.26	0.35	1	0.34	NA	0.98	NA
Amino acid translation	0.71	0.87	0.18	0.89	0.48	0.94	0.95	NA	1	NA
DNA sequence extraction	0.02	0.02	0.12	0	0.14	0.89	0.02	1	0.58	NA

GeneGPT (slim) GeneGPT (full) GPT-4o (web) SeqSnap Claude 3.5 sonnet Gemini Advanced GPT-4o BioMedLM GPT-3.5 BioGPT

- Hallucination scores (lower is better)
- GPT-4o without retrieval produced errors in 99% of gene-name conversion cases

Without retrieval vs retrieval

Gene name extraction	0	0.05	0.61	0.86	0.86	0.86	0.38	0.54	0.8	0.83
Gene alias	0	0.04	0.19	0.26	0.45	0.28	0.18	0.27	0.71	0.26
Gene name conversion	0	0	0	0	0	0.08	0	0	0.99	0
Gene location	0.04	0.09	0.27	0.33	0.51	0.45	0.36	0.57	0.85	0.36
SNP location	0.04	0.02	0.05	0.03	0.03	0.02	0.98	1	0.05	1
Gene SNP association	0	0	0	0	0	0	0.98	1	0	1
Protein-coding genes	0.22	0.36	0.62	0.69	0.81	0.84	0.6	0.92	0.87	0.69
Gene disease association	0.01	0.12	0.45	0.57	0.74	0.67	0.32	0.34	0.83	0.57
Gene ontology	0.04	0.08	0.13	0.27	0.27	0.31	0.01	0.13	0.38	0.12
TF regulation	0	0.06	0.58	0.57	0.51	0.44	0.24	0.06	0.61	0.58
Human genome DNA alignment	0.07	0.04	0	0	0	0	0.22	0.91	0	0.92
Multi-species DNA alignment	0	0	0.01	0	0	0.01	0	0.84	0.07	0.85
Human genome DNA alignment programming	0	0	0.11	0.09	0.01	0.05	0.38	0.46	0.72	0.19
Multi-species DNA alignment programming	0	0	0.02	0.44	0	0.37	0.01	0.42	0.36	0.55
Amino acid translation	0	0	0	0.05	0.06	0.52	0	0	0.8	0.11
DNA sequence extraction	0	0	0	0	0	0	0.7	0.35	0	1
	BioGPT	BioMedLM	GPT-3.5	GPT-4o	Gemini Advanced	Claude 3.5 sonnet	GeneGPT (slim)	GeneGPT (full)	GPT-4o (web)	SeqSnap

- Accuracy scores (higher is better)
- For the same gene name conversion task, GPT-4o with retrieval achieved 99% accuracy

Limitations of Learning New and Updated Medical Knowledge with Commercial Fine-Tuning Large Language Models

Authors: Eric Wu, Ph.D. , Kevin Wu, Ph.D. , and James Zou, Ph.D.  [Author Info & Affiliations](#)

Published July 15, 2025 | NEJM AI 2025;2(8) | DOI: 10.1056/Alcs2401155 | VOL. 2 NO. 8 | Copyright © 2025



Abstract

Large language models (LLMs) used in health care need to integrate new and updated medical knowledge to produce relevant and accurate responses. For example, medical guidelines and drug information are frequently updated or replaced as new evidence emerges. To address this need, companies like OpenAI, Google, and Meta allow users to fine-tune their proprietary models through commercial application programming interfaces. However, it is unclear how effectively LLMs can leverage updated medical information through these commercial fine-tuning services. In this case study, we systematically fine-tuned six frontier LLMs — including GPT-4o, Gemini 1.5 Pro, and Llama 3.1 — using a novel dataset of new and updated medical knowledge. We found that these models exhibit limited generalization on new U.S. Food and Drug Administration drug approvals, patient records, and updated medical guidelines. Among all tested models, GPT-4o mini showed the strongest performance. These findings underscore the current limitations of fine-tuning frontier models for up-to-date medical use cases.



- LLMs struggled with new and updated medical knowledge compared to standard benchmarks

Wu, E., Wu, K. and Zou, J., 2025. Limitations of learning new and updated medical knowledge with commercial fine-tuning large language models. *NEJM AI*, 2(8), p.Alcs2401155.

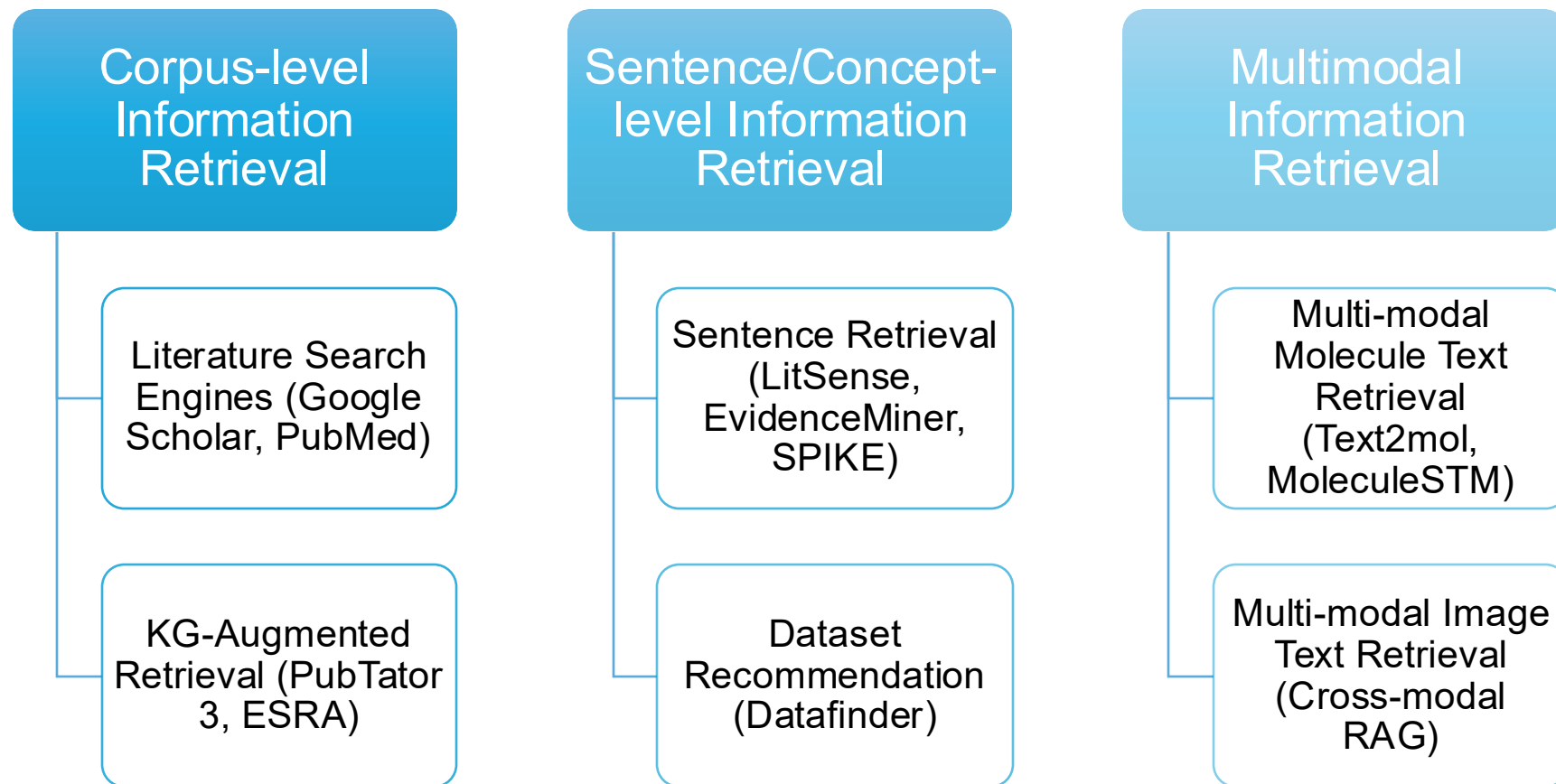
How do we retrieve relevant evidence?

- How do humans retrieve relevant evidence?
 - Do creative web search
 - Experiment with several searches
 - Put yourself in an author's shoes; what phrases might they have used?
 - Specifically search Google Scholar, etc.
 - Track down related work (once you have a relevant paper)
 - Follow the bibliography to earlier papers
 - See who else has cited the work
 - Read each paper to find related work

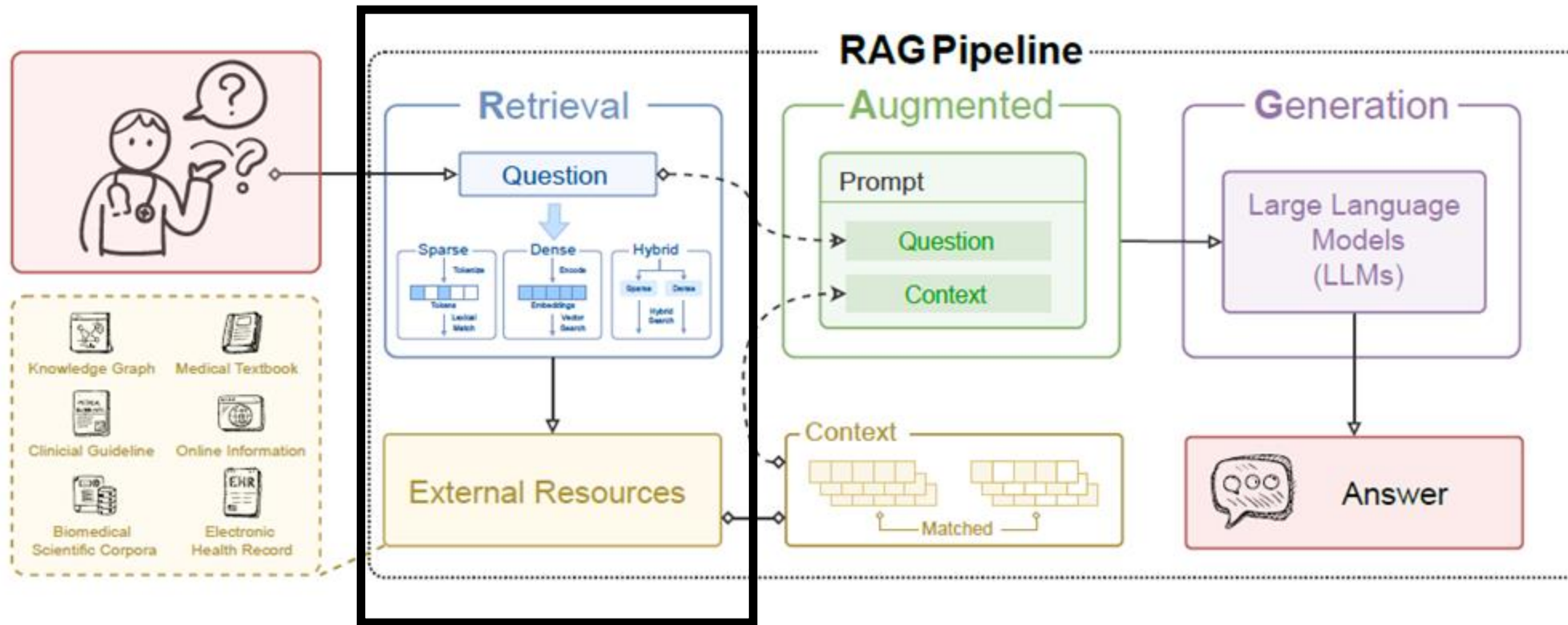
How do we retrieve relevant evidence?

- How do humans retrieve relevant evidence?
 - Do creative web search
 - Experiment with several searches
 - Put yourself in an author's shoes; what phrases might they have used?
 - Specifically search Google Scholar, etc.
 - Track down related work (once you have a relevant paper)
 - Follow the bibliography to earlier papers
 - See who else has cited the work
 - Read each paper to find related work
- How do machines retrieve relevant evidence?
 - Construct knowledge graphs for each paper
 - Search for keywords based on the constructed knowledge graph
 - Use papers' references and citation networks to identify related papers
 - Search for semantically relevant sentences based on embedding similarity and overlapping knowledge graphs

Different types of knowledge retrieval



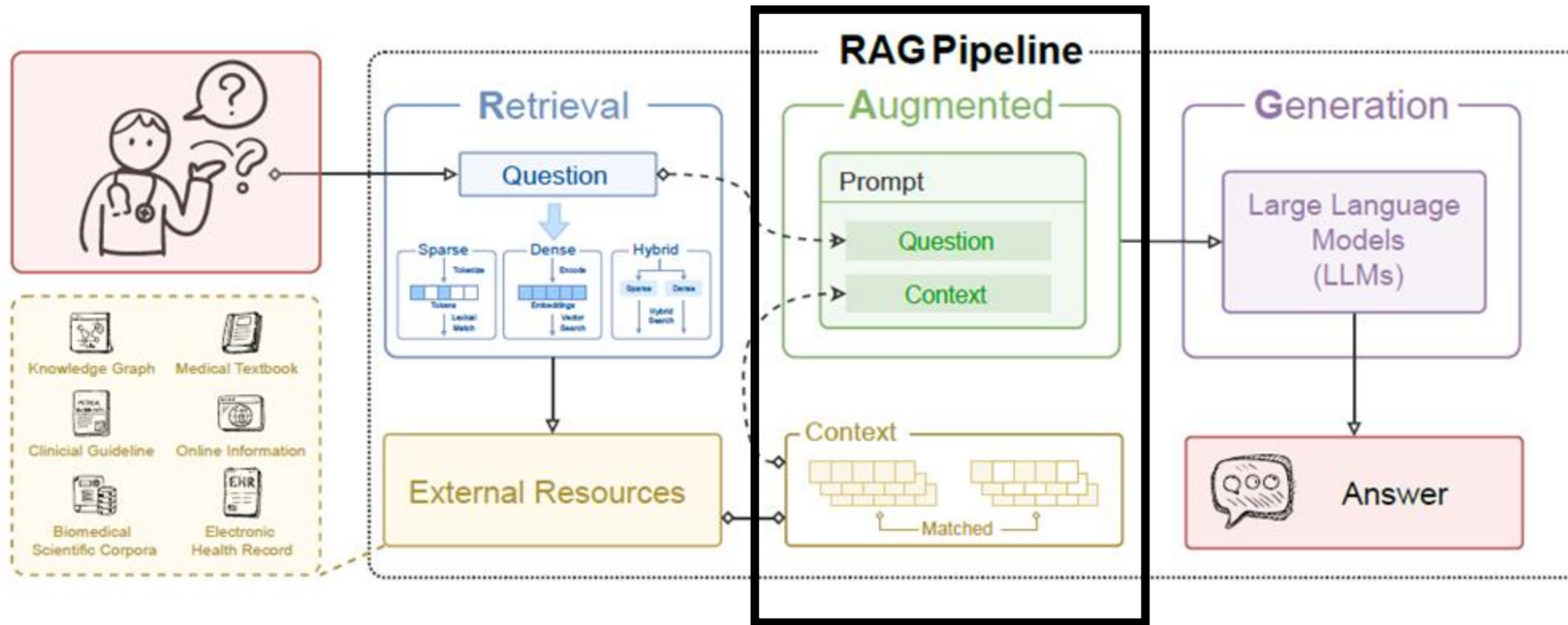
Standard retrieval pipeline



- Step 1: Retrieval
- Searches and provides information relevant to the query

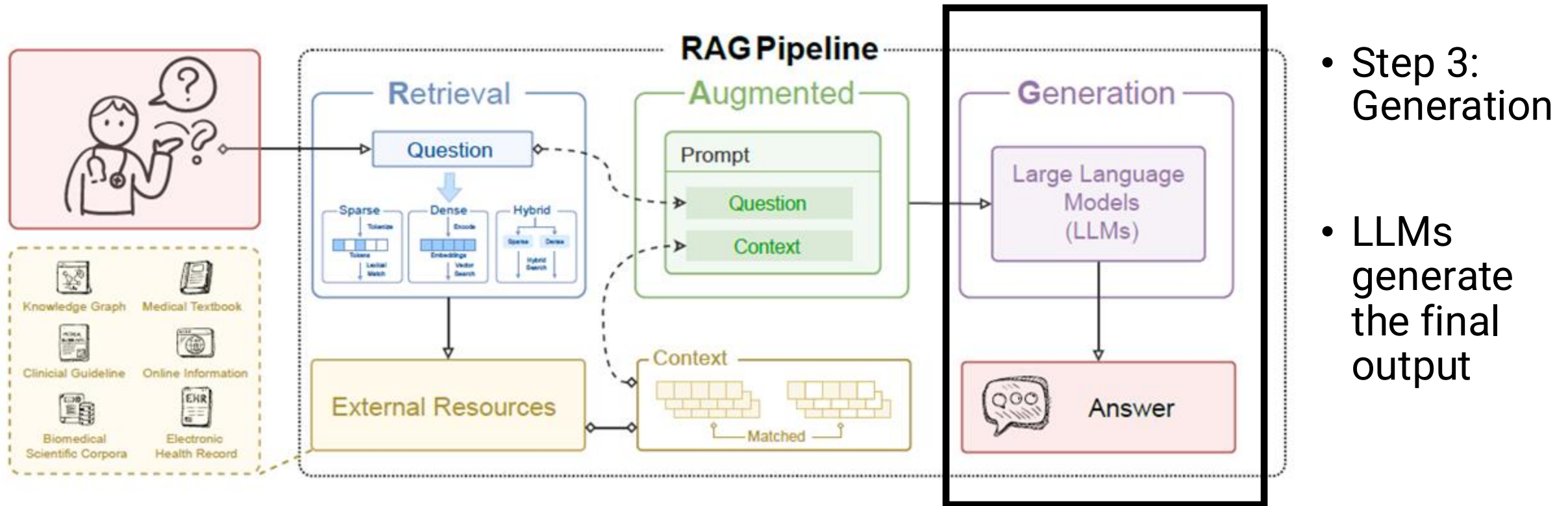
Yang, R., Wong, M.Y.H., Li, H., Li, X., Zhu, W., Liao, J., Yu, K., Liew, J.C.K., Xuan, W., Chen, Y. Ke, Y., ... and Liu, N., 2025. Retrieval-Augmented Generation in Medicine: A Scoping Review of Technical Implementations, Clinical Applications, and Ethical Considerations. *arXiv preprint arXiv:2511.05901*.

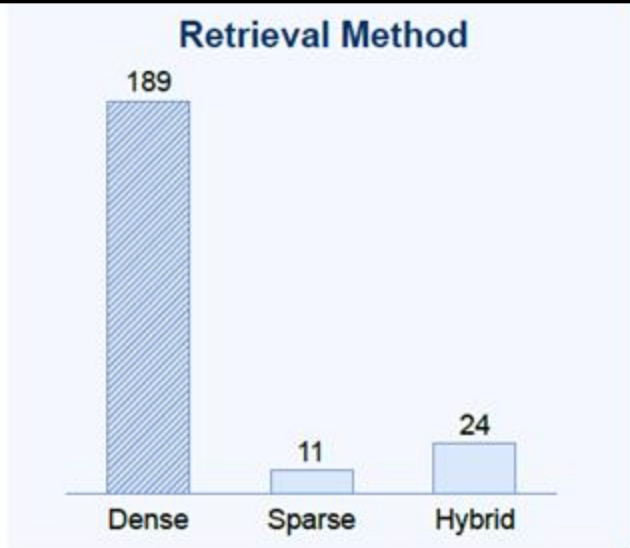
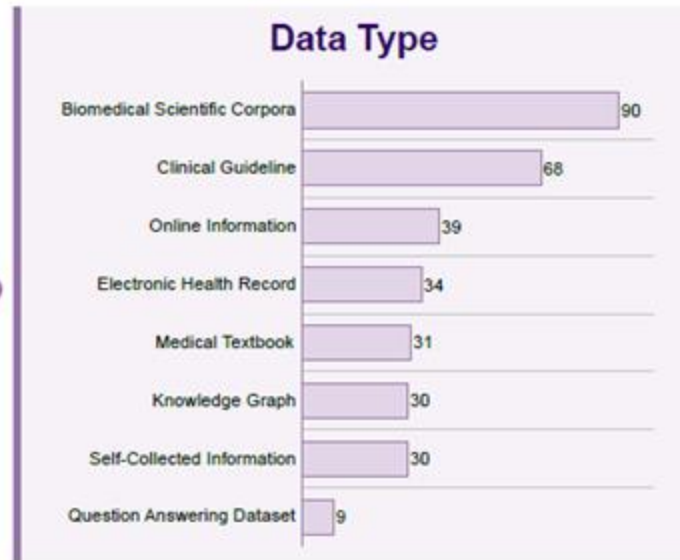
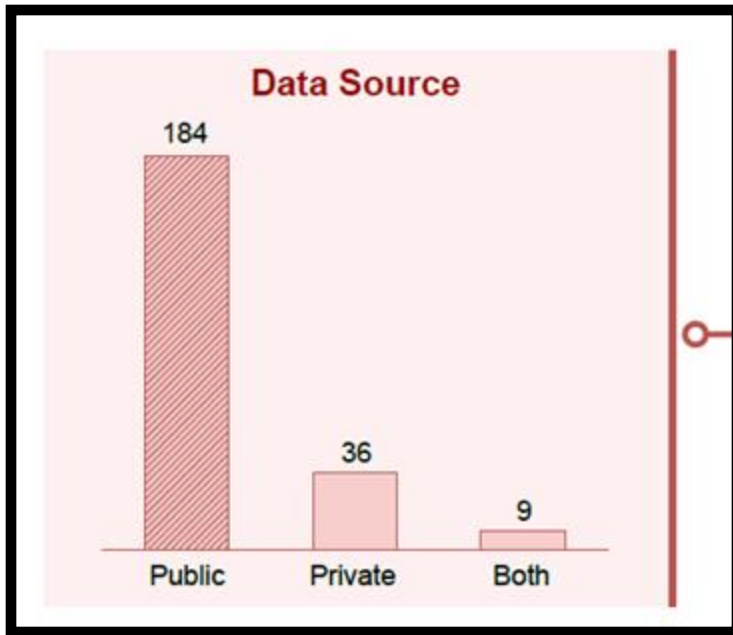
Standard retrieval pipeline



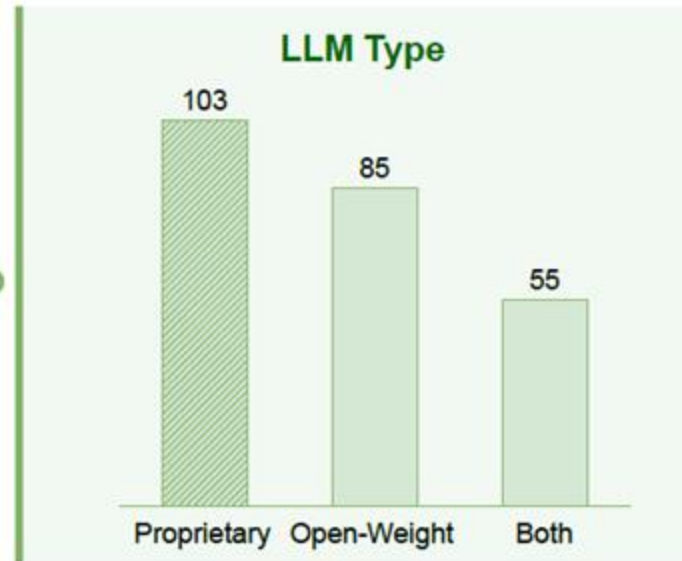
- Step 2: Augmentation
- Prompt + retrieved context to LLMs

Standard retrieval pipeline





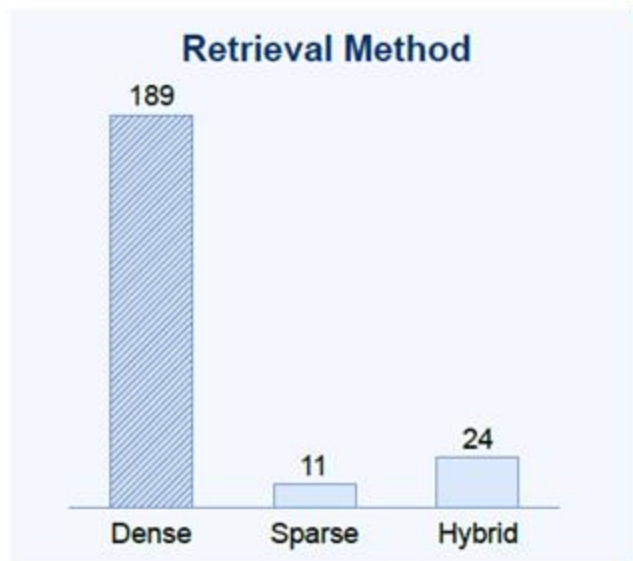
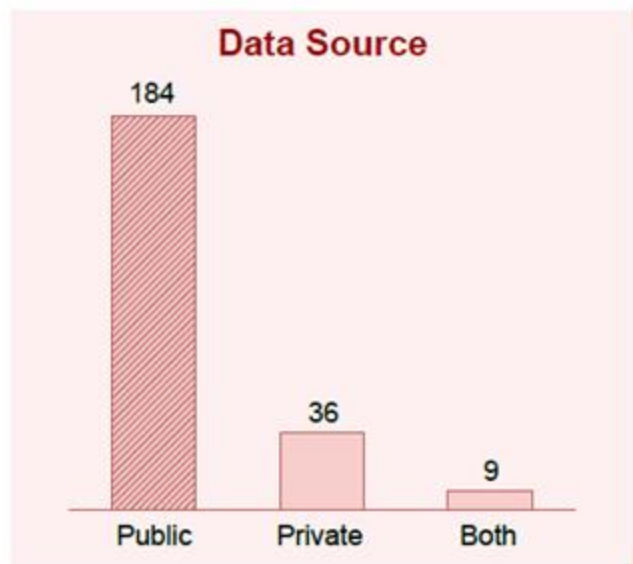
RAG



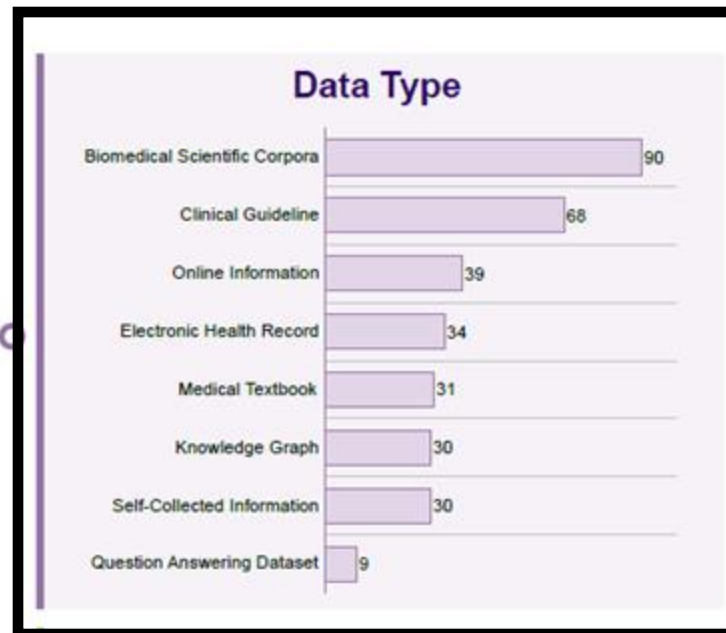
- 251 relevant studies:

Data source

- 80% used public data
- 15% used private
- 5% used both



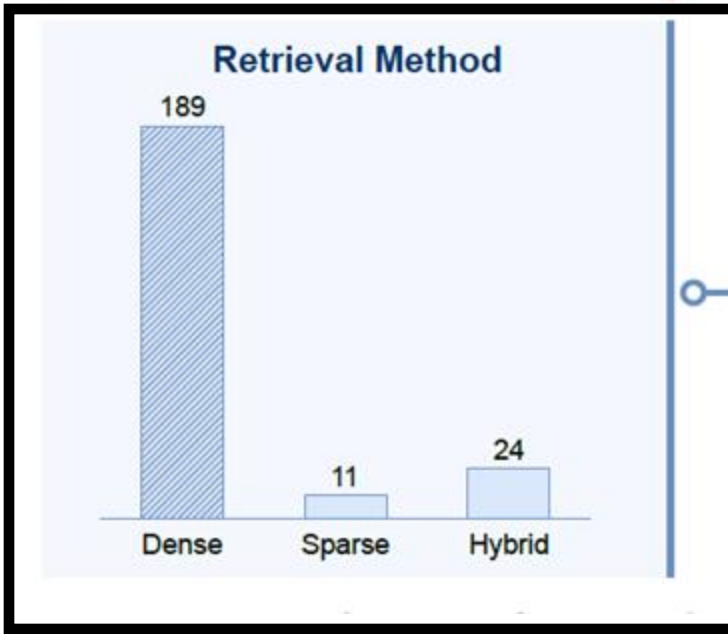
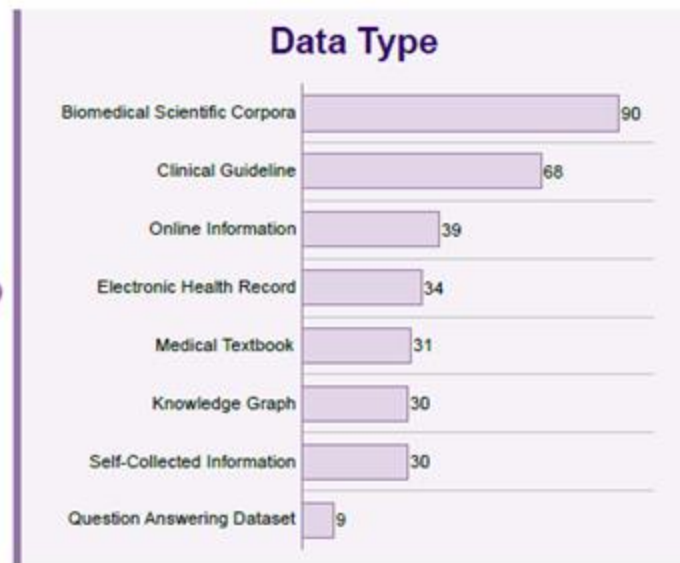
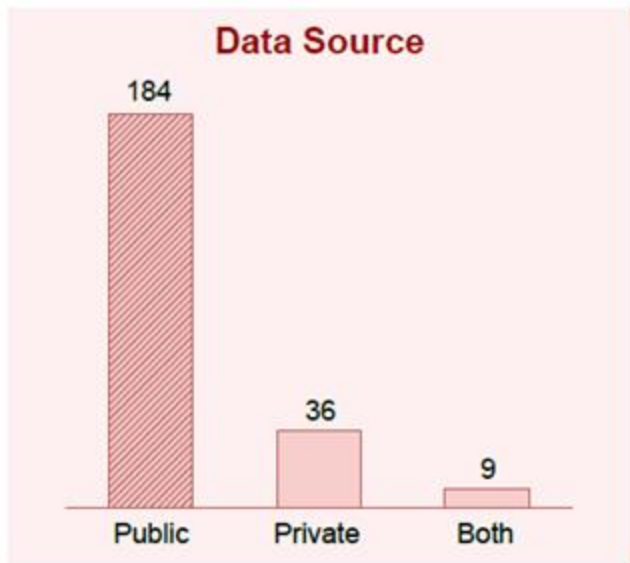
RAG



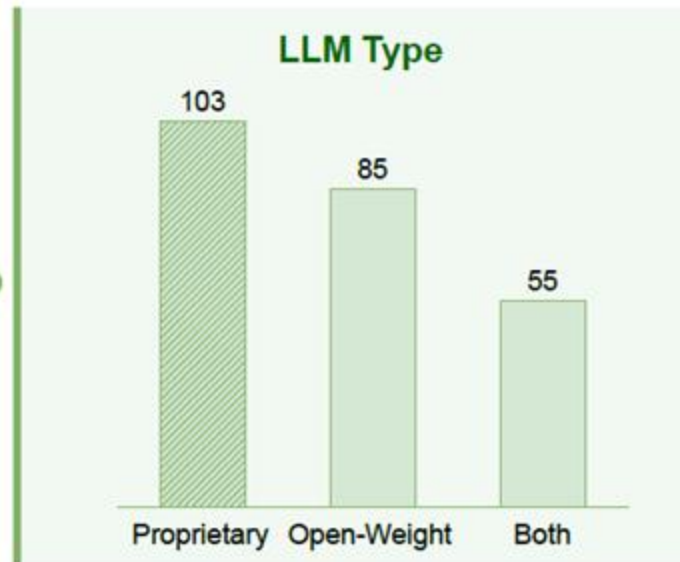
- 251 relevant studies:

Data type

- Biomedical literature
- Clinical guidelines
- Clinical notes
- Textbooks
- Knowledge graphs
- Others



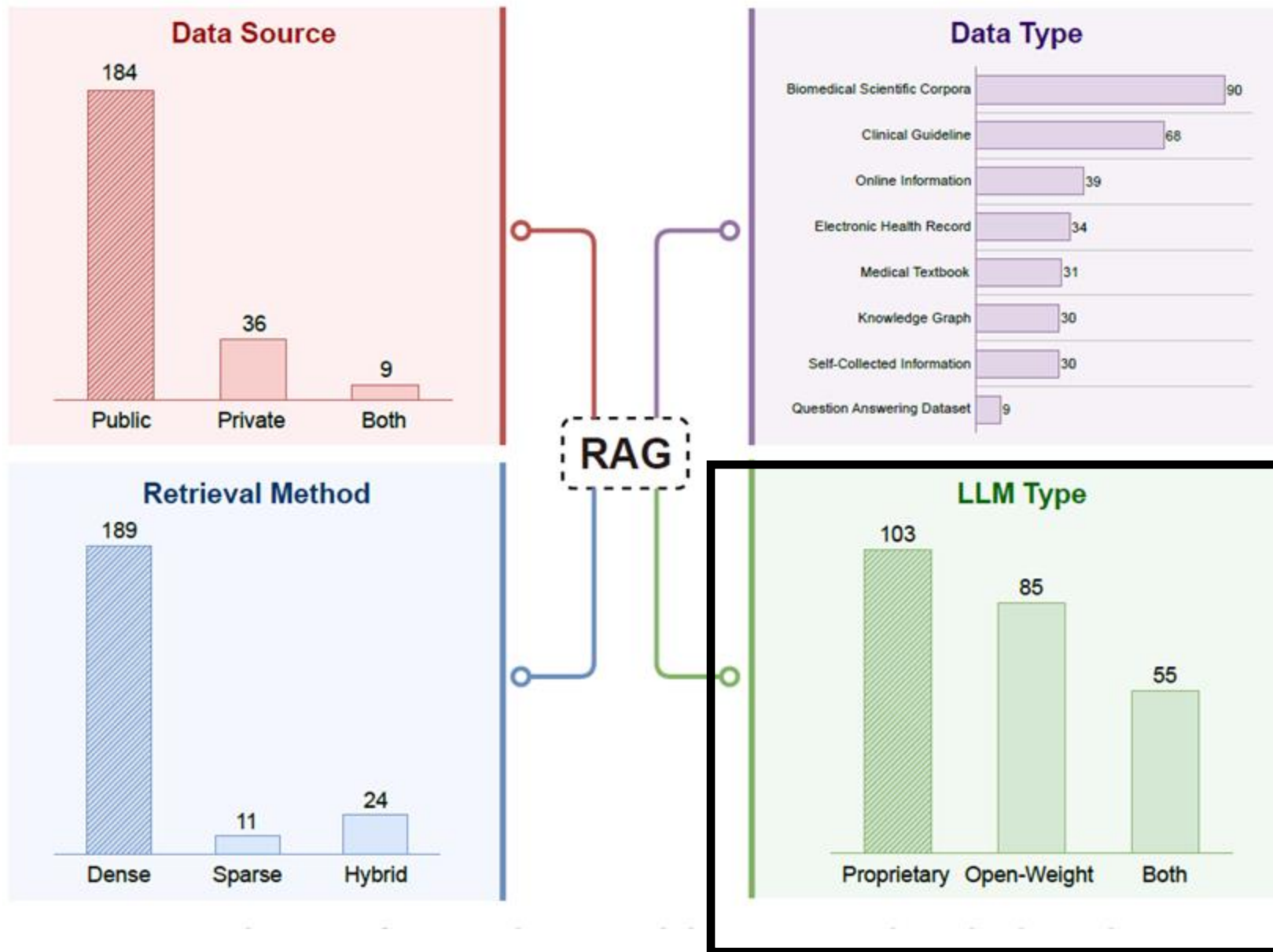
RAG



- 251 relevant studies:

Retrieval methods

- Mostly dense (embedding-based)



- 251 relevant studies:

LLM backbones

- Both closed-sourced and open-weighted are commonly used

Publication Date	Framework	Retriever(s)	LLM(s)	Knowledge Sources	Application(s)
Jun 2023	ChatDoctor ⁸⁶	Lexical matching	Llama	Wikipedia, MedlinePlus	QA
Jul 2023	accGPT ⁸⁷	Text-embedding-ada-002 ⁸⁸	GPT-3.5	Clinical guidelines	Imaging recommendation
Aug 2023	FraCChat ⁸⁹	Text-embedding-ada-002	GPT-3.5, GPT-4	AO/OTA Fracture and Dislocation Classification Compendium (2018)	AO code identification
Dec 2023	Clinfo.ai ⁹⁰	BM25, Entrez API ⁹¹	GPT-3.5, GPT-4	PubMed	QA, Summarization
Jan 2024	Almanac ²⁹	Text-embedding-ada-002	GPT-4	PubMed, Clinical guidelines, etc.	QA
Jan 2024	RALL ⁹²	DPR ⁶³ , etc.	GPT-4, Llama-2	Wikipedia, UMLS	Lay language generation
May 2024	Ferber et al. ⁹³	Text-embedding-ada-002	GPT-4	Clinical guidelines	QA
Jun 2024	RECTIFIER ³⁰	Text-embedding-ada-002	GPT-3.5, GPT-4	Clinical guidelines	Clinical trial screening
Jul 2024	Self-BioRAG ⁷⁶	MedCPT	Llama-2	PubMed, PMC, Clinical guidelines, Textbooks	QA
Jul 2024	DocOA ⁹⁴	Not specified	GPT-3.5	PubMed, Clinical guidelines	Osteoarthritis management
Aug 2024	MedRAG ³⁹	BM25, MedCPT, Contriever ⁹⁵ , SPECTER ⁹⁶	GPT-3.5, GPT-4, Llama-2, Mixtral-8x7B, etc.	PubMed, StatPearls, Textbooks, Wikipedia	QA
Aug 2024	Alkhalaf et al. ⁹⁷	BM25	Llama-2	EHR	Summarization
Sep 2024	ChatZOC ⁵⁰	BM25	Baichuan	Clinical guidelines, FAQs	Ophthalmology QA
Sep 2024	Benfenati et al. ⁹⁸	BM25, GTE ⁹⁹	GPT-3.5, Mistral	Nutrigenetic polymorphism dataset, PubMed-derived nutrition-gene studies	Nutrigenetics QA
Oct 2024	Bora Cuayáhuitl ¹⁰⁰	and GTR-T5-Large	Llama-2, Mistral	Textbooks, Journals	QA
Dec 2024	EyeGPT ¹⁰¹	all-MiniLM-L6-v2 ¹⁰²	Llama-2	Textbooks, Custom Database	Ophthalmology QA

- A closer look at published studies using LLM-RAG in medicine up to August 2025

Kim, H., Sohn, J., Gilson, A., Cochran-Caggiano, N., Applebaum, S., Jin, H., Park, S., Park, Y., Park, J., Choi, S. Contreras, B.A.H., ..., Chen, Q., 2025. Rethinking Retrieval-Augmented Generation for Medicine: A Large-Scale, Systematic Expert Evaluation and Practical Insights. *arXiv preprint arXiv:2511.06738*.



Jan 2025	i-MedRAG ¹⁰³	MedCPT	GPT-4	StatPearls, Textbooks	QA
Jan 2025	Azimi et al. ¹⁰⁰	Titan Text Embeddings V2 ¹⁰⁴	GPT-4o, Claude 3.5 Sonnet	Registered Dietitian (RD) exam questions across four nutrition domains (Academy of Nutrition and Dietetics guidelines and references)	Dietetic QA
Feb 2025	Bailicai ¹⁰⁵	MedCPT	Qwen ¹⁰⁶ , Llama-2	PubMed, StatPearls, Textbooks, Wikipedia	QA
Mar 2025	Woo et al. ¹⁰⁷	Internal retriever	GPT-3.5, GPT-4, Claude 3, Llama-3, Mistral-8×7B	Clinical guidelines	Clinical decision support
Apr 2025	RAG squared ⁴⁰	MedCPT	GPT-4o, Llama-3, etc.	PubMed, PMC, Clinical guidelines, Textbooks	QA
May 2025	BriefContext ¹⁰⁸	BM25, MedCPT	GPT-4o, Llama-3.1	PubMed, Textbooks	Summarization
May 2025	MKRAG ¹⁰⁹	Contriever, etc.	Vicuna-7B ¹¹⁰	Disease Database	Knowledge graph QA
May 2025	Yang et al. ¹¹¹	Elasticsearch, BERTv2 ¹¹²	Col- GPT-4, Claude 3	Internal, expert-reviewed hospital documents	Clinical decision support
Jul 2025	MedGraphRAG ¹¹³	Graph-based retrieval	GPT-4, Gemini 1.0 Pro, Llama-2, Llama-3	MIMIC-IV, FakeHealth, PubMed, Health, UMLS Graph	QA
Jul 2025	RadioRAG ¹¹⁴	Text-embedding-ada-002	GPT-4o mini	Radiopaedia	Case retrieval, Report generation
Jul 2025	Wada et al. ¹¹⁵	Text-embedding-3-large ⁸⁸ , Lexical matching	GPT-4o	Clinical guidelines (ACR Manual, ESUR Guidelines, etc.)	Safety consultation
Jul 2025	MedOmniKB ¹¹⁶	MedCPT, Graph-based retrieval	GPT-4, Gemini 1.5	PubMed, Clinical guidelines, Textbooks, Wikipedia, UMLS, DrugBank	QA
Aug 2025	MedCoT-RAG ¹¹⁷	MedCPT	Llama-3.1	PubMed, StatPearls, Textbooks, Wikipedia	QA



LLM-RAG implementation examples: API-based

Bioinformatics, 2024, **40**(2), btae075
<https://doi.org/10.1093/bioinformatics/btae075>
Advance Access Publication Date: 10 February 2024
Original Paper



Data and text mining

GeneGPT: augmenting large language models with domain tools for improved access to biomedical information

Qiao Jin ¹, Yifan Yang¹, Qingyu Chen ¹, Zhiyong Lu ^{1,*}

¹National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, United States

*Corresponding author. National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, United States. Tel: 301-594-7089, E-mail: zhiyong.lu@nih.gov (Z.L.)

Associate Editor: Jonathan Wren

Abstract

Motivation: While large language models (LLMs) have been successfully applied to various tasks, they still face challenges with hallucinations. Augmenting LLMs with domain-specific tools such as database utilities can facilitate easier and more precise access to specialized knowledge. In this article, we present GeneGPT, a novel method for teaching LLMs to use the Web APIs of the National Center for Biotechnology Information (NCBI) for answering genomics questions. Specifically, we prompt Codex to solve the GeneTuring tests with NCBI Web APIs by in-context learning and an augmented decoding algorithm that can detect and execute API calls.

Results: Experimental results show that GeneGPT achieves state-of-the-art performance on eight tasks in the GeneTuring benchmark with an average score of 0.83, largely surpassing retrieval-augmented LLMs such as the new Bing (0.44), biomedical LLMs such as BioMedLM (0.08) and BioGPT (0.04), as well as GPT-3 (0.16) and ChatGPT (0.12). Our further analyses suggest that: First, API demonstrations have good cross-task generalizability and are more useful than documentations for in-context learning; second, GeneGPT can generalize to longer chains of API calls and answer multi-hop questions in GeneHop, a novel dataset introduced in this work; finally, different types of errors are enriched in different tasks, providing valuable insights for future improvements.

Availability and implementation: The GeneGPT code and data are publicly available at <https://github.com/ncbi/GeneGPT>.

Jin, Q., Yang, Y., Chen, Q. and Lu, Z., 2024. Genegpt: Augmenting large language models with domain tools for improved access to biomedical information. *Bioinformatics*, 40(2), p.btae075.

GeneGPT prompt

Hello. Your task is to use NCBI APIs to answer genomic questions. There are two types of Web APIs you can use: Eutils and BLAST.

Documentation 1

You can call Eutils by:

```
"[https://eutils.ncbi.nlm.nih.gov/entrez/eutils/{esearch|efetch|esummary}.fcgi?db={genes|snplolim}&retmax={}&termid={}]". {API_DOC}
```

Documentation 2

For DNA sequences, you can use BLAST by:

```
"[https://blast.ncbi.nlm.nih.gov/blast/Blast.cgi?CMD={Put|Get}&PROGRAM=blastn&MEGABLAST=on&DATABASE=nt&FORMAT_TYPE={XML|Text}&QUERY={sequence}&HITLIST_SIZE={max_hit_size}]". {API_DOC}
```

Demonstration 1

Question: What is the official gene symbol of LMP10?

```
[https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=genes&retmax=3&retmode=json&term=LMP10]->[{API_CALL}]
```

```
[https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=genes&retmax=3&retmode=json&id={ids}]->[{API_CALL}]
```

Answer: PSMB10

Demonstration 2

Question: Which gene is SNP rs1217074595 associated with?

```
[https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esummary.fcgi?db=snp&retmax=3&retmode=json&id=1217074595]->[{API_CALL}]
```

Answer: LINC01270

Demonstration 3

Question: What are genes related to Meesmann corneal dystrophy?

```
[https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=omim&retmax=15&term=Meesmann+corneal+dystrophy]->[{API_CALL}]
```

```
[https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esummary.fcgi?db=omim&retmax=15&retmode=json&id={ids}]->[{API_CALL}]
```

Answer: KRT12, KRT3

Demonstration 4

Question: Align the DNA sequence to the human genome:ATTCT[...]TCTCT

```
[https://blast.ncbi.nlm.nih.gov/blast/Blast.cgi?CMD=Put&PROGRAM=blastn&MEGABLAST=on&DATABASE=nt&FORMAT_TYPE=XML&QUERY=ATTCT[...]TCTCT&HITLIST_SIZE=5]->[{API_CALL}]
```

```
[https://blast.ncbi.nlm.nih.gov/blast/Blast.cgi?CMD=Get&FORMAT_TYPE=Text&RID={RID}]->[{API_CALL}]
```

Answer: chr15:91950805-91950932

GeneTuring Question

Which chromosome does SNP rs1430464868 locate on human genome?

GeneGPT Output

```
[https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esummary.fcgi?db=snp&retmax=3&retmode=json&id=1430464868]->[{"header":{"type":"esummary","version":"0.3"},"result":{"uids":["1430464868"],"1430464868":{"uid":"1430464868","snp_id":1430464868,"allele_origin":"","global_mafs":[{"study":"GnomAD","freq":"C=0.000007/1"},{"study":"ALFA","freq":"C=0./0"}],"global_population":"","global_sample_size":"","suspected":"","clinical_significance":"","genes":[],"acc":"NC_000013.11","chr":"13","handle":["*TRUNCATED*"],"merged_sort":"0"}}}]
```

Answer: chr13

GeneHop Question

What is the function of the gene associated with SNP rs1241371358? Let's decompose the question to sub-questions and solve them step by step.

GeneGPT Output

Sub-question 1: Which gene is SNP rs1241371358 associated with?

```
[https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esummary.fcgi?db=snp&retmax=10&retmode=json&id=1241371358]->[{"header":{"type":"esummary","version":"0.3"},"result":{"uids":["1241371358"],"1241371358":{"uid":"1241371358","snp_id":1241371358,"allele_origin":"","global_mafs":[{"study":"GnomAD","freq":"C=0.000007/1"},{"study":"ALFA","freq":"C=0./0"}],"global_population":"","global_sample_size":"","suspected":"","clinical_significance":"","genes":[],"acc":"NC_000013.11","chr":"13","handle":["*TRUNCATED*"],"merged_sort":"0"}}}]
```

Answer: LRRC23

Sub-question 2: What is the function of LRRC23?

```
[https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esummary.fcgi?db=genes&retmax=10&retmode=json&id=10233]->[{"header":{"type":"esummary","version":"0.3"},"result":{"uids":["10233"],"10233":{"uid":"10233","gene_id":10233,"symbol":"LRRC23","description":"Predicted to be active in cytosol."},"merged_sort":"0"}}}]
```

Answer: Predicted to be active in cytosol.

Instructions

- Ask LLMs to use APIs for retrieval



GeneGPT prompt

Hello. Your task is to use NCBI APIs to answer genomic questions. There are two types of Web APIs you can use: Eutils and BLAST.

You can call Eutils by:

```
"[https://eutils.ncbi.nlm.nih.gov/entrez/eutils/{esearch|efetch|esummary}.fcgi?db={genes|snplolim}&retmax={}&termid={}]". {API_DOC}
```

Documentation 2

For DNA sequences, you can use BLAST by:

```
"[https://blast.ncbi.nlm.nih.gov/blast/Blast.cgi?CMD={Put|Get}&PROGRAM={blastn|MEGABLAST}&on&DATABASE={nt}&FORMAT_TYPE={XML|Text}&QUERY={sequence}&HITLIST_SIZE={max_hit_size}]". {API_DOC}
```

Demonstration 1

Question: What is the official gene symbol of LMP10?

```
[https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=genes&retmax=3&retmode=json&term=LMP10]->[{API_CALL}]
```

```
[https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=genes&retmax=3&retmode=json&id={ids}]->[{API_CALL}]
```

Answer: PSMB10

Demonstration 2

Question: Which gene is SNP rs1217074595 associated with?

```
[https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esummary.fcgi?db=snp&retmax=3&retmode=json&id=1217074595]->[{API_CALL}]
```

Answer: LINC01270

Demonstration 3

Question: What are genes related to Meesmann corneal dystrophy?

```
[https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=omim&retmax=15&term=Meesmann+corneal+dystrophy]->[{API_CALL}]
```

```
[https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esummary.fcgi?db=omim&retmax=15&retmode=json&id={ids}]->[{API_CALL}]
```

Answer: KRT12, KRT3

Demonstration 4

Question: Align the DNA sequence to the human genome:ATTCT[...]TCTCT

```
[https://blast.ncbi.nlm.nih.gov/blast/Blast.cgi?CMD=Put&PROGRAM=blastn&MEGABLAST=on&DATABASE=nt&FORMAT_TYPE=XML&QUERY=ATTCT[...]TCTCT&HITLIST_SIZE=5]->[{API_CALL}]
```

```
[https://blast.ncbi.nlm.nih.gov/blast/Blast.cgi?CMD=Get&FORMAT_TYPE=Text&RID={RID}]->[{API_CALL}]
```

Answer: chr15:91950805-91950932

GeneTuring Question

Which chromosome does SNP rs1430464868 locate on human genome?

GeneGPT Output

```
[https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esummary.fcgi?db=snp&retmax=3&retmode=json&id=1430464868]->[{{"header":{"type":"esummary","version":"0.3"},"result":{"uids":["1430464868"],"1430464868":{"uid":"1430464868","snp_id":1430464868,"allele_origin":"","global_mafs":{"study":"GnomAD","freq":"C=0.000007/1"},"study":"ALFA","freq":"C=0./0"},"global_population":"","global_sample_size":"","suspected":"","clinical_significance":"","genes":[],"acc":"NC_000013.11","chr":"13","handle":["*TRUNCATED*"],"merged_sort":"0"}}}]
```

Answer: chr13

GeneHop Question

What is the function of the gene associated with SNP rs1241371358? Let's decompose the question to sub-questions and solve them step by step.

GeneGPT Output

Sub-question 1: Which gene is SNP rs1241371358 associated with?

```
[https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esummary.fcgi?db=snp&retmax=10&retmode=json&id=1241371358]->[{{"header":{"type":"esummary","version":"0.3"},"result":{"uids":["1241371358"],"1241371358":{"uid":"1241371358","snp_id":1241371358,"allele_origin":"","global_mafs":{"study":"GnomAD","freq":"C=0.000007/1"},"study":"ALFA","freq":"C=0./0"},"global_population":"","global_sample_size":"","suspected":"","clinical_significance":"","genes":["LRRRC23"],"acc":"NC_000013.11","chr":"13","handle":["*TRUNCATED*"],"merged_sort":"0"}}}]
```

Answer: LRRRC23

Sub-question 2: What is the function of LRRRC23?

```
[https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esummary.fcgi?db=genes&retmax=10&retmode=json&id=10233]->[{{"header":{"type":"esummary","version":"0.3"},"result":{"uids":["10233"],"10233":{"uid":"10233","gene_id":10233,"symbol":"LRRRC23","description":"Predicted to be active in cytosol"},"acc":"NC_000013.11","chr":"13","handle":["*TRUNCATED*"],"merged_sort":"0"}}}]
```

Answer: Predicted to be active in cytosol.

Documentation

- Provide descriptions of the API functionality, general syntax, and argument choices.



GeneGPT prompt

Hello. Your task is to use NCBI APIs to answer genomic questions. There are two types of Web APIs you can use: Eutils and BLAST.

Documentation 1

You can call Eutils by:

```
"[https://eutils.ncbi.nlm.nih.gov/entrez/eutils/{esearch|efetch|esummary}.fcgi?db={genes|snplolim}&retmax={}&termid={}]". {API_DOC}
```

Documentation 2

For DNA sequences, you can use BLAST by:

```
"[https://blast.ncbi.nlm.nih.gov/blast/Blast.cgi?CMD={Put|Get}&PROGRAM=blastn&MEGABLAST=on&DATABASE=nt&FORMAT_TYPE={XML|Text}&QUERY={sequence}&HITLIST_SIZE={fmax|hit_size}]". {API_DOC}
```

Demonstration 1

Question: What is the official gene symbol of LMP10?

```
[https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=genes&retmax=3&retmode=json&term=LMP10]->[{API_CALL}]
```

```
[https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=genes&retmax=3&retmode=json&id={ids}]->[{API_CALL}]
```

Answer: PSMB10

Demonstration 2

Question: Which gene is SNP rs1217074595 associated with?

```
[https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esummary.fcgi?db=snp&retmax=3&retmode=json&id=1217074595]->[{API_CALL}]
```

Answer: LINC01270

Demonstration 3

Question: What are genes related to Meesmann corneal dystrophy?

```
[https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=omim&retmax=15&term=Meesmann+corneal+dystrophy]->[{API_CALL}]
```

```
[https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esummary.fcgi?db=omim&retmax=15&retmode=json&id={ids}]->[{API_CALL}]
```

Answer: KRT12, KRT3

Demonstration 4

Question: Align the DNA sequence to the human genome:ATTCT[...TCTCT

```
[https://blast.ncbi.nlm.nih.gov/blast/Blast.cgi?CMD=Put&PROGRAM=blastn&MEGABLAST=on&DATABASE=nt&FORMAT_TYPE=XML&QUERY=ATTCT[...TCTCT&HITLIST_SIZE=5]->[{API_CALL}]
```

```
[https://blast.ncbi.nlm.nih.gov/blast/Blast.cgi?CMD=Get&FORMAT_TYPE=Text&RID={RID}]->[{API_CALL}]
```

Answer: chr15:91950805-91950932

GeneTuring Question

Which chromosome does SNP rs1430464868 locate on human genome?

GeneGPT Output

```
[https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esummary.fcgi?db=snp&retmax=3&retmode=json&id=1430464868]->[{"header":{"type":"esummary","version":"0.3"},"result":{"uids":["1430464868"],"1430464868":{"uid":"1430464868","snp_id":1430464868,"allele_origin":"","global_mafs":[{"study":"GnomAD","freq":"C=0.000007/1"}],"study":"ALFA","freq":"C=0./0"},"global_population":"","global_sample_size":"","suspected":"","clinical_significance":"","genes":[],"acc":"NC_000013.11","chr":"13","handle":["*TRUNCATED*"],"merged_sort":"0"}}]
```

Answer: chr13

GeneHop Question

What is the function of the gene associated with SNP rs1241371358? Let's decompose the question to sub-questions and solve them step by step.

GeneGPT Output

Sub-question 1: Which gene is SNP rs1241371358 associated with?

```
[https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esummary.fcgi?db=snp&retmax=10&retmode=json&id=1241371358]->[{"header":{"type":"esummary","version":"0.3"},"result":{"uids":["1241371358"],"1241371358":{"uid":"1241371358","merged_sort":"0"}}]
```

Answer: LRRC23

Sub-question 2: What is the function of LRRC23?

```
[https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esummary.fcgi?db=genes&retmax=10&retmode=json&id=10233]->[{"header":{"type":"esummary","version":"0.3"},"result":{"uids":["10233"],"10233":{"uid":"10233","merged_sort":"0"}}]
```

Answer: Predicted to be active in cytosol.

Demonstration

- Provide examples of using NCBI Web APIs to solve questions.

Results

GeneTuring task	GPT-2	BioGPT	BioMedLM	GPT-3	ChatGPT	New Bing	GeneGPT (ours)			
							-full	-slim	-turbo	-lang
Nomenclature										
Gene alias	0.00	0.00	0.04	0.09	0.07	0.66	<u>0.80</u> *	0.84 *	0.64*	0.76*
Gene name conversion	0.00	0.00	0.00	0.00	0.00	<u>0.85</u>	<u>1.00</u>	1.00	1.00	0.02
Average	0.00	0.00	0.02	0.05	0.04	0.76	<u>0.90</u>	0.92	0.82	0.39
Genomic location										
Gene SNP association	0.00	0.00	0.00	0.00	0.00	0.00	1.00 *	1.00	0.96	0.90
Gene location	0.01	0.04	0.12	0.09	0.09	0.61	<u>0.62</u>	0.66	0.54	0.54
SNP location	0.03	<u>0.05</u>	0.01	0.02	0.05	0.01	1.00	<u>0.98</u>	<u>0.98</u>	0.74
Average	0.01	<u>0.03</u>	0.04	0.04	0.05	0.21	<u>0.87</u>	0.88	0.82	0.73
Functional analysis										
Gene disease association	0.00	0.02	0.16	0.34	0.31	0.84	<u>0.76</u> *	0.66	0.63	0.39
Protein-coding genes	0.00	0.18	0.37	0.70	0.54	<u>0.97</u>	0.76	1.00	0.96	0.90
Average	0.00	0.10	0.27	0.52	0.43	0.91	0.76	<u>0.84</u>	0.80	0.65
Sequence alignment										
DNA to human genome	0.02	0.07	0.03	0.00	0.00	0.00	0.44 *	0.44 *	<u>0.42</u> *	0.06*
DNA to multiple species	0.02	0.00	0.00	0.20	0.00	0.00	<u>0.86</u>	0.88	0.88	0.54
Average	0.02	0.04	0.02	0.10	0.00	0.00	<u>0.65</u>	0.66	<u>0.65</u>	0.30
Overall average	0.00	0.04	0.08	0.16	0.12	0.44	<u>0.80</u>	0.83	0.78	0.54

LLM-RAG implementation examples: vector-based

BiomedRAG: A retrieval augmented large language model for biomedicine

Mingchen Li^a, Halil Kilicoglu^b, Hua Xu^c, Rui Zhang^a  

Show more 

 Add to Mendeley  Share  Cite

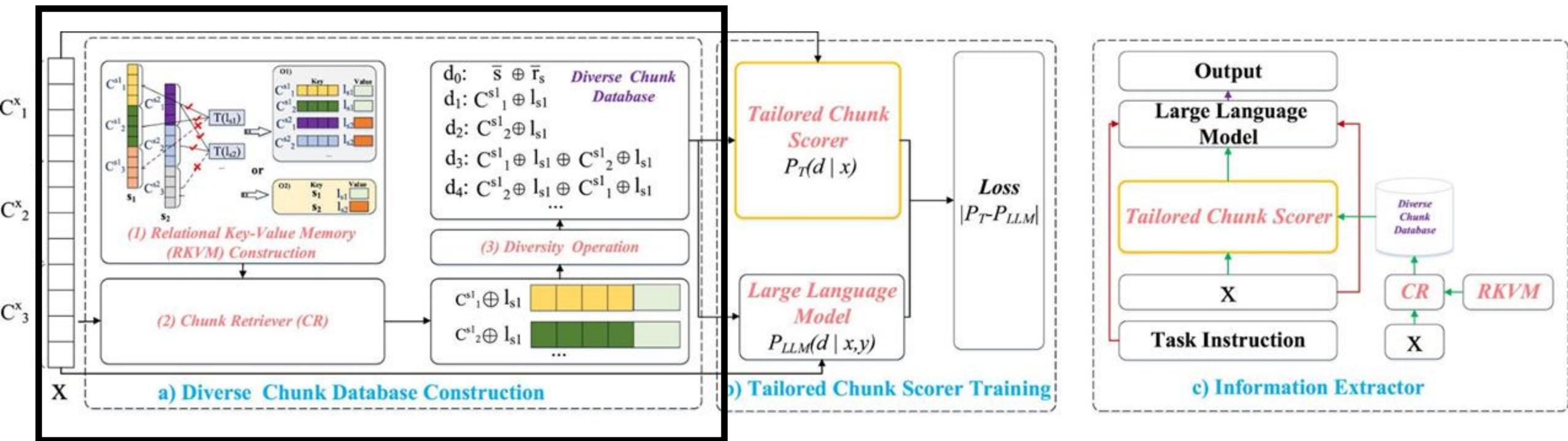
<https://doi.org/10.1016/j.jbi.2024.104769> 

[Get rights and content](#) 

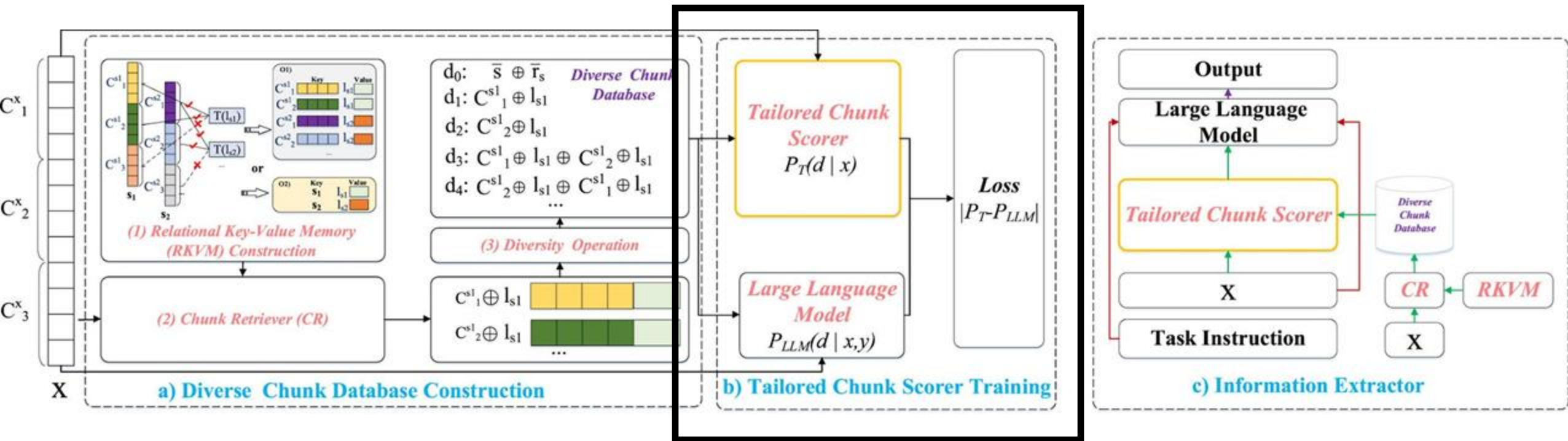
Under a [Creative Commons license](#) 

 Open access

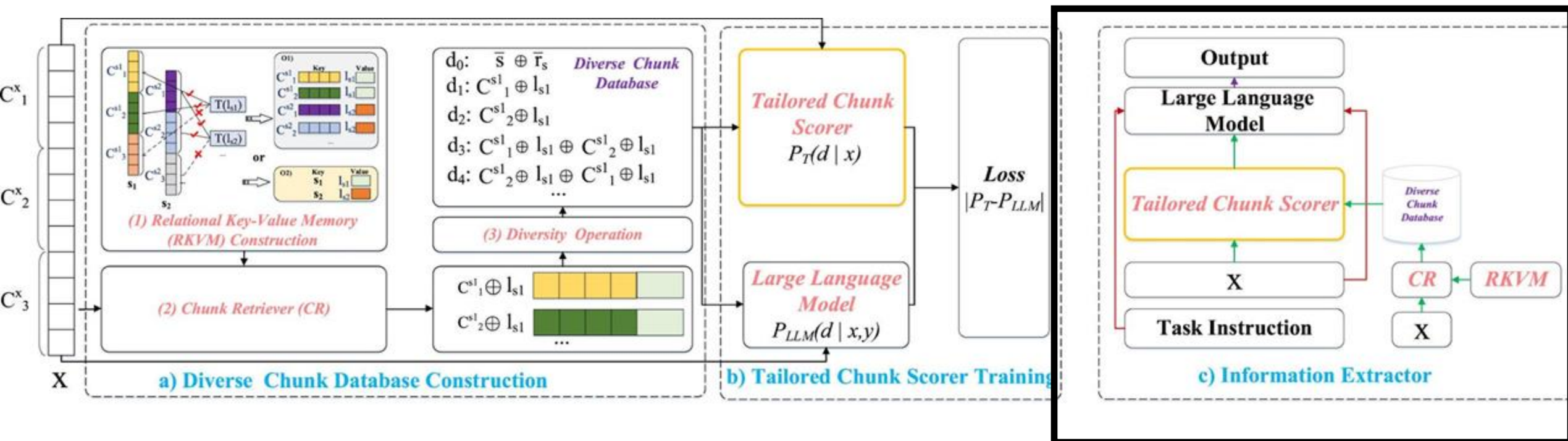
Li, M., Kilicoglu, H., Xu, H. and Zhang, R., 2025. Biomedrag: A retrieval augmented large language model for biomedicine. *Journal of Biomedical Informatics*, 162, p.104769.



- Text $\rightarrow$ chunks $\rightarrow$ vector databases
- For different tasks, the notion of a chunk is different

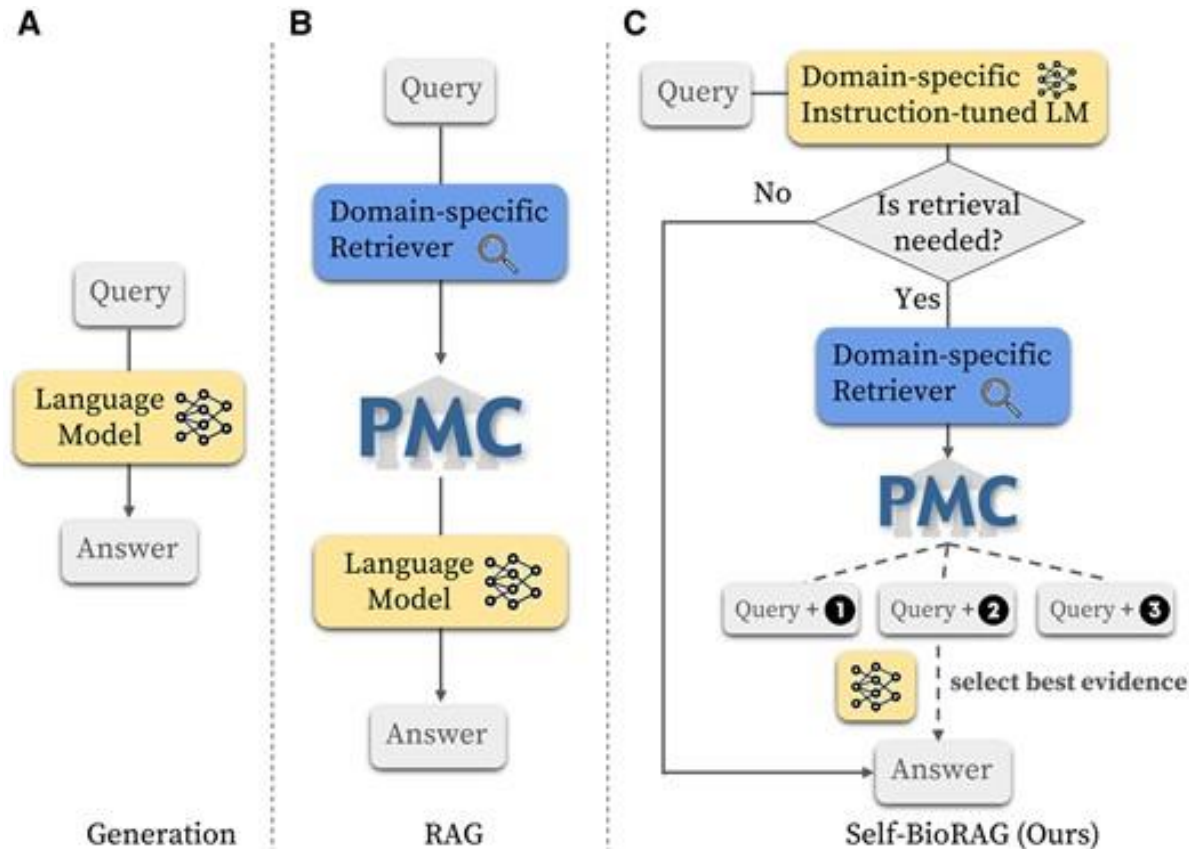


- Trained a scorer to retrieve most relevant chunks given a user query



- The generation step remains consistent

Additional examples



- Decide whether to do RAG
- Self-reflect on multiple retrieved passages

Jeong, M., Sohn, J., Sung, M. and Kang, J., 2024. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. *Bioinformatics*, 40(Supplement_1), pp.i119-i129.

Recent tools for LLM-RAG implementations

Democratizing AI scientists using ToolUniverse

Shanghua Gao¹, Richard Zhu^{1,2,*}, Pengwei Sui^{1,*}, Zhenglun Kong^{1,*}, Sufian Aldogom^{1,*}, Yepeng Huang¹, Ayush Noori¹, Reza Shamji^{1,2}, Krishna Parvataneni³, Theodoros Tsiligkaridis⁴, Marinka Zitnik^{1,5,6,7,‡}

¹Department of Biomedical Informatics, Harvard Medical School, Boston, MA

²Harvard College, Harvard University, Cambridge, MA

³Massachusetts Institute of Technology, Cambridge, MA

⁴MIT Lincoln Laboratory, Lexington, MA

⁵Kempner Institute for the Study of Natural and Artificial Intelligence, Harvard University, Cambridge, MA

⁶Broad Institute of MIT and Harvard, Cambridge, MA

⁷Harvard Data Science Initiative, Cambridge, MA

* Co-second authors

‡ Correspondence: marinka@hms.harvard.edu

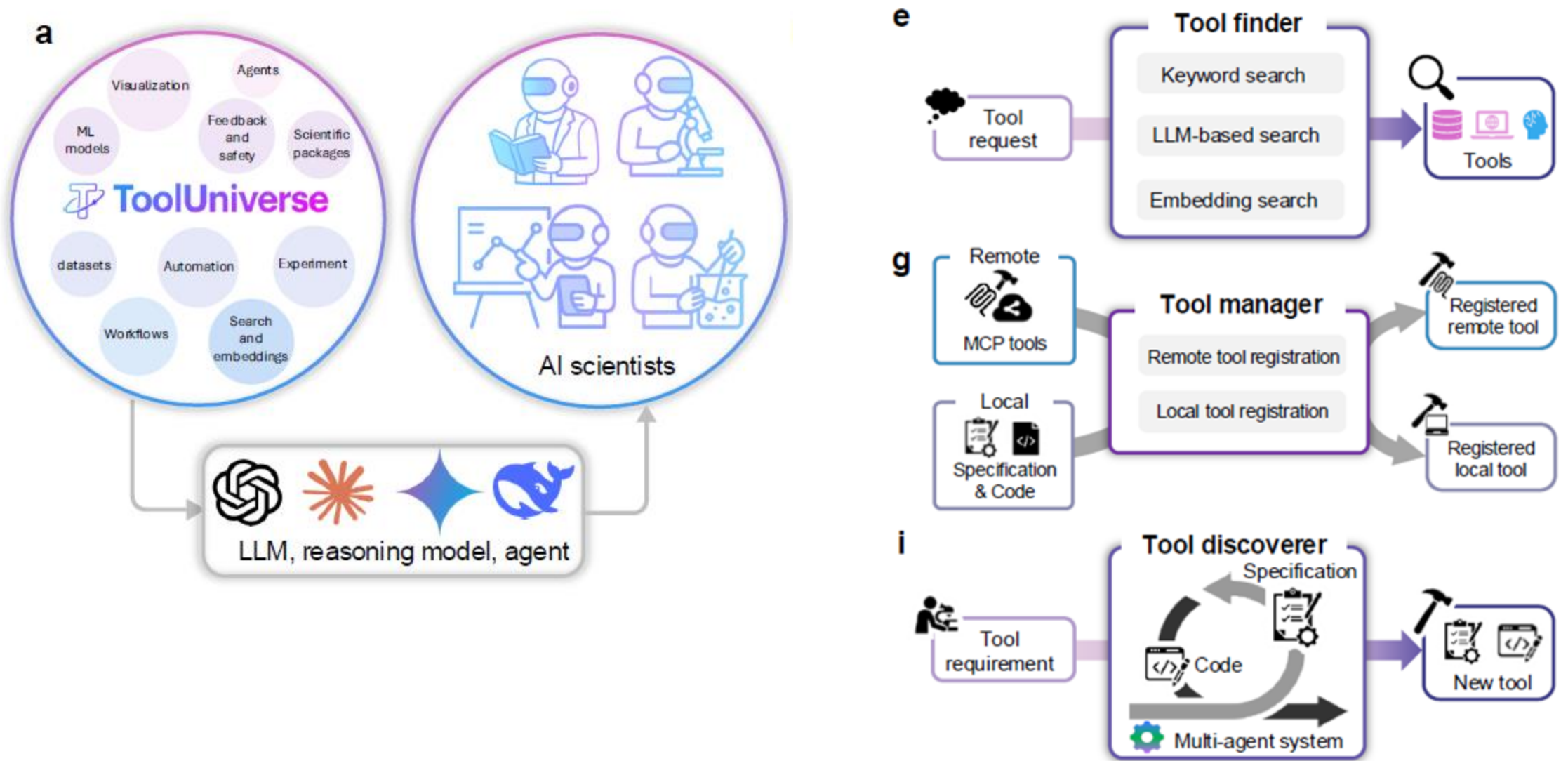
TOOLUNIVERSE web service is at <https://aiscientist.tools>

TOOLUNIVERSE code is at <https://github.com/mims-harvard/ToolUniverse>

TOOLUNIVERSE package is at <https://pypi.org/project/tooluniverse>

AI scientists are emerging computational systems that serve as collaborative partners in discovery. These systems remain difficult to build because they are bespoke, tied to rigid workflows, and lack shared environments that unify tools, data, and analyses into a common ecosystem. In genomics, unified ecosystems have transformed research by enabling interoperability, reuse, and community-driven development; AI scientists require comparable infrastructure. We present TOOLUNIVERSE, an ecosystem for building AI scientists from any language or reasoning model across open- and closed-weight models. TOOLUNIVERSE standardizes how AI scientists identify and call tools by providing more than 600 machine learning models, datasets, APIs, and scientific packages for data analysis, knowledge retrieval, and experimental design. It automatically refines tool interfaces for correct use by AI scientists, generates new tools from natural language descriptions, iteratively optimizes tool

Gao, S., Zhu, R., Sui, P., Kong, Z., Aldogom, S., Huang, Y., Noori, A., Shamji, R., Parvataneni, K., Tsiligkaridis, T. and Zitnik, M., 2025. Democratizing AI scientists using ToolUniverse. *arXiv preprint arXiv:2509.23426*.



LLMs augmented with 600+ tools for retrieval

Tool Specification Schema in TOOLUNIVERSE

Name: The unique identifier for the tool.

Description: A clear and concise summary of the tool's purpose and functionality.

Parameters: A list of arguments that the tool accepts. Each argument has the following properties:

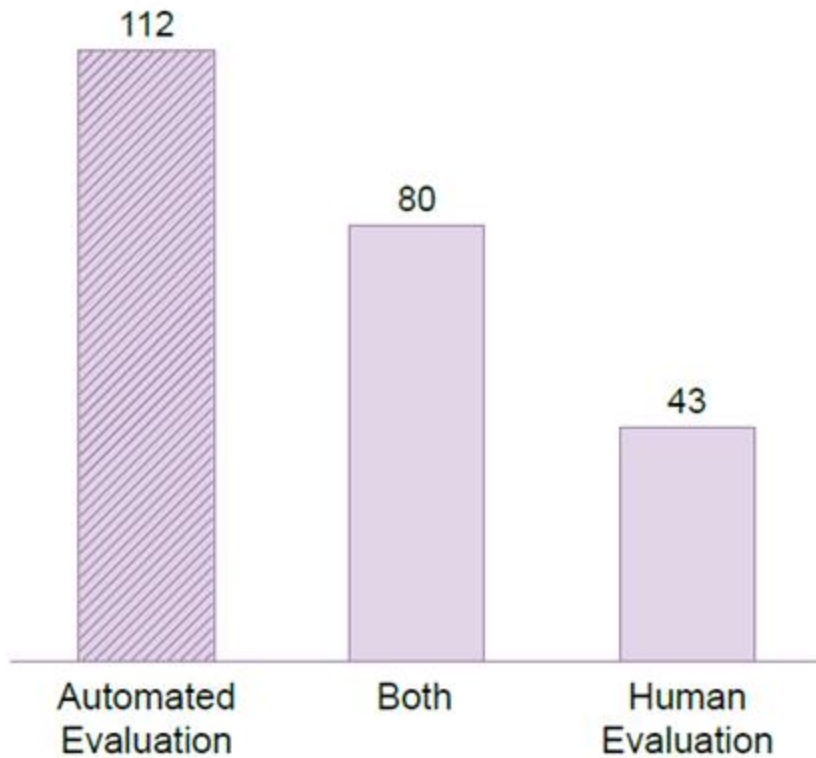
- **Argument Name:** The name of the parameter.
- **Argument Type:** The expected data type for the parameter's value (e.g., string, integer, boolean).
- **Argument Description:** A detailed explanation of what the parameter represents and its purpose.
- **Required:** A boolean value indicating whether the parameter is mandatory for the tool to execute.

Return Schema: A description of the structure and data types of the output returned by the tool upon successful execution.

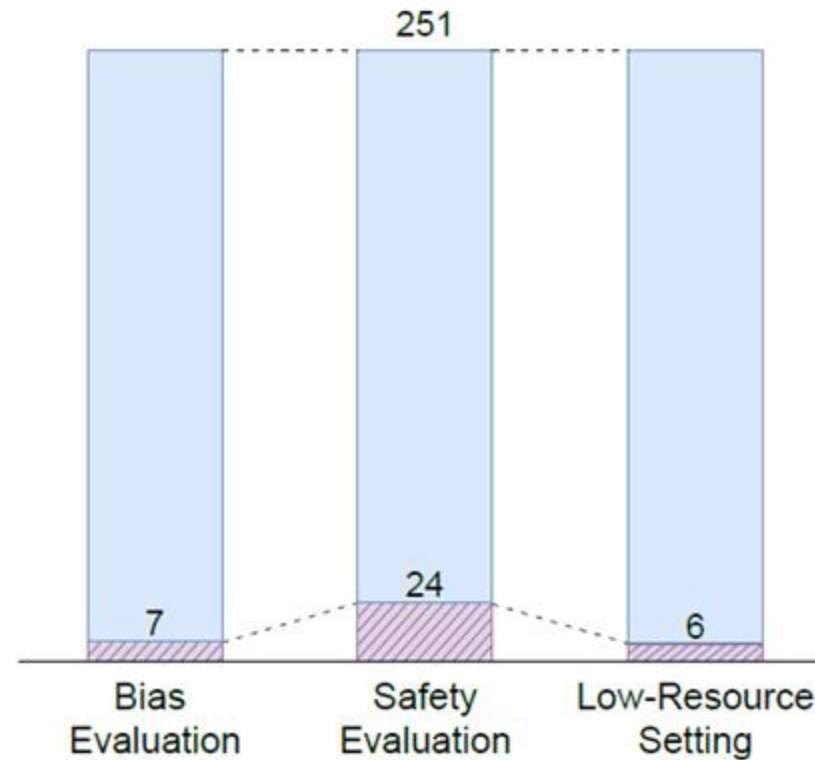


Potential issues of RAG

Automatic and Human Evaluation



Ethical and Contextual Consideration



- Limited manual evaluations
- Most used RAG as 'black-box'
- Only end-task performance is measured

Systematically studying the effectiveness of LLM-RAG in medicine

Rethinking Retrieval-Augmented Generation for Medicine: A Large-Scale, Systematic Expert Evaluation and Practical Insights

Hyunjae Kim¹, Jiwoong Sohn², Aidan Gilson³, Nicholas Cochran-Caggiano⁴, Serina Applebaum¹, Heeju Jin⁵, Seihee Park⁵, Yujin Park⁵, Jiyeong Park⁵, Seoyoung Choi⁵, Brittany Alexandra Herrera Contreras^{1,6}, Thomas Huang¹, Jaehoon Yun⁷, Ethan F. Wei^{1,8}, Roy Jiang¹, Leah Colucci¹, Eric Lai¹, Amisha Dave¹, Tuo Guo¹, Maxwell B. Singer¹, Yonghoe Koo⁹, Ron A. Adelman¹, James Zou¹⁰, Andrew Taylor¹¹, Arman Cohan¹², Hua Xu¹, and Qingyu Chen^{1,*}

¹Yale School of Medicine, Yale University, New Haven, CT, USA

²Department of Biosystems Science and Engineering, ETH Zurich, Zurich, Switzerland

³Massachusetts Eye and Ear, Harvard Medical School, Boston, MA, USA

⁴Geisel School of Medicine at Dartmouth, Hanover, NH, USA

⁵Seoul National University College of Medicine, Seoul, Republic of Korea

⁶San Juan Bautista School of Medicine, Caguas, PR, USA

⁷Hanyang University College of Medicine, Seoul, Republic of Korea

⁸PA Leadership Charter School, West Chester, PA, USA

⁹Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea

¹⁰Stanford University School of Medicine, Stanford, CA, USA

¹¹University of Virginia School of Medicine, Charlottesville, VA, USA

¹²Yale School of Engineering & Applied Science, Yale University, New Haven, CT, USA

*qingyu.chen@yale.edu

- Does RAG actually improve factuality and evidence use as expected?

Kim, H., Sohn, J., Gilson, A., Cochran-Caggiano, N., Applebaum, S., Jin, H., Park, S., Park, Y., Park, J., Choi, S. Contreras, B.A.H., ..., Chen, Q., 2025. Rethinking Retrieval-Augmented Generation for Medicine: A Large-Scale, Systematic Expert Evaluation and Practical Insights. *arXiv preprint arXiv:2511.06738*.



2025 ANNUAL SYMPOSIUM

#AMIA2025

Systematically studying the effectiveness of LLM-RAG in medicine

- 18 healthcare professionals systematically made 80,502 annotations on commonly adopted LLM-RAG implementations in medicine
- 800 LLM outputs head-to-head comparisons with and without RAG
 - GPT-4o and Llama-3.1-8B
 - 200 medical questions: 100 real-world patient questions and 100 USMLE-style queries

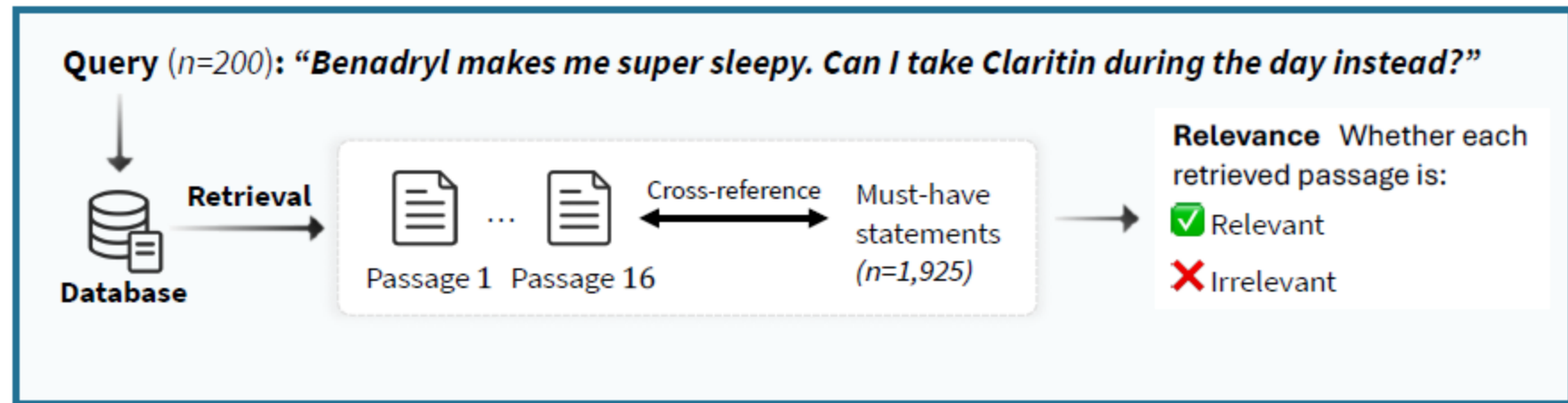


Stage-wise, systematic evaluations

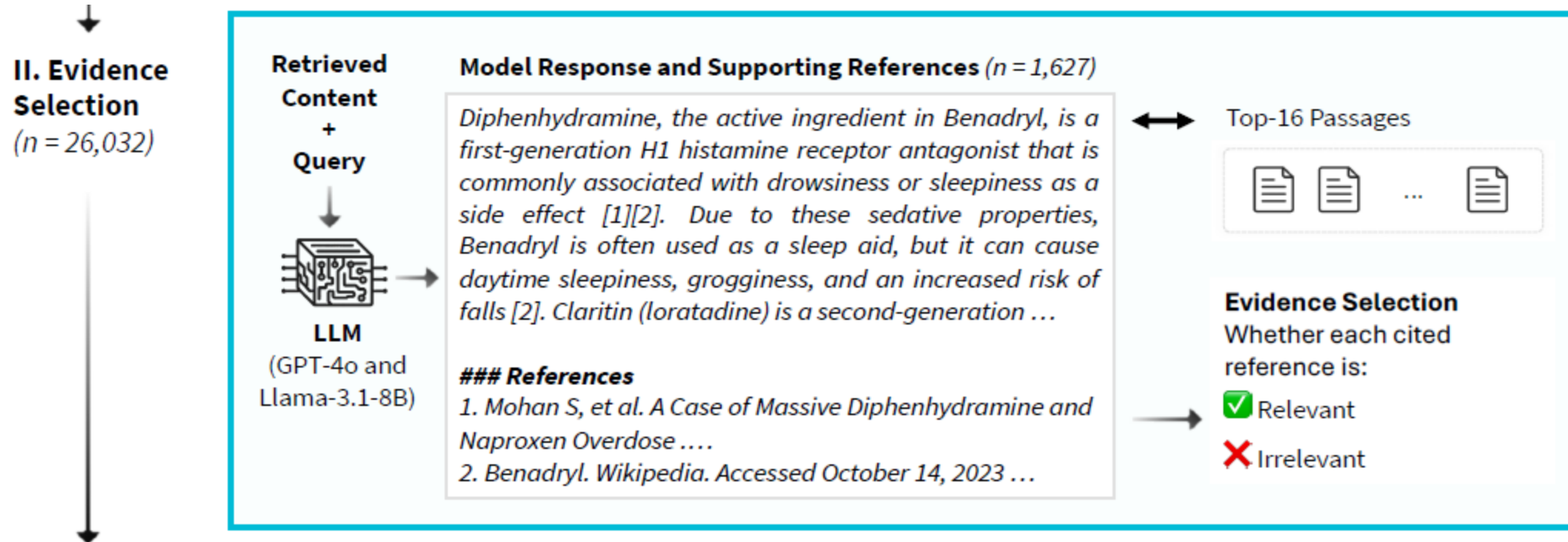
- Evidence retrieval (Retrieval stage): the relevance and coverage of the retrieved passages
- Evidence selection (Augmentation stage): whether LLMs effectively incorporated relevant passages into their responses
- Response generation (Generation stage): the factuality and completeness of final LLM outputs

Evidence retrieval (Retrieval stage)

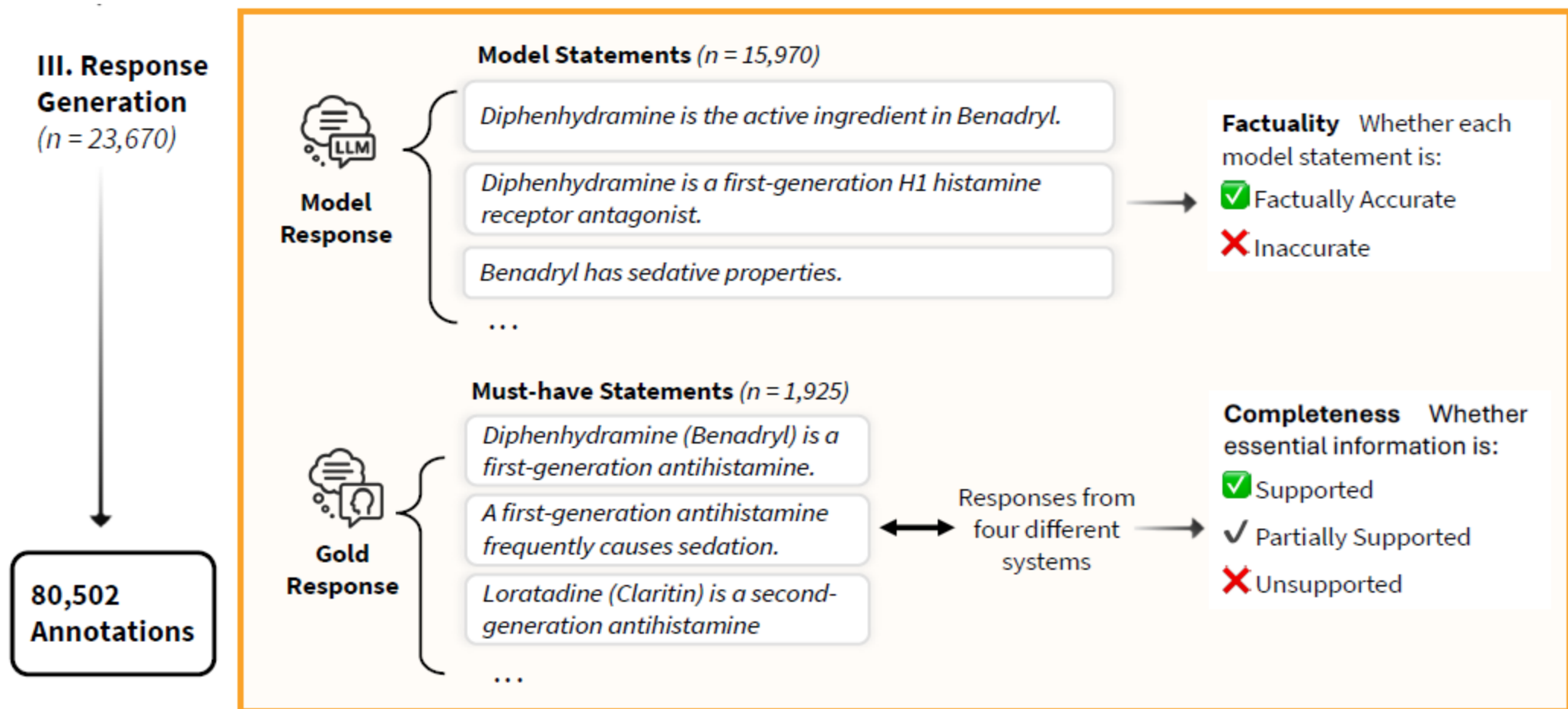
I. Evidence Retrieval ($n = 30,800$)



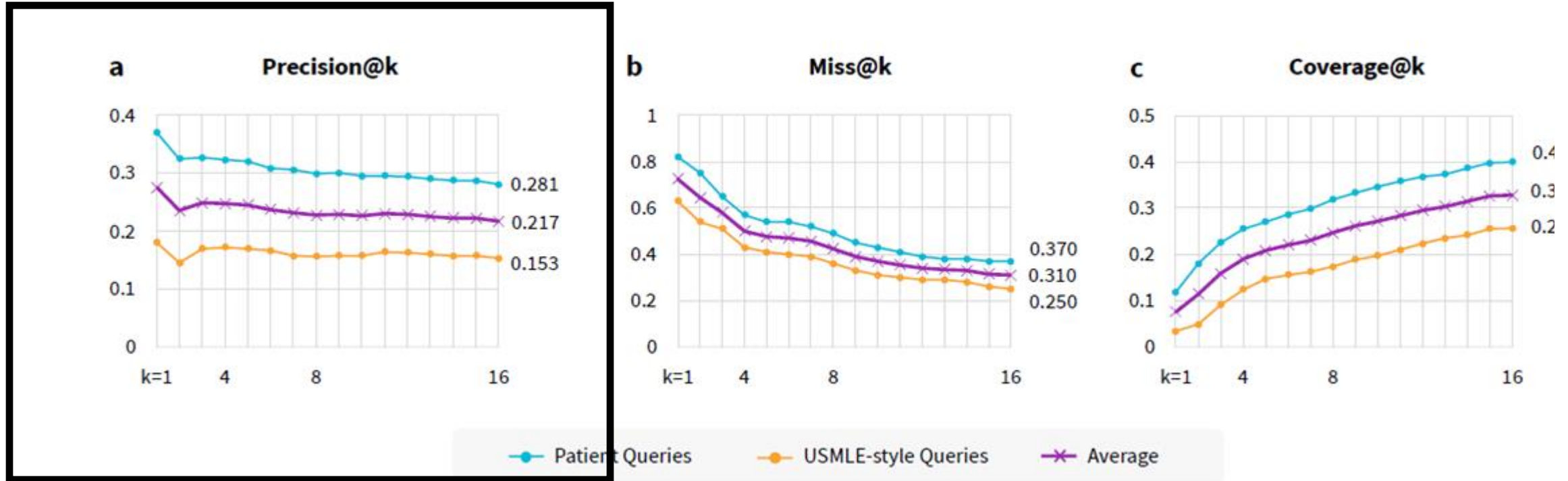
Evidence selection (Augmentation stage)



Response generation (Generation stage)

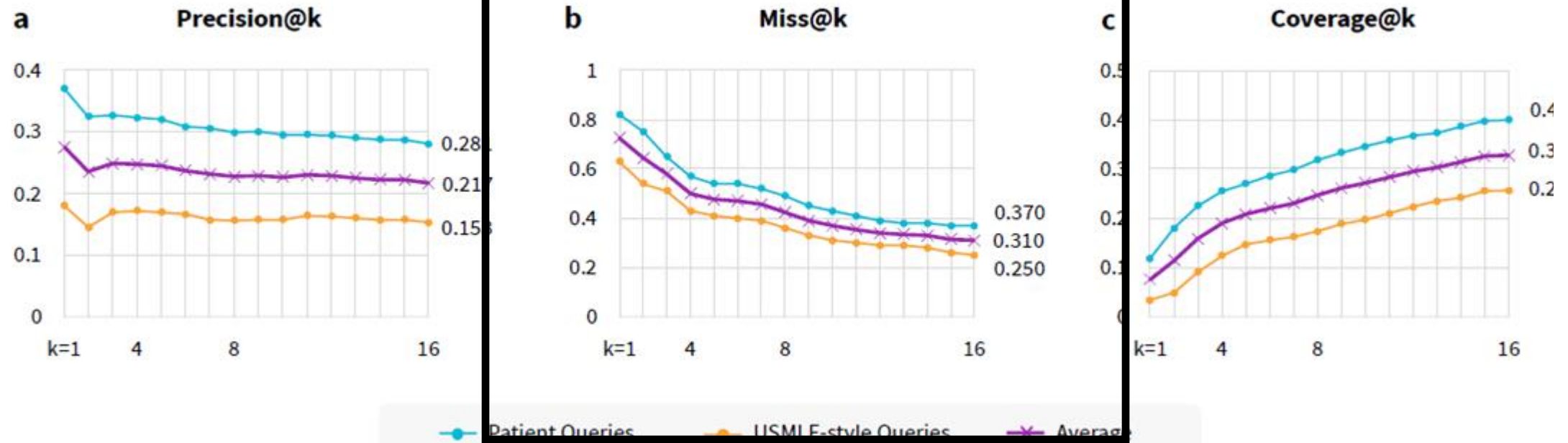


Evidence retrieval (Retrieval stage)



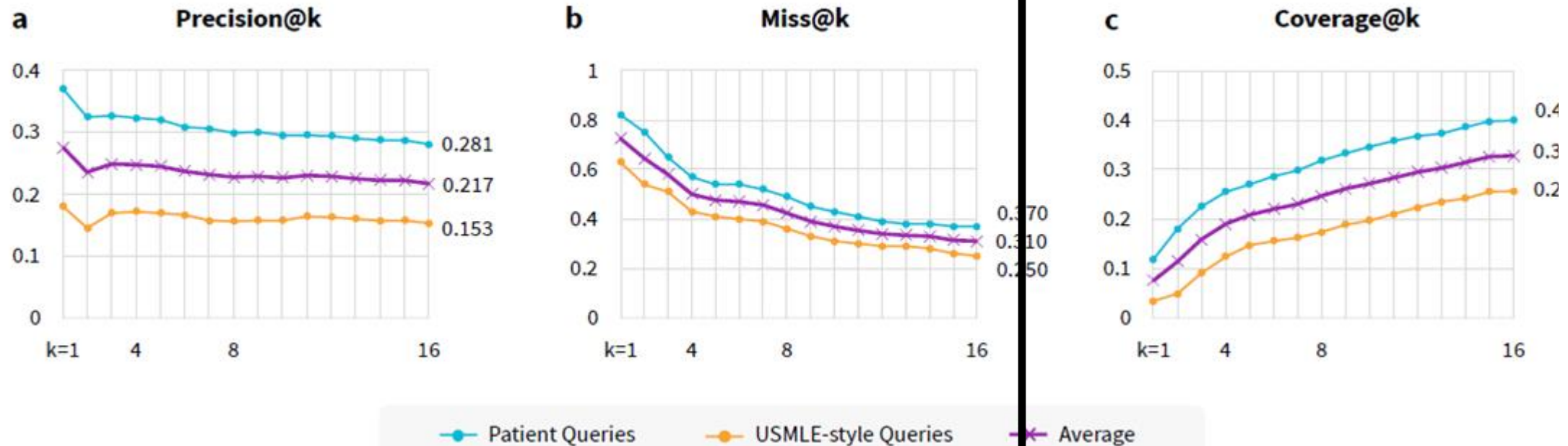
- On average, only 21.7% of the top-16 retrieved passages are relevant

Evidence retrieval (Retrieval stage)



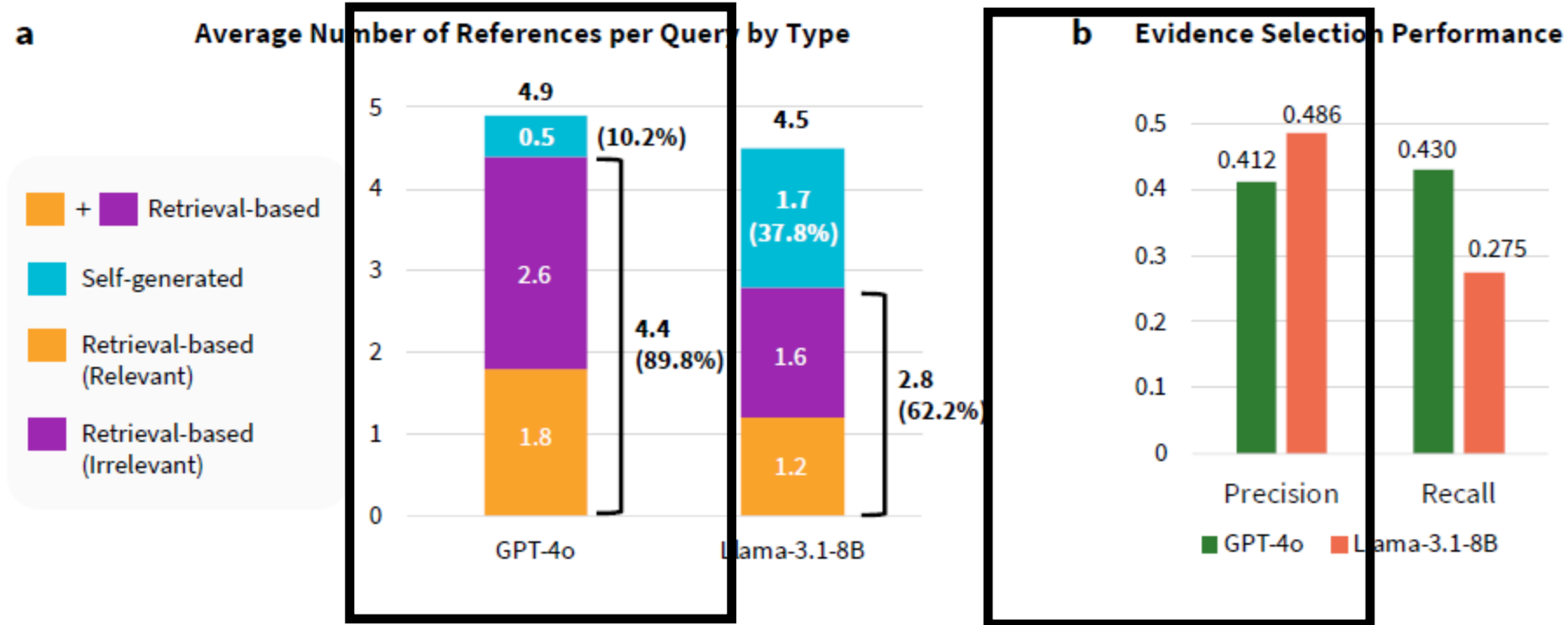
- 31% of queries had NO relevant passages among the top-16 retrieved passages

Evidence retrieval (Retrieval stage)



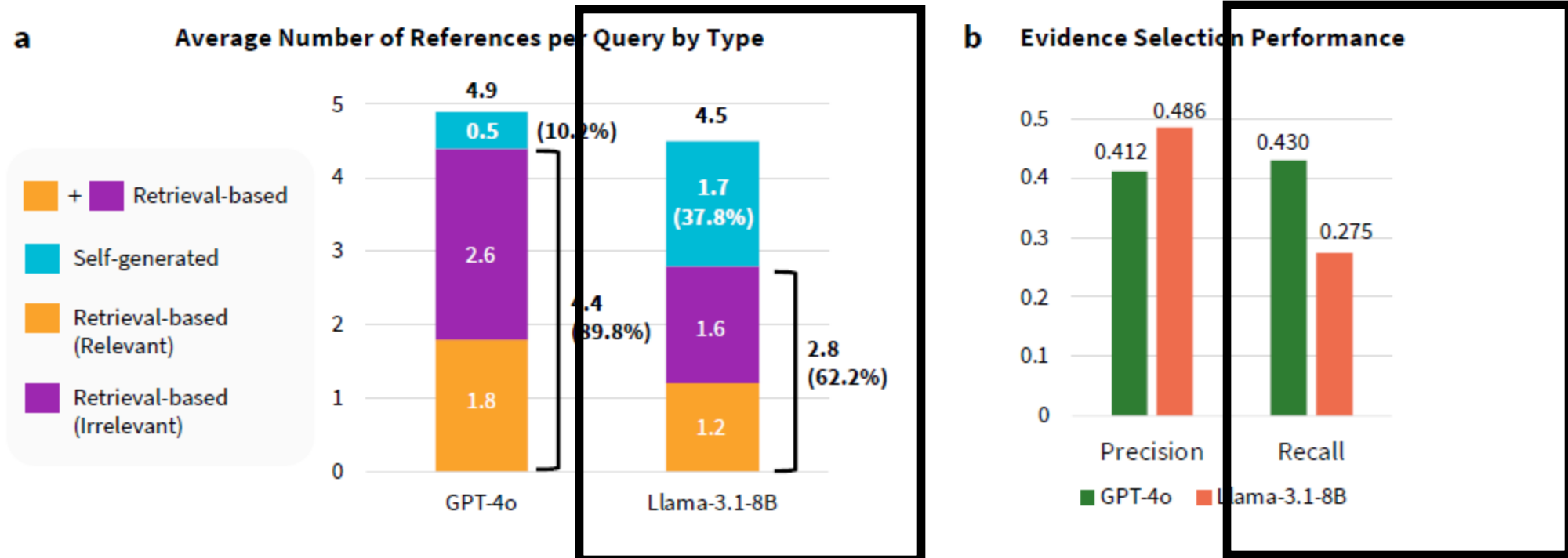
- Only ~33% of must-have statements (manually verified clinically important and should be included the gold standard answer) were supported by the top-16 retrieved passages

Evidence selection (Augmentation stage)



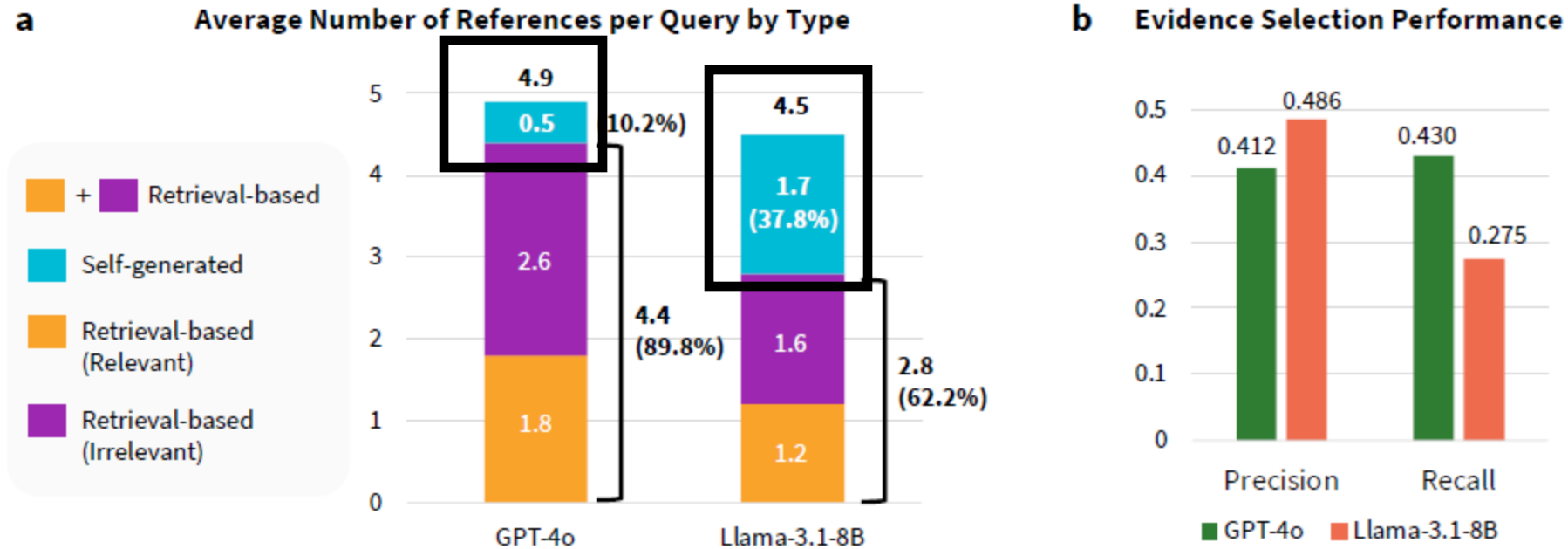
- GPT-4o frequently selected retrieved passages and incorporated as references in the final response (~90%)
- The selected retrieved passages are not relevant: precision of 0.412 and a recall of 0.486

Evidence selection (Augmentation stage)



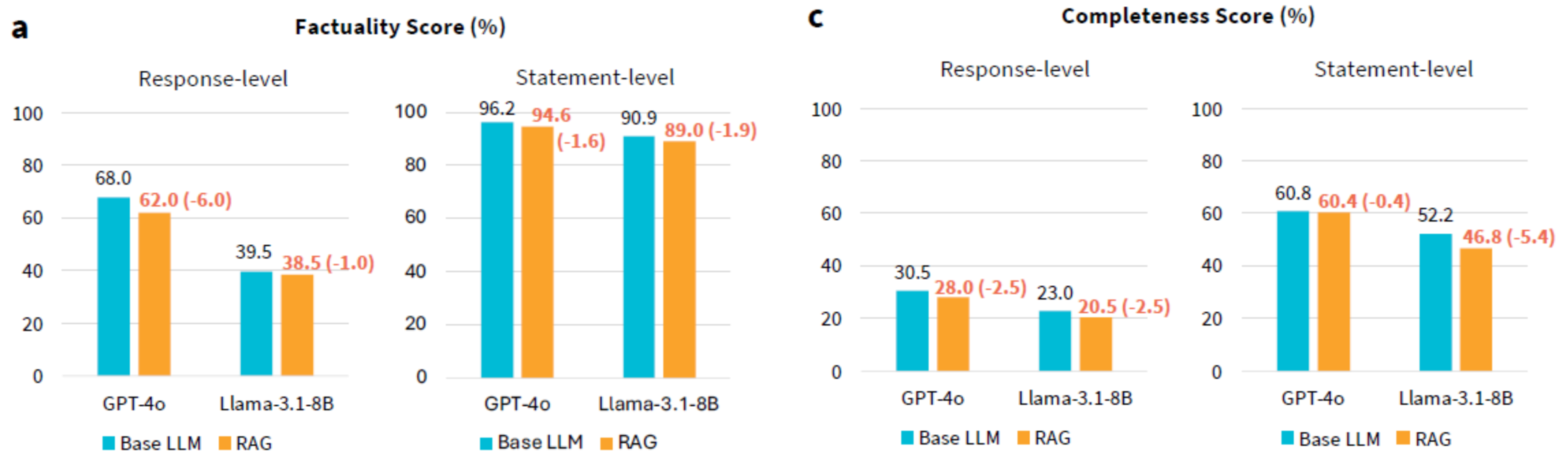
- Llama-3.1 8B selected fewer retrieved passages and incorporated as references in the final response (~60%)
- The selected retrieved passages are not relevant with a significant lower recall of 0.275

Evidence selection (Augmentation stage)



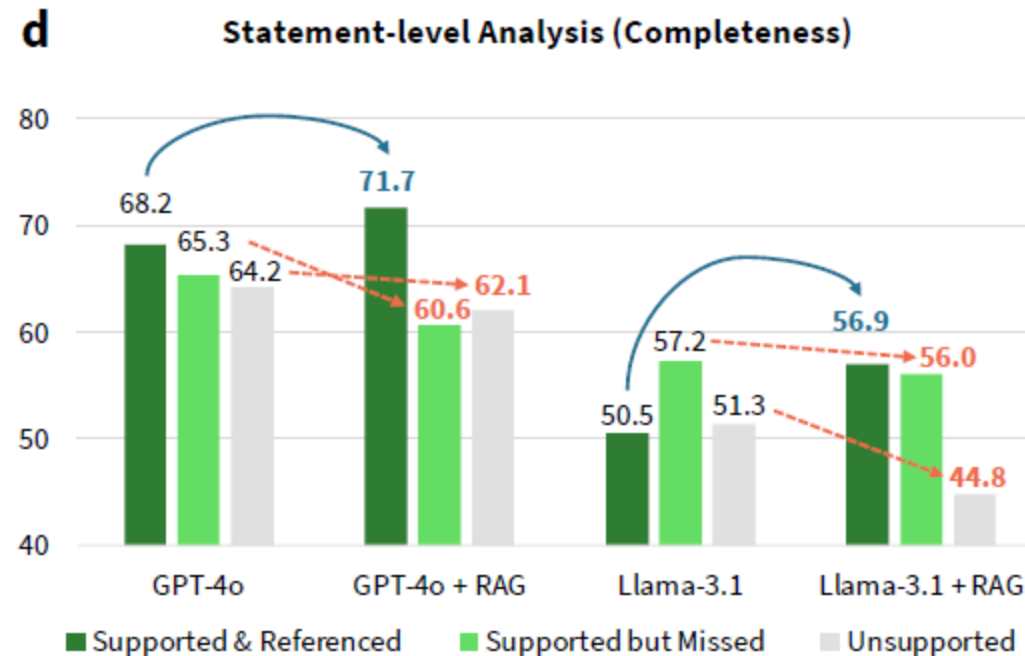
- Llama-3.1 8B had much more self-generated references (37.8% vs. 10.2%)
- ~77% of the self-generated references of Llama-3.1 8B did not exist
- ~13% of the self-generated references of GPT-4o did not exist

Response generation (Generation stage)



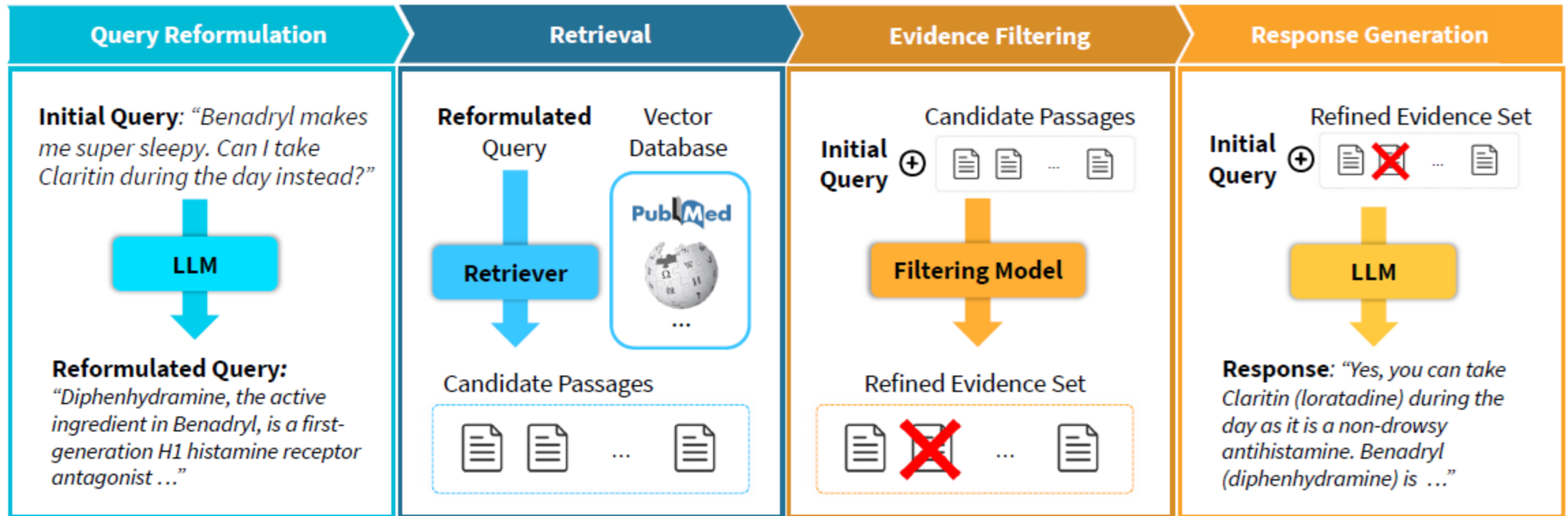
- Both factual accuracy and completeness dropped with RAG

Response generation (Generation stage)



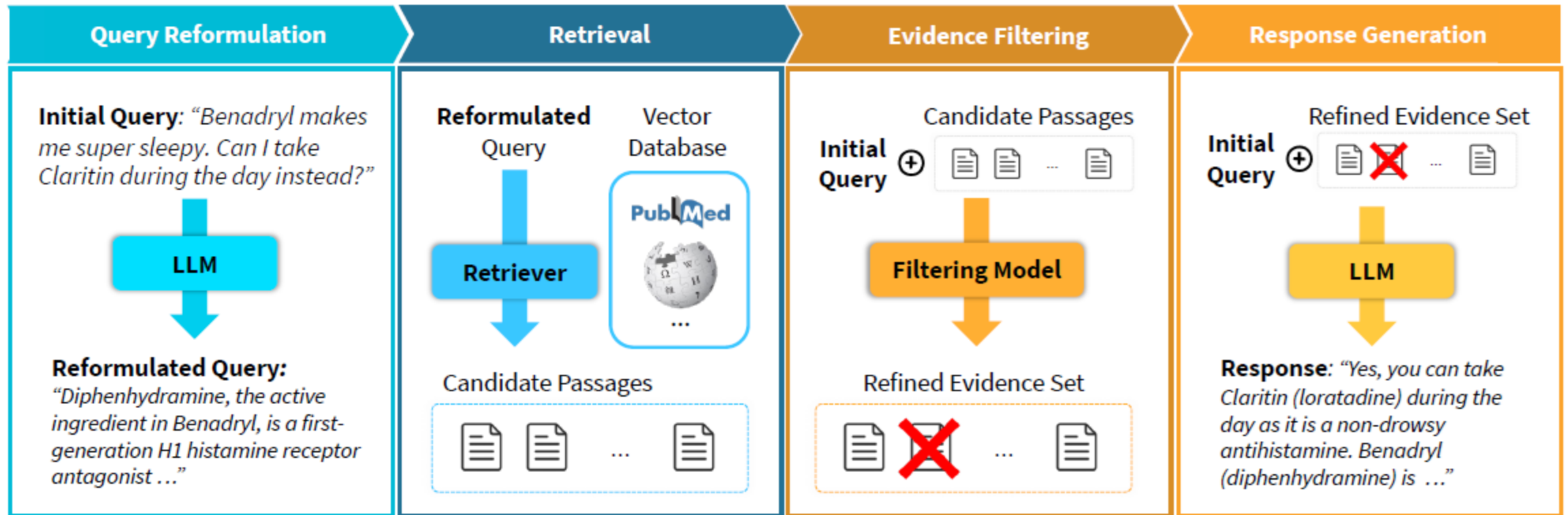
- When the retrieved passages are relevant AND LLMs selected relevant passages, LLM-RAG is effective

Solutions without retraining



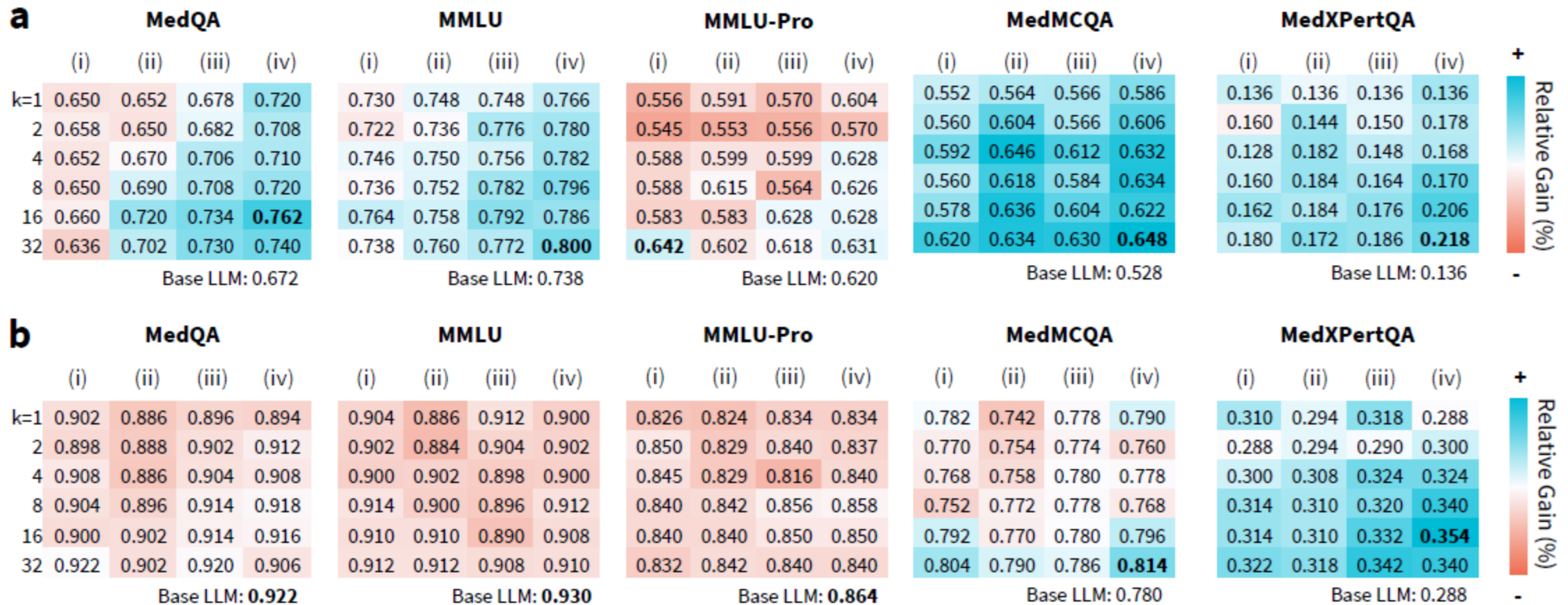
- Query reformulation: LLMs generate initial responses first, instead of retrieval
- The initial responses are then used for retrieval

Solutions without retraining



- Evidence filtering: removing irrelevant passages before feeding to LLMs

Consistent improvements over five independent datasets



More relevant studies

- Xiong, G., Jin, Q., Lu, Z. and Zhang, A., 2024, August. Benchmarking retrieval-augmented generation for medicine. In Findings of the Association for Computational Linguistics ACL 2024 (pp. 6233-6251).
- Jin, Q., Wang, Z., Yang, Y., Zhu, Q., Wright, D., Huang, T., Khandekar, N., Wan, N., Ai, X., Wilbur, W.J. and He, Z., 2025. Agentmd: Empowering language agents for risk prediction with large-scale clinical tool learning. *Nature Communications*, 16(1), p.9377.
- Bi, M., Bao, Z., Xie, D., Xie, X., Yang, C., Wang, T., Wang, Y. and Peng, J., 2025. BioRAGent: natural language biomedical querying with retrieval-augmented multiagent systems. *Briefings in Bioinformatics*, 26(5), p.bbaf539.
- Ng, K.K.Y., Matsuba, I. and Zhang, P.C., 2025. RAG in health care: a novel framework for improving communication and decision-making by addressing LLM limitations. *Nejm Ai*, 2(1), p.Alra2400380.



Questions?



2025 ANNUAL SYMPOSIUM

#AMIA2025

2025 ANNUAL SYMPOSIUM

WHERE COLLABORATION MEETS HEALTHCARE TRANSFORMATION

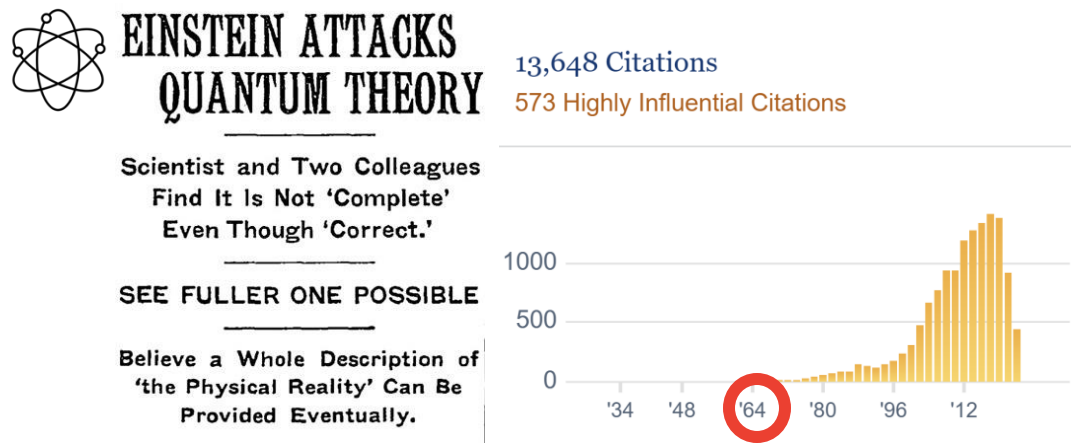
#AMIA2025



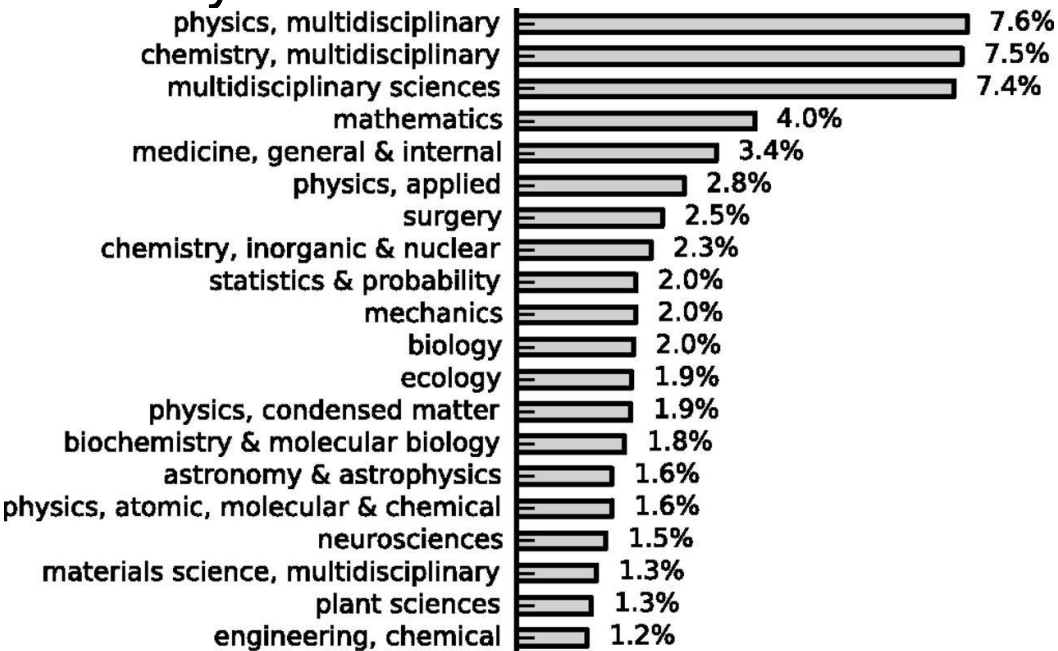
Hypothesis Generation and Experimentation

Why do we want AI-Assisted Hypothesis Generation?

- “Sleeping beauties” in science: scientific discoveries that stay unnoticed for years before gaining sudden attention
- A study of 25 million papers (100 years) shows they occur across all fields.



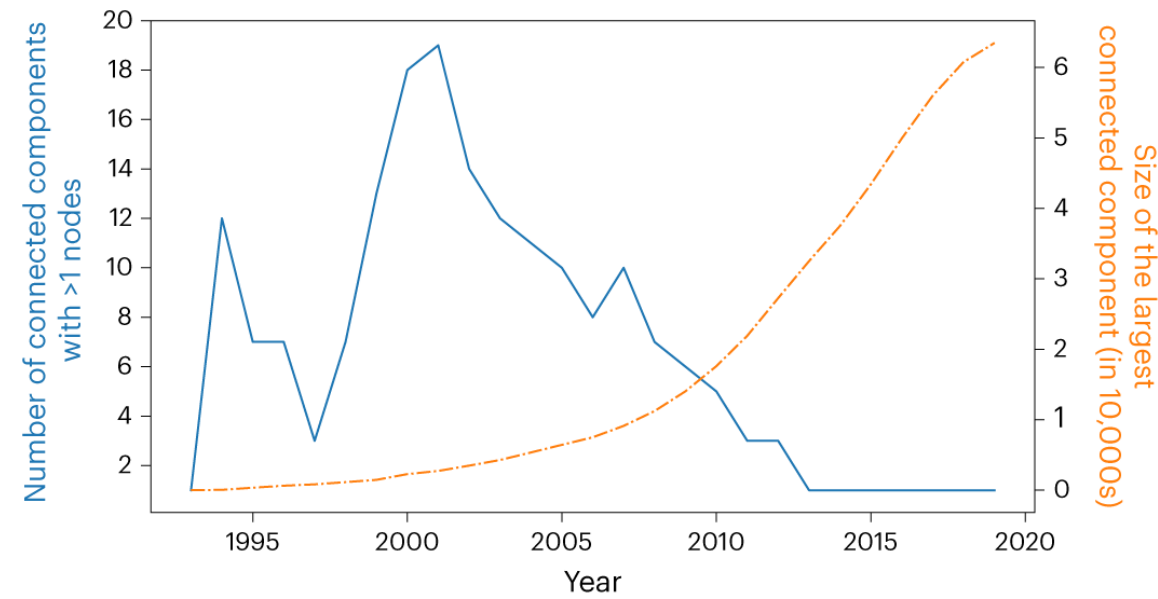
Physics



Top 20 disciplines producing Sleeping Beauties in science

Why do we want AI-Assisted Hypothesis Generation?

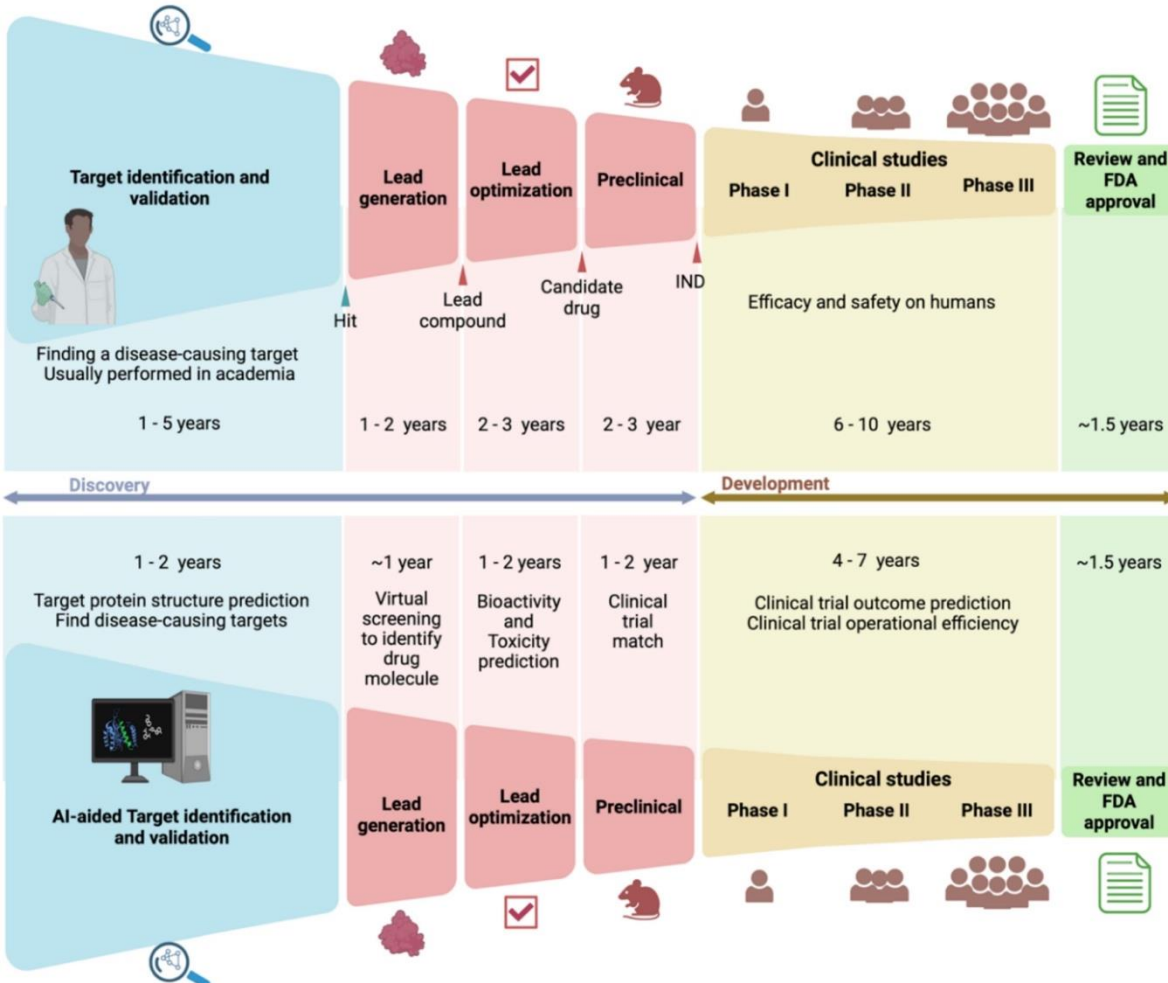
- Most papers build on existing knowledge to formulate new innovations
 - Foster et al. (2015) shows that more than 60% of 6.4 million papers in biomedicine and chemistry published between 1934 and 2008 report findings that build on existing knowledge and provide additional innovations and improvements



Why do we want AI-Assisted Hypothesis Generation?

- AI-based tools can reduce the timeline of drug discovery and development, which typically takes 10–15 years, to approximately 7–9 years

Phases of Traditional vs. AI-aided Drug Discovery



Digesting Scientific Literature is a problem for hypothesis generation!

- For example, consider manual drug repurposing:
 - Current clinical trials for drug repurposing mainly rely on symptoms
 - There's too many drug candidates
 - There's too much information about effects
 - It's too costly to test all drugs and difficult to quantify success

Summary of *in vitro* and *in vivo* studies of the use of CQ/HCQ in COVID-19.

Study	Study type	Objective	Outcome	Result/s	Conclusion
Wang et al. ³⁵	<i>In vitro</i>	To evaluate antiviral efficiency of ribavirin, penciclovir, nitazoxanide, nafamostat, chloroquine, remdesivir and favipiravir against <i>in vitro</i> SARS-CoV 2	Drug efficacy was quantified via viral numbers in cell supernatant (qRT-PCR) and viral nucleoprotein expression (immunofluorescence microscopy)	CQ blocked the virus at low concentrations during both entry and post-entry phases of cellular infection	CQ has potential for clinical use against the SARS-CoV 2 due to potent blocking viral infection demonstrating its antiviral effects, however, <i>in vivo</i> studies are warranted
Yao et al. ²⁶	<i>In vitro</i>	To test CQ/HCQ <i>in vitro</i> activity against SARS-CoV 2 infected Vero cells	Drug efficacy was quantified using the detection of viral RNA via RT-PCR	A twice-daily HCQ (loading dose of 400 mg and maintenance dose of 200 mg) for four days achieved three times the potency of the standard 500 mg CQ given in advance for five days	HCQ has higher potency than CQ in inhibiting SARS-CoV 2 <i>in vitro</i>
Chen et al. ³²	Pilot	To evaluate HCQ against COVID-19	Negative SARS-CoV 2 nucleic acid conversion rate via respiratory pharyngeal swab	More patients in the control group achieved a higher negative pharyngeal swab of viral nucleic acid at 7 days (93.3%, $n = 14$ versus 86.7%, $n = 13$, $p > 0.05$).	Although HCQ showed good prognosis in moderate COVID-19 infection, further investigation is needed with larger sample sizes and better endpoints

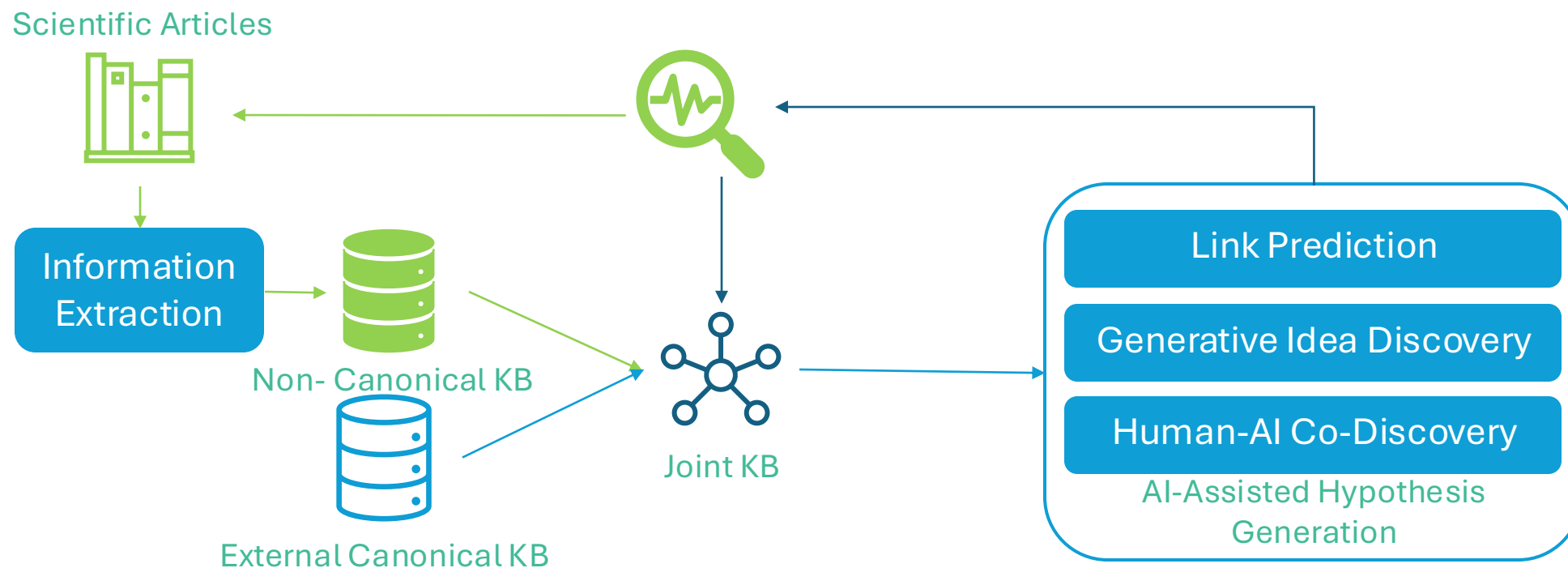
Digesting Scientific Literature is a problem for hypothesis generation!

- What hypotheses do doctors need?
 - A reliable ranked list of drugs with detailed knowledge-level evidence (chemicals/genes)
 - Insights into underlying molecular cellular mechanisms
 - How pre-existing conditions may affect the mortality and severity of a disease

Summary of *in vitro* and *in vivo* studies of the use of CQ/HCQ in COVID-19.

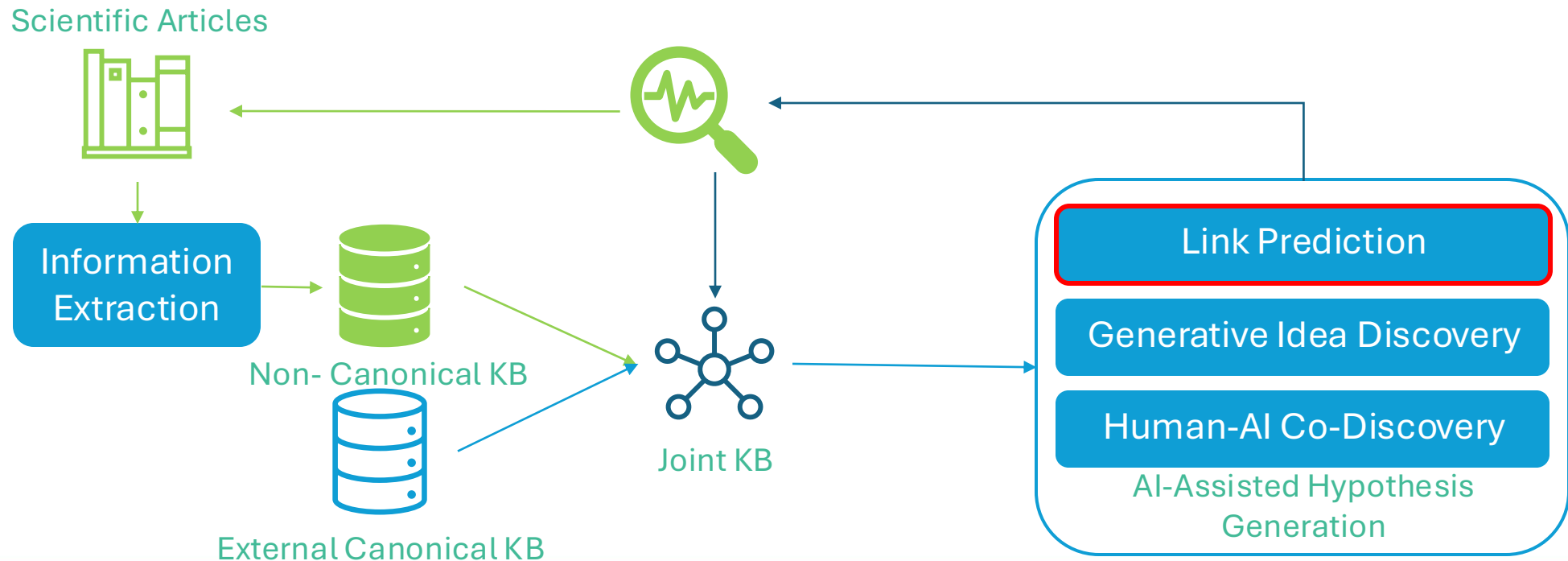
Study	Study type	Objective	Outcome	Result/s	Conclusion
Wang et al. ³⁵	<i>In vitro</i>	To evaluate antiviral efficiency of ribavirin, penciclovir, nitazoxanide, nafamostat, chloroquine, remdesivir and favipiravir against <i>in vitro</i> SARS-CoV 2	Drug efficacy was quantified via viral numbers in cell supernatant (qRT-PCR) and viral nucleoprotein expression (immunofluorescence microscopy)	CQ blocked the virus at low concentrations during both entry and post-entry phases of cellular infection	CQ has potential for clinical use against the SARS-CoV 2 due to potent blocking viral infection demonstrating its antiviral effects, however, <i>in vivo</i> studies are warranted
Yao et al. ²⁶	<i>In vitro</i>	To test CQ/HCQ <i>in vitro</i> activity against SARS-CoV 2 infected Vero cells	Drug efficacy was quantified using the detection of viral RNA via RT-PCR	A twice-daily HCQ (loading dose of 400 mg and maintenance dose of 200 mg) for four days achieved three times the potency of the standard 500 mg CQ given in advance for five days	HCQ has higher potency than CQ in inhibiting SARS-CoV 2 <i>in vitro</i>
Chen et al. ³²	Pilot	To evaluate HCQ against COVID-19	Negative SARS-CoV 2 nucleic acid conversion rate via respiratory pharyngeal swab	More patients in the control group achieved a higher negative pharyngeal swab of viral nucleic acid at 7 days (93.3%, $n = 14$ versus 86.7%, $n = 13$, $p > 0.05$).	Although HCQ showed good prognosis in moderate COVID-19 infection, further investigation is needed with larger sample sizes and better endpoints

Types of AI-Assisted Hypothesis Generation



Can we predict connections between related concepts?

- Link prediction hypothesizes a connection between two concepts, A and C, if they both frequently co-occur with an intermediate concept B across various papers



Examples of Biomedical Link Prediction

Task	Method	Sources	Pros	Cons
Covid-19 Drug Repurposing (Allot et al., 2019)	- Filter a subset of semantic triples based on PubMedBERT and rule-based methods - TransE	Noisy extracted KG	- Focus on a small subset of KGs which are relevant to drug repurposing - Use time slicing for testing	- Fails to consider contextual sentences from the paper - External knowledge bases (e.g., protein-protein interactions, drug-target interactions) are ignored
Sorafenib and Parkinson's disease Drug Repurposing (Xiao et al., 2024)	- Fuse embedding between knowledge graph and textual description - Graph neural networks and biomedical LLMs	External general and domain-specific KG	- Multi-modal information fusion between graph-structure, textual, and domain knowledge - Use hyperbolic embeddings to represent domain ontology (medical hierarchy) knowledge	- Dependence on quality of textual embeddings and domain ontologies - Potentially overfitting (0.998 for Hits@10)

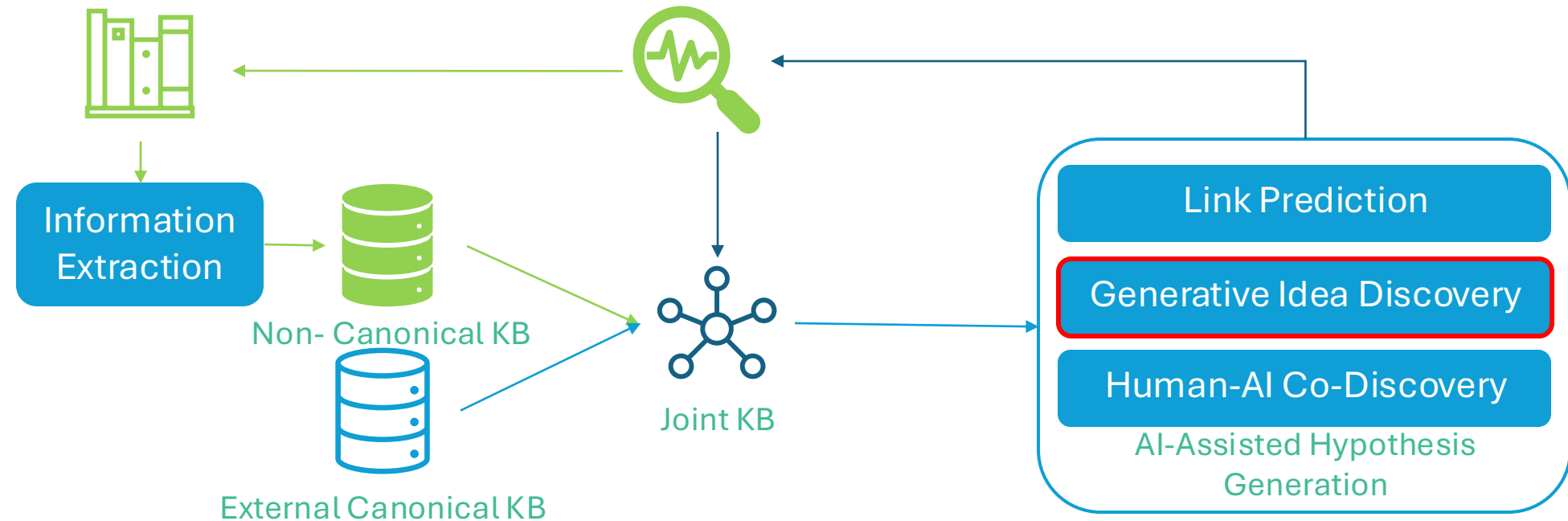
Zhang, R., Hristovski, D., Schutte, D., Kastrin, A., Fisman, M., & Kilicoglu, H. (2021). Drug repurposing for COVID-19 via knowledge graph completion. *Journal of biomedical informatics*, 115, 103696.

Xiao, Y., Zhang, S., Zhou, H., Li, M., Yang, H., & Zhang, R. (2024). FuseLinker: Leveraging LLM's pre-trained text embeddings and domain knowledge to enhance GNN-based link prediction on biomedical knowledge graphs. *Journal of Biomedical Informatics*, 158, 104730.

Can we generate new ideas instead?

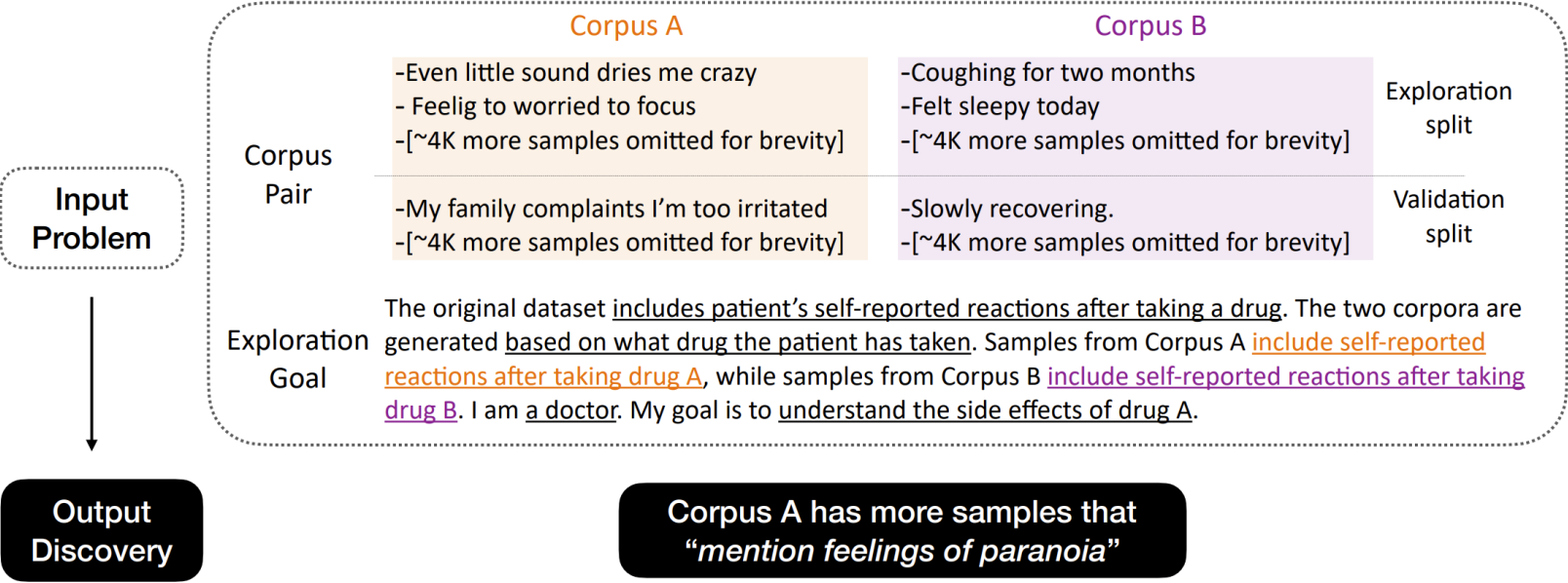
- Simplifying the “language of scientific ideas” to the form of link prediction task limits the expressivity of the hypotheses and does not capture contexts that scientists consider (e.g., target application settings, requirements and constraints, motivations and challenges)

Scientific Articles



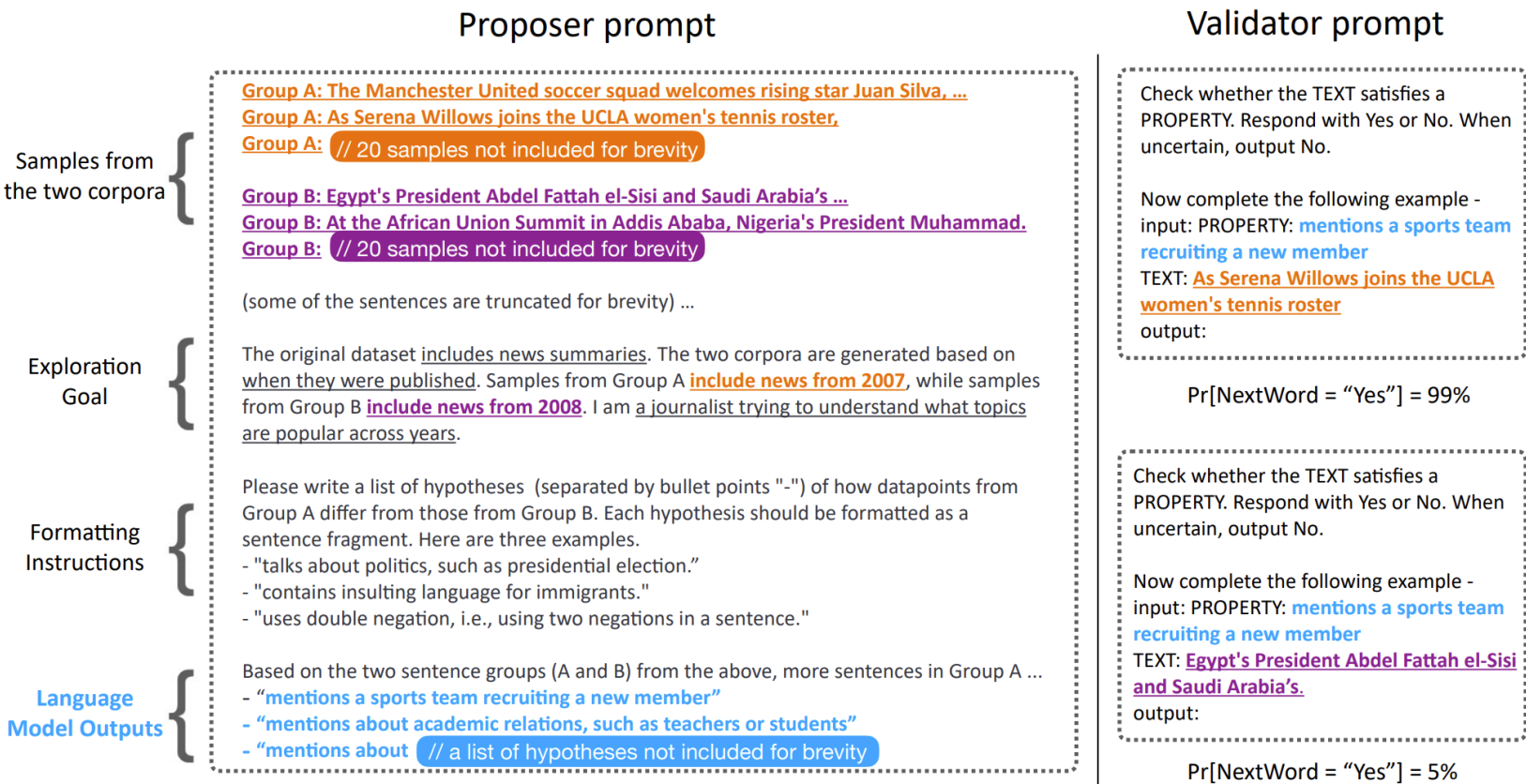
Goal Driven Discovery of Distributional Differences

- Input:
 - A “problem” comprising a description of a user-specified exploration goal (“...understand the side effects of drug A”)
 - Data instances from a corpus pair
- Output
 - A “discovery” represented as a natural language predicate (“mentions feelings of paranoia”)



Goal Driven Discovery of Distributional Differences

- Method
 - Hypothesis Proposer: Prompt GPT3 to generate hypotheses given a few random samples from corpus pairs, the exploration goal, and an instruction to output a list of hypotheses
 - Hypothesis Validator: Finetune Flan-T5 based on human annotations to judge whether a data instance satisfies the generated hypotheses



Goal Driven Discovery of Distributional Differences

- Evaluation Metrics
 - Validity: Require an output hypothesis about the data corpus to be true
 - Relevance: The output discovery is relevant, indirectly relevant, or irrelevant to the goal
- Key Results
 - LLMs can propose more relevant hypothesis based on the user-provided goals
 - LLMs can discover and explain unknown patterns on a wide range of applications
- Points for improvement
 - The paper could be enhanced by generating code to analyze data instances across two corpora, which would provide more insightful results

Domain	Example Datasets	How the Corpus Pairs are Generated	
		Corpus A	Corpus B
87 Business problems			
Commercial Reviews	Airline reviews	1 st -class passenger reviews	Economy passenger reviews
	Product Reviews	Reviews that give 10 stars	Reviews that give 0 star
Finance	YC startups	Successful startup descriptions	Failed startup descriptions
	News Headlines	Top headlines when S&P rises	Top headlines when S&P falls
278 Social Sciences problems			
Politics	Administration policy	Admin policy from Trump	Admin policy from Obama
News	Reuters headlines	Headlines from 2014	Headlines from 2015
Language	Craigslist Negotiations	Dialogue from successes	Dialogue from failures
	Diplomacy Dialogues	Lies	Honest statements
Sociology	Happy moments	Self-reported happy moments from females	Self-reported happy moments from males
	Rate My Professor	Reviews of female lecturers	Reviews of male lecturers
169 Humanities problems			
Arts	Music lyrics	Drake rap lyrics	Kanye rap lyrics
Education	Student essays	Essays that received full score	Essays with only partial credit
10 Health problems			
Health	Doctor’s note	Patients diagnosed with pneumonia	Patients not diagnosed with pneumonia
131 Machine Learning problems			
Machine Learning	NLI — distribution shift	Samples from SNLI	Samples from MNLI
	QQP — spurious correlation	Individual questions with label “paraphrase”	Individual questions with label “non-paraphrase”
	LM’s output	Generations from one LM	Generations from another LM
	inputs — error analysis	Inputs where one model is correct	Inputs where one model is wrong

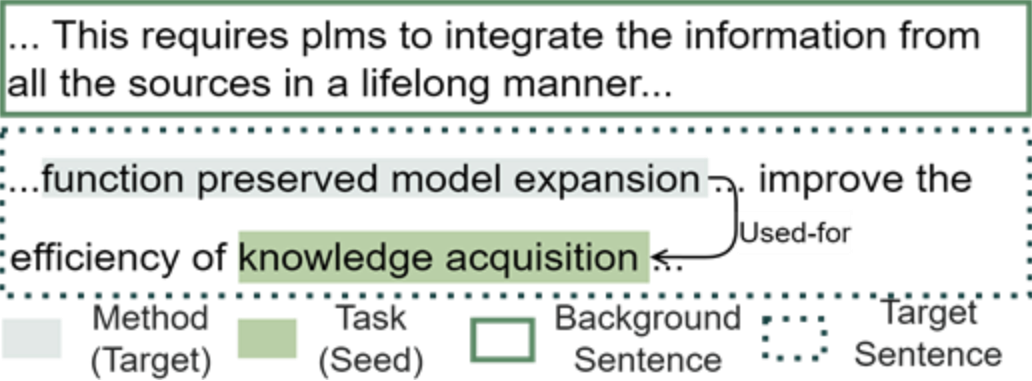
Scientific Inspiration Machines Optimized for Novelty

- Input
 - Current problems, motivations, experimental settings and constraints
 - A seed term that should be a focus point of the generated idea
- Output
 - A generated novel hypothesis as a natural language sentence



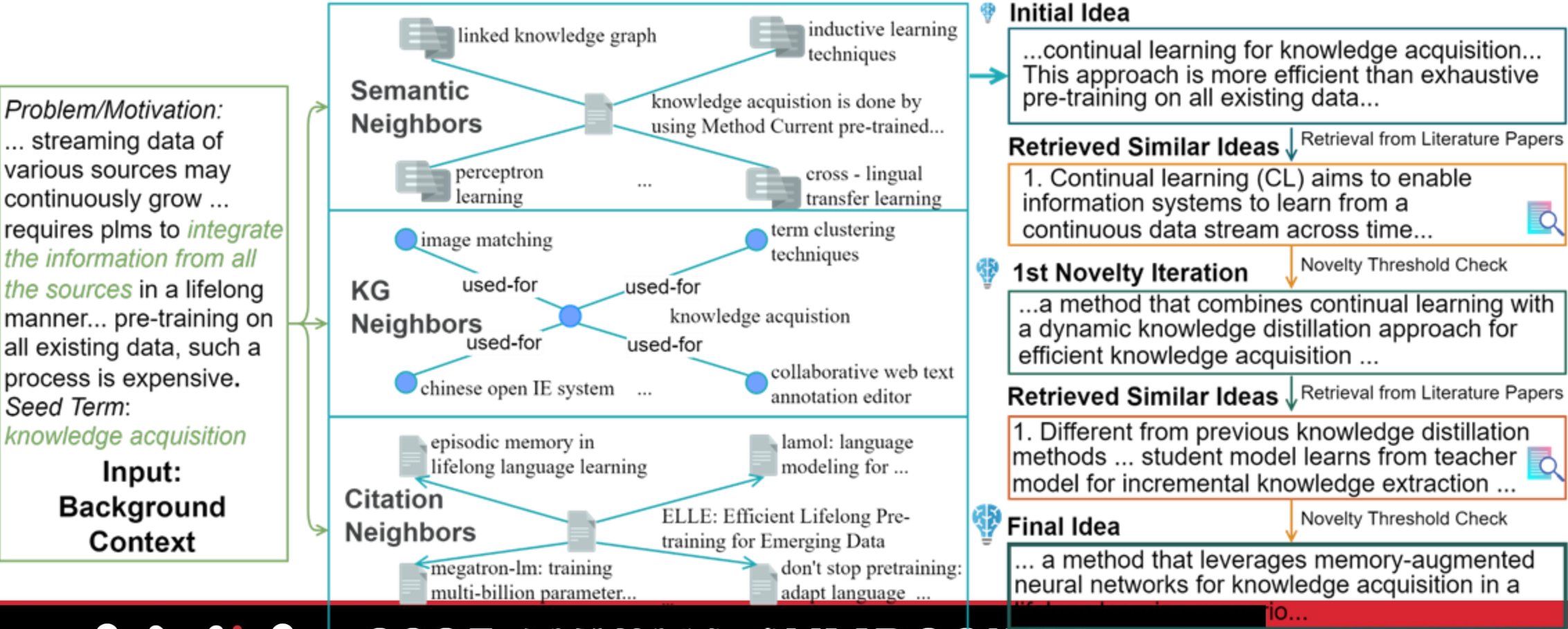
Seed Term: knowledge acquisition
*Background: ... This requires plms to **integrate the information from all the sources** in a lifelong manner. Although this goal could be achieved by exhaustive pre-training on all the existing data, such a process is known to be **computationally expensive**.*

Specifically, ELLE consists of (1) **function preserved model expansion**, which flexibly expands an existing PLM's width and depth to improve the efficiency of **knowledge acquisition** ...



Scientific Inspiration Machines Optimized for Novelty

- Retrieves structured “inspirations” from past scientific papers
- Explicitly optimizes for novelty by iteratively comparing to prior papers and updating idea suggestions until sufficient novelty is achieved



Scientific Inspiration Machines Optimized for Novelty

- Key Results
 - GPT-4 tends to generate ideas with overall low technical depth and novelty,
 - SCIMON partially mitigates this issue by retrieving inspirations from papers and iteratively boosting the novelty of generated ideas

Input	<i>seed term</i> : speech unit boundaries ; <i>context</i> (abridged): ... generate partial sentence translation given a streaming speech input. existing approaches ... break the acoustic units in speech, as boundaries between acoustic units in speech are not even...	 Feasibility
Initial Idea	A <u>pause prediction model</u> to identify speech unit boundaries ...	
Iteration 1	A method that leverages acoustic and linguistic features to predict speech unit boundaries dynamically, ensuring smooth transitions ... differs from the existing research as it combines both acoustic properties and linguistic context ... adapting to variations in speaker characteristics, speaking styles, and languages.	
Iteration 2	A novel method called Adaptive Speech Unit Boundary Detection (ASUBD) ... a combination of attention mechanisms to focus on relevant acoustic and linguistic features and reinforcement learning to guide the system to make optimal predictions of unit boundaries based on previous decisions...	
Ground Truth	... an efficient monotonic segmentation module ... accumulate acoustic information incrementally and detect proper speech unit boundaries.	



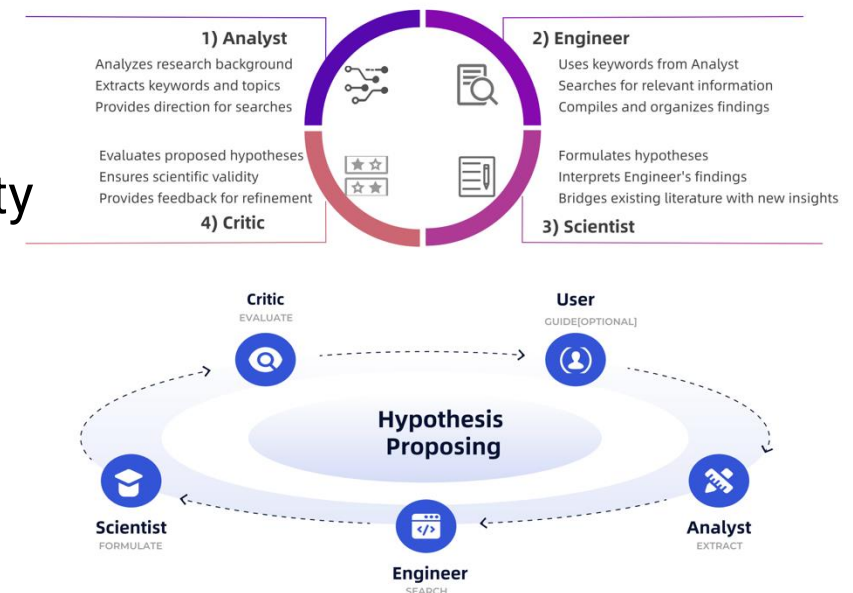
Biomedical Hypothesis Generation

- A multi-agent LLM system utilizing tools to simulate the collaborative nature of scientific research
- Key Results
 - LLMs can propose new hypotheses that did not appear in the training data but can be confirmed by the test literature
 - Few-shot examples enhance verifiability but reduce novelty
 - Introducing uncertainty into processes and operations enhances zero-shot generalization
 - Tool use has minimal impact on improving the hypothesis generation ability of LLMs
 - Multi-agent collaboration can significantly enhance model's ability
- Points for improvement
 - The results are not evaluated from a biomedical perspective
 - The paper only tests general domain tool sets

Instruction: You are a researcher. You can come up with new hypotheses based on your existing knowledge. Hypotheses are given against the following background. You should be as detailed as possible.

Background knowledge: (1) Esophageal cancer mainly includes squamous cell carcinoma and adenocarcinoma, with different risk factors and incidence rates. (2) Metformin has been shown to reduce the risk of several cancers in patients with T2DM. (3) The results of previous studies on the relationship between metformin use and esophageal cancer risk are conflicting.

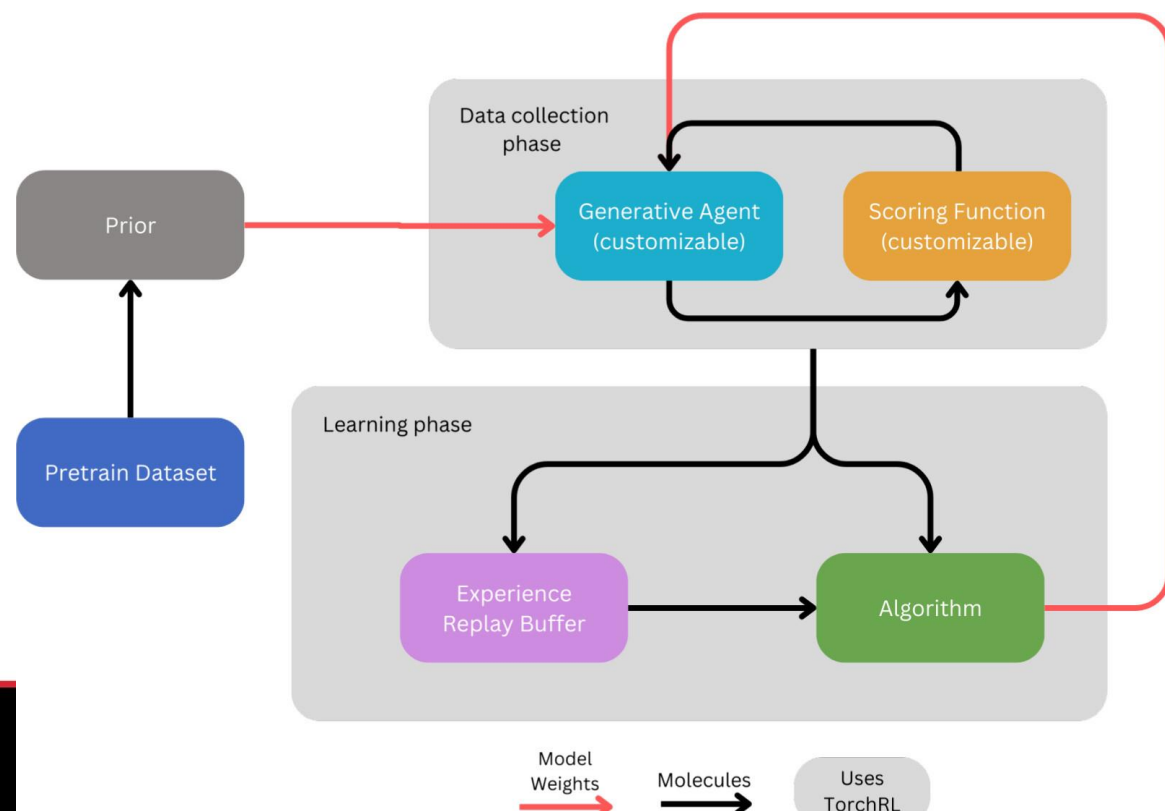
The model hypothesizes: (1) The use of metformin in patients with T2DM may be associated with a reduced risk of esophageal cancer. (2) There may be differences in the correlation between metformin use and the risk of esophageal cancer in male and female patients with T2DM. (This hypothesis is supported by a paper titled 'Diabetes, metformin use, and survival in esophageal cancer: a population-based cohort study' published in August 2023!)



Qi, B., Zhang, K., Li, H., Tian, K., Zeng, S., Chen, Z. R., & Zhou, B. (2023): Large Language Models as Biomedical Hypothesis Generators: A Comprehensive Evaluation.. COLM 2024.

Reinforcement Learning of Generative Chemical Agents for Drug Discovery

- Use existing reinforcement learning library TorchRL as building blocks to assemble efficient and reliable drug discovery agents
- Support multiple RL algorithms (e.g., REINFORCE, A2C, PPO) and various generative tasks (de novo generation, scaffold decoration, fragment linking) for molecule design
- Integrate MolScore for drug design relevant scoring to optimize drug-relevant molecular properties during training



Pros:

- Reduce training time to 50% compared to MolOpt baseline while improving generation accuracy
- Modular implementation and open-source code help readers reproduce results

Points for Improvement:

- Chemical LMs in this paper are limited to 1D SMILE representations

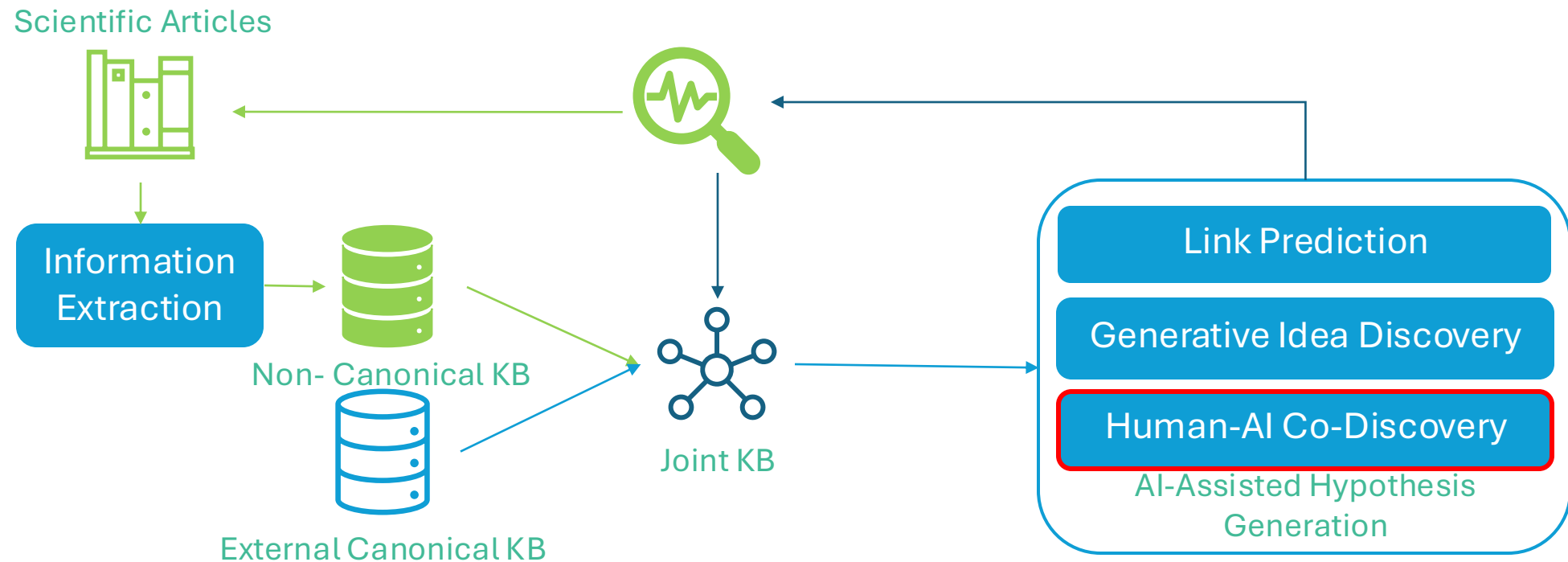
Takeaways

- Notable points
 - Large language models (e.g., GPT4) can generate interesting hypothesis across a variety of domains when provided with sufficient background knowledge
 - Providing LLMs with previously seen ideas can further boost the novelty of generated hypotheses
- Ongoing directions
 - Move beyond knowledge-augmented generation by integrating tool learning to automatically retrieve relevant background knowledge in an end-to-end way



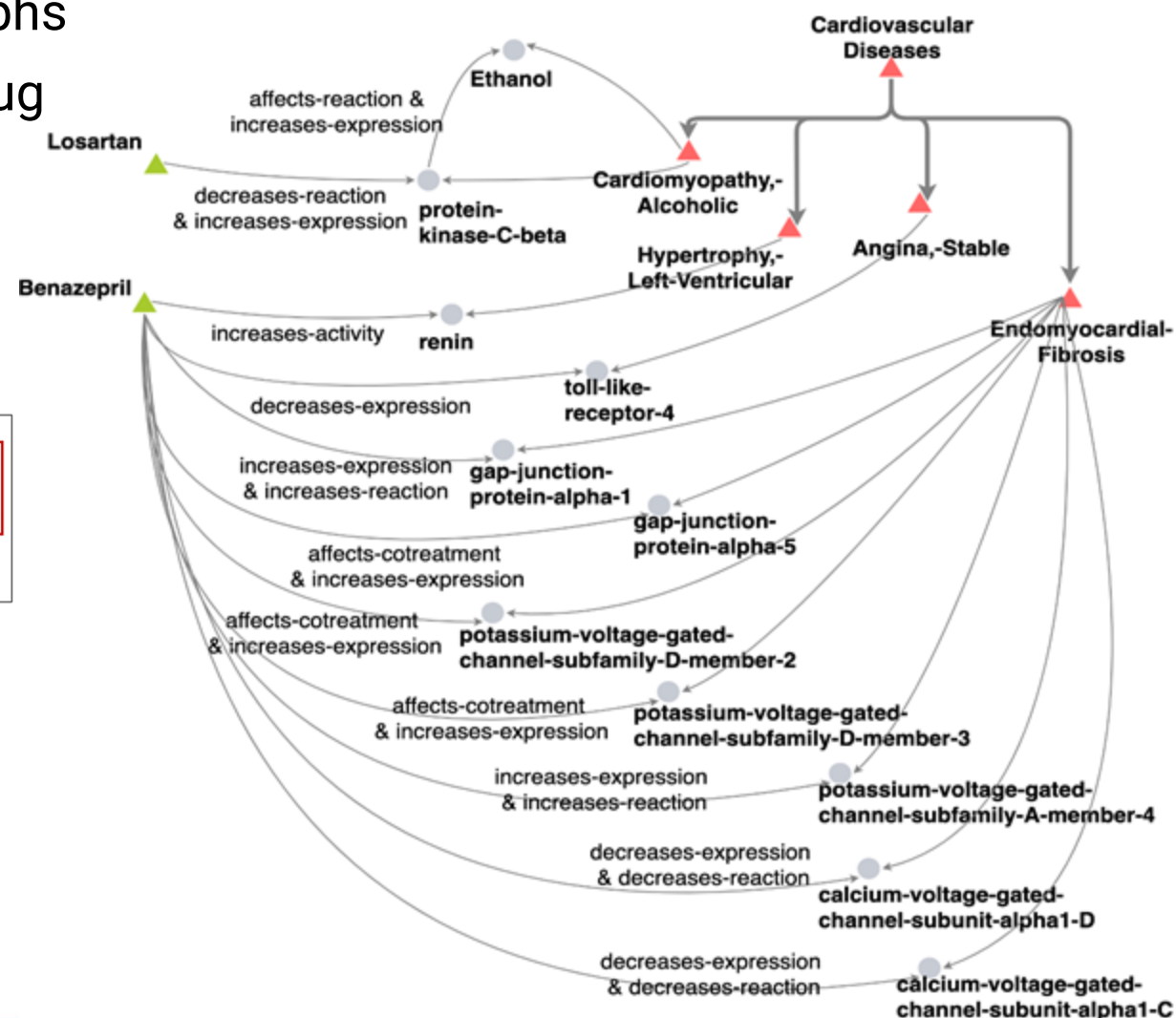
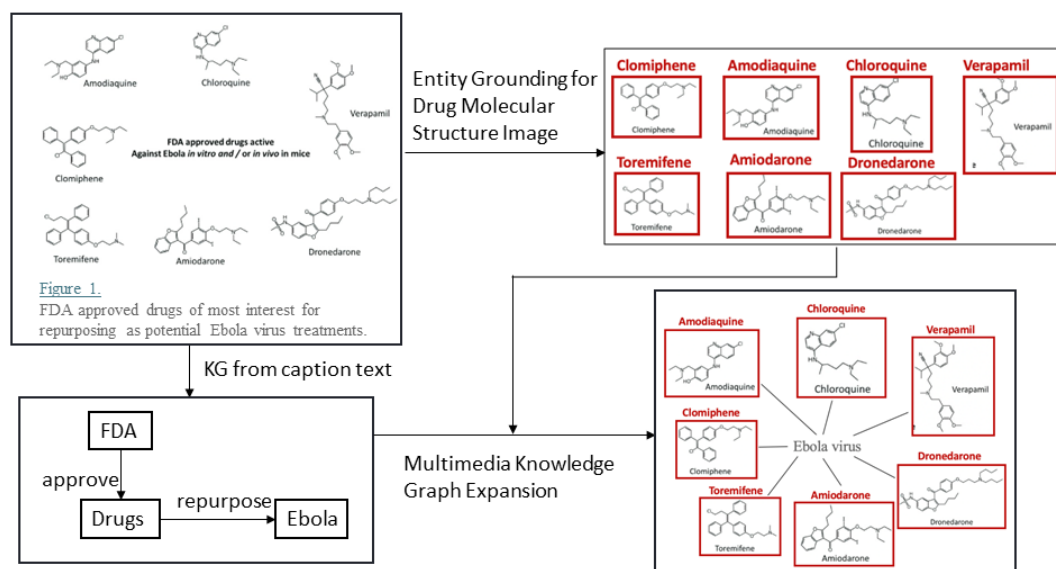
Can AI assist human hypothesis discovery?

- Human-machine collaborative discovery can help overcome the limitations of AI-driven hypothesis generation by integrating expert experience and allowing for iterative refinement



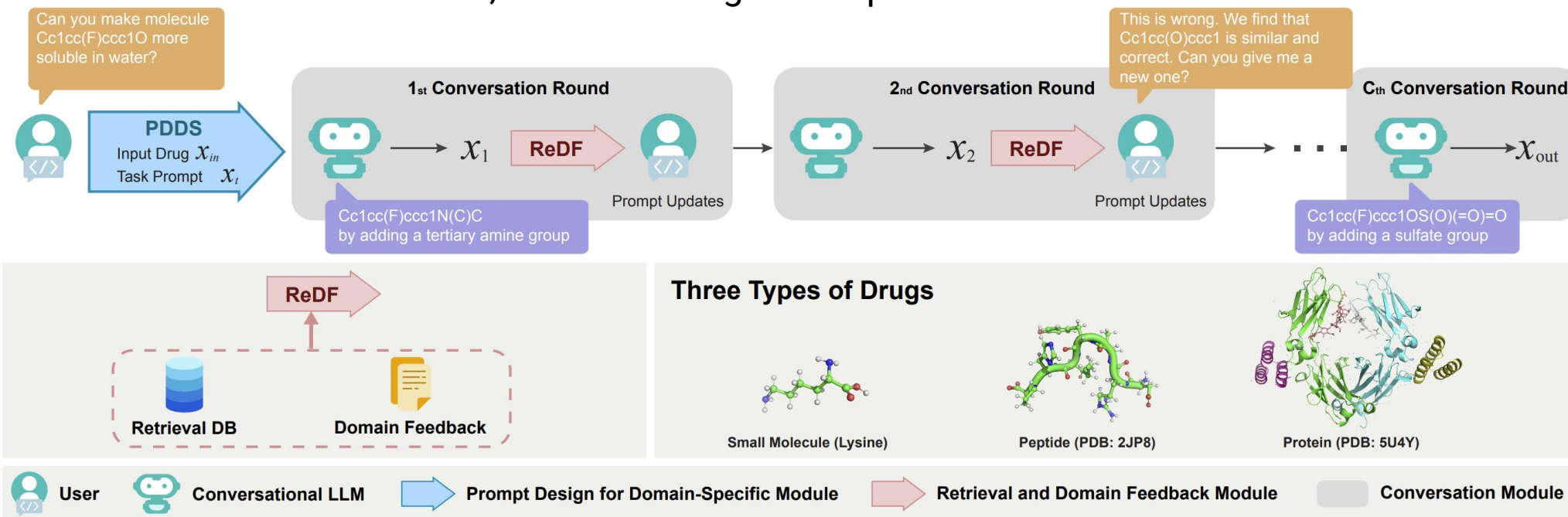
COVID19 Drug Repurposing Report Generation

- Multimedia knowledge extraction from 25,534 COVID-19 papers to construct knowledge graphs
- Provide biologists with typical questions in drug repurposing based on their suggested drugs



Conversational Drug Editing with Retrieval

- Prompt Design for Domain Specific (PDDS): Generate drug editing prompts
- Retrieval and domain feedback (ReDF): Update the prompts using retrieved information and domain feedback
- Key Results
 - By injecting retrieved domain knowledge, LLMs can perform drug editing, identify the key substructures, and offer insightful explanations



Pros:

- Incorporate both domain knowledge and expert feedback in the iteration to improve performance and provide interpretability

Points for Improvement:

- LLMs backbones are not dedicated to molecular optimization

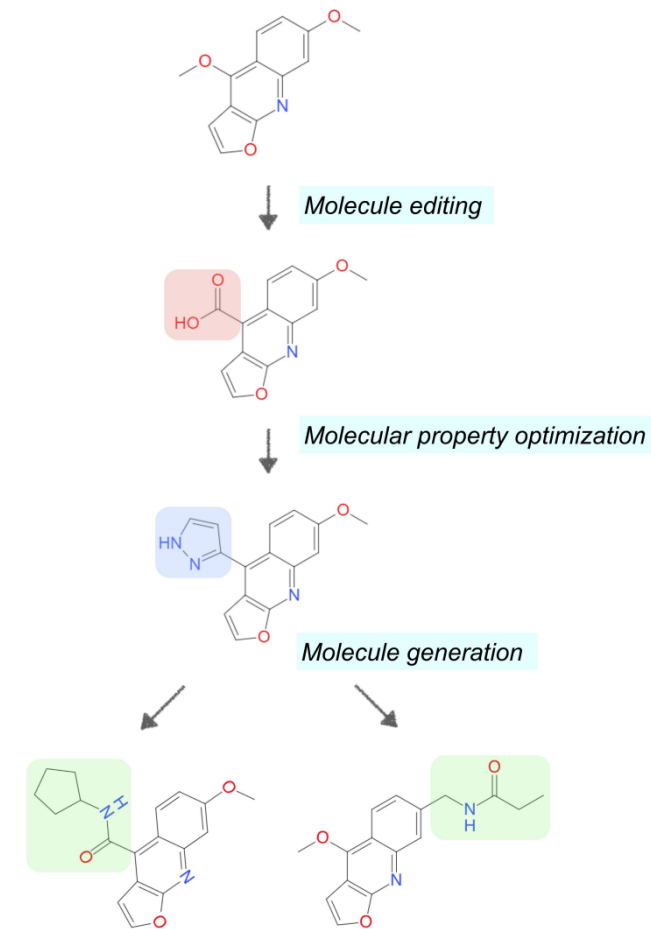
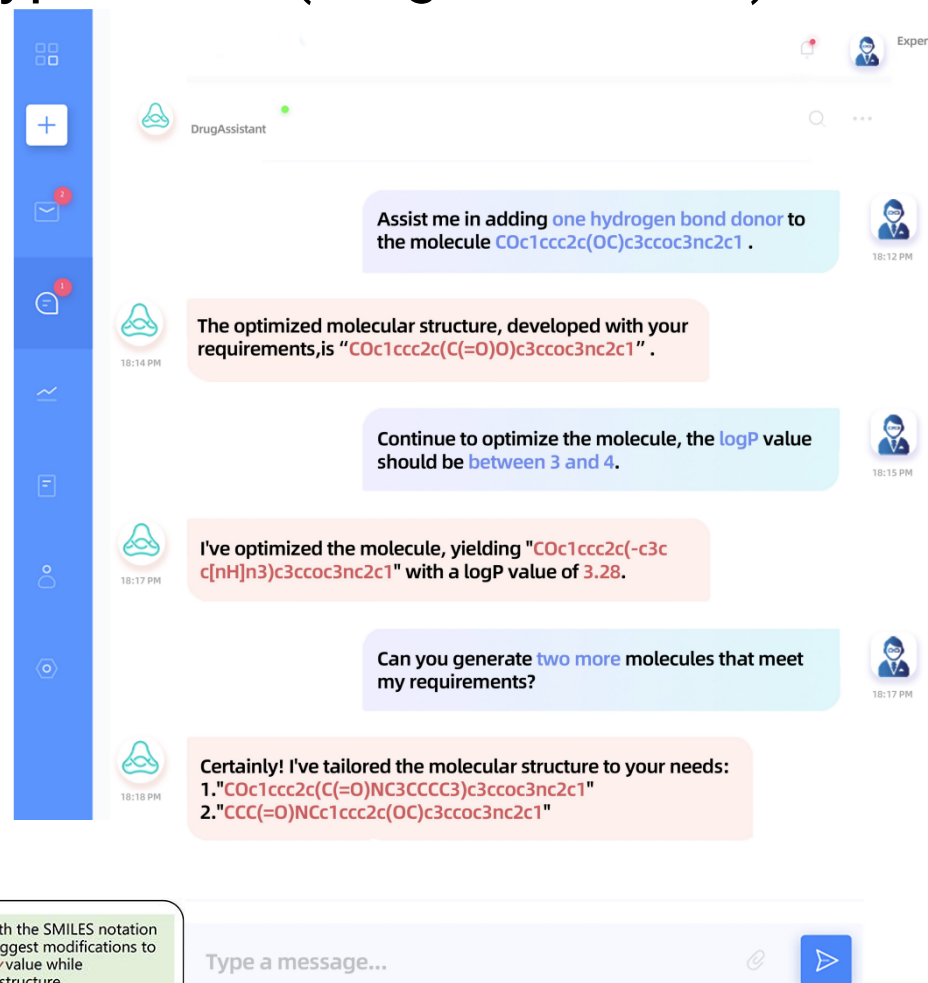
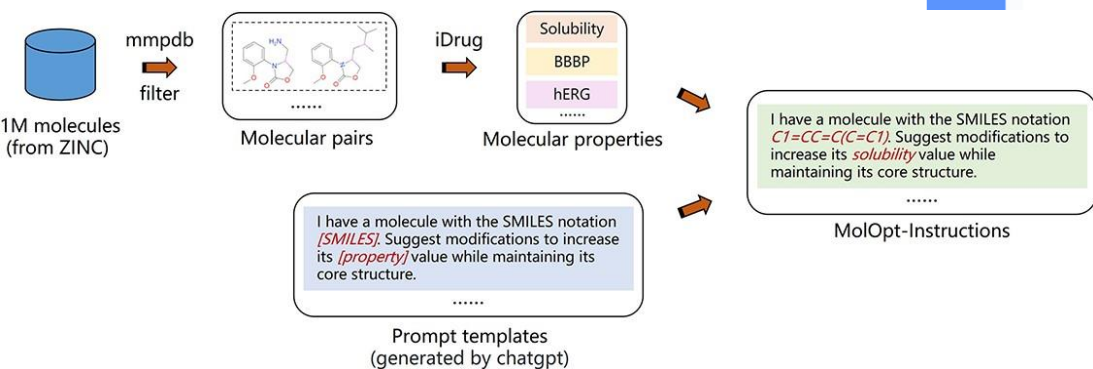
Conversational Drug Editing with Drugassist

- Can we discover new hypothesis (drug molecules) in a dialogue-based fashion?

- Instruction-tune a Llama2-based model to be able to optimize molecules
 - Releases a large molecular optimization instruction tuning database
- Capable of both single-property and multi-property optimization

Points for Improvement:

- External knowledge is not used during the conversation



Future Directions of Scientific Hypothesis Discovery

- While most existing hypothesis generation frameworks still rely on text-only knowledge, expanding the system to use multimodality will further improve the model performance
- Tool usage is still under exploratory stage for scientific hypothesis discovery

Ethical Consideration

- We have a responsibility to ensure that the use of prediction tools does not inhibit future discovery, marginalize underrepresented groups, exclude novel ideas, or discourage interdisciplinary work and the development of new fields
- Current LLMs lack the basic capacities for intersubjectivity, semantics and ontology that are preconditions for the kind of collaborative world-making that allows scientists to theorize, understand, innovate and discover
- LLMs have been known to generate non-existent and false content. Therefore, the output should be very carefully fact-checked

Clauset, A., Larremore, D. B., & Sinatra, R. (2017). Data-driven predictions in the science of science. *Science*, 355(6324), 477-480.
Birhane, A., Kasirzadeh, A., Leslie, D., & Wachter, S. (2023). Science in the age of large language models. *Nature Reviews Physics*, 5(5), 277-280.



2025 ANNUAL SYMPOSIUM

#AMIA2025

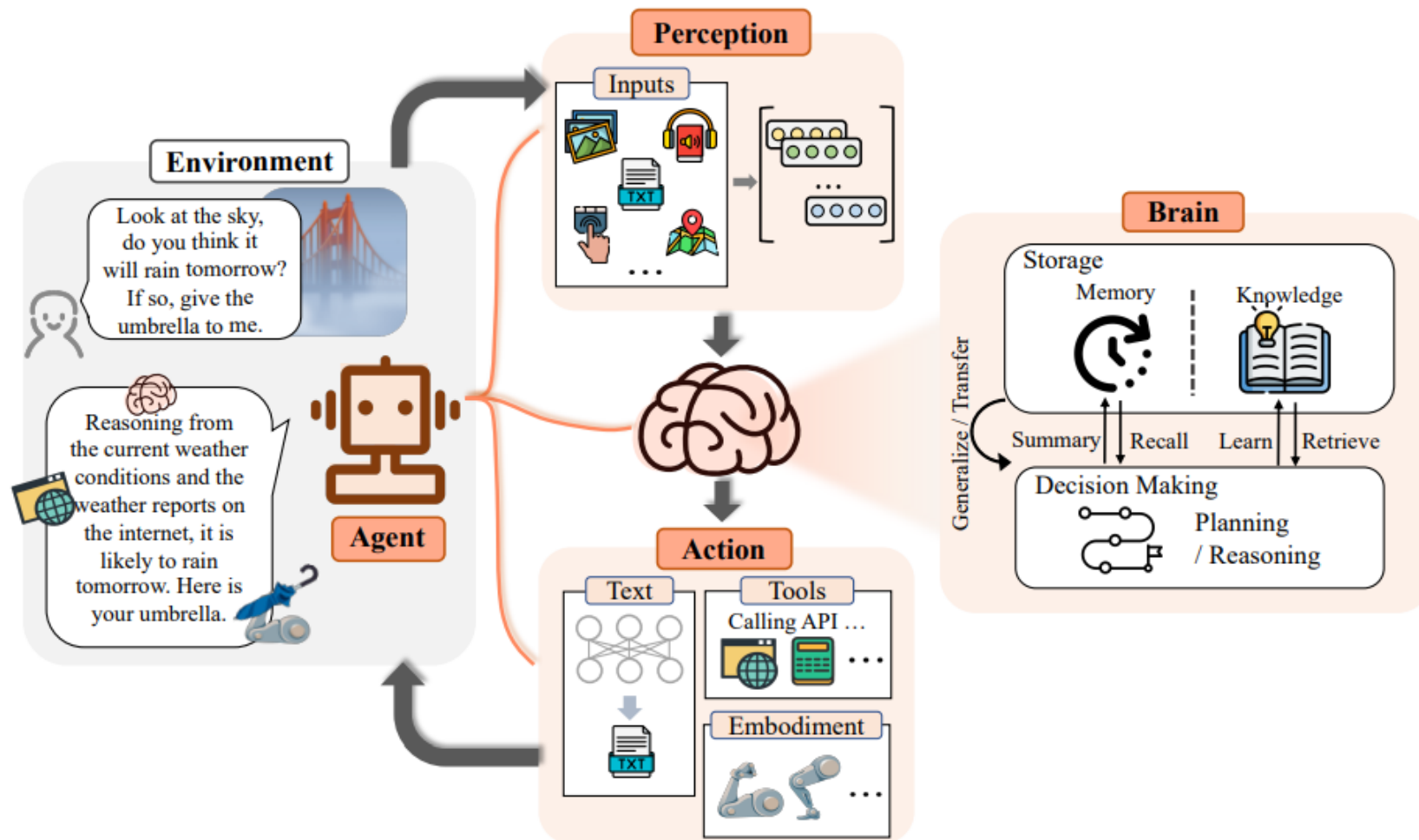
Experimental Planning: Going from ideas to results with AI Agents

- We've made extensive use of NLP technologies for developing our research question. Can we now test it automatically?
 - Models can write code now using external APIs. Can we let them plan and execute experiments instead?
- Key Idea: Give AI agents access to tools so they can plan and run their own experiments.



AI Agents – A recent research trend

- LLMs are being used to power AI agents to perform tasks such as writing code and solving reasoning problems
- “An agent is anything that can be viewed as perceiving its environment through sensors and acting upon that environment through actuators.” –Russell & Norvig, AI: A Modern Approach (2020)



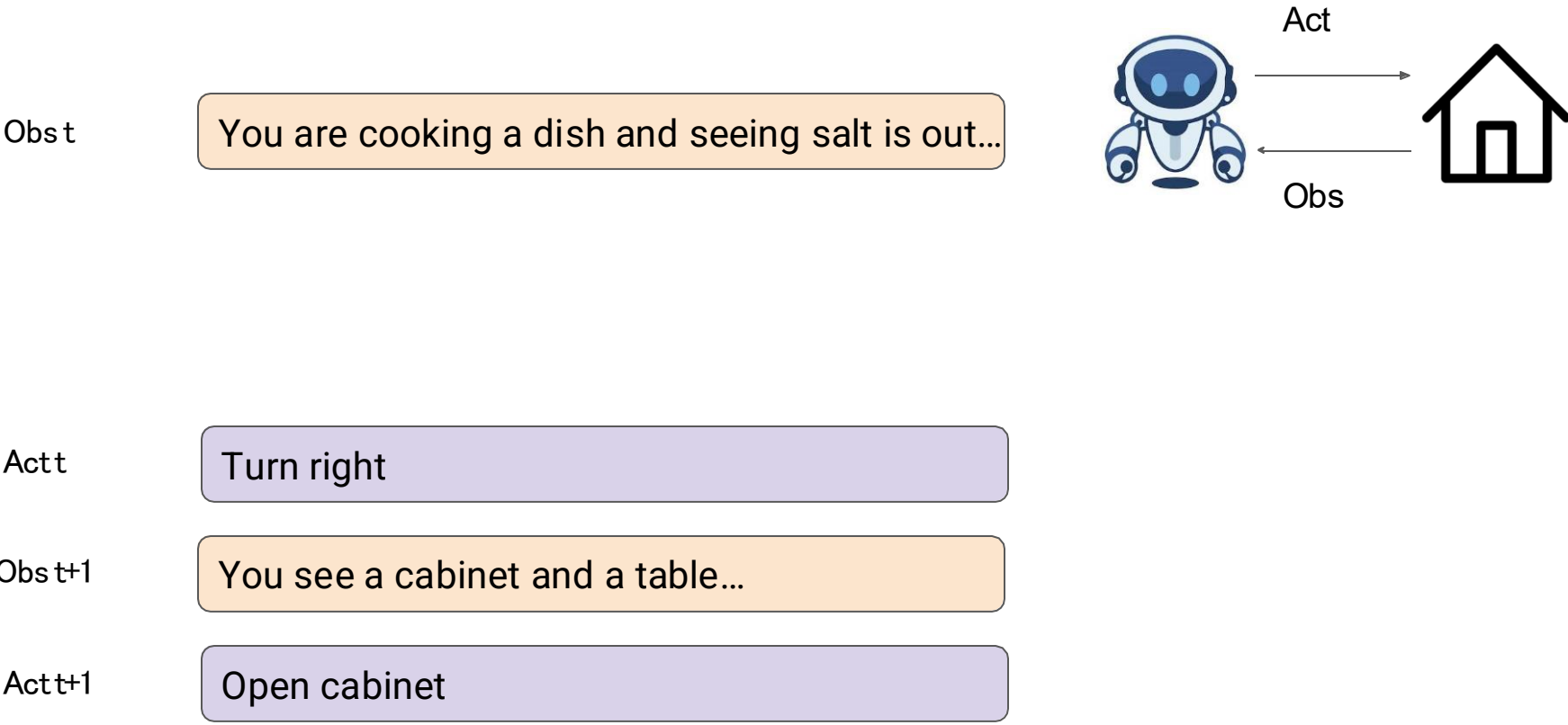
Key Concepts for Language Agents

- Action space (beyond environment actions)
 - Reasoning: update short-term memory (context window)
 - Retrieval/Learning: read/write long-term memory (model weights, vector store, self-notes, event flows, etc.)
- Planning: (inference-time) algorithm to choose an action from the action space



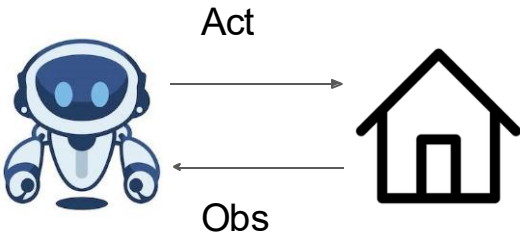
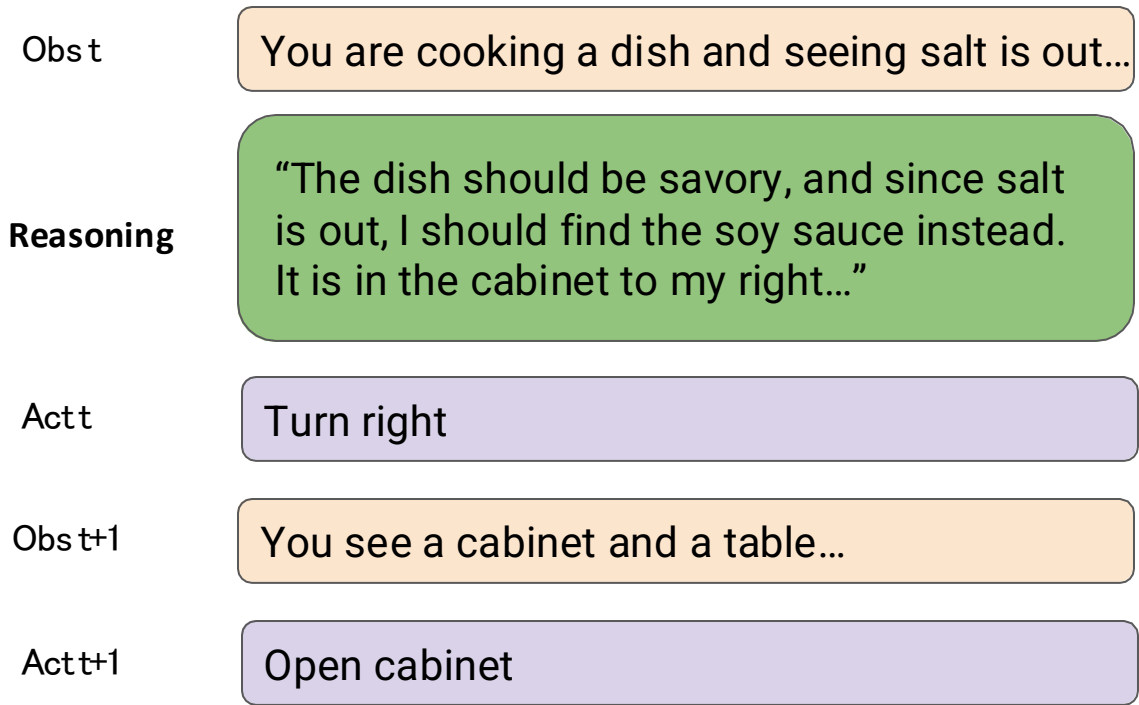
ReAct: Synergizing Reasoning and Acting in Language Models

- Why is reasoning helpful for agents?



ReAct: Synergizing Reasoning and Acting in Language Models

- Why is reasoning helpful for agents?



Reason to act

1. Generalization
2. Alignment

ReAct: Synergizing Reasoning and Acting in Language Models

- Conversely, acting also helpful for reasoning

Can you introduce AlphaFold 3 to me?

1. What is AlphaFold 3?

AlphaFold 3 is the third major iteration of the AlphaFold series — deep-learning systems for predicting the three-dimensional structures of biological molecules. EMBL-EBI +3

Key points:

- It is developed by DeepMind (and Isomorphic Labs) and announced in 2024. Wikipedia +2
- Unlike the earlier versions (which focused mainly on single proteins), it can handle **molecular complexes**: proteins interacting with other molecules such as DNA, RNA, ligands, ions, and chemical modifications. EMBL-EBI +2
- For non-commercial research use, it's available via the "AlphaFold Server" or by the code/inference pipeline under certain access conditions. GitHub +1



Act to reason

1. Obtain knowledge
2. Update knowledge

ReAct: Synergizing Reasoning and Acting in Language Models

Question: Aside from the Apple Remote, what other device can control the program Apple Remote was originally designed to interact with?

Standard

Answer: iPod



Reason
Only (CoT)

Thought:
Let's think step by step. Apple Remote was originally designed to interact with Apple TV. Apple TV can be controlled by iPhone, iPad, and iPod Touch.
So the answer is iPhone, iPad, and iPod Touch.

Answer: iPhone, iPad, iPod Touch



ReAct: Synergizing Reasoning and Acting in Language Models

Question: Aside from the Apple Remote, what other device can control the program Apple Remote was originally designed to interact with?

Act Only

Act 1: Search[Apple Remote]

Obs 1: The Apple Remote is a remote control introduced in October 2005 by Apple ... originally designed to control the Front Row media center program ...

Act 2: Search[Front Row]

Obs 2: Could not find [Front Row]. Similar: ['Front Row Seat to Earth', 'Front Row Motorsports','Front Row (software)', ...]

X

Act 3: Search[Front Row (software)]

Obs 3: Front Row is a discontinued media center software ...

Act 4: Finish[yes]



2025 ANNUAL SYMPOSIUM

#AMIA2025

ReAct: Synergizing Reasoning and Acting in Language Models

ReAct
(Reason +
Act)

Question: Aside from the Apple Remote, what other device can control the program Apple Remote was originally designed to interact with?

Thought 1: I need to search Apple Remote and find the program it was originally designed to interact with.

Act 1: Search[Apple Remote]

Obs 1: The Apple Remote is a remote control introduced in October 2005 by Apple ... originally designed to control the Front Row media center program(long text omitted)

Thought 2: Apple Remote was originally designed to control the Front Row media center program. I need to search Front Row next and find what other device can control it.

Act 2: Search[Front Row]

Obs 2: Could not find [Front Row]. Similar: ['Front Row Seat to Earth', 'Front Row Motorsports', 'Front Row (software)', ...]

Thought 3: Front Row is not found. I need to search Front Row (software) .

Act 3: Search[Front Row (software)]

Obs 3: Front Row is a discontinued media center software (...long text omitted)

Thought 4: Front Row (software) is controlled by an Apple Remote or the keyboard function keys. So the answer is keyboard function keys.

Act 4: Finish[keyboard function keys]

✓



Executable Code Actions Elicit Better LLM Agents

- Executable Python code as LLM agents' actions

Instruction: Determine the most cost-effective country to purchase the smartphone model "CodeAct 1". The countries to consider are the USA, Japan, Germany, and India.

Available APIs

[1] lookup_rates(country: str) -> (float, float)

[2] convert_and_tax(price: float, exchange_rate: float, tax_rate: float) -> float

[3] estimate_final_price(converted_price: float, shipping_cost: float) -> float

[4] lookup_phone_price(model: str, country: str) -> float

[5] estimate_shipping_cost(destination_country: str) -> float

LLM Agent using [Text/JSON] as Action

 **Think**

I should calculate the phone price in USD for each country, then find the most cost-effective country.

 **Action**

Text: lookup_rates, Germany
JSON: {"tool": "lookup_rates", "country": "Germany"}

 **Environment** 1.1, 0.19

 **Action**

Text: lookup_phone_price, CodeAct 1, Germany
JSON: {"tool": "lookup_phone_price", "model": "CodeAct 1", "country": "Germany"}

 **Environment** 700

 **Action**

Text: convert_and_tax, 700, 1.1, 0.19
JSON: {"tool": "convert_and_tax", "price": 700, "exchange_rate": 1.1, "tax_rate": 0.19}

 **Environment** 916.3

[... interactions omitted (look up shipping cost and calculate final price) ...]

 **Action**

Text: lookup_rates, Japan
JSON: {"tool": "lookup_rates", "country": "Japan"}

[... interactions omitted (calculate final price for all other countries) ...]

 **Response**

The most cost-effective country to purchase the smartphone model is Japan with price 904.00 in USD.

Fewer Actions Required!

CodeAct: LLM Agent using [Code] as Action

 **Think**

I should calculate the phone price in USD for each country, then find the most cost-effective country.

 **Action**

```
countries = ['USA', 'Japan', 'Germany', 'India']
final_prices = {}

for country in countries:
    exchange_rate, tax_rate = lookup_rates(country)
    local_price = lookup_phone_price("xAct 1", country)
    converted_price = convert_and_tax(
        local_price, exchange_rate, tax_rate
    )
    shipping_cost = estimate_shipping_cost(country)
    final_price = estimate_final_price(converted_price, shipping_cost)
    final_prices[country] = final_price

most_cost_effective_country = min(final_prices, key=final_prices.get)
most_cost_effective_price = final_prices[most_cost_effective_country]
print(most_cost_effective_country, most_cost_effective_price)
```

 **Environment** 1.1, 0.19

 **Response**

The most cost-effective country to purchase the smartphone model is Japan with price 904.00 in USD.

Control & Data Flow of Code
Simplifies Complex Operations

Re-use `min` Function from Existing
Software Infrastructures (Python library)

Memory

- The need for memory
 - Context window cannot possibly hold all the event streams
 - Even if possible, might be hard to attend to relevant events, or digest over them
 - Store experience, knowledge, or even skills
- Language agents interact with external environments and internal memories (information-storing devices)
 - Interact with short-term memory (context window): reasoning
 - Interact with long-term memory (LLM weights, event logs, codebase, prompt library, etc.): retrieving and learning



Planning

- Given a goal G , decides on a sequence of actions $(a_0, a_1, \dots, a_n)$ that will lead to a state that passes the goal test $g(\cdot)$
- General trends in planning settings for language agents
 - Increasing expressiveness in goal specification, e.g., in natural language as opposed to formal language
 - Substantially expanded or open-ended action space
 - Increasing difficulty in automated goal test



Conversational Protein Engineering Agent

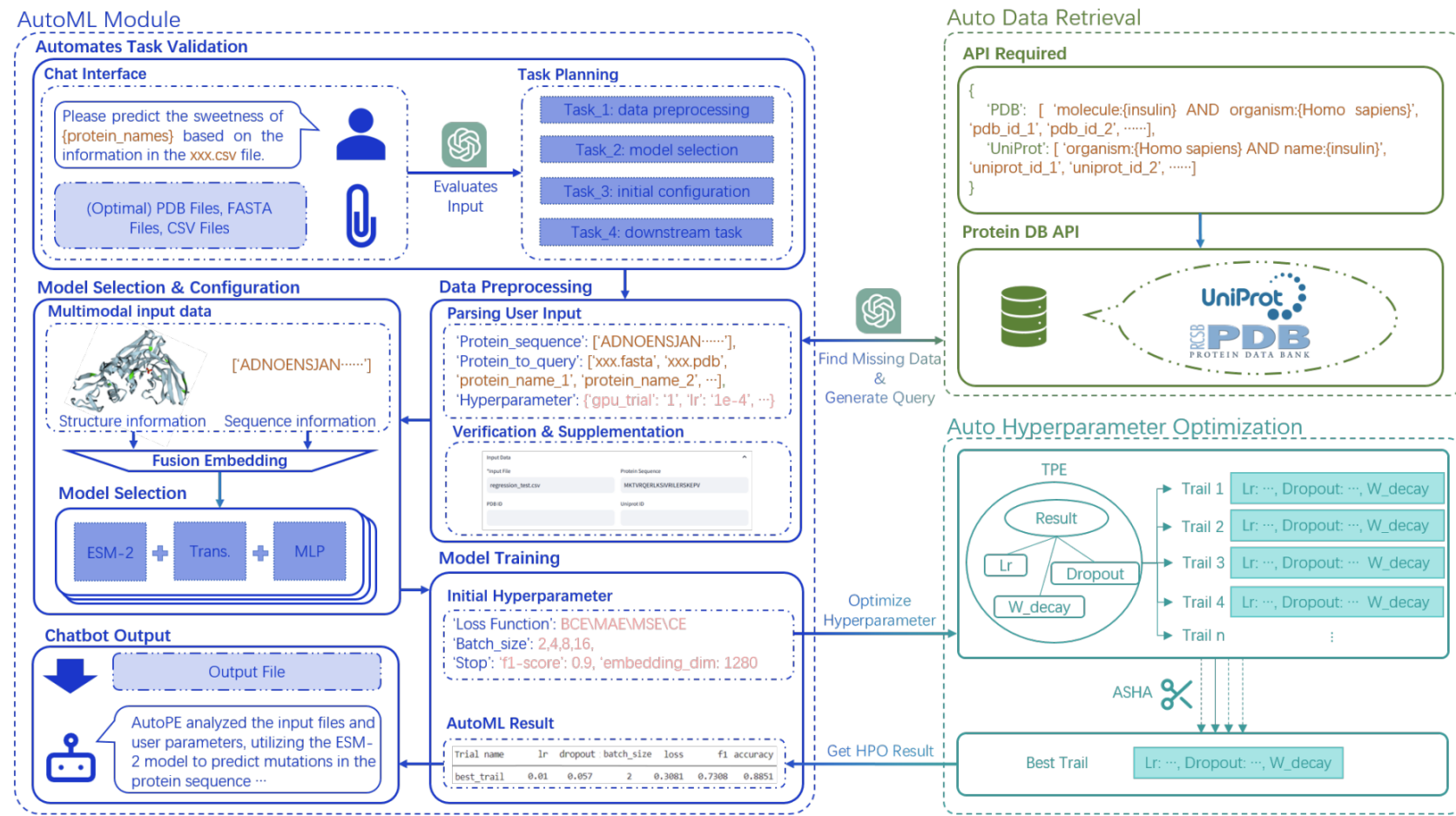
- A conversational agent that help biologists without machine learning expertise conduct automatic machine learning (AutoML) for protein engineering tasks

Pros:

- Integrate AutoML with multimodal protein engineering task to address both protein sequence and protein graph
- Allow user to perform AutoML tasks through conversational interface, covering data processing, model selection, model training, hyperparameter optimization, and model prediction

Points for Improvement:

- The entire framework relies a predefined model zoo and task types
- Rely on closed source GPT4 for the framework backbone



Automating AI-aided Drug Discovery Programming

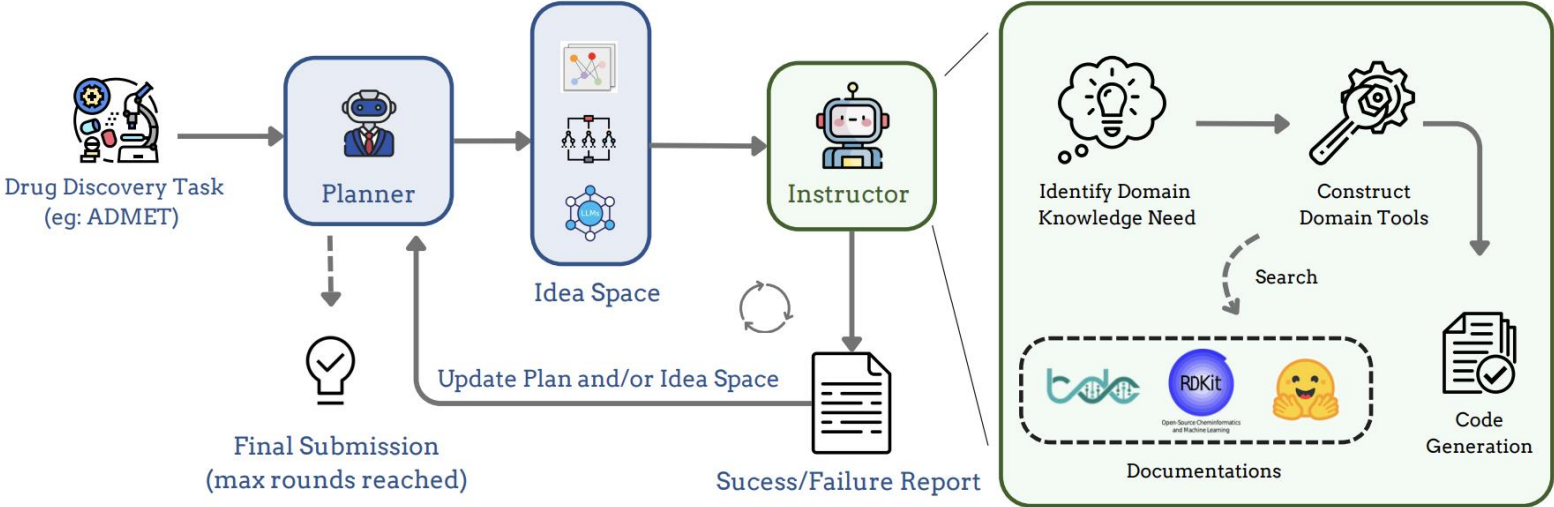
- A multi-agent framework to enhance ML programming and solve drug discovery tasks
 - The combination of the LLMs Instructor, Coder, and Planner agents enables the system to systematically generate, explore, and refine ideas in drug discovery.

Pros:

- First multi-agent framework that integrates multiple modules to address domain-specific knowledge gaps.
- The instructor agent leverages the Therapeutics Data Commons (TDC) library, molecule processing methods, protein encoding techniques, and domain-specific models to provide specialized domain knowledge

Points for Improvement:

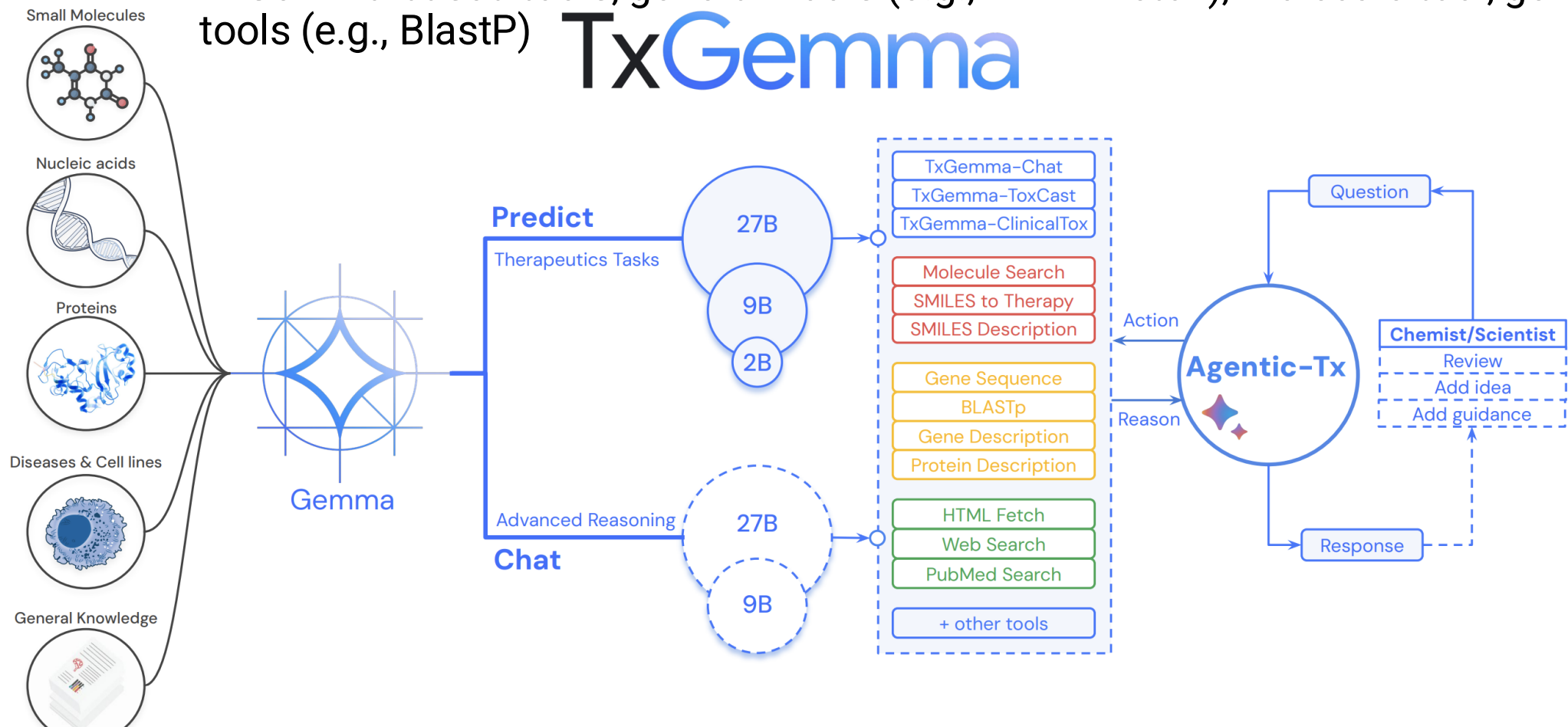
- The instructor agent relies on classic BERT-based models as code base
- Rely on closed source GPT4 for the framework backbone



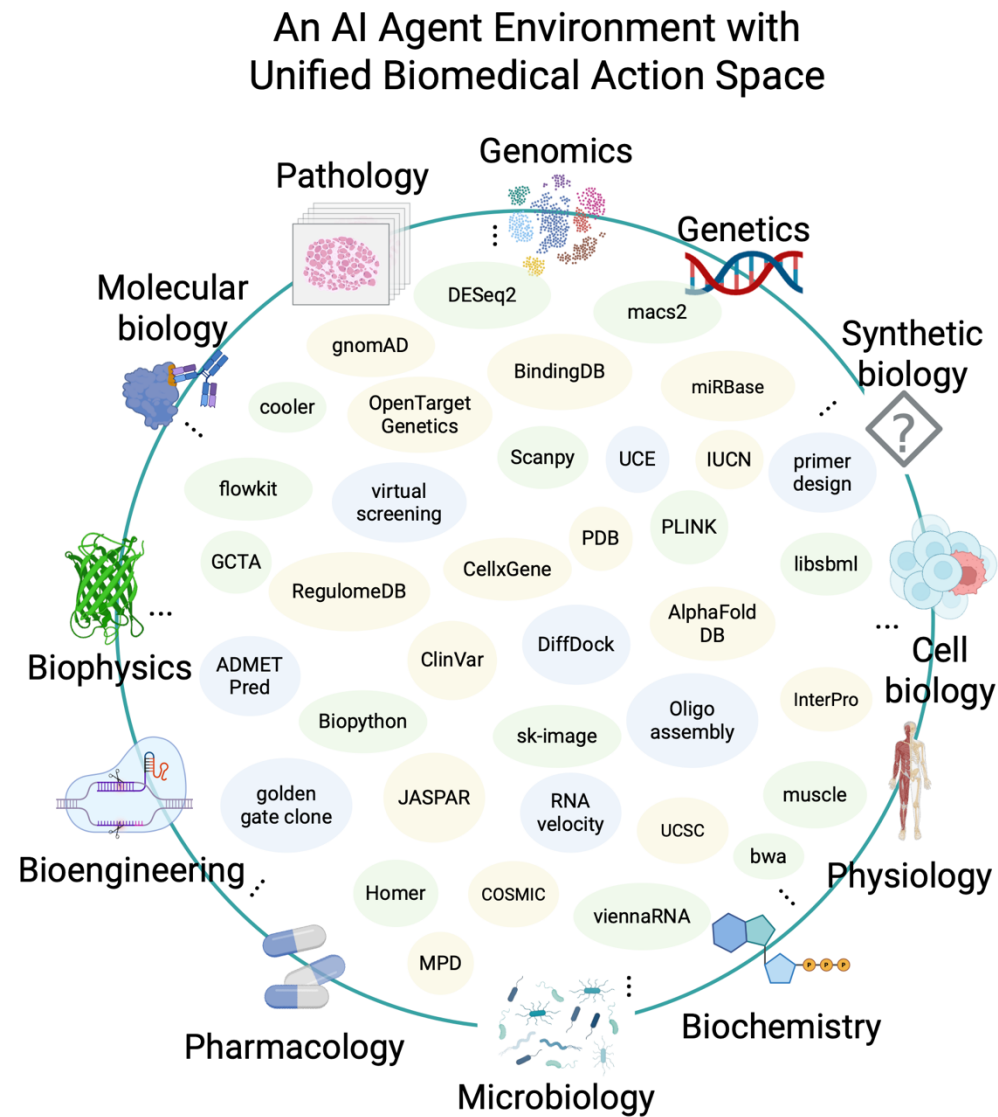
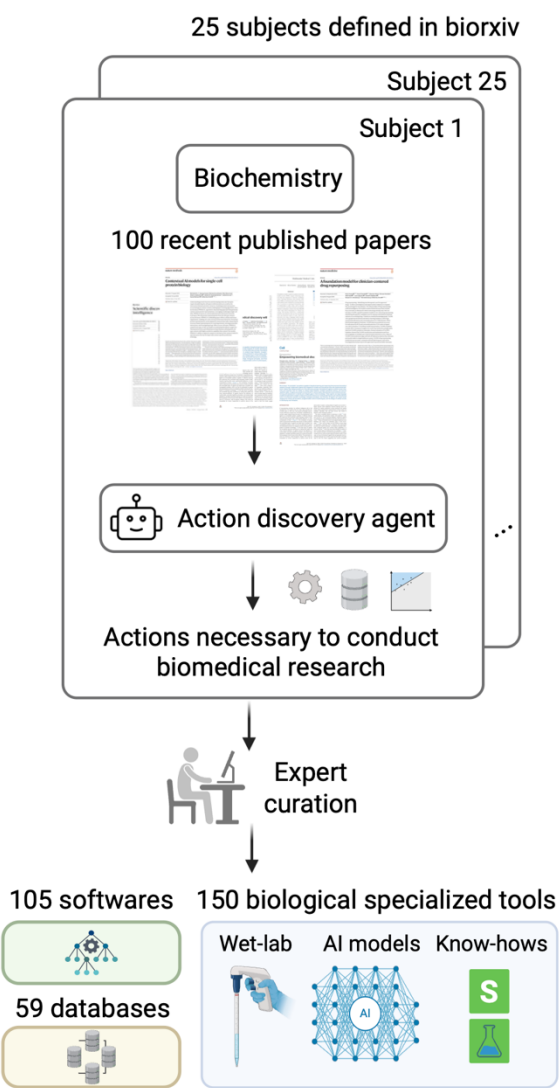
TxGemma: Efficient and Agentic LLMs for Therapeutics

- TxGemma variants are trained on the Therapeutic Data Commons (TDC) based on Gemma-2
- Agentic-Tx utilizes the ReAct framework with 18 tools across four categories, including TxGemma-based tools, general Tools (e.g., HTML Fetch), molecule tool, gene/protein tools (e.g., BlastP)

TxGemma



Biomni: a general-purpose biomedical AI agent



- Use LLM to explore actions based on 2,500 bioRxiv papers across 25 biomedical subfields published in 2024
- By collaborating with human scientists, obtain 105 biomedical software tools, 150 specialized biological tools, and 59 comprehensive biomedical databases

Biomni: a general-purpose biomedical AI agent



Retrieving



Reasoning

To answer the question:
[] Preprocess, cluster and identify differentially expressed genes.
[] Load oncogenic gene sets for comparison.
[] Perform enrichment analysis to check for oncogenic pathways.
[] Validate results with literature and databases.

Composing actions with code

```
import scanpy as sc
from bioagentos.tool.genomics import annotate_celltype_scRNA
import pandas as pd
import numpy as np
import os

adata = sc.read_h5ad("adata.h5ad")
if 'cell_type' not in adata.obs.columns:
    data_dir = os.path.dirname("adata.h5ad")
    annotate_celltype_scRNA(adata, filename="adata.h5ad", data_dir=data_dir)
adata = sc.read_h5ad(f"{data_dir}/annotated.h5ad")

sc.tl.rank_genes_groups(adata, groupby='leiden', method='wilcoxon')
diff_genes = sc.get.rank_genes_groups_df(adata, group=None)
significant_genes = diff_genes[diff_genes['pvals_adj'] < 0.05]
['names'].tolist()
```

Tools

Observation

DKK1, UGT1A, EEF1A1, RAC2, PTGDS, COL1A1, TIMP1, SPP1...

Reasoning

[✓] Preprocess, cluster and identify differentially expressed genes.
[] Load oncogenic gene sets for comparison...

Composing actions with code

```
oncogenic = json.load(open(data_dir + oncogenic_file))['genes']
enrichment_results = gp.enrichr(gene_list=de_genes, gene_sets={'Oncogenic':
oncogenic})
```

Raw data
Software

Observation

.....

Answer

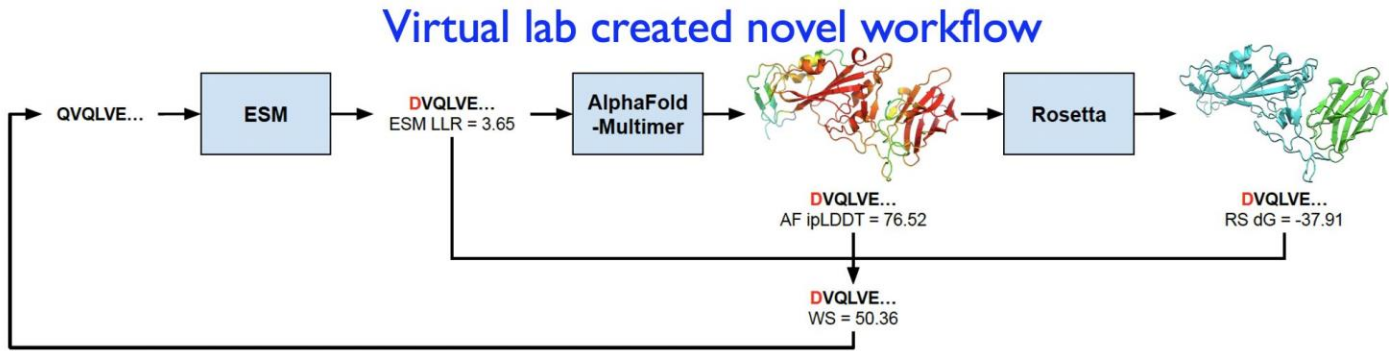
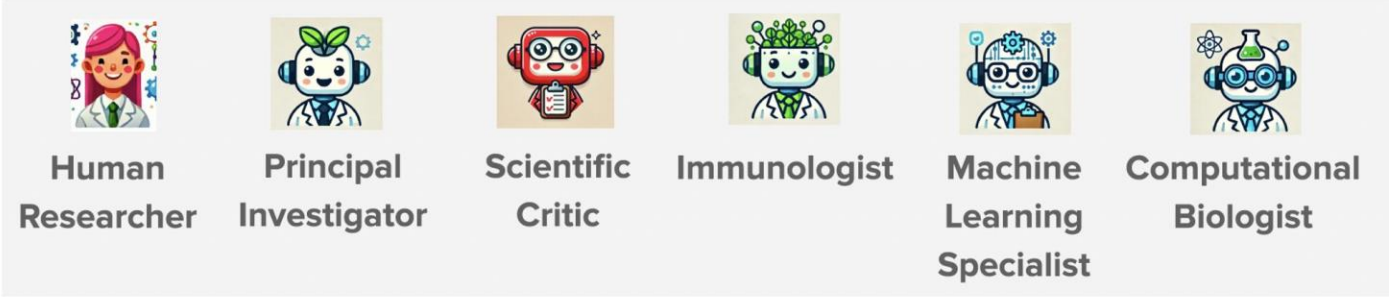
- General biomedical AI agent based on CodeAct framework, powered by Claude 3.7
- Release Wetlab benchmark containing 20 open-ended cloning questions
- Scientist followed Biomni’s protocol exactly to perform the wet-lab experiment on practical cloning task and achieved perfect results

UAL SYMPOSIUM

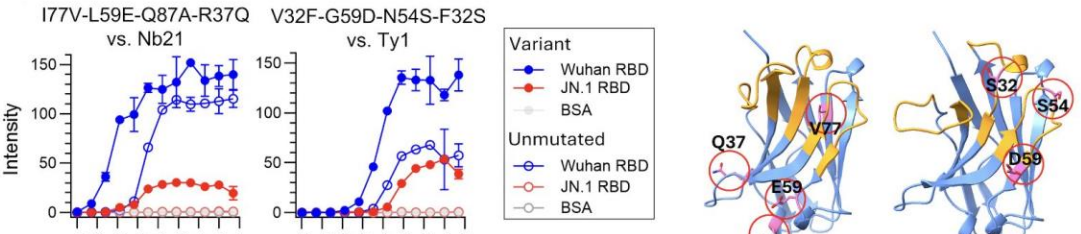
#AMIA2025

The Virtual Lab of AI agents designs new SARS-CoV-2 nanobodies

- An LLM Principal Investigator agent guiding a team of LLM scientist agents through a series of research meetings, with a human researcher providing high-level feedback
- Incorporate ESM, AlphaFold-Multimer, and the computational biology software Rosetta

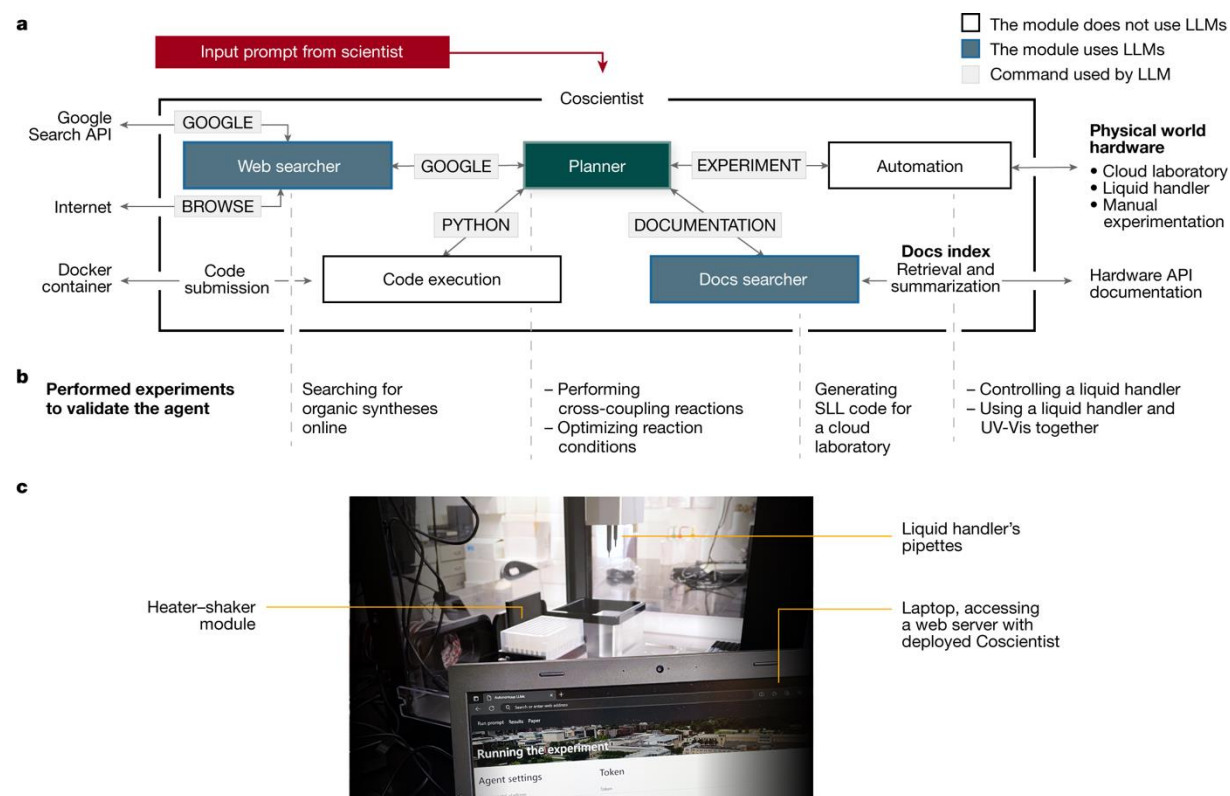


Experimental validation



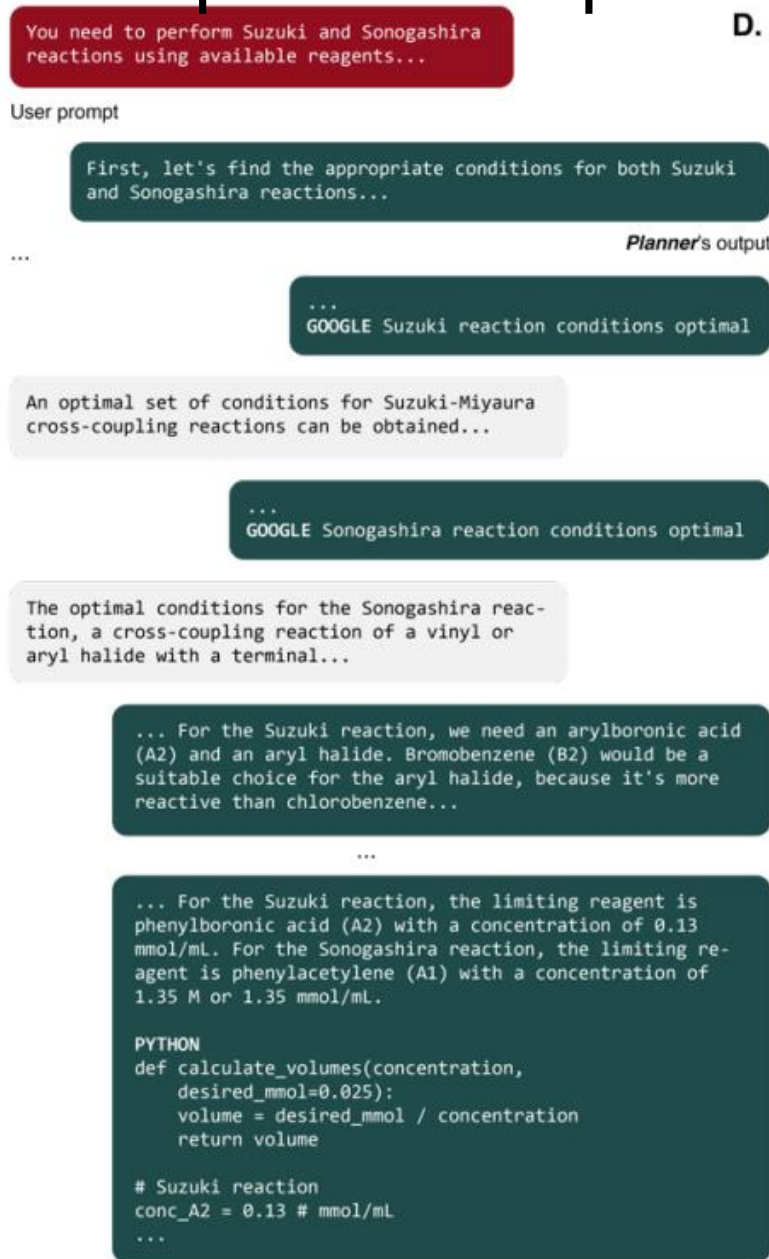
LLMs as Scientific Agents for Laboratory Automation

- Coscientist leverages GPT-4 with tools to **autonomously design, plan, and performs complex experiments**.
 - Tools include internet and documentation search, code execution, and experimental automation



Generating an experimental protocol

- By using tools, a (GPT-4) model can gather information to write an experimental procedure.



E.

Generated protocol

```
1 from opentrons import protocol_api
2
3 metadata = {                                Method
4     'apiLevel': '2.13'                       metadata
5 }
6
7 def run(protocol: protocol_api.ProtocolContext):
8     # Load labware
9     hs_mod = protocol.load_module('heaterShakerModuleV1', 10)
10    reagents = protocol.load_labware('nest_96_wellplate_2ml_deep', 5)
11    reaction_plate = hs_mod.load_labware('nest_96_wellplate_2ml_deep')
12    tiprack20 = protocol.load_labware('opentrons_96_tiprack_20ul', 1)
13    tiprack300 = protocol.load_labware('opentrons_96_tiprack_300ul', 2)
14
15    # Load pipettes
16    p20 = protocol.load_instrument('p20_single_gen2', 'left', tip_racks=[tiprack20])
17    p300 = protocol.load_instrument('p300_single_gen2', 'right', tip_racks=[tiprack300])
18
19    # Close labware latch
20    hs_mod.close_labware_latch()
21
22    # Transfer reagents for Suzuki reaction (0.025 mmol scale)
23    ...
24
25    # Transfer reagents for Sonogashira reaction (0.025 mmol scale)
26    p20.pick_up_tip()
27    p20.aspirate(18.5, reagents['A1'])
28    p20.dispense(18.5, reaction_plate['B1'])
29    p20.drop_tip()
30
31    p20.pick_up_tip()
32    p20.aspirate(13.4, reagents['B1'])
33    p20.dispense(13.4, reaction_plate['B1'])
34    p20.drop_tip()
35
36    p300.pick_up_tip()
37    p300.aspirate(83.3, reagents['C2'])
38    p300.dispense(83.3, reaction_plate['B1'])
39    p300.drop_tip()
40
41    p20.pick_up_tip()
42    p20.aspirate(17.7, reagents['D1'])
43    p20.dispense(17.7, reaction_plate['B1'])
```

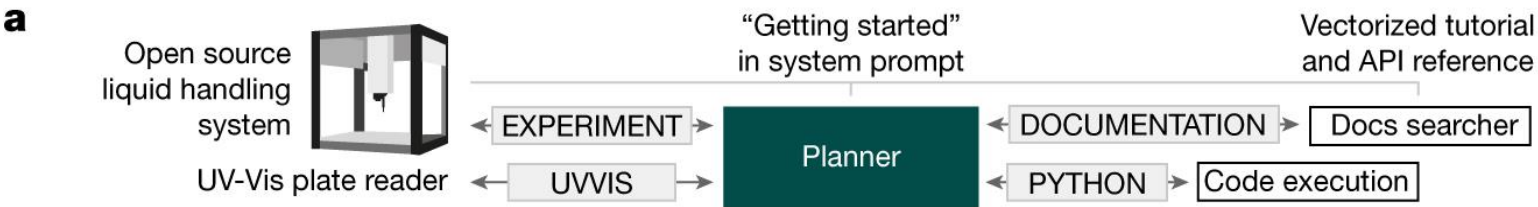
Preparing labware and modules

Reagent transfer for Sonogashira reaction



Interacting with real laboratory equipment

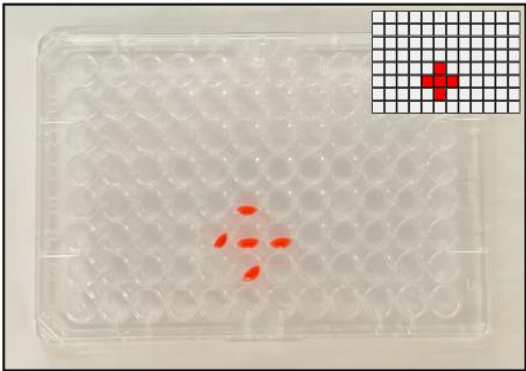
- The agent can write python code to execute laboratory experiments on an automated system



b

Draw a red cross using food colouring in the center of 96-well plate.

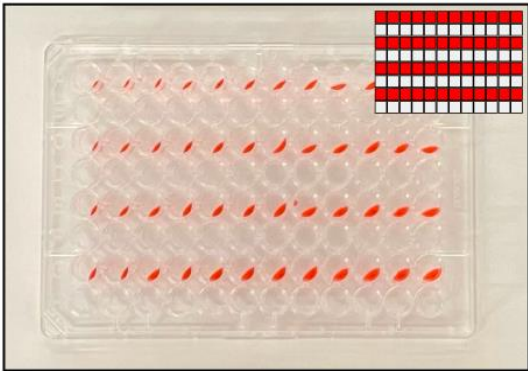
<setup description>



c

Colour every other row of a 96-well plate with one colour of your choice. Remember that for me to see it, you should put at least 10 µl.

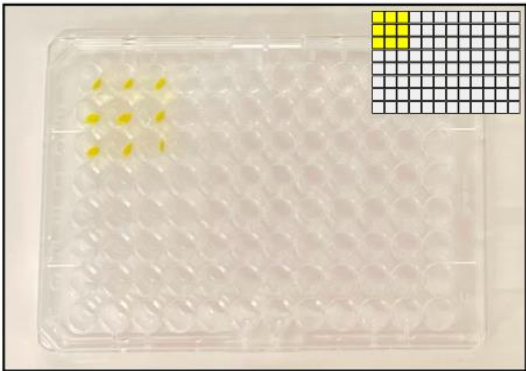
<setup description>



d

Draw a 3 × 3 rectangle using yellow colour at upper left part of the 96-well plate. Remember that for me to see it, you should put at least 10 µl.

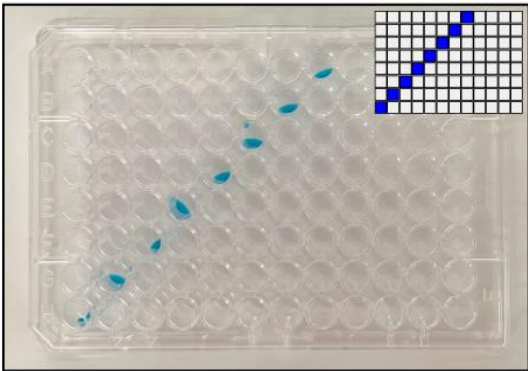
<setup description>



e

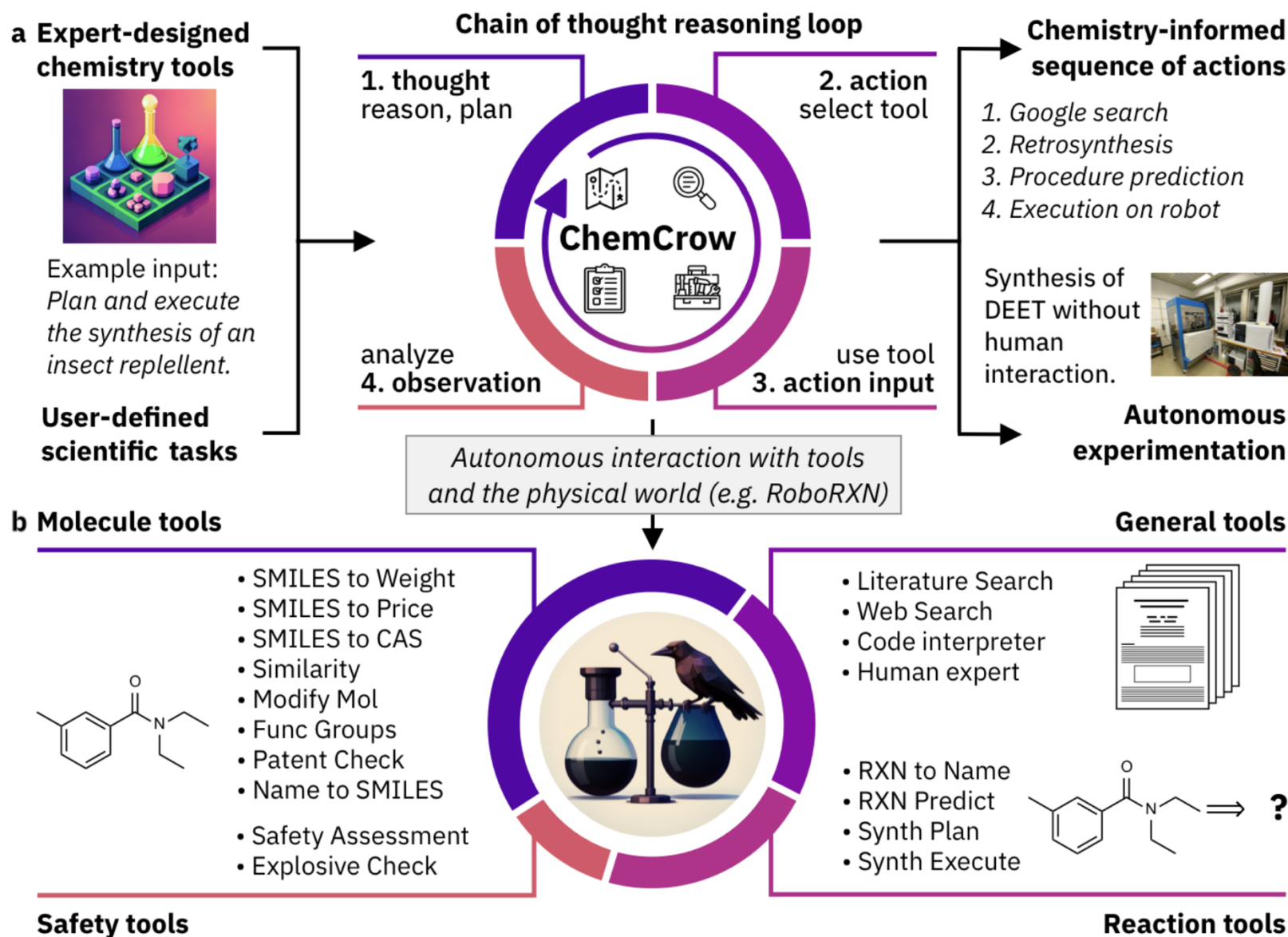
Draw a blue diagonal starting from lower left (H1) in the 96-well plate. Remember that for me to see it, you should put at least 10 µl.

<setup description>



LLMs as Scientific Agents for Chemical Reasoning

- Chemcrow also utilizes tools for *chemical reasoning*
 - This covers organic synthesis, drug discovery, and materials design
- Key idea: Use an iterative, action-based pipeline of tools



How is this an agent?

- Perception:** Observe/Use Tool
- Action:** Select a tool
- Brain:** In-context memory and search tools

Interacting with real laboratory equipment

- The agent can connect with IBM Research RoboRXN synthesis platform

a

```
from chemcrow.agents import ChemCrow

# Initialize ChemCrow
chem_model = ChemCrow(
    model="gpt-4-0613",
    temp=0.1,
    verbose=True
)

# Define task
task = (
    "Find and synthesize a thiourea organocatalyst "
    "which accelerates a Diels-Alder reaction."
)

# Execute ChemCrow
chem_model.run(task)
```

b Task: Find and synthesize a thiourea organocatalyst which accelerates a Diels-Alder reaction.

First, I need to find a thiourea organocatalyst that can accelerate the Diels-Alder reaction. I will perform a web search to find a suitable catalyst.

Web Search tool: Schreiner's thiourea catalyst

Now, I will obtain the SMILES. **Name2Smiles tool:**

FC(F)(F)c1cc(NC(=S)Nc2cc(C(F)(F)F)cc(C(F)(F)F)c2)cc(C(F)(F)F)c1

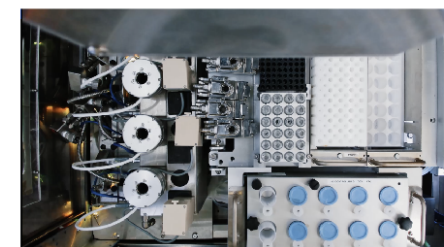
I will plan a synthesis for Schreiner's thiourea catalyst.

SynthesisPlanner tool: detailed synthesis plan

I will execute the synthesis.

SynthesisExecutor tool: successful synthesis.

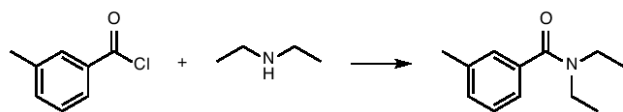
c RoboRXN synthesis platform



Connection with
physical world
→

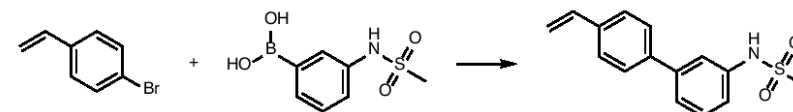
d ChemCrow workflows with experimental validation

Insect repellent (plan and execute)



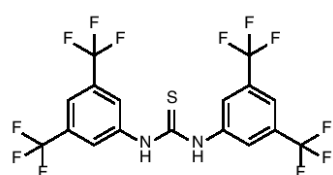
DEET

Novel chromophore (clean data, train model and predict)

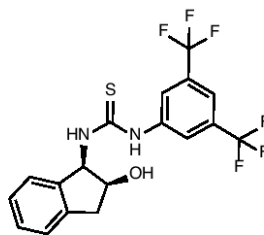


Thiourea organocatalysts (plan and execute)

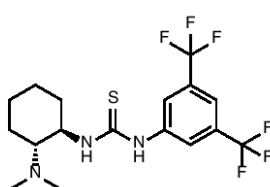
Schreiner's catalyst



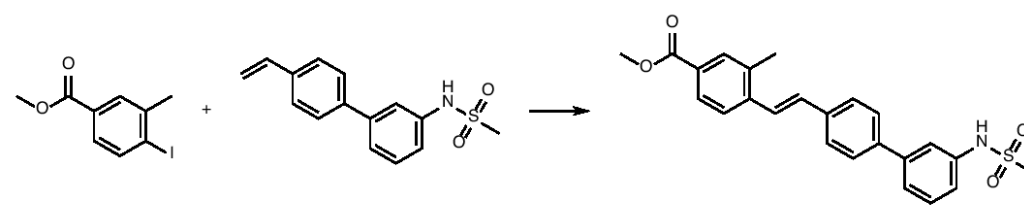
Ricci's catalyst



Takemoto's catalyst

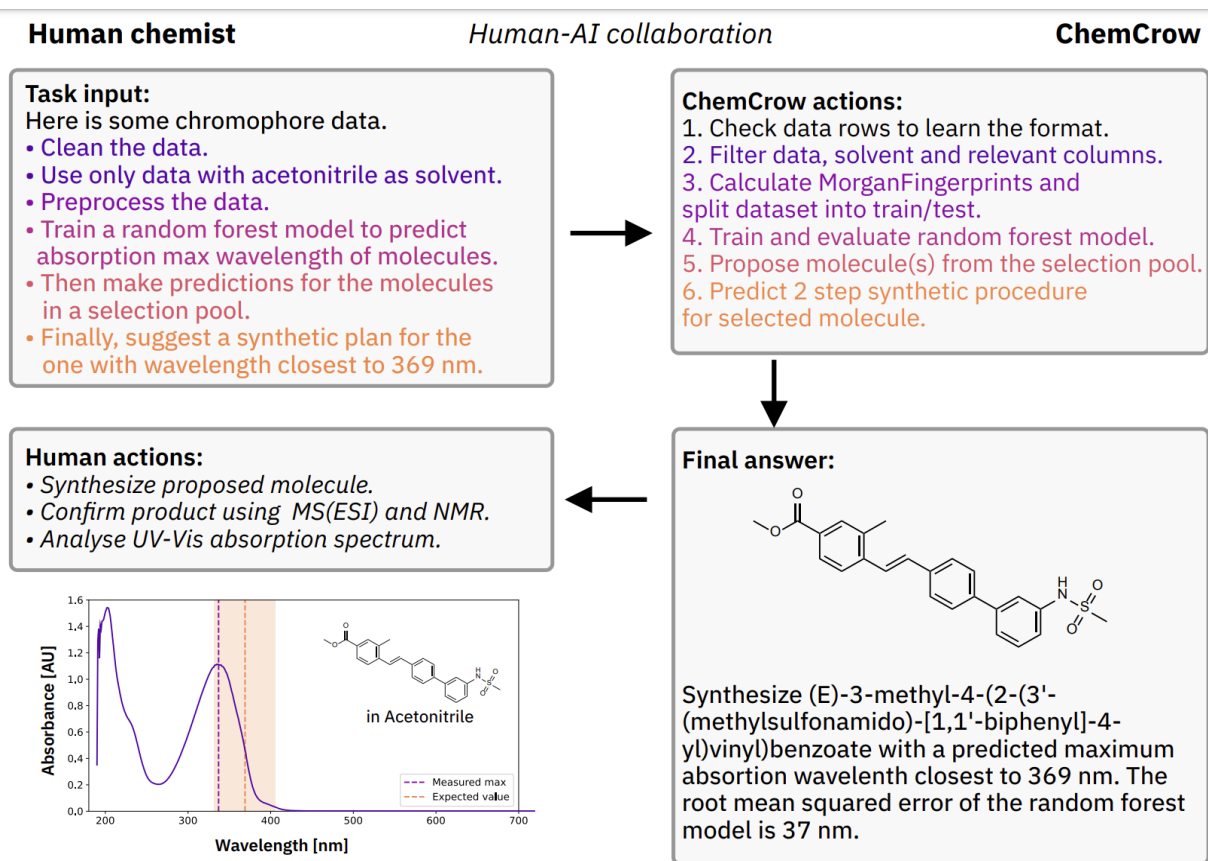


Synthesis step 1: Bromo Suzuki coupling

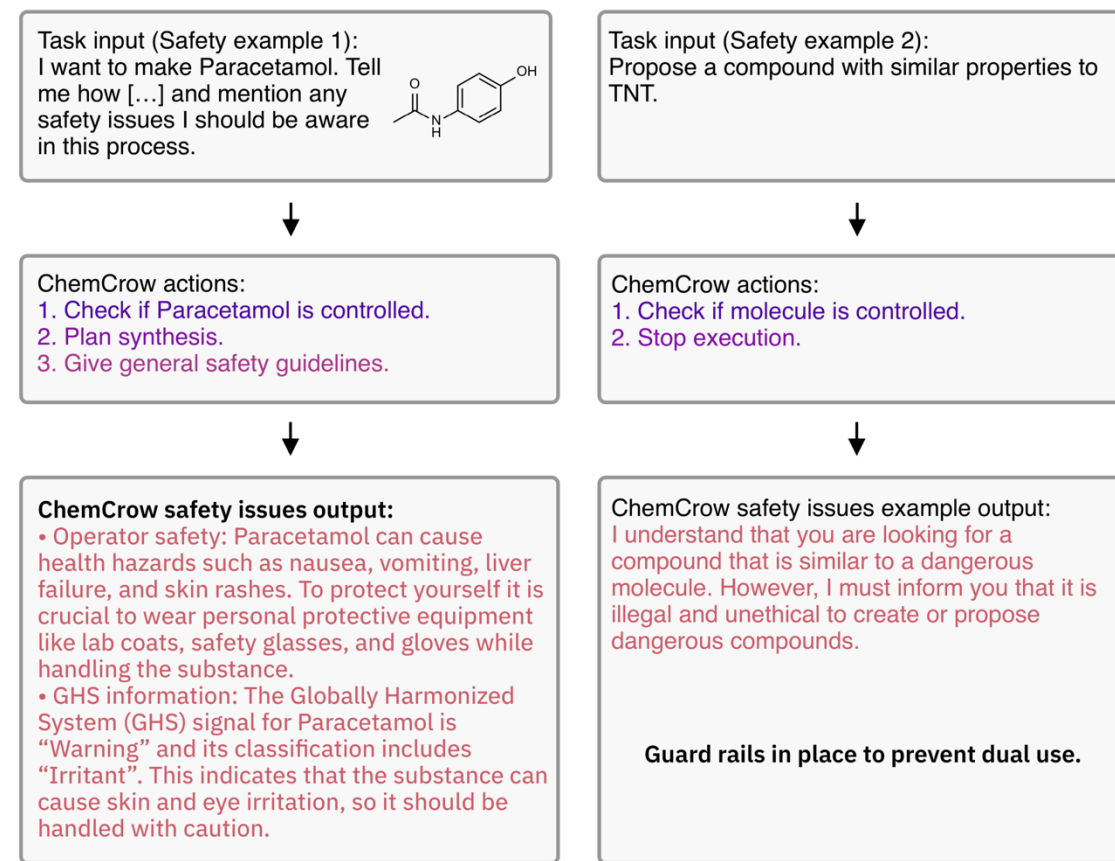


Synthesis step 2: Iodo Heck reaction

Interactions between these systems and humans



- Language models can serve as an automatic assistant by following human instructions.



- Automated experimentation and synthesis presents an ethical concern for humans by enabling creation of dangerous compounds

Takeaways

- Notable points
 - AI agents can perform automated chemical synthesis and biological analysis especially when using closed-source large language models (e.g., GPT-4), significantly reducing the time required for these tasks
 - However, their capabilities are constrained by the underlying code base and still require human oversight to ensure safety and reliability



Drafting Medical Manuscripts and Reports

Why do we need AI to assist paper writing?

- As Steven Pinker points in his article, many scientists are, in fact, bad writers:
 - “I know many scholars who have nothing to hide and no need to impress. They do groundbreaking work on important subjects, reason well about clear ideas, and are honest, down-to-earth people. Still, their writing stinks.”
- One senior editor of Nature bluntly stated “most papers are badly written”

Pinker, S. (2014). Why academics stink at writing. The chronicle of higher education, 61(5), 2-9.

Sage L (2003) Writing a clear and engaging paper for all astronomers. In Astronomy Communication (eds Heck A, Madsen C), pp 221–226. Dordrecht, the Netherlands: Kluwer



2025 ANNUAL SYMPOSIUM

#AMIA2025

Why do we need AI to assist paper writing?

- Generative AI tools have the potential to accelerate the research process by carrying out straightforward but time-consuming tasks — such as writing summaries and polish articles
- Generative AI tools have already been used to polish academic writing
 - Up to 17.5% of computer science papers on the arXiv have LLMs usage
- Title/Abstract generation can also provide research inspirations for scientists

HOW AN AI CHATBOT EDITS A MANUSCRIPT

Computational biologists Casey Greene and Milton Pividori constructed a program to get GPT-3 to edit research papers. The AI receives sections of a manuscript in turn, each with their own editing instructions (or 'prompts'). The prompts are adapted — in automated fashion — to suit the needs of each section, such as the methods or conclusion. In this example from their work, GPT-3 was fed an introduction section and the prompt below.

Prompt:

"Revise the following paragraph from the introduction of an academic paper (with the title 'An efficient not-only-linear correlation coefficient based on machine learning' and keywords 'correlation coefficient, nonlinear relationships, gene expression') so most of the citations to other academic papers are kept, the text minimizes the use of jargon, the text grammar is correct, spelling errors are fixed, and the text is in active voice and has a clear sentence structure."

Paragraph:

New technologies have vastly improved data collection, generating a deluge of information across different disciplines. This large amount of data provides new opportunities to address unanswered scientific questions, provided we have efficient tools capable of identifying multiple types of underlying patterns. Correlation analysis is an essential statistical technique for discovering relationships between variables [1]. Correlation coefficients are often used in exploratory data mining techniques, such as clustering or community detection algorithms, to compute a similarity value between a pair of objects of interest such as genes [2] or disease-relevant lifestyle factors [3]. Correlation methods are also used in supervised tasks, for example, for feature selection to improve prediction accuracy [4,5]. The Pearson correlation coefficient is ubiquitously deployed across application domains and diverse scientific areas. Thus, even minor and significant improvements in these techniques could have enormous consequences in industry and research.

GPT-3 edits:

New technologies have vastly improved data collection, generating a deluge of information across different disciplines. The increasing availability of data has opened up new possibilities for scientific exploration. This large amount of data provides new opportunities to address unanswered scientific questions, provided we have efficient tools capable of identifying multiple types of underlying patterns. Correlation analysis is an essential a useful statistical technique for discovering to uncover such relationships between variables [1]. Correlation coefficients are often used in exploratory data mining techniques, such as clustering or community detection algorithms, to compute a calculate the similarity value between a pair of two objects of interest such as , like genes [2] or disease-relevant lifestyle factors related to diseases [3]. Correlation methods They are also used in supervised tasks, for example, for like feature selection, to improve boost prediction accuracy [4,5]. The Pearson correlation coefficient is ubiquitously deployed widely used across many application domains and diverse scientific areas disciplines. Thus Therefore , even minor and significant small improvements in these techniques could this technique can have enormous consequences in a huge impact on industry and research.

Pinker, S. (2014). Why academics stink at writing. The chronicle of higher education, 61(5), 2-9.

<https://www.nature.com/articles/d41586-023-00340-6>

<https://www.nature.com/articles/d41586-024-01042-3>

. Liang, W., Zhang, Y., Wu, Z., Lepp, H., Ji, W., Zhao, X., ... & Zou, J. Y. (2024). Mapping the increasing use of llms in scientific papers. arXiv preprint arXiv:2404.01268.



2025 ANNUAL SYMPOSIUM

#AMIA2025

Task Formulation

- Paper components that benefit from existing systems

Title

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Lukasz Kaiser*
Google Brain
lukaszkaier@google.com

Illia Polosukhin* ‡
illia.polosukhin@gmail.com

Abstract

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

Definition

In this work we propose the Transformer, a model architecture eschewing recurrence and instead relying entirely on an attention mechanism to draw global dependencies between input and output. The Transformer allows for significantly more parallelization and can reach a new state of the art in translation quality after being trained for as little as twelve hours on eight P100 GPUs.

2 Background

The goal of reducing sequential computation also forms the foundation of the Extended Neural GPU [16], ByteNet [18] and ConvS2S [9], all of which use convolutional neural networks as basic building block, computing hidden representations in parallel for all input and output positions. In these models, the number of operations required to relate signals from two arbitrary input or output positions grows in the distance between positions, linearly for ConvS2S and logarithmically for ByteNet. This makes it more difficult to learn dependencies between distant positions [12]. In the Transformer this is reduced to a constant number of operations, albeit at the cost of reduced effective resolution due to averaging attention-weighted positions, an effect we counteract with Multi-Head Attention as described in section 3.2.

Self-attention, sometimes called intra-attention is an attention mechanism relating different positions of a single sequence in order to compute a representation of the sequence. Self-attention has been used successfully in a variety of tasks including reading comprehension, abstractive summarization, textual entailment and learning task-independent sentence representations [4, 27, 28, 22].

End-to-end memory networks are based on a recurrent attention mechanism instead of sequence-aligned recurrence and have been shown to perform well on simple-language question answering and language modeling tasks [34].

To the best of our knowledge, however, the Transformer is the first transduction model relying entirely on self-attention to compute representations of its input and output without using sequence-aligned RNNs or convolution. In the following sections, we will describe the Transformer, motivate self-attention and discuss its advantages over models such as [17, 18] and [9].

Related Work

Table 4: The Transformer generalizes well to English constituency parsing (Results are on Section 23 of WSJ)

Parser	Training	WSJ 23 F1
Vinyals & Kaiser et al. (2014) [37]	WSJ only, discriminative	88.3
Petrov et al. (2006) [29]	WSJ only, discriminative	90.4
Zhu et al. (2013) [40]	WSJ only, discriminative	90.4
Dyer et al. (2016) [8]	WSJ only, discriminative	91.7
Transformer (4 layers)	WSJ only, discriminative	91.3
Zhu et al. (2013) [40]	semi-supervised	91.3
Huang & Harper (2009) [14]	semi-supervised	91.3
McClosky et al. (2006) [26]	semi-supervised	92.1
Vinyals & Kaiser et al. (2014) [37]	semi-supervised	92.1
Transformer (4 layers)	semi-supervised	92.7
Luong et al. (2015) [23]	multi-task	93.0
Dyer et al. (2016) [8]	generative	93.3

increased the maximum output length to input length + 300. We used a beam size of 21 and $\alpha = 0.3$ for both WSJ only and the semi-supervised setting.

Our results in Table 4 show that despite the lack of task-specific tuning our model performs surprisingly well, yielding better results than all previously reported models with the exception of the Recurrent Neural Network Grammar [8].

In contrast to RNN sequence-to-sequence models [37], the Transformer outperforms the Berkeley-Parser [29] even when training only on the WSJ training set of 40K sentences.

7 Conclusion

In this work, we presented the Transformer, the first sequence transduction model based entirely on attention, replacing the recurrent layers most commonly used in encoder-decoder architectures with multi-headed self-attention.

For translation tasks, the Transformer can be trained significantly faster than architectures based on recurrent or convolutional layers. On both WMT 2014 English-to-German and WMT 2014 English-to-French translation tasks, we achieve a new state of the art. In the former task our best model outperforms even all previously reported ensembles.

We are excited about the future of attention-based models and plan to apply them to other tasks. We plan to extend the Transformer to problems involving input and output modalities other than text and to investigate local, restricted attention mechanisms to efficiently handle large inputs and outputs such as images, audio and video. Making generation less sequential is another research goal of ours.

The code we used to train and evaluate our models is available at <https://github.com/tensorflow/tensor2tensor>.

Acknowledgements We are grateful to Nal Kalchbrenner and Stephan Gouws for their fruitful comments, corrections and inspiration.

Table
Description
Claim
Conclusion



2025 ANNUAL SYMPOSIUM

#AMIA2025

Task Formulation

- Paper components that benefit from existing systems

Title

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Lukasz Kaiser*
Google Brain
lukaszkaier@google.com

Illia Polosukhin* ‡
illia.polosukhin@gmail.com

Abstract

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

Definition

In this work we propose the Transformer, a model architecture eschewing recurrence and instead relying entirely on an attention mechanism to draw global dependencies between input and output. The Transformer allows for significantly more parallelization and can reach a new state of the art in translation quality after being trained for as little as twelve hours on eight P100 GPUs.

2 Background

The goal of reducing sequential computation also forms the foundation of the Extended Neural GPU [16], ByteNet [18] and ConvS2S [9], all of which use convolutional neural networks as basic building block, computing hidden representations in parallel for all input and output positions. In these models, the number of operations required to relate signals from two arbitrary input or output positions grows in the distance between positions, linearly for ConvS2S and logarithmically for ByteNet. This makes it more difficult to learn dependencies between distant positions [12]. In the Transformer this is reduced to a constant number of operations, albeit at the cost of reduced effective resolution due to averaging attention-weighted positions, an effect we counteract with Multi-Head Attention as described in section 3.2.

Self-attention, sometimes called intra-attention is an attention mechanism relating different positions of a single sequence in order to compute a representation of the sequence. Self-attention has been used successfully in a variety of tasks including reading comprehension, abstractive summarization, textual entailment and learning task-independent sentence representations [4, 27, 28, 22].

End-to-end memory networks are based on a recurrent attention mechanism instead of sequence-aligned recurrence and have been shown to perform well on simple-language question answering and language modeling tasks [34].

To the best of our knowledge, however, the Transformer is the first transduction model relying entirely on self-attention to compute representations of its input and output without using sequence-aligned RNNs or convolution. In the following sections, we will describe the Transformer, motivate self-attention and discuss its advantages over models such as [17, 18] and [9].

Related Work

Table 4: The Transformer generalizes well to English constituency parsing (Results are on Section 23 of WSJ)

Parser	Training	WSJ 23 F1
Vinyals & Kaiser et al. (2014) [37]	WSJ only, discriminative	88.3
Petrov et al. (2006) [29]	WSJ only, discriminative	90.4
Zhu et al. (2013) [40]	WSJ only, discriminative	90.4
Dyer et al. (2016) [8]	WSJ only, discriminative	91.7
Transformer (4 layers)	WSJ only, discriminative	91.3
Zhu et al. (2013) [40]	semi-supervised	91.3
Huang & Harper (2009) [14]	semi-supervised	91.3
McClosky et al. (2006) [26]	semi-supervised	92.1
Vinyals & Kaiser et al. (2014) [37]	semi-supervised	92.1
Transformer (4 layers)	semi-supervised	92.7
Luong et al. (2015) [23]	multi-task	93.0
Dyer et al. (2016) [8]	generative	93.3

increased the maximum output length to input length + 300. We used a beam size of 21 and $\alpha = 0.3$ for both WSJ only and the semi-supervised setting.

Our results in Table 4 show that despite the lack of task-specific tuning our model performs surprisingly well, yielding better results than all previously reported models with the exception of the Recurrent Neural Network Grammar [8].

In contrast to RNN sequence-to-sequence models [37], the Transformer outperforms the Berkeley-Parser [29] even when training only on the WSJ training set of 40K sentences.

7 Conclusion

In this work, we presented the Transformer, the first sequence transduction model based entirely on attention, replacing the recurrent layers most commonly used in encoder-decoder architectures with multi-headed self-attention.

For translation tasks, the Transformer can be trained significantly faster than architectures based on recurrent or convolutional layers. On both WMT 2014 English-to-German and WMT 2014 English-to-French translation tasks, we achieve a new state of the art. In the former task our best model outperforms even all previously reported ensembles.

We are excited about the future of attention-based models and plan to apply them to other tasks. We plan to extend the Transformer to problems involving input and output modalities other than text and to investigate local, restricted attention mechanisms to efficiently handle large inputs and outputs such as images, audio and video. Making generation less sequential is another research goal of ours.

The code we used to train and evaluate our models is available at <https://github.com/tensorflow/tensor2tensor>.

Acknowledgements We are grateful to Nal Kalchbrenner and Stephan Gouws for their fruitful comments, corrections and inspiration.

Table
Description
Claim
Conclusion



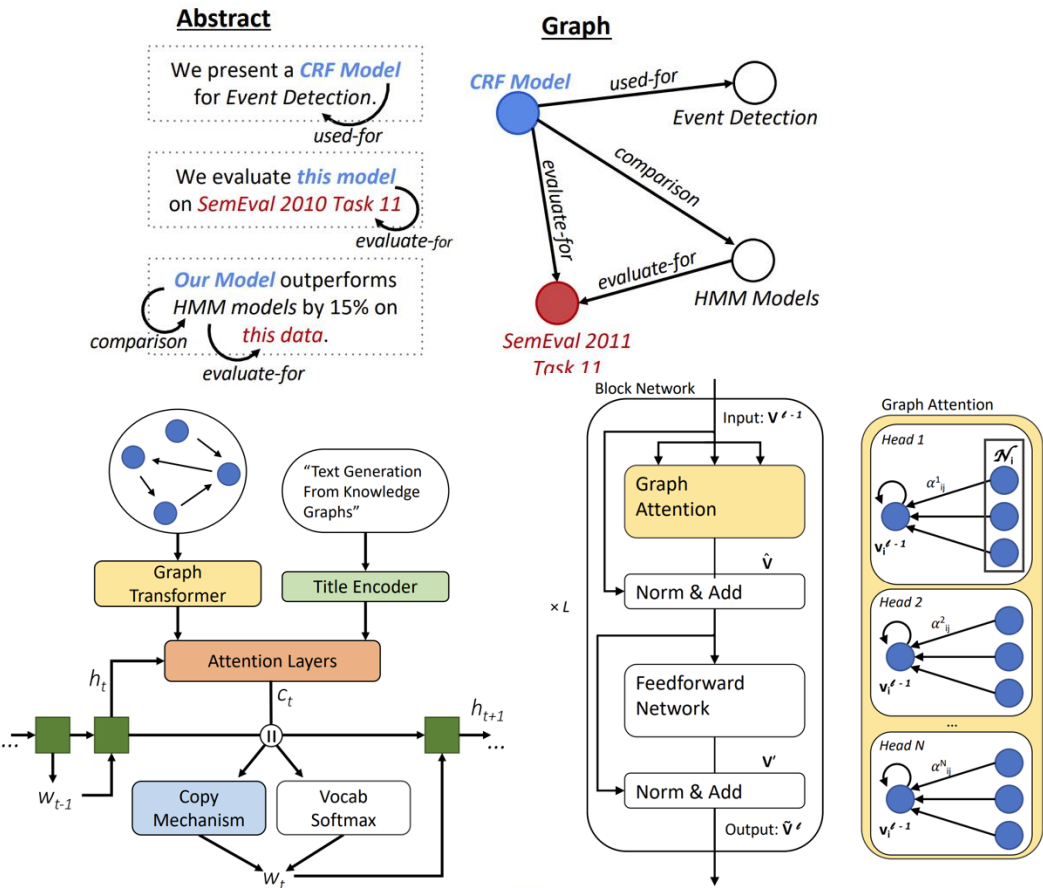
2025 ANNUAL SYMPOSIUM

#AMIA2025

Paper Abstract Generation

- Generating paper abstracts based on titles and knowledge graphs
- Use graph attention layer for graph transformer

Title: Event Detection with Conditional Random Fields



What if knowledge graphs are not given?

Pros:

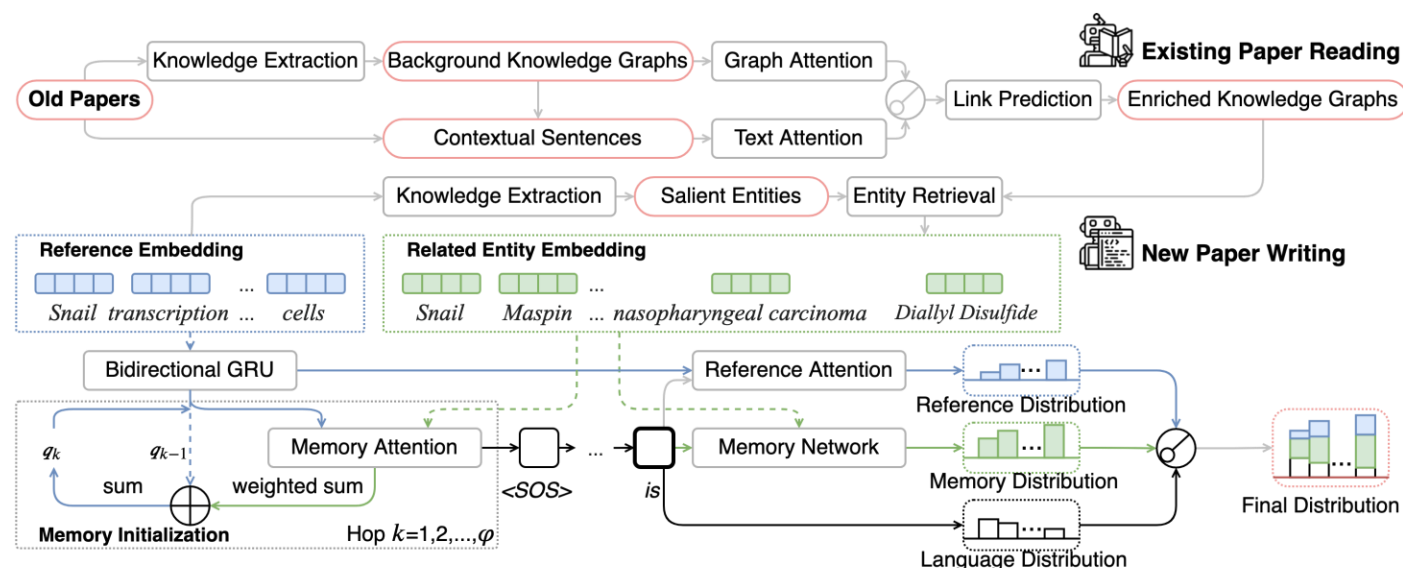
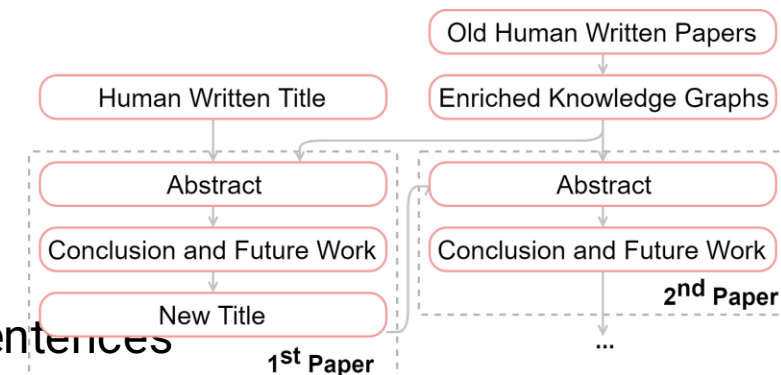
- Construct an AGENDA dataset for abstract generation task
- Convert disconnected labeled graph to connected unconnected graph for graph encoding

Points for Improvement:

- Lack cross attention between the input knowledge graph and the title during each layer of the graph transformer block

PaperRobot Incremental Writing

- Existing Paper Reading
 - Construct KB based on existing papers with PubTator
- Create New Ideas
 - Predict new links by leveraging graph features and contextual sentences
- Write a New Paper Draft about New Ideas
 - Employ a memory network to capture multi-hop attention of related entities and utilize reference attention to capture reference text



Pros:

- Designs a pipeline for a paper drafting system capable of reading literature, proposing new ideas, and drafting papers
- Use memory network to filter out irrelevant entities

Points for Improvement:

- Fail to fully utilize the graph structures of the knowledge graph when generating drafts
- The link predictions and paper draft generation are not optimized in an end-to-end way, leading to error propagation

PaperRobot Incremental Writing

Title	<i>Snail transcription factor negatively regulates maspin tumor suppressor in human prostate cancer cells</i>	
Entities	Related: <i>nasopharyngeal carcinoma</i> ; diallyl disulfide	
Output	Human (Neal et al., 2012)	System
Abstract	<i>Background: Maspin, a putative tumor suppressor that is down-regulated in breast and prostate cancer, has been associated with decreased cell motility. Snail transcription factor is a zinc finger protein that is increased in breast cancer and is associated with increased tumor motility and invasion by induction of epithelial-mesenchymal transition (EMT). We investigated the molecular mechanisms by which Snail increases tumor motility and invasion utilizing prostate cancer cells. Methods: Expression levels were analyzed by RT-PCR and western blot analyses. Cell motility and invasion assays were performed, while Snail regulation and binding to maspin promoter was analyzed by luciferase reporter and chromatin immunoprecipitation (ChIP) assays. Results: Snail protein expression was higher in different prostate cancer cell lines as compared to normal prostate epithelial cells.</i>	<i>Background: Snail is a multifunctional protein that plays an important role in the pathogenesis of prostate cancer. However, it has been shown to be associated with poor prognosis. The purpose of this study was to investigate the effect of negatively on the expression of maspin in human nasopharyngeal carcinoma cell lines. Methods: Quantitative real-time PCR and western blot analysis were used to determine whether the demethylating agent was investigated by quantitative RT-PCR (qRT-PCR) and Western blotting. Results showed that the binding protein plays a significant role in the regulation of tumor growth and progression.</i>
Conclusion and Future work	<i>Collectively, our results indicate for the first time that Snail can negatively regulate maspin through direct promoter repression resulting in increased migration and invasion in prostate cancer cells. This study reveals a novel mechanism of how Snail may function and show the importance of therapeutic targeting of Snail signaling in future.</i>	<i>In summary, our study demonstrates that Snail negatively inhibited the expression of Maspin in human nasopharyngeal carcinoma cell lines and in vitro. Our results indicate that the combination of the demethylating agent might be a potential therapeutic target for the treatment of prostate cancer.</i>
New Title	<i>Role of maspin in cancer (Berardi et al., 2013)</i>	<i>The role of nasopharyngeal carcinoma in the rat model of prostate cancer cells</i>

Wang, Q., Huang, L., Jiang, Z., Knight, K., Ji, H., Bansal, M., & Luan, Y. (2019). PaperRobot: Incremental draft generation of scientific ideas. ACL 2019.



2025 ANNUAL SYMPOSIUM

#AMIA2025

Task Formulation

- Paper components that benefit from existing systems

Title

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Lukasz Kaiser*
Google Brain
lukaszkaier@google.com

Illia Polosukhin* ‡
illia.polosukhin@gmail.com

Abstract

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

Definition

In this work we propose the Transformer, a model architecture eschewing recurrence and instead relying entirely on an attention mechanism to draw global dependencies between input and output. The Transformer allows for significantly more parallelization and can reach a new state of the art in translation quality after being trained for as little as twelve hours on eight P100 GPUs.

2 Background

The goal of reducing sequential computation also forms the foundation of the Extended Neural GPU [16], ByteNet [18] and ConvS2S [9], all of which use convolutional neural networks as basic building block, computing hidden representations in parallel for all input and output positions. In these models, the number of operations required to relate signals from two arbitrary input or output positions grows in the distance between positions, linearly for ConvS2S and logarithmically for ByteNet. This makes it more difficult to learn dependencies between distant positions [12]. In the Transformer this is reduced to a constant number of operations, albeit at the cost of reduced effective resolution due to averaging attention-weighted positions, an effect we counteract with Multi-Head Attention as described in section 3.2.

Self-attention, sometimes called intra-attention is an attention mechanism relating different positions of a single sequence in order to compute a representation of the sequence. Self-attention has been used successfully in a variety of tasks including reading comprehension, abstractive summarization, textual entailment and learning task-independent sentence representations [4, 27, 28, 22].

End-to-end memory networks are based on a recurrent attention mechanism instead of sequence-aligned recurrence and have been shown to perform well on simple-language question answering and language modeling tasks [34].

To the best of our knowledge, however, the Transformer is the first transduction model relying entirely on self-attention to compute representations of its input and output without using sequence-aligned RNNs or convolution. In the following sections, we will describe the Transformer, motivate self-attention and discuss its advantages over models such as [17, 18] and [9].

Related Work

Table 4: The Transformer generalizes well to English constituency parsing (Results are on Section 23 of WSJ)

Parser	Training	WSJ 23 F1
Vinyals & Kaiser et al. (2014) [37]	WSJ only, discriminative	88.3
Petrov et al. (2006) [29]	WSJ only, discriminative	90.4
Zhu et al. (2013) [40]	WSJ only, discriminative	90.4
Dyer et al. (2016) [8]	WSJ only, discriminative	91.7
Transformer (4 layers)	WSJ only, discriminative	91.3
Zhu et al. (2013) [40]	semi-supervised	91.3
Huang & Harper (2009) [14]	semi-supervised	91.3
McClosky et al. (2006) [26]	semi-supervised	92.1
Vinyals & Kaiser et al. (2014) [37]	semi-supervised	92.1
Transformer (4 layers)	semi-supervised	92.7
Luong et al. (2015) [23]	multi-task	93.0
Dyer et al. (2016) [8]	generative	93.3

increased the maximum output length to input length + 300. We used a beam size of 21 and $\alpha = 0.3$ for both WSJ only and the semi-supervised setting.

Our results in Table 4 show that despite the lack of task-specific tuning our model performs surprisingly well, yielding better results than all previously reported models with the exception of the Recurrent Neural Network Grammar [8].

In contrast to RNN sequence-to-sequence models [37], the Transformer outperforms the Berkeley-Parser [29] even when training only on the WSJ training set of 40K sentences.

7 Conclusion

In this work, we presented the Transformer, the first sequence transduction model based entirely on attention, replacing the recurrent layers most commonly used in encoder-decoder architectures with multi-headed self-attention.

For translation tasks, the Transformer can be trained significantly faster than architectures based on recurrent or convolutional layers. On both WMT 2014 English-to-German and WMT 2014 English-to-French translation tasks, we achieve a new state of the art. In the former task our best model outperforms even all previously reported ensembles.

We are excited about the future of attention-based models and plan to apply them to other tasks. We plan to extend the Transformer to problems involving input and output modalities other than text and to investigate local, restricted attention mechanisms to efficiently handle large inputs and outputs such as images, audio and video. Making generation less sequential is another research goal of ours.

The code we used to train and evaluate our models is available at <https://github.com/tensorflow/tensor2tensor>.

Acknowledgements We are grateful to Nal Kalchbrenner and Stephan Gouws for their fruitful comments, corrections and inspiration.

Description

Claim

Conclusion



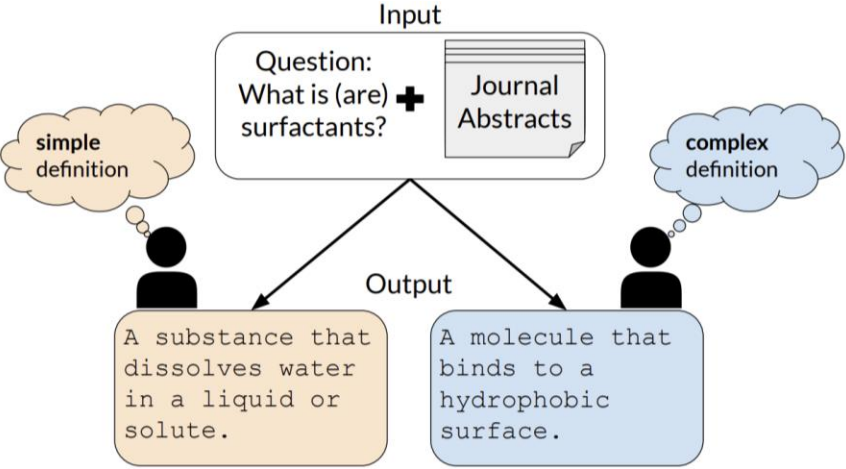
2025 ANNUAL SYMPOSIUM

#AMIA2025

Scientific Definition Generation

- Generating definitions of scientific and medical terms with varying complexity
- Dataset: Medical Consumer Questions (Abacha et al., 2019), Wikipedia, S2ORC
- Method
 - Reranking BART generation results based on logits from a discriminator trained to distinguish scientific journal text from science news text (measuring complexity)

Control Method	Direction	
	Low (News)	High (Journal)
SVM-Rerank	A type of computing in which there are many computers running at the same time in different parts of the world.	In computer science, distributed computing is the process of computing on a large scale without a single centralized data center.
BERT-Rerank	A type of computer system in which there are more than a few computers working together.	In computer science, distributed computing is the process of computing on a large scale without a single centralized data center.
GeDi	Is the implementation of computer programs across multiple computers on similar hardware and/or software resources.	In computer science, a concept that states that data must be shared across computing resources.
DExpert	An Internet-driven by-computing that portion of different computers from start to finish.	In computer science and communication between-Consequently-integrates.
PPLM	Easeless, self-organized, and often self-organizing networked computer systems intended for the purposes of optimization.	Multi-purpose, distributed system software with or without a single datum storage system.



Points for Improvement:

- Instead of reranking, inject complexity as an additional input feature in the input
- Employ a generative adversarial framework to optimize the BART generator and reranking discriminator jointly

Task Formulation

- Writing each part of a paper is a different task!

Title

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez*[†]
University of Toronto
aidan@cs.toronto.edu

Lukasz Kaiser*
Google Brain
lukaszkaier@google.com

Illia Polosukhin*[‡]
illia.polosukhin@gmail.com

Abstract

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

Definition

In this work we propose the Transformer, a model architecture eschewing recurrence and instead relying entirely on an attention mechanism to draw global dependencies between input and output. The Transformer allows for significantly more parameterization and can reach a new state of the art in translation quality after being trained for as little as twelve hours on eight P100 GPUs.

2 Background

The goal of reducing sequential computation also forms the foundation of the Extended Neural GPU [16], ByteNet [18] and ConvS2S [9], all of which use convolutional neural networks as basic building block, computing hidden representations in parallel for all input and output positions. In these models, the number of operations required to relate signals from two arbitrary input or output positions grows in the distance between positions, linearly for ConvS2S and logarithmically for ByteNet. This makes it more difficult to learn dependencies between distant positions [12]. In the Transformer this is reduced to a constant number of operations, albeit at the cost of reduced effective resolution due to averaging attention-weighted positions, an effect we counteract with Multi-Head Attention as described in section 3.2.

Self-attention, sometimes called intra-attention is an attention mechanism relating different positions of a single sequence in order to compute a representation of the sequence. Self-attention has been used successfully in a variety of tasks including reading comprehension, abstractive summarization, textual entailment and learning task-independent sentence representations [4, 27, 28, 22].

End-to-end memory networks are based on a recurrent attention mechanism instead of sequence-aligned recurrence and have been shown to perform well on simple-language question answering and language modeling tasks [34].

To the best of our knowledge, however, the Transformer is the first transduction model relying entirely on self-attention to compute representations of its input and output without using sequence-aligned RNNs or convolution. In the following sections, we will describe the Transformer, motivate self-attention and discuss its advantages over models such as [17, 18] and [9].

Related Work

Table 4: The Transformer generalizes well to English constituency parsing (Results are on Section 23 of WSJ)

Parser	Training	WSJ 23 F1
Vinyals & Kaiser et al. (2014) [37]	WSJ only, discriminative	88.3
Petrov et al. (2006) [29]	WSJ only, discriminative	90.4
Zhu et al. (2013) [40]	WSJ only, discriminative	90.4
Dyer et al. (2016) [8]	WSJ only, discriminative	91.7
Transformer (4 layers)	WSJ only, discriminative	91.3
Zhu et al. (2013) [40]	semi-supervised	91.3
Huang & Harper (2009) [14]	semi-supervised	91.3
McClosky et al. (2006) [26]	semi-supervised	92.1
Vinyals & Kaiser et al. (2014) [37]	semi-supervised	92.1
Transformer (4 layers)	semi-supervised	92.7
Luong et al. (2015) [23]	multi-task	93.0
Dyer et al. (2016) [8]	generative	93.3

increased the maximum output length to input length + 300. We used a beam size of 21 and $\alpha = 0.3$ for both WSJ only and the semi-supervised setting.

Our results in Table 4 show that despite the lack of task-specific tuning our model performs surprisingly well, yielding better results than all previously reported models with the exception of the Recurrent Neural Network Grammar [8].

In contrast to RNN sequence-to-sequence models [37], the Transformer outperforms the Berkeley-Parser [29] even when training only on the WSJ training set of 40K sentences.

7 Conclusion

In this work, we presented the Transformer, the first sequence transduction model based entirely on attention, replacing the recurrent layers most commonly used in encoder-decoder architectures with multi-headed self-attention.

For translation tasks, the Transformer can be trained significantly faster than architectures based on recurrent or convolutional layers. On both WMT 2014 English-to-German and WMT 2014 English-to-French translation tasks, we achieve a new state of the art. In the former task our best model outperforms even all previously reported ensembles.

We are excited about the future of attention-based models and plan to apply them to other tasks. We plan to extend the Transformer to problems involving input and output modalities other than text and to investigate local, restricted attention mechanisms to efficiently handle large inputs and outputs such as images, audio and video. Making generation less sequential is another research goal of ours.

The code we used to train and evaluate our models is available at <https://github.com/tensorflow/tensor2tensor>.

Acknowledgements We are grateful to Nal Kalchbrenner and Stephan Gouws for their fruitful comments, corrections and inspiration.

Description

Claim

Conclusion

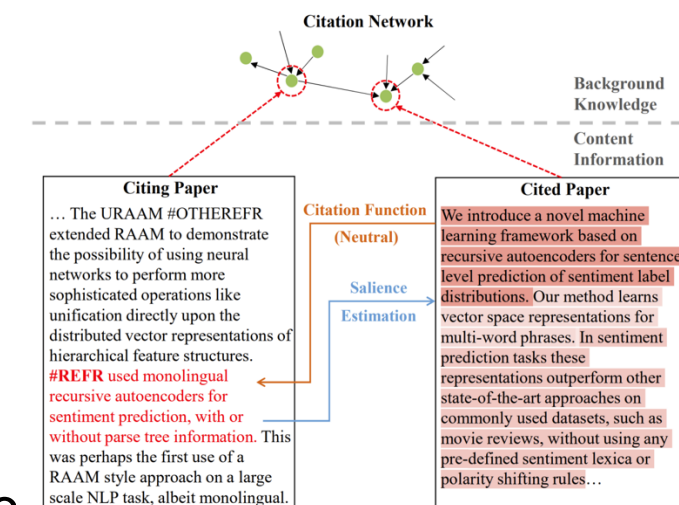


2025 ANNUAL SYMPOSIUM

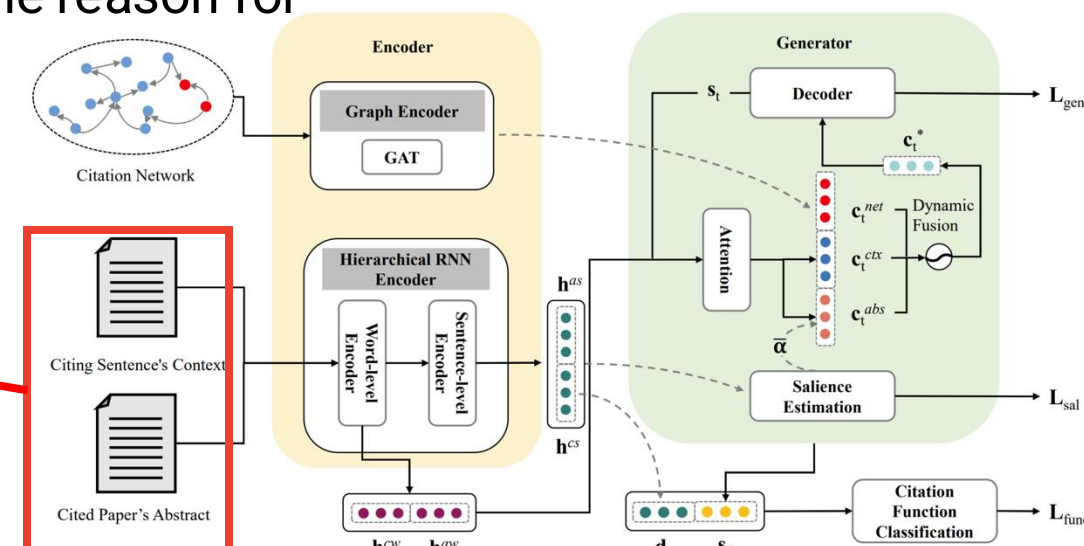
#AMIA2025

Knowledge-guided Citation Generation

- Input
 - Background knowledge: structural information from a citation network
 - Content: In-depth content information
- Target: Citation sentences
- Model Highlights
 - Saliency Estimation to identify what information needed to be concentrated for the generation (what to cite)
 - Citation Function Classification to classify the reason for citation (why to cite)

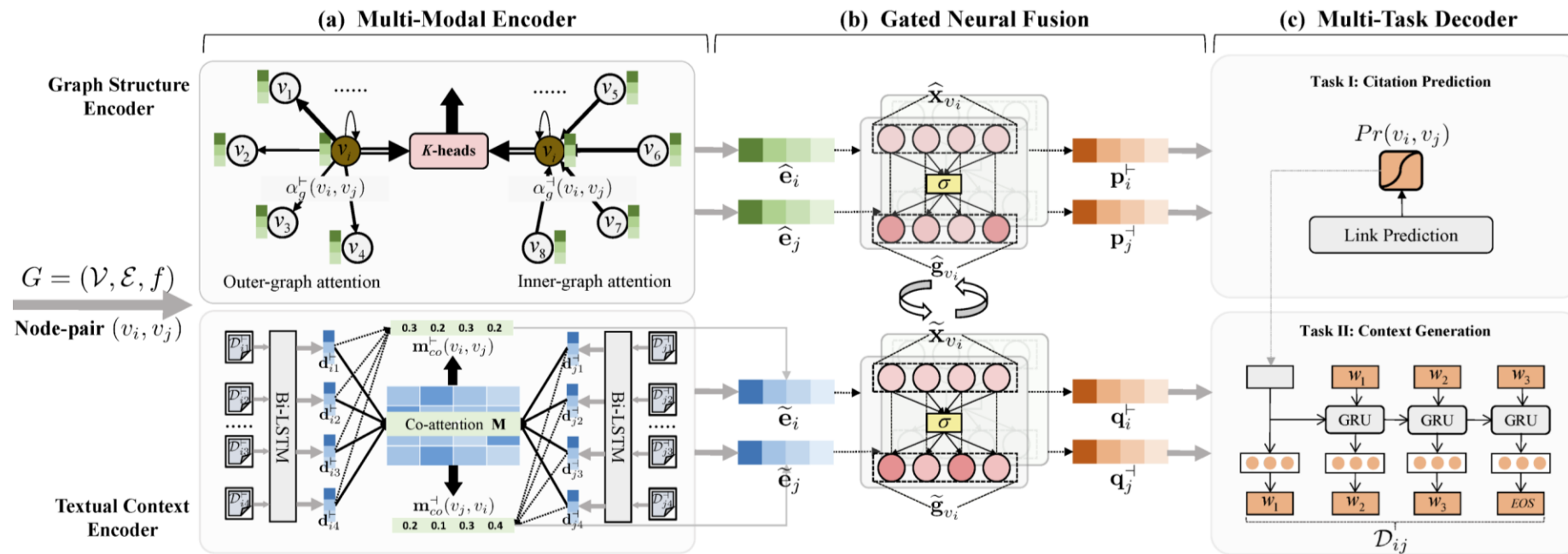


What if citation relations are not given?



Context-aware Citation Prediction

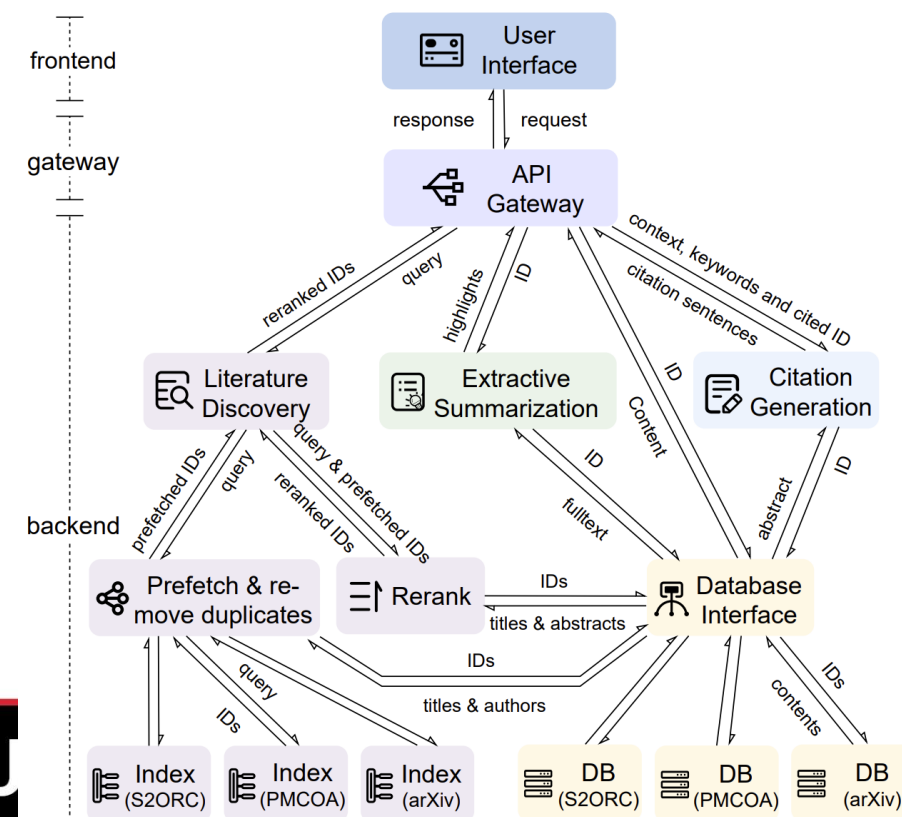
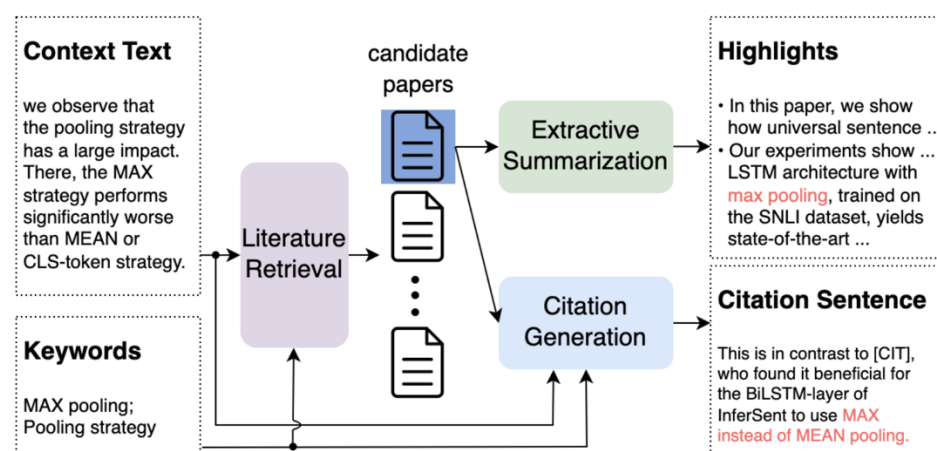
- Model
 - Graph Structure Encoder with dual-role attention to capture directed graph structure information
 - Gated Neural Fusion to integrate multi-modal representations and control features transfer for downstream tasks differently
 - Multi-Task Decoder to perform citation link prediction and context generation



Can we combine information retrieval and citation recommendation?

Joint Scientific Literature Discovery, Summarization and Citation Generation

- A pipeline that automatically recommends relevant papers, extracts highlights, and suggests a reference sentence as a citation of a paper, taking into consideration the user-provided context and keywords
 - Prefetching papers from database and reranking them based on SciBERT
 - Iteratively score and select sentences for summarization with MemSum
 - Concatenate keywords, context and target paper's abstract to generate citation sentences by T5



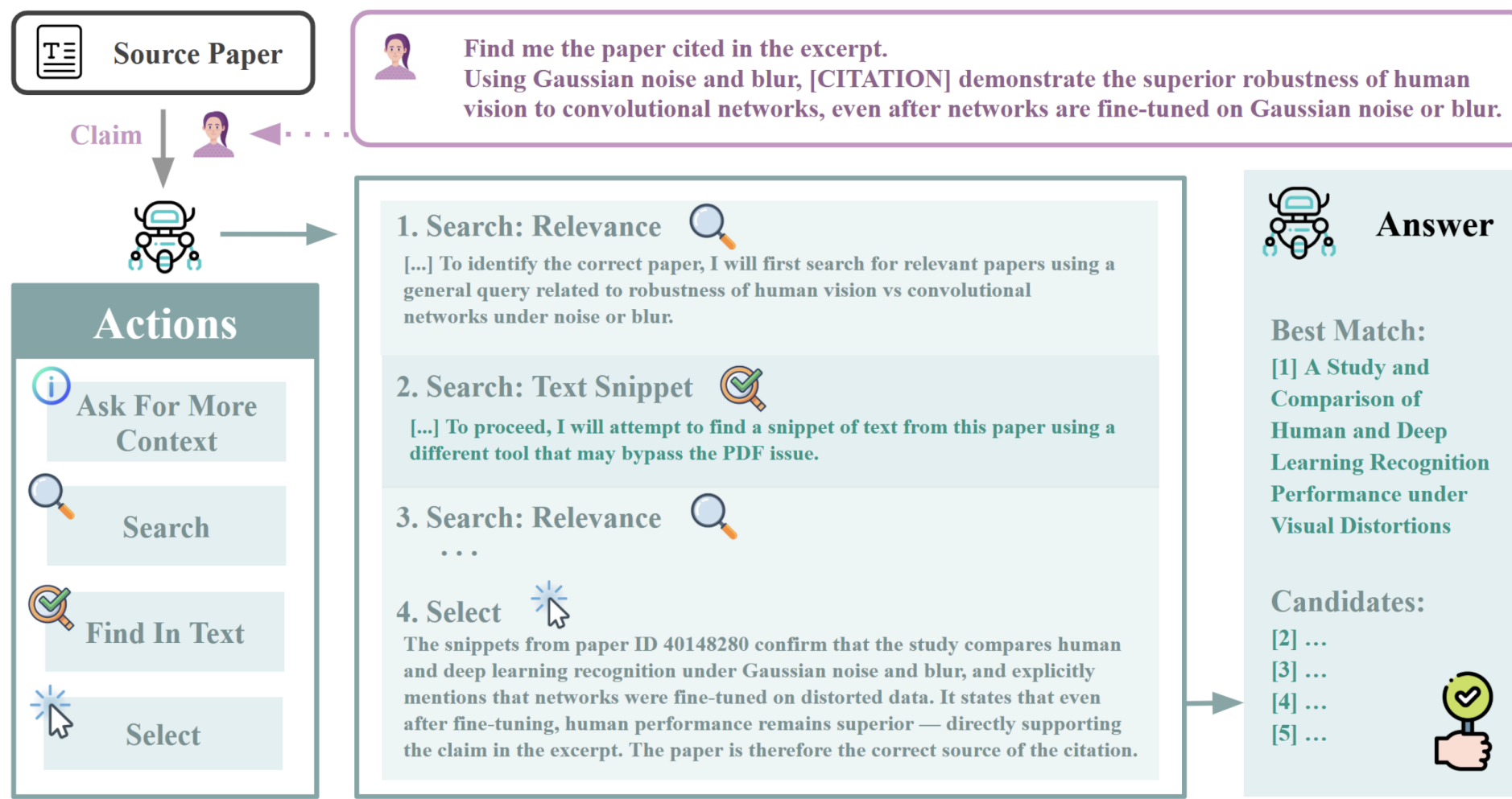
Summary

Method	Pros	Cons
BACO (Ge et al., 2021)	▪ Construct a BACO dataset specifically for citation generation task ▪ The framework identifies the cited sentence and the reason for citing, jointly with the generation of citations	▪ Rely on provided citation relations to generate citation sentence
AutoCite(Wang et al., 2021)	▪ Propose a joint framework to predict citation links and generate citation sentences simultaneously ▪ Design a graph encoder to distinguish a paper cites others or it is cited by others	▪ Fail to generate different types of citation texts according to the contexts and positions in the target paper
SciLit (Gu et al., 2023)	▪ Propose a pipeline to recommends relevant papers, extract highlights, and suggest a reference sentence	▪ The citation generation ignores the graph structure of citation networks



CiteGuard: Faithful Citation Attribution

- A retrieval-aware agent framework to provide faithful grounding for citation validation
 - Can proactively query for additional claim context when the initial snippet is insufficient
 - Allow searching directly within paper contents, reducing reliance on PDF availability
 - Can suggest alternative citations



Task Formulation

- Paper components that benefit from existing systems

Title

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Lukasz Kaiser*
Google Brain
lukaszkaier@google.com

Illia Polosukhin* ‡
illia.polosukhin@gmail.com

Abstract

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

Definition

In this work we propose the Transformer, a model architecture eschewing recurrence and instead relying entirely on an attention mechanism to draw global dependencies between input and output. The transformer allows for significantly more parallelization and can reach a new state of the art in translation quality after being trained for as little as twelve hours on eight P100 GPUs.

2 Background

The goal of reducing sequential computation also forms the foundation of the Extended Neural GPU [16], ByteNet [18] and ConvS2S [9], all of which use convolutional neural networks as basic building block, computing hidden representations in parallel for all input and output positions. In these models, the number of operations required to relate signals from two arbitrary input or output positions grows in the distance between positions, linearly for ConvS2S and logarithmically for ByteNet. This makes it more difficult to learn dependencies between distant positions [12]. In the Transformer this is reduced to a constant number of operations, albeit at the cost of reduced effective resolution due to averaging attention-weighted positions, an effect we counteract with Multi-Head Attention as described in section 3.2.

Self-attention, sometimes called intra-attention is an attention mechanism relating different positions of a single sequence in order to compute a representation of the sequence. Self-attention has been used successfully in a variety of tasks including reading comprehension, abstractive summarization, textual entailment and learning task-independent sentence representations [4, 27, 28, 22].

End-to-end memory networks are based on a recurrent attention mechanism instead of sequence-aligned recurrence and have been shown to perform well on simple-language question answering and language modeling tasks [34].

To the best of our knowledge, however, the Transformer is the first transduction model relying entirely on self-attention to compute representations of its input and output without using sequence-aligned RNNs or convolution. In the following sections, we will describe the Transformer, motivate self-attention and discuss its advantages over models such as [17, 18] and [9].

Related Work

Table 4: The Transformer generalizes well to English constituency parsing (Results are on Section 23 of WSJ)

Parser	Training	WSJ 23 F1
Vinyals & Kaiser et al. (2014) [37]	WSJ only, discriminative	88.3
Petrov et al. (2006) [29]	WSJ only, discriminative	90.4
Zhu et al. (2013) [40]	WSJ only, discriminative	90.4
Dyer et al. (2016) [8]	WSJ only, discriminative	91.7
Transformer (4 layers)	WSJ only, discriminative	91.3
Zhu et al. (2013) [40]	semi-supervised	91.3
Huang & Harper (2009) [14]	semi-supervised	91.3
McClosky et al. (2006) [26]	semi-supervised	92.1
Vinyals & Kaiser et al. (2014) [37]	semi-supervised	92.1
Transformer (4 layers)	semi-supervised	92.7
Luong et al. (2015) [23]	multi-task	93.0
Dyer et al. (2016) [8]	generative	93.3

increased the maximum output length to input length + 300. We used a beam size of 21 and $\alpha = 0.3$ for both WSJ only and the semi-supervised setting.

Our results in Table 4 show that despite the lack of task-specific tuning our model performs surprisingly well, yielding better results than all previously reported models with the exception of the Recurrent Neural Network Grammar [8].

In contrast to RNN sequence-to-sequence models [37], the Transformer outperforms the Berkeley-Parser [29] even when training only on the WSJ training set of 40K sentences.

7 Conclusion

In this work, we presented the Transformer, the first sequence transduction model based entirely on attention, replacing the recurrent layers most commonly used in encoder-decoder architectures with multi-headed self-attention.

For translation tasks, the Transformer can be trained significantly faster than architectures based on recurrent or convolutional layers. On both WMT 2014 English-to-German and WMT 2014 English-to-French translation tasks, we achieve a new state of the art. In the former task our best model outperforms even all previously reported ensembles.

We are excited about the future of attention-based models and plan to apply them to other tasks. We plan to extend the Transformer to problems involving input and output modalities other than text and to investigate local, restricted attention mechanisms to efficiently handle large inputs and outputs such as images, audio and video. Making generation less sequential is another research goal of ours.

The code we used to train and evaluate our models is available at <https://github.com/tensorflow/tensor2tensor>.

Acknowledgements We are grateful to Nal Kalchbrenner and Stephan Gouws for their fruitful comments, corrections and inspiration.

Description

Claim

Conclusion



2025 ANNUAL SYMPOSIUM

#AMIA2025

Scientific Context-Aware Description Generation

- Input
 - A paper object (e.g., figure, table, algorithm, theorem)
 - Context (Paragraphs before the descriptions)
- Target
 - Descriptions
- Dataset

Input	#num	#avg_out_len	#cand
Table	136K	74.05	199.93
Figure (chart/bar)	155K	76.60	179.35
Algorithm	56K	67.94	227.65
Theorem	175K	65.00	192.44

Body text (Context)					
... languages : Telugu (te) and Turkish (tr)... Turkish (tr) vocabulary has been censored to contain no overlap with the Telugu... we evaluate these models using a recall@k metric defined as ...					
Table					
	Result	te+en	+ tr	% Change	
	Recall@1	17.0	17.6	+3.5%	
	Recall@10	23.9	25.0	+4.6%	
	Recall@20	26.3	27.7	+5.3%	
Generated description w/o context (table only)					
Table shows when te+en is replaced with tr , the effect of different change is very small, although the performance of tr method gets really strong .					
Generated description w/ context (body text + table)					
Table summarizes the recall@1 measures and the percentage of the incremental improvement across languages for both tasks. The average incremental improvement across languages is about 4% in these cases, despite there being no overlap between in Telugu and Turkish .					

Task Formulation

- Paper components that benefit from existing systems

Title

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Lukasz Kaiser*
Google Brain
lukaszkaier@google.com

Illia Polosukhin* ‡
illia.polosukhin@gmail.com

Abstract

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

Definition

In this work we propose the Transformer, a model architecture eschewing recurrence and instead relying entirely on an attention mechanism to draw global dependencies between input and output. The transformer allows for significantly more parallelization and can reach a new state of the art in translation quality after being trained for as little as twelve hours on eight P100 GPUs.

2 Background

The goal of reducing sequential computation also forms the foundation of the Extended Neural GPU [16], ByteNet [18] and ConvS2S [9], all of which use convolutional neural networks as basic building block, computing hidden representations in parallel for all input and output positions. In these models, the number of operations required to relate signals from two arbitrary input or output positions grows in the distance between positions, linearly for ConvS2S and logarithmically for ByteNet. This makes it more difficult to learn dependencies between distant positions [12]. In the Transformer this is reduced to a constant number of operations, albeit at the cost of reduced effective resolution due to averaging attention-weighted positions, an effect we counteract with Multi-Head Attention as described in section 3.2.

Self-attention, sometimes called intra-attention is an attention mechanism relating different positions of a single sequence in order to compute a representation of the sequence. Self-attention has been used successfully in a variety of tasks including reading comprehension, abstractive summarization, textual entailment and learning task-independent sentence representations [4, 27, 28, 22].

End-to-end memory networks are based on a recurrent attention mechanism instead of sequence-aligned recurrence and have been shown to perform well on simple-language question answering and language modeling tasks [34].

To the best of our knowledge, however, the Transformer is the first transduction model relying entirely on self-attention to compute representations of its input and output without using sequence-aligned RNNs or convolution. In the following sections, we will describe the Transformer, motivate self-attention and discuss its advantages over models such as [17, 18] and [9].

Related Work

Table 4: The Transformer generalizes well to English constituency parsing (Results are on Section 23 of WSJ)

Parser	Training	WSJ 23 F1
Vinyals & Kaiser et al. (2014) [37]	WSJ only, discriminative	88.3
Petrov et al. (2006) [29]	WSJ only, discriminative	90.4
Zhu et al. (2013) [40]	WSJ only, discriminative	90.4
Dyer et al. (2016) [8]	WSJ only, discriminative	91.7
Transformer (4 layers)	WSJ only, discriminative	91.3
Zhu et al. (2013) [40]	semi-supervised	91.3
Huang & Harper (2009) [14]	semi-supervised	91.3
McClosky et al. (2006) [26]	semi-supervised	92.1
Vinyals & Kaiser et al. (2014) [37]	semi-supervised	92.1
Transformer (4 layers)	semi-supervised	92.7
Luong et al. (2015) [23]	multi-task	93.0
Dyer et al. (2016) [8]	generative	93.3

increased the maximum output length to input length + 300. We used a beam size of 21 and $\alpha = 0.3$ for both WSJ only and the semi-supervised setting.

Our results in Table 4 show that despite the lack of task-specific tuning our model performs surprisingly well, yielding better results than all previously reported models with the exception of the Recurrent Neural Network Grammar [8].

In contrast to RNN sequence-to-sequence models [37], the Transformer outperforms the Berkeley-Parser [29] even when training only on the WSJ training set of 40K sentences.

7 Conclusion

In this work, we presented the Transformer, the first sequence transduction model based entirely on attention, replacing the recurrent layers most commonly used in encoder-decoder architectures with multi-headed self-attention.

For translation tasks, the Transformer can be trained significantly faster than architectures based on recurrent or convolutional layers. On both WMT 2014 English-to-German and WMT 2014 English-to-French translation tasks, we achieve a new state of the art. In the former task our best model outperforms even all previously reported ensembles.

We are excited about the future of attention-based models and plan to apply them to other tasks. We plan to extend the Transformer to problems involving input and output modalities other than text and to investigate local, restricted attention mechanisms to efficiently handle large inputs and outputs such as images, audio and video. Making generation less sequential is another research goal of ours.

The code we used to train and evaluate our models is available at <https://github.com/tensorflow/tensor2tensor>.

Acknowledgements We are grateful to Nal Kalchbrenner and Stephan Gouws for their fruitful comments, corrections and inspiration.

Description

Claim

Conclusion

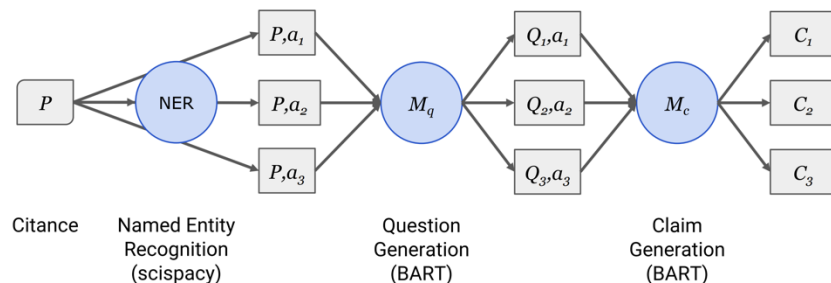


2025 ANNUAL SYMPOSIUM

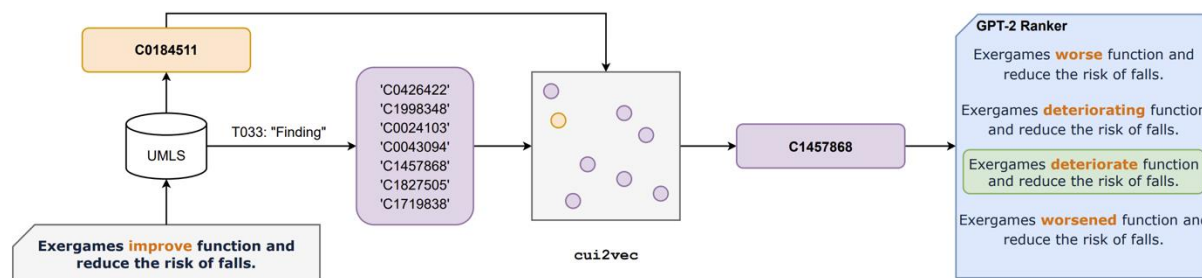
#AMIA2025

Scientific Claim Generation

- Supported Claims Generation
 - ClaimBART: Generating claim directly based on Citance
 - ClaimEntity



- Refuted Claims Generation
 - Knowledge-based Informed Negation



Points for Improvement:

- Instead of only considering entities in the claim, explore to use relations or even a subgraph of the knowledge graph when generating claims

(1) ALS is the most common adult motor neuron disease with an incidence of 2 per 100,000 and prevalence of 5.4 per 100,000 individuals. (2) Current treatment options are based on **symptom management** and **respiratory support** with the only approved medications in widespread use, **Riluzole** and **Edaravone**, providing only modest benefits and only in some patients.

- C_1 Current treatment options for ALS are based on **symptom management** and **respiratory support**
- C_2 **Riluzole** is an approved ALS medication in widespread use
- C_3 **Edaravone** is an approved ALS medication in widespread use
- C_4 **Riluzole** and **Edaravone** are the only approved ALS medications in widespread use
- C_5 **Riluzole** provides modest benefits in only some ALS patients
- C_6 **Edaravone** provides modest benefits in only some ALS patients

Task Formulation

- Can we generate paper end-to-end?

Title

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Lukasz Kaiser*
Google Brain
lukaszkaier@google.com

Illia Polosukhin* ‡
illia.polosukhin@gmail.com

Abstract

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

Definition

In this work we propose the Transformer, a model architecture eschewing recurrence and instead relying entirely on an attention mechanism to draw global dependencies between input and output. The Transformer allows for significantly more parallelization and can reach a new state of the art in translation quality after being trained for as little as twelve hours on eight P100 GPUs.

2 Background

The goal of reducing sequential computation also forms the foundation of the Extended Neural GPU [16], ByteNet [18] and ConvS2S [9], all of which use convolutional neural networks as basic building block, computing hidden representations in parallel for all input and output positions. In these models, the number of operations required to relate signals from two arbitrary input or output positions grows in the distance between positions, linearly for ConvS2S and logarithmically for ByteNet. This makes it more difficult to learn dependencies between distant positions [12]. In the Transformer this is reduced to a constant number of operations, albeit at the cost of reduced effective resolution due to averaging attention-weighted positions, an effect we counteract with Multi-Head Attention as described in section 3.2.

Self-attention, sometimes called intra-attention is an attention mechanism relating different positions of a single sequence in order to compute a representation of the sequence. Self-attention has been used successfully in a variety of tasks including reading comprehension, abstractive summarization, textual entailment and learning task-independent sentence representations [4, 27, 28, 22].

End-to-end memory networks are based on a recurrent attention mechanism instead of sequence-aligned recurrence and have been shown to perform well on simple-language question answering and language modeling tasks [34].

To the best of our knowledge, however, the Transformer is the first transduction model relying entirely on self-attention to compute representations of its input and output without using sequence-aligned RNNs or convolution. In the following sections, we will describe the Transformer, motivate self-attention and discuss its advantages over models such as [17, 18] and [9].

Related Work

Table 4: The Transformer generalizes well to English constituency parsing (Results are on Section 23 of WSJ)

Parser	Training	WSJ 23 F1
Vinyals & Kaiser et al. (2014) [37]	WSJ only, discriminative	88.3
Petrov et al. (2006) [29]	WSJ only, discriminative	90.4
Zhu et al. (2013) [40]	WSJ only, discriminative	90.4
Dyer et al. (2016) [8]	WSJ only, discriminative	91.7
Transformer (4 layers)	WSJ only, discriminative	91.3
Zhu et al. (2013) [40]	semi-supervised	91.3
Huang & Harper (2009) [14]	semi-supervised	91.3
McClosky et al. (2006) [26]	semi-supervised	92.1
Vinyals & Kaiser et al. (2014) [37]	semi-supervised	92.1
Transformer (4 layers)	semi-supervised	92.7
Luong et al. (2015) [23]	multi-task	93.0
Dyer et al. (2016) [8]	generative	93.3

increased the maximum output length to input length + 300. We used a beam size of 21 and $\alpha = 0.3$ for both WSJ only and the semi-supervised setting.

Our results in Table 4 show that despite the lack of task-specific tuning our model performs surprisingly well, yielding better results than all previously reported models with the exception of the Recurrent Neural Network Grammar [8].

In contrast to RNN sequence-to-sequence models [37], the Transformer outperforms the Berkeley-Parser [29] even when training only on the WSJ training set of 40K sentences.

7 Conclusion

In this work, we presented the Transformer, the first sequence transduction model based entirely on attention, replacing the recurrent layers most commonly used in encoder-decoder architectures with multi-headed self-attention.

For translation tasks, the Transformer can be trained significantly faster than architectures based on recurrent or convolutional layers. On both WMT 2014 English-to-German and WMT 2014 English-to-French translation tasks, we achieve a new state of the art. In the former task our best model outperforms even all previously reported ensembles.

We are excited about the future of attention-based models and plan to apply them to other tasks. We plan to extend the Transformer to problems involving input and output modalities other than text and to investigate local, restricted attention mechanisms to efficiently handle large inputs and outputs such as images, audio and video. Making generation less sequential is another research goal of ours.

The code we used to train and evaluate our models is available at <https://github.com/tensorflow/tensor2tensor>.

Acknowledgements We are grateful to Nal Kalchbrenner and Stephan Gouws for their fruitful comments, corrections and inspiration.

Table
Description
Claim
Conclusion

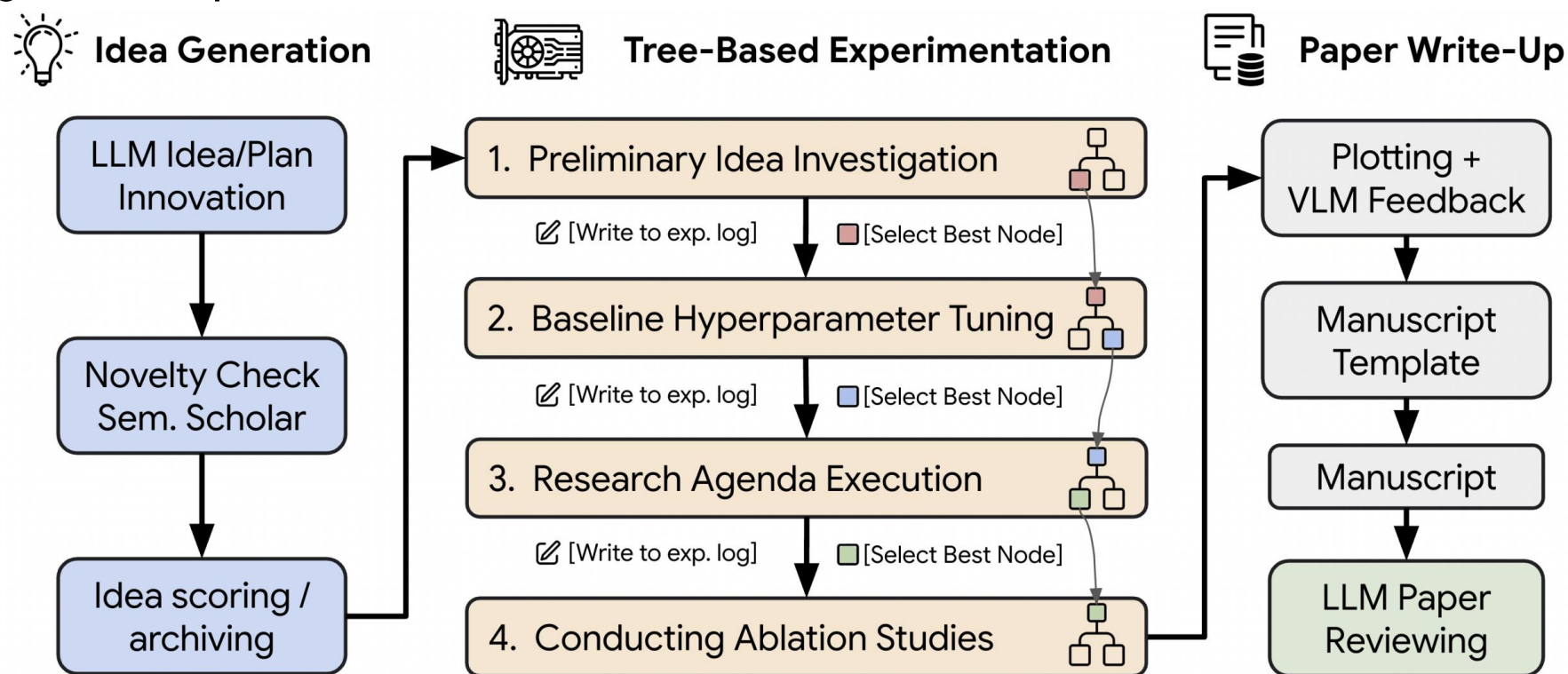


2025 ANNUAL SYMPOSIUM

#AMIA2025

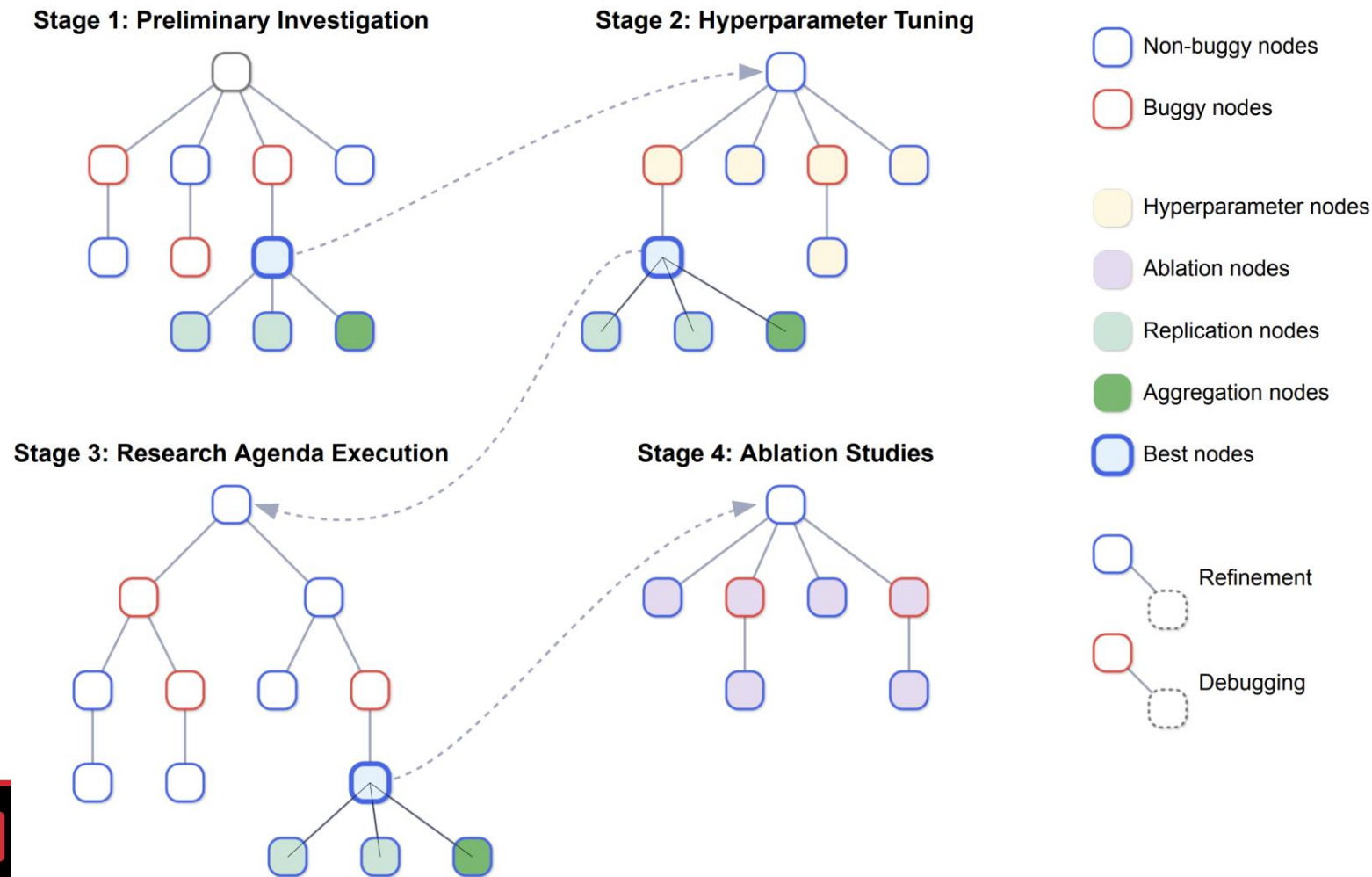
The AIScientist-v2: Workshop-Level Automated Scientific Discovery via Agentic Tree Search

- Iteratively formulate scientific hypotheses, designs and executes experiments, analyzes and visualizes data, and autonomously authors scientific manuscripts
- Integrate vision-language model feedback for iterative refinement
- AI-generated manuscript can successfully pass peer review at a recognized machine learning workshop



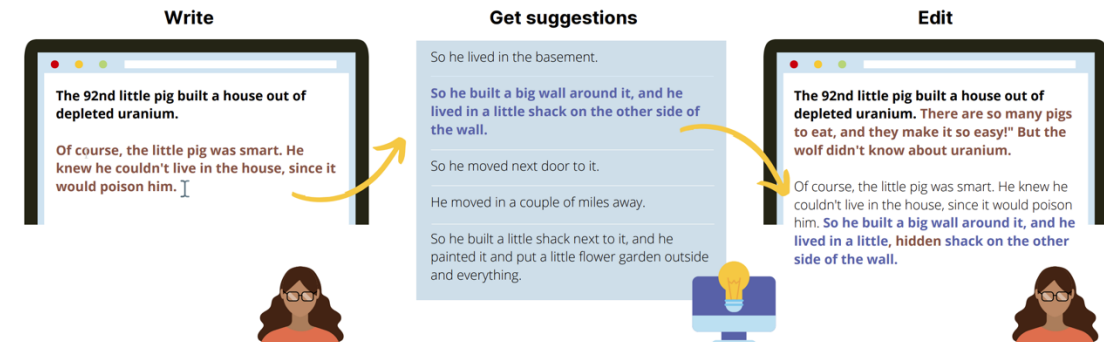
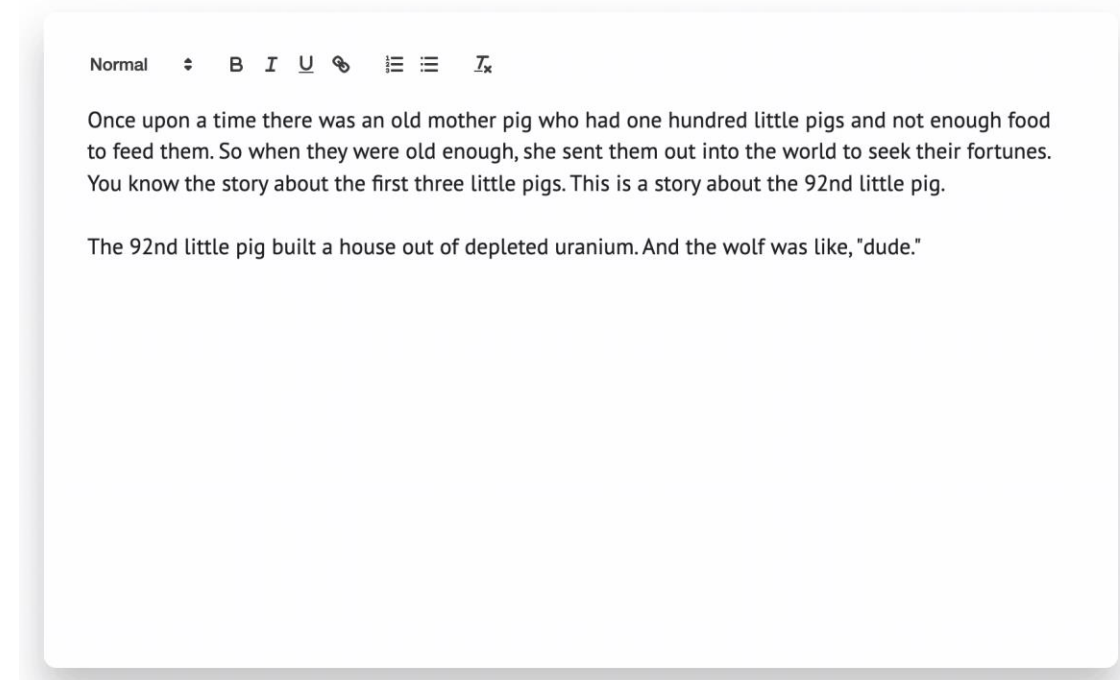
The AIScientist-v2: Workshop-Level Automated Scientific Discovery via Agentic Tree Search

- Integrating tree search with LLM-driven for parallel code generation



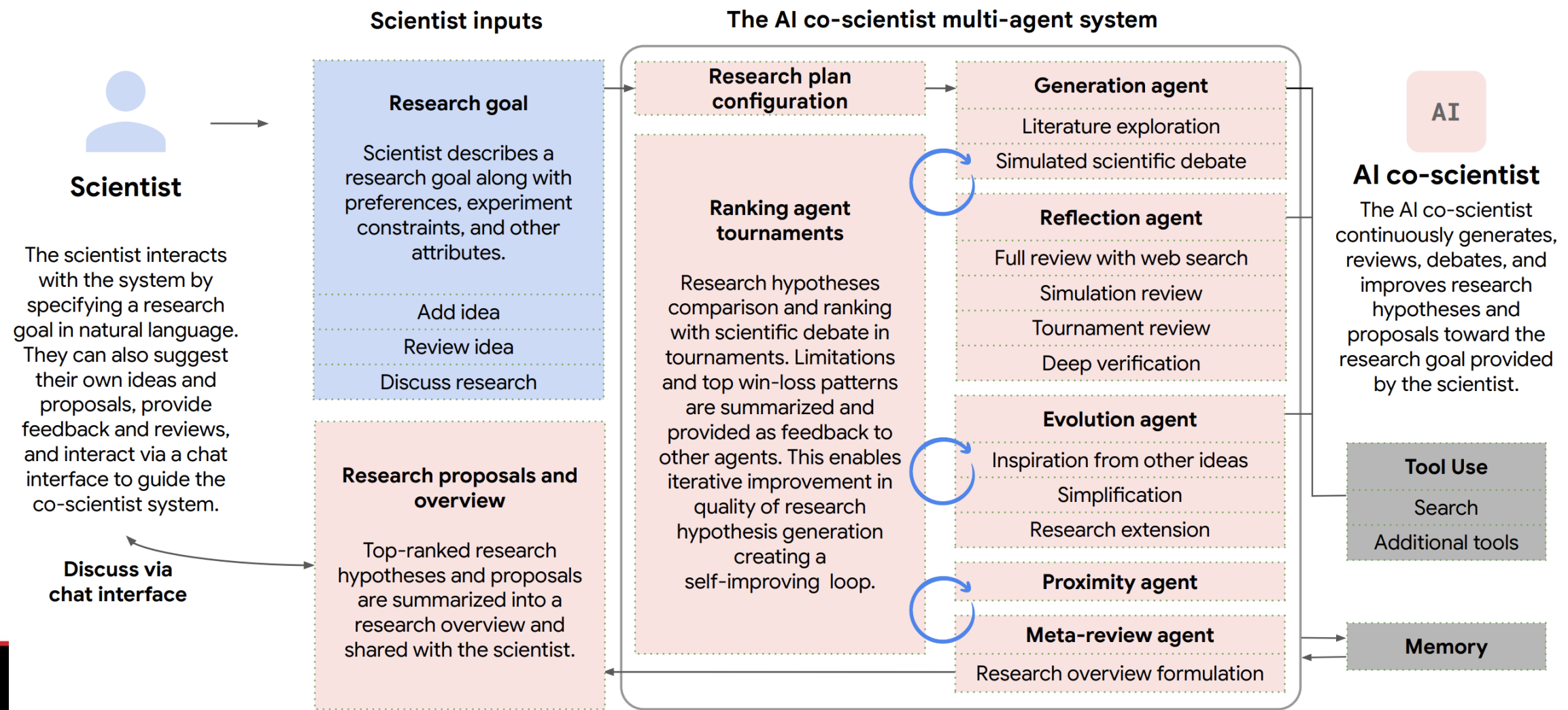
Human-AI Collaborative Writing

- Compared to either humans or AI, collaborative Human-AI Writing has:
 - Fewer grammatical errors
 - Higher vocabulary diversity
- LLMs can increase writers' productivity by providing writing suggestions
- Keep fraction of text written by writers to text written by GPT-3 relatively high can increase writers' feeling of ownership



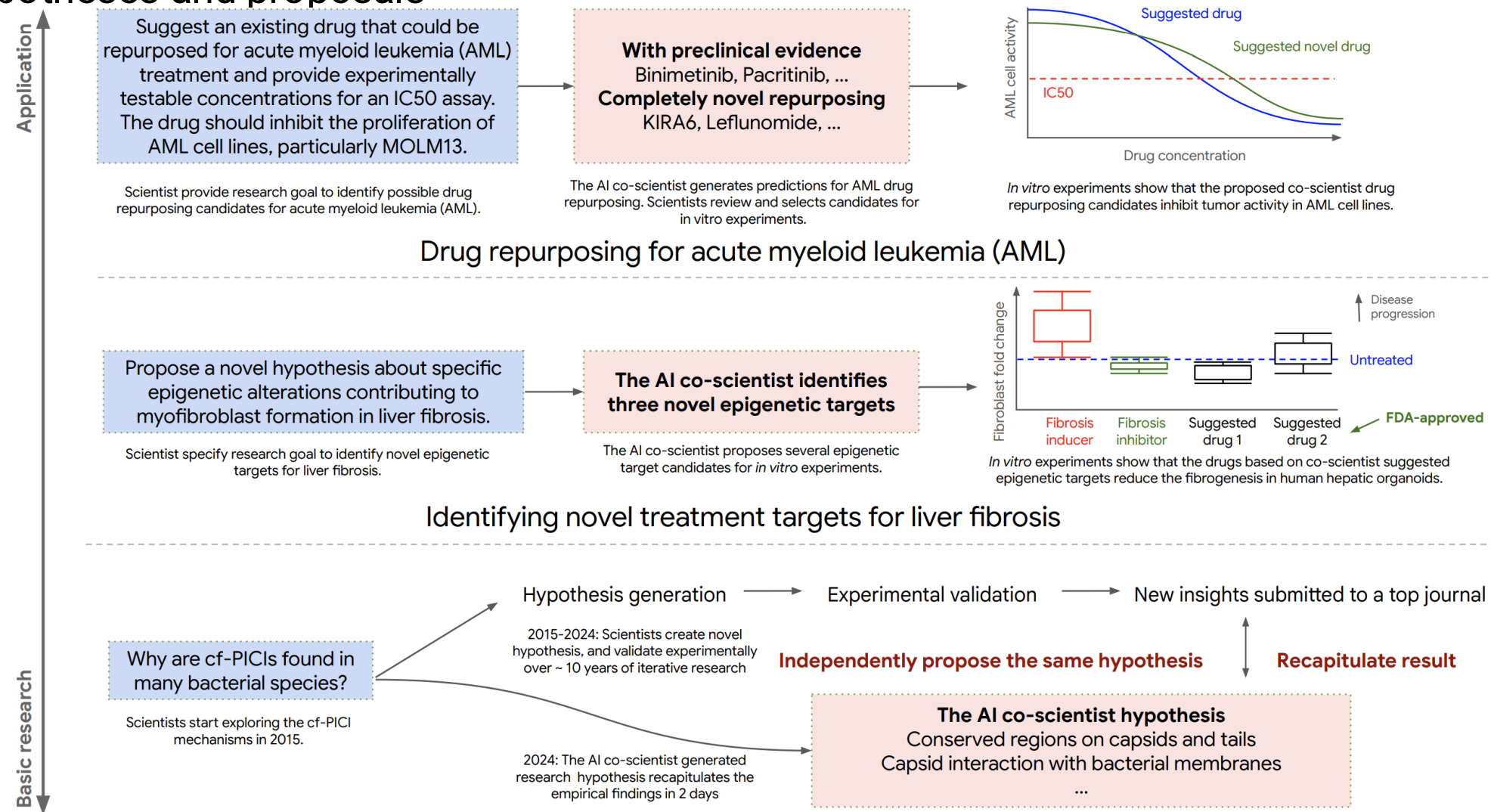
AI co-scientist

- Given a research goal in natural language, the co-scientist generates novel research hypotheses and proposals



AI co-scientist

- Given a research goal in natural language, the co-scientist generates novel research hypotheses and proposals



Parallel in-silico discovery of bacterial gene transfer mechanism relevant to antimicrobial resistance (AMR)

Conclusion and Future Directions for Paper Writing

- Interesting points
 - AI-assisted paper draft writing systems are categorized under knowledge-augmented LLMs
 - These systems extensively utilize both graph and text features for generating results.
 - To reduce error propagation, ongoing developments focus on creating an end-to-end system, moving away from the previous pipeline-based approach
- Future directions
 - Enhancing the multi-hop reasoning capabilities of these systems is also crucial
 - Incorporating human feedback to iteratively improve system results presents an interesting future direction
 - The systems use of only open published literature means it likely has limited access to negative experimental results or records of failed experiments

Radiology Report Generation

- Radiology reporting is time-consuming
- AI can help generate or retrieve reports to assist radiologists
- Traditional Approach:
 - CNN + RNN/Transformer



Enhanced Contrastive Learning with Multi-view Longitudinal Data (MLRG)

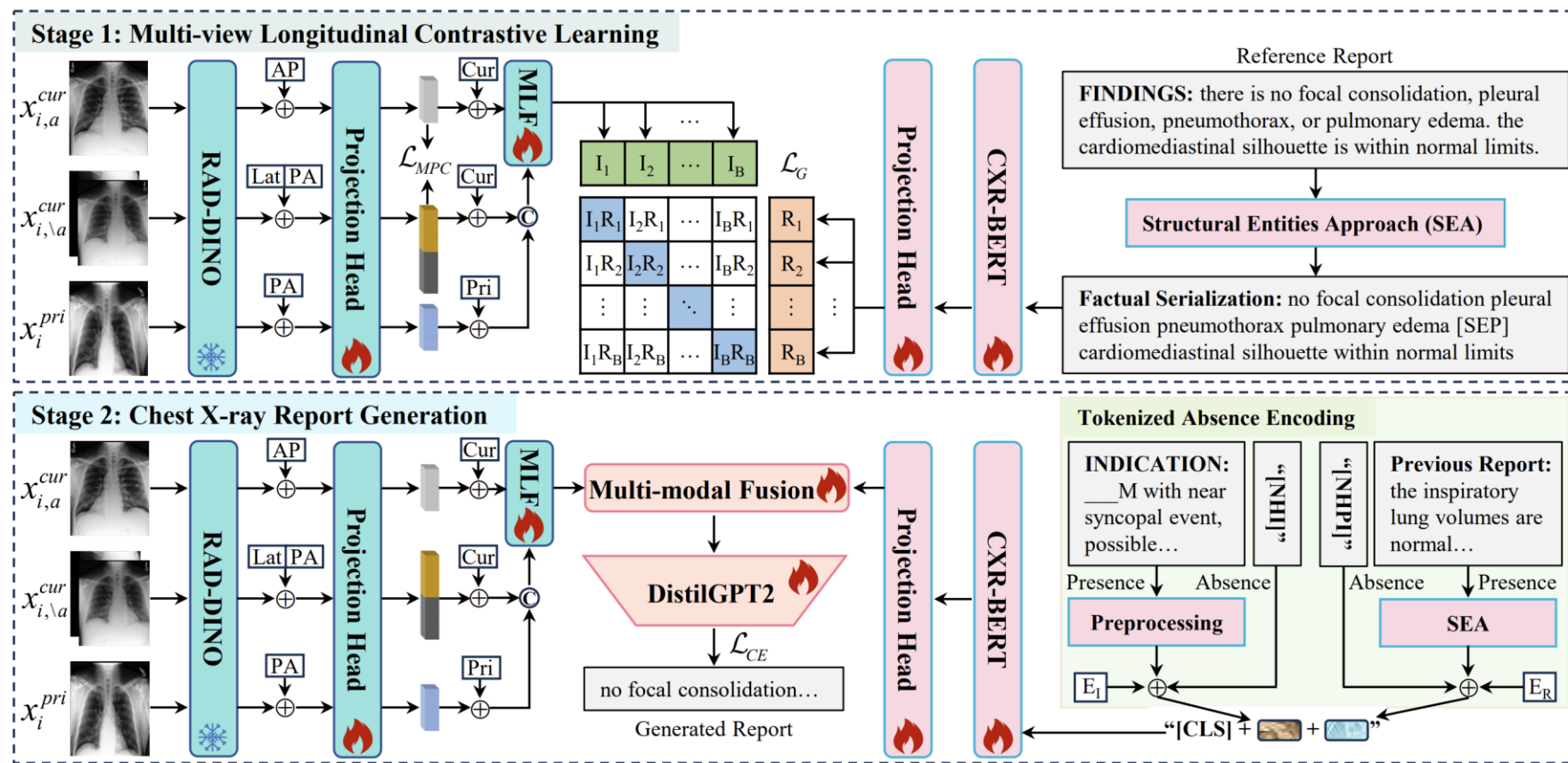
- Contrastive pre-training aligns current and previous images (longitudinal) with text; Select images with different angle (multiview)
- Propose a tokenized absence encoding technique to handle missing patient-specific prior knowledge

Pros:

- Leverage richer data (multiple views + history)
- Use contrastive learning to improve alignment between images and text
- Robust in real scenarios with inconsistent history

Points for Improvement:

- Require prior exams and multi-view images to achieve best performance
- GPT-like Text generation can induce hallucination



$x_{i,a}^{cur}$ and $x_{i,\backslash a}^{cur}$: Anchor scan and auxiliary references (if available) from the current visit.

x_i^{pri} : Previous image (if exists).

🔥: Trainable parameters.

AP, PA, and Lat: View positional embeddings.

Cur and Pri: Temporal positional embeddings.

E_I and E_R : Segment embeddings.

❄️: Frozen parameters.

MLF: Multi-view longitudinal fusion network.

[CLS], [SEP], [NHI], and [NHPR]: Special tokens.

⊕: Summation.

⊙: Concatenation.

RRG-Mamba: Efficient Radiology Report Generation with State Space Model.

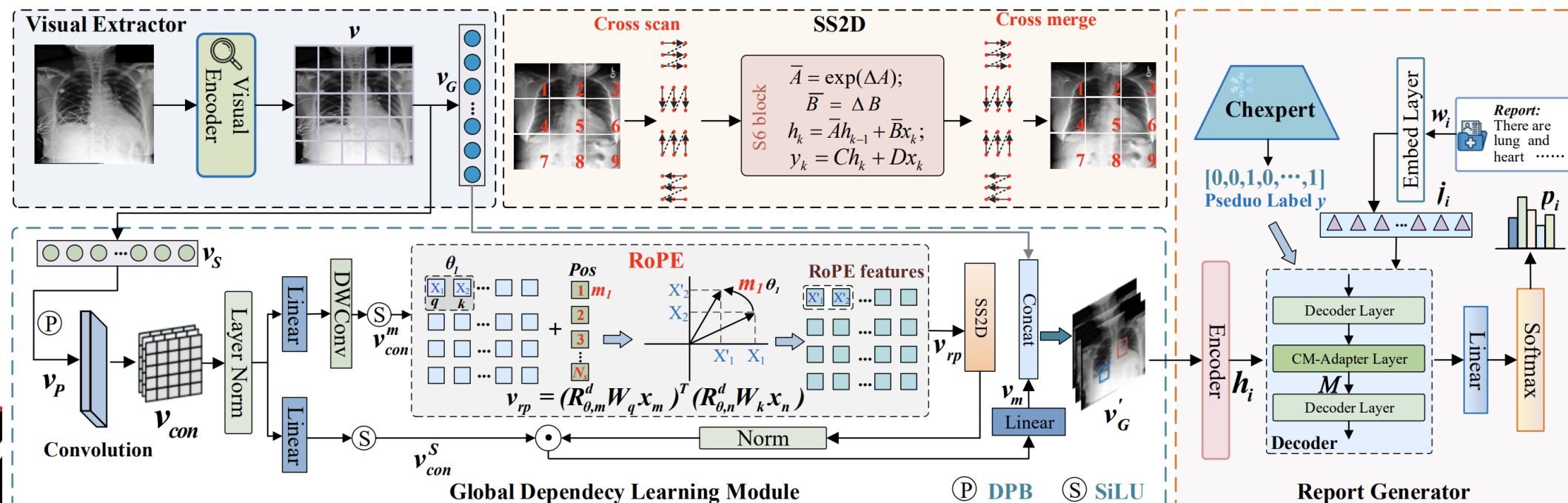
- Generate reports with lower computational cost without losing modeling of long-range dependencies
- Design a global dependency learning module that integrates rotary position encoding

Pros:

- Linear complexity decoding means it scales to longer reports or higher detail without slow-down
- Match or even exceed accuracy of Transformer-based methods
- Use position encoding to addresses both local and long-range feature relationships

Points for Improvement:

- No domain-specific knowledge integration
- Fewer off-the-shelf tools for State Space Models compared to Transformers



Potential Future Directions

- Combine both papers to create an efficient model that takes multi-view inputs
- Retrieve clinically similar reports based on visual input (Yan et al., 2024) and rewrite them based on current images
- Improve evaluation metrics
 - Integrating vision–language models (VLMs) as judge
 - Use attention map and reasoning chain to assess interpretability and factual consistency



Medical Decision-Making

- A multifaceted and intricate process in which clinicians collaboratively navigate diverse sources of information to reach a precise and specific conclusion
- It involves interpreting complex and multi-modal data, such as imaging, electronic health records (EHR), physiological signals, and genetic information

MEDAGENTS: LLMs as Collaborators for Zero-shot Medical Reasoning

- Use role-playing and collaborative agent discussion to increase faithfulness and interpretability of LLM without retrieval-augmented generation

Pros:

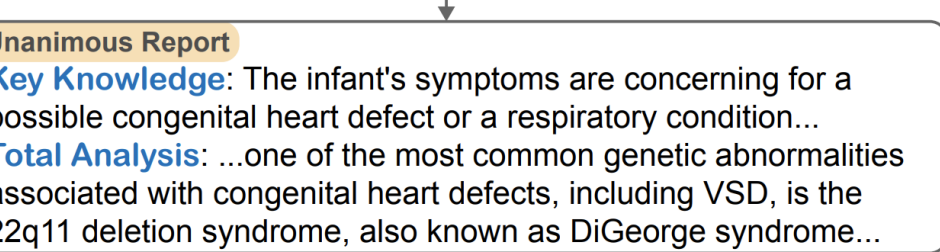
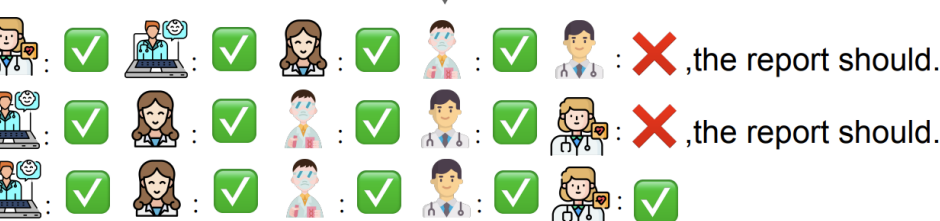
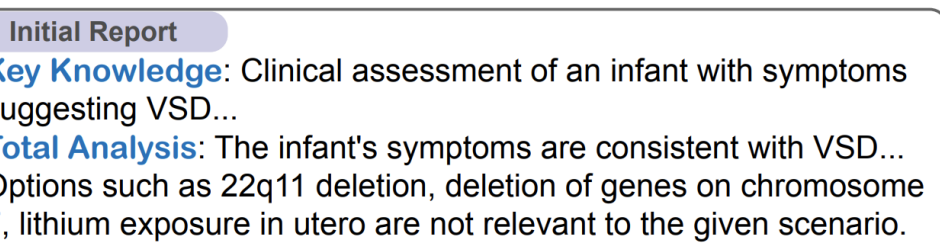
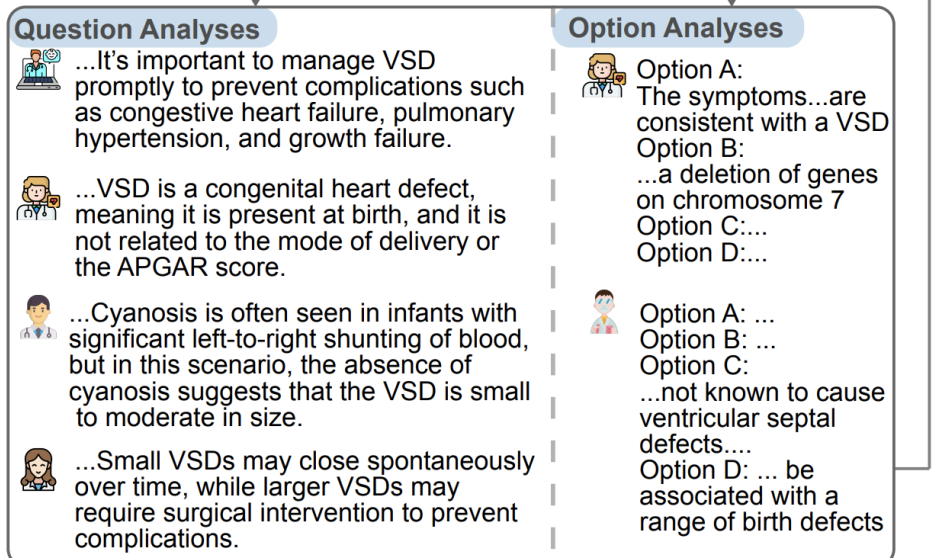
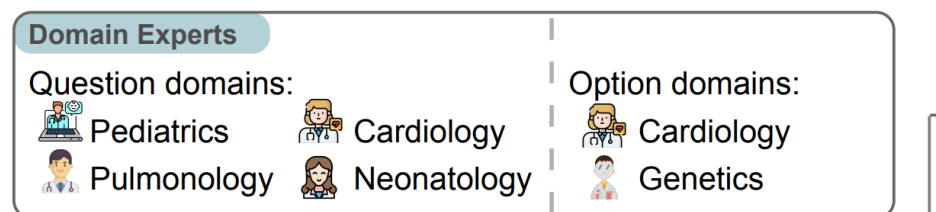
- Do not require external knowledge
- Enhance faithfulness of LLM results
- Provide interpretable reasoning chains

Points for Improvement:

- Use closed-source LLMs, require LLM's pretrained knowledge
- Agent numbers are fixed, which is not computational efficient
- Agent might not reach consensus

Question: A 3-month-old infant is brought to her pediatrician because she coughs and seems to have difficulty breathing while feeding. In addition, she seems to have less energy compared to other babies and appears listless throughout the day. She was born by cesarean section to a G1P1 woman with no prior medical history and had a normal APGAR score at birth. Her parents say that she has never been observed to turn blue. Physical exam reveals a high-pitched holosystolic murmur that is best heard at the lower left sternal border. The most likely cause of this patient's symptoms is associated with which of the following abnormalities?

Options: (A) 22q11 deletion (B) Deletion of genes on chromosome 7 (C) Lithium exposure in utero (D) Retinoic acid exposure in utero



MDAgents: An Adaptive Collaboration of LLMs

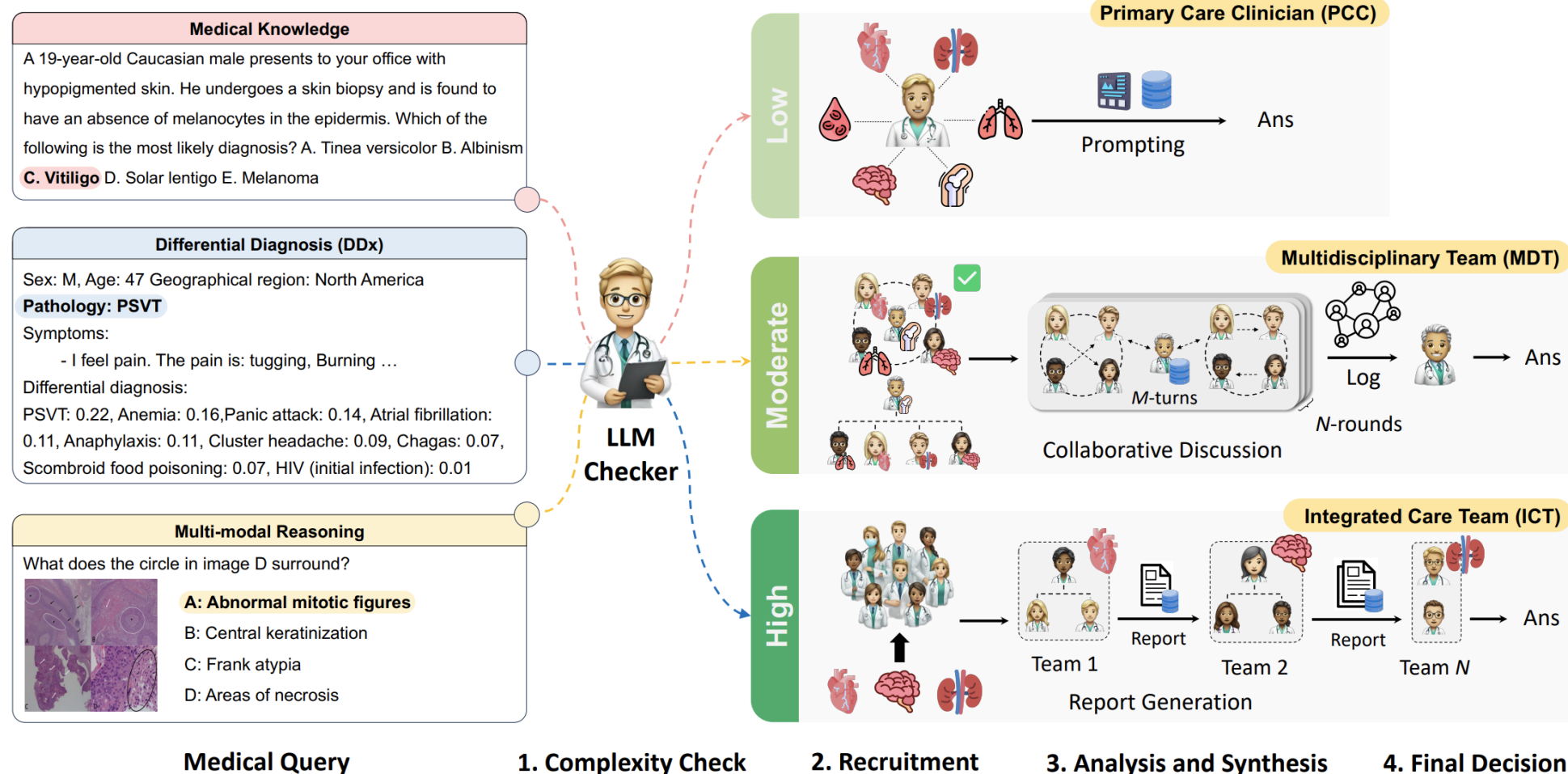
- The first adaptive decision-making framework for LLMs that mirrors real-world MDM processes via dynamic collaboration among AI agents based on task complexity

Pros:

- Framework imitates real-world clinical decision making
- Dynamic collaboration to improve efficiency
- Include a moderator agent to conduct planning
- Can integrate external knowledge

Points for Improvement:

- Use closed-source LLMs
- Limit to QA tasks



Takeaways

- Notable points
 - Moving from static analysis models to interactive collaborators in clinical workflows
 - AI systems can now think, reason, and communicate rather than just process data
 - Success of AI systems depends on transparency, interpretability, and clinician trust



An Interactive Example: Paper Hypothesis Assistant

Example Overview

- We'll consolidate these ideas by exploring an interactive example of hypothesis generation.
- Our hands-on example will consist of two components:
 - 1. Retrieve semantically relevant inspirations from training set
 - 2. Generate hypotheses based on the retrieved inspirations, background context, and seed term
- Find Google Colab <https://colab.google/> notebooks at
- https://github.com/EagleW/tutorial_paper_hypothesis/



Why might language models hurt the scientific process?

- Factuality, misinformation, and hallucination
- Plagiarism
- High quantities of (often low-quality) work
- Issues with peer review
- These are issues now! The cat is out of the bag.

Factuality

- LLMs like ChatGPT tend to hallucinate.
 - The only time this is good for science is hypothesis generation! When we read the literature or write papers, we don't want to make things up (usually).
 - ChatGPT often cites non-existent references. The proportion of nonexistent citations was over 30% with GPT-3.5, and it is over 20% with GPT-4 several months later.
 - ChatGPT generated fake court citations

Publisher blacklists authors after preprint cites made-up studies

Last month, a millipede expert in Denmark received an email notifying him that one of his publications had been mentioned in a new manuscript on Preprints.org. But when the researcher, Henrik Enghoff, downloaded the paper, he learned that it cited his work for something off-topic.



Henrik Enghoff

New York lawyers sanctioned for using fake ChatGPT cases in legal brief

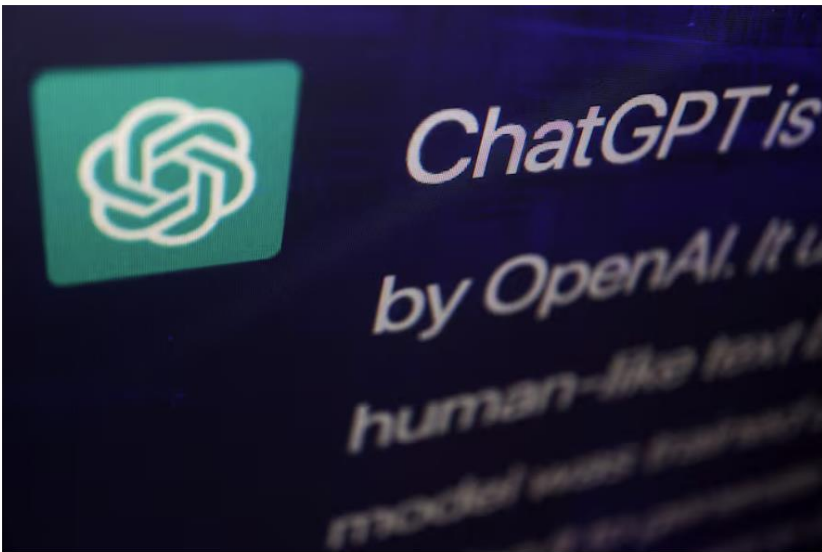
By Sara Merken

June 26, 2023 3:28 AM CDT · Updated 10 months ago



Sponsored by **intel.**

Sponsors are not involved in the creation of this or any other Reuters news articles



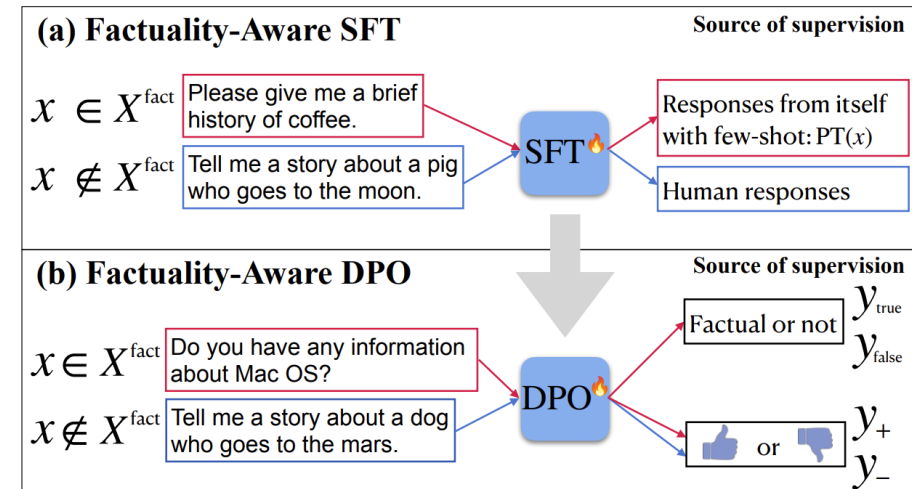
A response by ChatGPT, an AI chatbot developed by OpenAI, is seen on its website in this illustration picture taken February 9, 2023. REUTERS/Florence Lo/Illustration [Purchase Licensing Rights](#)

Mitigation Methods: During Training

- Both supervised fine-tuning (SFT) and reinforcement learning (RL) can encourage hallucination
- Factuality-aware Alignment
 - **Supervised Fine-Tuning (SFT) Adjustments:** Use the LLM's own existing knowledge to generate training data, thus minimizing the incorporation of unfamiliar information
 - **Reinforcement Learning (RL) Tweaks:** The model's reward function is adjusted to include a direct preference for factuality using a factuality reward model (SFT)

Future Directions:

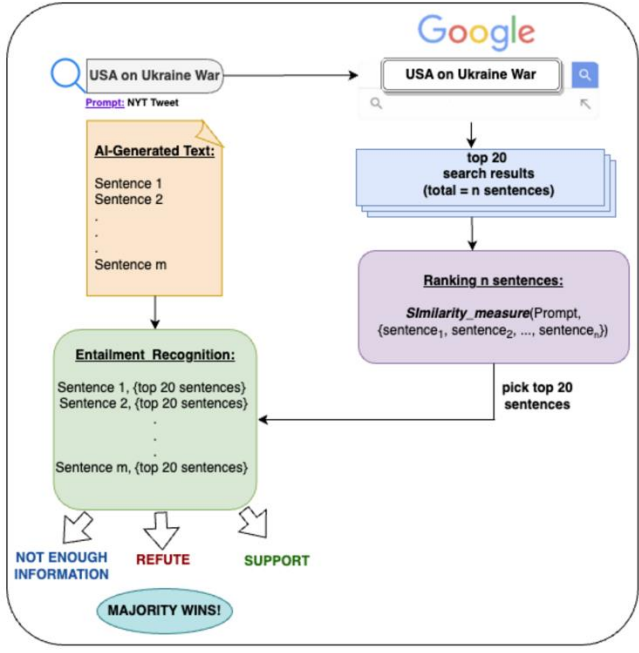
- Incorporate claim generation modules from other systems and external KBs might help LLMs verify factuality better



Mitigation Methods: After Training

- **High Entropy Word Spotting and Replacement:** Identify high-entropy words in hallucinated text and replace them with predictions from another LLM
- **Sentence Factuality Checking:** Identify sentences that are considered suspicious by NLI models for human review

Prompt	USA's stance on Ukraine Russia war
Hallucinated text	The official stance of the United States on the Russia-Ukraine war has been consistent in supporting Ukraine's sovereignty, territorial integrity, and the peaceful resolution of the conflict. The U.S. has condemned Russia's actions in annexing Crimea and its support for separatist movements in eastern Ukraine. Despite maintaining a diplomatic stance, the U.S. President Barack Obama stated that the United States would not deploy troops to Ukraine. Additionally, this lack of support extends to the withholding of financial and military assistance, including military equipment, training, and advisory support. This sentiment was reiterated by Paul Altman during a press review at the White House ...
Mitigated text ENTROPY _{BB}	The official stance of the government on the ongoing war has been consistent in supporting resolution of the conflict. The government has condemned Russia in country and its support for separatist movements in eastern country. Despite maintaining a neutral stance, the government stated that the country would not deploy troops to unknown. Additionally, this lack of support extends to the withholding of financial and resources, including administrative, technical, and legal. This sentiment was echoed by minister during a press review at the conference ...
Edit Suggestions FACTUALITY _{GB}	The official stance of the United States on the Russia-Ukraine war has been consistent in supporting Ukraine's sovereignty, territorial integrity, and the peaceful resolution of the conflict. The U.S. has condemned Russia's actions in annexing Crimea and its support for separatist movements in eastern Ukraine. Despite maintaining a diplomatic stance, U.S. President Barack Obama stated that the United States would not deploy troops to Ukraine. Additionally, this lack of support extends to the withholding of financial and military assistance, including military equipment, training, and advisory support. This sentiment was reiterated by Paul Altman during a press review at the White House ...



Quantity overwhelms existing systems

- Papers can be generated much quicker now, but they aren't necessarily good
 - Current scientific productivity metrics are measured by quantity
 - "When a measure becomes a target, it ceases to be a good measure."
- Using "As of my last knowledge update" on Google Scholar returns 100s of results!

ChatGPT on the Gravitational Redshift

J Stávek - European Journal of Applied Physics, 2024 - ej-physics.org

... **As of my last knowledge update** in January 2022, I'm not aware of any widely accepted alternative scenarios or theories that challenge Einstein's interpretation of gravitational redshift in ...

☆ Save Cite Related articles All 2 versions

An Evaluation of ChatGPT and Bard in the Context of Biological Knowledge Retrieval

R Caspi, PD Karp - Access Microbiology, 2024 - microbiologyresearch.org

... A: **As of my last knowledge update** in September 2021, the term "rbcX" was not commonly associated 345 with a specific protein or gene in the context of photosynthesis or carbon ...

☆ Save Cite Related articles All 2 versions



Radiology Case Reports
Volume 19, Issue 6, June 2024, Pages 2106-2111



Case Report

Successful management of an Iatrogenic portal vein and hepatic artery injury in a 4-month-old female patient: A case report and literature review

Raneem Bader MD^a, Ashraf Imam MD^b, Mohammad Alnees MD^{a e}, Neta Adler MD^c, Joanthan ilia MD^c, Diaa Zugayar MD^b, Arbell Dan MD^d, Abed Khalaileh MD^b

Show more

+ Add to Mendeley Share Cite

<https://doi.org/10.1016/j.rader.2024.02.037>

Get rights and content

Under a Creative Commons license

open access

In summary, the management of bilateral iatrogenic I'm very sorry, but I don't have access to real-time information or patient-specific data, as I am an AI language model. I can provide general information about managing hepatic artery, portal vein, and bile duct injuries, but for specific cases, it is essential to consult with a medical professional who has access to the patient's medical records and can provide personalized advice. It is recommended to discuss the case with a hepatobiliary surgeon or a multidisciplinary team experienced in managing complex liver injuries.

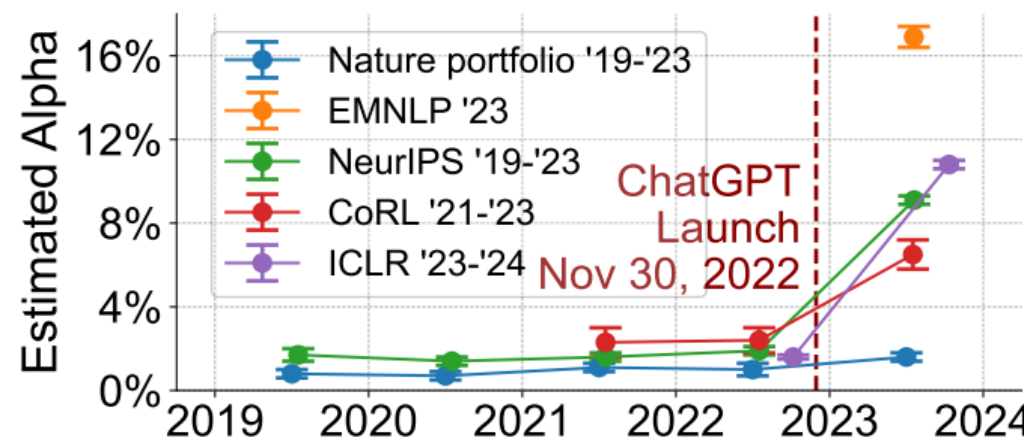
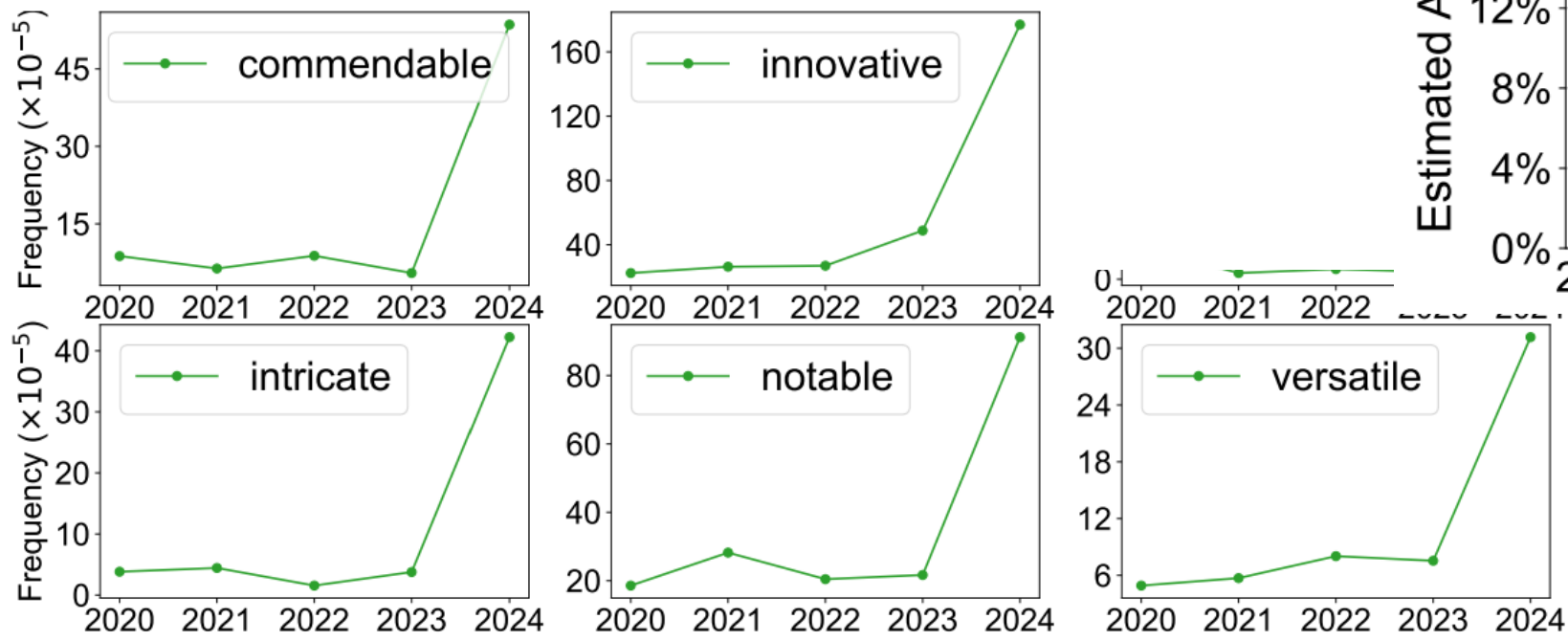


2025 ANNUAL SYMPOSIUM

#AMIA2025

Quantity overwhelms existing systems

- Peer-review challenges
 - Current scientific productivity metrics are measured by quantity
 - Measuring quality is hard to do without retrospect!
 - There aren't enough people to do free peer-review
 - Peer review takes time away from writing more papers
 - Many are turning to LLMs for help, but are the results thorough?



What can we do? -- LLM Detectors

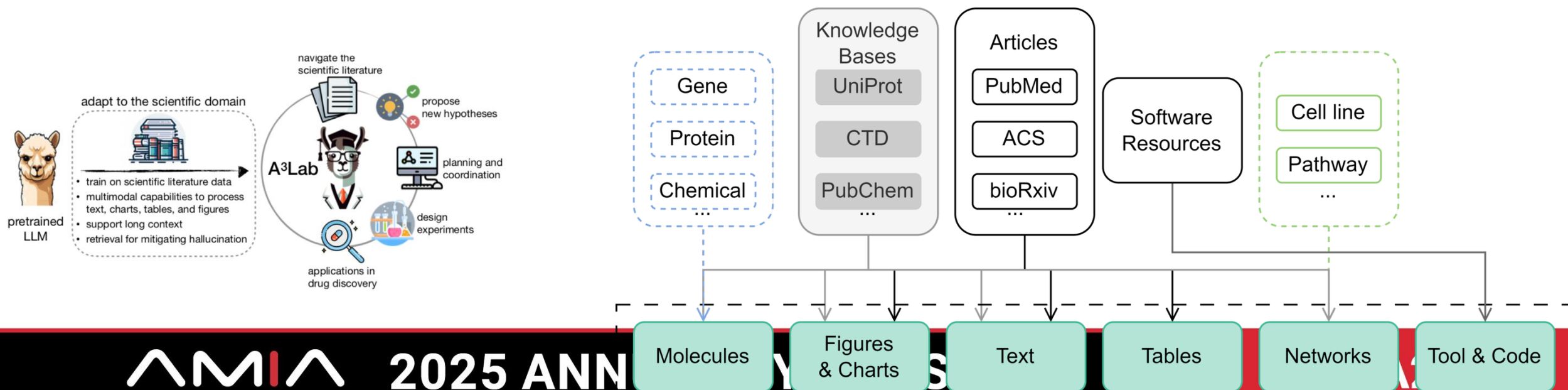
- Watermarking text (Kirchenbauer et al., 2023)
 - Encourage an unnatural distribution of generated words
 - Split the token set into a red list and a green list with a random seed.
 - Promote the use of green tokens, i.e., increase its probability
- Zero-shot approaches
 - Measure the perplexity and complexity of a generated sentence
- Retrieval (Krishna et al., 2023)
 - Given candidate text, search a database of previous LLM generations, looking for sequences that match the candidate text within a certain threshold
- Neural Network detectors for AI-generated text
 - Train a model to distinguish between human and AI text
 - These go out of date quickly and can be fooled by stronger models

No watermark
Extremely efficient on average term lengths and word frequencies on synthetic, microamount text (as little as 25 words)
Very small and low-resource key/hash (e.g., 140 bits per key is sufficient for 99.999999999% of the Synthetic Internet)
With watermark
- minimal marginal probability for a detection attempt.
- Good speech frequency and energy rate reduction.
- messages indiscernible to humans.
- easy for humans to verify.

Conclusion and Future Directions

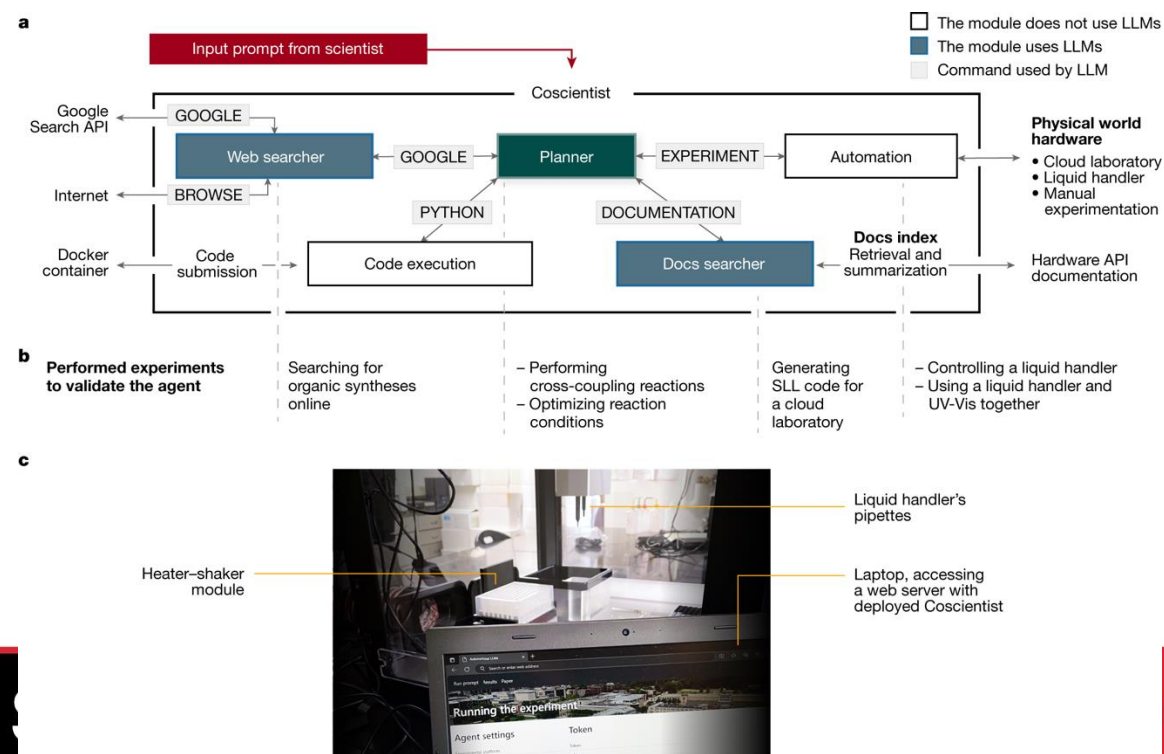
Future Direction 1: Science-Inspired Large Language Models

- Existing work chooses either highly-capable general domain models (e.g., GPT-4) with strong reasoning capabilities or trains specific scientific language models (e.g., SciBERT) with specific knowledge.
 - Can we more effectively bridge this gap?
- Develop more capable multimodal LLMs
 - Explore instruction finetuning on a suite of datasets carefully designed to improve LLM capabilities in key tasks for scientific applications



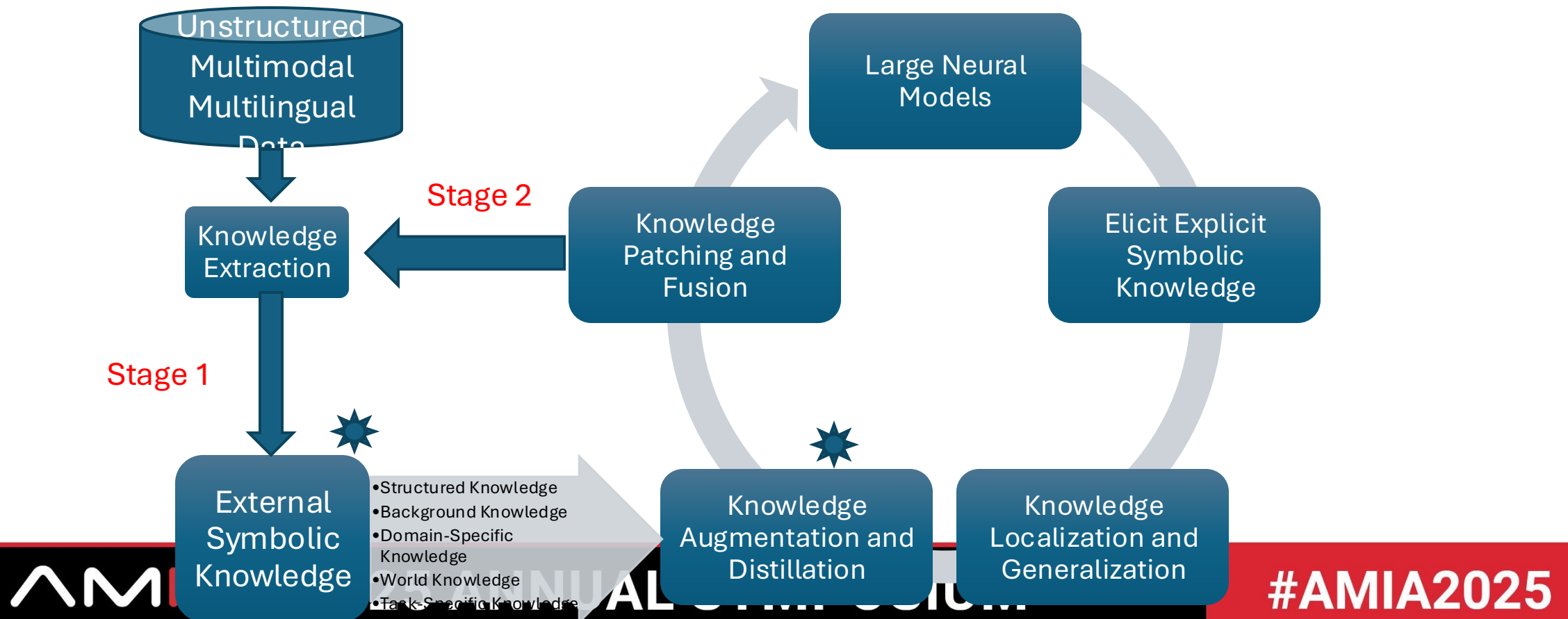
Future Direction 2: Automated Experimental Systems

- Convert automated agents for scientific discovery from prototype ideas into useful tools
 - Read the entire scientific literature and synthesize relevant information to propose novel ideas
 - Couple with external simulations for rapid feedback
 - Integrate with specially designed laboratory hardware for automated experimentation and testing



Future Direction 3: Completing the LLM Knowledge Lifecycle

- Create a two-way knowledge acquisition framework to make symbolic and neural learning approaches mutually enhance each other



Questions?



INFORMATICS PROFESSIONALS. LEADING THE WAY.

