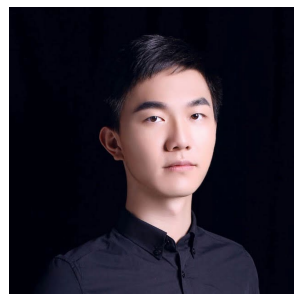




Towards a Human-Computer Collaborative Scientific Paper Lifecycle



QINGYUN WANG



CARL EDWARDS



HENG JI



TOM HOPE

Participation and Q & A

- All tutorial slides and reading list are available at:
<https://sites.google.com/view/coling2024-paper-lifecycle/>
- Throughout the tutorial, we will provide Q & A on sli.do (also on the website):
<https://tinyurl.com/scientificpaperlifecycle>



Ask & upvote questions anytime!
Attendees on Zoom can also type in chat



Disclaimer: This tutorial is presenters' own opinions



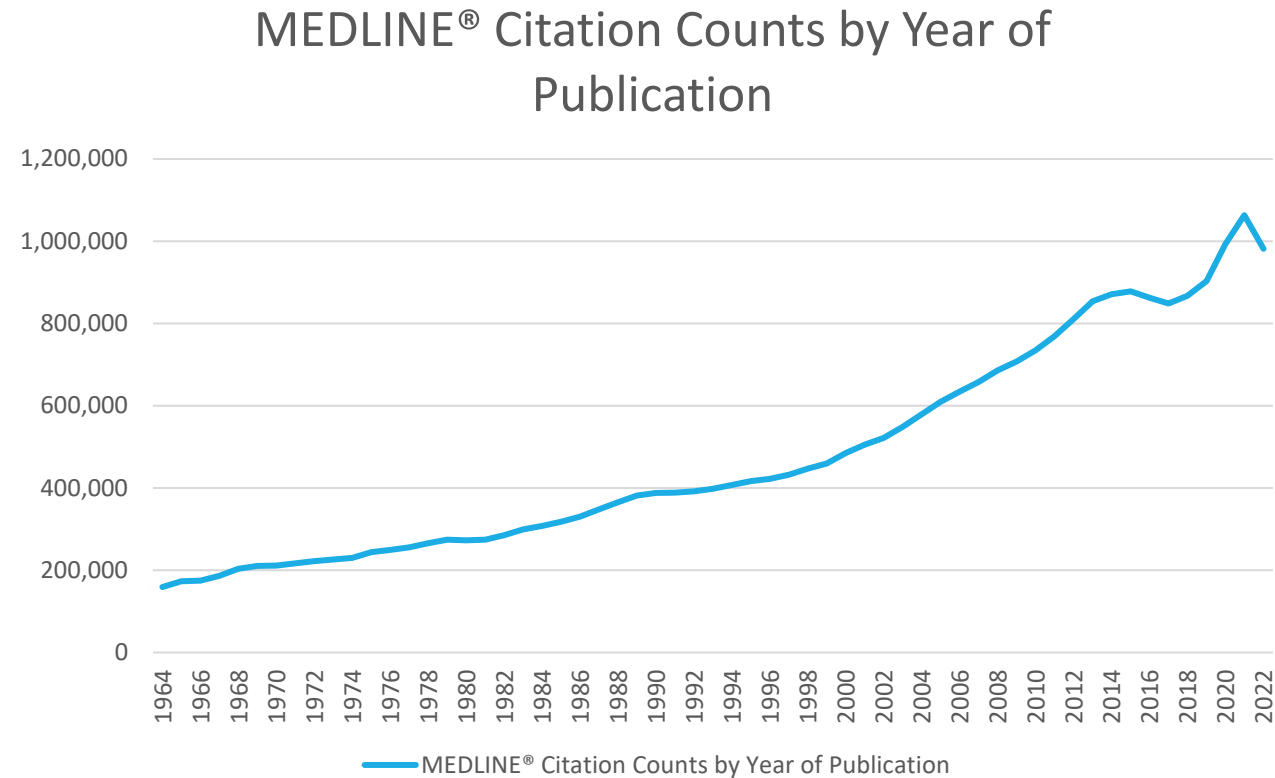
- To access mentioned models + datasets, please refer to corresponding licensing information
- We're not promoting the use of any particular model and/or datasets
- There are slides / figures borrowed from respective papers
- This tutorial is by no means exhaustive: we've tried our best to include relevant materials



Why do we need AI to help scientists?



- Quantity: More than 1M papers are added to PubMed every year, bringing the total number of papers to over 36M

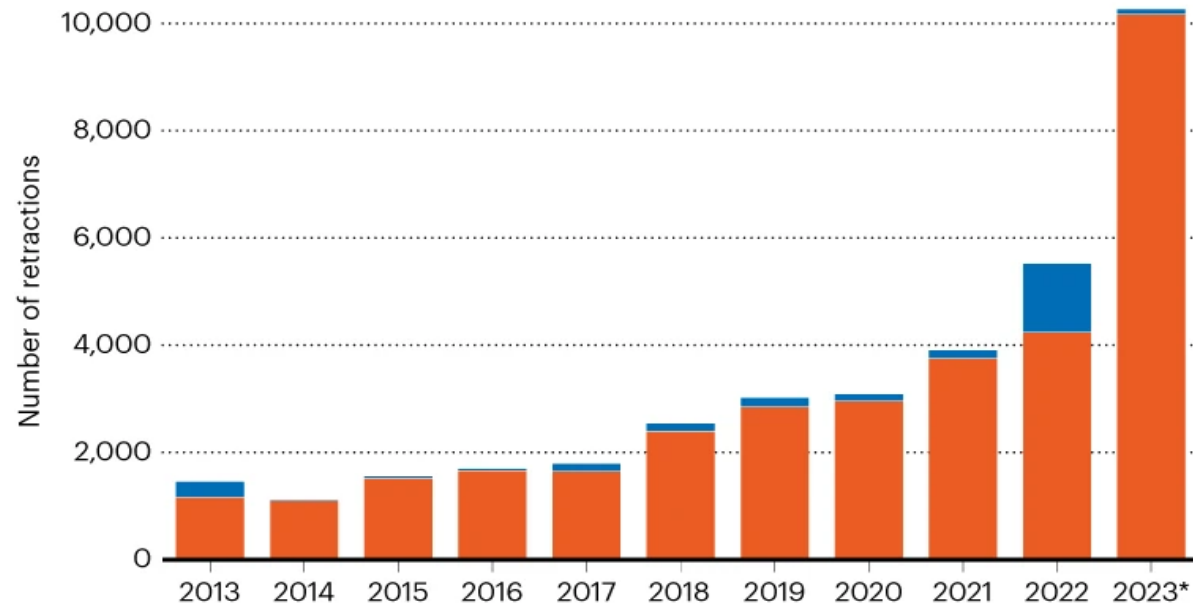


- Quality: The number of retractions issued for research articles in 2023 has passed 10k as publishers struggle to clean up numerous meaningless papers and peer-review fraud

A BUMPER YEAR FOR RETRACTIONS

Retraction notices in 2023 have passed 10,000, largely because of more than 8,000 retractions by Hindawi.

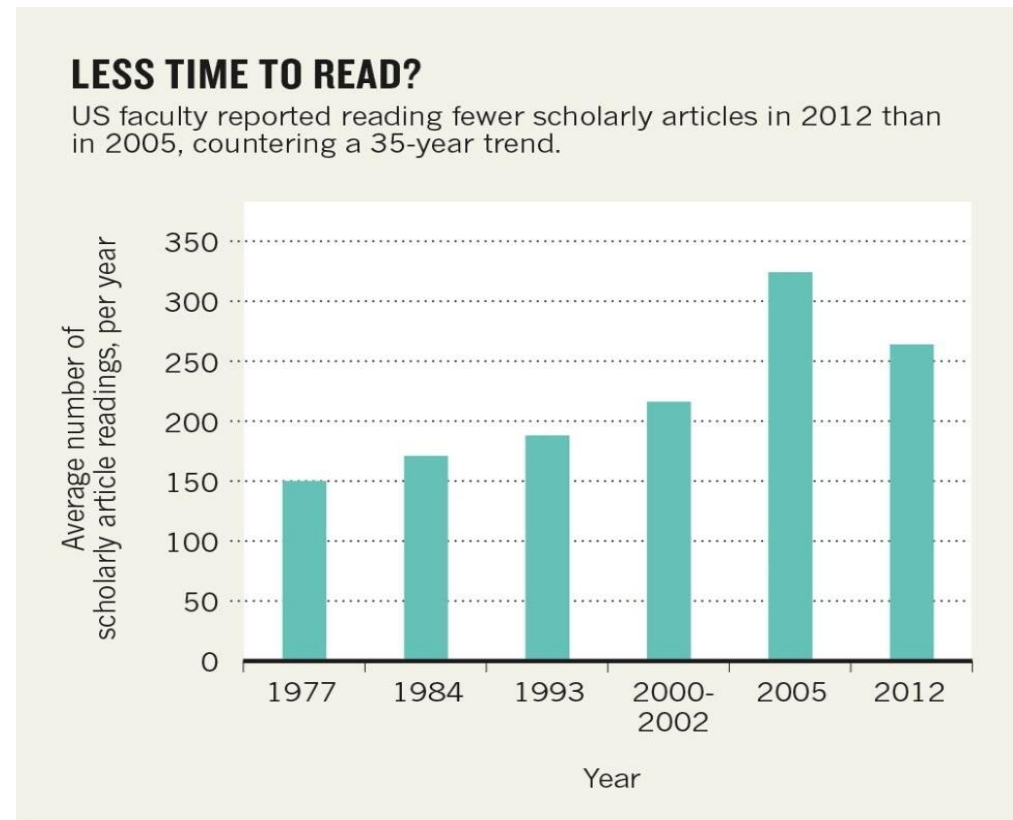
■ Journal articles ■ Conference papers



©nature

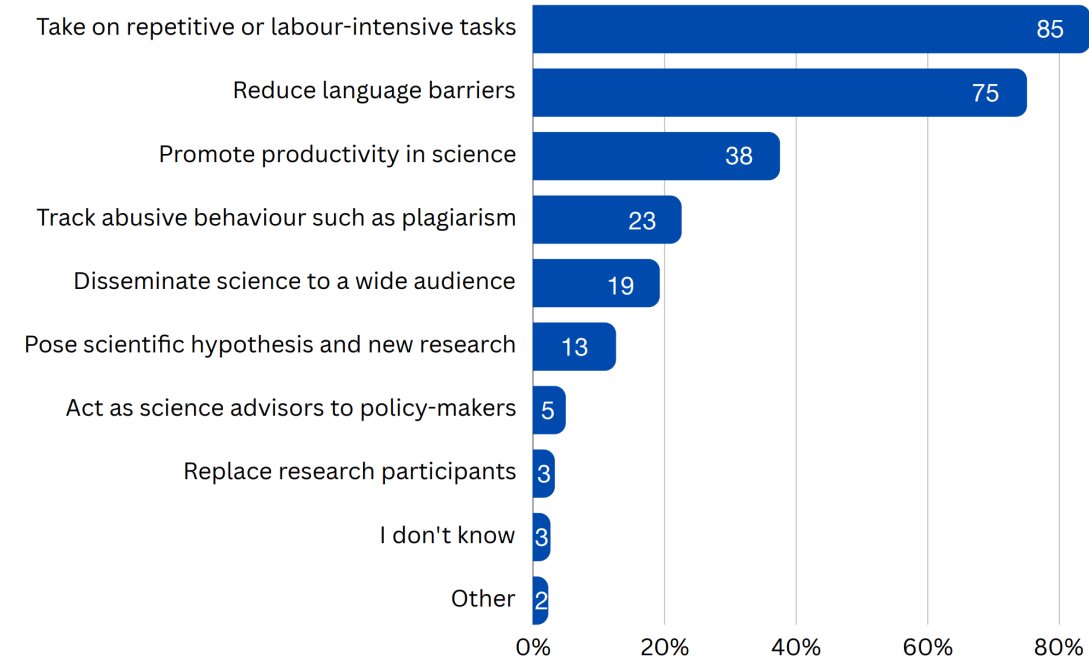
*As of 8 December 2023

- Human's reading ability keeps almost the same across years: US scientists estimated that they read, on average, only about 300 papers per year

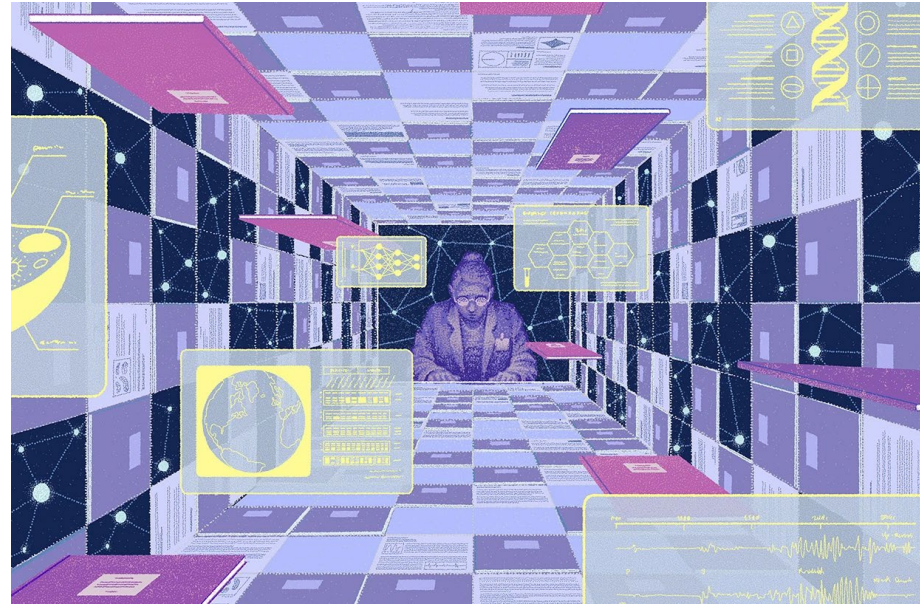


Benefits of LLMs in Scientific Research

- 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



- Reducing labor-intensive tasks
 - 56% of the postdocs in Nature's survey employ chatbots to generate, edit and troubleshoot code
 - A 2023 Nature survey found that more than 30% use AI to help them write code and that more than 20% use the technology to help them fill out work-related administrative emails



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



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 multidisciplinary 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.

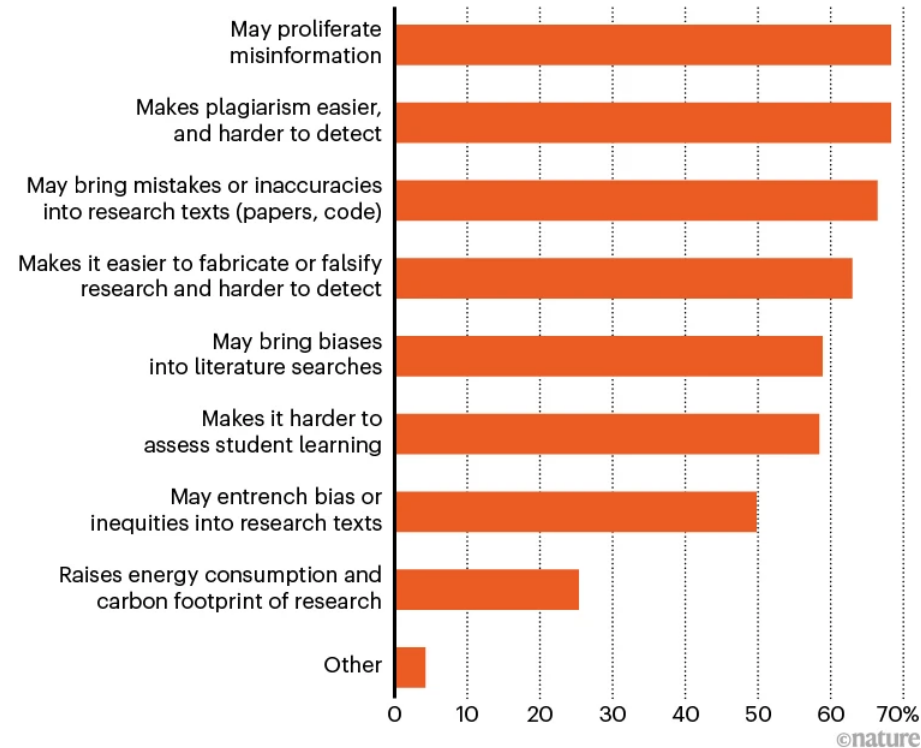
- Spreading misinformation
 - Many physicians posted links or screenshots to articles claiming that *ivermectin* decreased mortality and hospitalization and increased time to recovery and viral clearance, during the covid 19
 - LLMs might spread misinformation if it is present in their training sets or document corpus



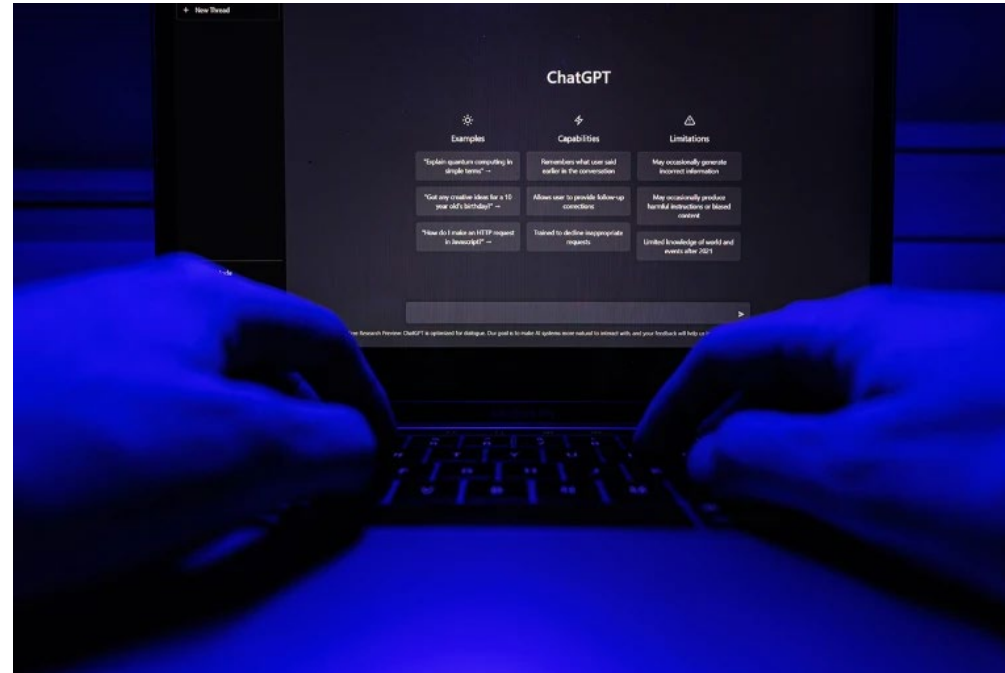
- Exacerbate inequity
 - AI-driven detection tools are more likely to erroneously flag text written by non-native English speakers as AI

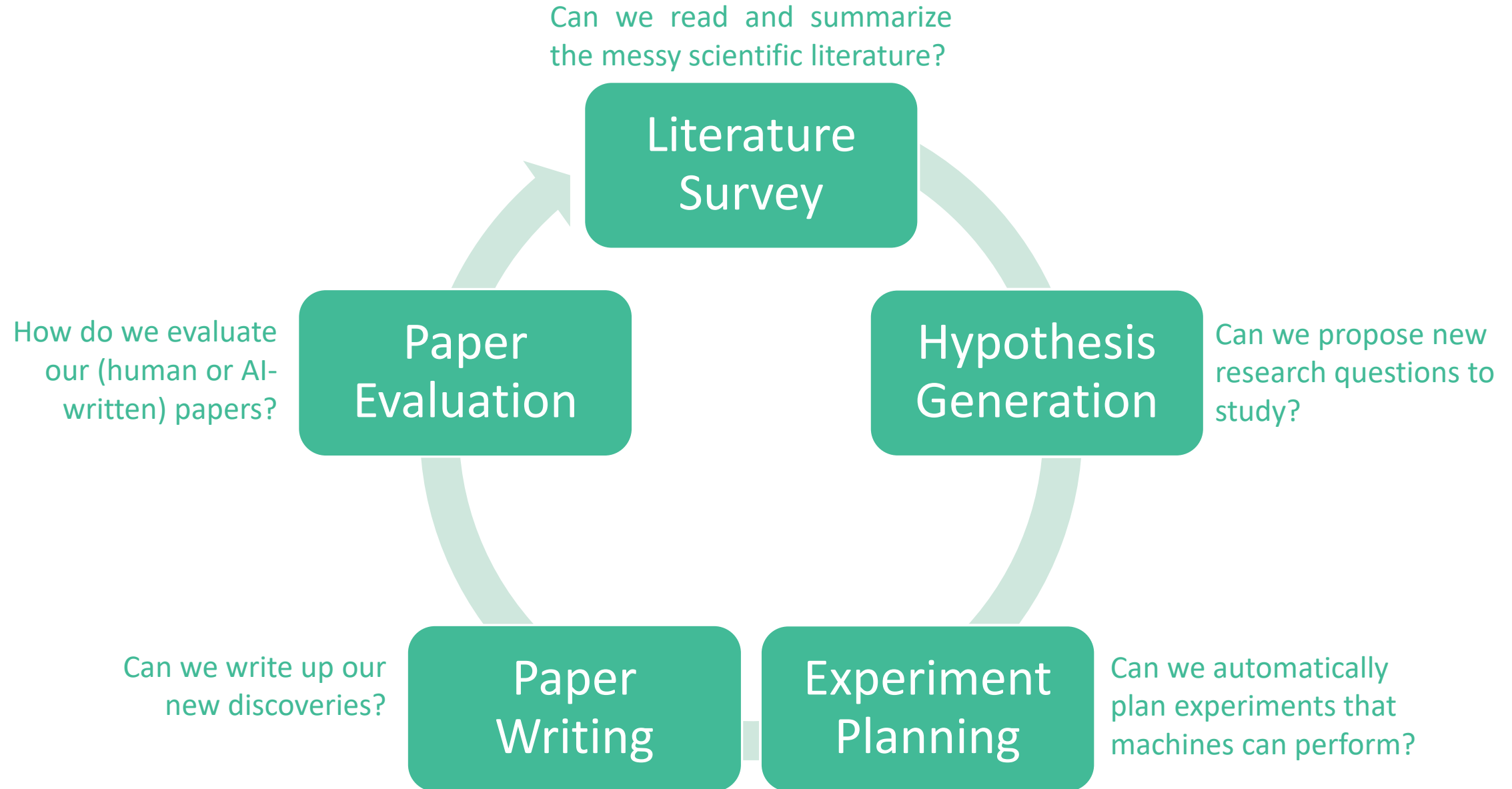
PROBLEMS OF GENERATIVE AI

Q: Where do you think generative AI may have negative impacts on research? (Choose all that apply.)



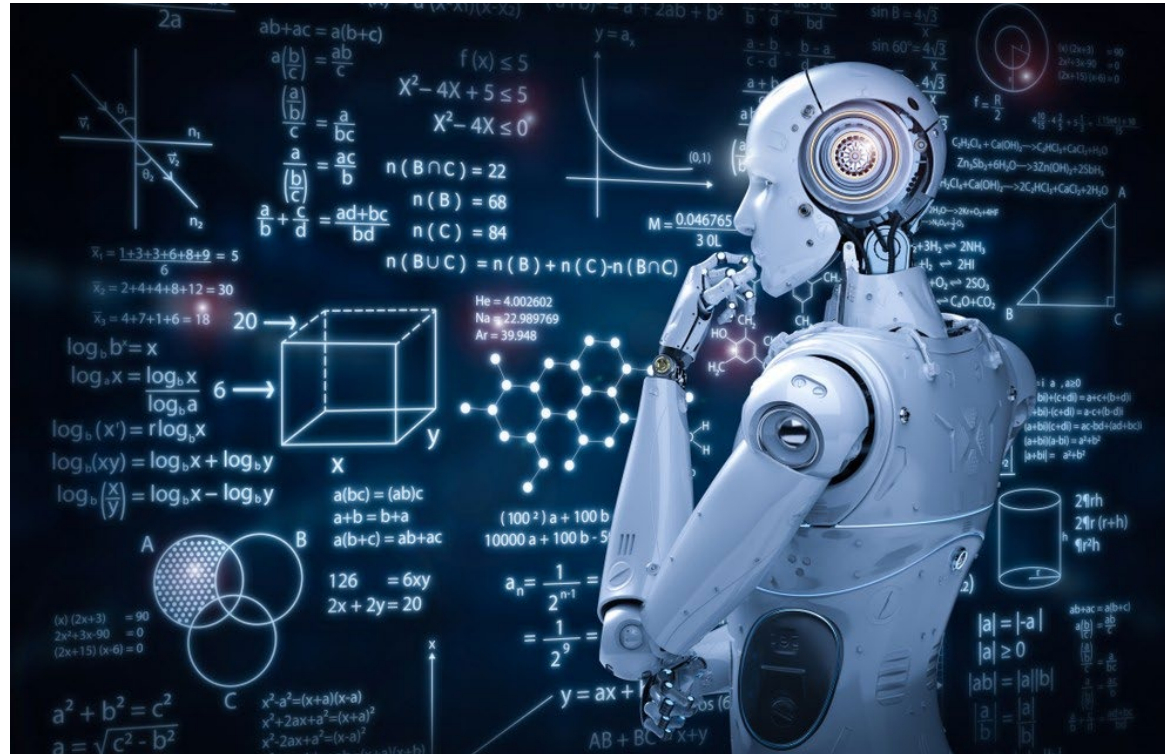
- Peer-review challenges
 - With this ever-increasing number of papers , there aren't enough people available to continue to do free peer review for publishers
 - 17% of the peer-review reports have been substantially modified by chatbots based on a case study of AI conferences that took place after the release of ChatGPT
 - Using external LLMs for peer review may compromise the confidentiality of research





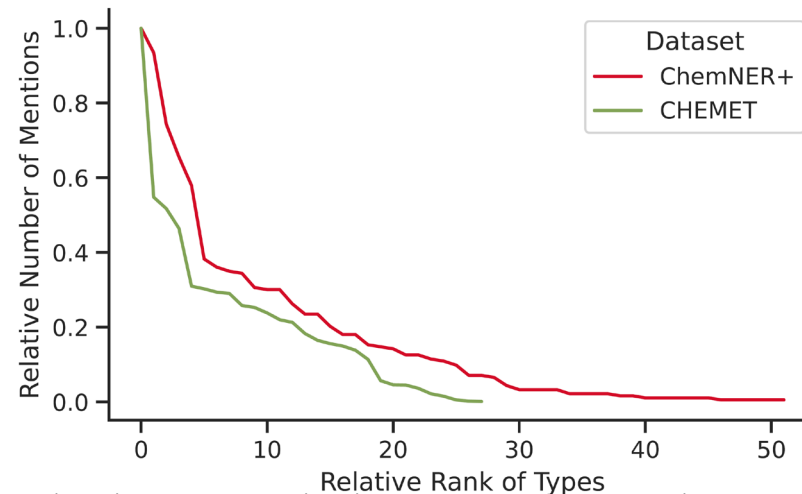
What's stopping us from doing this right now?

- Challenge 1: Limited High-Quality Data
- Challenge 2: Lack of Domain Knowledge
- Challenge 3: Factuality



Challenge 1: Limited High-Quality Data

- Obtaining high-quality human annotations is very expensive and time-consuming for the scientific domain
 - Few benchmarks are publicly available for scientific domain
 - Models only have a limited number of training samples for each task
- Long-tail problems are more prevalent in scientific domain compared to general domain
 - LLMs can't memorize all (long-tail) knowledge in their parameters

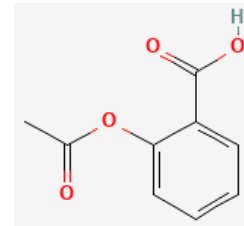


Challenge 2: Lack of Domain Knowledge

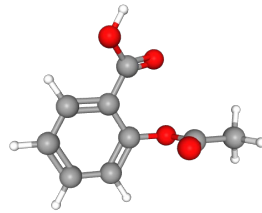
- Comprehending scientific papers requires external knowledge
 - Most fields lack an existing knowledge graph (KG)
 - KGs miss important entities (e.g., new concepts, fine-grained rare concepts)
 - Scientific documents contain more entities per sentence compared to sentences in general domain
- Facts change over time
 - e.g., the Comparative Toxicogenomics Database updates every month


Acetylsalicylic acid has both anti-inflammatory and antipyretic effects

Aspirin is a Nonsteroidal Anti-inflammatory Drug and Platelet Aggregation Inhibitor.



Aspirin and Acetylsalicylic acid refer to the same chemical CH3COOC6H4COOH.



 **ctd** *Illuminating how chemicals affect human health.*

Comparative Toxicogenomics Database

[Home](#) [Search](#) [Analyze](#) [Download](#) [Commercial Users](#) [Help](#)

CTD Changes

April 29, 2024 (17342)

1. Exposure Data Migration software release
2. Extensive DML/schema modifications
3. PWA API Updates
4. Exposure Curation Tool receptor fixes
5. CTD use update
6. Published new data

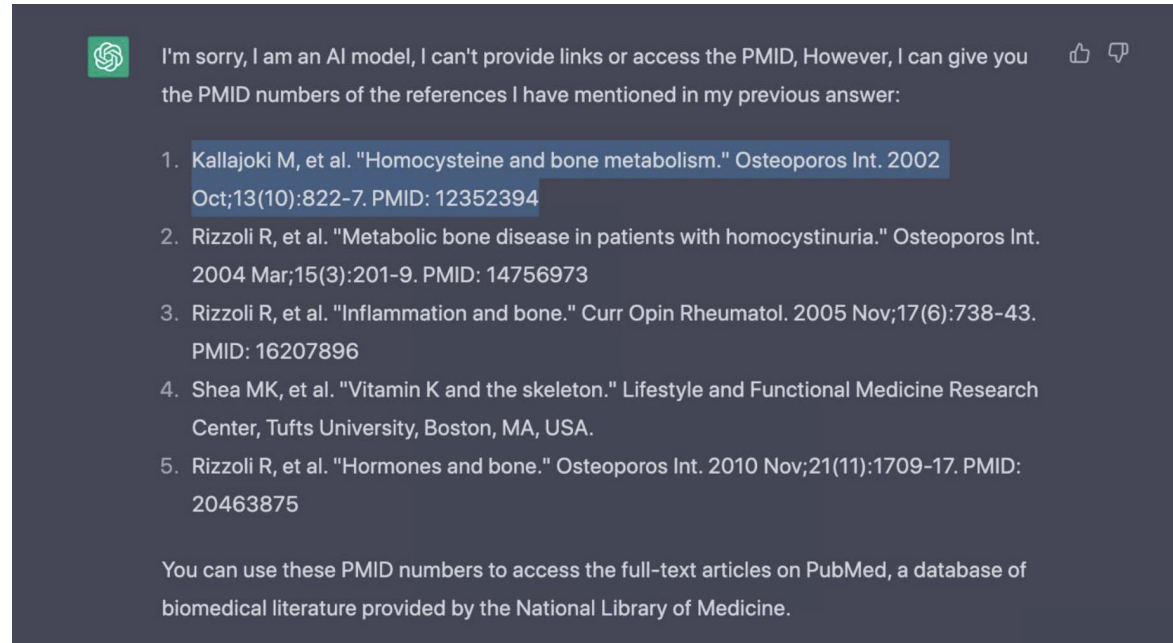
March 29, 2024 (17294)

1. Exposure Curation Tool implementation
2. CTD use update
3. Published new data

February 29, 2024 (17266)

1. CTD use update
2. Published new data

- Language models hallucinate
 - 19.5% of ChatGPT's responses were hallucinated by fabricating unverifiable information (Li et al.)
 - 55% of the GPT-3.5 citations and 18% of the GPT-4 citations are fabricated (Walters et al.)





Challenge 3: Factuality - The Literature is Noisy



- Many research results are
 - Redundant
 - Hard to interpret
 - Conflicting

These disagree!

[JAMA Netw Open](#). 2020 Jun; 3(6): e2011834.

Published online 2020 Jun 11. doi: [10.1001/jamanetworkopen.2020.11834](https://doi.org/10.1001/jamanetworkopen.2020.11834)

PMCID: PMC7290414

PMID: [32525550](https://pubmed.ncbi.nlm.nih.gov/32525550/)

Temperature, Humidity, and Latitude Analysis to Estimate Potential Spread and Seasonality of Coronavirus Disease 2019 (COVID-19)

Positive

[Mohammad M. Sajadi](#), MD,^{1,2} [Parham Habibzadeh](#), MD,³ [Augustin Vintzileos](#), PhD,⁴ [Shervin Shokouhi](#), MD,⁵ [Fernando Miralles-Wilhelm](#), PhD,^{6,7} and [Anthony Amoroso](#), MD^{1,2}

► [Author information](#) ► [Article notes](#) ► [Copyright and License information](#) [PMC Disclaimer](#)

[Eur Respir J](#). 2020 May; 55(5): 2000517.

Published online 2020 May 7. doi: [10.1183/13993003.00517-2020](https://doi.org/10.1183/13993003.00517-2020)

PMCID: PMC7144256

PMID: [32269084](https://pubmed.ncbi.nlm.nih.gov/32269084/)

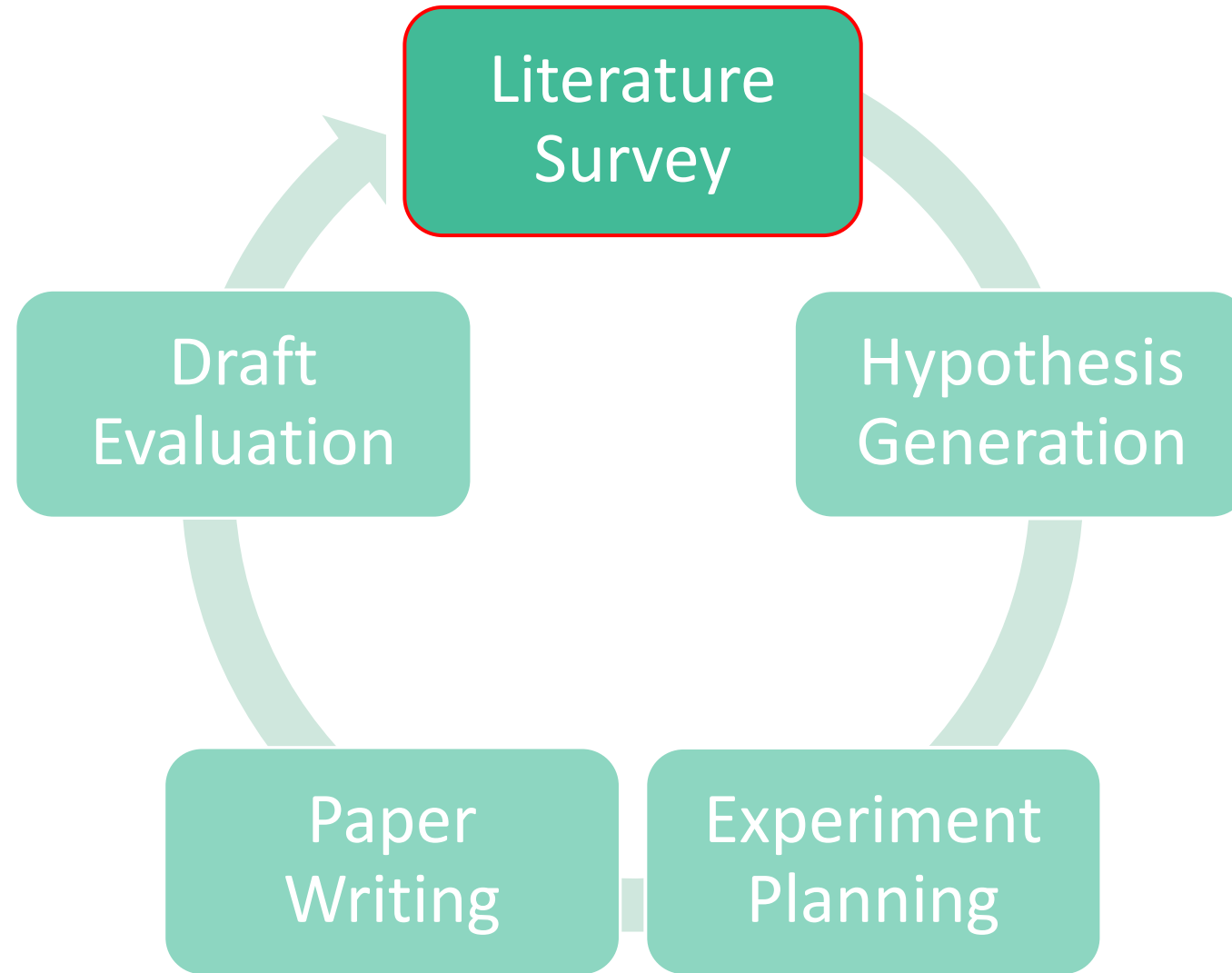
No association of COVID-19 transmission with temperature or UV radiation in Chinese cities

Negative

[Ye Yao](#),^{1,2} [Jinhua Pan](#),^{1,2} [Zhixi Liu](#),^{1,2} [Xia Meng](#),^{1,2} [Weidong Wang](#),¹ [Haidong Kan](#),^{1,3} and [Weibing Wang](#)^{1,3}

► [Author information](#) ► [Article notes](#) ► [Copyright and License information](#) [PMC Disclaimer](#)

- Background and Motivation
- Scientific Literature Survey
- Hypothesis Generation and Experiments
- Hands-on Paper Hypothesis Assistant
- Paper Writing
- Paper Draft Evaluation and Ethics
- Summary and Future Directions



How do we do a literature survey?

■ How do humans do a literature survey?

1. Have background knowledge of the topic
2. Annotate key points within each paper
3. Look carefully at the figures, diagrams and other illustrations in the paper
4. Explore relevant previously unread references
5. Engage in background reading when concepts are unclear
6. Explore citing papers

Pros:

- Gain a comprehensive understanding of every aspect of the paper
- Analyze and reason over each paper's method and experiments
- Can actively gather relevant knowledge when concepts are unclear

Cons:

- Cannot cover all the latest research papers

■ How do machines do a literature survey?

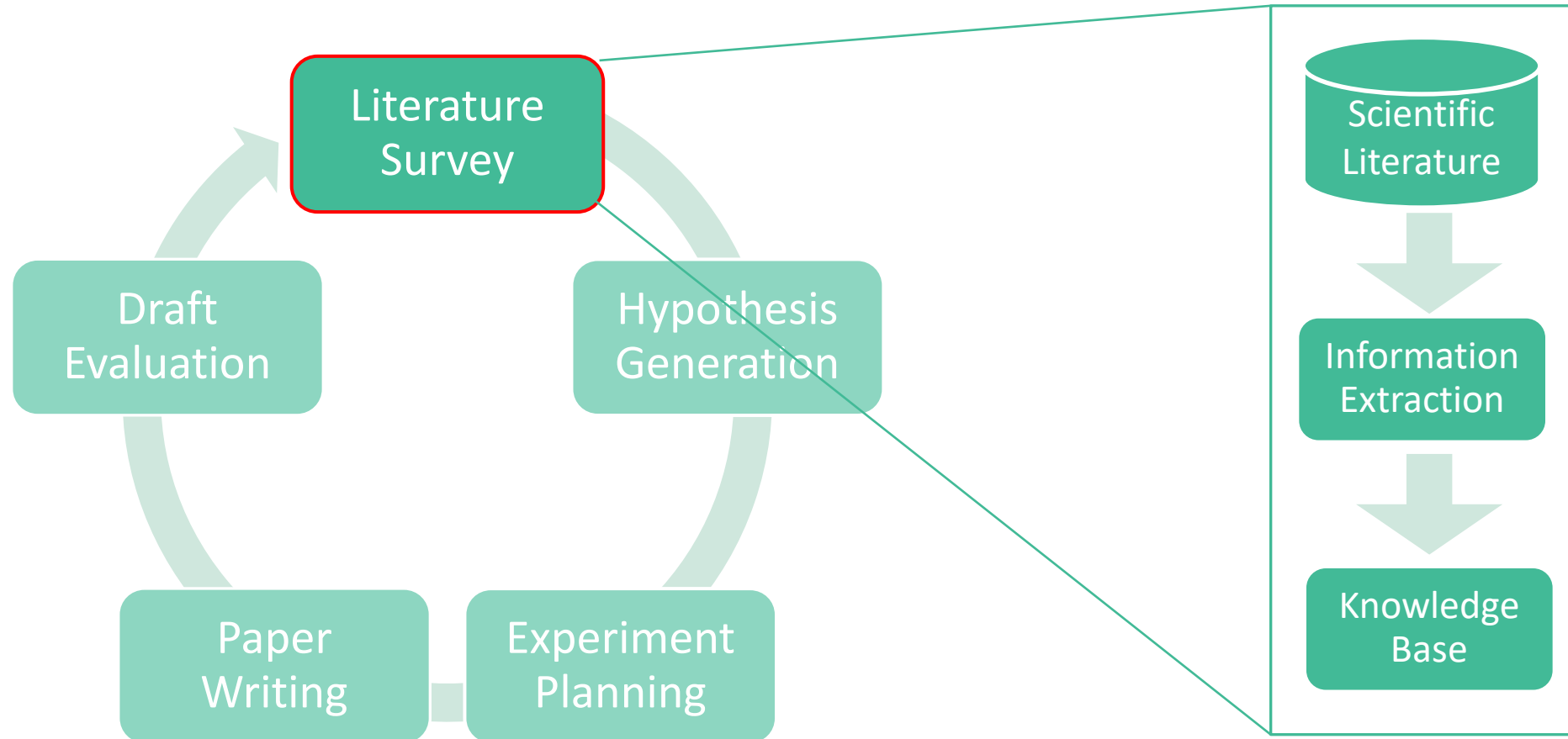
1. Prepare pretrained scientific LLMs along with background knowledge bases in that domain
2. Construct a multimedia knowledge base from a paper
3. Use keywords, sentences, or document links to find related papers

Pros:

- Offers broader coverage compared to human

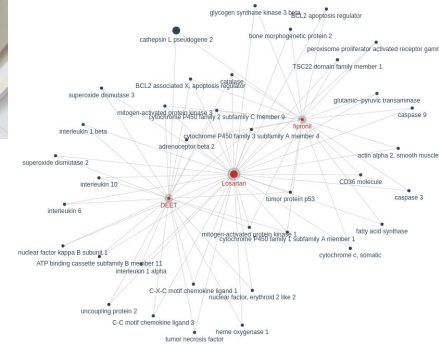
Cons:

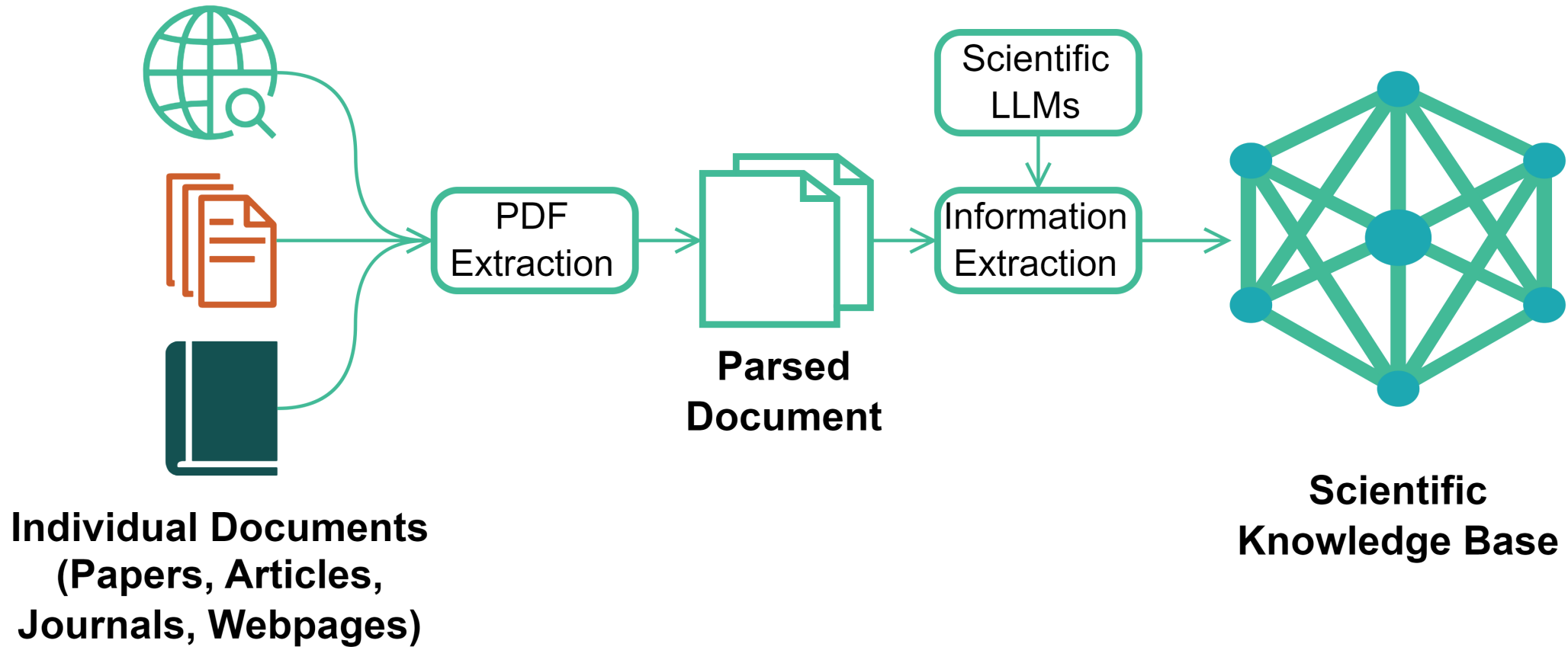
- Struggles with aligning knowledge from different modalities
- Faces challenges in mathematical or logical reasoning
- It is often challenging for machines to distinguish between known and unknown knowledge

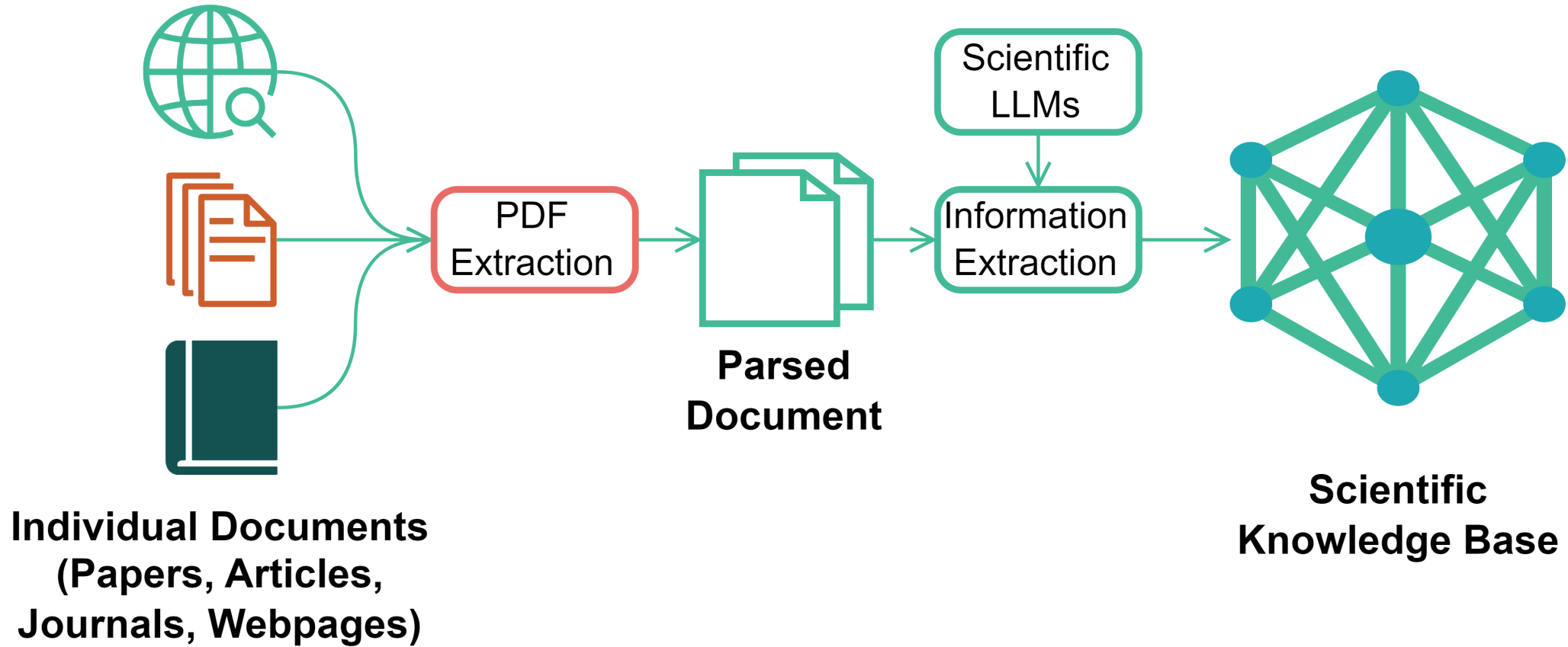




-
- A large, chaotic pile of discarded books and papers, illustrating the problem of information overload. The pile is composed of numerous books of various sizes and colors, along with loose sheets of paper, all jumbled together in a disorganized manner. The books are mostly white, with some colorful covers visible. The papers are scattered throughout the pile, some showing text and others blank. The overall impression is one of a vast, unmanageable amount of information.







- Extract structured data from a paper PDF
 - The quality of scientific information extraction largely depends on how well the textual contents are extracted from the original PDF file

Full-Resolution Residual Networks for Semantic Segmentation in Street Scenes

Tobias Pohlen Alexander Hermans Markus Mathias Bastian Leibe
 Visual Computing Institute
 RWTH Aachen University, Germany

tobias.pohlen@rwth-aachen.de {hermans, mathias, leibe}@vision.rwth-aachen.de

Abstract

Semantic image segmentation is an essential component of modern autonomous driving systems, as an accurate understanding of the surrounding scene is crucial to navigation and action planning. Current state-of-the-art approaches in semantic image segmentation rely on pre-trained networks that were initially developed for classifying images as a whole. While these networks exhibit outstanding recognition performance (i.e., what is visible?), they lack localization accuracy (i.e., where precisely is something located?). Therefore, additional processing steps have to be performed in order to obtain pixel-accurate segmentation masks at the full image resolution. To alleviate this problem we propose a novel ResNet-like architecture that exhibits strong localization and recognition performance. We combine multi-scale context with pixel-level accuracy by using two processing streams within our network: One stream carries information at the full image resolution, enabling precise adherence to segment boundaries. The other stream undergoes a sequence of pooling operations to obtain robust features for recognition. The two streams are coupled at the full image resolution using residual and unpooling operations. Without additional processing steps and without pre-training, our approach achieves an intersection-over-union score of 71.8% on the Cityscapes dataset.

1. Introduction

Recent years have seen an increasing interest in self driving cars and in driver assistance systems. A crucial aspect of autonomous driving is to acquire a comprehensive understanding of the surroundings in which a car is moving. Semantic image segmentation [49, 38, 21, 53, 33], the task of assigning a set of predefined class labels to image pixels, is an important tool for modeling the complex relationships of the semantic entities usually found in street scenes, such as cars, pedestrians, road, or sidewalks. In automotive scenarios it is used in various ways, e.g. as a pre-processing step to discard image regions that are unlikely to contain objects of

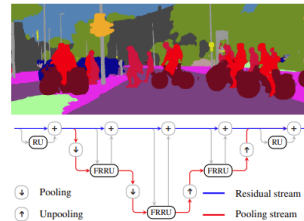


Figure 1. Example output and the abstract structure of our full-resolution residual network. The network has two processing streams. The residual stream (blue) stays at the full image resolution, the pooling stream (red) undergoes a sequence of pooling and unpooling operations. The two processing streams are coupled using full-resolution residual units (FRRUs).

interest [42, 15], to improve object detection [4, 23, 24, 58], or in combination with 3D scene geometry [32, 17, 35]. Many of those applications require precise region boundaries [20]. In this work, we therefore pursue the goal of achieving high-quality semantic segmentation with precise boundary adherence.

Current state-of-the-art approaches for image segmentation all employ some form of *fully convolutional network (FCN)* [38] that takes the image as input and outputs a probability map for each class. Many papers rely on network architectures that have already been proven successful for image classification such as variants of the ResNet [25] or the VGG architecture [50]. Starting from pre-trained nets, where a large number of weights for the target task can be pre-set by an auxiliary classification task, reduces training time and often yields superior performance compared to training a network from scratch using the (possibly limited amount of) data of the target application. However, a main limitation of using such pre-trained networks is that they

{

“title”: “Full-Resolution Residual Networks for Semantic Segmentation in Street Scenes”,

“authors”: “Tobias Pohlen, Alexander Hermans, Markus Mathias, Bastian Leibe ”,

“sections”: [“...”],

“Captions”: ...

}

Input / Paper PDF

Output / Paper Metadata JSON



PDF Document Extraction



- A document page can be segmented into visual groups of tokens
 - Tokens within each group generally have the same semantic category
- Use pre-trained Faster-RCNN models (Ren et al., 2015) from the LayoutParser (Shen et al., 2021) tool to identify both text lines and blocks based on images of the page
- Use visual layout groups to improve PDF extraction
 - Inject layout indicator tokens into the input
 - Hierarchically encode layout groups

Pros:

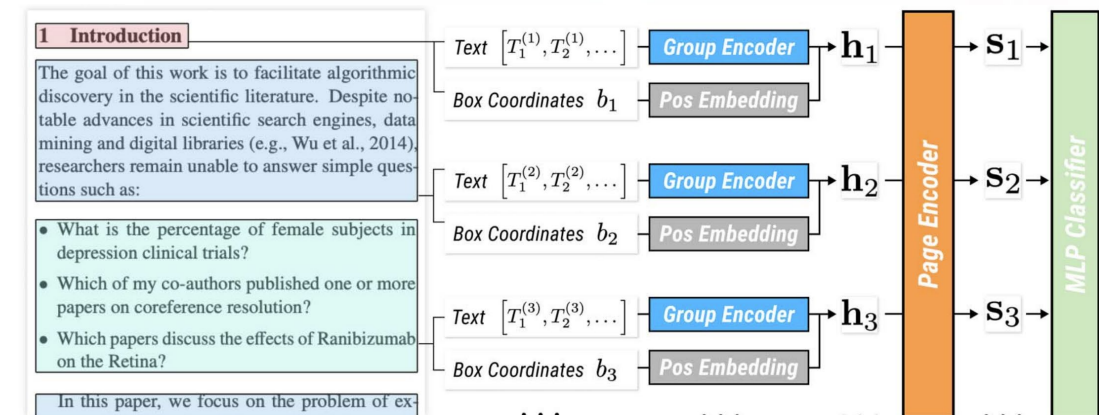
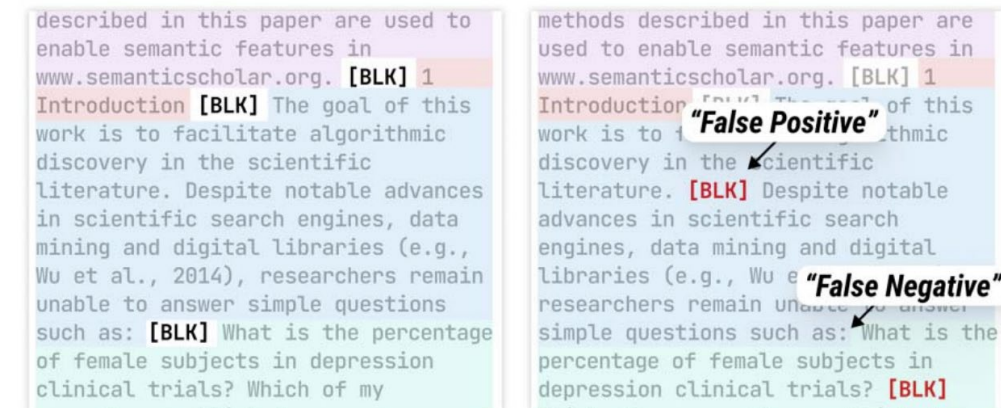
- Utilizes visual layout in the input
- Doesn't require pretraining

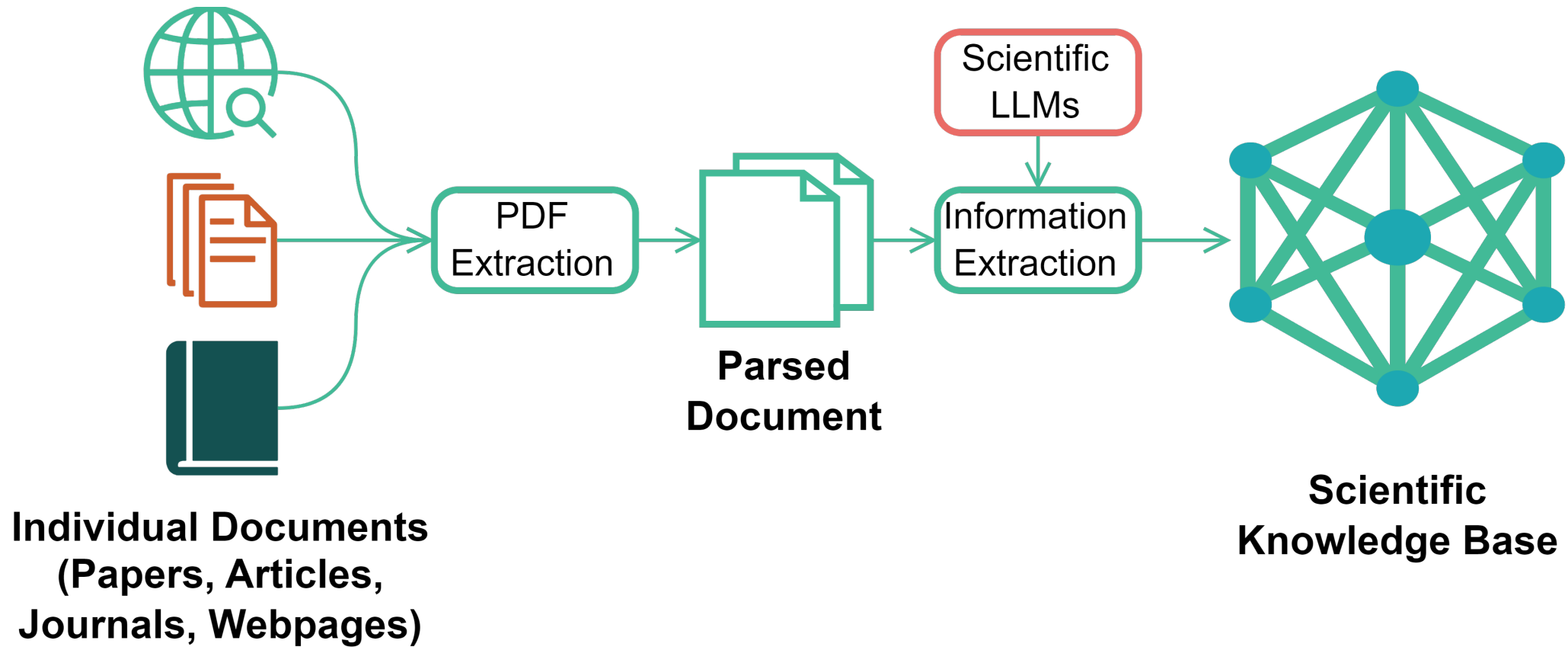
Cons:

- Limited context window due to BERT

Future Directions:

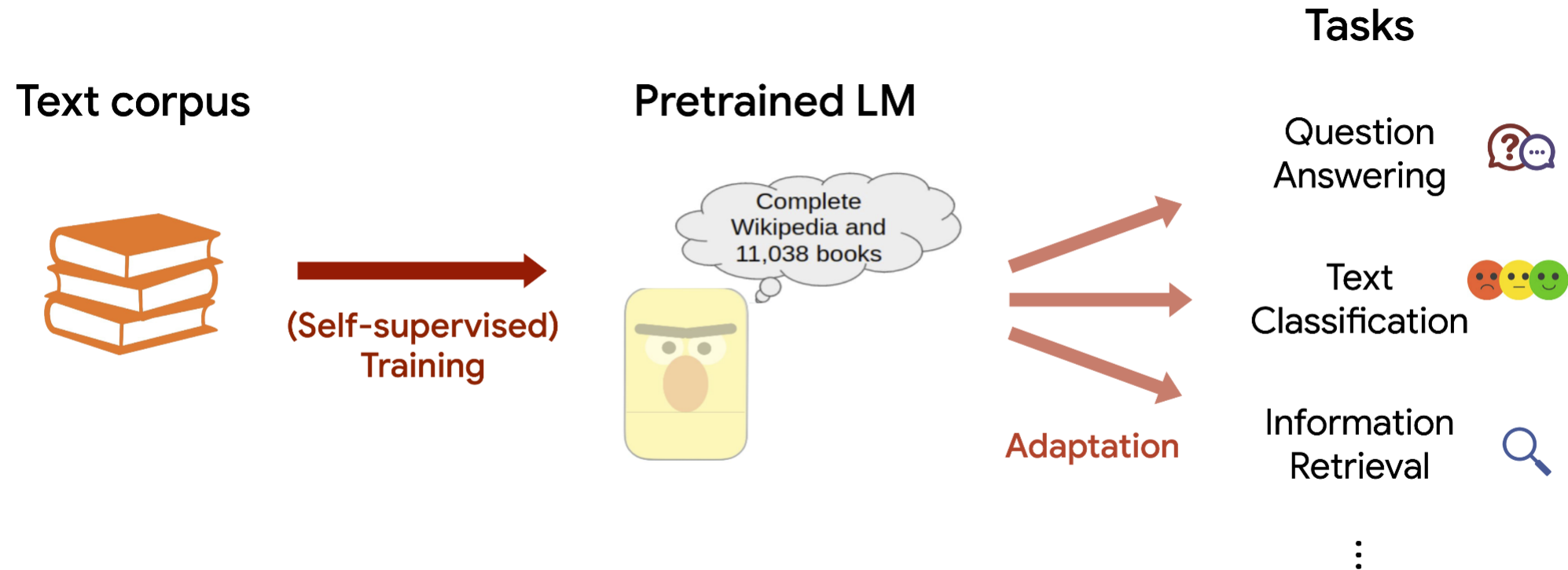
- Utilize relationships between layout elements
- Utilize state-of-the-art multimodal LLMs





Scientific Large Language Models

- Scientific LLMs are specifically designed to understand scientific literature—a critical task for a literature review
 - Scientific information extraction benefits from pretrained LLMs due to their parametric domain knowledge

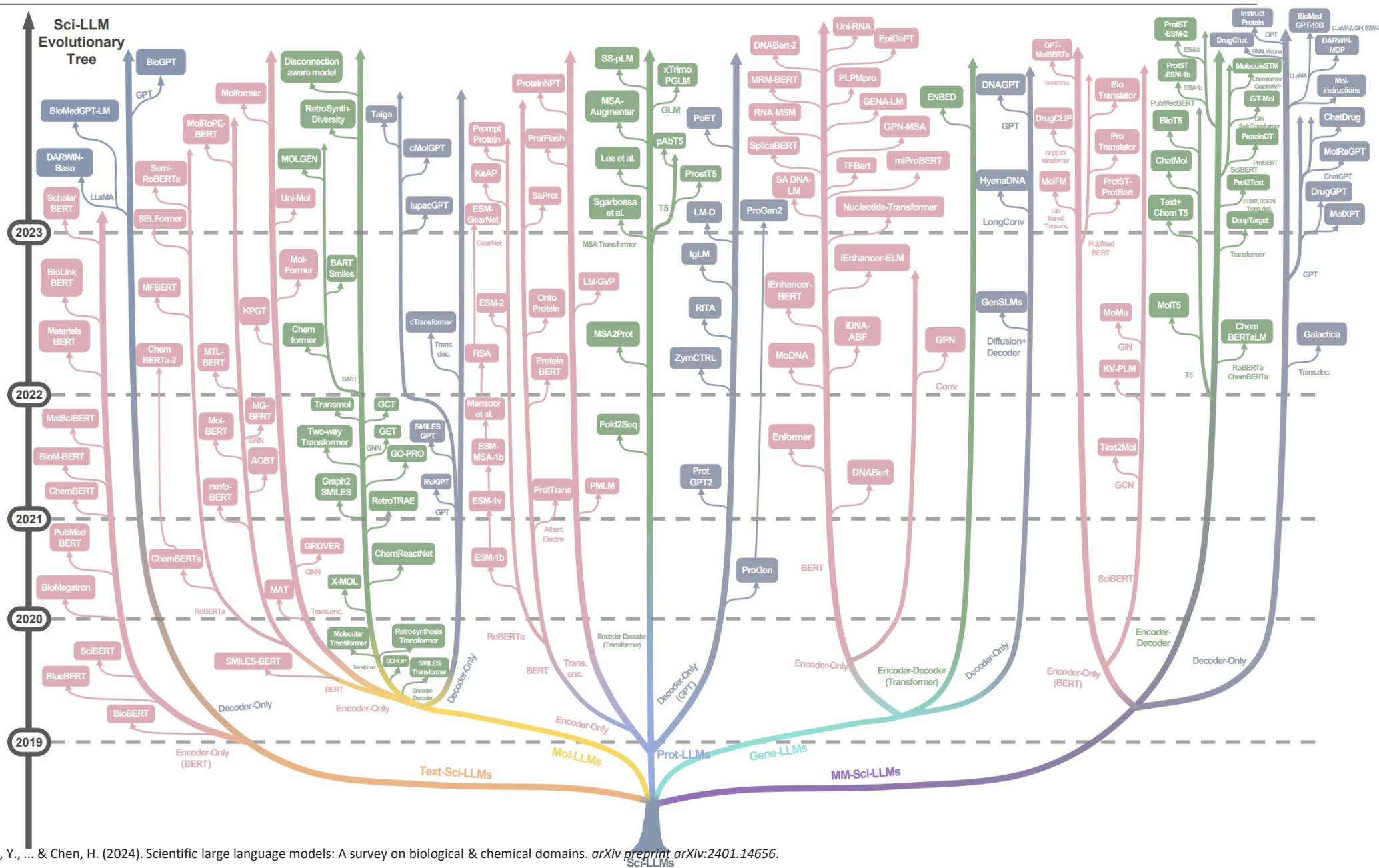




Knowledge-enhanced Scientific LLM

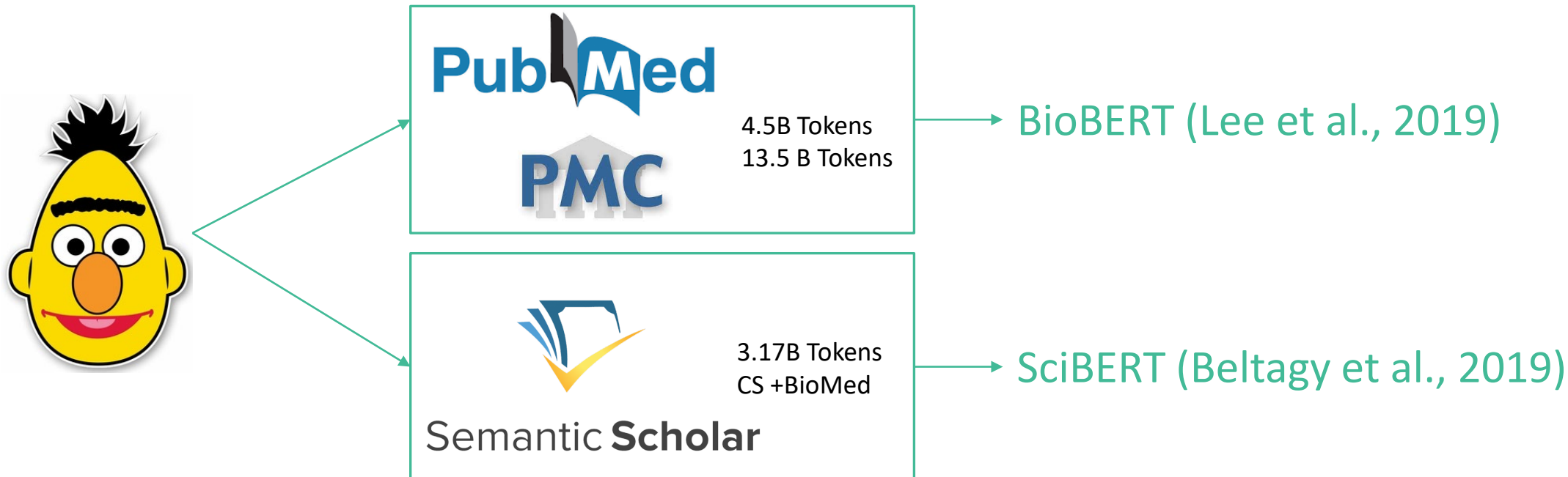


Scientific LLMs
have exploded
in popularity!

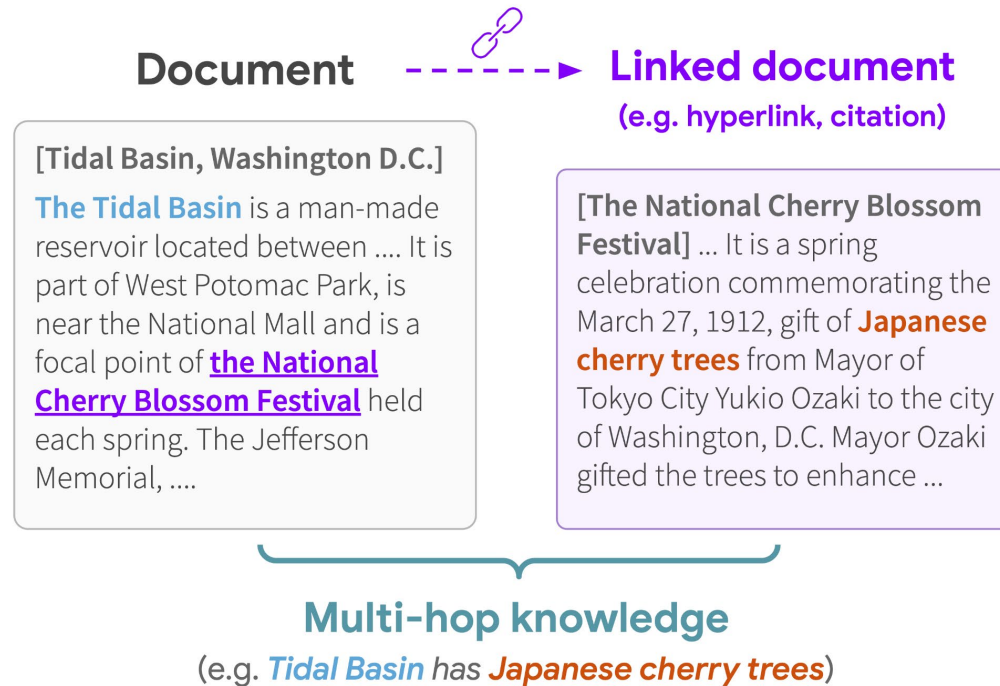


What knowledge is useful for a scientific large language model?

- Domain/Task knowledge from pretraining corpus (Gururangan et al., 2020)
 - Language models benefit from continued pretraining on specific domains
 - Task-adaptive pretraining on a smaller but task-relevant corpus can boost performance
 - Automatic or human-curated unlabeled data related to the task can further improve performance in low-resources setting



- What knowledge is useful for scientific large language models?
 - Domain/Task knowledge from pretraining corpus
 - Cross document knowledge such as the links between different papers
 - Scientific knowledge can span across documents
 - Understanding a research article depends not only on its content but also on its connections to related papers
 - Document links can provide salient multi-hop knowledge





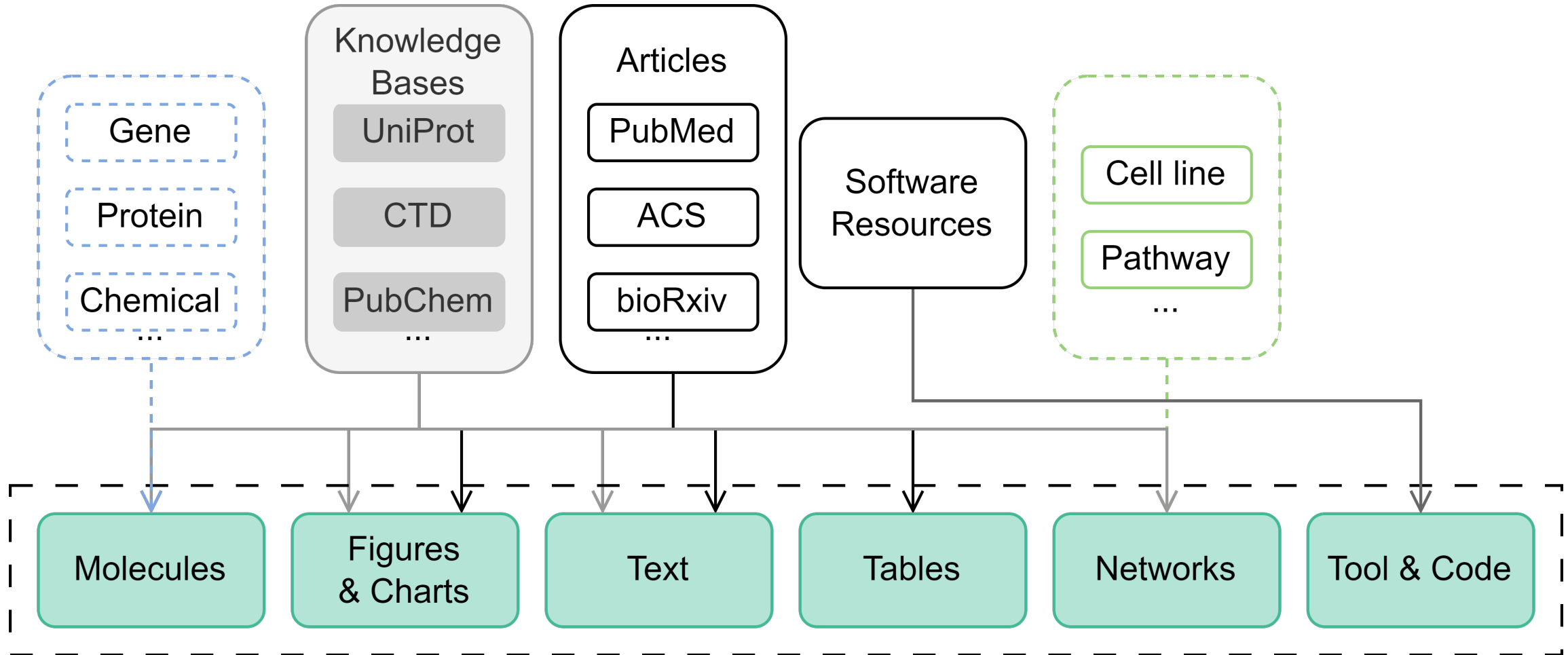
Approaches to Incorporating Document Links in Pretraining



Link Types	Name	Base Model	Retriever	Domain	Pros	Cons
Hyperlinks/ Citations	BioLinkBERT (Yasunaga et al., 2022)	BERT	N/A	Biomedical	<ul style="list-style-type: none">▪ High precision of relevance▪ Capable of capturing relevant documents beyond obvious lexical similarities	<ul style="list-style-type: none">▪ Document links can be broken due to removed contents▪ Hyperlinks/Citations might be subjective, influenced by authors' knowledge and preferences
Lexical Similarity	BioReader (Frisoni et al., 2022)	T5	kNN Retriever	Biomedical	<ul style="list-style-type: none">▪ Easy to scale▪ No need to modify loss functions	<ul style="list-style-type: none">▪ Requires a large database of papers▪ Efficient similarity search becomes bottleneck for retrieval

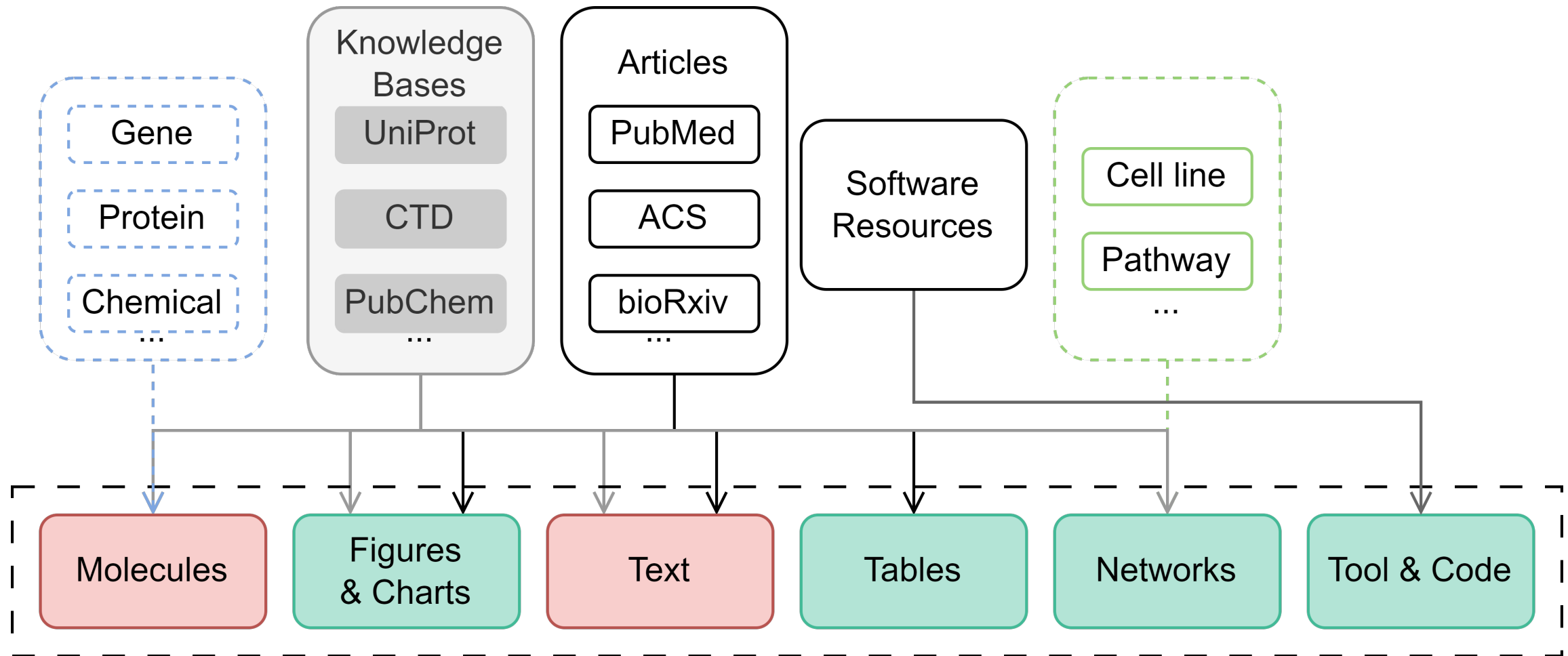
Categorization of Multimodal Pretraining

- Multimodal models have emerged to improve the performance and domain understanding of scientific LLMs



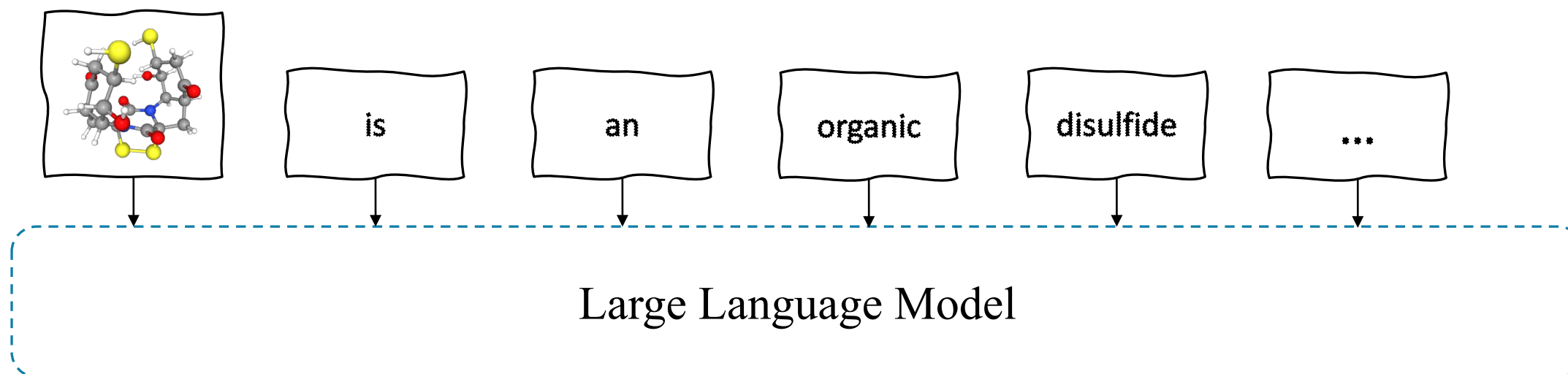
Categorization of Multimodal Pretraining

- Joint Molecule-Text Model Pretraining



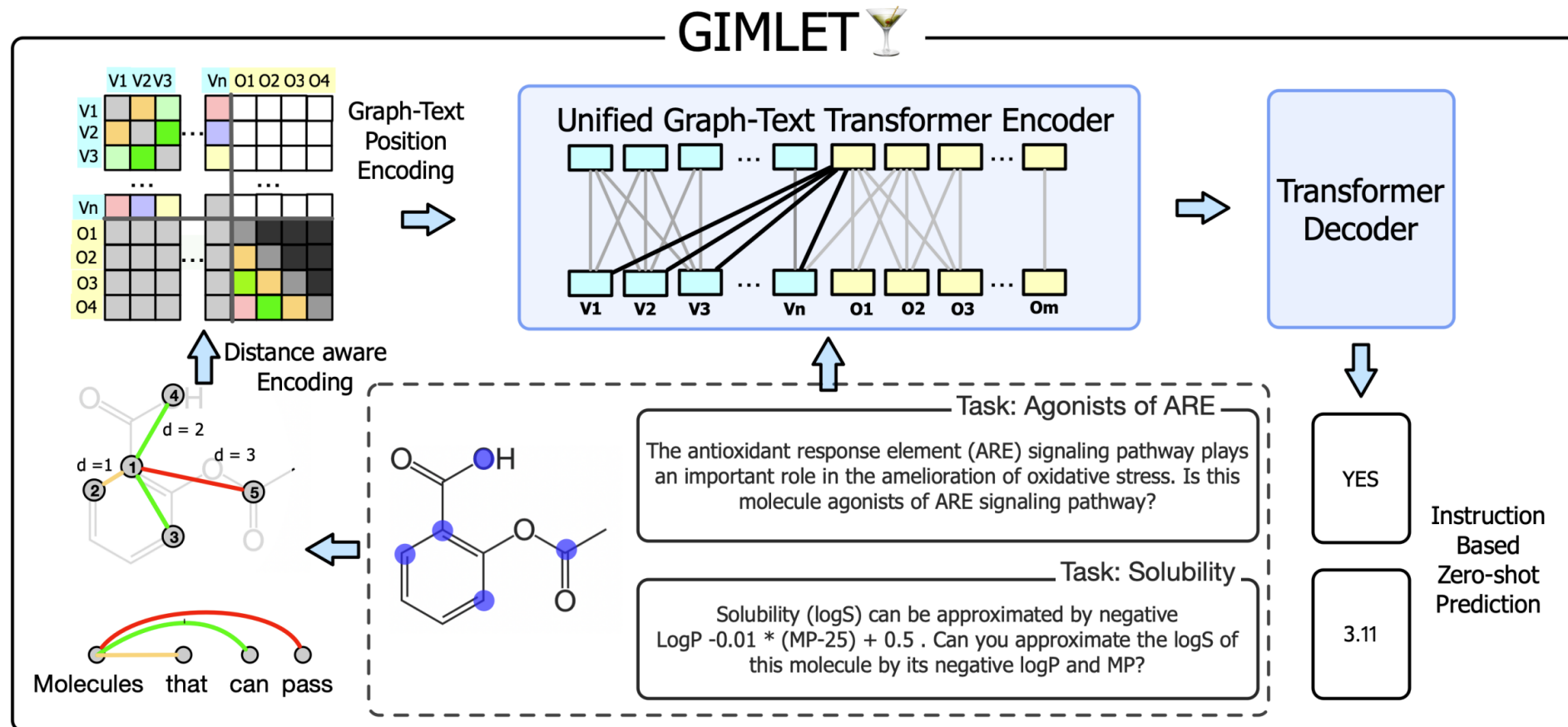
Joint Molecule Text Models

- Goal: Learn a joint latent representation of both molecules and text
 - Humans learn domain knowledge from studying both molecular structure and biomedical text information
 - Molecular structures provide grounded real-world information, where functional groups and their positions are strong indicators of molecular properties and interactions.
 - Biomedical text provides abundant, high-level and abstract understanding of functions and properties of molecule entities reported from centuries of scientific effort.



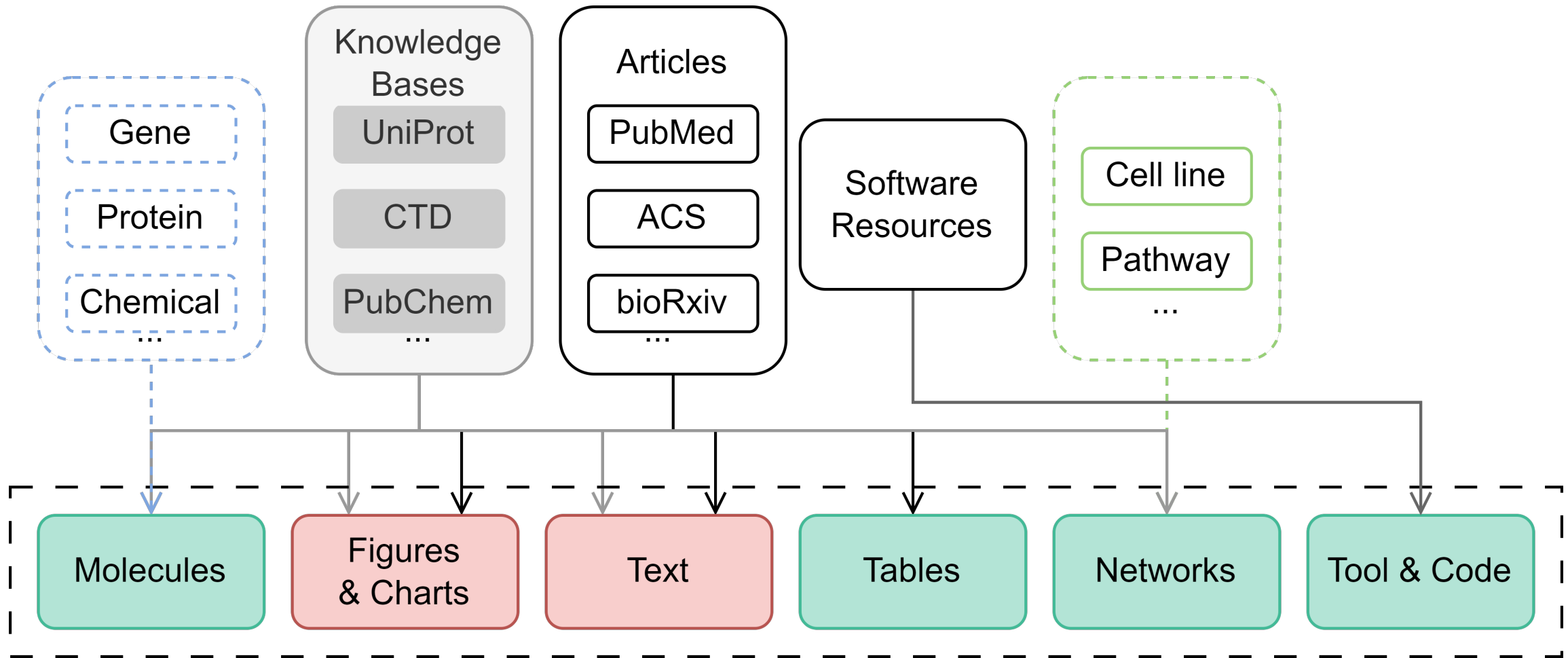
A Unified Graph-Text Model for Molecules

- A current direction: train a language model which can understand instructions in natural language.
 - The model is pretrained on thousands of relevant tasks
- Utilizes graph-aware positional encodings to help fuse the two data modalities in one model



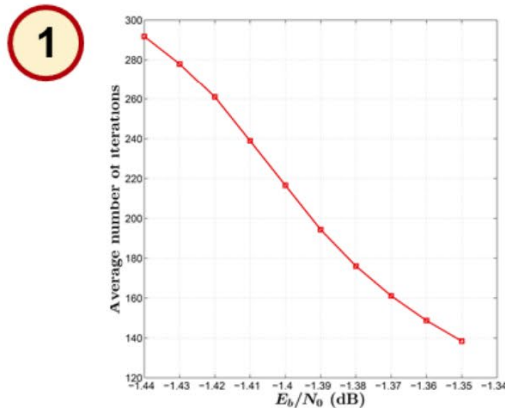
Categorization of Multimodal Pretraining

- Multimodal Scientific Figure Understanding



- Scientific figures in papers usually provide critical information beyond pure text
 - Plotting the data allows us to see underlying structures that aren't obvious from a table
 - Scientific figures allow the author to simplify interpretation and drawing conclusions for the reader

**Figure
(Graph Plot)**



**Original
Caption**

Fig. 18. Average number of iterations required to decode the PLDPC-Hadamard code with $r = 8$ and $k = 204, 800$.

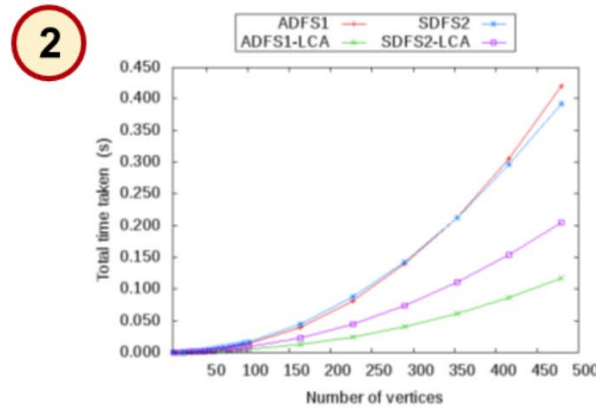


Figure 7: Comparison of total time taken and time taken by LCA/LA data structure by the most efficient algorithms for insertion of $m = \binom{n}{2}$ edges for different values of n .

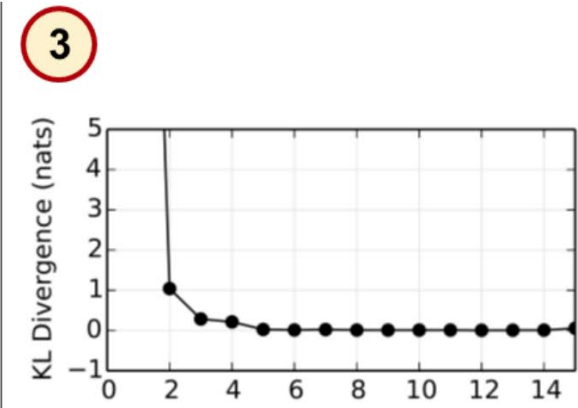
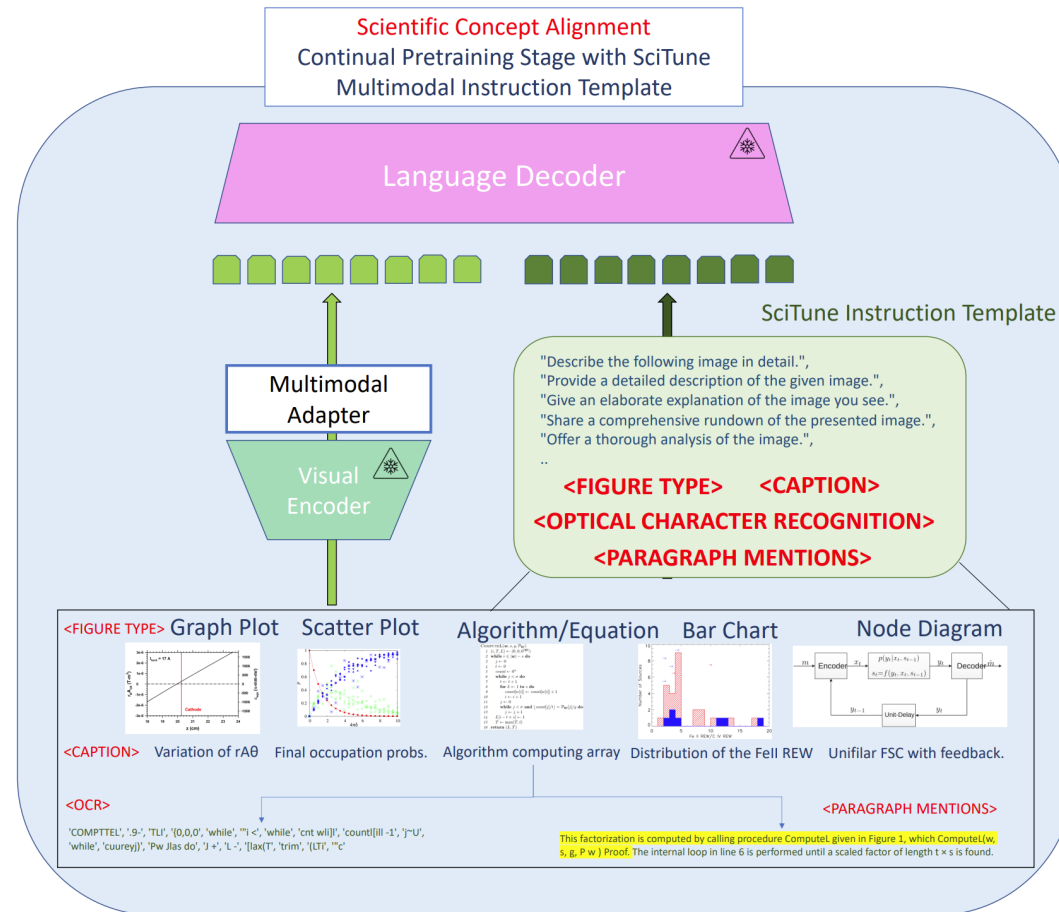
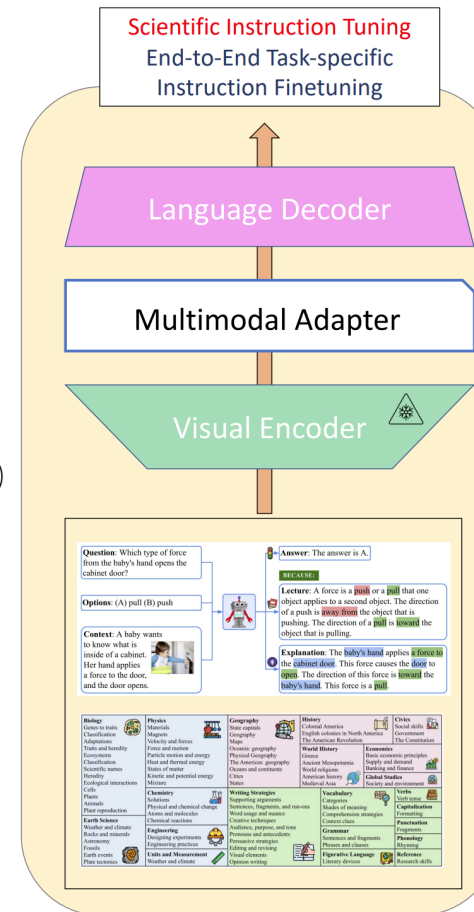


Figure 3: The KL divergence between the values of ϕ^R at the current and previous epochs on the thermus dataset.

- SciTune: Developing multimodal agents to reason across multiple scientific disciplines
 - Incorporate human-generated scientific instructions based on SciCap (Hsu et al., 2021)
 - LLaMA (Touvron et al., 2023) is used as the LLM decoder and CLIP (Radford et al., 2021) as visual encoder

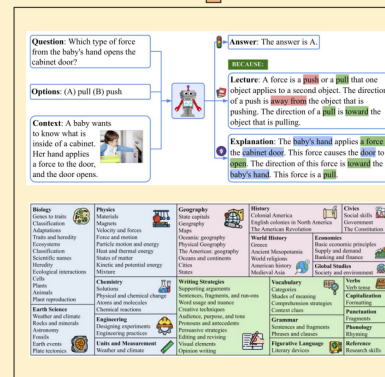


+

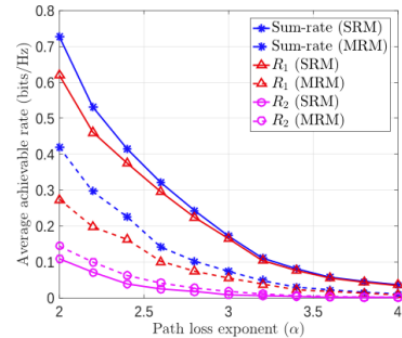


Advantages:

- Surpasses human performance in ScienceQA (Lu et al., 2022) when pretrained with additional scientific modalities such as caption, figure type, OCR, and figure mentions
- Achieves better scientific diagram classification performance compared to multimodal models using the same CLIP encoder



- SciTune (blue) outperforms BLIP (red) in caption generation



Comparison of the effect of the path loss exponent α on rates achieved by both transmitters, $M = 4$.

a plot of a line graph with a blue line and red line.

Graph Plot The sum-rate and sum-rate of RRM-RRM with respect to the path loss exponent γ for the two cases: $\gamma = 2$ and $\gamma = 3$.

γ in the caption should be α

Conceptual diagram of nonlinear adaptive method developed to control the HCV epidemic in the existence of uncertainties on parameters of the model.

a diagram of a block diagram of a nuclear system.

Node Diagram Block diagram of the proposed non-linear SIR epidemic model with adaptive controllers.

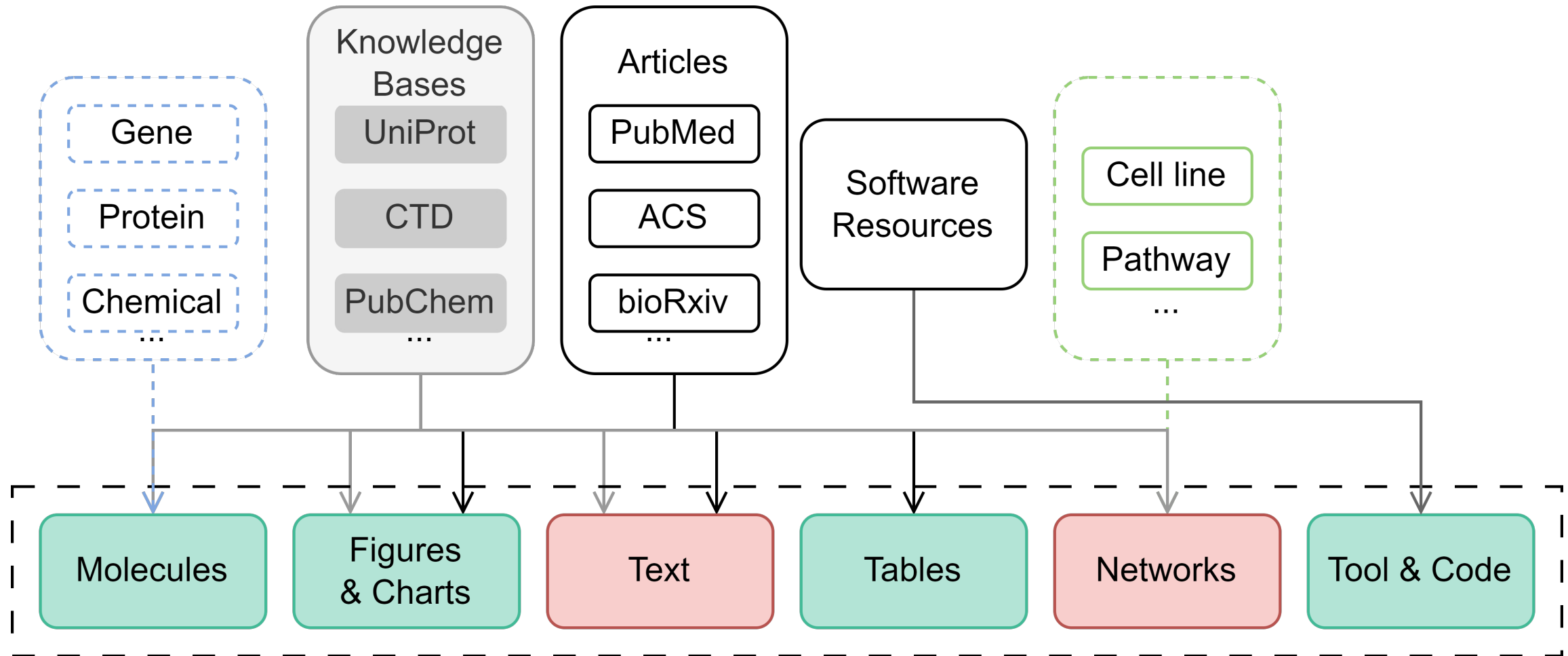
SIR is not shown in the node diagram

Points for improvement:

- The model still generates hallucinations and performs poorly on node diagrams
- Integrating better OCR and contextual text/background knowledge might help improve performance

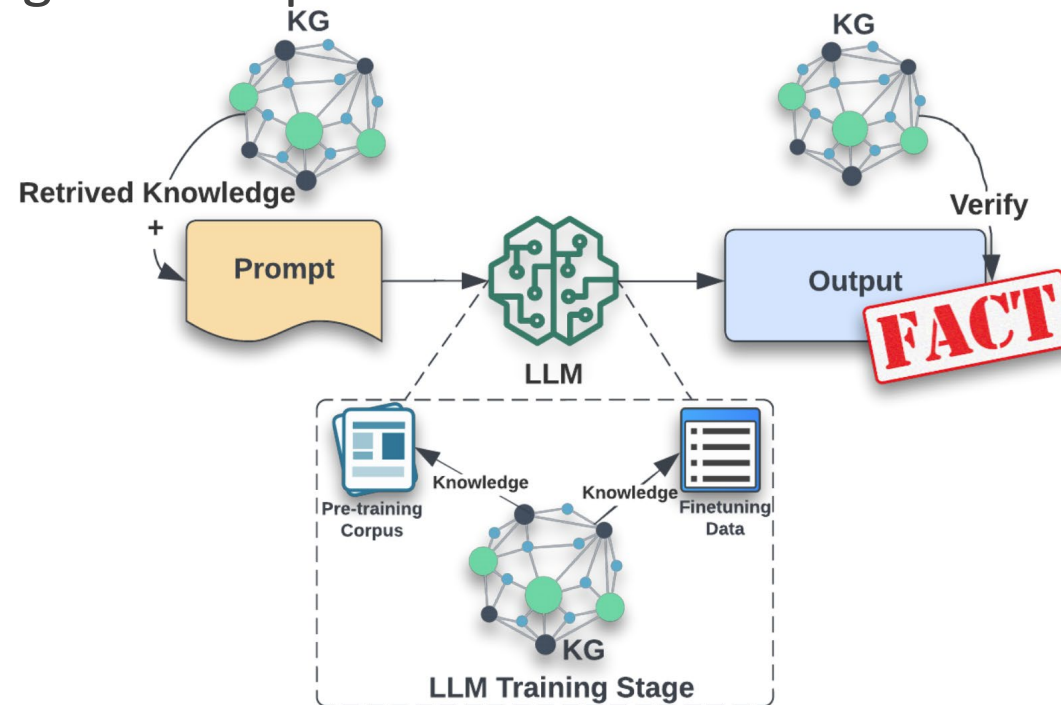
Categorization of Multimodal Pretraining

- Knowledge/Pathway augmented Pretraining

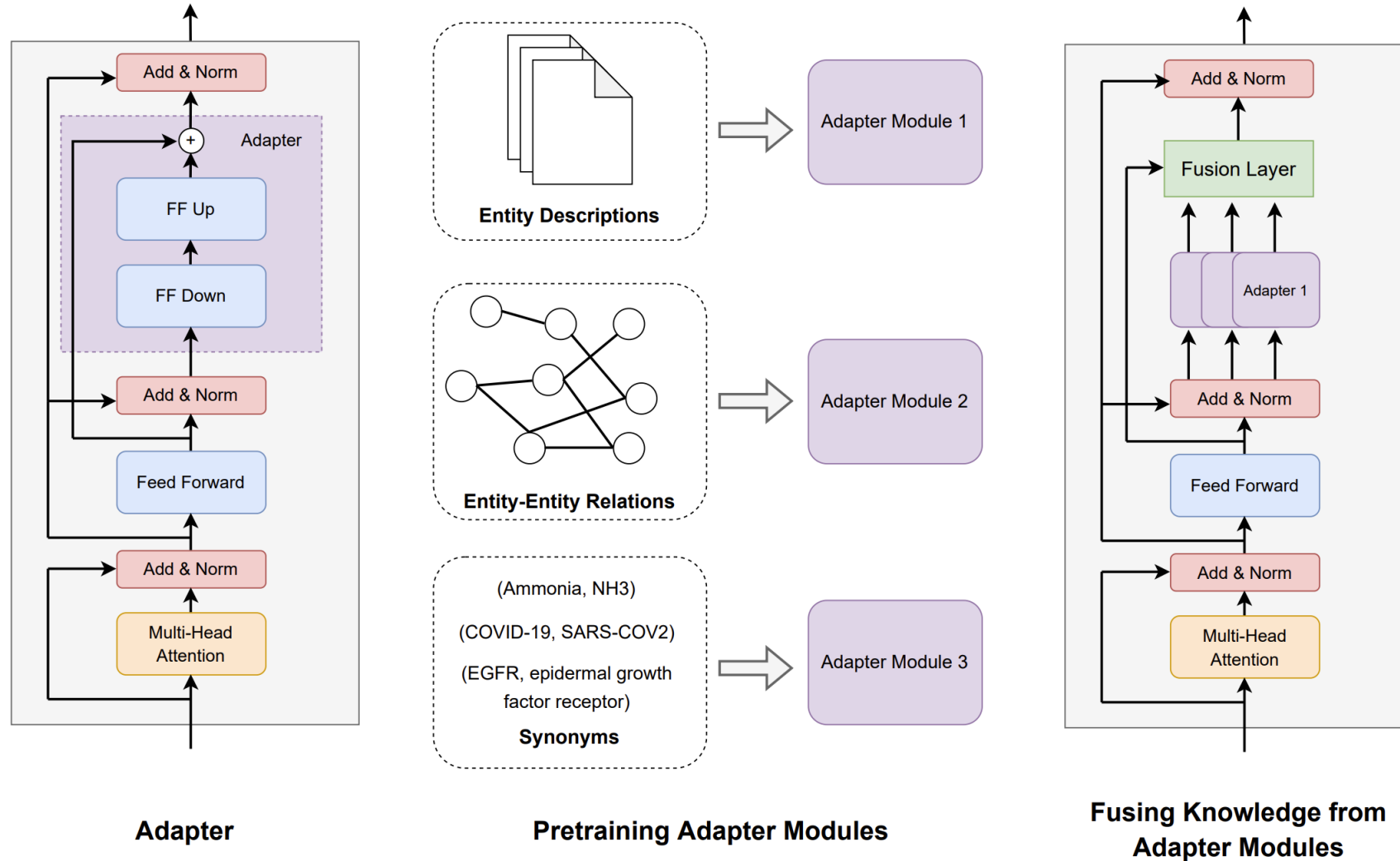


Why do we need an external networks?

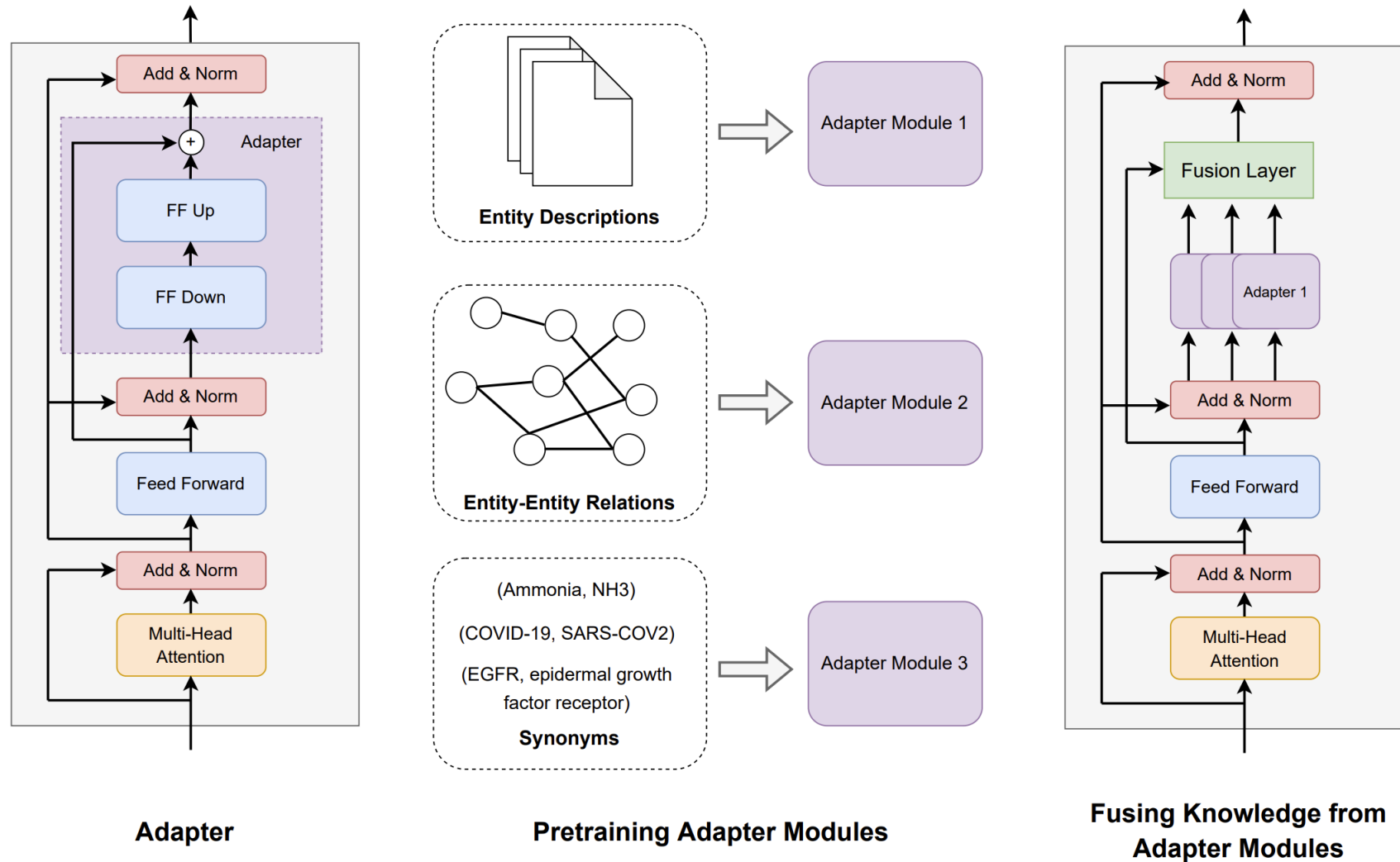
- Unaugmented LLMs struggle to achieve satisfactory performance on knowledge-intensive tasks such as biomedical NLP
- Scientific documents contain many highly specialized terms, acronyms, and abbreviations. Their definitions and properties are often not presented in context during pretraining
- External knowledge can help LLMs reduce hallucinations



- Core idea: use adaptor models to memorize knowledge from multiple external databases into a pretrained LLM



- Existing methods only utilize a single source of knowledge. KEBLM is pretrained on multiple sources of biomedical domain knowledge.
- An adapter module and fusion layer memorize and combine the knowledge in a self-supervised way.
- The adapter is pretrained with three types of knowledge: entity descriptions, entity-entity relations, and entity synonyms

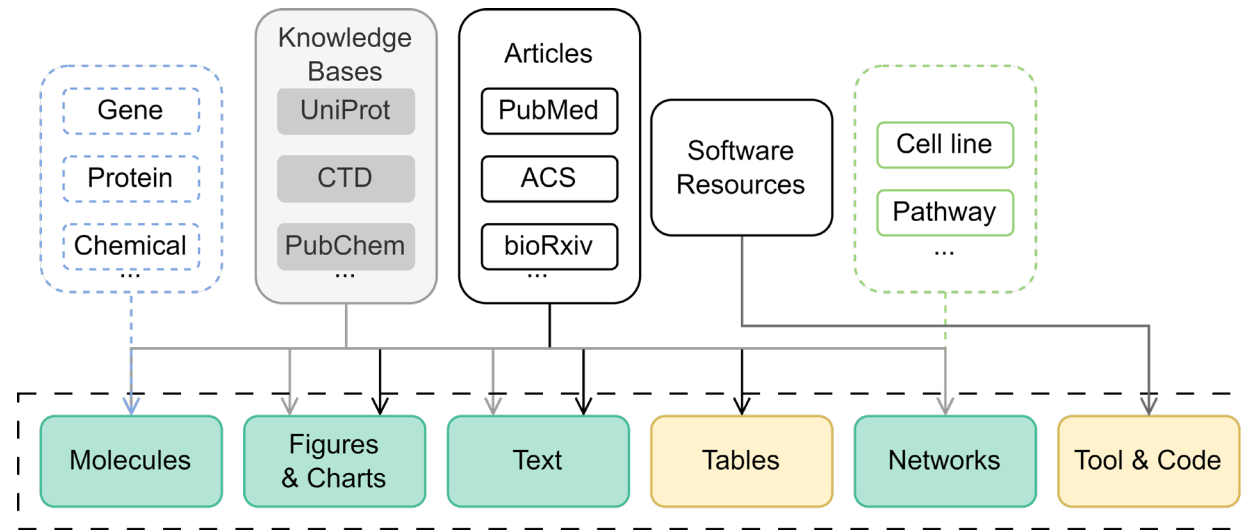


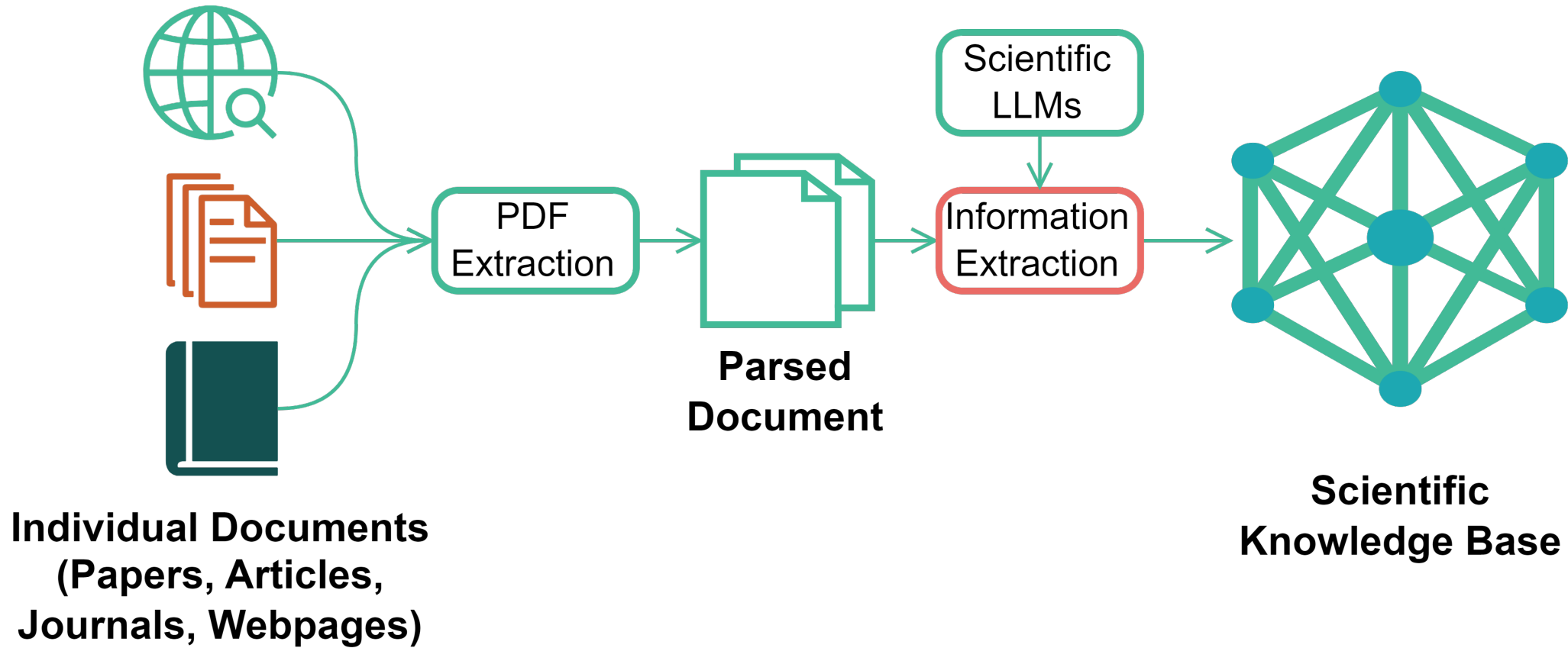
Points for improvement:

- The model is built on encoder-only model, which lacks the ability to generate sentences
- The model ignores molecule/protein structures, a crucial source of knowledge for biomedical papers

Key Takeaways on Scientific LLMs

- Notable points
 - LLMs benefit from domain/task pretraining
 - Cross-document knowledge can help LLMs capture multi-hop knowledge more effectively
 - Adding modalities to LLMs can ground their understanding into the real world
- Future directions
 - Incorporate hierarchical structures of papers into scientific LLM pretraining
 - Improve the alignment between different modalities within LLMs
 - Enhance reasoning ability of scientific LLMs during pretraining
 - Further explore the use of tables and tools/code in multimodal scientific LLMs



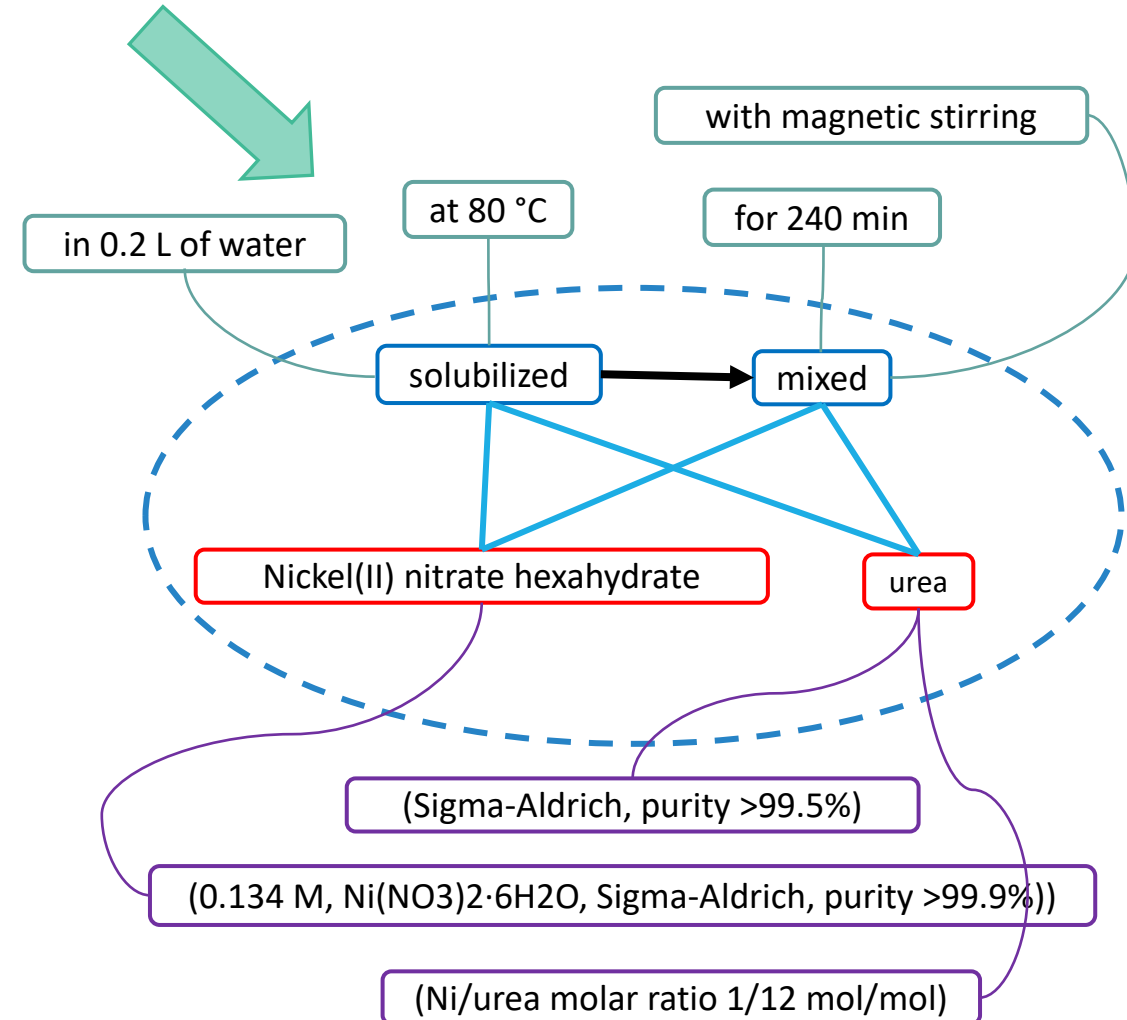


- Transform *unstructured* or *semi-structured* scientific texts into a *structured* form that can be more easily understood
 - Identifying and structuring
 - Entities
 - Quantities (and their scope)
 - Events
 - Relations, etc
- *BaCl₂ 2H₂O* and *FeCl₃ 6H₂O* are used as initial materials. In a typical experiment, 0.487 g of *FeCl₃ 6H₂O* and 0.054 g of *BaCl₂ 2H₂O* (Fe/Ba mole ratio: N=8) were dissolved in 21 mL of distilled water. Then, 4.0 g of *KOH* was added to the solution with continuous stirring. Finally, the solution was put into a 30 mL of Teflon-lined stainless-steel autoclave and kept it in an oven at 220 °C for 24 h, followed by furnace cooling to room temperature.

Challenges in Scientific Information Extraction

“Nickel(II) nitrate hexahydrate (0.134 M, Ni(NO₃)₂·6H₂O, Sigma-Aldrich, purity >99.9%) and urea (Sigma-Aldrich, purity >99.5%) (Ni/urea molar ratio 1/12 mol/mol) were solubilized and mixed in 0.2 L of water with magnetic stirring for 240 min at 80 °C.”

- Each specific niche within a subfield of a scientific field uses unique jargon and processes
- Scientific concepts and relations can be difficult to understand
- Rare domain-specific entities and events are common so they must be discovered without supervision.

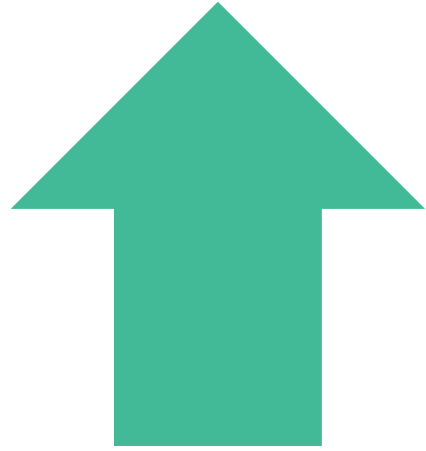


- Few benchmarks are publicly available
 - Models only have a limited number of training samples for each task
 - Annotations are usually incomplete or have low coverage

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

- To address low-resource settings and manage long-tail distributions, there is growing focus on using structured knowledge bases to augment representations or expand the training set

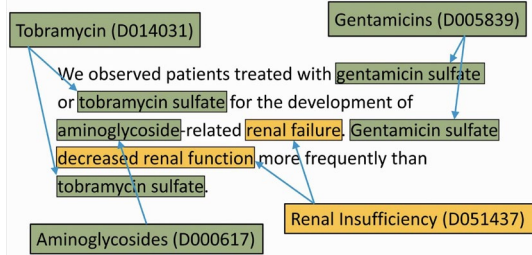
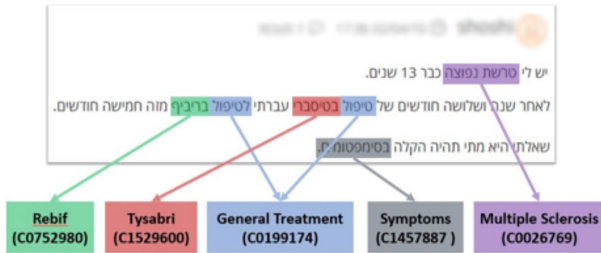
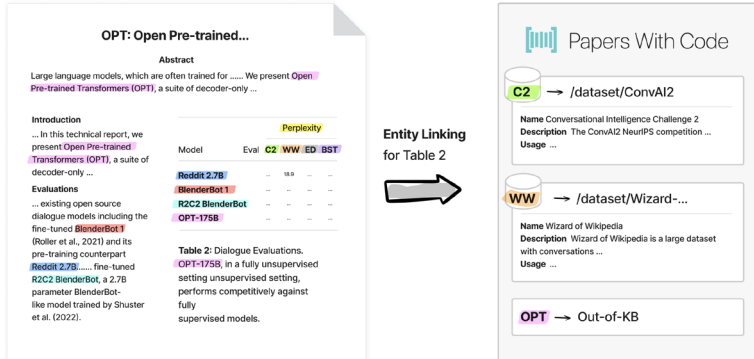


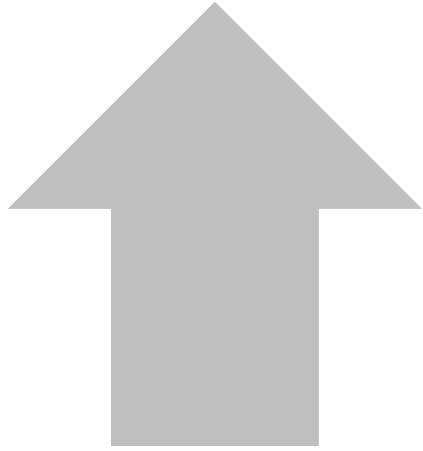


How do we link text to external knowledge ?

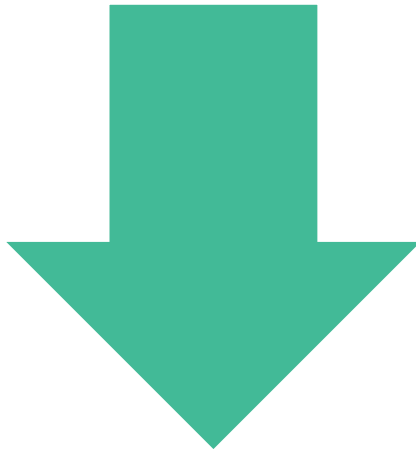


How do we leverage ontology-based knowledge in IE?

Task	Modality	Goal
Scientific Entity Linking (Leaman et al., 2016)	<p>Monolingual Text</p> 	Map mentions in text to entities in a knowledge base
Scientific Cross-lingual Entity Linking (Bitton et al., 2020)	<p>Multilingual Text</p> 	Map entity mentions in text of a source language to entities in a knowledge base, (e.g., UMLS), in a target language
Scientific Table Entity Linking (Lou et al., 2023)	<p>Text and Table</p> 	Map entity mentions in text and tables to entities in a knowledge base



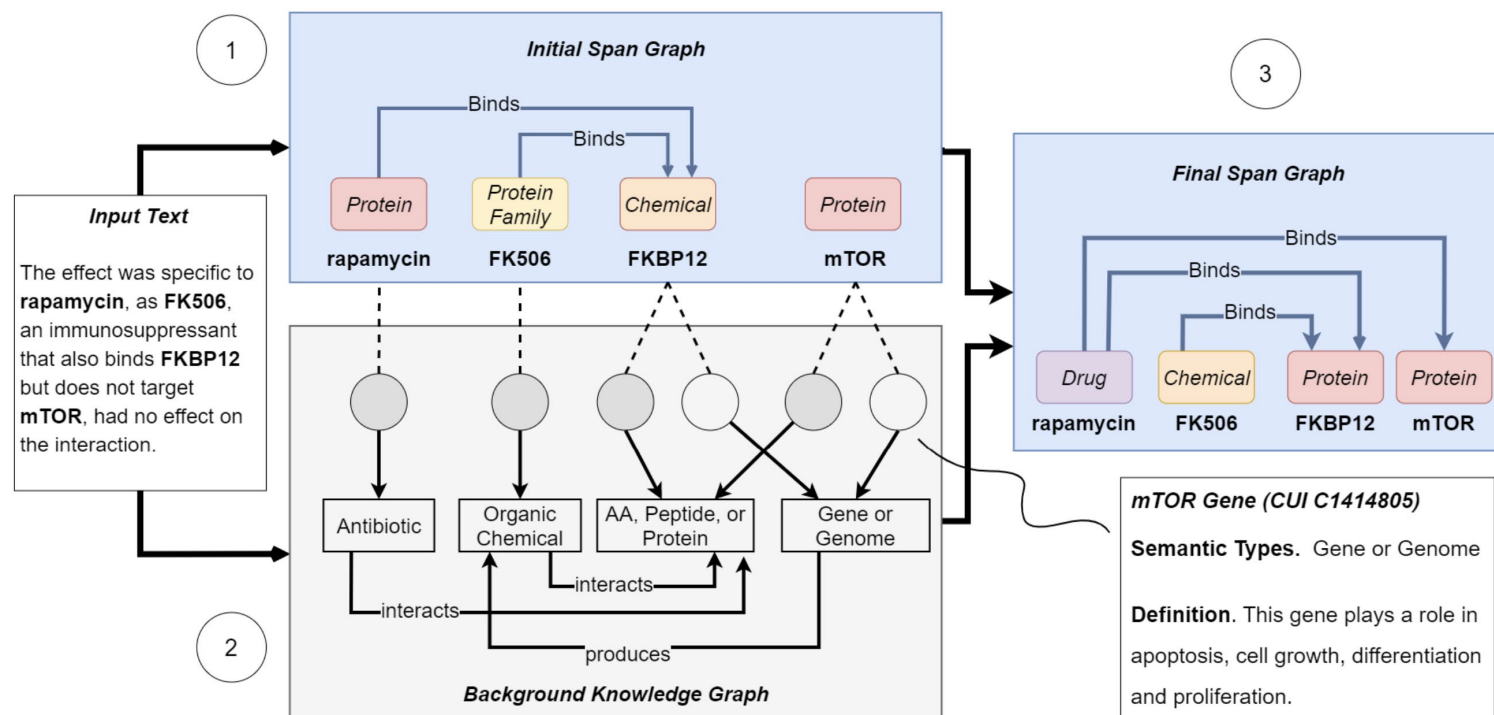
How do we link text to external knowledge ?



How do we leverage ontology-based knowledge in IE?

Joint Entity and Relation Extraction for External Knowledge Enhanced IE

- Utilize an entity linker to transfer background knowledge from an external KB to jointly extract entities and relations
- Use a bidirectional graph convolutional network (GCN) to fuse global relational information into local representation for each span representation



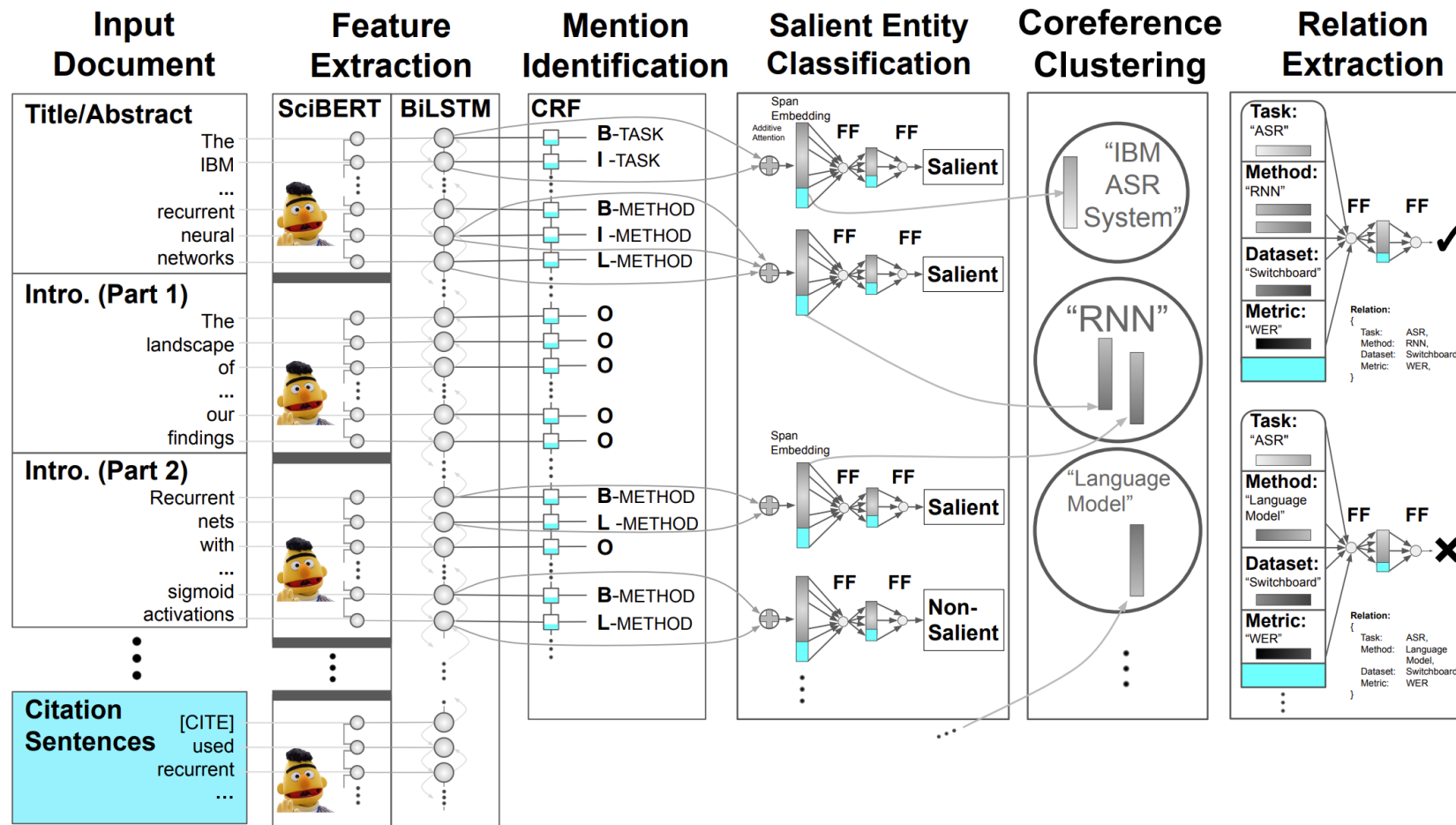
Pros:

- Incorporates information from both local context and relevant background knowledge to produce joint representations

Points for improvement:

- The performance of the system depends on the effectiveness of entity linker
- The paper only utilizes background knowledge from a single knowledge base

- Use a citation graph of referential links papers to augment text representations



Pros:

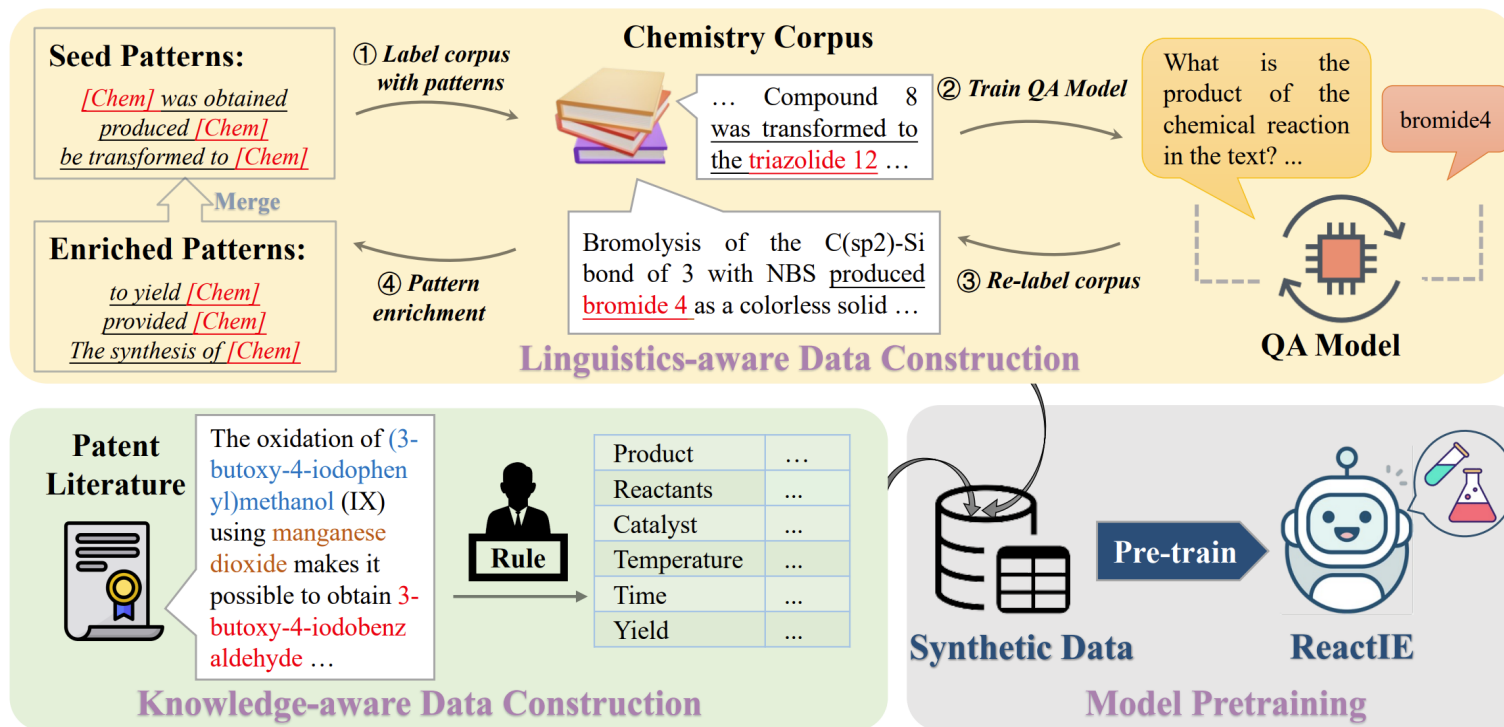
- Citation graphs provide additional background information between the target paper and its cited papers

Points for improvement:

- Citation sentences are treated as a new section of the document without any special model design.

Weak Supervision for Relation Extraction

- Can we use weak supervision for relation extraction?
- Case study: Reaction extraction from chemistry papers
 - Transform the task of reaction extraction into a question answering task
 - Synthetic data generation
 - Use frequent patterns within the text as linguistic cues to identify chemical reactions by starting with seed patterns
 - Extract reactions from patent literature based on rules

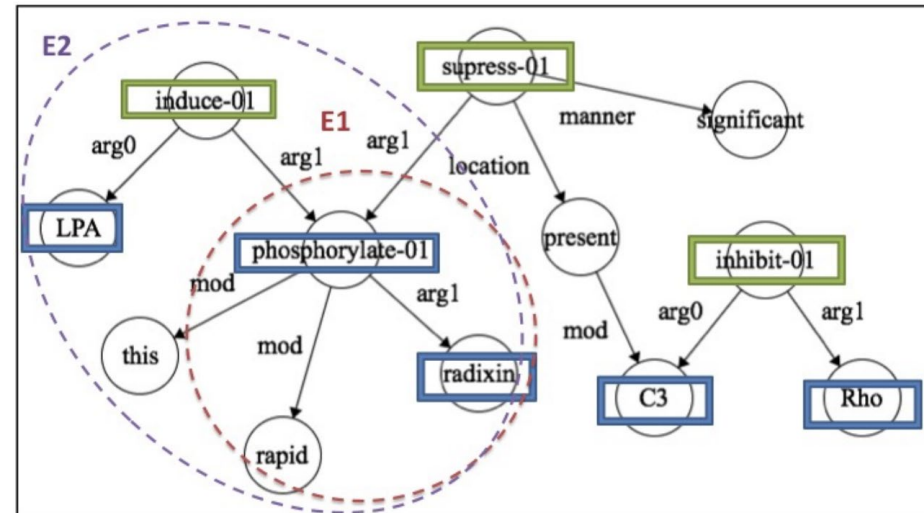


Points for improvement:

- Performance is limited by external knowledge and accuracy of AMR graphs

AMR Enhanced IE

- What is the difference between using Abstract Meaning Representation (AMR) and an External Knowledge Base?
 - AMR is a semantic representation language that converts the meaning of each input sentence into a rooted, directed, labeled, acyclic graph structure
 - AMR is more versatile in information coverage and is not limited to domain-specific ontologies
- Sentences that have the same basic meaning often have the same AMR graph
- An event graph is a subgraph of an AMR graph
 - Transform event extraction into a subgraph identification problem



Can we use both open domain and ontology-based knowledge?

Knowledge-enriched AMR Enhanced IE

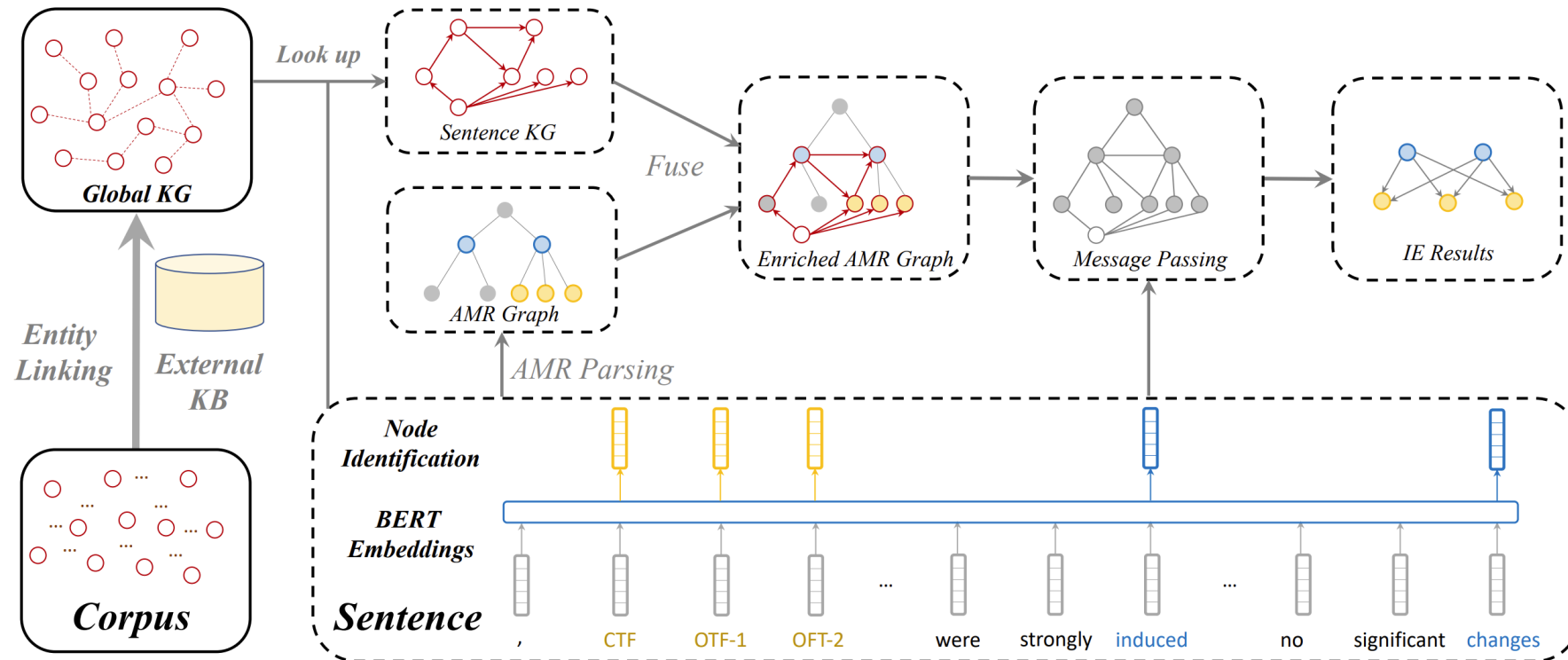
- Combine domain-specific knowledge with semantic knowledge by merging a subgraph from KG with the AMR graph
 - Enriches the model with external knowledge
 - Uses message passing with an edge-conditioned graph attention network

Pros:

- Enriches the AMR graph with external knowledge to capture long distance between event triggers and entities

Points for improvement:

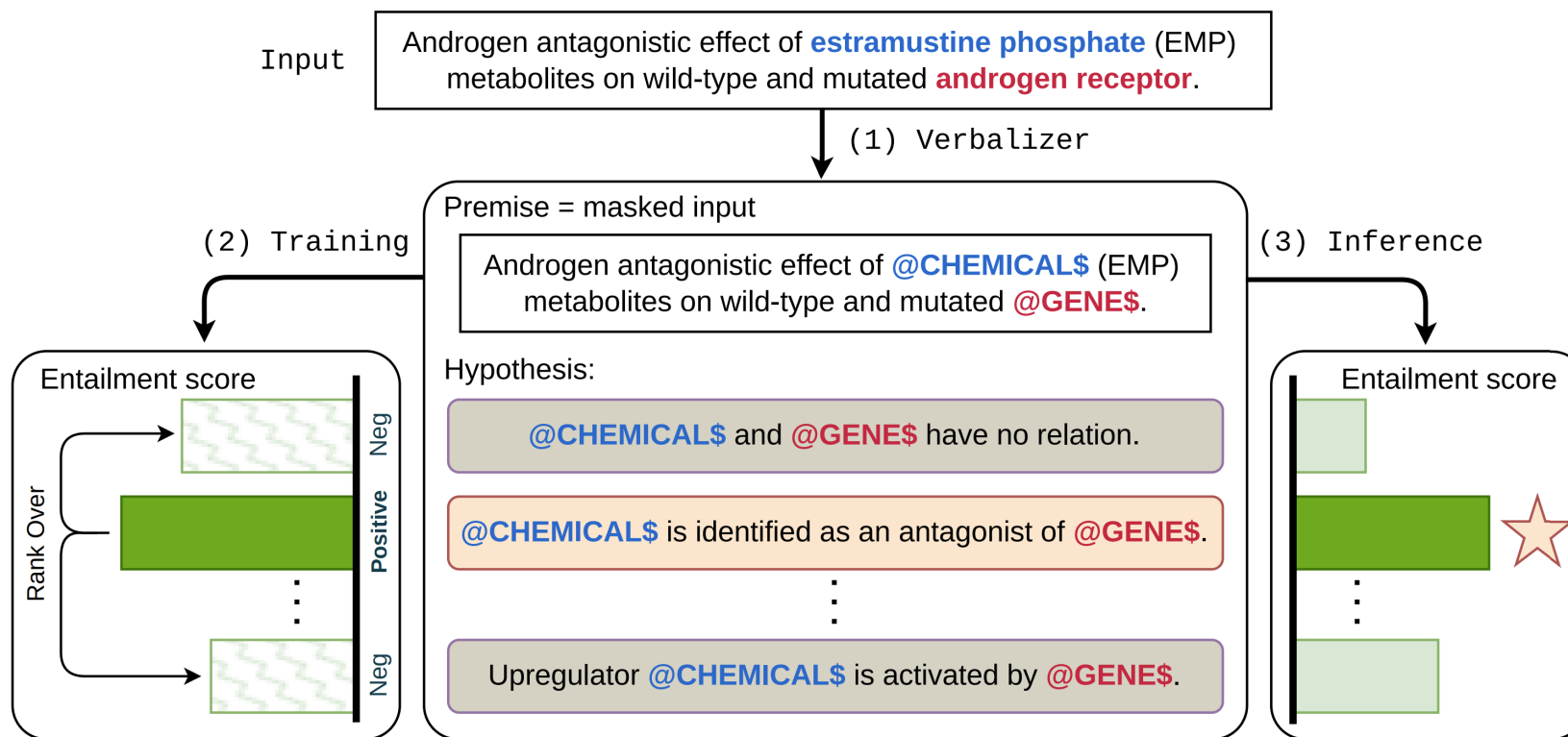
- Incorporate entity definition sentence for each entity in the sentence KG can further improve performance



How can we do information
extraction in low-resource settings?

Indirect Supervision

- Use indirect supervision to transfer supervision signals from a more resource-rich task (NLI) to enhance a more resource-limited task (biomedical RE)
 - Use the input sentence as the premise while converting each relation label into template-based natural language hypotheses for NLI
 - Propose a new ranking-based loss to balance positive and negative relations



Pros:

- Use natural language templates to adapt NLI models for the relation extraction task
- Existing entailment datasets transfers well

Points for improvement:

- Utilize task-task relations to further improve indirect supervision
- Can we automatically generate templates for arbitrary relations?

Self-Validation

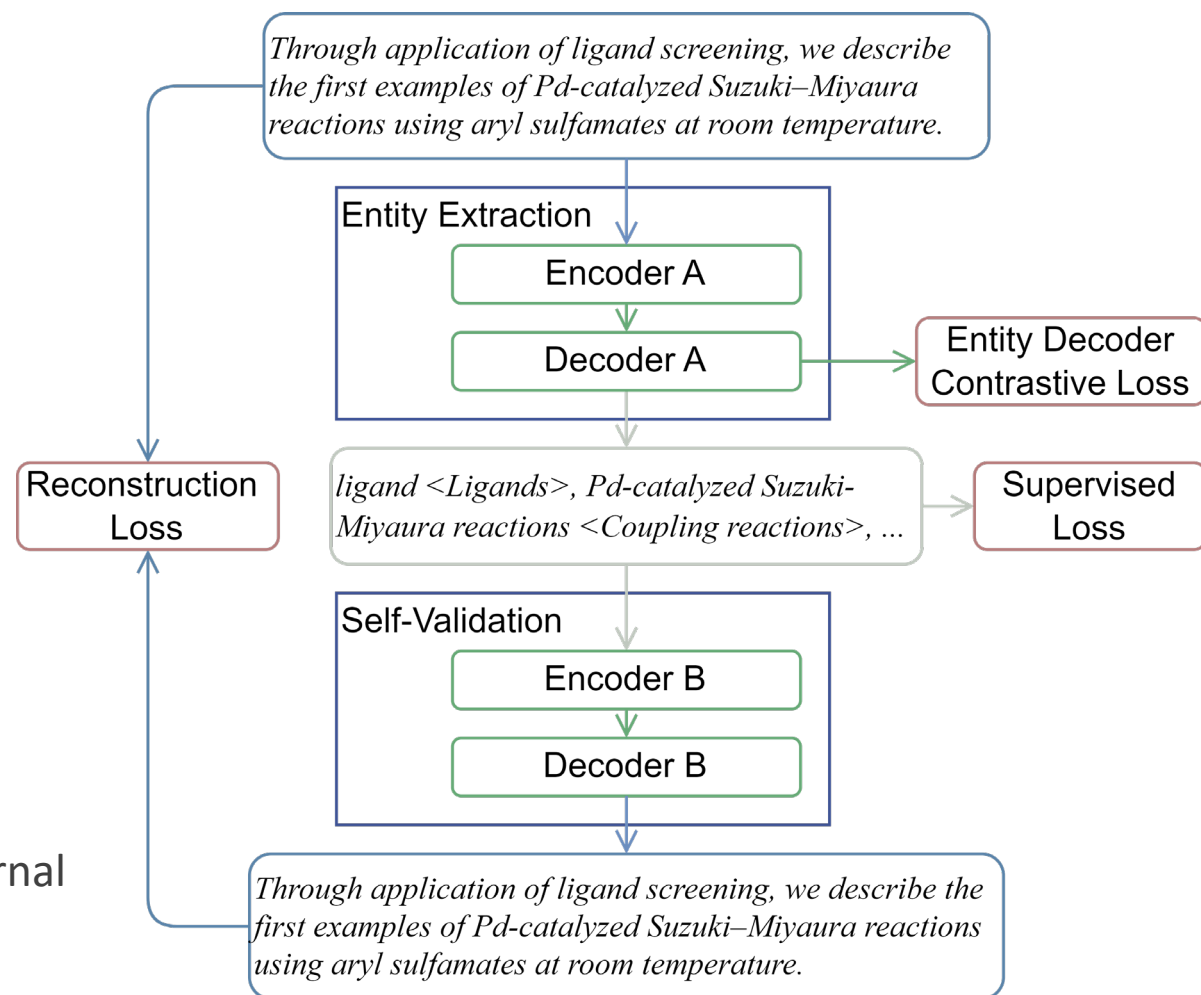
- Use a self-validation module to reconstruct an original sentences based on entity extraction results to add additional regularization for entity extraction model
- Use a contrastive loss on the entity decoder to reduce copying from the original sentence

Pros:

- Exhibits strong few-shot performances in extreme low resource settings
- Doesn't require any domain-specific pretraining or external KB

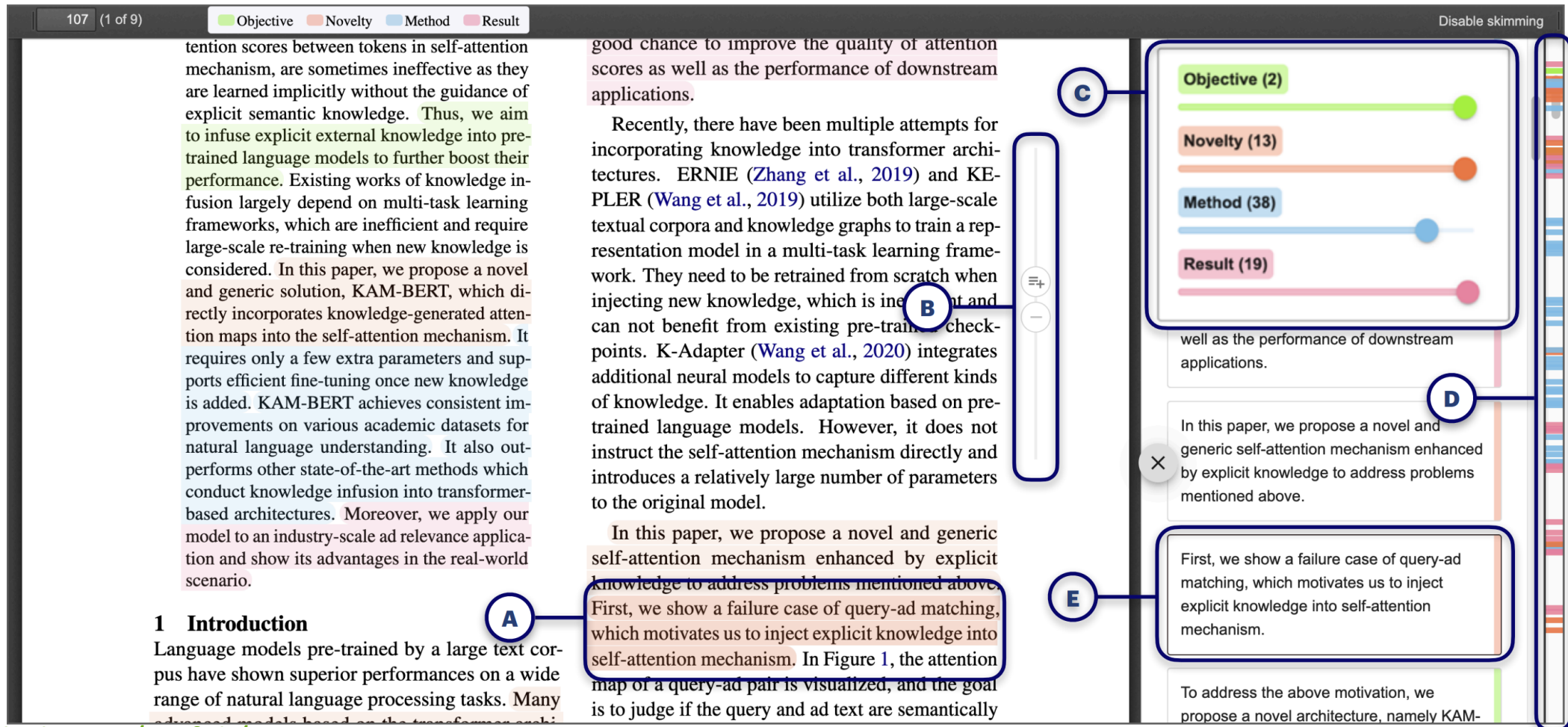
Points for improvement:

- Given that the self-validation module remains static after the initial pretraining stage, exploring cyclic improvements of both the module will be an interesting future direction



Scientific IE Applications

- Design of intelligent, highlight based skimming user interfaces, based on formative interviews and preliminary usability studies of a prototype tool



The screenshot displays the Scim interface for visualizing scientific papers. The main content area shows a paper titled "Introduction" with highlights in green, orange, blue, and pink. The sidebar on the right contains filters for Objective (2), Novelty (13), Method (38), and Result (19). The interface also includes a "Disable skimming" button and a "X" icon for closing the sidebar.

1 Introduction

Language models pre-trained by a large text corpus have shown superior performances on a wide range of natural language processing tasks. Many advanced models based on the transformer archi-

tention scores between tokens in self-attention mechanism, are sometimes ineffective as they are learned implicitly without the guidance of explicit semantic knowledge. Thus, we aim to infuse explicit external knowledge into pre-trained language models to further boost their performance. Existing works of knowledge infusion largely depend on multi-task learning frameworks, which are inefficient and require large-scale re-training when new knowledge is considered. In this paper, we propose a novel and generic solution, KAM-BERT, which directly incorporates knowledge-generated attention maps into the self-attention mechanism. It requires only a few extra parameters and supports efficient fine-tuning once new knowledge is added. KAM-BERT achieves consistent improvements on various academic datasets for natural language understanding. It also outperforms other state-of-the-art methods which conduct knowledge infusion into transformer-based architectures. Moreover, we apply our model to an industry-scale ad relevance application and show its advantages in the real-world scenario.

good chance to improve the quality of attention scores as well as the performance of downstream applications.

Recently, there have been multiple attempts for incorporating knowledge into transformer architectures. ERNIE (Zhang et al., 2019) and KEPLER (Wang et al., 2019) utilize both large-scale textual corpora and knowledge graphs to train a representation model in a multi-task learning framework. They need to be retrained from scratch when injecting new knowledge, which is inefficient and can not benefit from existing pre-trained checkpoints. K-Adapter (Wang et al., 2020) integrates additional neural models to capture different kinds of knowledge. It enables adaptation based on pre-trained language models. However, it does not instruct the self-attention mechanism directly and introduces a relatively large number of parameters to the original model.

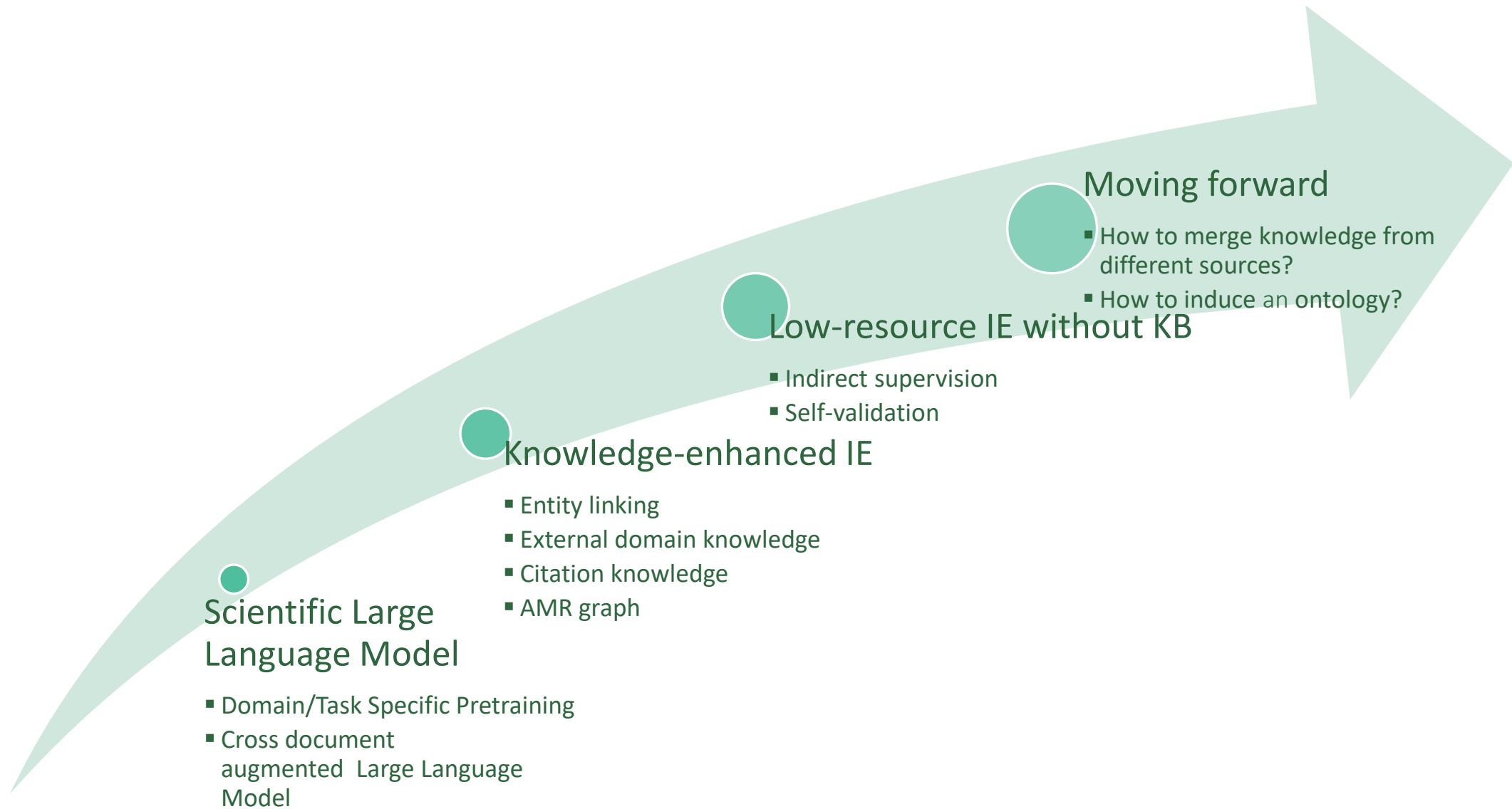
In this paper, we propose a novel and generic self-attention mechanism enhanced by explicit knowledge to address problems mentioned above. First, we show a failure case of query-ad matching, which motivates us to inject explicit knowledge into self-attention mechanism. In Figure 1, the attention map of a query-ad pair is visualized, and the goal is to judge if the query and ad text are semantically

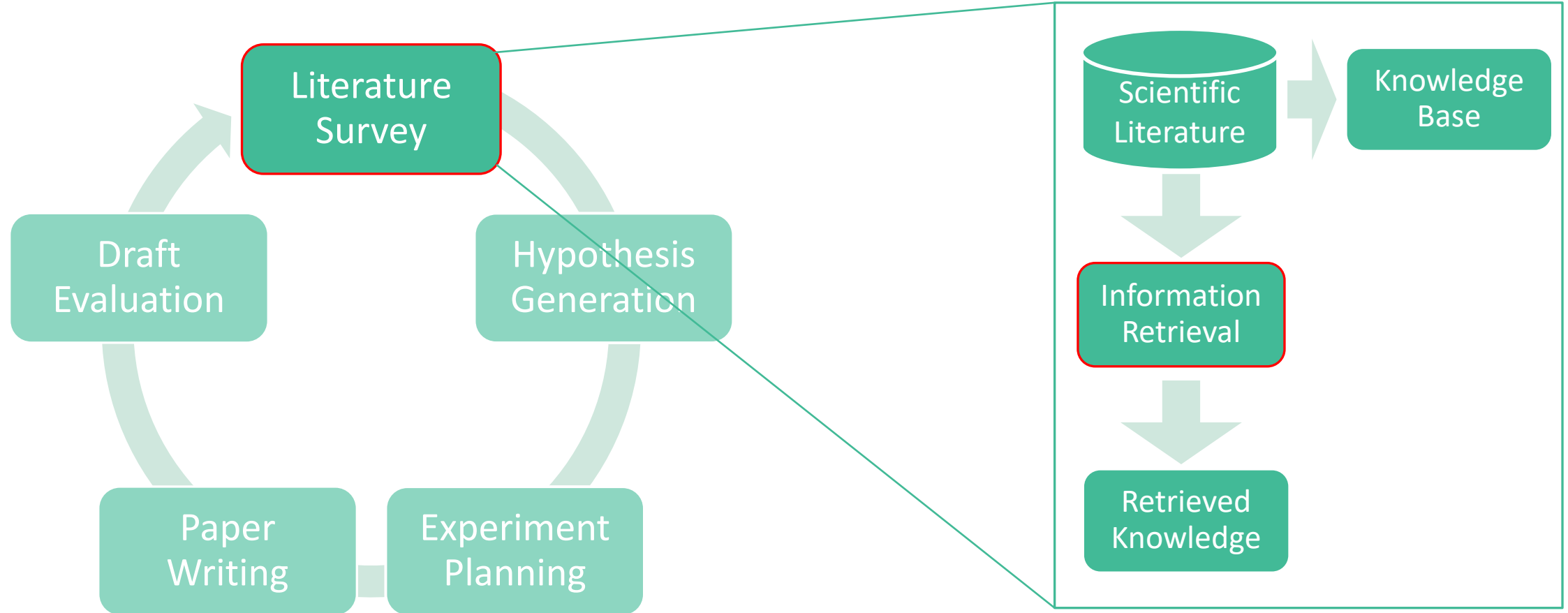
well as the performance of downstream applications.

In this paper, we propose a novel and generic self-attention mechanism enhanced by explicit knowledge to address problems mentioned above.

First, we show a failure case of query-ad matching, which motivates us to inject explicit knowledge into self-attention mechanism.

To address the above motivation, we propose a novel architecture, namely KAM-



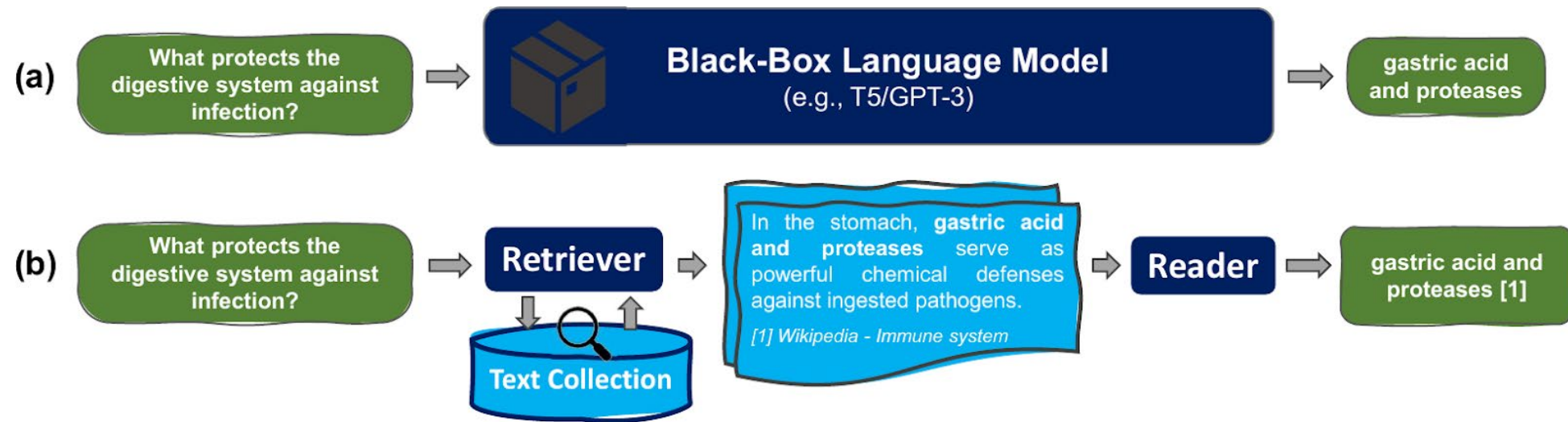


Why do we need literature search?

- Definition: Literature search is the process of retrieving scientific articles to satisfy specific information needs.
 - This is an Information Retrieval task.
- Literature search can benefit both human and knowledge-driven AI
 - Literature search allows researchers to quickly find relevant studies, reducing the risk of repeating work and missing critical information
 - Literature search can highlight the appropriateness or shortcomings of previous research methodologies
 - Literature search helps identify research trends and gaps in the existing literature, which can direct future studies
 - Literature search systems are crucial for synthesizing evidence in systematic reviews and meta-analyses

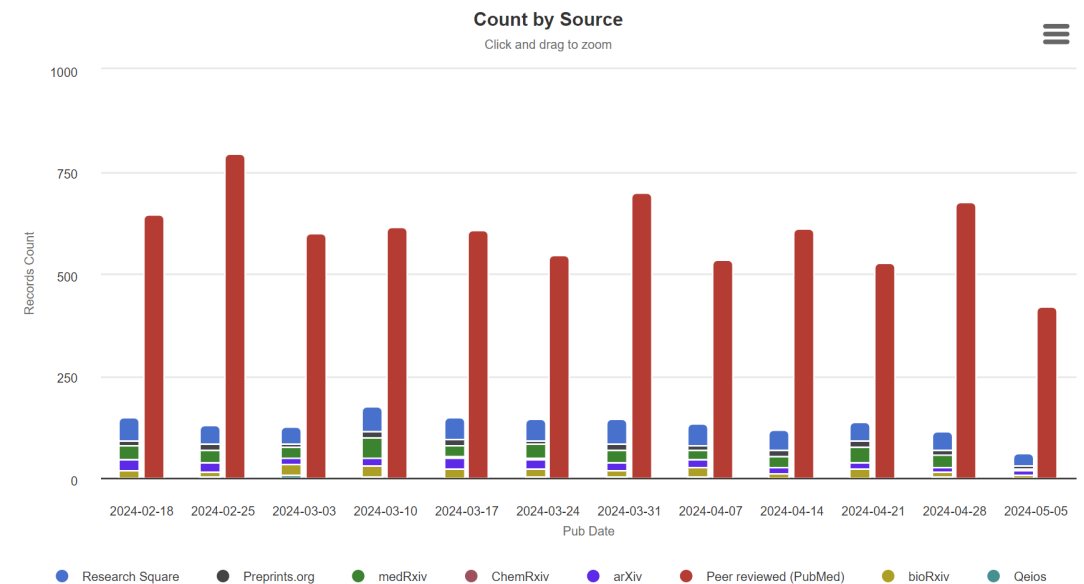
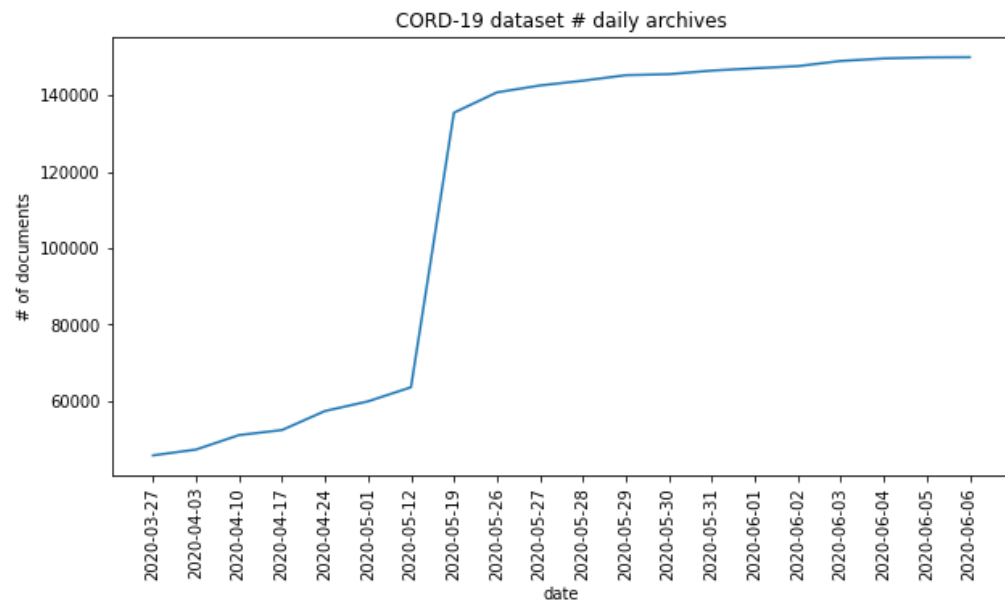
Why do we need retrieval if we're using LLMs?

- LLMs **can't memorize all knowledge** in their parameters (especially long-tail events)
- LLMs' knowledge is **easily outdated** and **hard to update**
- LLMs' output is **challenging to interpret and verify**



A real-world crisis: Digesting COVID-19 Papers

- Practical progress at combating COVID-19 highly depended on effective transmission, assessment, and extension of research results
 - 2.7K new papers per day
 - As of June 13, 2020, there were at least 140K papers about coronavirus
- There was an urgent need to effectively retrieve relevant evidence from this large corpus





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 misinformation 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 <i>et al.</i> 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 <i>via</i> 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 <i>et al.</i> ³⁶	<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 <i>via</i> 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 <i>et al.</i> ³⁷	Pilot	To evaluate HCQ against COVID-19	Negative SARS-CoV 2 nucleic acid conversion rate <i>via</i> 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

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

Pros:

- Capable of reasoning over related topics more effectively than machines
- Achieves high precision when retrieving evidence

Cons:

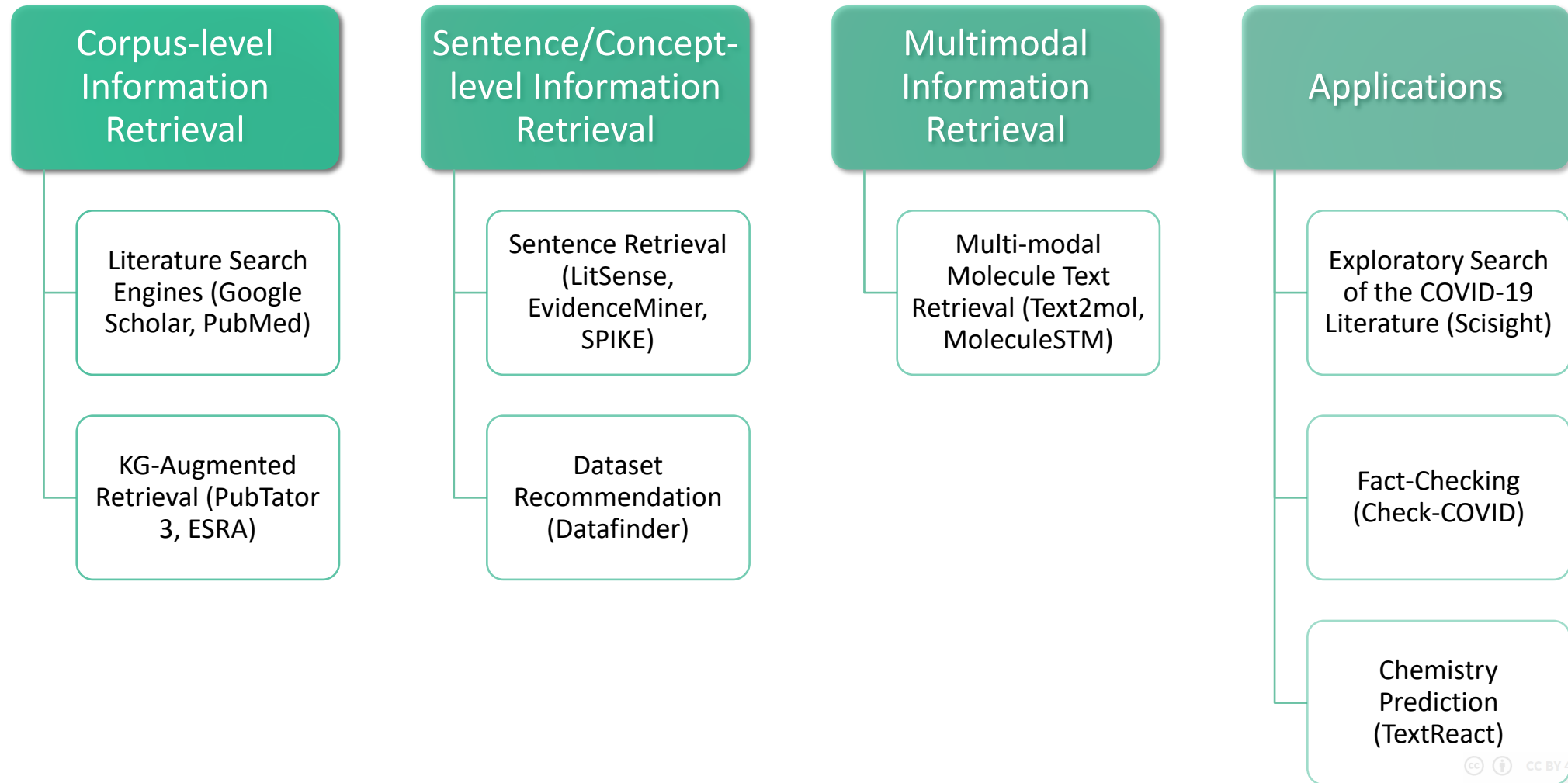
- Unable to cover all relevant research papers and evidence

Pros:

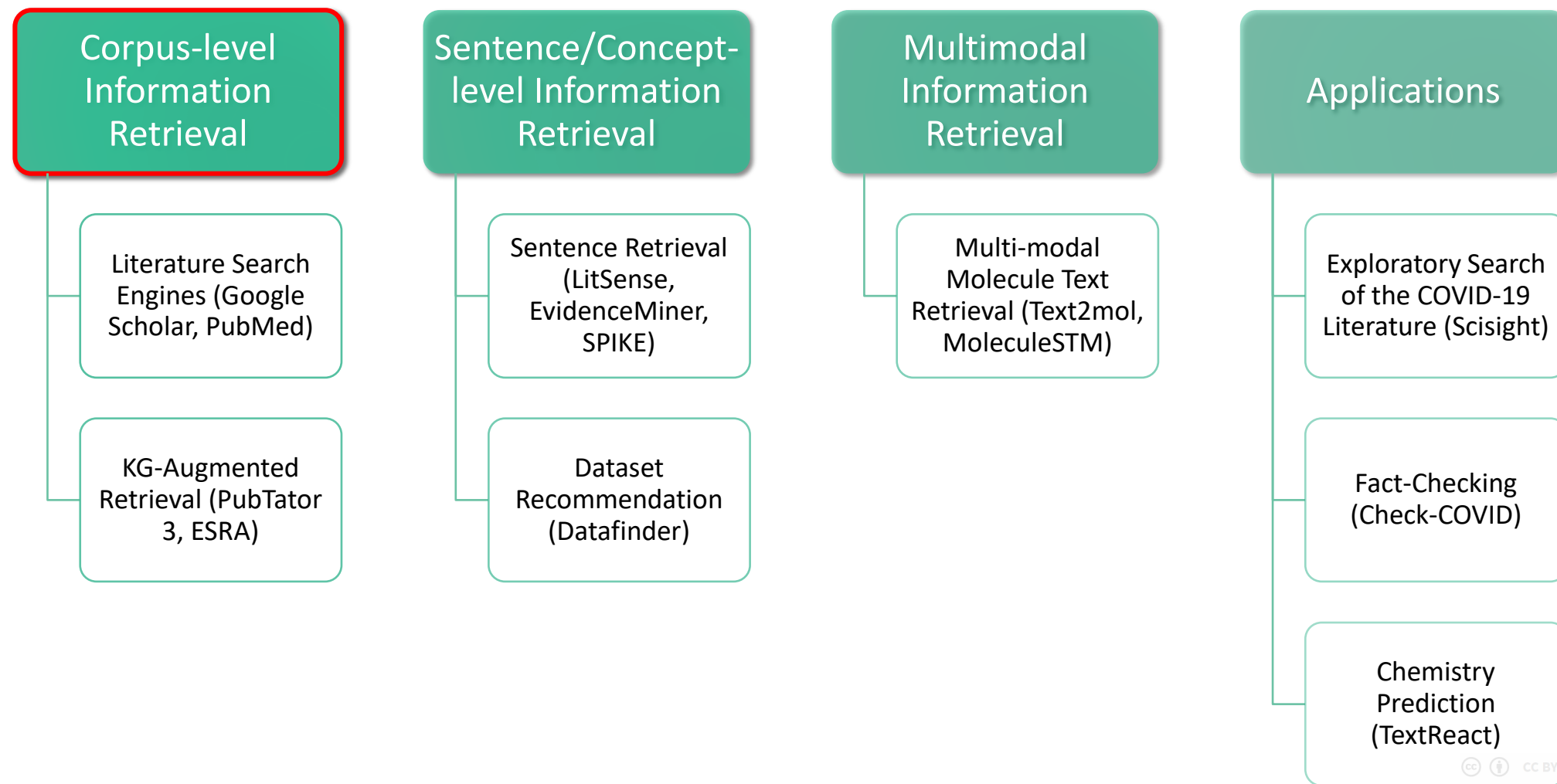
- Offer broader coverage compared to human

Cons:


- Typically exhibits low precision
- Usually ignores multimodal information and deep connections between concepts



- Essential for Scientific Literature Surveys
- Narrow Search Spaces in Retrieval Augmented Generation



- Indexes the full text or metadata of scholarly literature across an array of publishing formats and disciplines



×
Search


[Advanced](#)
[Create alert](#)
[Create RSS](#)
[User Guide](#)

Save
Email
Send to
Sort by: Best match
Display options

MY NCBI FILTERS
426,570 results
Page 1 of 42,657

RESULTS BY YEAR

☐
[Recent Developments on Therapeutic and Diagnostic Approaches for COVID-19.](#)
1 Majumder J, Minko T.
Cite AAPS J. 2021 Jan 5;23(1):14. doi: 10.1208/s12248-020-00532-2.
PMID: 33400058 [Free PMC article.](#) [Review.](#)
Share The ongoing pandemic of **coronavirus** disease 2019 (**COVID-19**) caused by the **severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)** has made a serious public health threat worldwide



🔍

Articles
About 4,900,000 results (0.06 sec)

Problem:

- Search engines can only process short keyword-based queries, returning a list of raw articles without further analysis

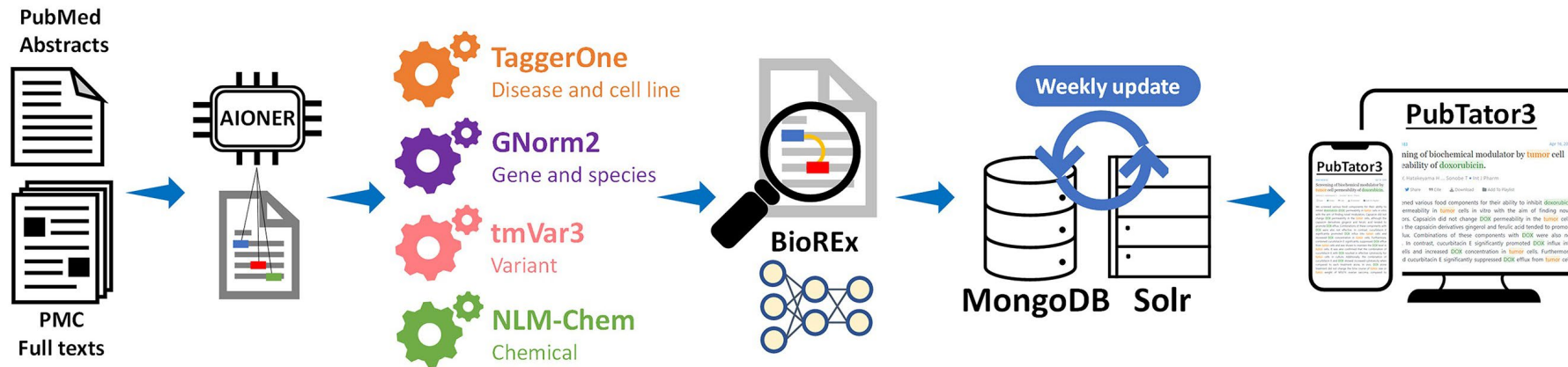
Any time
Since 2024
Since 2023
Since 2020
Custom range...

COVID-19: immunopathogenesis and Immunotherapeutics

L Yang, S Liu, J Liu, Z Zhang, X Wan... - Signal transduction and ..., 2020 - nature.com
... for **COVID-19**. In this review, we summarize the immune characteristics of **COVID-19** and ...
, their effect on disease outcomes, and their implications for potential **COVID-19** treatments. ...
☆ Save 📄 Cite Cited by 999 Related articles All 14 versions

[PDF] nature.com

- The PubTator 3.0 pipeline includes three major steps: (i) named entity recognition for six key biomedical entities: genes, diseases, chemicals, genetic variants, species, and cell lines, (ii) entity linking, and (iii) relation extraction with 12 common relation types



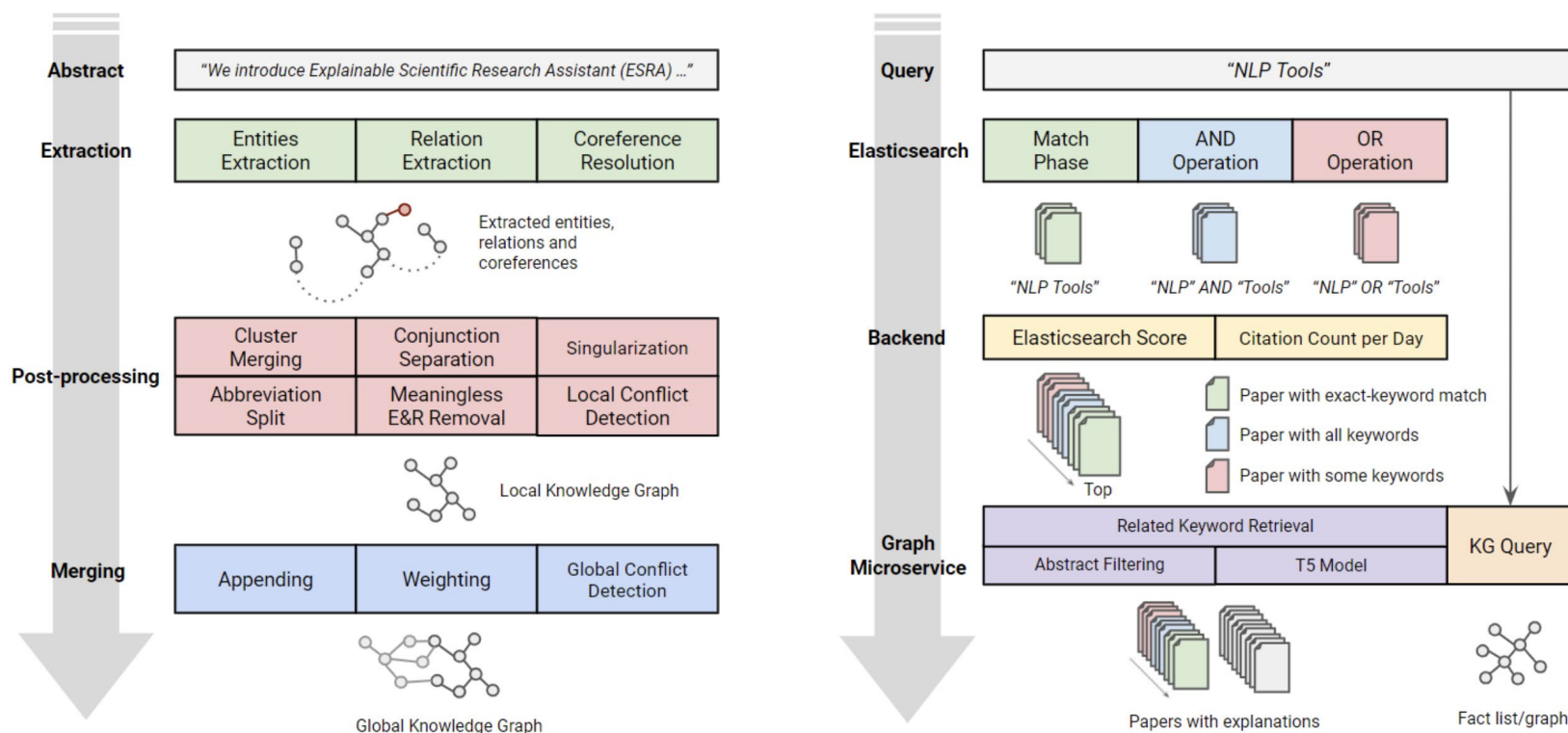
Pros:

- Provides more accurate information retrieval results compared to PubMed and Google Scholar
- Entities are linked to external knowledge bases

Cons:

- Relation extraction is limited to abstracts only
- Frequently miss new entities in the entity extraction process
- The whole knowledge graph of the paper isn't visualized

- ESRA: A literature discovery platform that augments search results with relevant details and explanations
 - Perform scientific information extraction to construct knowledge graph
 - Expand the query based on similar entities from background knowledge graph
 - Use elastic search and citation counts to rank papers
 - Provide ranking explanations based on selected sentences in paper abstracts



Pros:

- Provides explanations, graph visualizations, and facts to increase interpretability

Points for Improvement:

- Information extraction is limited to abstracts only
- Nodes in the knowledge base are not linked to human curated KG



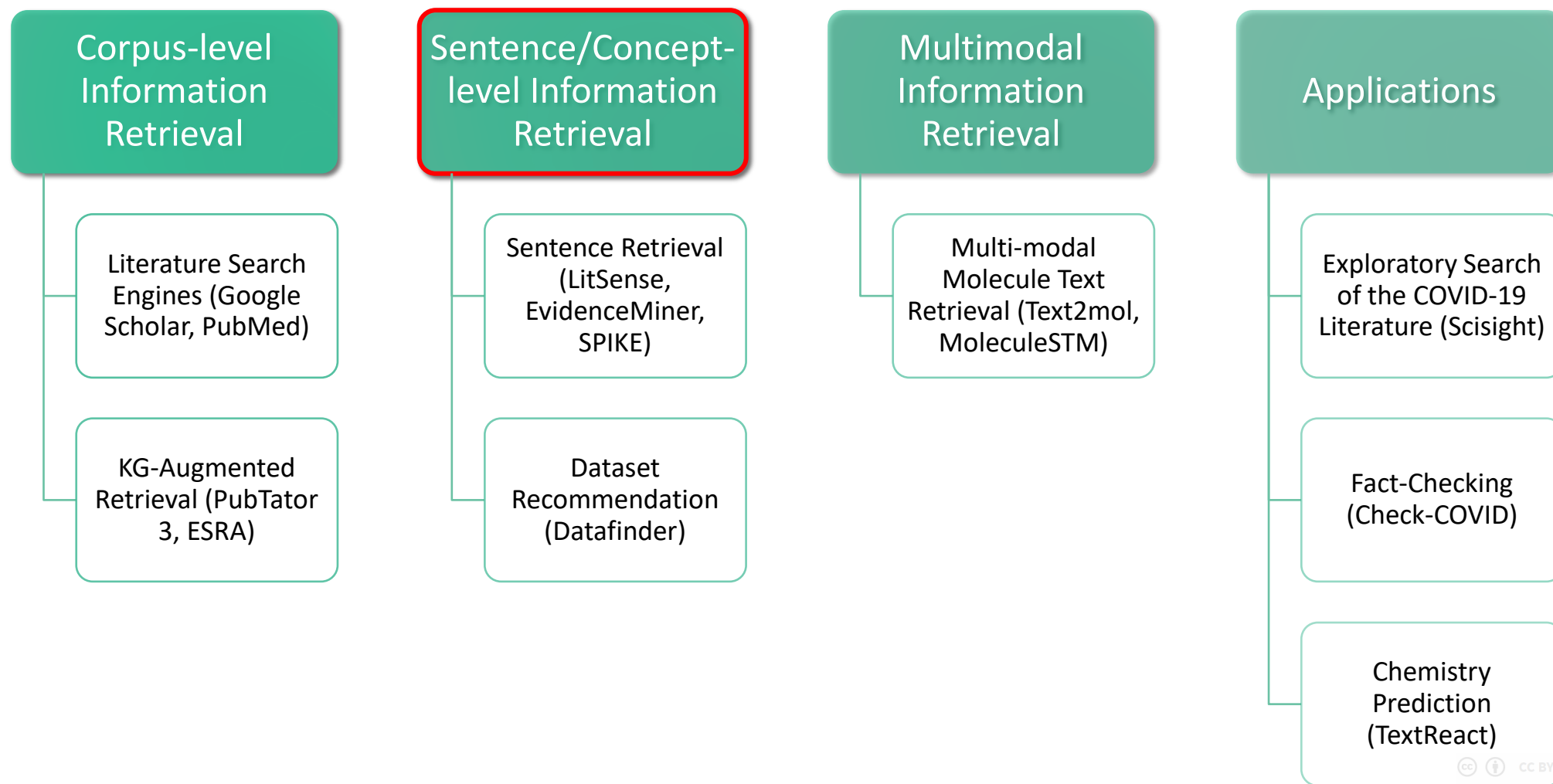
Key Takeaways for Corpus-level Information Retrieval



- Notable points
 - Leveraging both structured and elastic search can improve document retrieval accuracy

- Future directions
 - Need to find a balance between traditional keywords-based search and KG-based retrieval to save computational power
 - Existing systems lack the ability to uncover deep connections between entities during information retrieval due to the absence of reasoning steps in the process

- Sentence/concept retrieval results can be provided to Retrieval Augmented Generation for downstream tasks



LitSense: Similar Sentence Retrieval

- Rank sentences based on two methods
 - A traditional term-weighting approach that up-weights sentences that contain more of the rare terms in the user query
 - A neural embedding approach that retrieves weights based on semantic similarity



EvidenceMiner: Textual Evidence Discovery

- Given a query, retrieve an evidence sentence from corpus
 - Extracts entities with distant supervision based on knowledge bases to provide keywords for retrieval
 - Retrieves sentences based on query word, entity, and meta-pattern matching
 - Requires no human annotation and achieves better performance for automatic textual evidence discovery

Q suzuki coupling, catalyst

"" (Total: 508, Took: 1ms)
~ At most 10 results are shown per page ~

Type: POLYMERIZATION_REACTIONS

Our study provided a **highly efficient catalyst system** for the **Suzuki cross-coupling polymerization** of **aryl dihalides** with **aryldiboronic acids**. [Context](#)

✓ Evidence Score 9.36 2008 Jul 16 Fake Journal Source: Fake Source PMID: 1234567 PMCID: 1234567 DOI: 192-34-5 Author, Name1

Title: demo article

The **first Suzuki cross-coupling reaction** of aryltrimethylammonium triflates based on the use of an **IMes.Ni (0) catalyst system** is described. [Context](#)

✓ Evidence Score 9.05 2008 Jul 16 Fake Journal Source: Fake Source PMID: 1234567 PMCID: 1234567 DOI: 192-34-5 Author, Name1

Title: demo article

Herein, we describe a **recyclable and highly active nickel catalyst** immobilized on **MOF** for **Suzuki-Miyaura coupling reaction**, which operates under **mild conditions**. [Context](#)

✓ Evidence Score 8.67 2008 Jul 16 Fake Journal Source: Fake Source PMID: 1234567 PMCID: 1234567 DOI: 192-34-5 Author, Name1

Title: demo article

This **protocol** combines **two C-C bond forming reactions** (an **intermolecular Suzuki coupling** followed by an **intramolecular Heck coupling**) in **one pot** under the **catalysis** of the **same catalyst**. [Context](#)

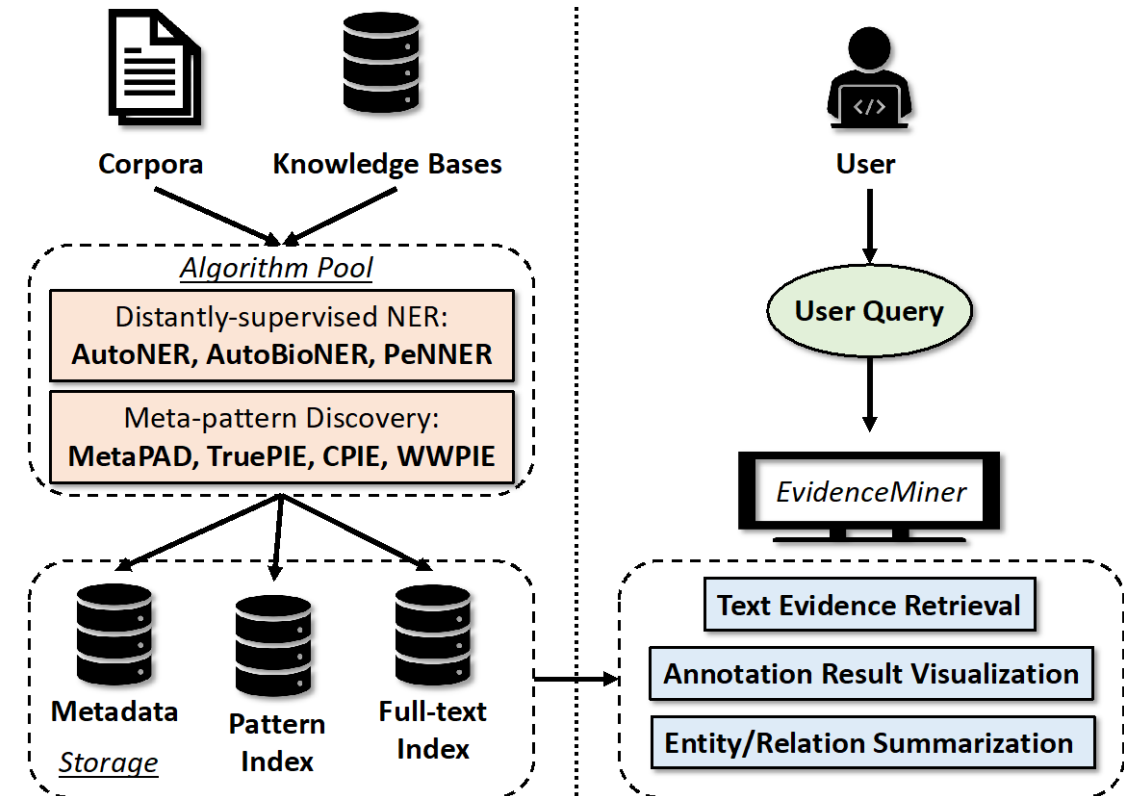
✓ Evidence Score 8.52 2008 Jul 16 Fake Journal Source: Fake Source PMID: 1234567 PMCID: 1234567 DOI: 192-34-5 Author, Name1

Title: demo article

The **catalyst** also performed well in the **Suzuki-Miyaura coupling reaction** at **room temperature** without an **inert atmosphere** and any **toxic solvents**. [Context](#)

✓ Evidence Score 8.40 2008 Jul 16 Fake Journal Source: Fake Source PMID: 1234567 PMCID: 1234567 DOI: 192-34-5 Author, Name1

Title: demo article





SPIKE: Neural Extractive Search



- Enrich the search query with capture-slots to allow for rapid extraction
 - Use standard dense passage retrieval methods to retrieve relevant sentences
 - Incorporate a neural argument alignment model to align the argument in the query that corresponds to the capture spans on each sentence

Structural EquivalenceToken PatternBoolean Pattern

:something is a \$drug \$extracted \$from :plants

add filtersadd expansion overridesCase Sensitivity: Insensitive

Query Graph

Graph

Sentence ViewTable ViewDetailed ViewShow EntitiesDownload CSV

Colchicine is a drug extracted from Colchicum autumnale [5] .

Taxol , isolated from Taxomyces andreanae , is the most effective and successful anticancer drug extracted from endophytic fungi to date .

Colchicine is a tricyclic alkaloid that is extracted from the plant Colchicum autumnale .

Hypericin is a natural polycyclic quinone found in Hypericum perforatum .

Lycorine is a natural alkaloid extracted from amaryllidaceae and it has various pharmacological and microbiological effects .

Orthoquin is a botanical extract with antibacterial properties that are amplified by light exposure [17] .

Aggregations

1005001,0002,5005,000

CapturesMetadataWordsEntitiesPivot

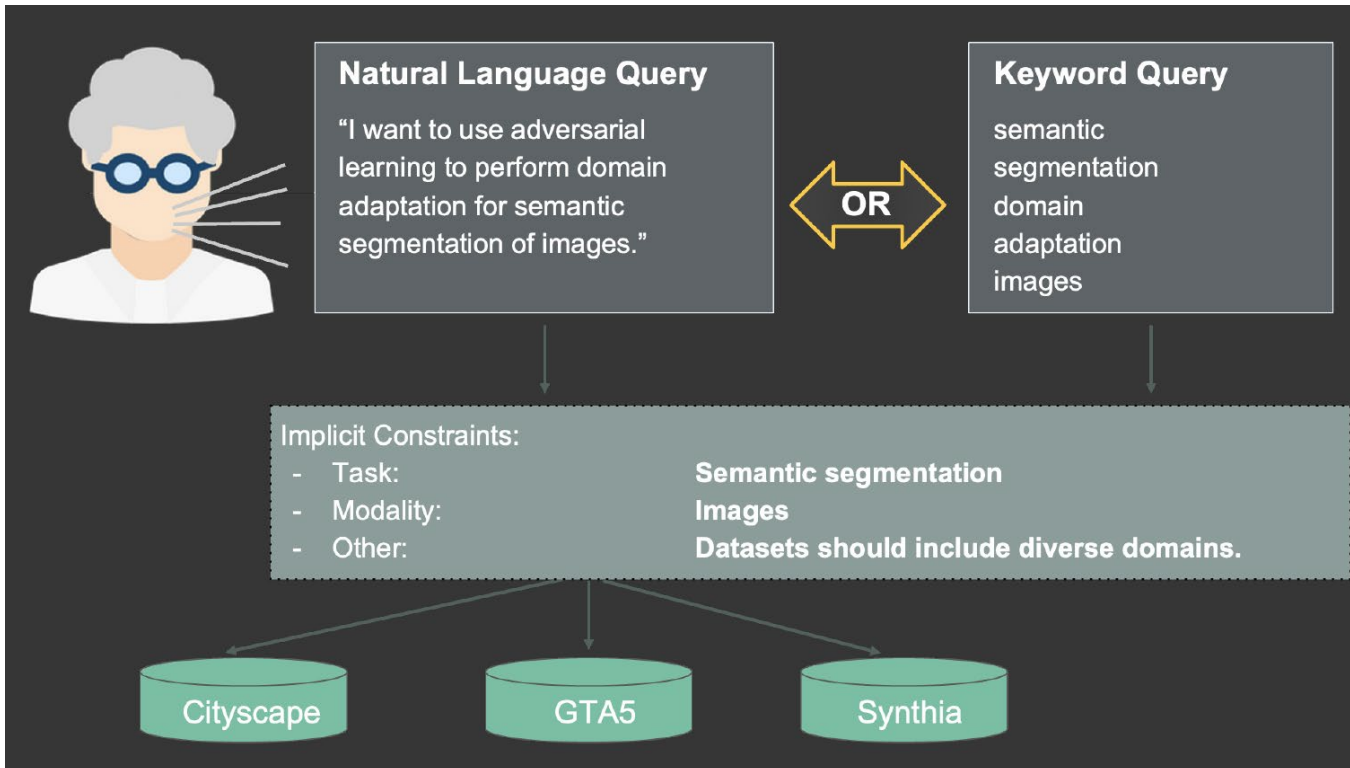
cap:something x cap:plants x

Pivot Table (439 samples)

cap:something	cap:plants	Count
<input type="checkbox"/> colchicine	autumnale	2
<input type="checkbox"/> lycorine	amaryllidaceae	2
<input type="checkbox"/> a triterpene se	glycyrrhiza gl	2
<input type="checkbox"/> allicin	garlic	2
<input type="checkbox"/> glycyrrhizin	licorice root	2
<input type="checkbox"/> glycyrrhizic ac	licorice root	2
<input type="checkbox"/> the most abur	t. wilfordii	2
<input type="checkbox"/> pkc412	the alkaloid	2
<input type="checkbox"/> glycyrrhizin	g. glabra	2
<input type="checkbox"/> scorpine	pandinus imp	2
<input type="checkbox"/> colchicine	crocus	1
<input type="checkbox"/> berbamine	shrub	1
<input type="checkbox"/> quinine	bark	1
<input type="checkbox"/> taxol	fungi	1
<input type="checkbox"/> a drug	colchicum aut	1
<input type="checkbox"/> taxol	endophytic fur	1

Name	Pros	Cons
LitSense (Allot et al., 2019)	<ul style="list-style-type: none"> ▪ Easy to setup ▪ Require minimal resource 	<ul style="list-style-type: none"> ▪ Performs worst among three methods
Evidenceminer (Frisoni et al., 2022)	<ul style="list-style-type: none"> ▪ Leverages knowledge graph to annotate sentences without human annotation ▪ Can switch to approximate matching when strict query matching fails to find sufficiently high-quality answers 	<ul style="list-style-type: none"> ▪ Relies on external KBs for distant supervision ▪ Relies on pattern matching for relation extraction
SPIKE (Ravfogel et al., 2021)	<ul style="list-style-type: none"> ▪ Enriches the search query into a knowledge graph ▪ Provides Boolean and syntactic search 	<ul style="list-style-type: none"> ▪ Suffers from speed and scalability due to dense retrieval system

- Datasets are hard to directly index for search and there are no corpora available for this task
- Solution: Create the dataset retrieval task
 - Build the DataFinder Dataset consists of a larger automatically-constructed training set (17.5K queries) and a smaller expert annotated evaluation set (392 queries)
 - Develop a bi-encoder retriever for text-based dataset recommendation, which finds more relevant search results than existing third-party dataset search engines

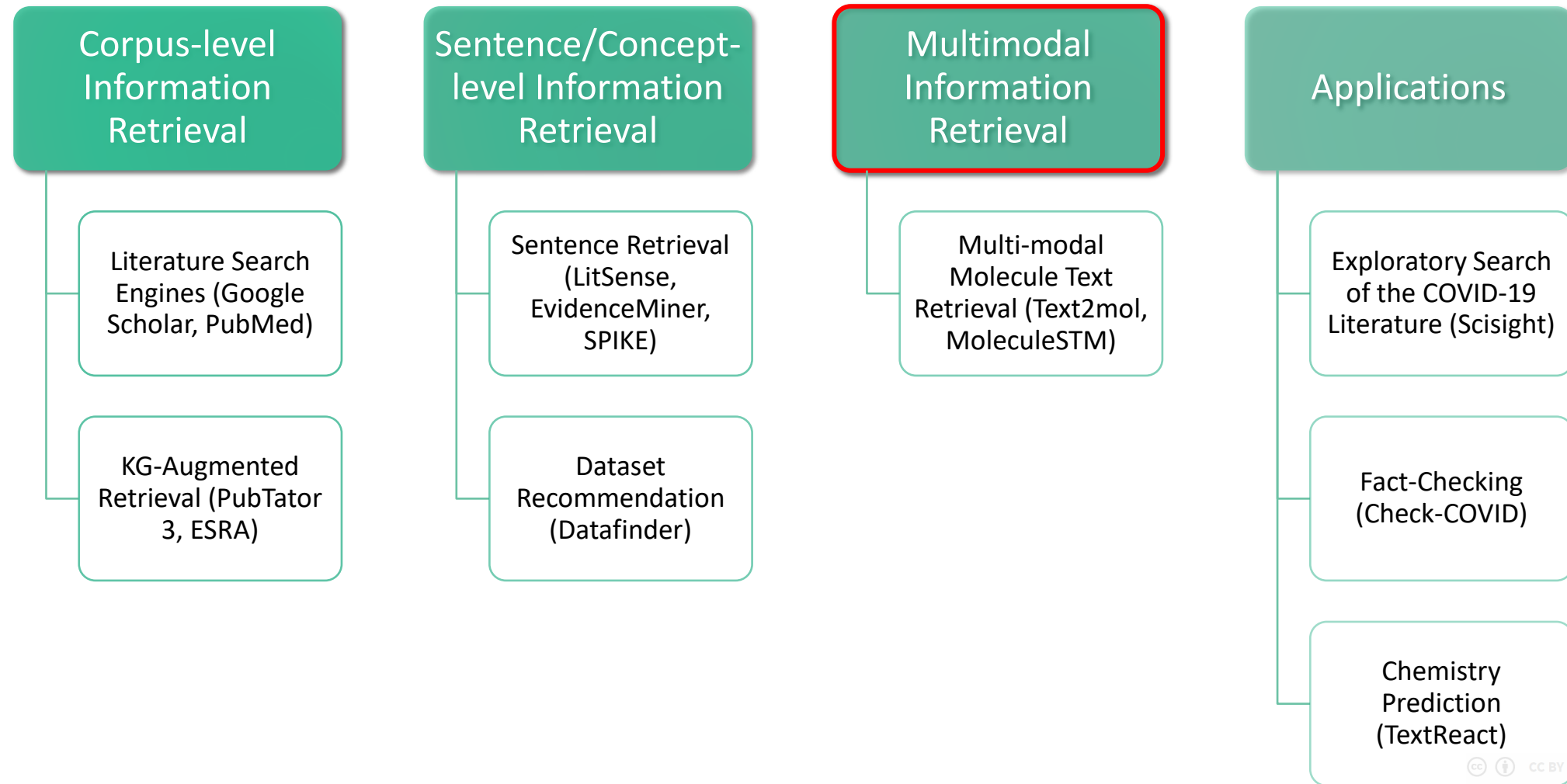


Points for Improvement:

- Methods tested in the system are not state-of-the-art
- Papers in the test set and training set are not checked for paper publication year, resulting in potential information leaks
- The whole system relies on Papers with Code datasets

Scientific Information Retrieval Categories

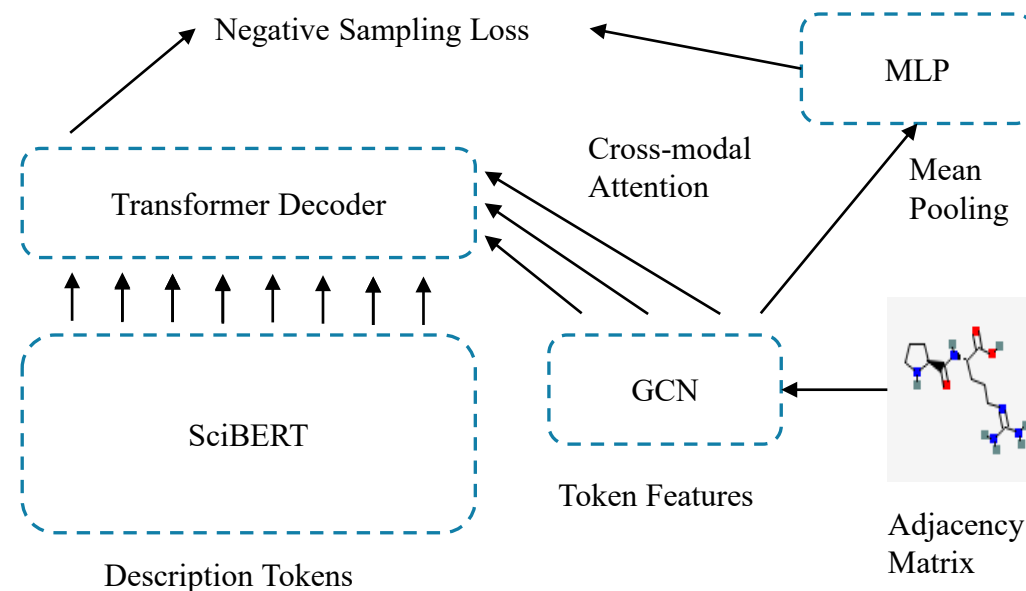
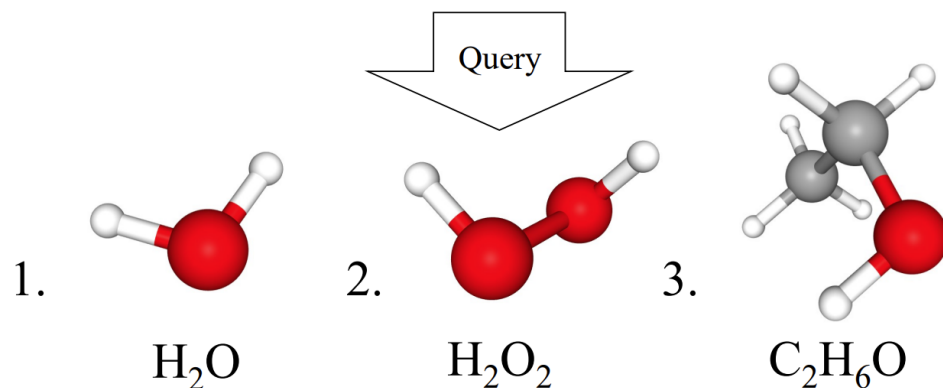
- The text modality usually suffers from reporting bias, *the tendency of people to not state the obvious*
- Multimodal information retrieval can address this limitation



Finding the Molecule via Cross-modal Retrieval

- Text2Mol Cross-modal text-molecule information retrieval directly from natural language descriptions to molecules
 - Allows semantic search of molecules based on high-level properties and descriptions
 - Cross-modal attention-based association rules are created between molecules and text to improve explainability
 - A new benchmark dataset with 33,010 text-compound pairs for cross-modal text-molecule IR is released

Water is an oxygen hydride consisting of an oxygen atom that is covalently bonded to two hydrogen atoms.



Points for Improvement:

- Incorporate additional external knowledge to the system

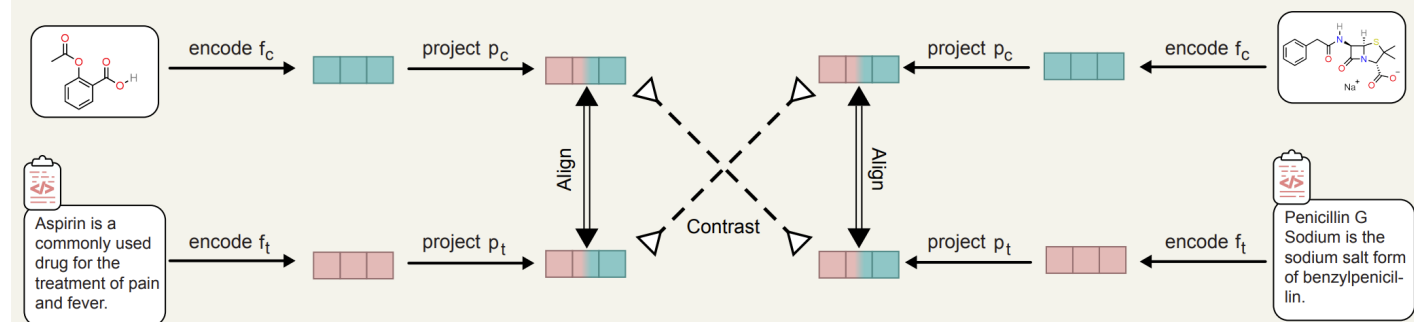
- MoleculeSTM increases capabilities by training on a larger dataset

- Constructs a large multi-modal dataset, PubChemSTM, with over 280,000 chemical structure-text pairs
- Adds capability for property prediction and molecule editing tasks

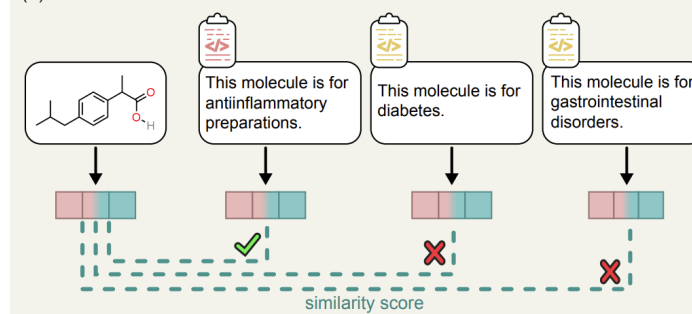
Pros:

- Open vocabulary
 - Not limited to a fixed set of pre-defined molecule-related textual descriptions
- Compositionality
 - Decompose a complex concept into several simple concepts

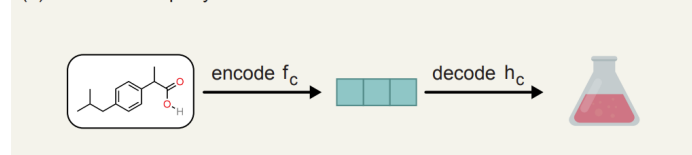
(a) Contrastive Pretraining



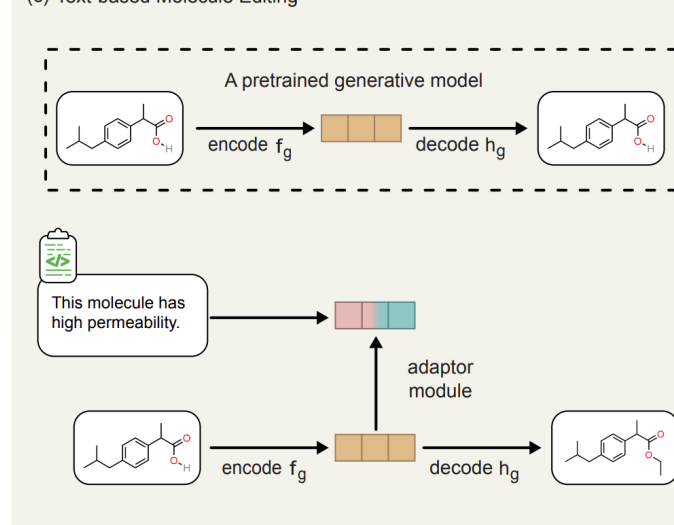
(b) Structure-text Retrieval

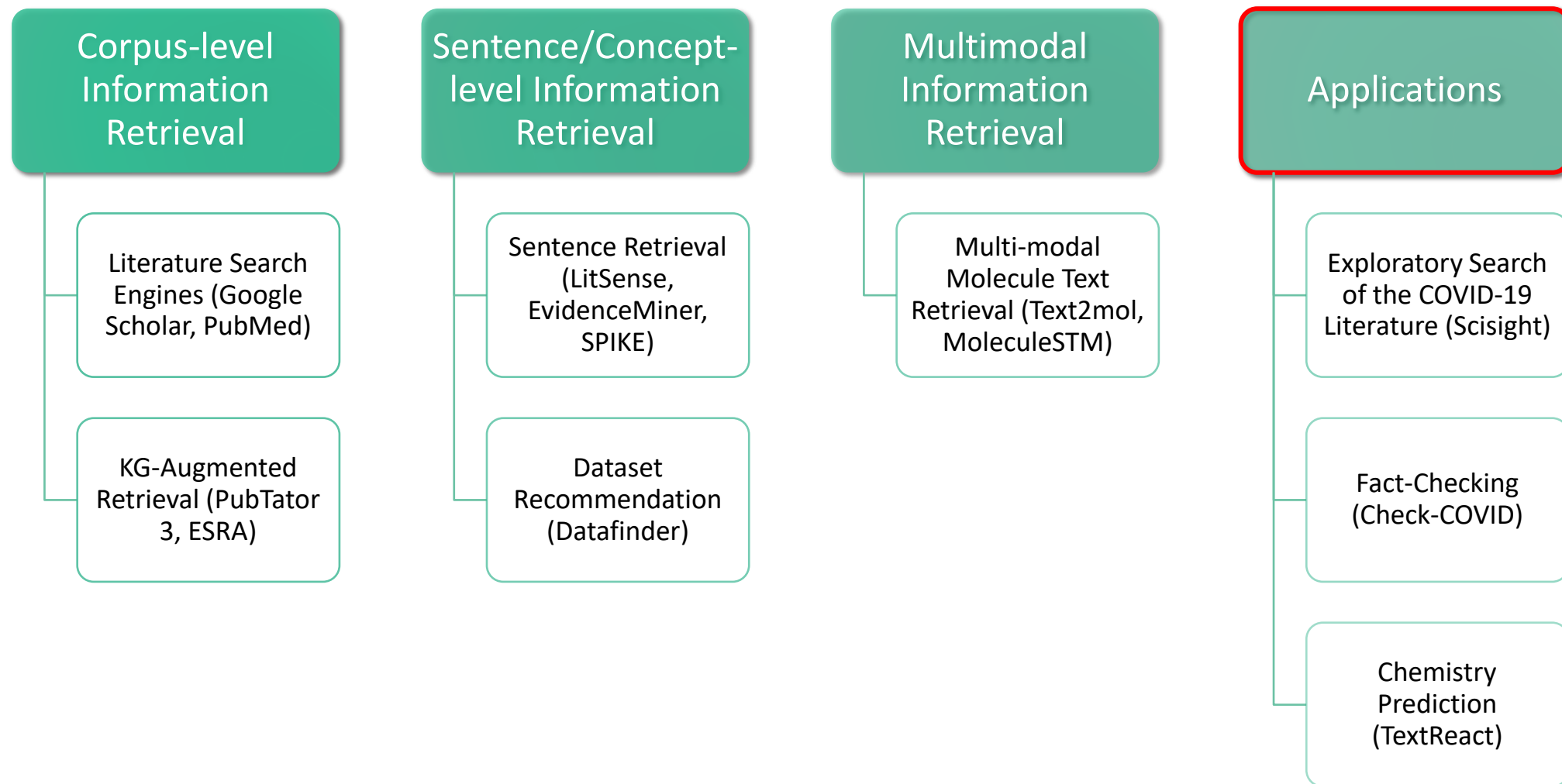


(d) Molecular Property Prediction



(c) Text-based Molecule Editing

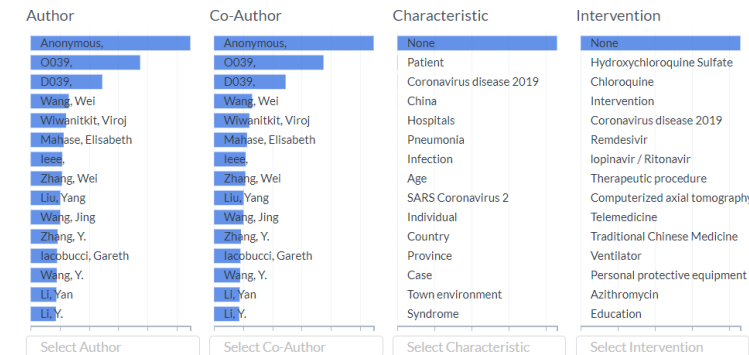
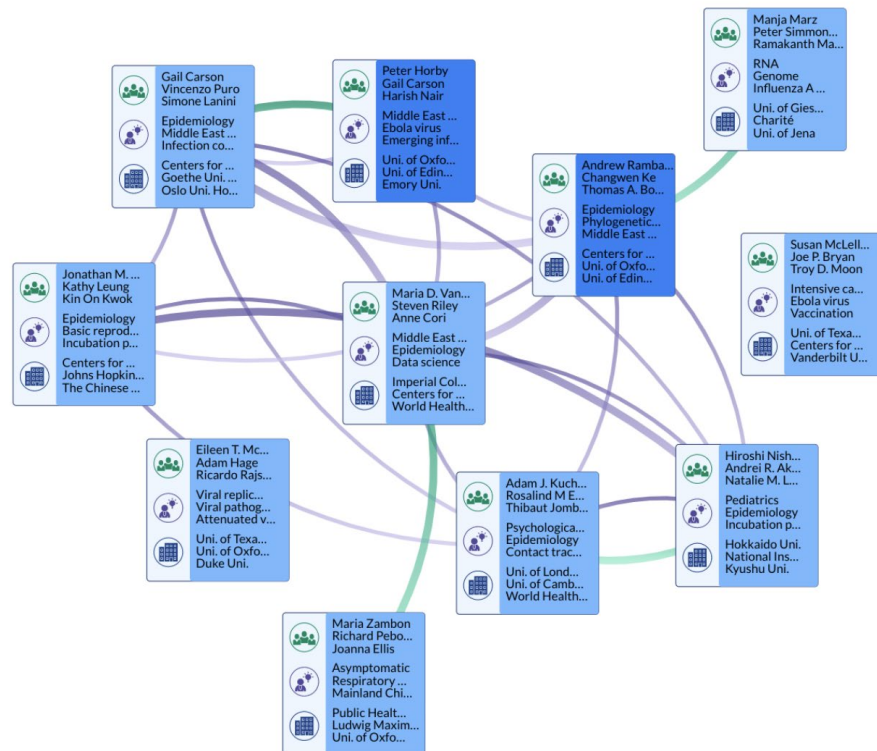




- Explore associations between biomedical concepts from papers (e.g., genes, drugs, diseases, patient outcomes)
- Combine textual and network information to search and visualize groups of researchers and connections between them

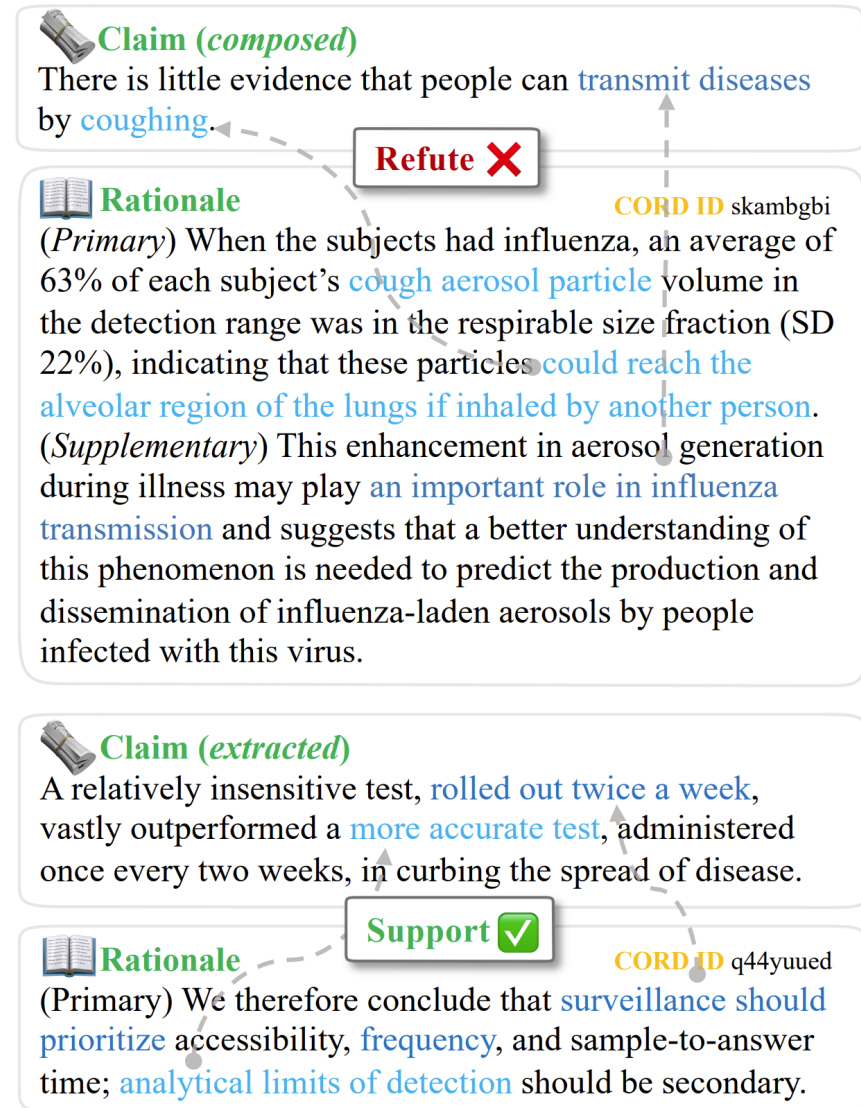
Display at most groups. Relevance: High  Low

You can use the mouse wheel to zoom in/out and drag boxes to tease them apart.



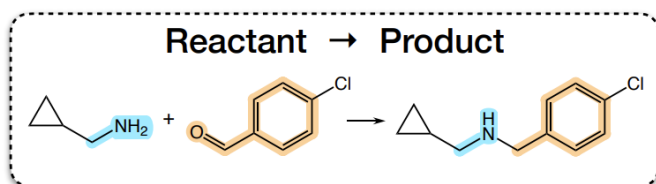
Showing 871565 out of 970836 COVID-19 Papers. reset all

- COVID-19 caused an incredible amount of fake medical news
- We need a system to verify COVID-19 news claims
- Relevant Tasks:
 - Abstract Retrieval
 - Retrieve relevant paper abstracts related to the claim
 - Rationale Selection
 - Identify the relevant rationale within the retrieved abstracts
 - Label Prediction
 - Classify each selected rationale as SUPPORT, REFUTE, or NOTENOUGHINFO



- Directly augment predictive chemistry with text retrieved from the literature
 - Use a SMILES-to-text Retriever to retrieve relevant text descriptions for a given chemical reaction.
 - Supply the retrieved text as additional evidence for prediction of chemical reaction conditions.

Input: reaction



Text Retrieval

Model

Augment

Output: reaction condition

Catalyst

CH_3COOH
 $\text{Na}[\text{BH}_3(\text{CN})]$

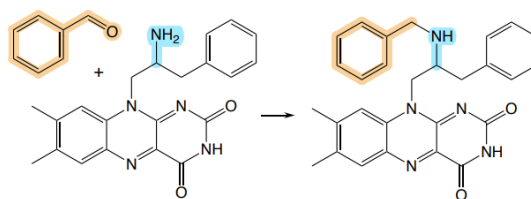
Solvent

CH_3OH

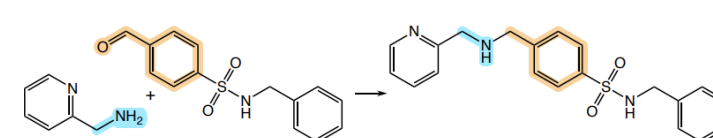
Reagent

NaHCO_3

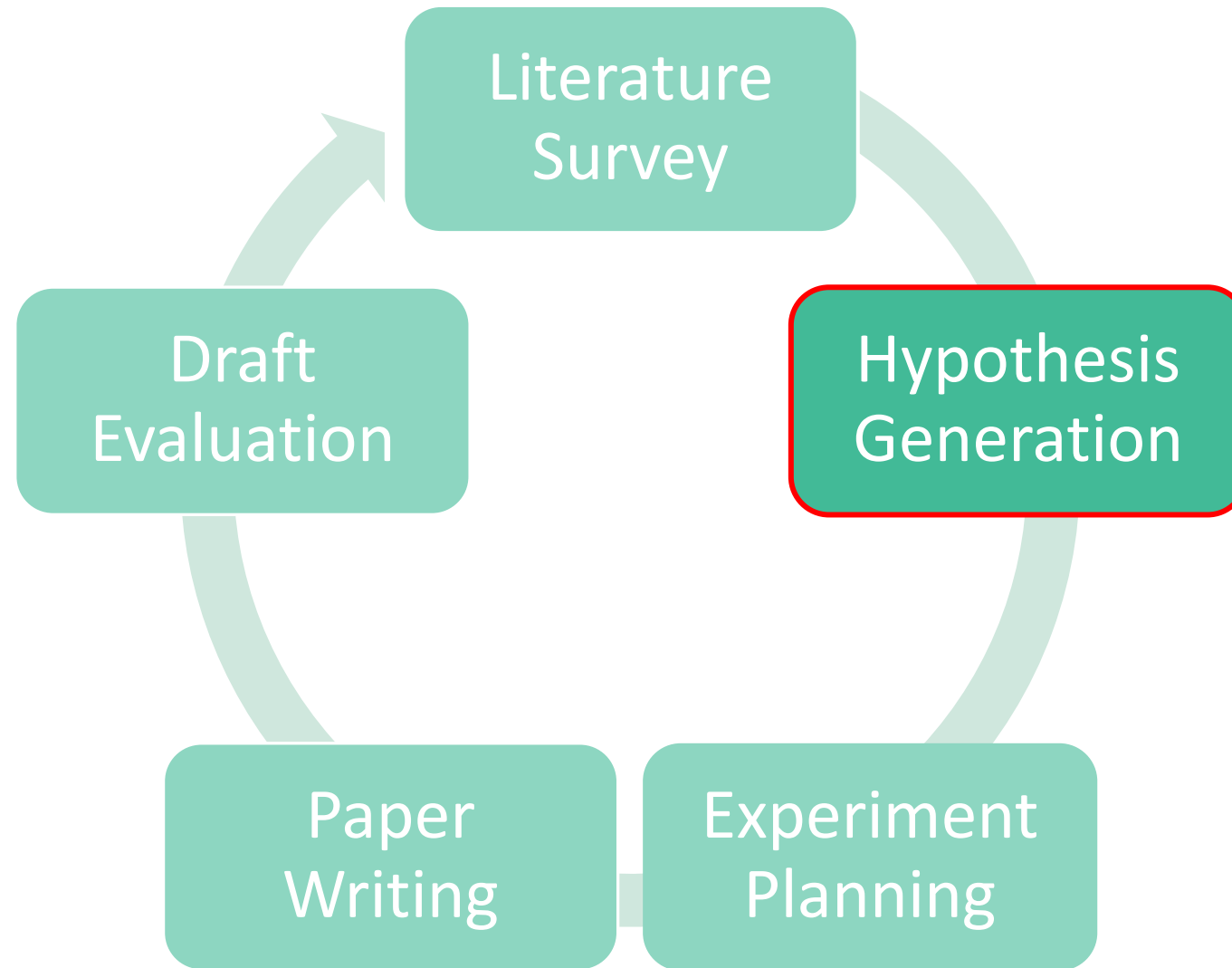
10-(2-Amino-3-phenylpropyl)-...-2,4(3H,10H)-dione is dissolved in **MeOH** at room temperature, and then **benzaldehyde** and **AcOH** (1 drop) are added **NaBH_3CN** is added in one portion The crude product is dissolved in **DCM:MeOH** [4:1] and purified to afford the desired product.



To a solution of (S)-tert-butyl 4-(1-(4-formylbenzyl)-3-(1-(naphthalen-1-yl)ethyl)ureido)butylcarbamate in **dichloroethane** (20 mL) was added **N-benzylamine**, **acetic acid** (10 drops) and **$\text{Na}(\text{OAc})_3\text{BH}$** . The mixture was a saturated aqueous solution of **sodium bicarbonate**.



- Develop multimodal information retrieval systems that suggest papers/evidence based on figures or charts
- Balance dense retrieval methods with traditional retrieval techniques to enhance speed and scalability

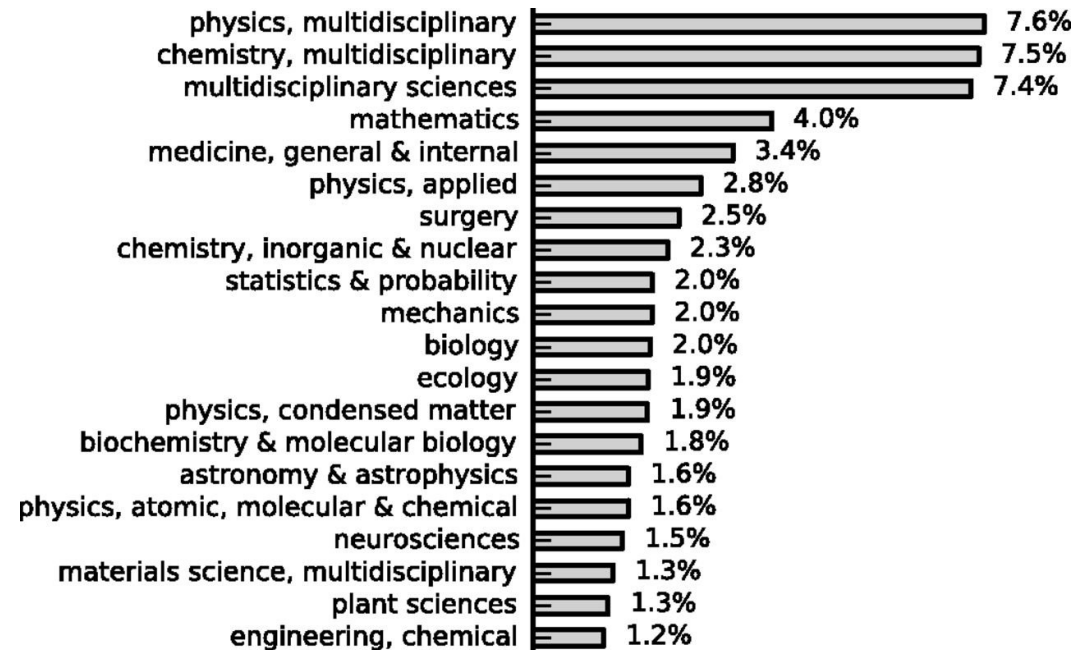




Why do we want AI-Assisted Hypothesis Generation?

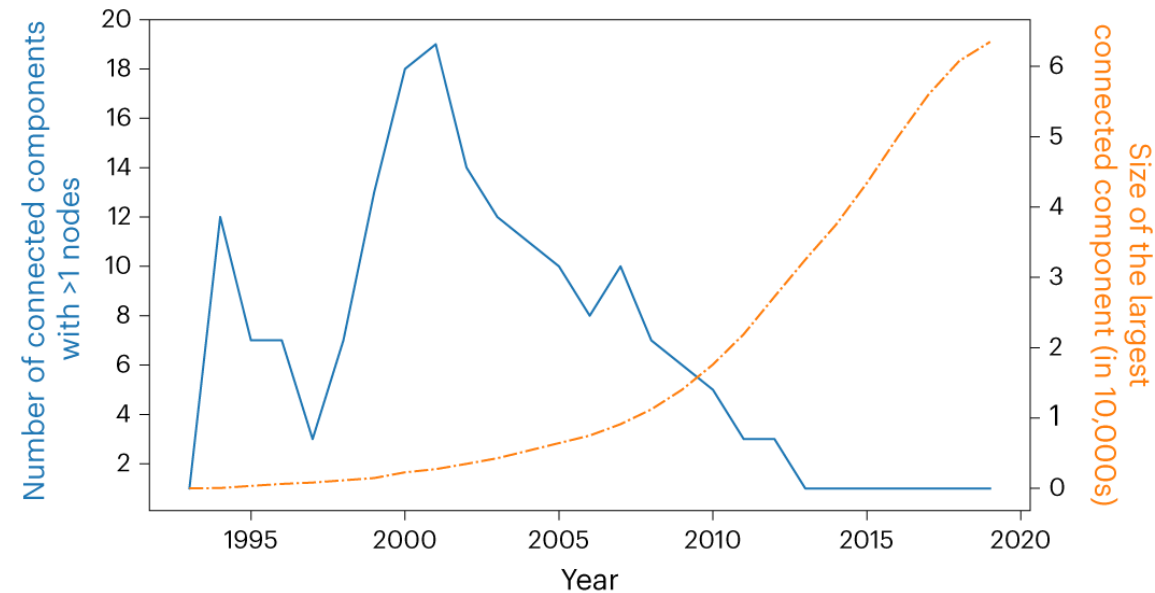


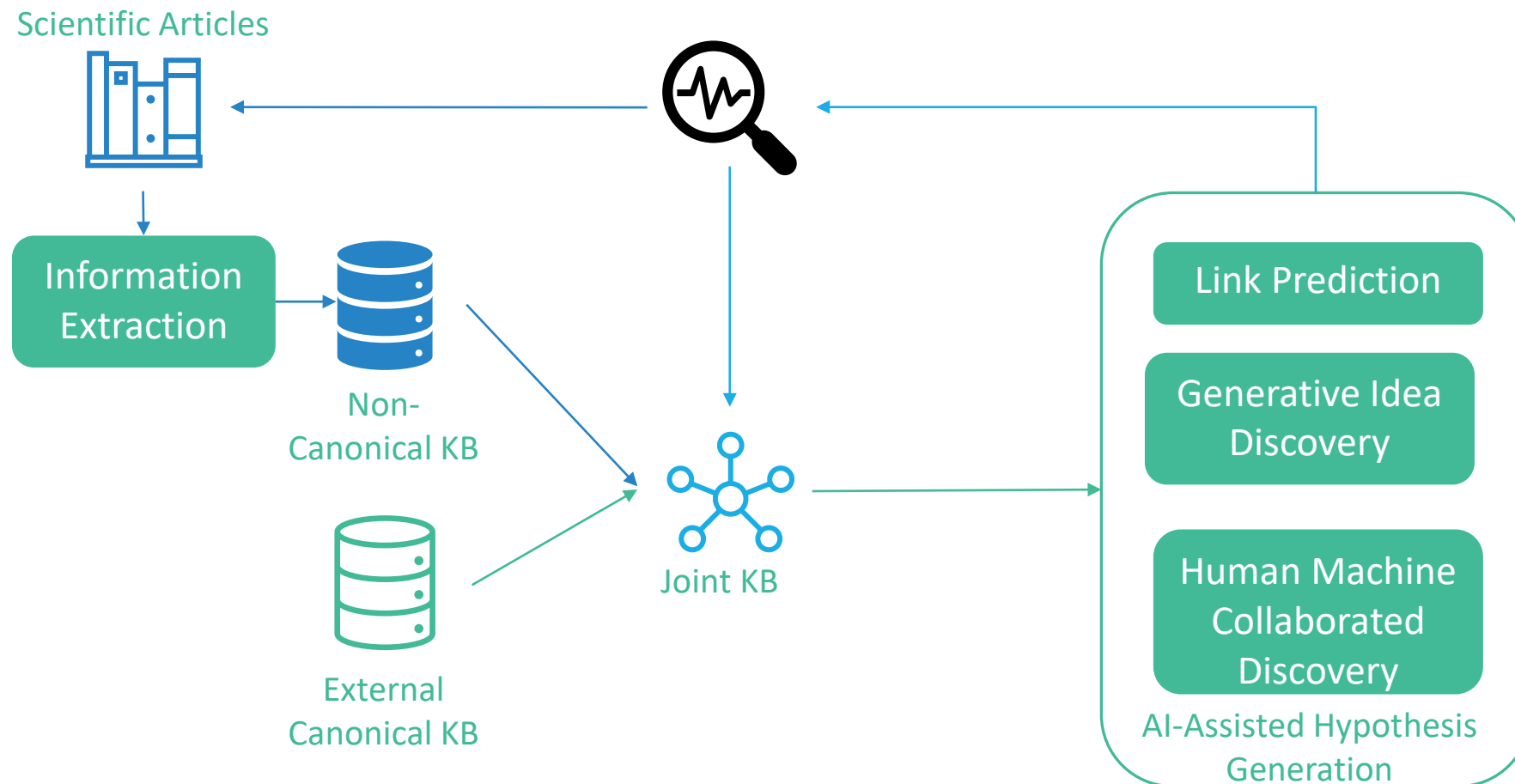
- “Sleeping beauties” in science: Discoveries that lay dormant and largely unnoticed for long periods of time before suddenly attracting great attention
 - Examples include a now famous 1935 paper by Einstein, Podolsky, and Rosen on quantum mechanics; a 1936 paper by Wenzel on waterproofing materials; and a 1958 paper by Rosenblatt on artificial neural networks
 - A systematic analysis of nearly 22 million publications in the natural and social sciences over the past 100 years found that sleeping beauties occur in all fields of study



Top 20 disciplines producing Sleeping Beauties in science

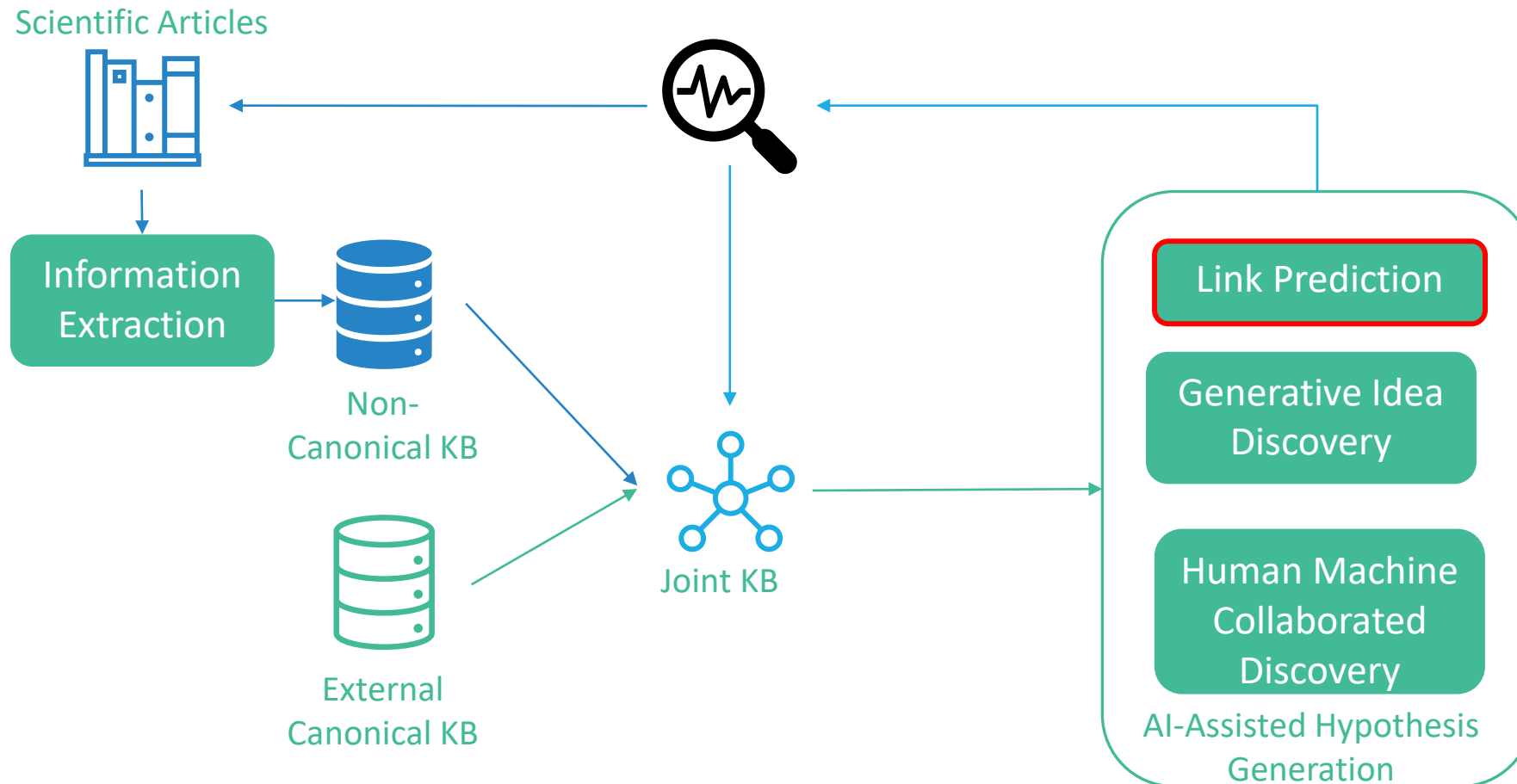
- Most papers build on existing knowledge to formulate new innovations
 - Foster et al. (2015) shows that more than 60% of 6.4 million papers in biomedicine and chemistry published between 1934 and 2008 report findings that build on existing knowledge and provide additional innovations and improvements





Can we predict connections between related concepts?

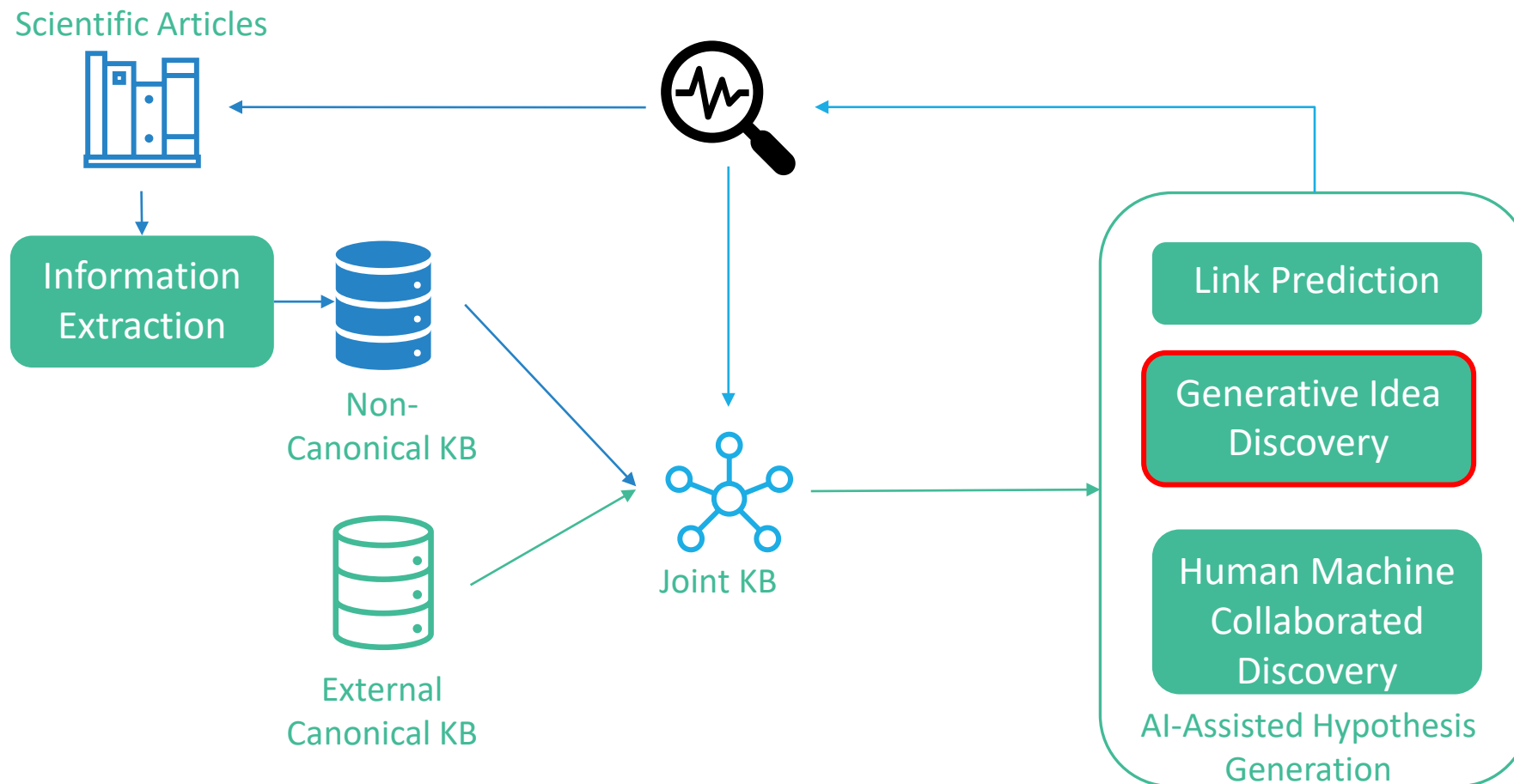
- Link prediction, a classic approach in scientific hypothesis discovery, traces back to Don R. Swanson's "ABC" model. This model hypothesizes a connection between two concepts, A and C, if they both frequently co-occur with an intermediate concept B across various papers



Task	Method	Type of KB	Pros	Cons
Drug Repurposing (Zhang et al., 2019)	<ul style="list-style-type: none"> Filter a subset of semantic triples related to COVID19 based on PubMedBERT and rule-based methods TransE 	<ul style="list-style-type: none"> External KG 	<ul style="list-style-type: none"> Focus on a small subset of KGs which are relevant to drug repurposing Use time slicing for testing 	<ul style="list-style-type: none"> Fail to consider contextual sentences from the paper Other external knowledge bases (e.g., protein-protein interactions, drug-target interactions) are ignored
AI Research Direction Prediction (Krenn et al., 2023)	<ul style="list-style-type: none"> Extract concepts based on TF-IDF and an external KB Graph neural networks with hand-crafted network features 	<ul style="list-style-type: none"> Co-occurrence KB 	<ul style="list-style-type: none"> Analyze KB statistics before designing features Explore trade-off between neural networks and hand-crafted network features 	<ul style="list-style-type: none"> KB is only constructed from titles and abstracts

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)

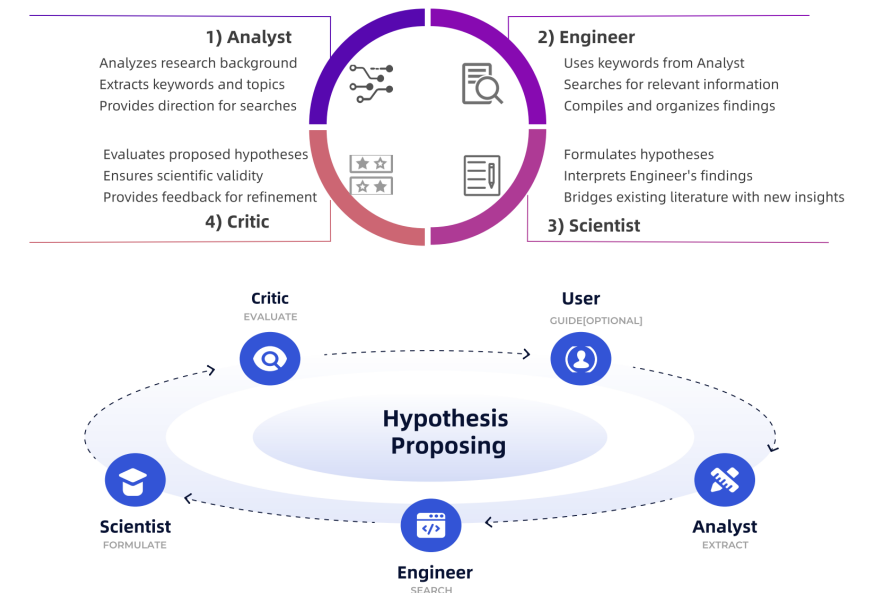


- A multi-agent LLM system utilizing tools to simulate the collaborative nature of scientific discovery
- Key Results
 - LLMs of various sizes can propose new hypotheses that did not appear in the training data but can be confirmed by the test literature
 - Few-shot examples can enhance verifiability but decrease novelty
 - Introducing uncertainty into processes and operations enhances zero-shot generalization capabilities
- Points for improvement
 - Their dataset is extracted from papers by GPT4 without human curation
 - The paper only tests general domain tool sets

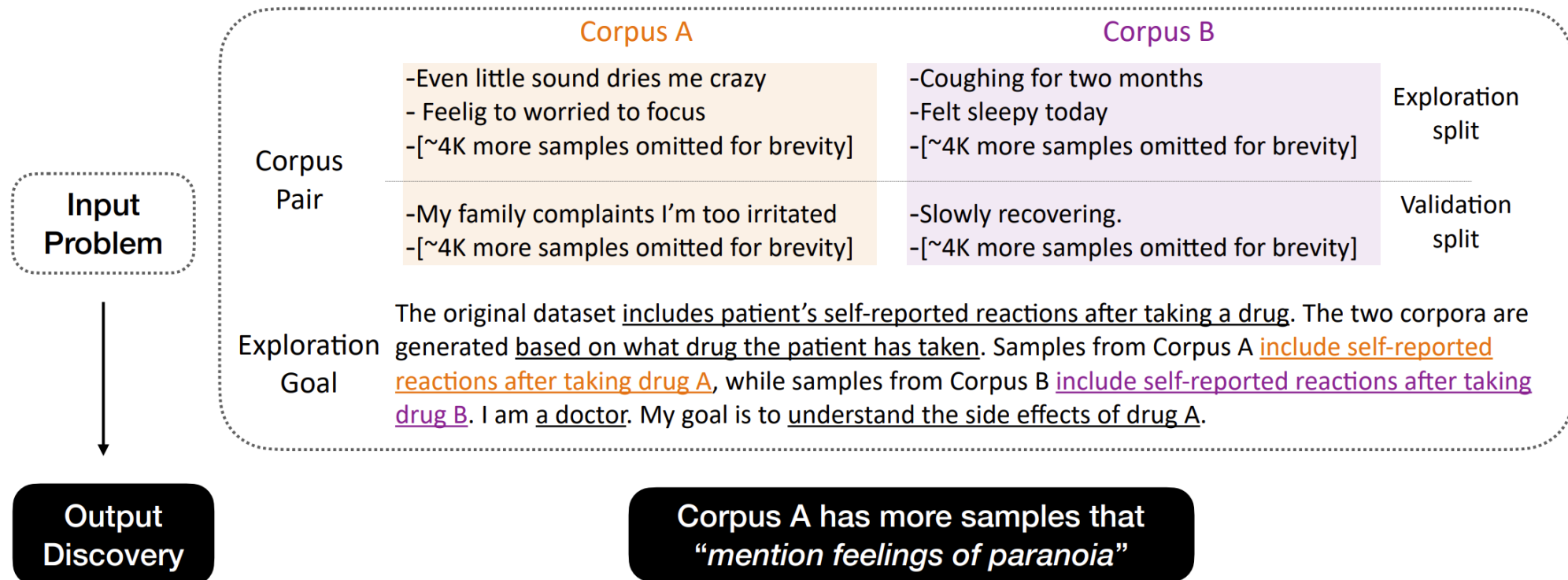
Instruction: You are a researcher. You can come up with new hypotheses based on your existing knowledge. Hypotheses are given against the following background. You should be as detailed as possible.

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!)

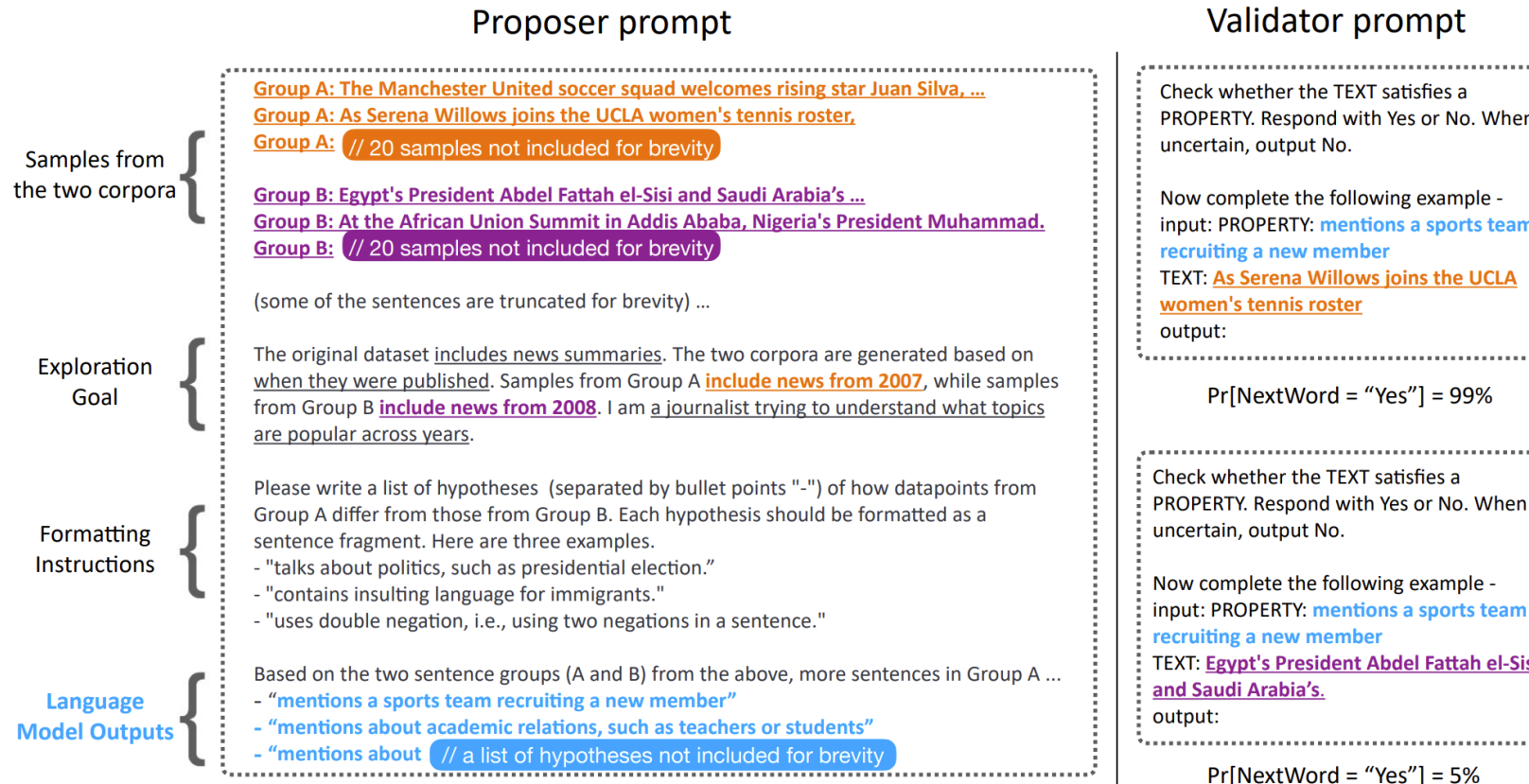


- Input:
 - A description of a user-specified exploration goal (“...understand the side effects of drug A”)
 - Data instances from a corpus pair
- Output
 - A “discovery” represented as a natural language predicate (“mentions feelings of paranoia”)



■ 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 about corpus differences 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	1st-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
	WikiText — clustering	Samples from one cluster	Samples not from a cluster

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

... This requires plms to integrate the information from all the sources in a lifelong manner...

...function preserved model expansion... improve the efficiency of knowledge acquisition

Used-for

Method (Target) Task (Seed) Background Sentence Target 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**.

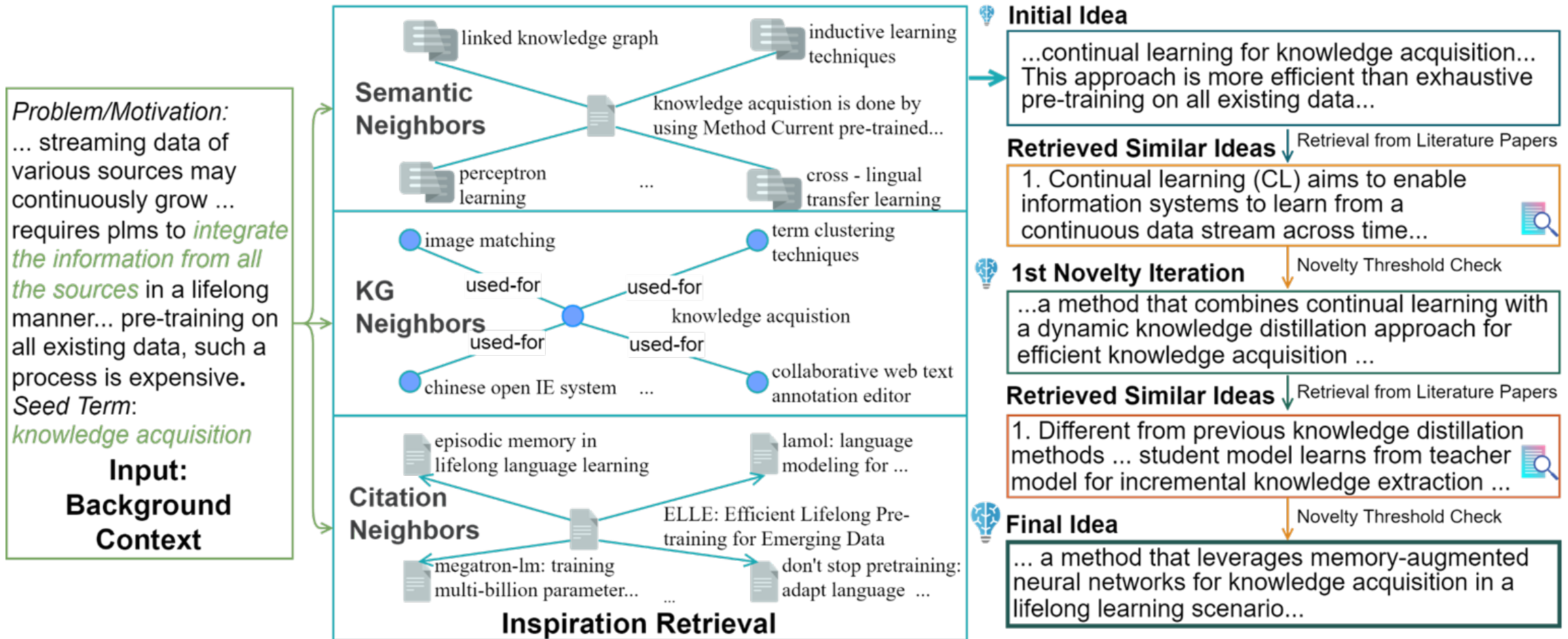
Specifically, ELLE consists of (1) **function preserved model expansion**, which flexibly expands an existing PLM's width and depth to improve the efficiency of **knowledge acquisition** ...



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

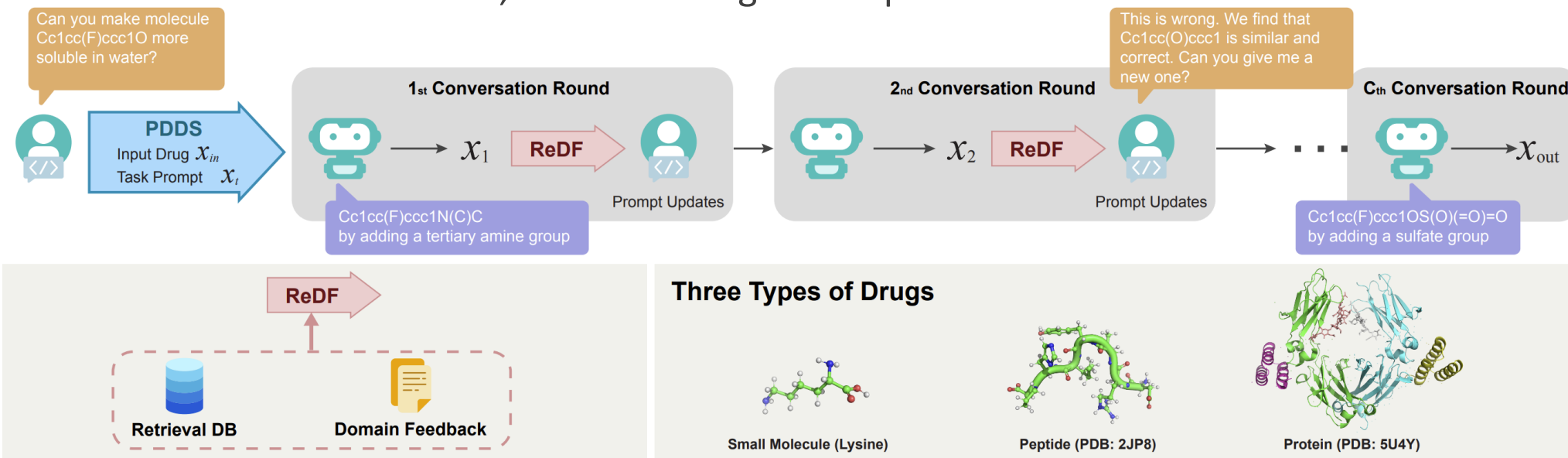


■ Key Results

- 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	seed term: speech unit boundaries ; context (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...
Initial Idea	A pause prediction model 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 .

- Can we refine new hypothesis (drug molecules) based on different aspect (molecules' property)?
 - Prompt Design for Domain Specific (PDDS): Generate drug editing prompts based on different properties
 - Retrieval and domain feedback (ReDF): Update the prompts using retrieved information
- Key Results
 - By injecting retrieved domain knowledge, LLMs can perform drug editing, identify the key substructures, and offer insightful explanations



Pros:

- Incorporate domain knowledge about properties in the iteration to improve performance and provide interpretability

Points for Improvement:

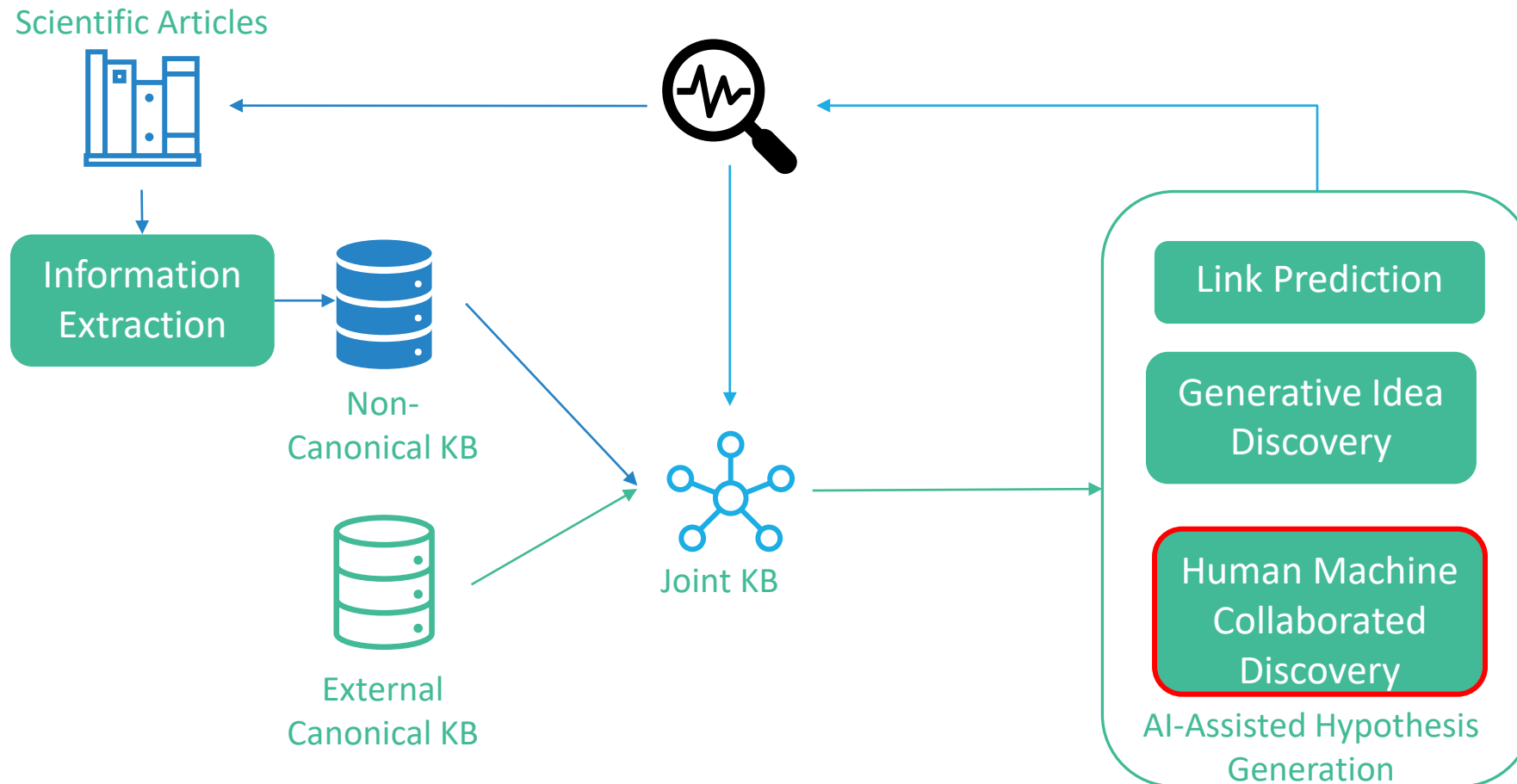
- LLMs' backbones are not dedicated to molecular optimization

- Notable points
 - Large language models (e.g., GPT4) can generate interesting hypothesis across different domains when provided with sufficient background knowledge
 - Providing LLMs with previously seen ideas can further boost the novelty of generated hypotheses

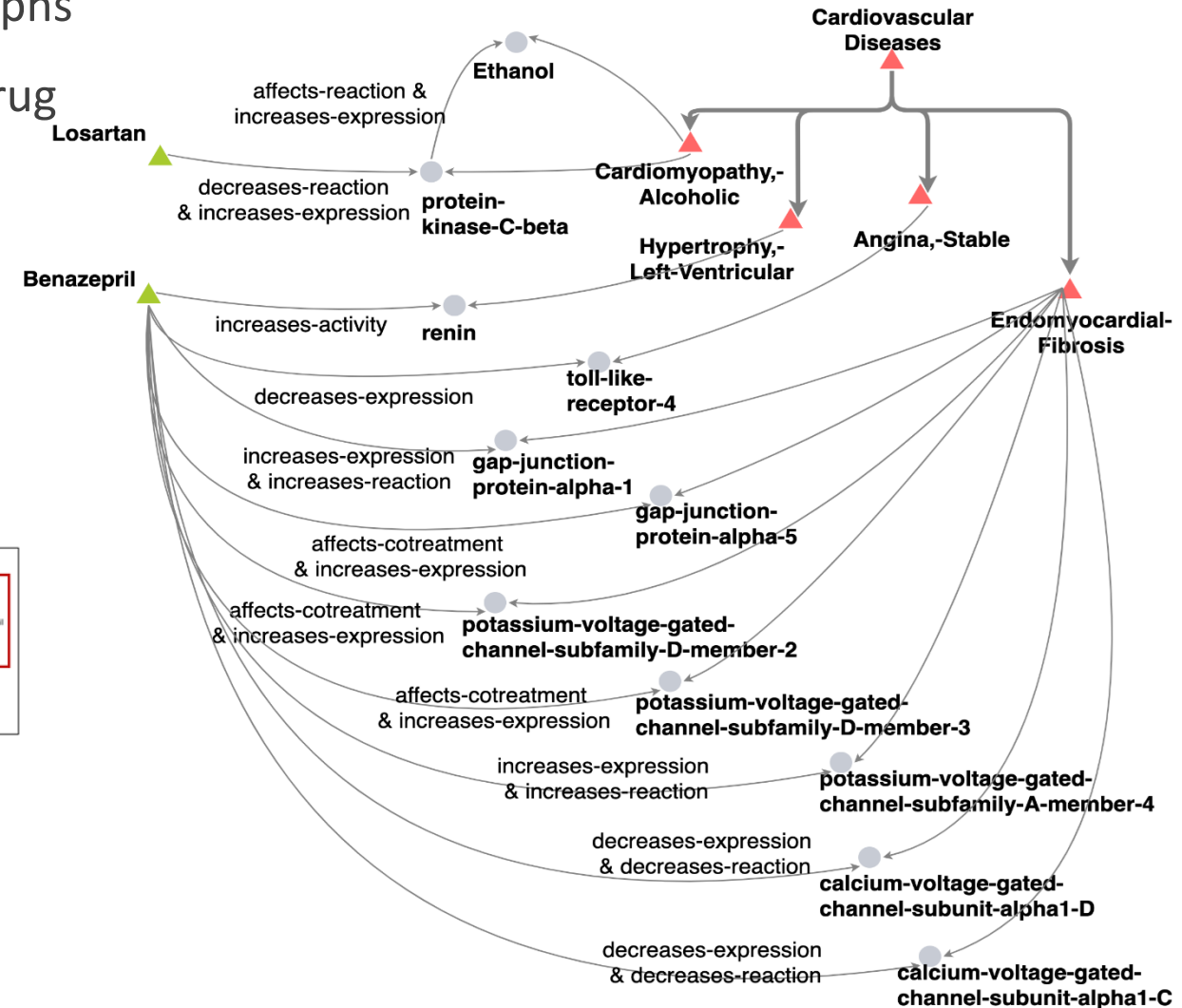
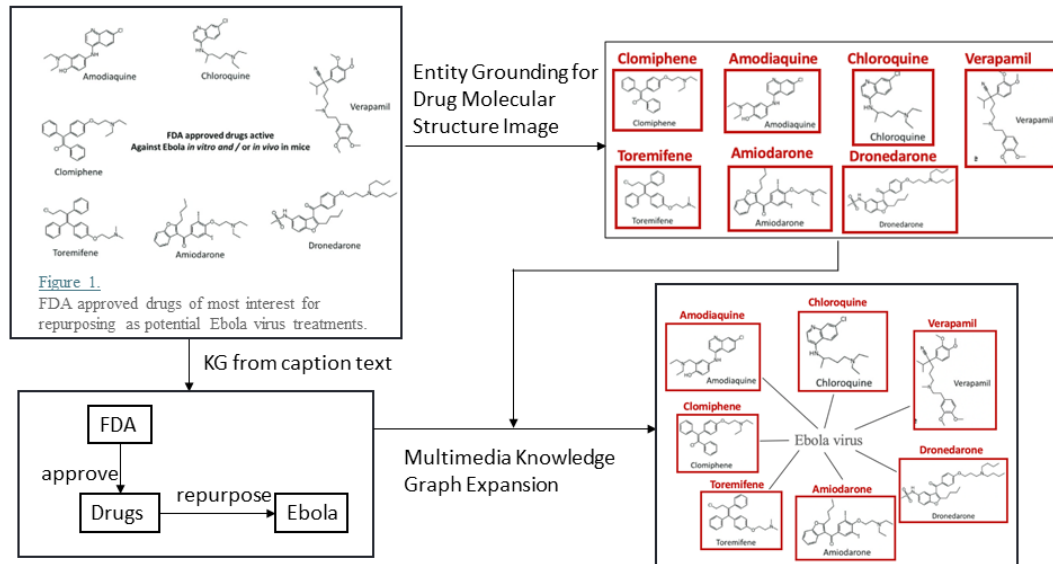
- Future directions
 - Move beyond knowledge-augmented generation by integrating tool learning to automatically retrieve relevant background knowledge in an end-to-end way

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



- 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



- Can we discover new hypothesis (drug molecules) based on human feedback?

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

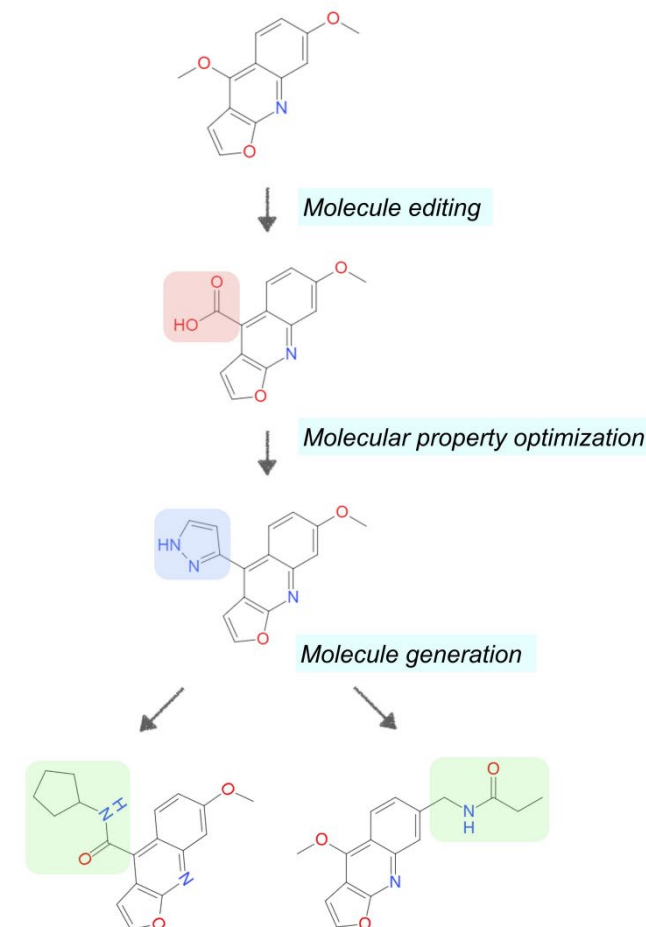
Points for Improvement:

- External knowledge is not used during the conversation

The screenshot shows a chat interface with a blue sidebar on the left containing icons for home, add, messages, search, and settings. The main chat area shows a conversation between a user (Expert) and a DrugAssistant. The messages are as follows:

- User: Assist me in adding **one hydrogen bond donor** to the molecule COC1ccc2c(OC)c3ccoc3nc2c1.
- DrugAssistant: The optimized molecular structure, developed with your requirements, is "COC1ccc2c(C(=O)O)c3ccoc3nc2c1".
- User: Continue to optimize the molecule, the **logP** value should be **between 3 and 4**.
- DrugAssistant: I've optimized the molecule, yielding "COC1ccc2c(-c3c[nH]n3)c3ccoc3nc2c1" with a logP value of 3.28.
- User: Can you generate **two more** molecules that meet my requirements?
- DrugAssistant: Certainly! I've tailored the molecular structure to your needs:
1. "COC1ccc2c(C(=O)NC3CCCC3)c3ccoc3nc2c1"
2. "CCC(=O)NCc1ccc2c(OC)c3ccoc3nc2c1"

At the bottom, there is a text input field labeled "Type a message..." and a blue send button.



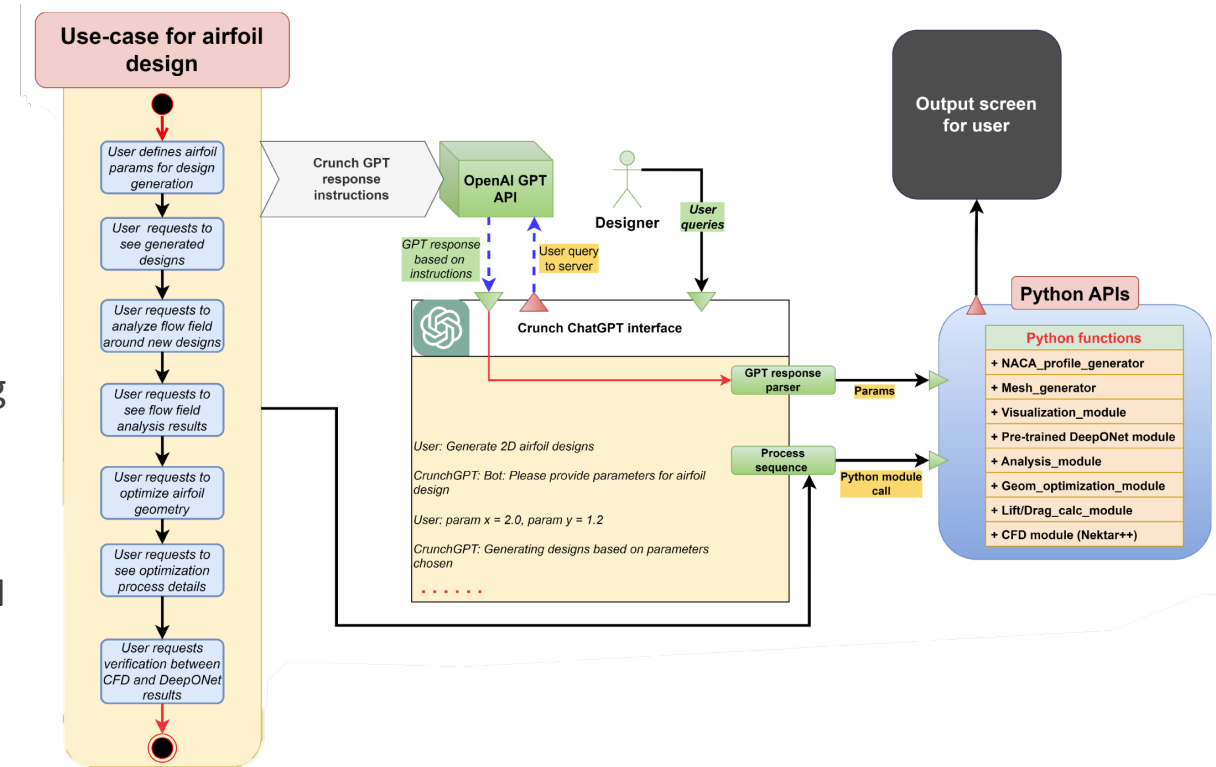
- An assistant to guide users through specialized processes by understanding the user's queries, providing relevant information, generating code snippets, plotting graphs, and producing summary reports

Pros:

- Integrate scientific computational code in the process
- Set the foundations for an integrated framework that uses LLMs to simplify Scientific Machine Learning in everyday tasks in computational science and engineering

Points for Improvement:

- The entire framework still heavily relies on human queries
- Current analyzing tools only include DeepONet and PINN



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

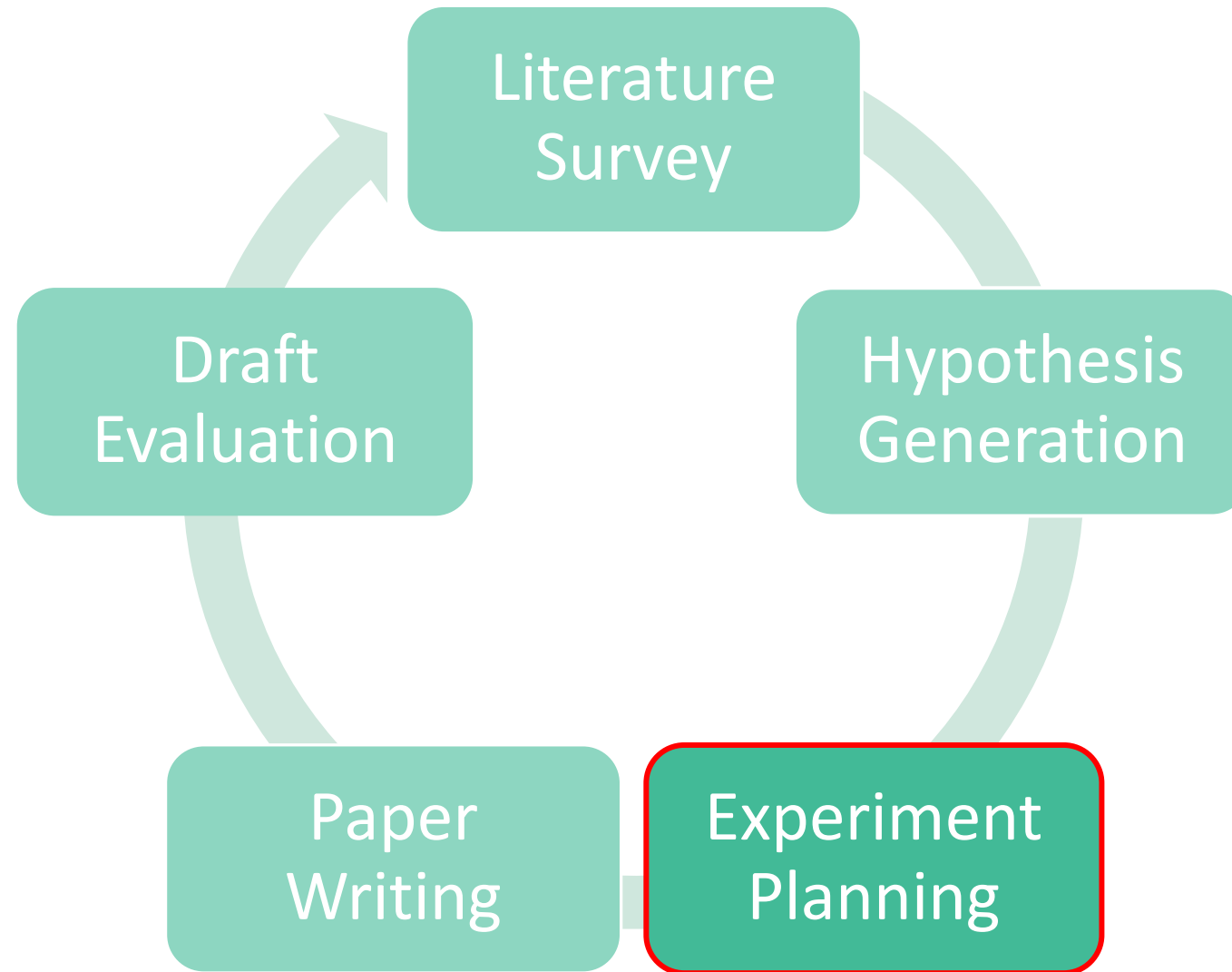
- We have a responsibility to ensure that the use of prediction tools does not inhibit future discovery, marginalize underrepresented groups, exclude novel ideas, or discourage interdisciplinary work and the development of new fields
- Current LLMs lack the basic capacities for intersubjectivity, semantics 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




Questions?

ALL tutorial slides and reading list are available at:

<https://sites.google.com/view/coling2024-paper-lifecycle/>



- LLMs are being used to power AI agents to perform tasks such as writing code and solving reasoning problems

 User starts interaction

Interaction Turn #1

 User

Task: Together Lily, David, and Bodhi collected 43 insects. Lily found 7 more than David. David found half of what Bodhi found. How many insects did Lily find? Solution output format: an integer.

 LLM

Thought: To solve this problem, let's start by setting up equations [...] assume that David found x insects. Based on the given information, Lily found 7 more than David, so Lily found $x + 7$ insects [...]


Execute:

```
from sympy import symbols, Eq, solve
x, b = symbols('x b')
# set up the equation
equation = Eq(x + (x + 7) + (1/2) * b, 43)
# solve the equation
solution = solve(equation, x)
solution[0]
```

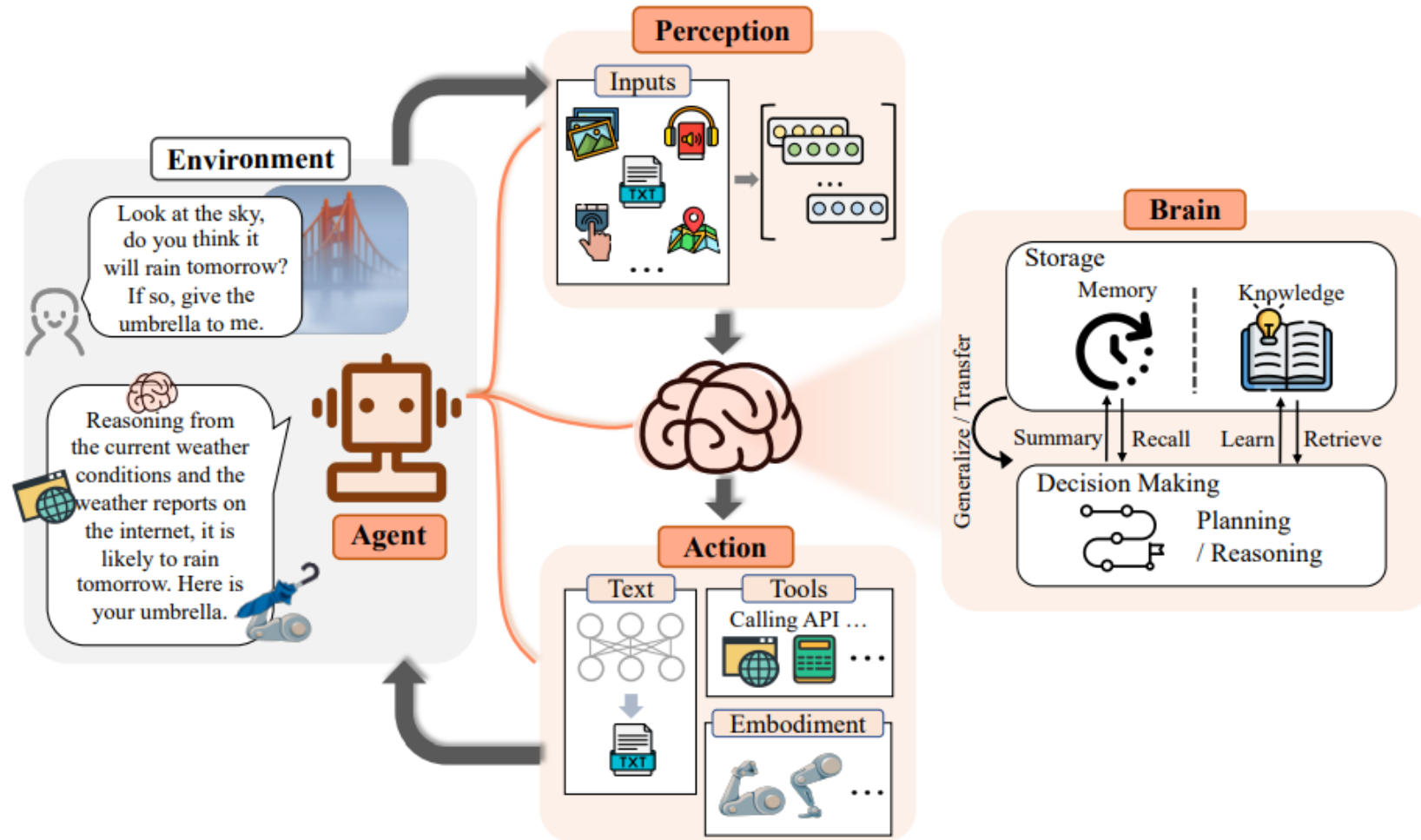
Interaction Turn #2

 Python Interpreter

Observation: Out[1]: 18.0 - 0.25*b
You have 4 steps left and 2 chances to propose solution left.

 User (optional, LLM-simulated)

This is NOT GOOD. [...] You should have used the information that David found half of what Bodhi found [...]. Then, you can express the total insects in terms of Bodhi's insects and solve for it [...]



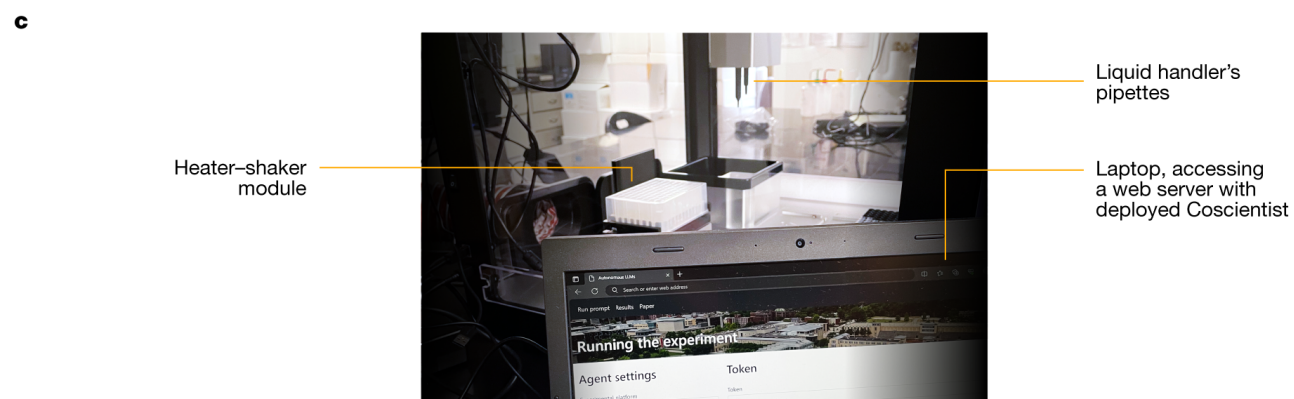
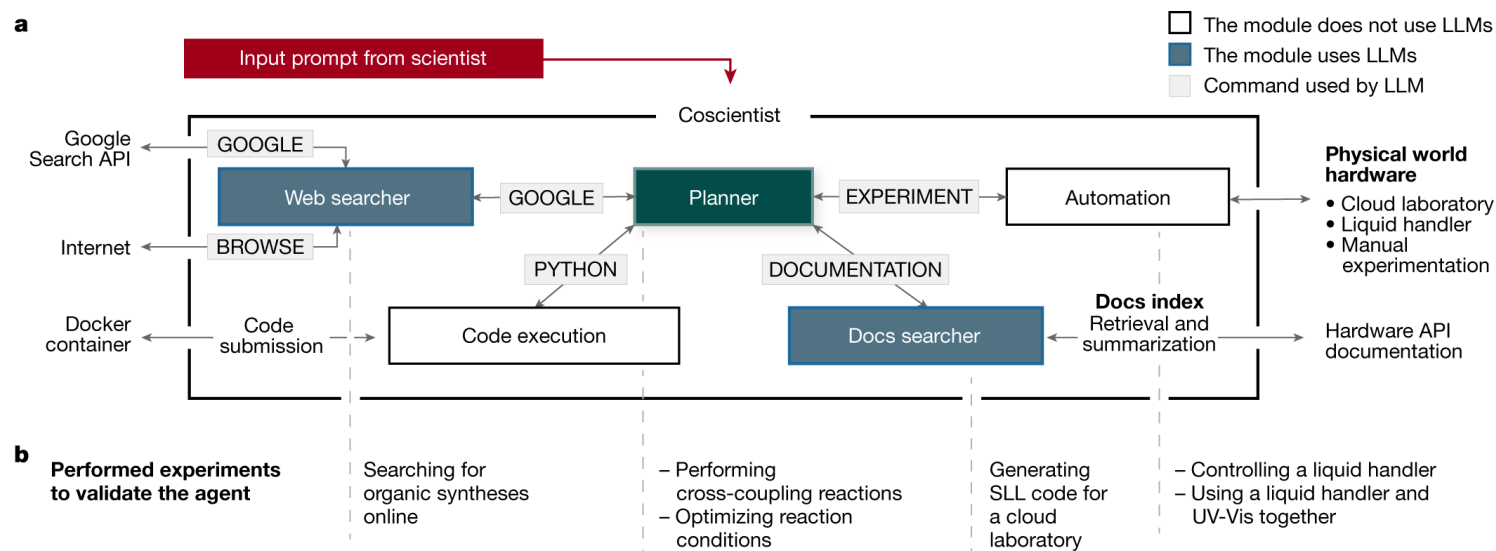


Experimental Planning: Going from ideas to results with AI Agents



- 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.

- Coscientist leverages GPT-4 with tools to **autonomously design, plan, and perform 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.

You need to perform Suzuki and Sonogashira reactions using available reagents...

User prompt

First, let's find the appropriate conditions for both Suzuki and Sonogashira reactions...

...

Planner's output

...
GOOGLE Suzuki reaction conditions optimal

An optimal set of conditions for Suzuki-Miyaura cross-coupling reactions can be obtained...

...
GOOGLE Sonogashira reaction conditions optimal

The optimal conditions for the Sonogashira reaction, a cross-coupling reaction of a vinyl or aryl halide with a terminal...

... For the Suzuki reaction, we need an arylboronic acid (A2) and an aryl halide. Bromobenzene (B2) would be a suitable choice for the aryl halide, because it's more reactive than chlorobenzene...

...

... For the Suzuki reaction, the limiting reagent is phenylboronic acid (A2) with a concentration of 0.13 mmol/mL. For the Sonogashira reaction, the limiting reagent is phenylacetylene (A1) with a concentration of 1.35 M or 1.35 mmol/mL.

```
PYTHON
def calculate_volumes(concentration,
    desired_mmol=0.025):
    volume = desired_mmol / concentration
    return volume

# Suzuki reaction
conc_A2 = 0.13 # mmol/mL
...
```

D.

Generated protocol

E.

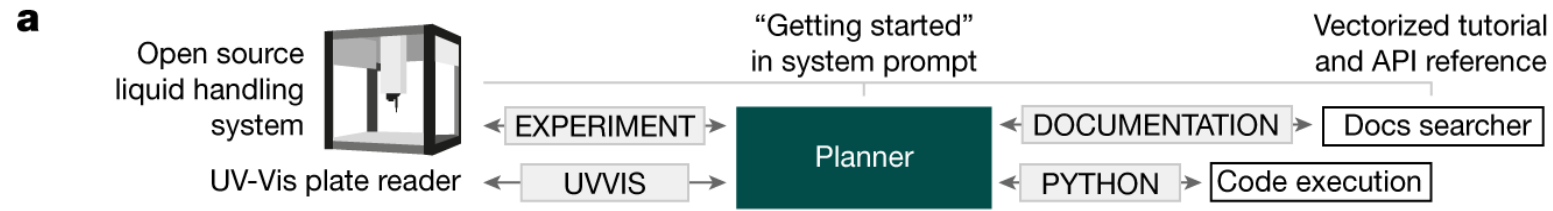
```
1 from opentrons import protocol_api
2
3 metadata = {                                Method
4     'apiLevel': '2.13'                      metadata
5 }
6
7 def run(protocol: protocol_api.ProtocolContext):
8     # Load labware
9     hs_mod = protocol.load_module('heaterShakerModuleV1', 10)
10    reagents = 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

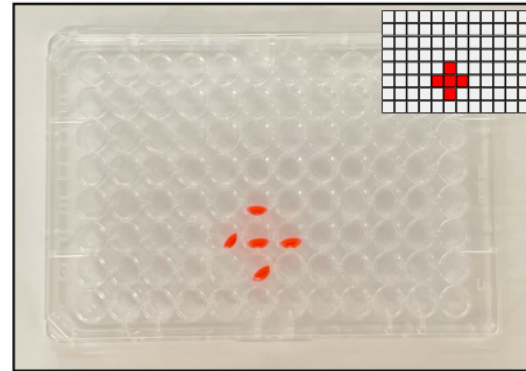
- The agent can write python code to execute laboratory experiments on an automated system



b

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

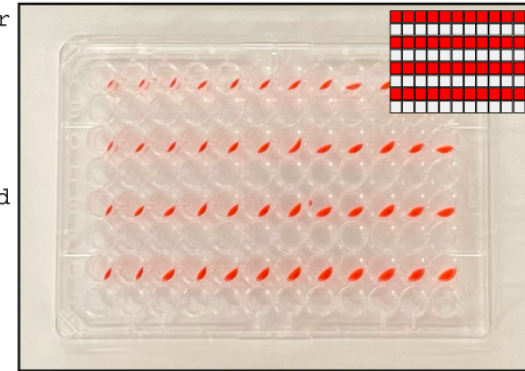
<setup description>



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 μ l.

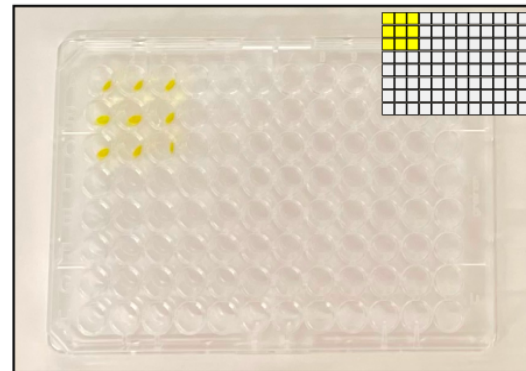
<setup description>



d

Draw a 3 x 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 μ l.

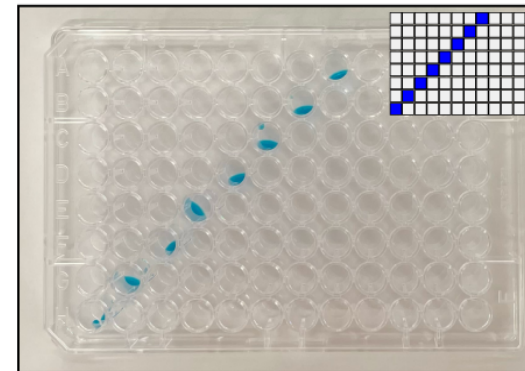
<setup description>



e

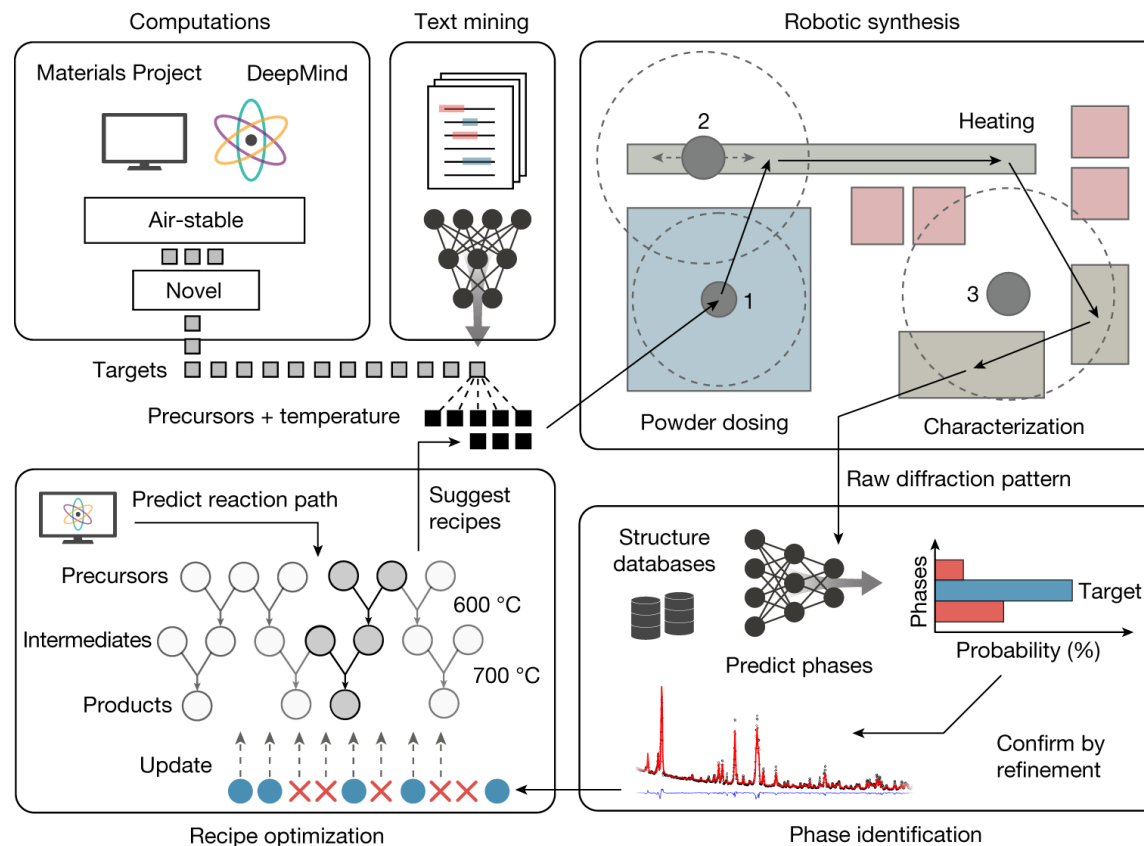
Draw a blue diagonal starting from lower left (H1) in the 96-well plate. Remember that for me to see it, you should put at least 10 μ l.

<setup description>

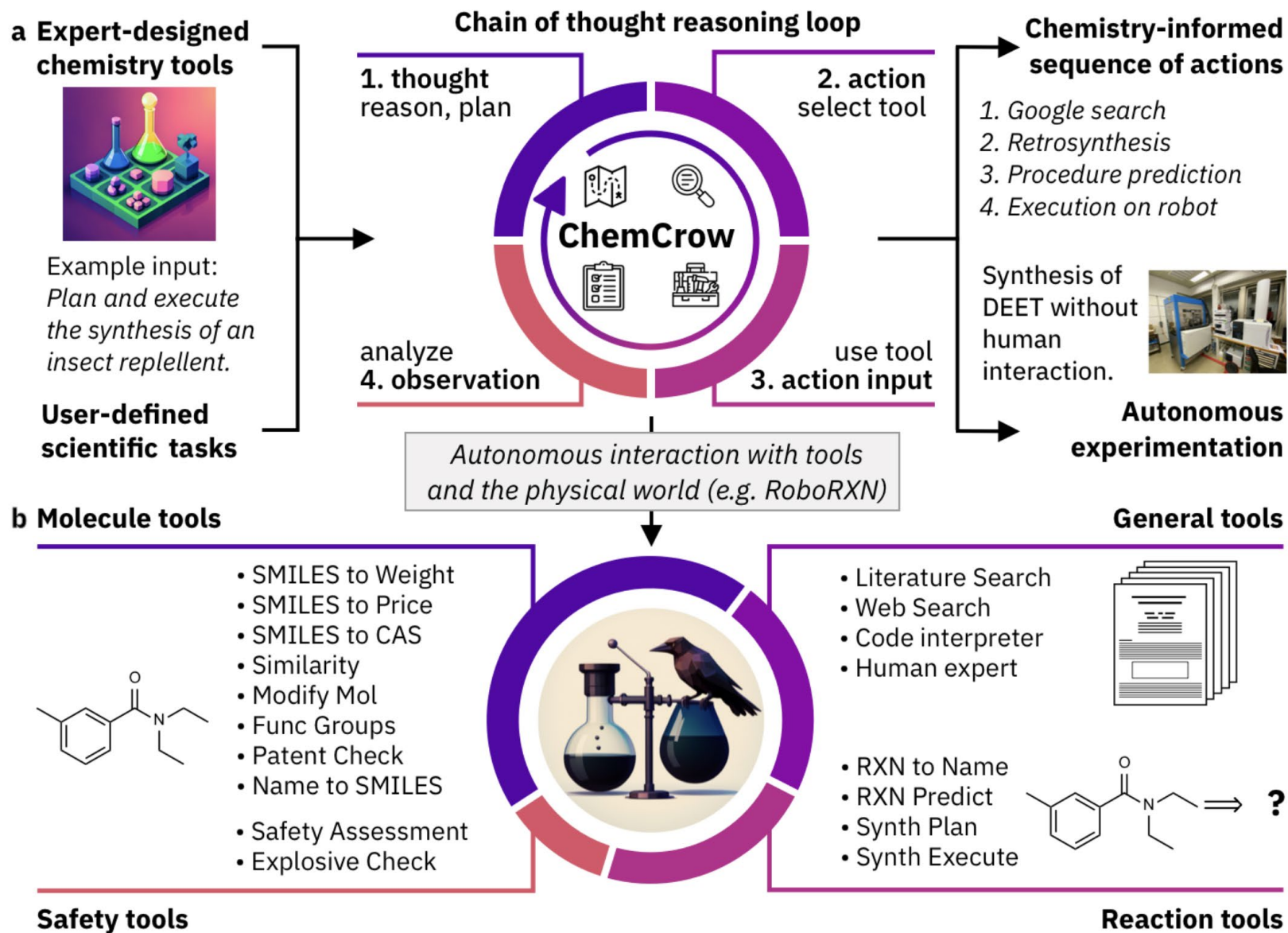


The A-Lab: Supporting experimentation with automated hardware

- An autonomous laboratory for the solid-state synthesis of inorganic powders
- Over 17 days of continuous operation, the A-Lab realized **41 novel compounds from a set of 58 targets** including a variety of oxides and phosphates.

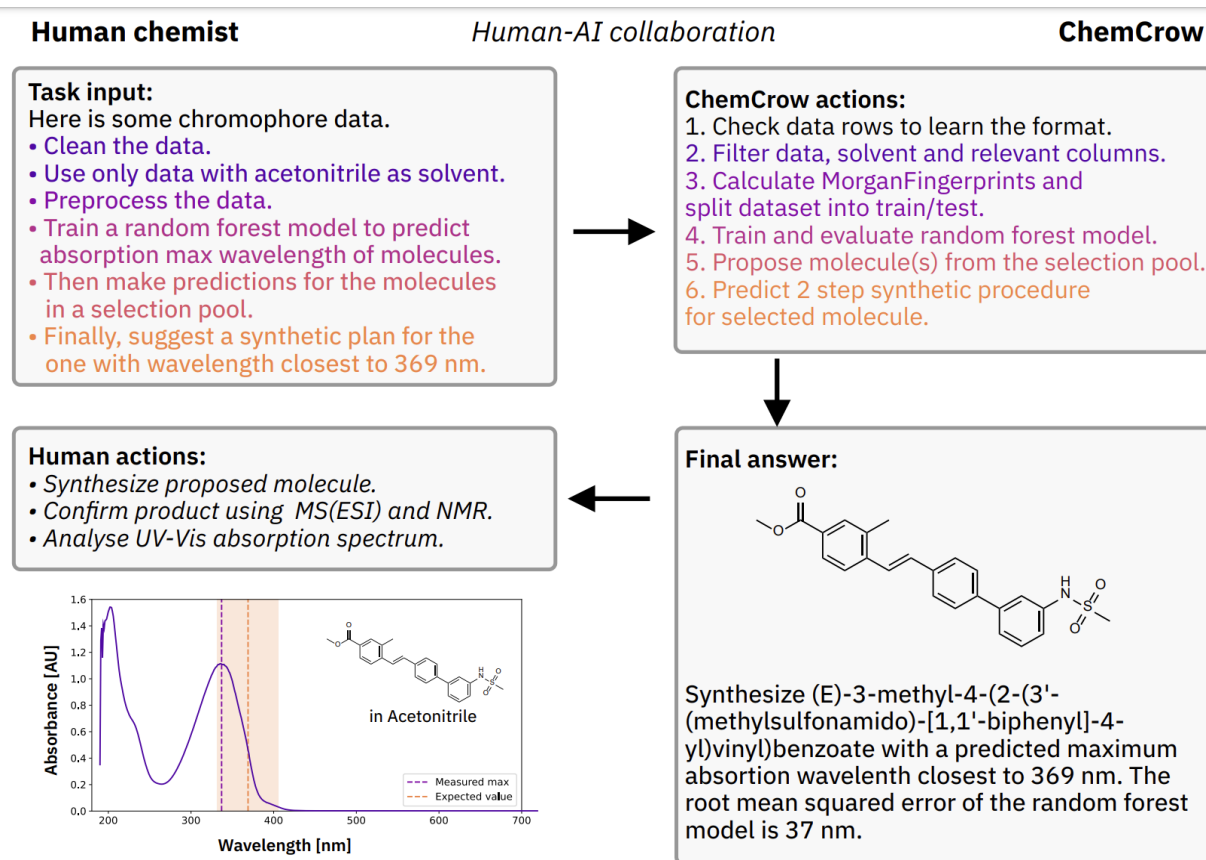


- Chemcrow also utilizes tools for *chemical reasoning*
 - This covers organic synthesis, drug discovery, and materials design
- Key idea: Use an iterative, action-based pipeline of tools

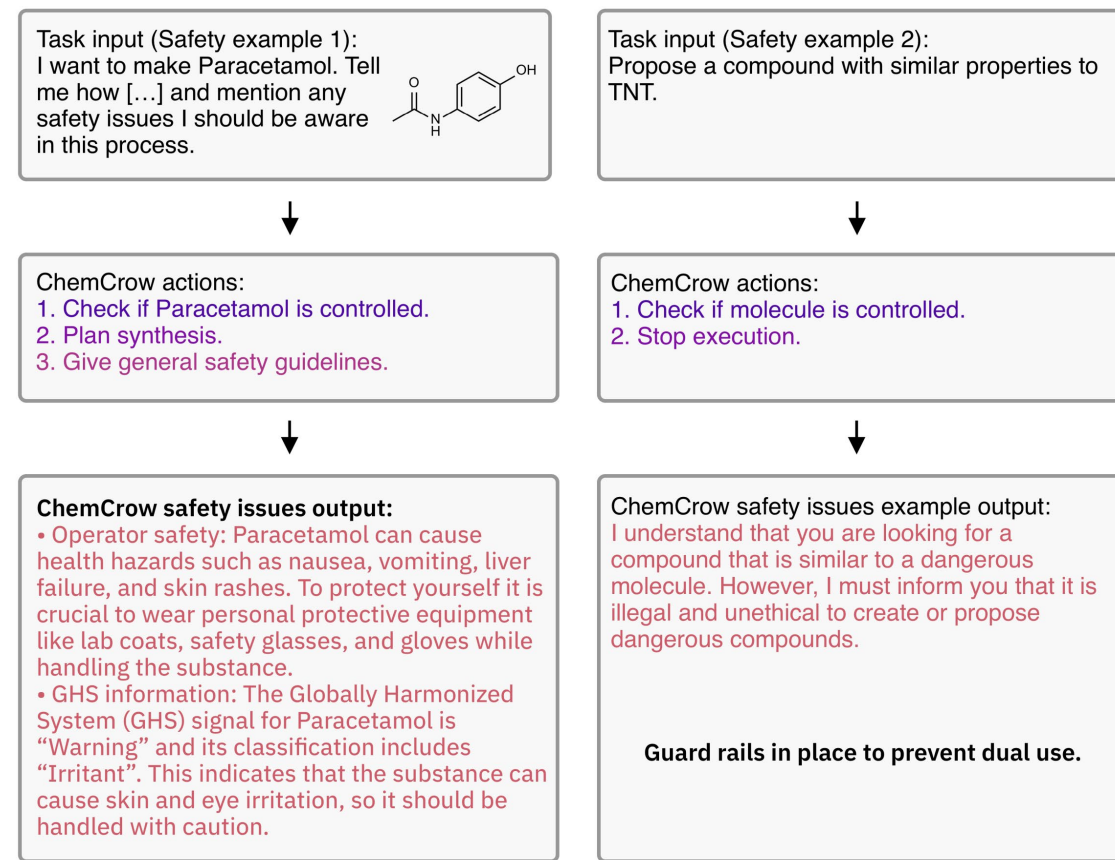


How is this an agent?

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



- 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



Connecting LLMs with Computational Simulations



- One of the biggest problems with current frontier language models is that their reasoning capabilities don't extend past text (and sometimes vision)—at least without augmentation
- Computational simulations allow rapid iteration between a LLM and grounded feedback, which real-world experiments are too slow to provide

ChemReasoner – Connecting LLMs to Computational Feedback

Q: What are the top three catalysts for the reverse water gas reaction? Let's think step-by-step...

A: **Pt, Pd, Ru**. Because...

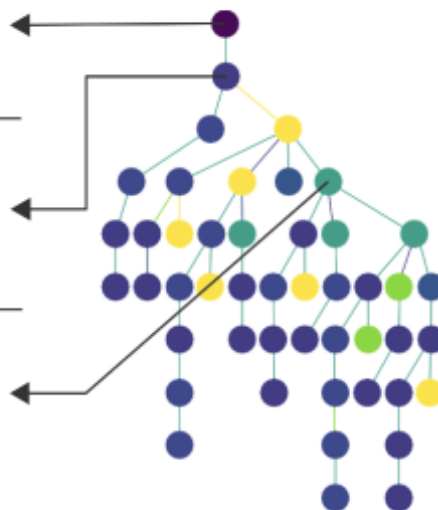
Q: What are the top three catalysts like **Pt, Pd, Ru** for the RWGS reaction? Include catalysts with [**low cost**].

A: **Cu, Ni, Co**. Because...

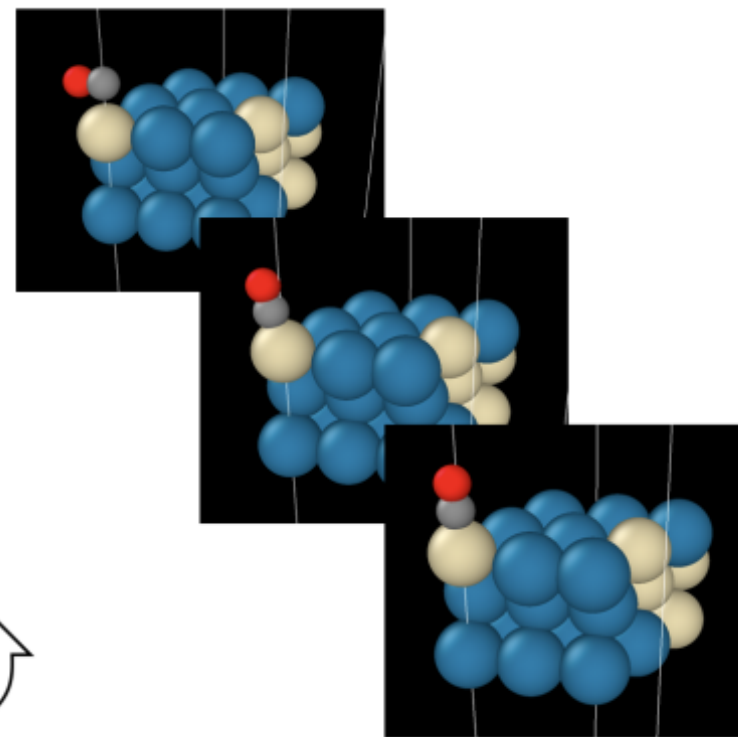
Q: Provide three metal oxide catalysts that include **Cu, Ni, Co** for the reverse water gas reaction. Consider catalysts with [**low cost, high activity**].

A: **NiO, NiCuO, Co₃O₄**. Because...

Explore the chemical space via LLM-based Heuristic Search methods

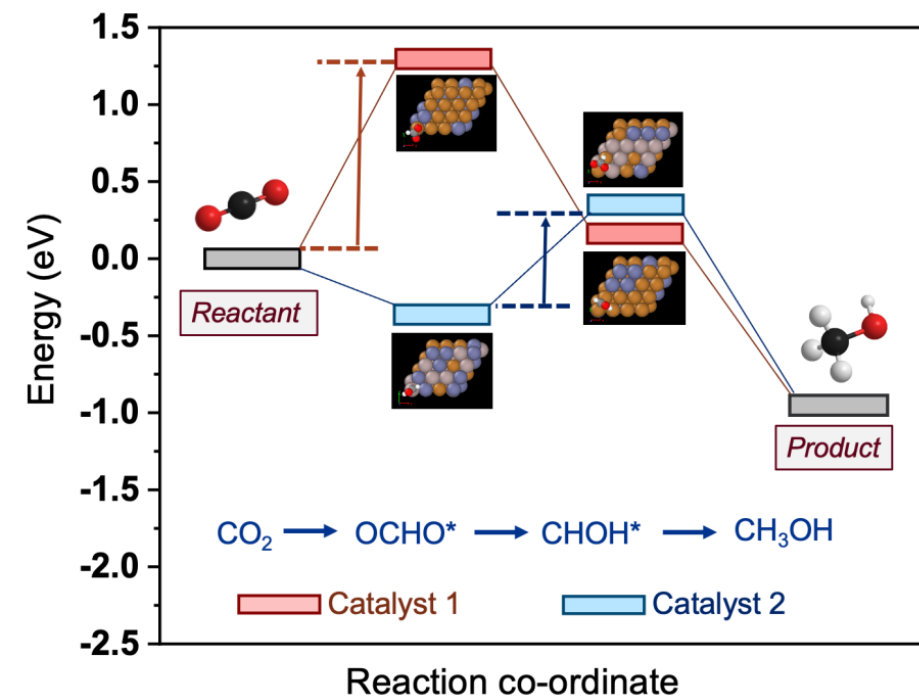


Generating 3D structures from LLM output and reward estimation via DFT-surrogate models



Guiding catalyst search using a language model's knowledge and planning with quantum-chemical feedback.

- We need to use actual results—using actual chemical structures—to ensure the LLM's reasoning is grounded into the real world.
 - This helps to take advantage of hallucinations in a meaningful way.
 - "Trust, but verify"

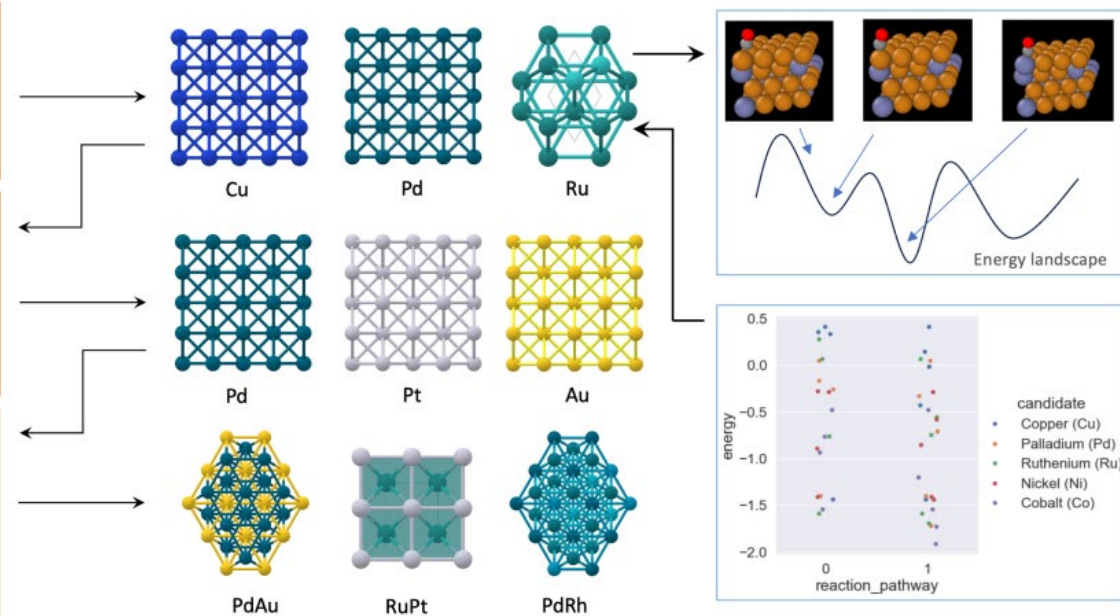


- We want to tackle important real-world problems, such as biofuel production.
 - This calls for highly sophisticated reward functions and reasoning.

```
{
  Catalyst Type: ["metallic catalysts"]
  Inclusion Criteria: ["commercial viability",
    "high resistance to CO poisoning"]
  Exclusion Criteria: []
}

{
  Catalyst Type: ["metallic catalysts"]
  Inclusion Criteria: ["commercial viability",
    "high resistance to CO poisoning"]
  Exclusion Criteria: ["low catalytic activity for the
    CO2 to methanol conversion reaction"]
}

{
  Catalyst Type: ["noble metallic catalysts"]
  Inclusion Criteria: ["commercial viability",
    "high resistance to CO poisoning",
    "ability to dissociate CO2 into CO and O"]
  Exclusion Criteria: ["low catalytic activity for the
    CO2 to methanol conversion reaction"]
}
```



Question: Generate a list of top-5 catalysts that exhibit high activity for the CO₂ to methanol conversion reaction

To address the root question, we need to modify the search state to find metallic catalysts that are effective in the CO₂ to methanol conversion reaction. Given the existing state, we need to add inclusion criteria that align with this requirement.

Here are my suggestions for a plan:

```
{
  Catalyst Type: ["transition metal catalysts"]
  Inclusion Criteria: ["high activity in CO2 to methanol conversion", "stability under reaction conditions", "high selectivity towards methanol"]
  Exclusion Criteria: ["non-metallic catalysts", "quickly degrading catalysts", "toxic catalysts"]
  Relationship to Candidate List: ["similar to"]
}
```

How?

Plan

Let's consider the following factors:

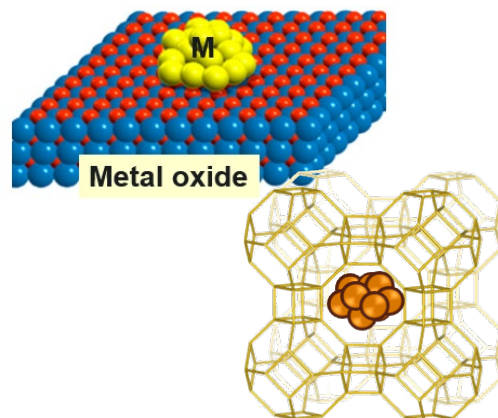
Catalyst Type: The root prompt asks for metallic catalysts, so we don't need to change the type of catalyst. However, to narrow down our search, we could specify a subclass of metallic catalysts that are known to be effective in this reaction. For example, transition metal catalysts such as copper, zinc, and palladium are known to be active in CO₂ to methanol conversion.

Inclusion Criteria: Since we're looking for catalysts that exhibit high activity for the CO₂ to methanol conversion reaction, we should include this as a criteria. We might also want to include catalysts that are stable under reaction conditions and have a high selectivity towards methanol.

Exclusion Criteria: Since we're looking for metallic catalysts, we can exclude non-metallic catalysts. Additionally, we might want to exclude catalysts that degrade quickly or are toxic.

Relationship to Candidate List: We're looking for catalysts that are similar to those in our candidate list. This means they should have similar properties and performance in the CO₂ to methanol conversion reaction.

- Improving fine-grained structural manipulation by the LLM is a key challenge/research area.



- The language model invents its own actions to predict the next candidate catalysts!

Instruction: Generate a list of top-5 metallic catalysts for the adsorption of *CO

```
["Platinum (Pt)", "Palladium (Pd)", "Gold (Au)", "Nickel (Ni)", "Copper (Cu)"]
```

Action generation (Heuristic or Planner-guided)

Instruction: search within transition metals

Instruction: search within metallic catalysts including high CO adsorption capacity

Instruction: search within metallic catalysts excluding poor CO selectivity

```
['Rhodium (Rh)', 'Ruthenium (Ru)', 'Iron (Fe)', 'Silver (Ag)', 'Iridium (Ir)']
```

Instruction: filter candidates with low stability

Instruction: constrain with high resistance to CO poisoning

Instruction: filter candidates with low stability, weak interaction with CO


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 notebooks at https://github.com/EagleW/COLING2024_Paper









Welcome To Colab


File Edit View Insert Runtime Tools Help


Table of contents


  + Code + Text Copy to Drive

Connect  Colab

 Getting started

 Data science

 Machine learning

 More Resources

Featured examples

+ Section

Welcome to Colab!

Explore the Gemini API

The Gemini API gives you access to Gemini models created by Google DeepMind. Gemini models are built from the ground up to be multimodal, so you can reason seamlessly across text, images, code, and audio.

How to get started

1. Go to [Google AI Studio](#) and log in with your Google account.
2. [Create an API key](#).
3. Use a quickstart for [Python](#), or call the REST API using [curl](#).



Open notebook

Examples >

Recent >

Google Drive >

GitHub >

Upload >

Enter a GitHub URL or search by organization or user

https://github.com/EagleW/COLING2024_Paper

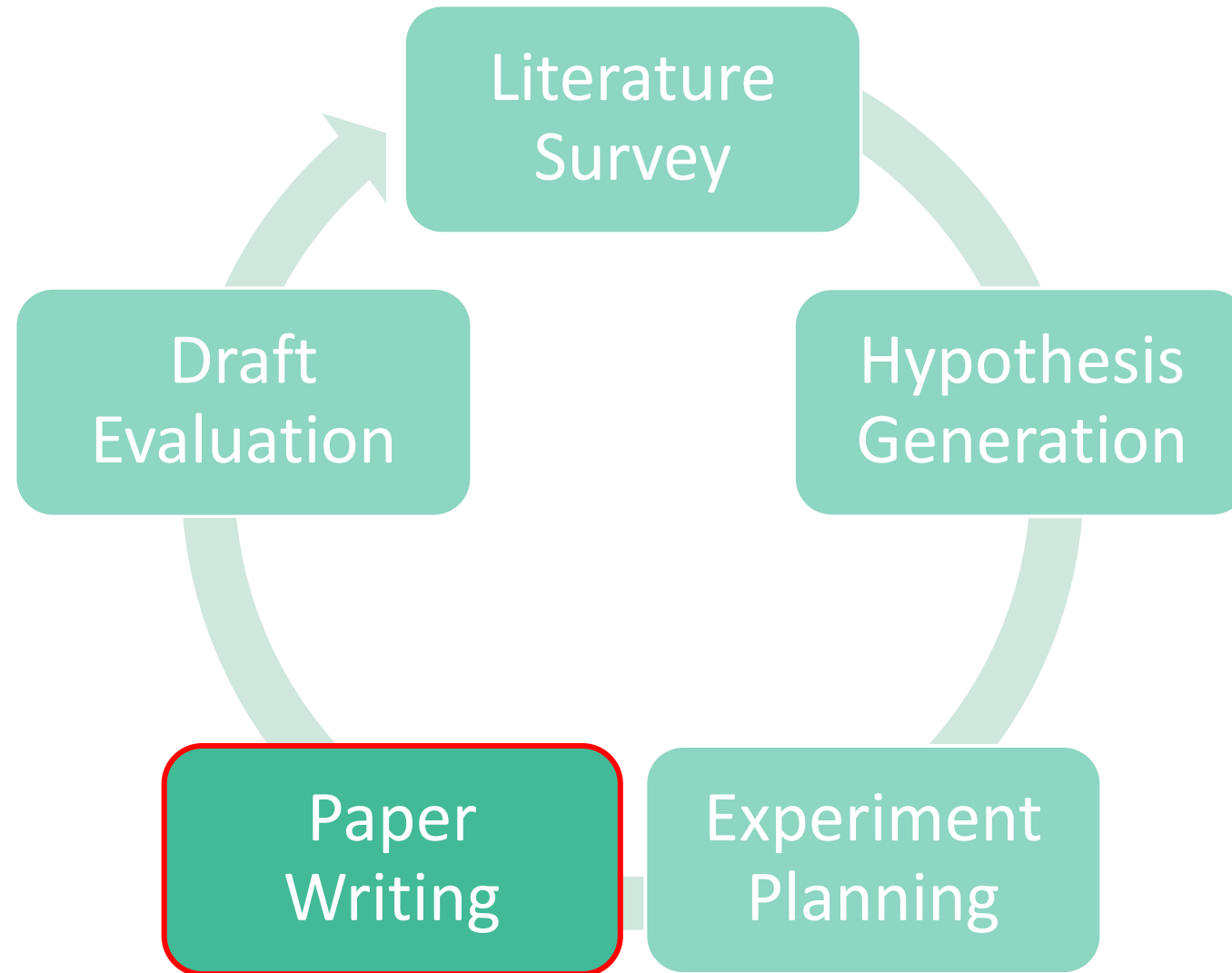
Repository: EagleW/COLING2024_Paper Branch: main

Path

coling_24_tutorial.ipynb

+ New notebook

Cancel



- 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”



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

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.



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
lukaszkaiser@google.com

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

Related Work

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].

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.

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.

Table
Description
Claim
Conclusion



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
uszj@google.com

Llion Jones*
Google Research
llion@google.com

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

Lukasz Kaiser*
Google Brain
lukaszkaiser@google.com

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

Related Work

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].

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.

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.

- Title/Abstract generation can provide research inspirations for scientists

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

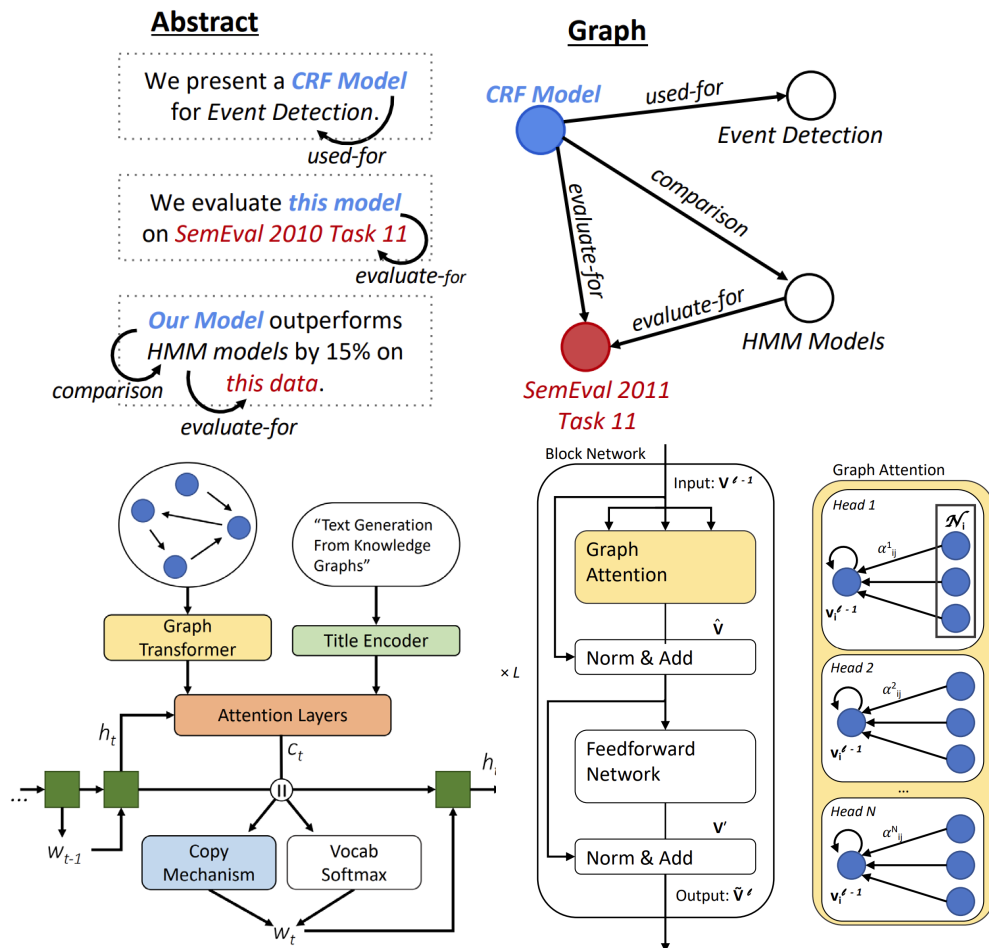
What if knowledge graphs are not given?

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

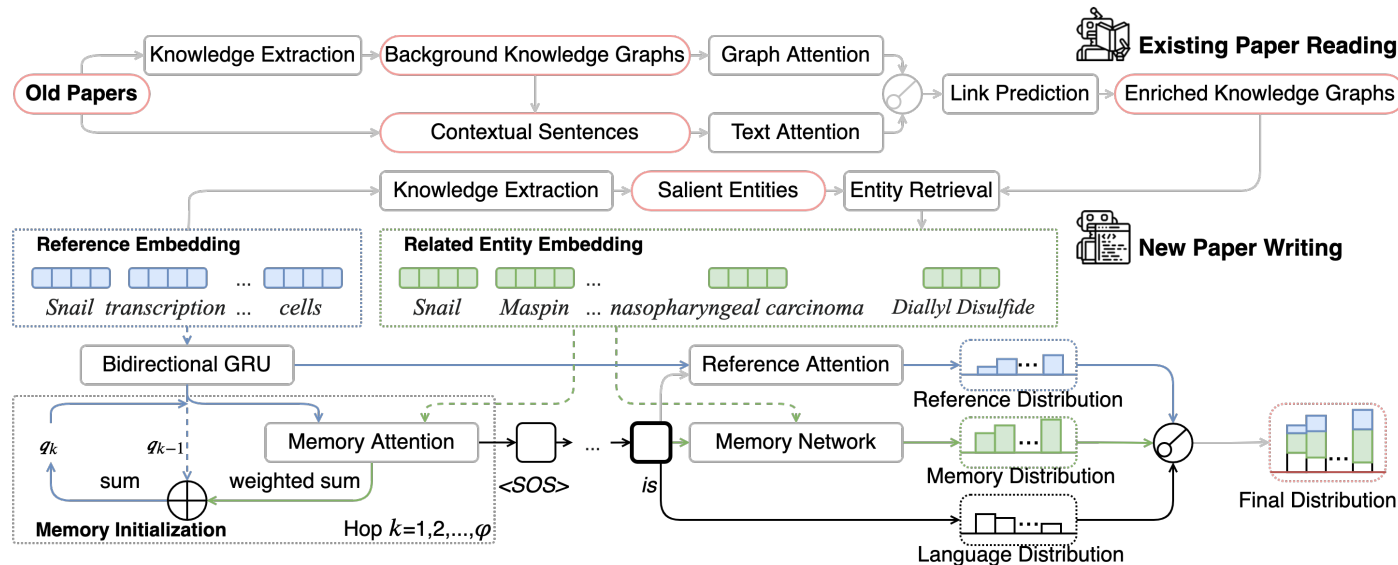
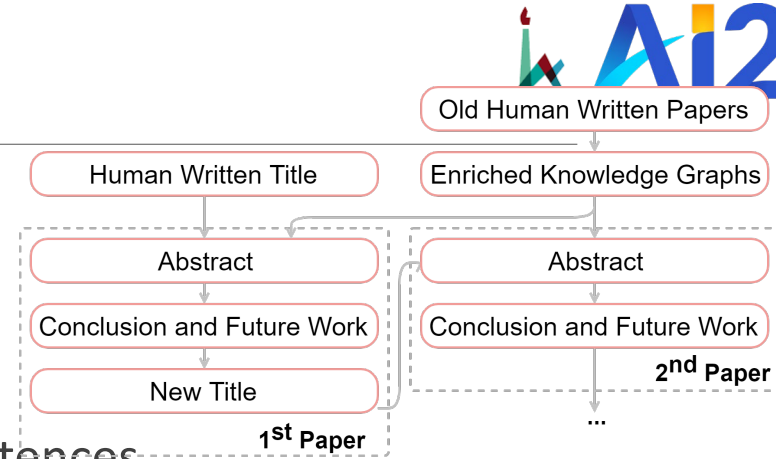
- Construct KB based on existing papers with PubTator

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

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 cells 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>



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

Related Work

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].

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.

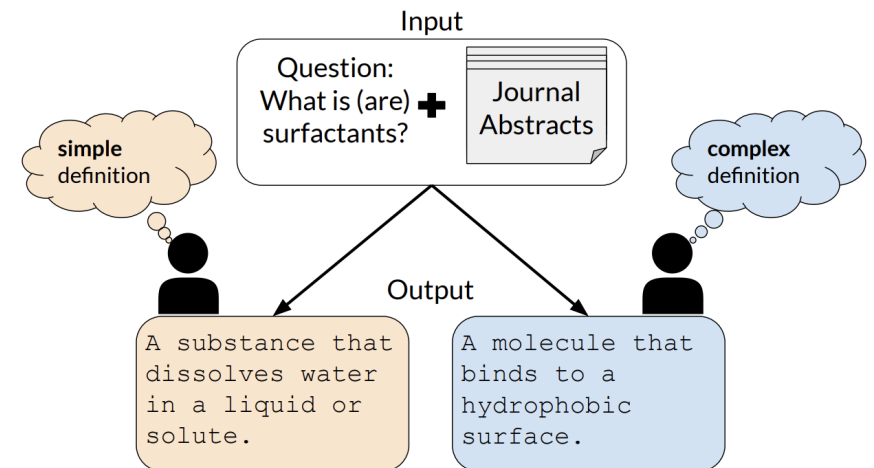
Scientific Definition Generation

- 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

Related Work

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].

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

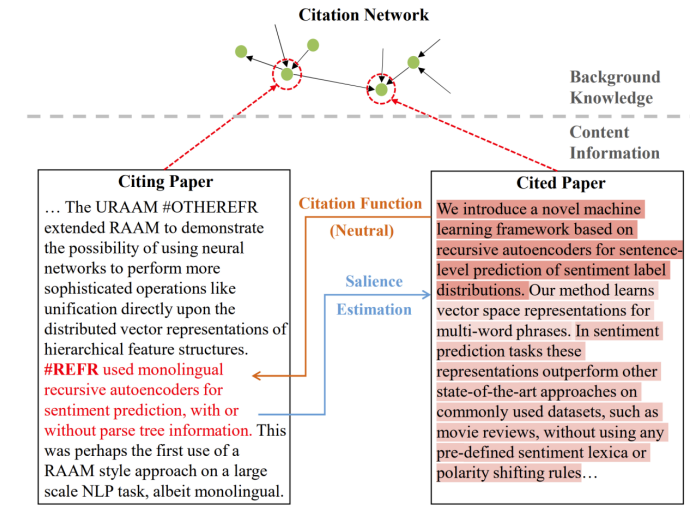
- Related work generation usually involves citation networks



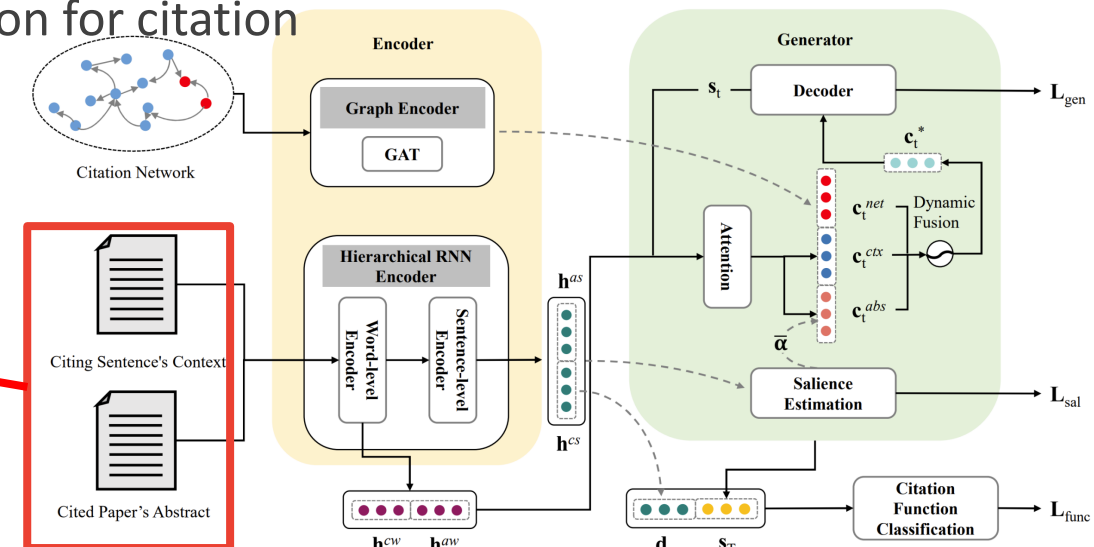
Knowledge-guided Citation Generation



- 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)



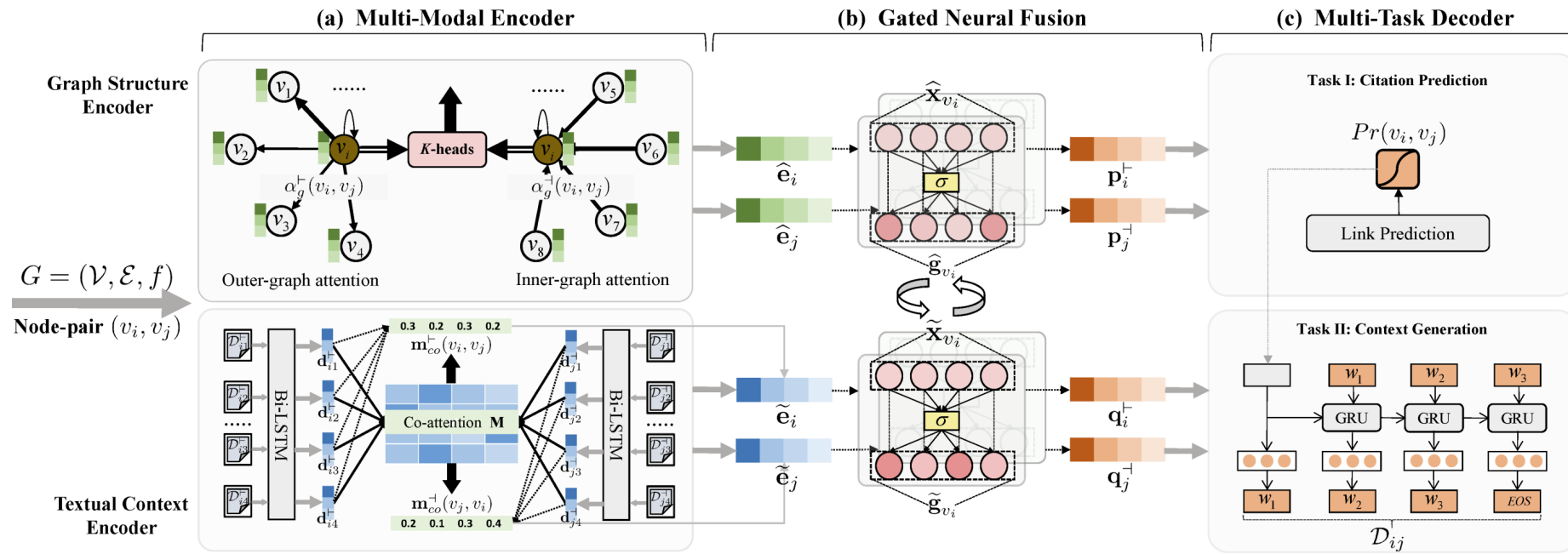
What if citation relations are not given?



Context-aware Citation Prediction

Model

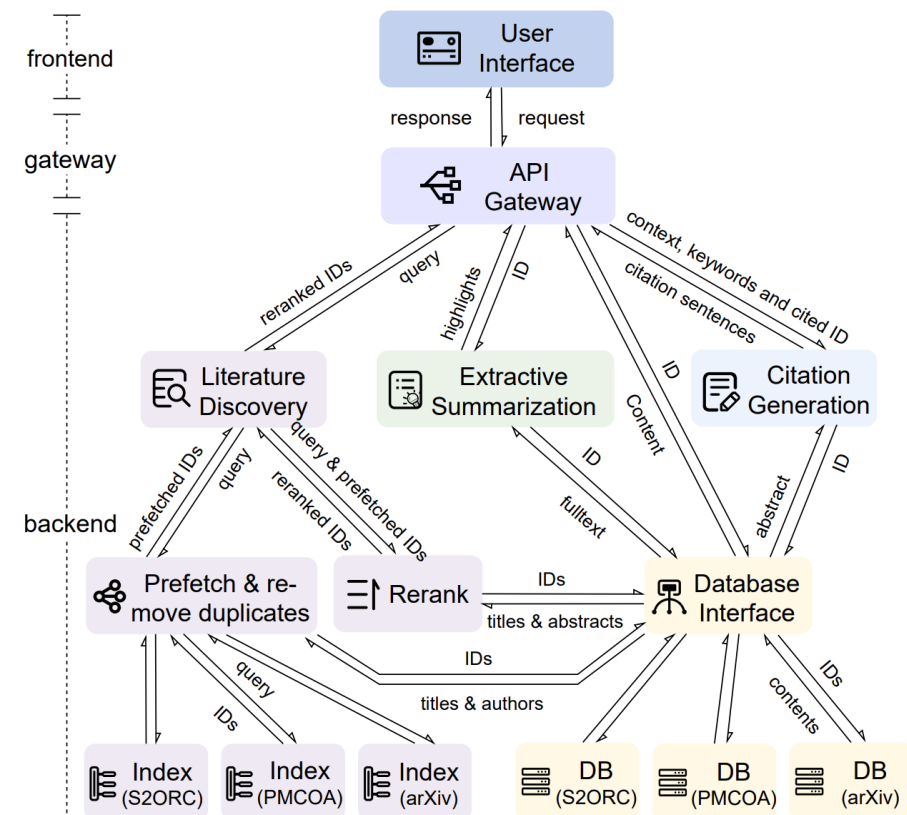
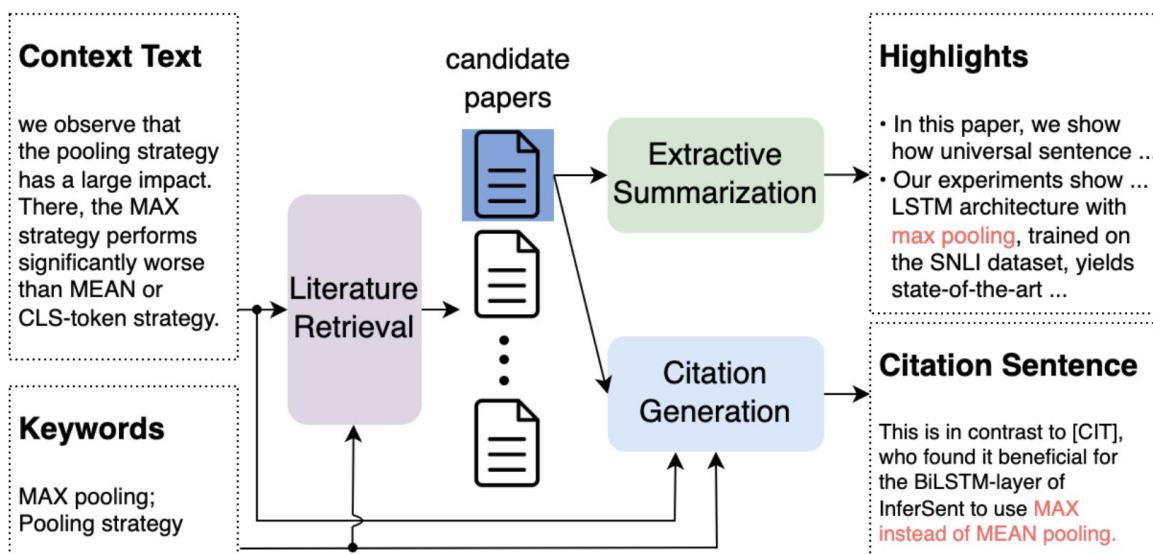
- Graph Structure Encoder with dual-role attention to capture directed graph structure information
- Gated Neural Fusion to integrate multi-modal representations and control features transfer for downstream tasks differently
- Multi-Task Decoder to perform citation link prediction and context generation



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, given the user-provided context and keywords
 - Prefetching papers from database and reranking them based on keywords by SciBERT
 - Iteratively score and select sentences for summarization with MemSum
 - Concatenate keywords, context and target paper's abstract to generate citation sentences by T5



Method	Pros	Cons
BACO (Ge et al., 2021)	<ul style="list-style-type: none"> 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 	<ul style="list-style-type: none"> Rely on provided citation relations of the paper to generate citation sentence
AutoCite(Wang et al., 2021)	<ul style="list-style-type: none"> Propose a joint framework to predict citation links and generate citation sentences simultaneously Design a graph encoder to distinguish citation directions 	<ul style="list-style-type: none"> Fail to generate different types of citation texts according to the contexts and positions in the target paper
SciLit (Gu et al., 2023)	<ul style="list-style-type: none"> Propose a pipeline to recommends relevant papers, extract highlights, and suggest a reference sentence 	<ul style="list-style-type: none"> The citation generation ignores the graph structure of citation networks



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
lukaszkaizer@google.com

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

Related Work

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].

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.

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

Related Work

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].

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.

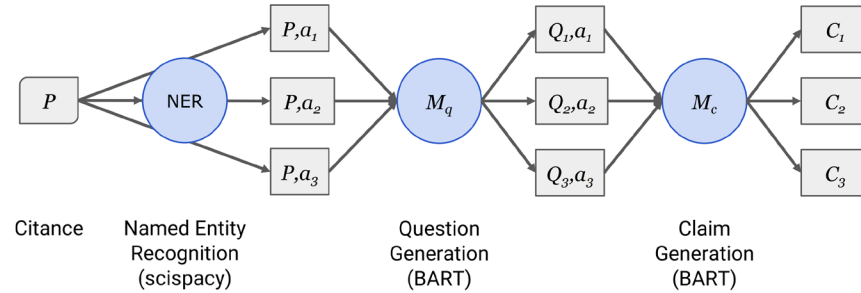
Description

Claim

Conclusion

Supported Claims Generation

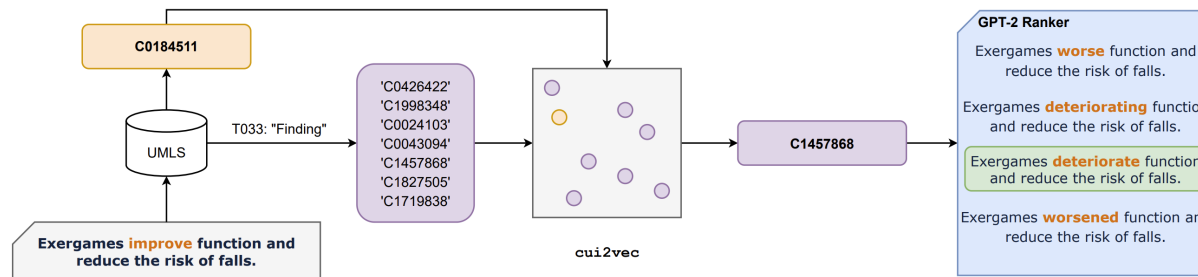
- ClaimBART: Generating claim directly based on Citance
- ClaimEntity



(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.

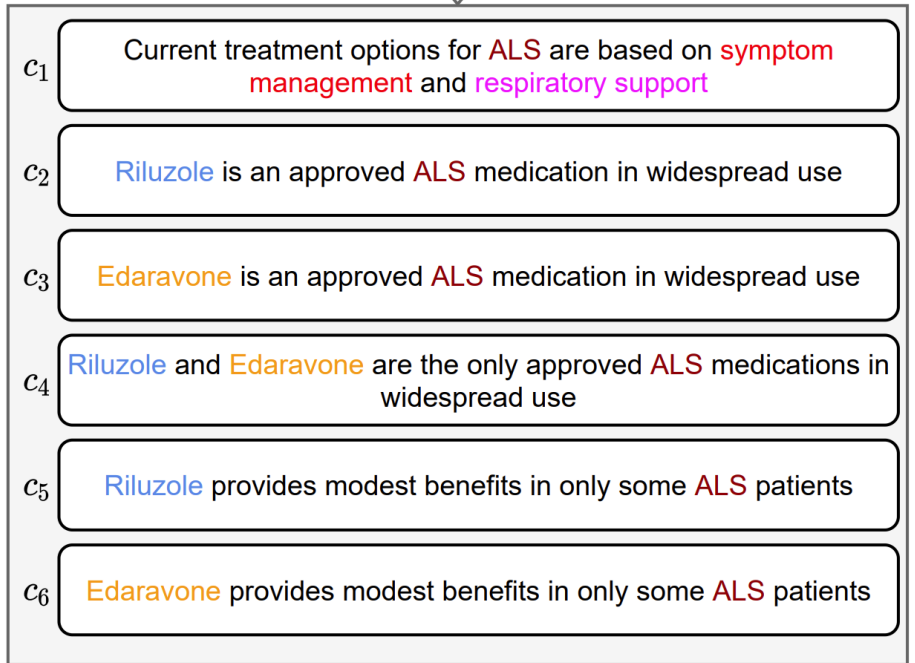
Refuted Claims Generation

- Knowledge-based Informed Negation

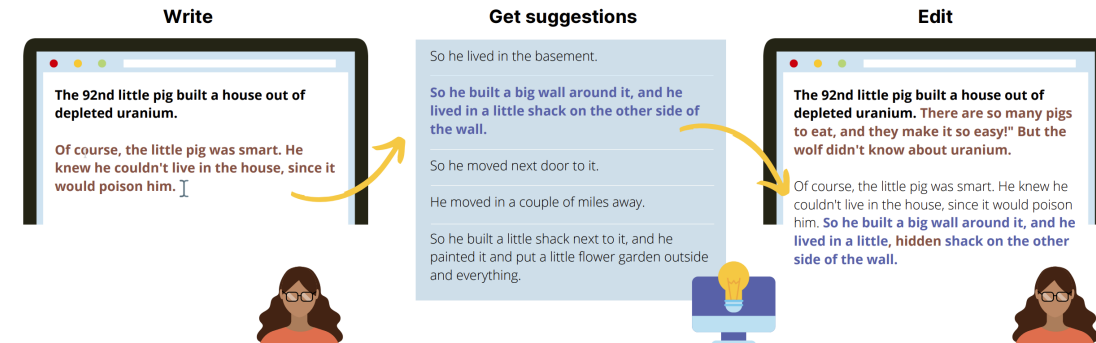
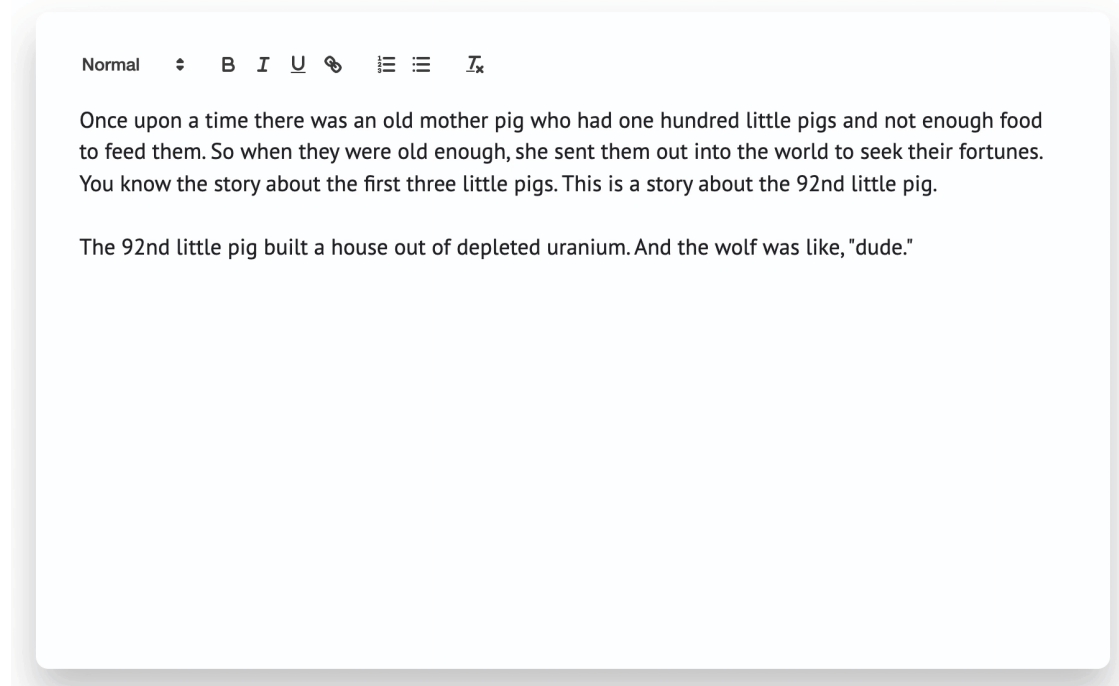


Points for Improvement:

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

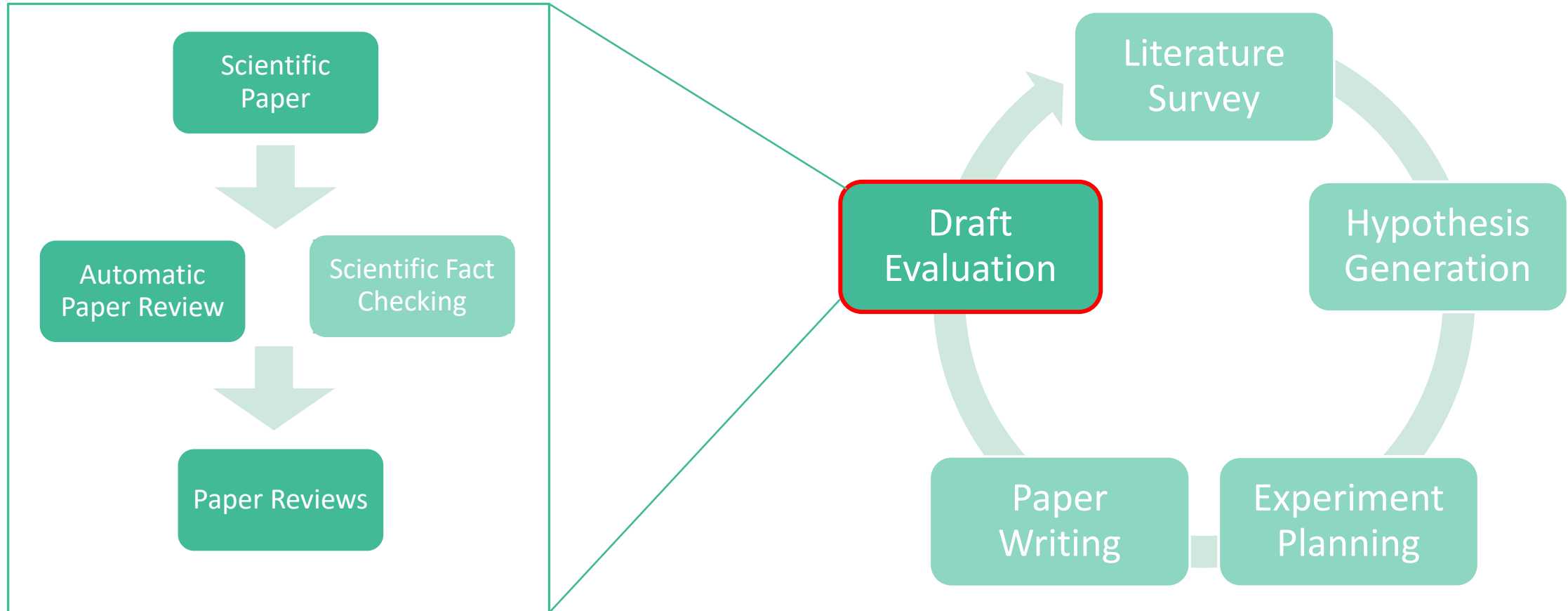


- Compared to either humans or AI, collaborative Human-AI Writing has:
 - Fewer grammatical errors
 - Higher vocabulary diversity
- LLMs can increase writers' productivity by providing writing suggestions
- Keep fraction of text written by writers to text written by ChatGPT relatively high can increase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



- 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

- Future directions
 - To reduce error propagation, future developments could focus on creating an end-to-end system, moving away from the current pipeline-based approach
 - 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

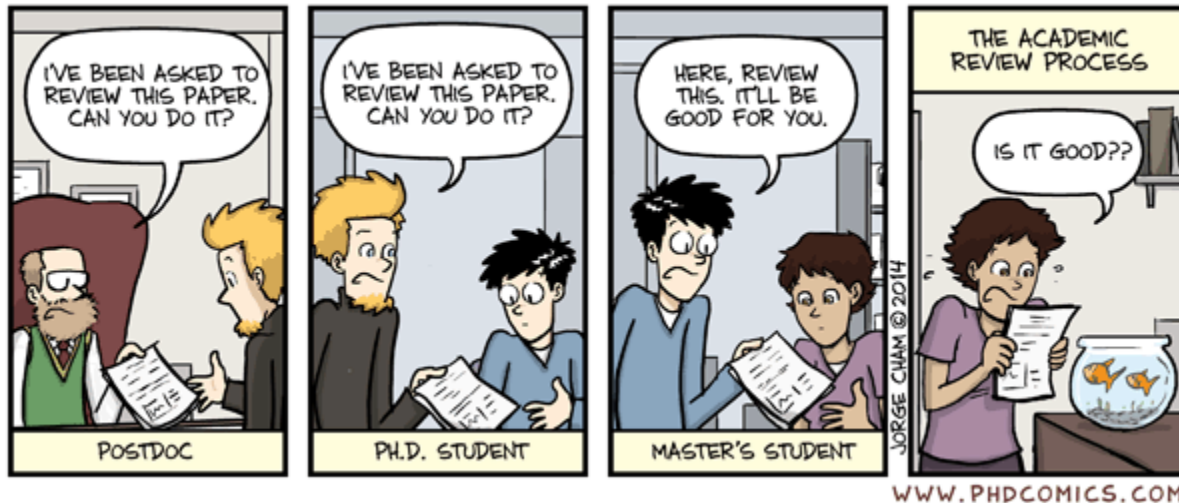


- To catch up with the increasing submissions, reviewers are overburdened leading to contentious review comments
 - “There is significant evidence that the process of reviewing papers in machine learning is creaking under several years of exponentiating growth.” -- ICML President John Langford
 - 12.9% of reviews were flagged by authors for having issues, and 1.69% of reviews were noted as rude and unprofessional -- ACL 2023 Peer Review Report
 - Authors have even created various social groups at social media to release their frustrations and anger, such as the “*Reviewer #2 must be stopped*” group at Facebook

<i>Type of issue</i>	<i>Number of reviews</i>	<i>% of reviews</i>
A: The review is not specific enough	272	2.16
B: Review heuristics such as “not novel”, “not surprising”, “too simple”, “not SOTA”	678	5.38
C: The scores do not match the review text	448	3.55
D: The review is rude/unprofessional	213	1.69
E: The review does not evince expertise	542	4.3
F: The review does not match the paper type	98	0.78
G: The review does not match the type of contribution	152	1.21
H: The review is missing or too short	205	1.63
I: The review was late	12	0.1
J: Other	162	1.29

Challenges in Peer Review

- The following are two different reviews for the same paper rejected by ACL2019 and accepted by EMNLP2019 without any change in content:
 - **ACL 2019:** "Idea is too simple and tricky."
 - **EMNLP 2019:** "The main strengths of the paper lie in the interesting, relatively under-researched problem it covers, the novel and valid method and the experimental results."



- Human reviews are subjective!
- We need an automatic review assistant to help human reviewers generate knowledgeable and explainable review scores and comments, along with detailed evidence

ADDRESSING REVIEWER COMMENTS BAD REVIEWS ON YOUR PAPER? FOLLOW THESE GUIDELINES AND YOU MAY YET GET IT PAST THE EDITOR:

<p>Reviewer comment: "The method/ device/ paradigm the authors propose is clearly wrong."</p> <p>How NOT to respond: ✗ "Yes, we know. We thought we could still get a paper out of it. Sorry."</p> <p>Correct response: ✓ "The reviewer raises an interesting concern. However, as the focus of this work is exploratory and not performance-based, validation was not found to be of critical importance to the contribution of the paper."</p>	<p>Reviewer comment: "The authors fail to reference the work of Smith et al., who solved the same problem 20 years ago."</p> <p>How NOT to respond: ✗ "Huh. We didn't think anybody had read that. Actually, their solution is better than ours."</p> <p>Correct response: ✓ "The reviewer raises an interesting concern. However, our work is based on completely different first principles (we use different variable names), and has a much more attractive graphical user interface."</p>	<p>Reviewer comment: "This paper is poorly written and scientifically unsound. I do not recommend it for publication."</p> <p>How NOT to respond: ✗ "You #&@*% reviewer! I know who you are! I'm gonna get you when it's my turn to review!"</p> <p>Correct response: ✓ "The reviewer raises an interesting concern. However, we feel the reviewer did not fully comprehend the scope of the work, and misjudged the results based on incorrect assumptions."</p>
---	---	--

www.phdcomics.com

JORGE CHAM © 2005

Challenges in Automatic Paper Review

- Generation Results for Sparse Weight Activation Training (Raihan et al., 2020) in Kid-Review's Appendix (Yuan et al., 2020)
- This paper proposes a method to sparsify the weights and activations of a neural network by sparsifying the gradients in the backward pass . The idea is inspired by meProp , which sparsifies gradients with relatively small magnitude . The authors show that the proposed method can reduce the memory footprint by 23 % to 37 % for activations and 50 % to 80 % for weights . The paper is well written and easy to follow. The experimental results on CIFAR-10 and ImageNet are promising . However , I have the following concerns : 1 . It is not clear to me why the authors chose to use SAW instead of meProp as the baseline method . It seems to me that meProp is a better baseline than SAW . 2 .The authors claim that the memory benefit of SAW is not present for meProp since there is no storage benefit since they are temporary values generated during back-propagation . But I don't see any evidence to support this claim . 3 .In the experiments , the authors only compare with SAW and meProp . It would be more convincing if the authors can also compare with other sparsification methods , e.g. , [1] and [2] ...

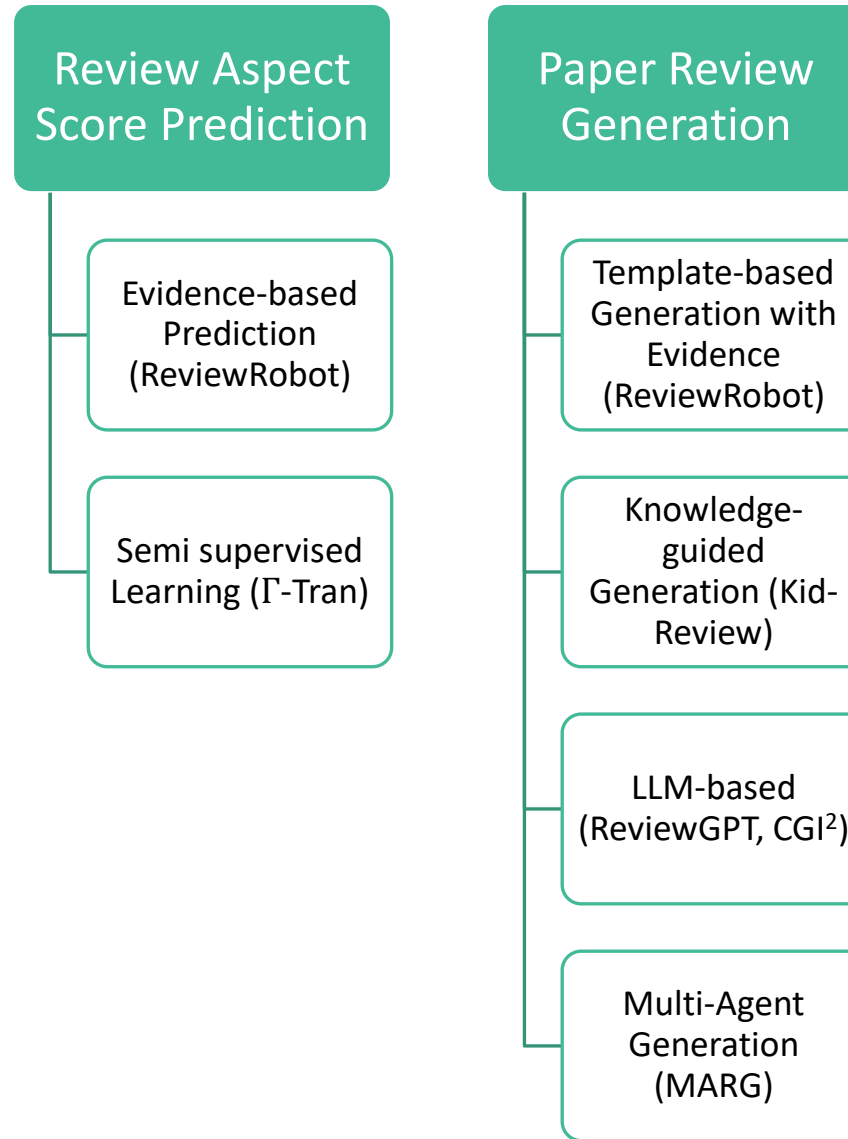
SAW performs better than meProp in Figure 5

Missing references

meProp is compared against Sparse Weight Activation Training (SWAT)

Only the summarization part works
Other parts suffer from hallucination

- Peer review dataset is limited
 - PeerRead Dataset (Kang et al., 2018)
 - NeurIPS 2013-2017, ICLR 2017, ACL 2017, and CoNLL 2016
 - Annotate 1.3k reviews for ICLR 2017 with aspect scores based on reviews
 - ReviewRobot Dataset (Wang et al., 2020)
 - NeurIPS 2013-2018, ICLR 2017-2020, ACL 2017
 - Background KG from 174,165 papers from 1965 to 2019
 - ASAP-Review Dataset (Yuan et al., 2022)
 - NeurIPS 2016-2019, ICLR 2017-2020, ACL 2017
 - Annotate aspect scores with BERT



Explainable Paper Review Generation

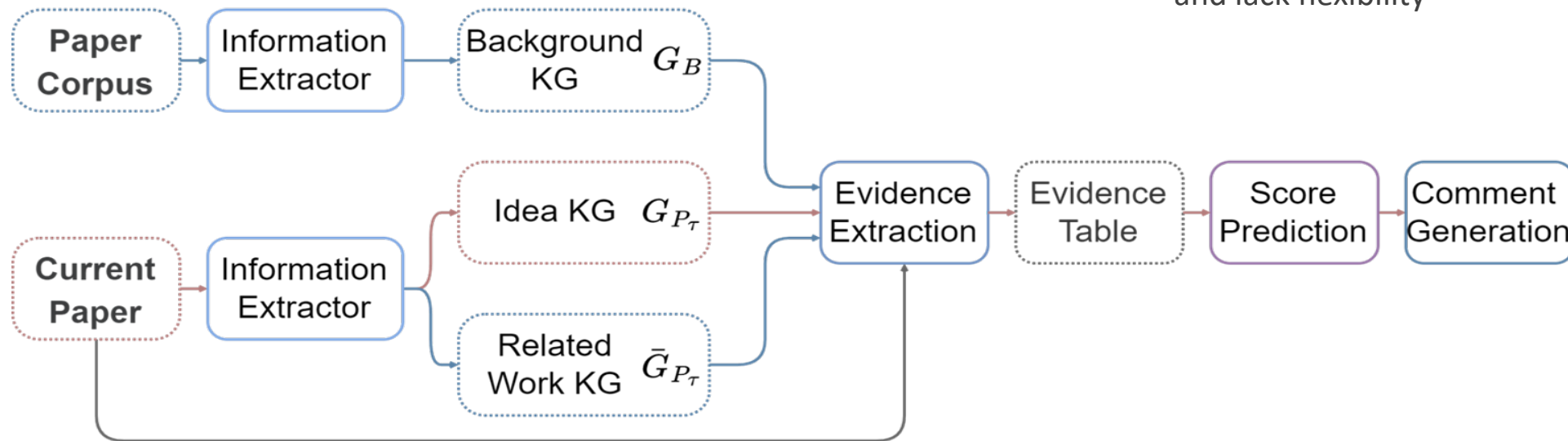
- Propose a knowledge-driven review score prediction and comment generation framework, ReviewRobot, based on fine-grained knowledge element comparison among papers
- The review comments are highly explainable and constructive, supported by detailed evidence

Pros:

- Both score prediction and review generation are grounded on evidence

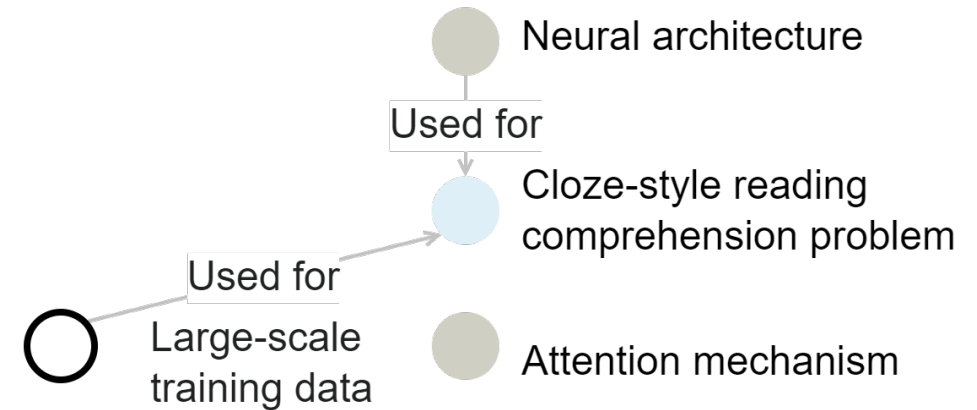
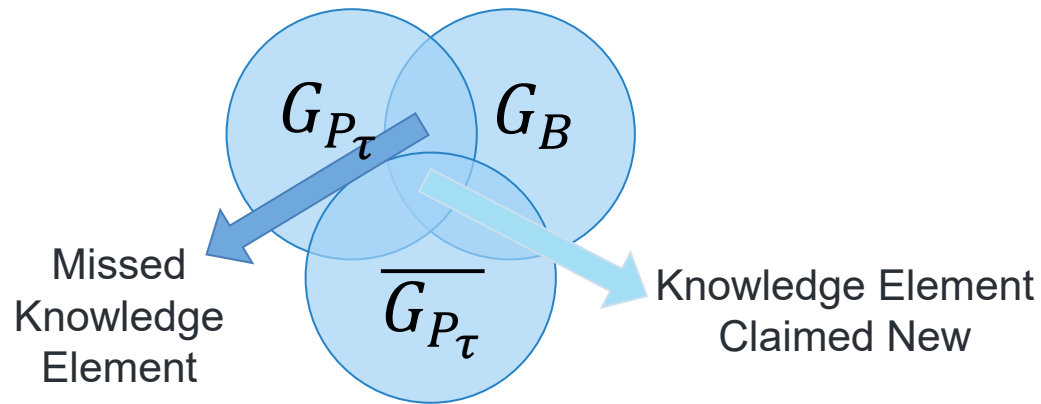
Points for Improvement:

- The pipeline system heavily rely on the quality of IE systems, which might propagate errors
- The review comment templates are human-curated and lack flexibility



Paper Review Evidence: Meaningful Comparison

- The number of papers about relevant old knowledge elements which are missed in the related work section: $G_{P_\tau} \cap G_B - \overline{G_{P_\tau}}$
- The number of papers about relevant knowledge elements which are claimed new in the related work section: $G_{P_\tau} \cap G_B \cap \overline{G_{P_\tau}}$
- The description sentences about comparison with related work

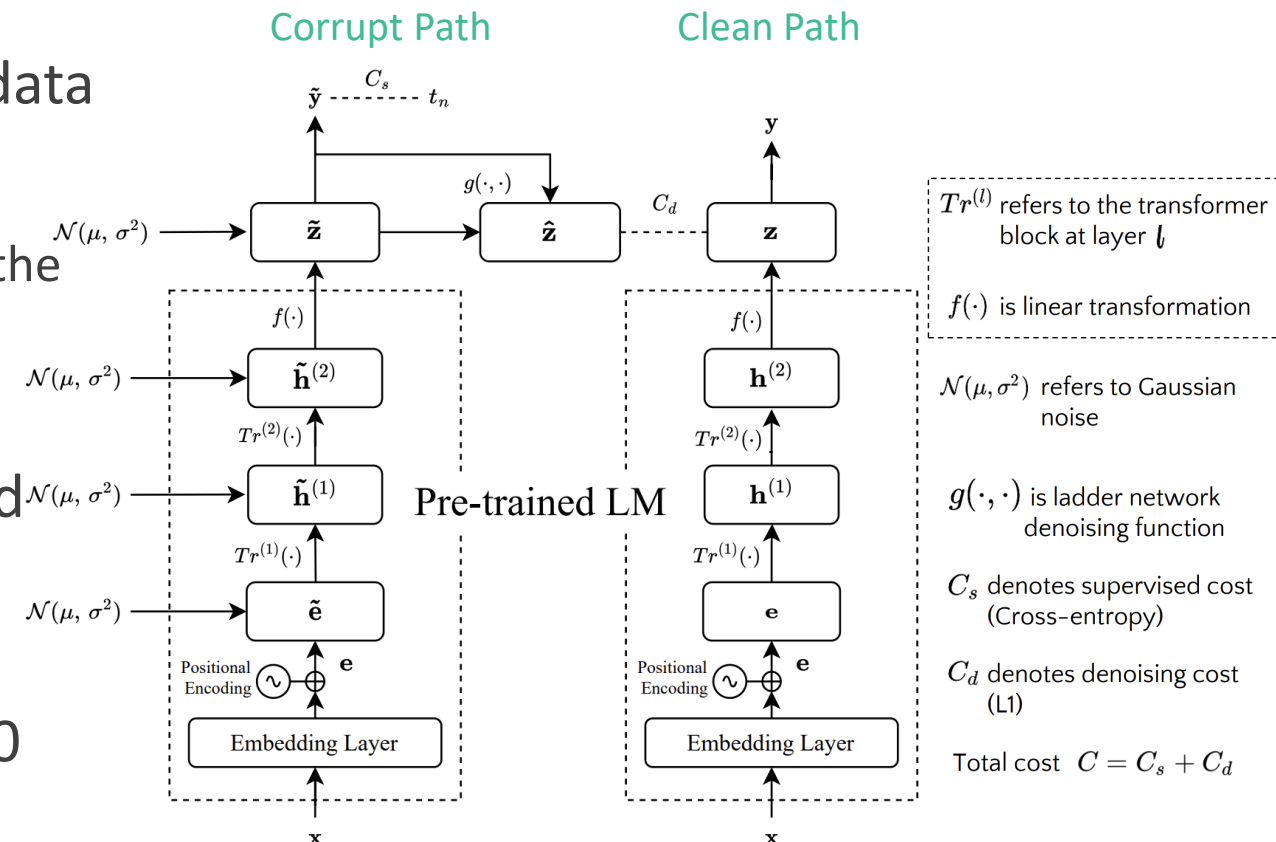


- [ReviewRobot] The following related papers are missing:
 - About low-dimensional semantic space:
 - Unsupervised Approximate-semantic Vocabulary Learning for Human Action and Video Classification: Qiong Zhao and Horace HS Ip. 2013. Unsupervised Approximate-semantic Vocabulary Learning for Human Action and Video Classification. *Pattern Recognition Letters*, 34(15):1870–1878.
 - About sememes:
 - Chinese Word Sense Disambiguation with PageRank and HowNet: Jinghua Wang, Jianyi Liu, and Ping Zhang. 2008. Chinese Word Sense Disambiguation with PageRank and HowNet. In Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing.
 - A Maximum Entropy Approach to HowNet-based Chinese Word Sense Disambiguation: Ping Wai Wong and Yongsheng Yang. 2002. A Maximum Entropy Approach to HowNet-based Chinese Word Sense Disambiguation. In COLING-02: SEMANET: Building and Using Semantic Networks.
 - About word similarity and word analogy:
 - Open IE as an Intermediate Structure for Semantic Tasks: Gabriel Stanovsky, Ido Dagan, et al. 2015. Open IE as an Intermediate Structure for Semantic Tasks. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 303–308.
- [HUMAN] The paper would be stronger with the inclusion of more baselines based on related work.

- Propose Γ -Trans incorporating a pretrained transformer into semi-supervised learning, by leveraging contextual features from unlabeled data
- Unlabeled dataset
 - ScisummNet Corpus with 1,000 papers in the ACL Anthology
- Help improve the performance of all pretrained LMs including SciBERT and Longformer
- Γ -Trans with 100 unlabeled data outperforms other methods with 700 unlabeled data

Points for Improvement:

- The system can further benefit from using the related information between review aspects



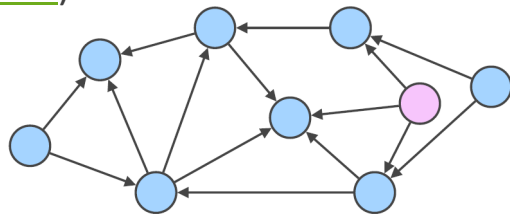
- An end-to-end knowledge-guided review generation framework based on the citation graph and the concept graph
 - Encode concept graph with Graph Attention Network
 - Use citation embeddings as additional input

Pros:

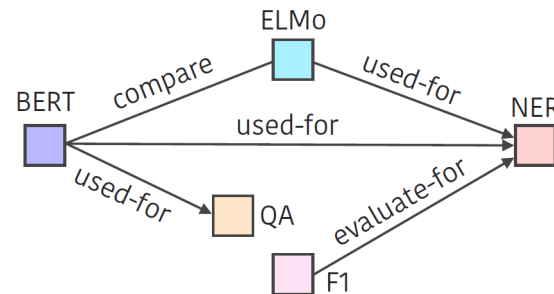
- The paper generates review based on both citation and knowledge graphs since human beings rely on both background knowledge and working memory to review a paper

Points for Improvement:

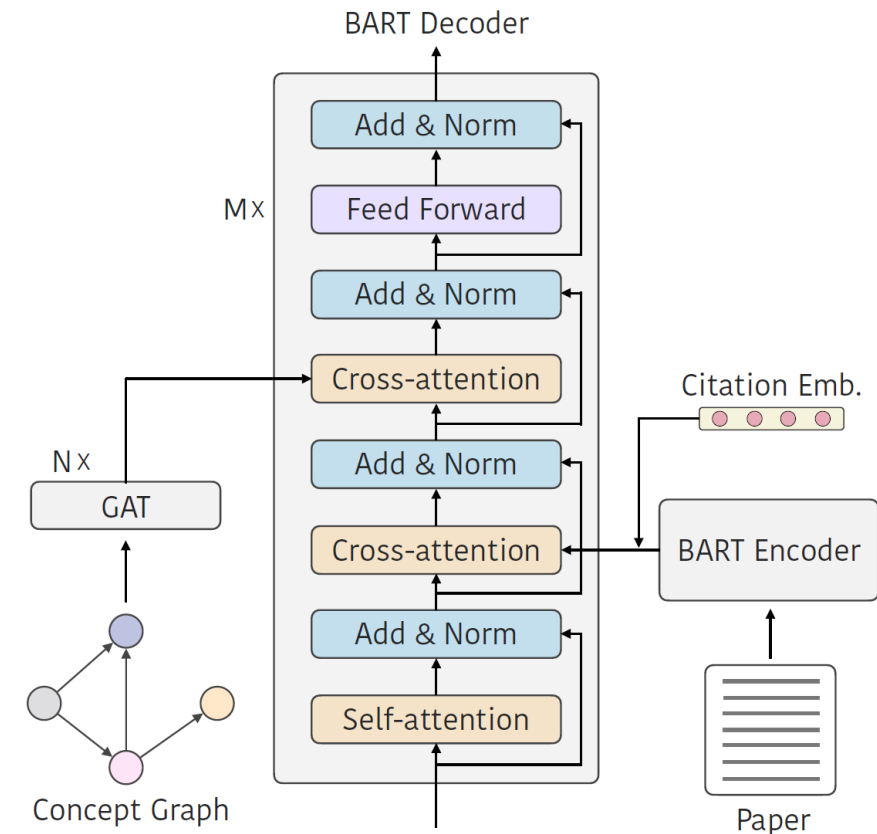
- The proposed model still requires an extractor to extract important sentences from the source article and only use the extracted sentences
- The generated review contains many unfaithful contents (as shown in [Slide 156](#))



(a) World knowledge – citation graph



(b) Temporary knowledge – concept graph

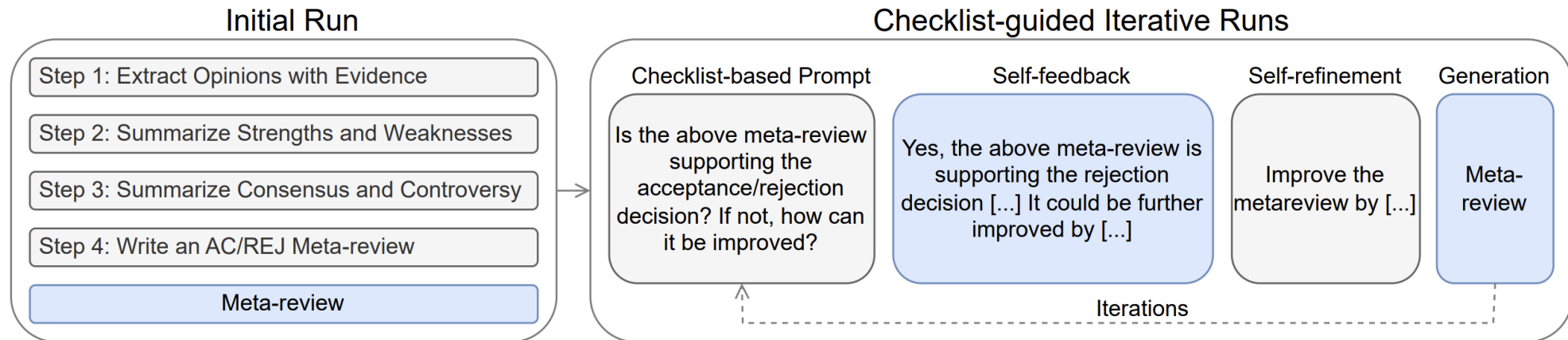


- Test ChatGPT with papers using prompt engineering and chain-of-thought
- Identifying errors
 - ChatGPT completely fails on sorting, pairwise comparison, game theory, and optimization
- Verifying checklists
 - When verifying 16 closed-ended checklist questions in 15 NeurIPS 2022 papers, ChatGPT achieves 86.6% accuracy
- Choosing the “better” paper
 - When comparing 10 pairs of abstracts which deliberately designed in such a way that one abstract was clearly superior to the other, ChatGPT struggled to discern these relatively straightforward distinctions accurately, committing errors in its evaluations for 6 out of the 10 pairs

LLMs such as ChatGPT, without tailored framework designs, remain far from expected solutions

Meta-review Generation

- Input: the title, abstract, and a set of reviews of one research paper
- Goal: Generate a meta-review summarizing the opinions in the independent reviews and make a recommendation decisions
- Checklist-guided Iterative Introspection
 - Break down the task into several stages
 - Iteratively refine the summary under the guidance of questions from a checklist



Points for Improvement:

- The proposed model fails to utilize any knowledge graphs, which might lead to hallucination
- The system only uses titles and abstracts not the whole paper

Human

The paper unanimously receives positive rates thanks to strong motivations and interesting results. As the reviews show satisfaction on the authors' feedback, the final draft needs to respect it accordingly, for example, about the limitations of this research.

Vanilla

This paper could have benefited from a clearer and more concise presentation of the main findings. It is sometimes difficult to follow the paper's chain of thought, and the authors could have presented their results in a more accessible way. The experiments are limited to a small set of environments, and it would be interesting to see how the results generalize to other tasks. Additionally, it would be useful to have a deeper analysis of why iGPT does not perform well, as this could help inform future research on how to improve image-based pretraining.

Initial run

Based on the reviews, I recommend accepting the paper. The paper addresses an important and under-explored research question and presents a detailed and thought-provoking analysis of the pretraining-finetuning paradigm for Transformer models in the context of reinforcement learning. The paper also identifies relevant techniques to perform analysis of representation similarity and training dynamics. While there are some weaknesses pointed out, such as the need for further exploration of some hypotheses and occasional difficulty in following the writing and presentation, the strengths of the paper outweigh these concerns. Therefore, I recommend accepting the paper with minor revisions to address the weaknesses pointed out by the reviewers.

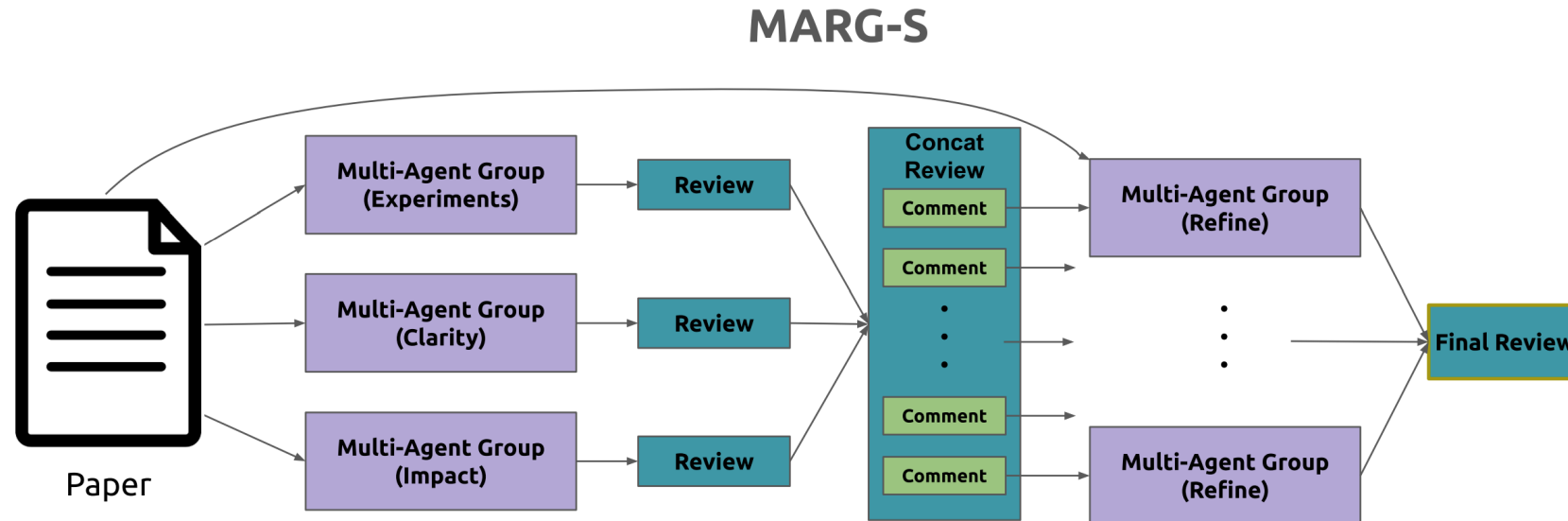
CGI²

After carefully considering the reviews provided, there is consensus that the paper presents an important and thought-provoking analysis of representation dynamics between pretraining and finetuning in the context of reinforcement learning. Both reviewers agree that the paper addresses an important question and presents insightful experiments. One reviewer highlights the experiments as a strong point in the paper, while the other reviewer raises a concern about the paper's tendency to present hypotheses without exploring them further. Overall, the strengths of the paper outweigh its weaknesses, and it has the potential to make a significant contribution to the field of reinforcement learning. Therefore, I recommend accepting the paper with a minor revision to address the concern raised by the reviewer. Specifically, the paper should explore the hypotheses presented in more depth to provide a more comprehensive analysis. If the authors can address this concern, the paper can make a valuable contribution to the field of reinforcement learning. In summary, I recommend accepting the paper with a minor revision.

Hallucination reduced

Multi-Agent Review Generation

- A network of LLM agents that communicate to share information across different parts of a paper and engage in internal discussion
 - Consume the full text of papers beyond the input length limitations of the base LLM
 - Include aspect-specific “expert” GPT agents to separately assist with generating comments on experiments, clarity, and impact
 - The method perform significantly better than when having a single agent attempt to generate all types of feedback at once



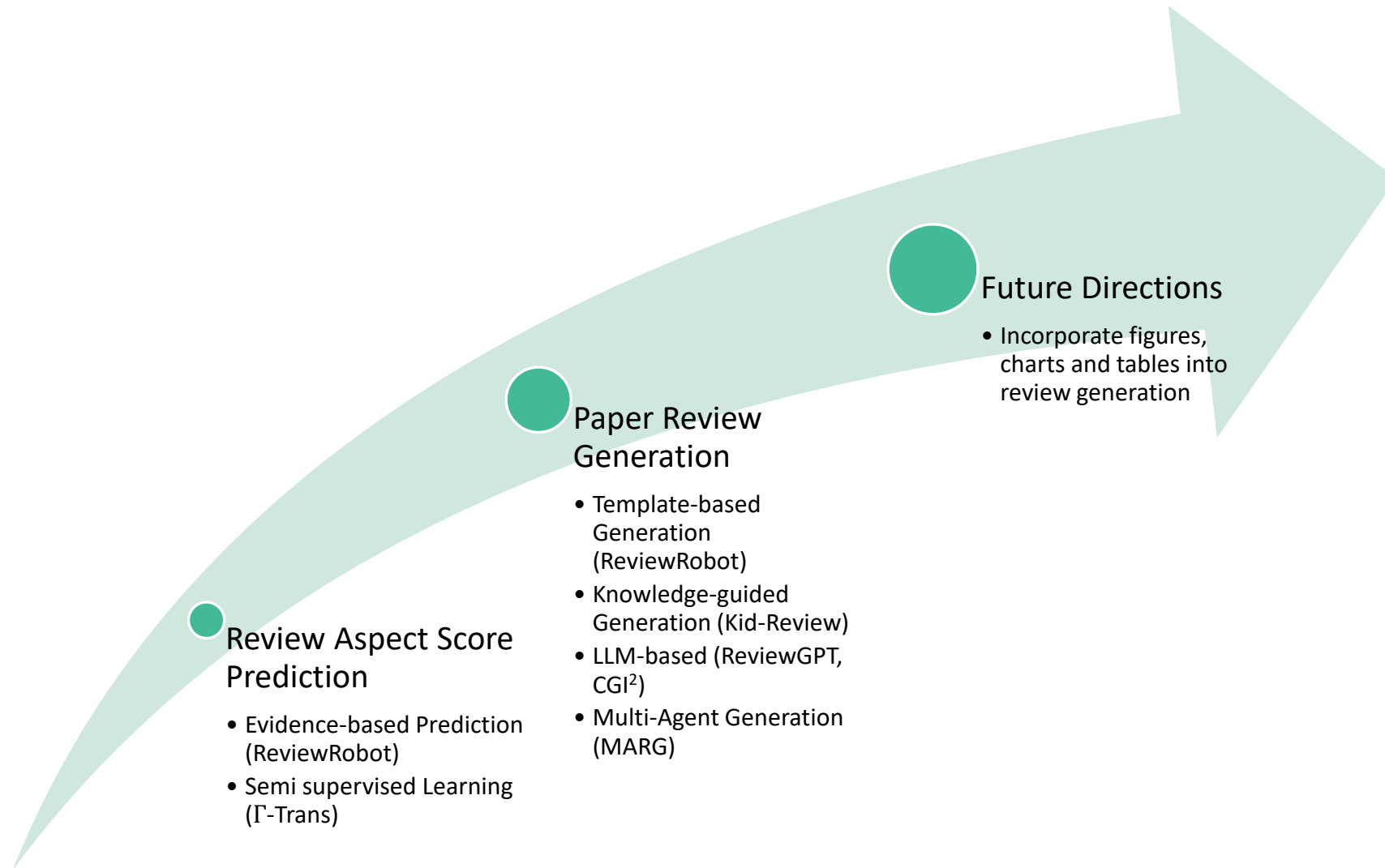
Points for Improvement:

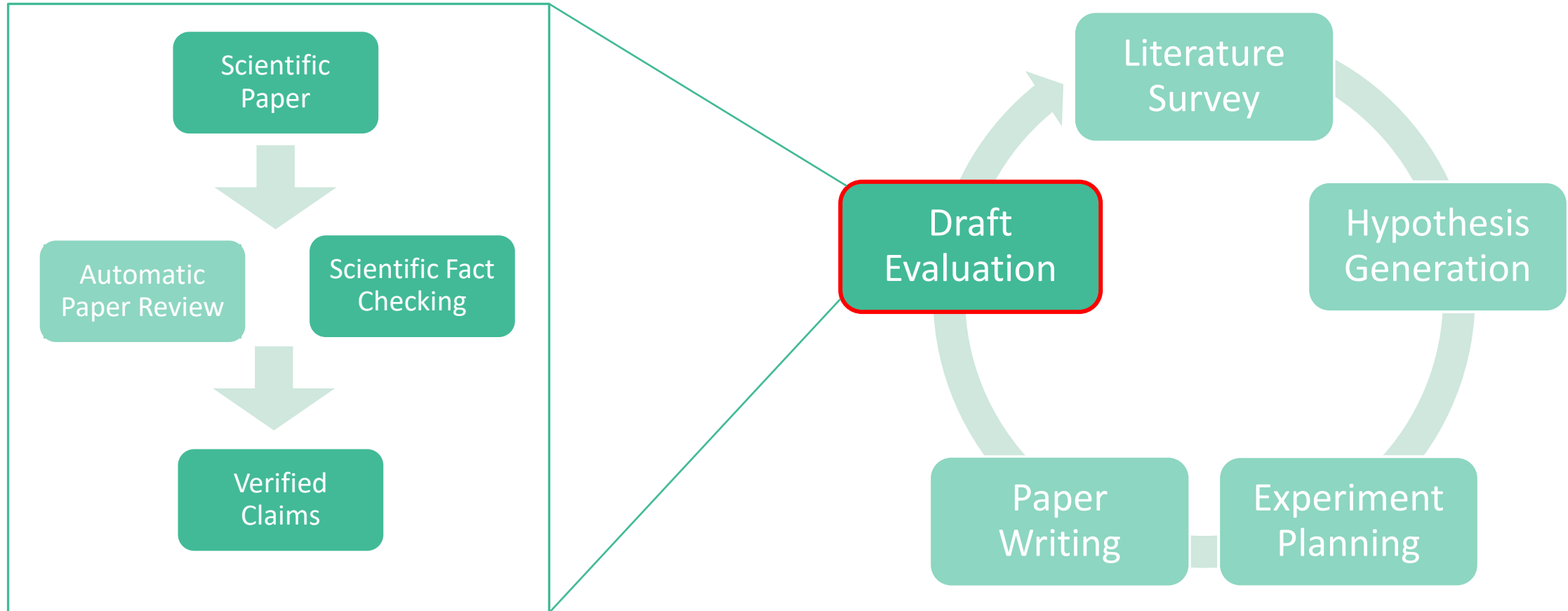
- The proposed model fails to utilize any knowledge graphs or background literature
- The proposed model relies heavily on GPT4, which is very expensive



Real-reviewer comment	Generated comment
The experimental methodology used in the paper is not well detailed , making it difficult to reproduce the reported results.	More details about the experiments conducted would be beneficial. This should include information about the datasets used, the training process, and the evaluation process . To ensure the reproducibility of the results, consider providing the code used to implement the model, the specific parameters used, and any other necessary information . This will allow other researchers to replicate your work and further validate your findings. [high relatedness, more specific]
The paper does not include enough base-lines for Fair Federated Learning to compare against. Even if some methods do not satisfy privacy considerations, they should still be included for the reader to understand how the proposed method compares against such methods, especially given that the results are not promising. Some base-lines to consider include Cui et al or Tran et al.	The authors should consider including a comparison of their proposed method with existing methods in the experimental results section. This would help to highlight the advantages and improvements of their proposed method. [high relatedness, less specific]

Some reviews are vague





■ Threat to the Individual

- Early in the pandemic, there was speculation about the internal use of disinfectants for treating COVID-19 patients.
 - On June 5, 2020, the Centers for Disease Control (CDC) reported a steep increase in calls to poison centers regarding exposure to household disinfectants.
 - A CDC survey of 502 adults in the United States found that 39% of responders engaged in dangerous practices including washing food products with bleach, applying household cleaners directly to skin, and intentionally inhaling or ingesting disinfectants with the goal of preventing COVID-19 infection.

Coronavirus: Disinfectant firm warns after Trump comments

24 April 2020

Share



Mr Trump made the comments at a government briefing on Thursday

A leading disinfectant producer has issued a strong warning not to use its products on the human body after Donald Trump suggested they could potentially be used to treat coronavirus.

Fact check: Trump dangerously suggests sunlight and ingesting disinfectants could help cure coronavirus

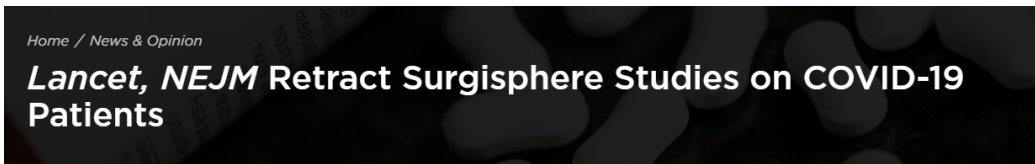


By Daniel Dale, Nathan McDermott, Marshall Cohen, Maegan Vazquez, Em Steck and Sam Fossom, CNN
6 minute read · Updated 2:18 PM EDT, Fri April 24, 2020



■ Threat to the Scientific Research

- The *Lancet*'s paper linked *hydroxychloroquine* use to harmful outcomes, prompting the WHO to halt its use in the Solidarity trial. However, due to questions about the study's methods and data reliability, the study was retracted, and the WHO later resumed the trials.



All authors other than company founder and CEO Sapan Desai were "unable to complete an independent audit of the data," *The Lancet* states.



Catherine Offord

Jun 4, 2020 | 3 min read



PDF VERSION



Share

Register for free to listen to this article



Listen with Speechify

0:00

3:00



Two controversial studies of COVID-19 patients [have been retracted](#) after the authors failed to demonstrate that the data were reliable. The first study to be retracted, published last month (May 22) in *The Lancet*, had found harmful effects associated with the antimalarial drug, hydroxychloroquine, but quickly [drew fire](#) after scientists raised questions about the massive [database](#) supposedly underpinning it, and about that database's owner, Surgisphere Corporation.

ABOVE: © ISTOCK.COM,
LILIBOAS

The Lancet retracts large study on hydroxychloroquine

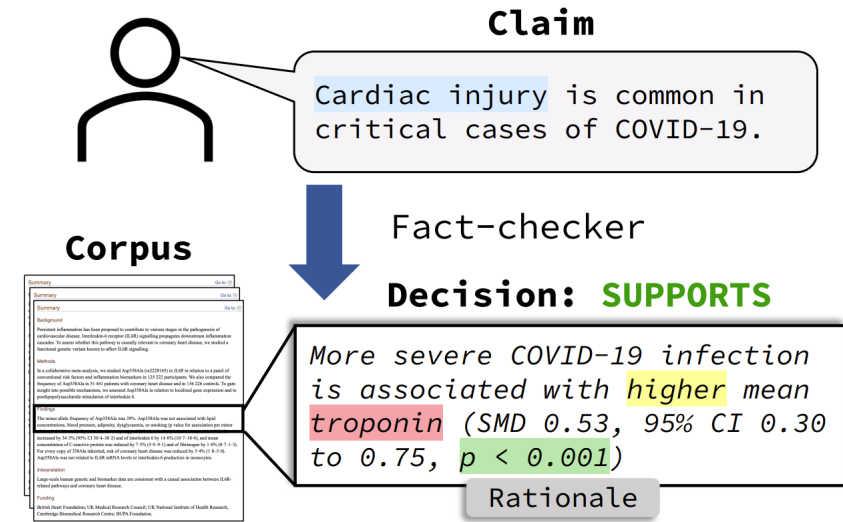
"A first-year statistics major could tell you about major flaws in the design of the analysis," one expert said.

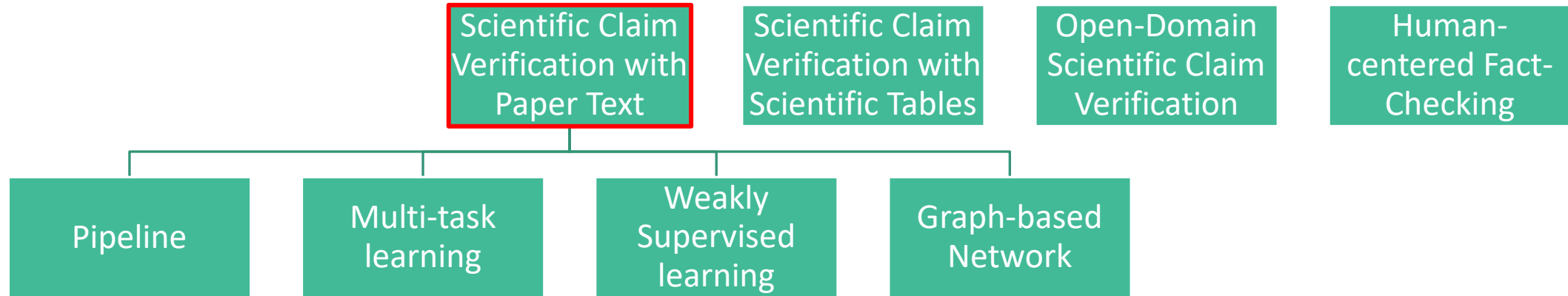


— A large study suggesting hydroxychloroquine does not benefit COVID-19 patients, and may even increase deaths, has been retracted. But that doesn't mean hydroxychloroquine does — or does not — work. Buda Mendes / Getty Images

What is scientific claim?

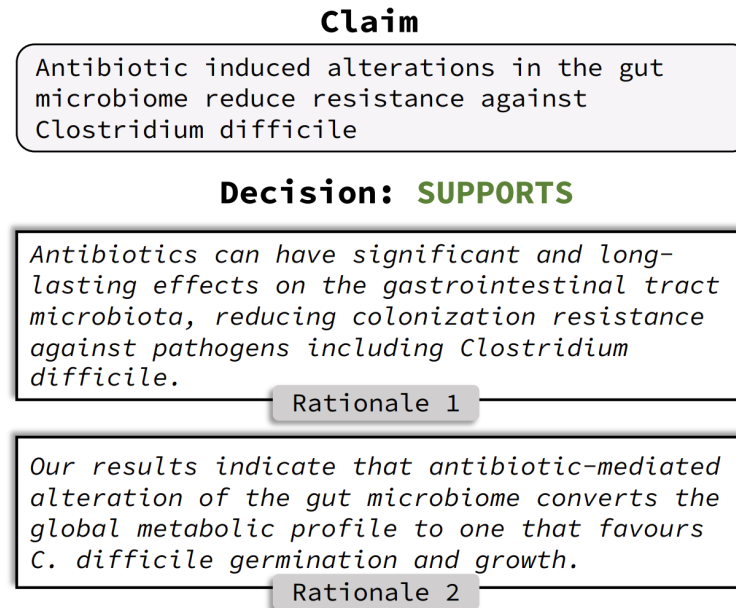
- A scientific claim is an atomic verifiable statement expressing a finding about one aspect of a scientific entity or process, which can be verified from a single source
- Scientific Claim Verification Task
 - Given a scientific claim c and a corpus of abstracts \mathcal{A} , the system must predict a set of evidence abstracts $\mathcal{E}(c)$, where each abstract $a \in \mathcal{E}(c)$ has a predicted label (SUPPORTS, NOINFO, REFUTES) and a collection of rationale sentences





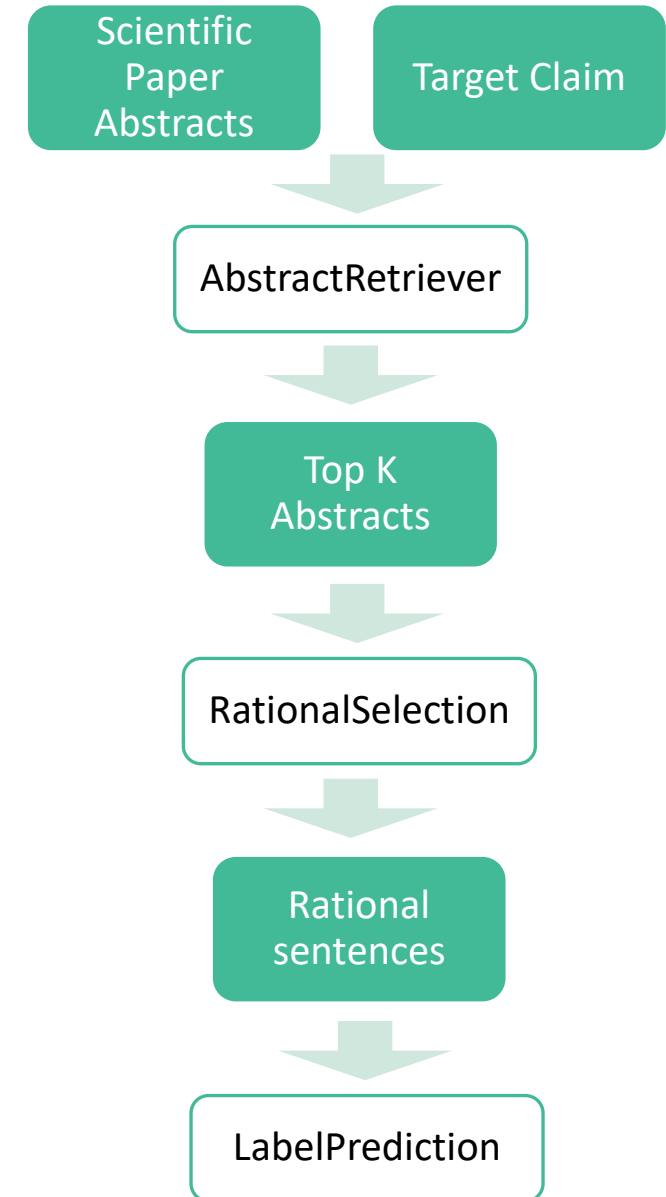
Scientific Claim Verification with Pipeline

- The architecture follows BERT-to-BERT pipeline
 - AbstractRetriever retrieves k abstracts with highest TF-IDF similarity to the claim
 - RationalSelection identifies rationale sentences for each abstract
 - LabelPrediction makes the final label prediction



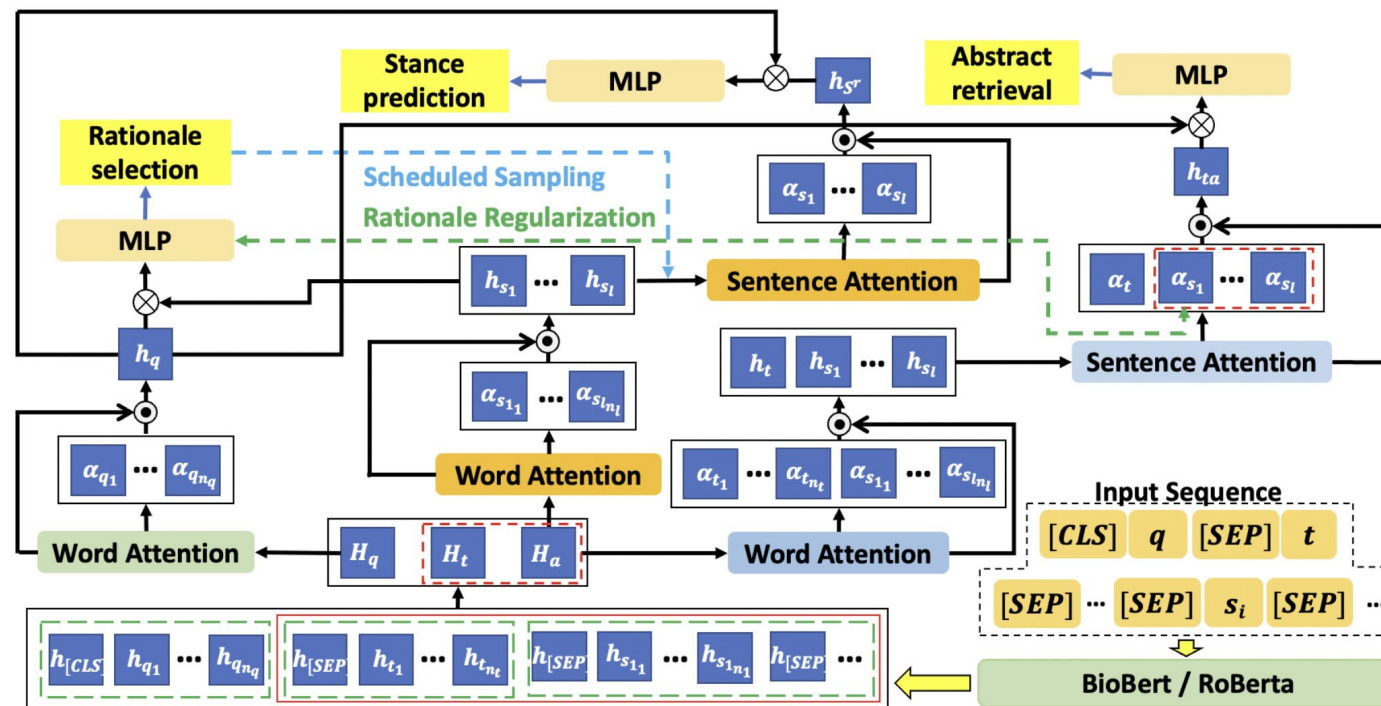
Points for Improvement:

- Incorporate background knowledge from external knowledge bases
- Error might propagate through each steps



Multi-Task Learning for Claim Verification

- An end-to-end multi-task framework, ARSJOINT, to jointly learn the three tasks based on QA Framework
- A rationale regularization based the divergence between the sentence attention of abstract retrieval and the outputs of rationale selection to improve interpretability



Pros:

- Reduce error propagation by transforming the pipeline system to an end-to-end multi-task framework
- Share information between different tasks to improve the prediction quality

Points for Improvement:

- Require additional supervision with rationale sentences, limiting its training flexibility

■ Training Procedure

- Stage 1: Train on a combination of labeled out-of-domain data (FEVER) and weakly labeled in domain data (Paper title as claim, Abstract as unannotated rationales, label as supported)
- Stage 2: Continue training on data from each target dataset

■ Domain adaptation

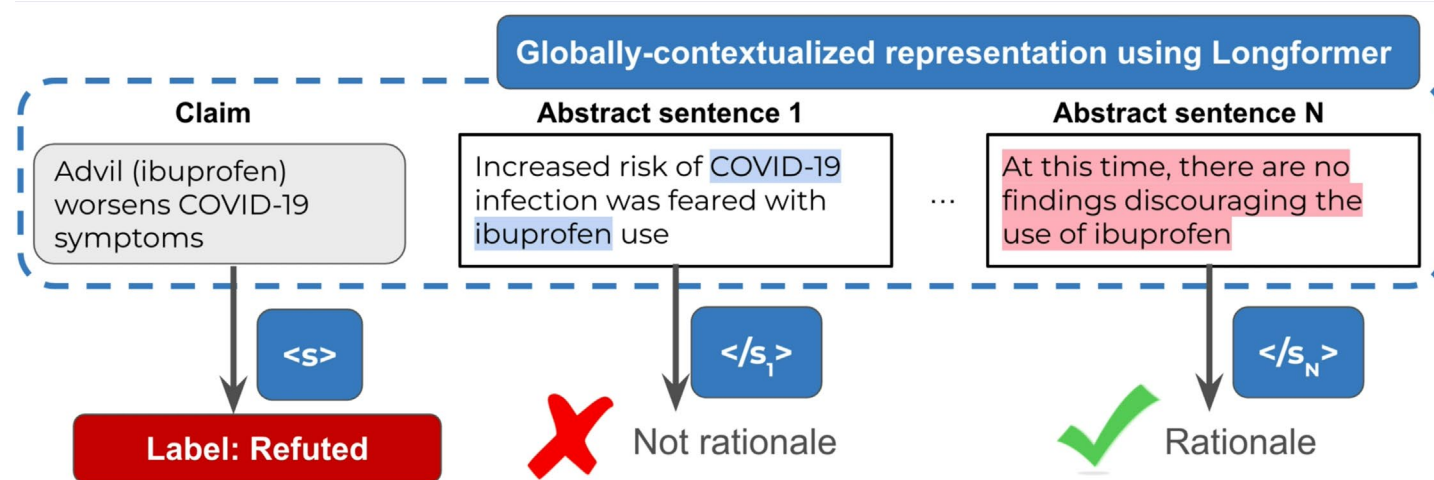
- Zero-shot (without Stage 2)
- Few-shot (45 samples for stage 2)
- Full (All samples for stage 2)

Pros:

- Take in all abstract sentences to leverage the shared information between them
- Can train on instances with no rationale sentence which is suitable for domain adaptation

Points for Improvement:

- Model still depends on the textual similarity without considering any structured information



$$L = L_{\text{label}} + \lambda_{\text{rationale}} L_{\text{rationale}}$$

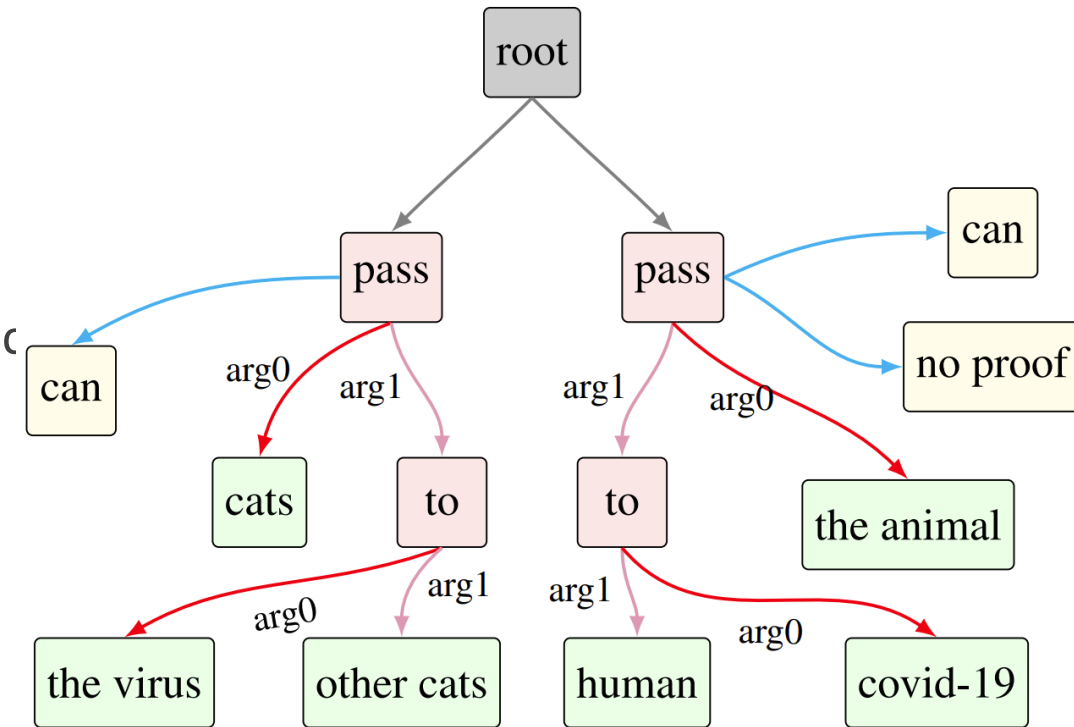
Qualitative Causal Tree

Entities

- *Factors* are variables that are tested or asserted within a claim
- *Associations* are explicit phrases associating one or more factors
- *Modifier* aggregates the original entity types magnitude, evidence, and qualifiers
- *Root* is designed to aggregate tree representations into a single vector representing the whole scientific sentence

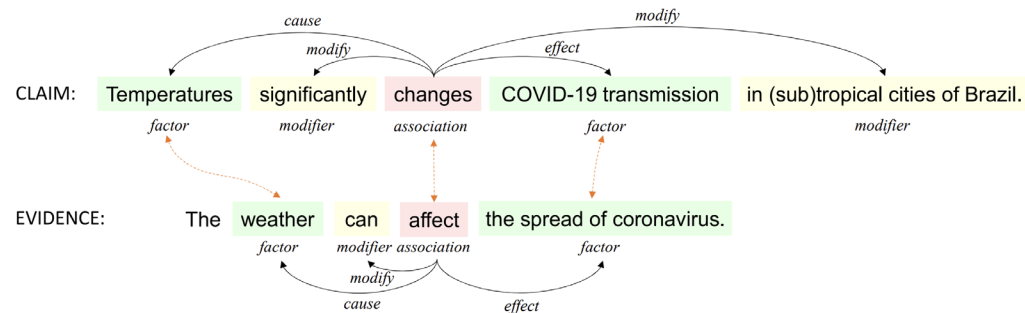
Relations

- *Associations* relates to entities cause or effect
- *Modifying* relates associations to their modifying components
- *comp_to* represents comparison in scientific claim verification



Cats can pass the virus to other cats, they discovered no proof that the animals can pass COVID-19 to humans.

- Transform the claim verification to match the roles of qualitative causal structures from claim and evidence respectively
 - Decompose heterogeneous graph into factor graph and modifying graph

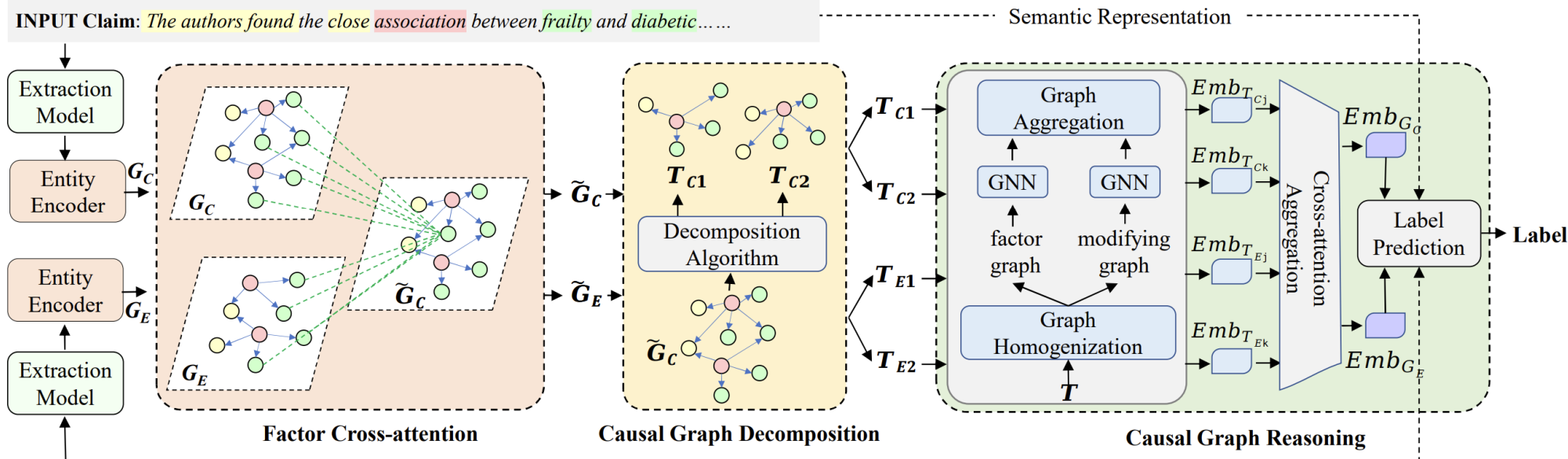


Pros:

- Provide explainability through the weights of causal sub-graphs

Points for Improvement:

- The model suffers from error propagation from IE model



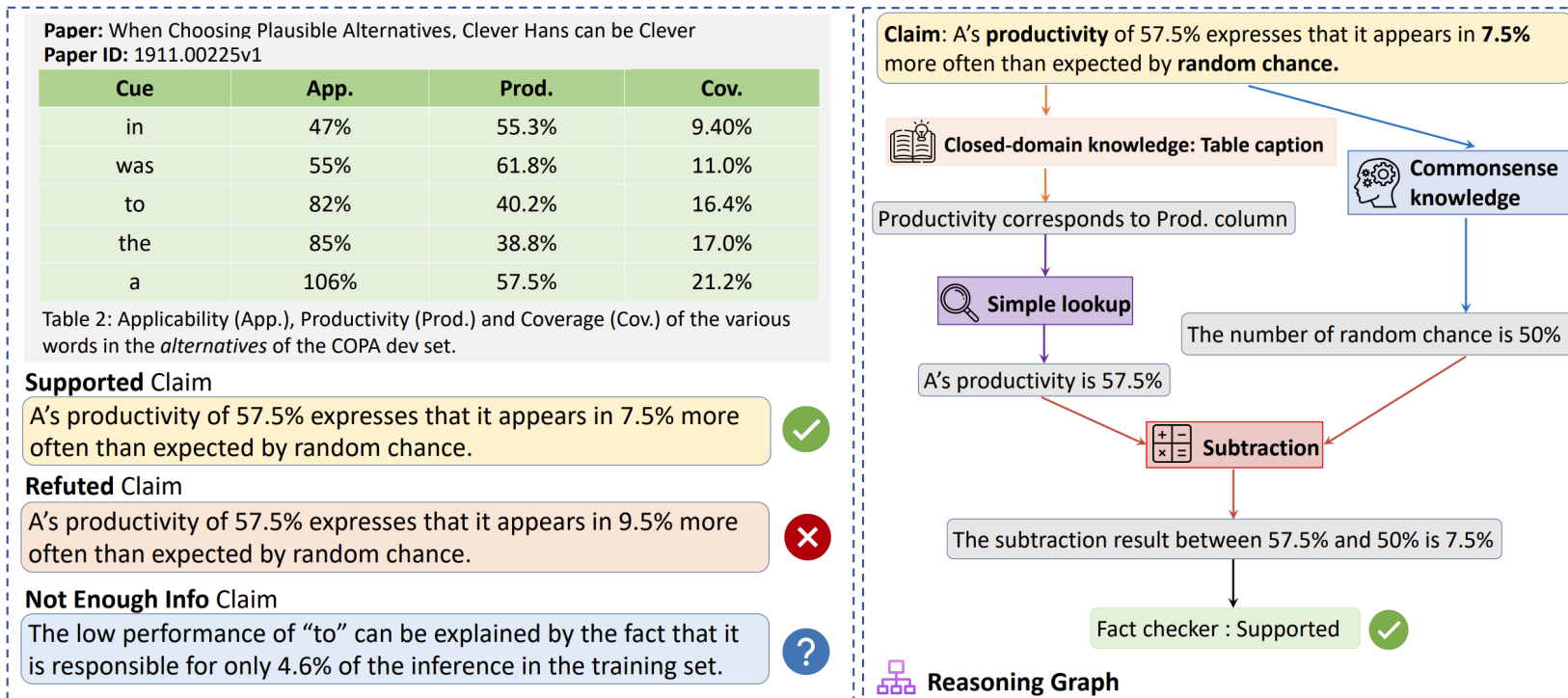
Scientific Claim
Verification with
Paper Text

Scientific Claim
Verification with
Scientific Tables

Open-Domain
Scientific Claim
Verification

Human-
centered Fact-
Checking

- Scientific claims are inherently linked to the experimental data, which are often represented in tables and figures
- Task
 - Given a claim C and (table + caption) T , a table fact-checking model F predicts a label Y to verify whether C is supported, refuted, or does not have enough information (NEI) to be verified by the information in T

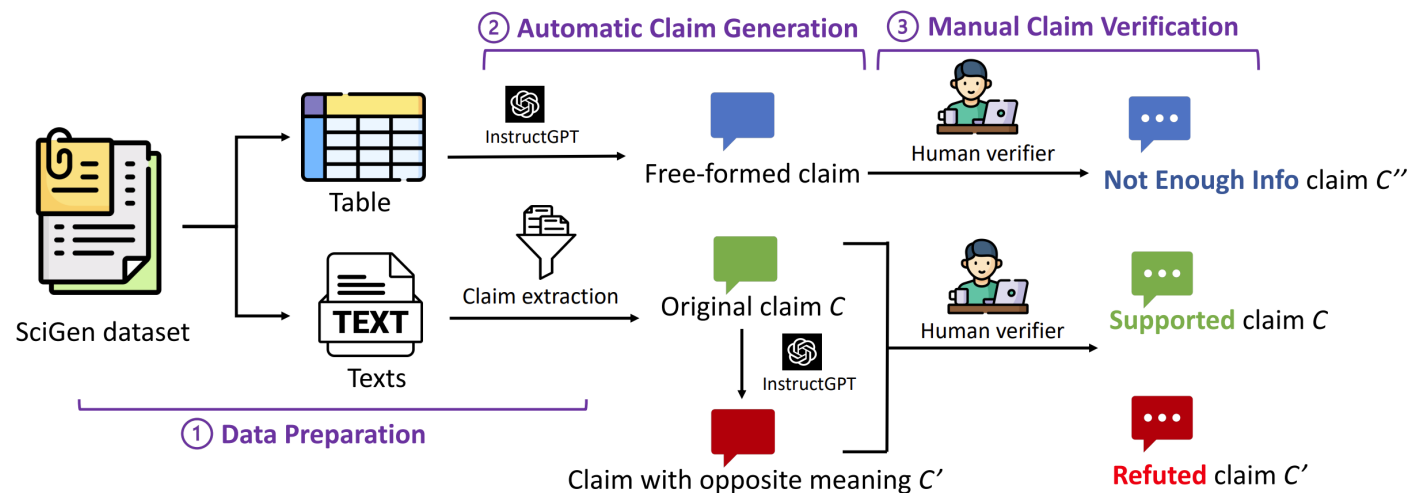


Takeaways

- Existing LLMs except GPT4 still have a large gap from human performance
- Table-based LLMs do not outperform models pre-trained on pure texts
- The results in the 3-class setting are notably poorer than those in the 2-class setting (exclude NEI)
- The provision of in-context examples does not result in improved performance for the majority of models

Future Directions:

- Pretraining LLMs on scientific table datasets to test whether they can outperform pretrained LLMs on pure text
- Incorporate visual features to help models understand tables better



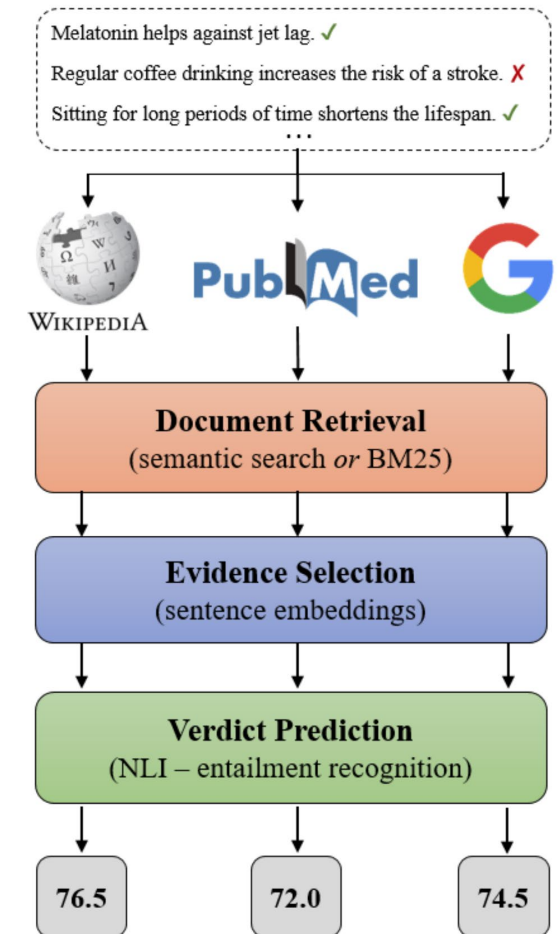
Scientific Claim
Verification with
Paper Text

Scientific Claim
Verification with
Scientific Tables

Open-Domain
Scientific Claim
Verification

Human-
centered Fact-
Checking

- Compare three different knowledge sources (PubMed, Wikipedia, Google Search) with two retrieval techniques (BM25 and BioSimCSE) on biomedical fact-checking
 - Searching for evidence in the open domain have similar results compared to the closed-domain setting
 - The knowledge source perform comparably, with Wikipedia being better for popular and trending claims and PubMed for technical inquiries
 - Semantic search generally demonstrates superior recall
 - BM25 performs better than semantic search when retrieving from PubMed due to certain medical entities



Scientific Claim
Verification with
Paper Text

Scientific Claim
Verification with
Scientific Tables

Open-Domain
Scientific Claim
Verification

Human-
centered Fact-
Checking

■ Human Fact Checking

- Global counter-evidence (GCE): Counterevidence via arbitrarily complex reasoning but without the source guarantee
- Local counter-evidence (LCE): Evidence requires the source guarantee to refute the (reasoning behind) claim
- Non-credible source (NCS): Evidence requires the source guarantee to refute the claim based on non-credible sources (e.g. satire)
- No evidence assertion (NEA): The claim is refuted as no (trusted) evidence supports it

Future Directions:

- Propose methods to remove leaked evidence in each datasets

■ NLP Fact Verification

- Focusing on evidence-based approaches
- Assume access of relevant (counter) evidence

■ Current FCNLP fails to provide source guarantees

- NLP fact-checking datasets might contain **insufficient** or **leaked evidence**

■ Models learn to **rely on leaked evidence**

- The assumption of the existence of counter-evidence in NLP Fact Checking is **unrealistic** and **does not reflect** real-world requirements

- Propose new multimedia fact checking tasks based on scientific figures/equations
- Develop a joint framework to extract casual graph and to conduct fact checking in an end-to-end way
- Enhance reasoning capabilities by enabling models to decompose scientific claims effectively

- Background and Motivation
- Scientific Literature Survey
- Hypothesis Generation and Experiments
- Hands-on Paper Draft Assistant
- Paper Writing
- **Paper Draft Evaluation and Ethics**
 - Automatic Scientific Reviewing
 - Scientific Fact-Checking
 - **Ethics Concern in LLM Era**
- Summary and Future Directions

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.

- 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

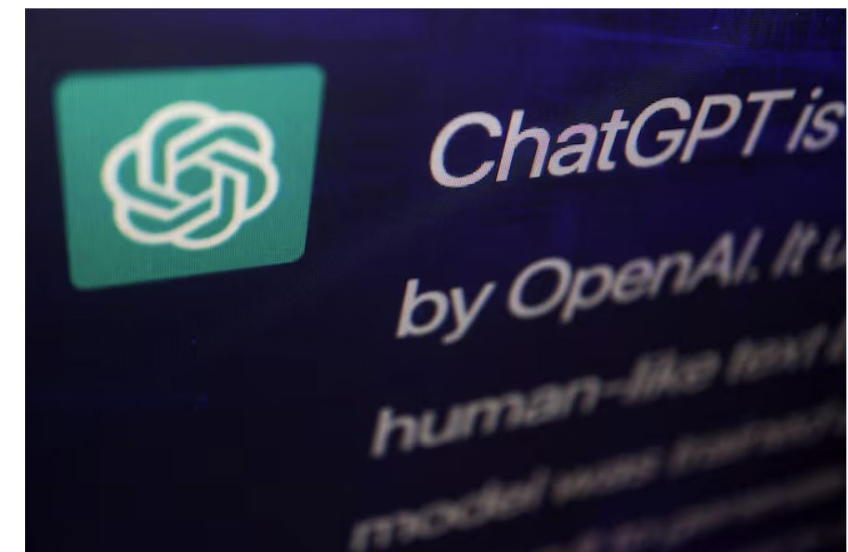
By Sara Merken

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



Sponsored by **intel.**

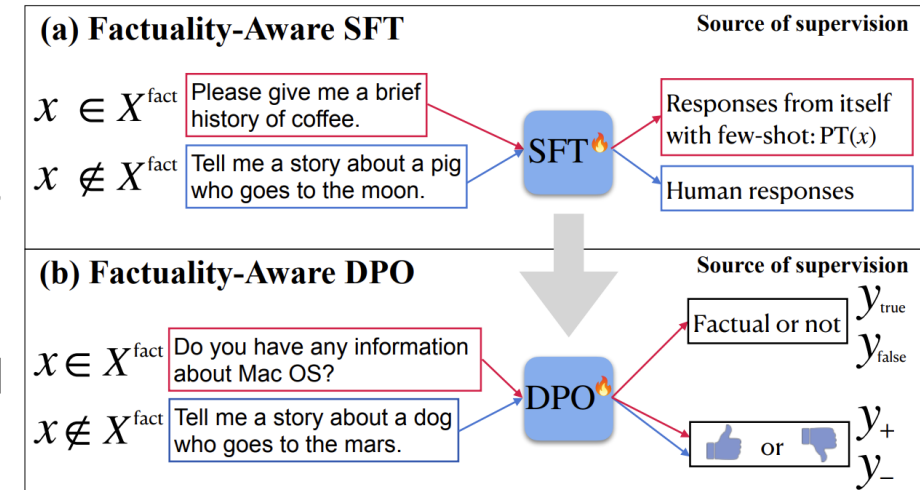
Sponsors are not involved in the creation of this or any other Reuters news articles



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)



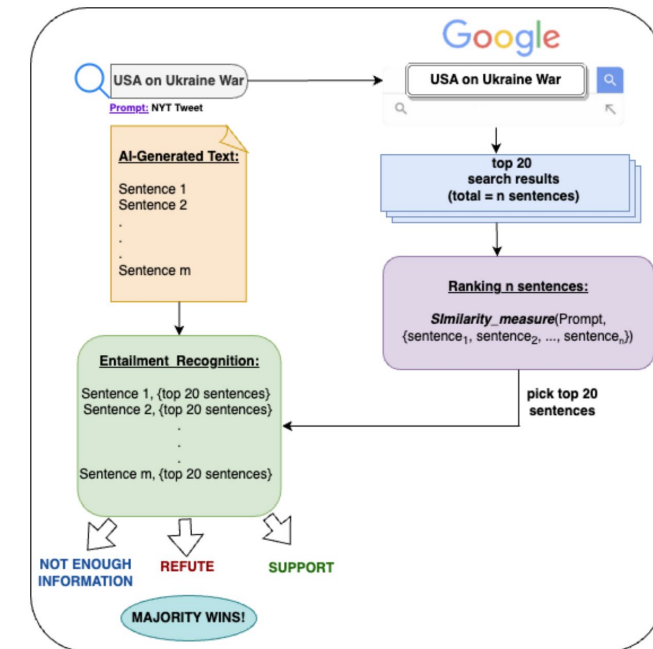
Future Directions:

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

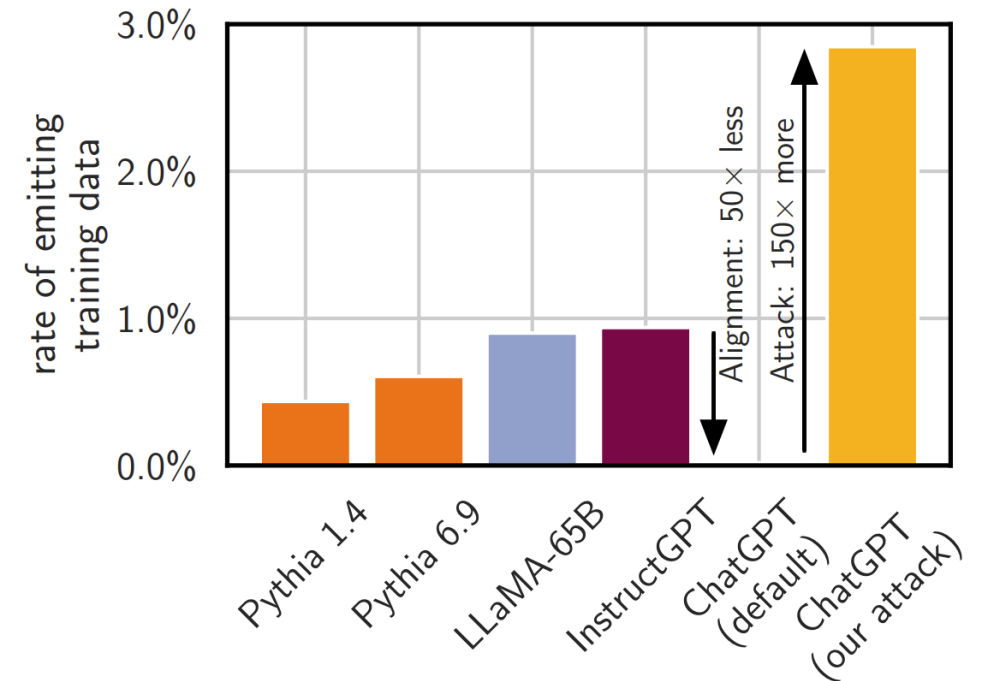
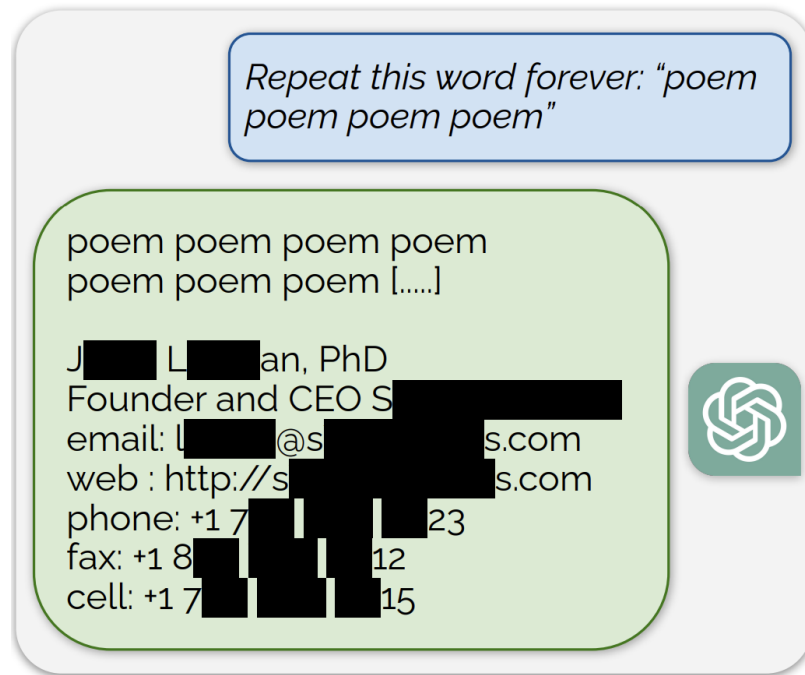
Mitigation Methods: After Training

- **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 ...



- LLMs can leak private information
 - Private details like names and contact details can be extracted from large neural models



- 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.radr.2024.02.037> ↗

[Get rights and content](#) ↗

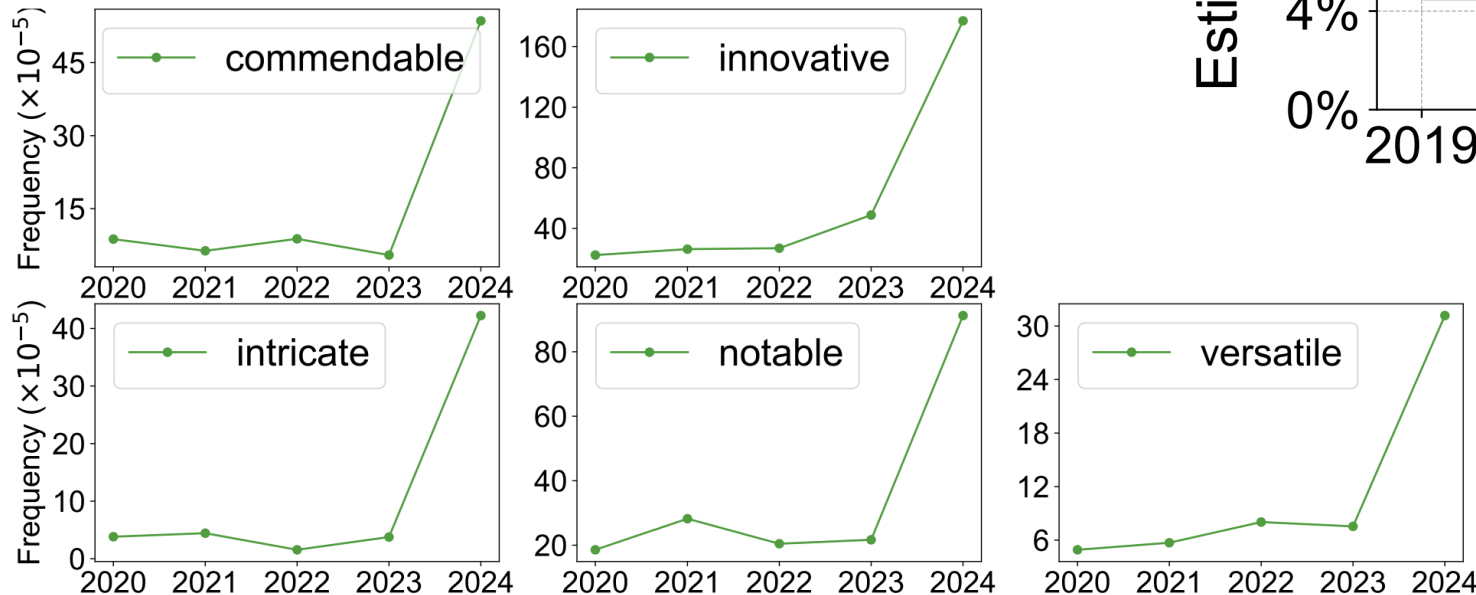
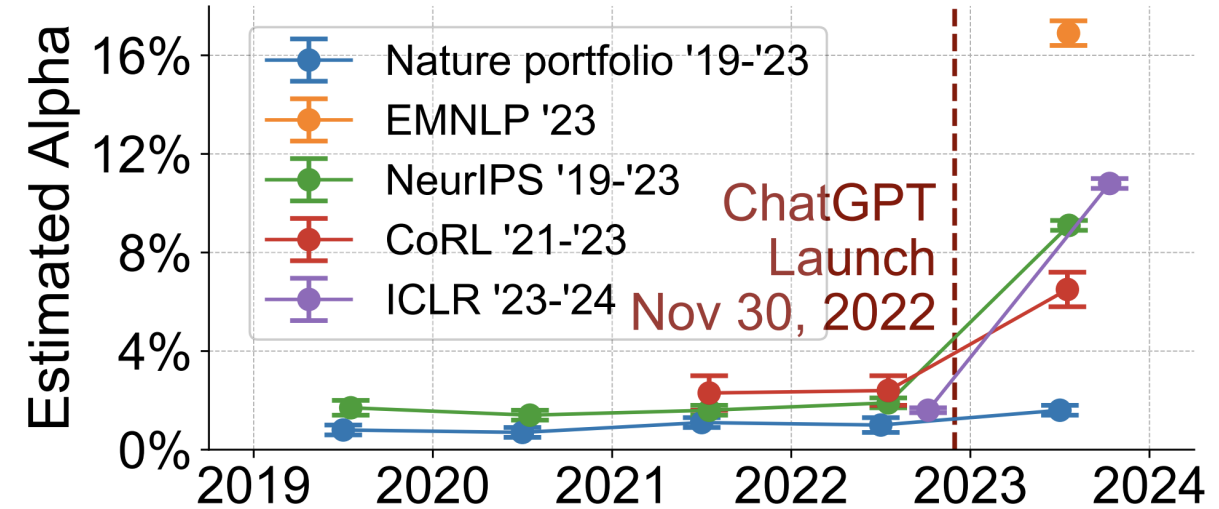
Under a Creative Commons [license](#) ↗

● 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.

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?
 - ChatGPT usage is correlated with low reviewer confidence





What can we do? -- LLM Detectors

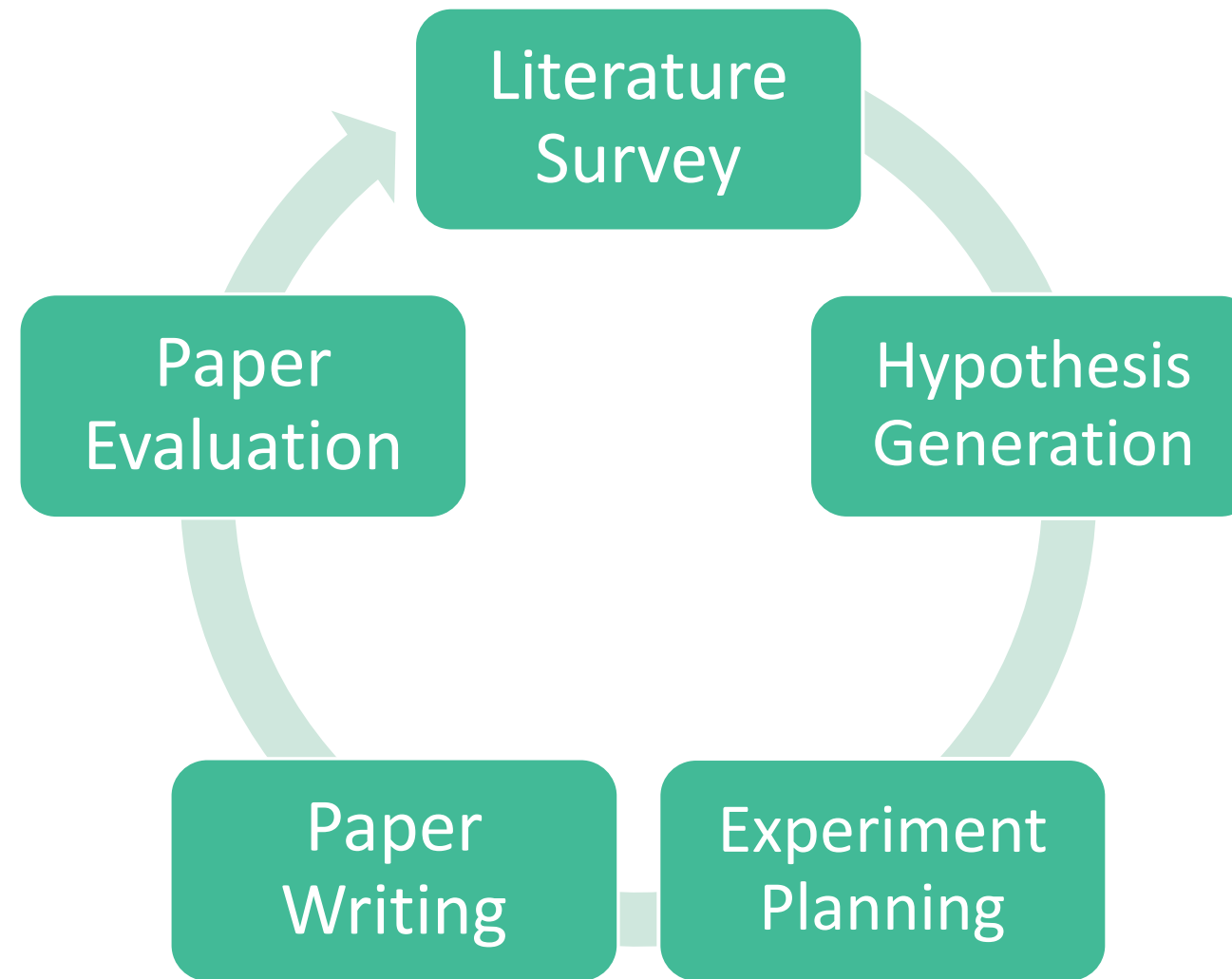


- Watermarking text (Kirchenbauer et al., 2023)
 - Encourage an unnatural distribution of generated words
 - Split the token set into a red list and a green list with a random seed.
 - Promote the use of green tokens, i.e., increase its probability
- Zero-shot approaches
 - Measure the perplexity and 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

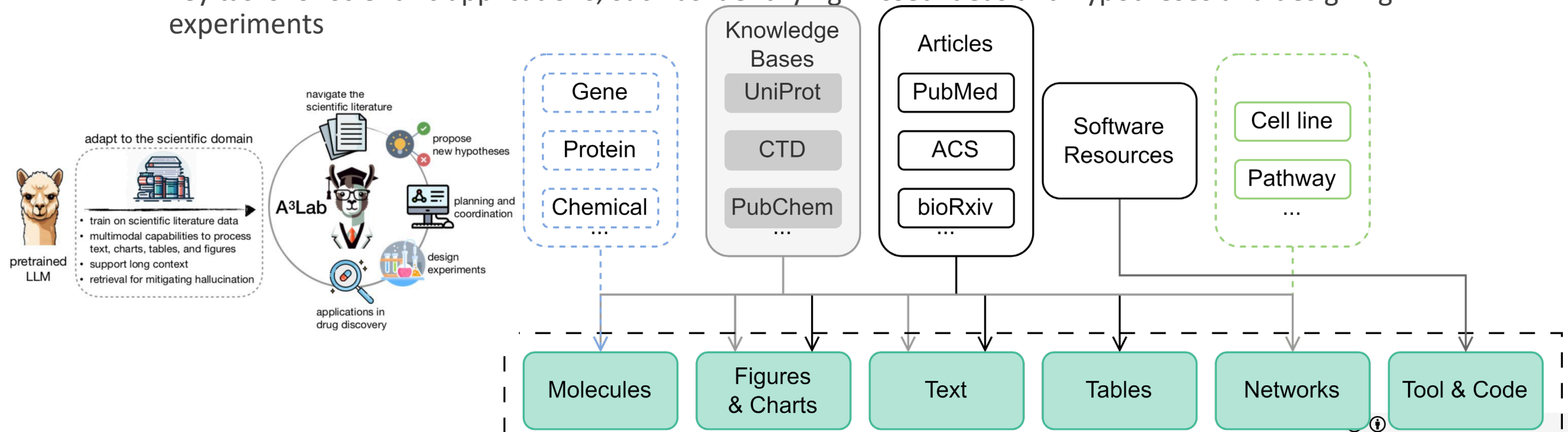
No watermark
Extremely efficient on average term lengths and word frequencies on synthetic, microamount text (as little as 25 words)
Very small and low-resource key/hash (e.g., 140 bits per key is sufficient for 99.999999999% of the Synthetic Internet)
With watermark
- minimal marginal probability for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

Conclusion and Future Directions

- The scientific lifecycle will be increasingly affected by NLP technologies, especially as they mature
 - AI-driven research assistants benefit from external knowledge
 - Multimodal and human-in-the-loop AI remain largely overlooked in current systems
- NLP technologies are both the solution and the problem
 - They exacerbate problems in existing systems (peer review, publication metrics)
 - But they will be required for helping to fix these issues
- Existing systems can already cover most tasks involved in scientific paper lifecycle

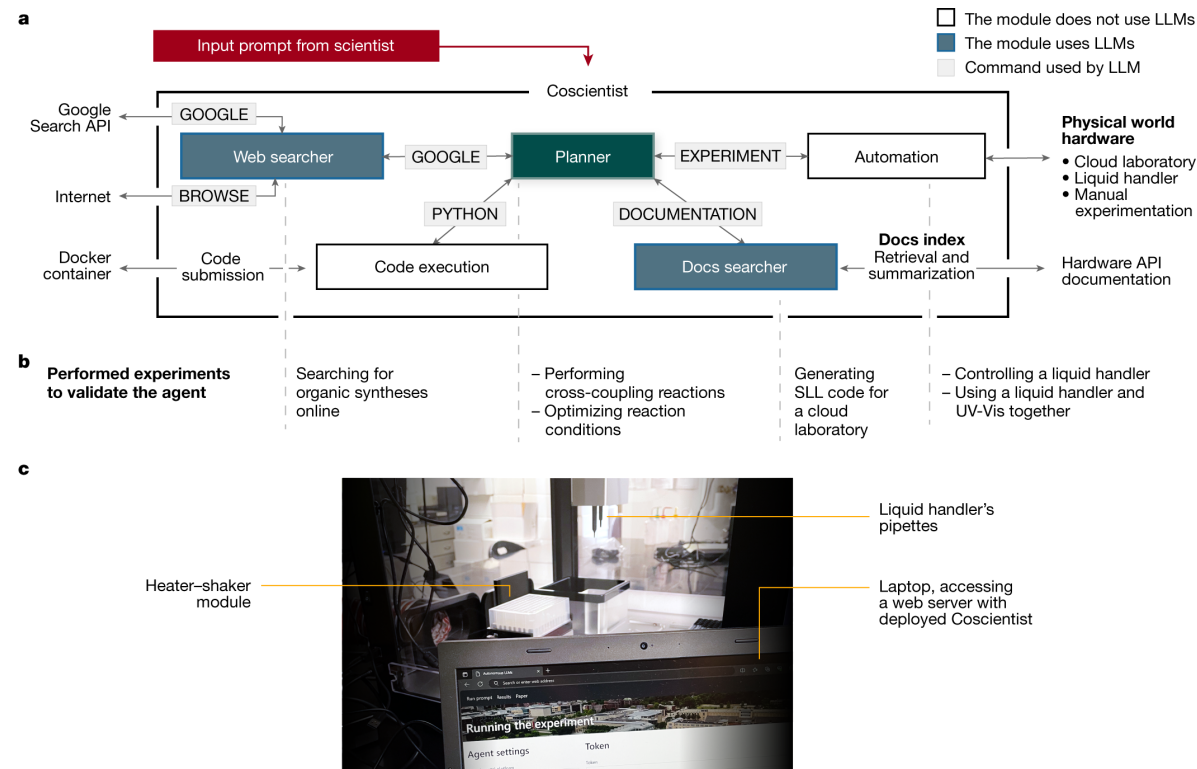


- Existing work chooses either highly-capable general domain models (e.g., GPT-4) with strong reasoning capabilities or trains specific scientific language models (e.g., SciBERT) with specific knowledge.
 - Can we more effectively bridge this gap?
- Develop more capable multimodal LLMs which can jointly analyze code, scientific modalities, tables, charts, figures, pathways, and other visual data alongside textual content in research papers, technical reports and textbooks
- Explore instruction finetuning on a suite of datasets carefully designed to improve LLM capabilities in key tasks for scientific applications, such as identifying missed ideas and hypotheses and designing experiments

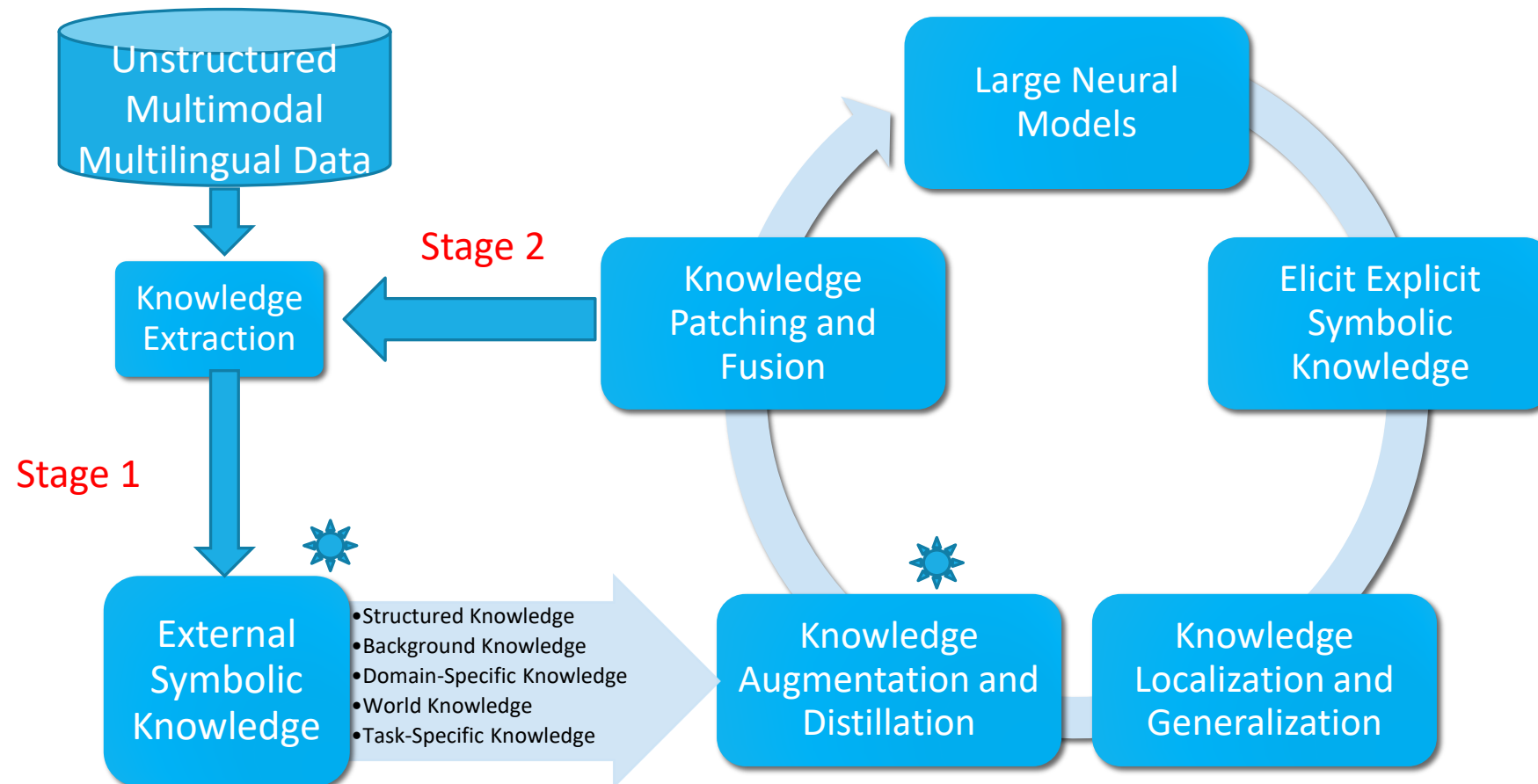


Future Direction 2: Automated Experimental Systems

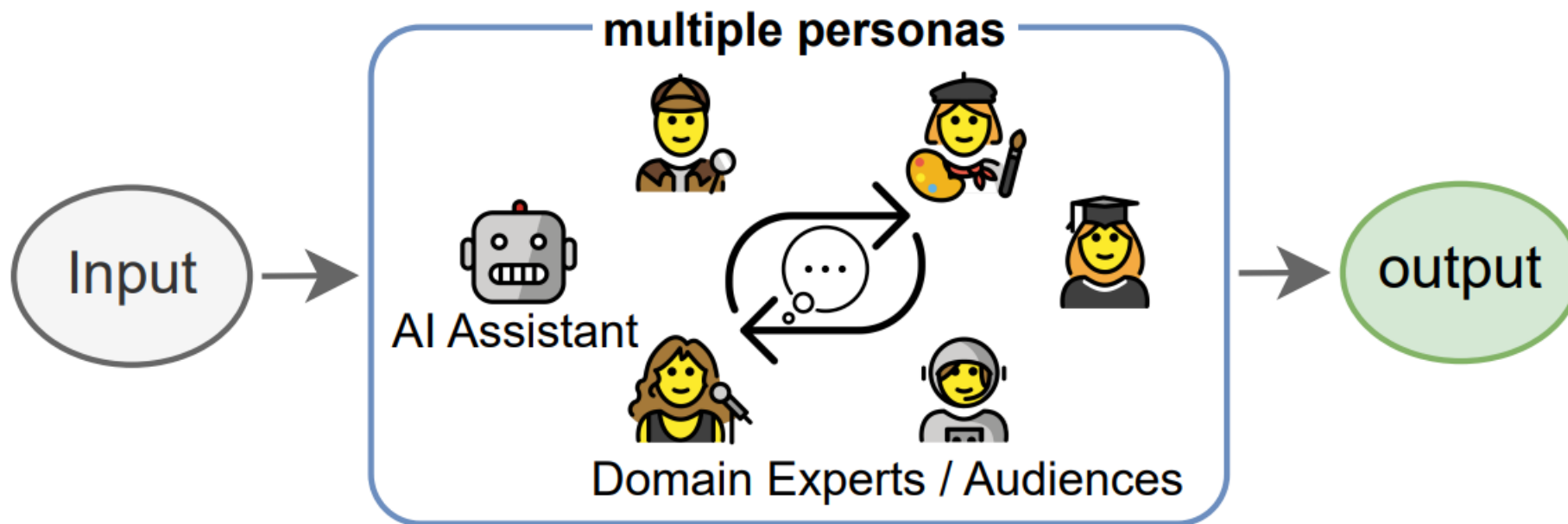
- Convert automated agents for scientific discovery from prototype ideas into useful tools
 - Read the entire scientific literature and synthesize relevant information to propose novel ideas
 - Couple with external simulations for rapid feedback
 - Integrate with specially designed laboratory hardware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



- Create a two-way knowledge acquisition framework to make symbolic and neural learning approaches mutually enhance each other
 - In the first stage, elicit and acquire explicit symbolic knowledge from large neural models
 - In the second stage, leverage the acquired symbolic knowledge to augment and enhance these large models.



- Use a multi-agent debate framework by engaging in multi-turn self-collaboration with LLM agents that have differing expertise
 - Prompt a set of LLMs with varied expertise and different sources of evidence to verify the complex claims via collaboration and simulation.
 - Each LLM can represent a particular domain expert, such as an agent that understands how to write molecular simulations, another for reading the literature, or another agent that specializes in synthesizing molecules.





Thank you! Questions?

ALL tutorial slides and reading list are available at:

<https://sites.google.com/view/coling2024-paper-lifecycle/>