LLMs for Research

Contents

LLMs in Research

- Literature Review
- Hypothesis Generation
- Experiment Design
- Data Collection
- Experiment
- Analysis
- Publication

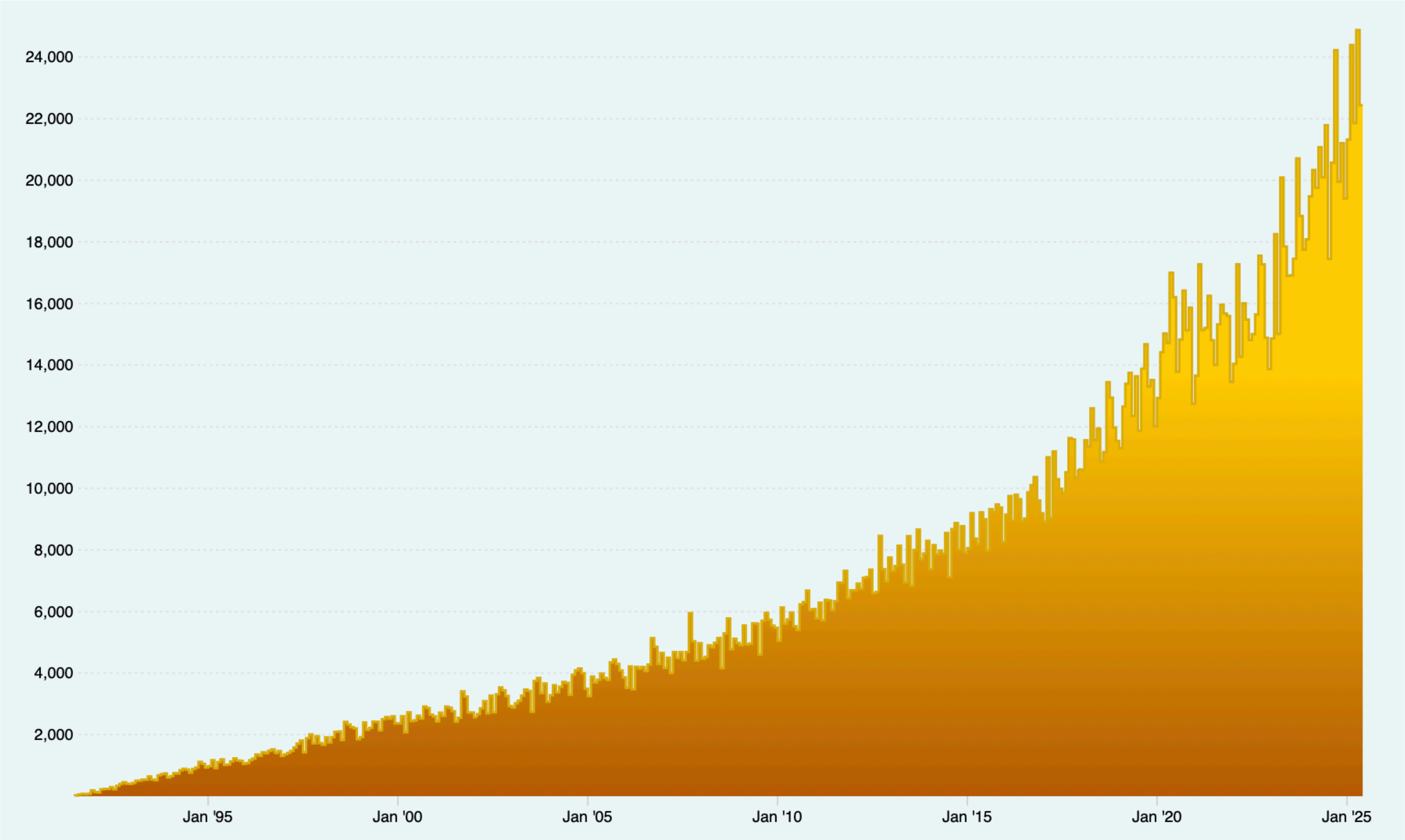
- Literature review is heavily **bottlenecked**
  - Researchers can manually process only ~50-100 papers per week, while relevant literature often spans hundreds of publications, possibly across multiple domains
  - New papers appear every week and it is extremely hard to be up to date
- Hypothesis generation is heavily **biased**
  - Relies heavily on individual researcher expertise and creativity
  - Identifying research gaps from limited personal knowledge can be limited as discoveries can emerge from cross-disciplinary connections and/or novel perspectives, which can be overlooked due to restriction to familiar domains / patterns
- Experiment planning involves complex decision-making about methodologies, parameters, and resource allocation with **limited prior experience**

LLMs in Research

- Scientific literature **grows exponentially** with hundreds of thousands papers published annually across all disciplines
- Knowledge fragmentation occurs as researchers struggle to connect insights across different (sub)fields; time constraints force researchers to focus on narrow specializations, missing potential cross-domain innovations
- Empirical analysis shows that the fraction of LLM-modified content in arXiv papers increased in several times between 2022-2024, with fastest growth in Computer Science [7]
- Higher LLM usage correlates strongly with more frequent preprint posting behavior and increased activity in crowded research areas, suggesting that researchers in rapidly evolving fields utilize LLMs to maintain competitive advantage and accelerate their research output [7]
- => LLMs are becoming integral tools for literature review, writing assistance, data analysis, and hypothesis generation

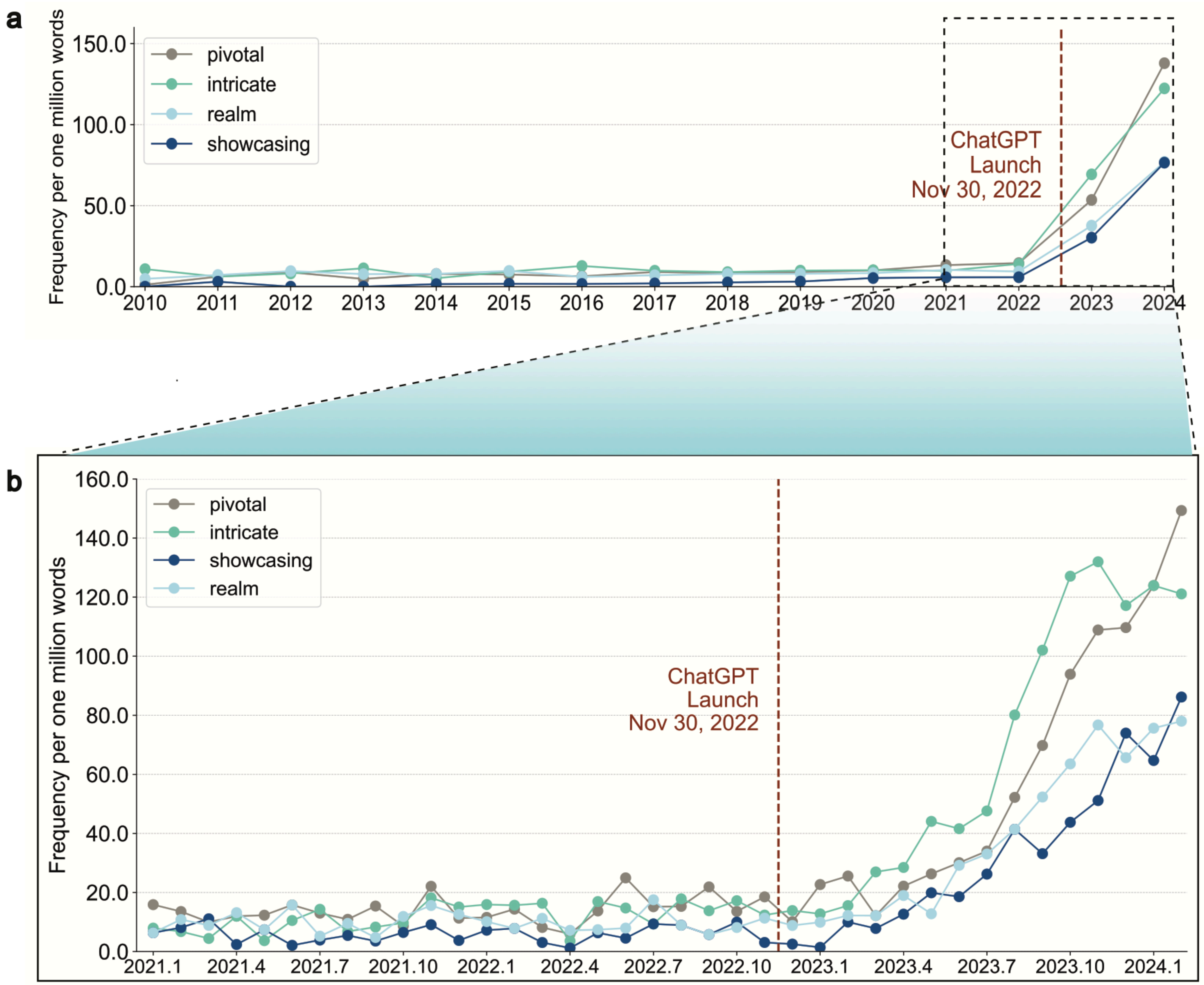
LLMs in Research

arXiv Monthly Submissions chart



LLMs in Research

Word frequency shift in arXiv abstracts (Computer Science) [7]



Literature Review

- LLMs don't have any physical limitations of humans: they can process and synthesize **thousands** of papers simultaneously [4]
- That also allows for **real-time** literature monitoring that keeps researchers updated with the latest developments without manual paper screening and selection
- Multi-document summarization techniques generate coherent overviews by identifying common themes, contradictions, and research gaps across multiple papers simultaneously => improved interconnection [4]
- Domain adaptation through fine-tuning on scientific texts improves understanding of technical terminology [4]
- Semantic search capabilities of LLM-based systems enable researchers to query literature using natural language questions rather than keyword-based searches, improving relevance and discovery [5]
- On top of that, interactive querying interfaces allow researchers to iteratively refine searches and ask follow-up questions in a chat manner

Literature Review

- For a more specific scientific domain, it becomes beneficial in terms of performance to pretrain and fine-tune a model on scientific corpora
  - *SciLitLLM* [4] was trained on 40M+ academic papers to improve technical language comprehension
  - Performance results show 15-25% improvement over general LLMs on scientific text understanding tasks [4]
- *Paper Copilot* [5] implements a **self-evolving** system that maintains personalized knowledge graphs for individual researchers based on their reading and current projects
  - Dynamic knowledge updating mechanism continuously ingests new publications, updates researcher profiles, and identifies emerging relevant research areas automatically
  - Recommendation engine combines collaborative filtering with content-based analysis to suggest papers, authors, and research directions aligned with user interests and goals [5]
  - Contextual summarization provides personalized paper summaries that highlight sections most relevant to user's current research questions and methodological approaches [5]

Hypothesis Generation

- Individual researchers generate **limited** hypothesis diversity due to cognitive biases, expertise boundaries, and tendency toward incremental research approaches [1]
  - Systematic exploration gaps occur when hypothesis spaces are incompletely searched
  - Feasibility assessment difficulties arise from incomplete understanding of experimental constraints, resource requirements, and methodological limitations [1]
- Different approaches to applying LLMs to hypothesis generation in (one-step setting) may include:
  - *Straightforward* application: "Given phenomenon X in domain Y, propose mechanisms that could explain…"
  - *Literature-conditioned* generation: provide the LLMs with relevant paper abstracts and asks them to identify gaps and propose novel research directions
  - *Cross-domain* analogy generation: ask the LLM to identify patterns from one research area and propose their application to different domains
  - *Contradiction-based* generation: present the LLM with conflicting research findings and ask it to propose hypotheses that could resolve observed inconsistencies

Hypothesis Generation

- *HypoGeniC* [1] shows that the LLMs can be a powerful backbone for hypothesis generation
- Iterative refinement based on labeled examples guides the LLM towards hypotheses with high explainability
  - The LLM generates initial hypotheses from small labeled examples, then iteratively refines them providing small portions of unseen data
  - Each hypothesis receives a reward based on how many existing data observations it can explain (namely, predicts correctly); the reward has an exploration coefficient to ensure balance between exploration and exploitation
  - Generates separate hypotheses for data subsamples that could not be predicted several times
- Iterative approach overcomes LLM context length limitations by starting with few examples and progressively incorporating more data, enabling analysis of arbitrarily large datasets (given enough budget)

Hypothesis Generation

- HypoGeniC [1] achieved a **significant increase** in performance (hypothesis-based reference): 31.7% accuracy improvement on synthetic data and 13.9-24.9% gains on real-world datasets compared to few-shot prompting; 3.3% on average on a OOD dataset!
- Generated hypotheses successfully replicated existing human-verified scientific theories, demonstrating that LLMs can rediscover established knowledge through pure data analysis
- Moreover, the system uncovered new patterns and insights not previously identified by human researchers, suggesting genuine scientific discovery potential
- A larger hypothesis pool generally led to better results => different portions of data may be affected by different factors
- Additionally, this research detected interchangeability of the LLMs for hypothesis generation vs reference

Hypothesis Generation

Some of HypoGeniC's hypotheses [1]

Dataset	Finding	Supported/Novel
DECEPTIVE REVIEWS	Deceptive reviews contain more emotional terms.	Li et al. (2014)
	Deceptive reviews are more likely to use superlatives.	Ott et al. (2011)
	Deceptive reviews contain hearsay or information that could not have been directly experienced.	Ott et al. (2011)
	Deceptive reviews tend to be more exaggerated.	Anderson and Simester (2014)
	Truthful reviews tend to use more balanced and objective tone.	Anderson and Simester (2014)
	Truthful reviews could mention the reviewer's purpose for staying at the hotel (e.g., business trip, vacation).	Novel
	Truthful reviews would mention weddings or special occasions.	Novel
	Truthful reviews may contain information about reviewer's expectations and previous hotel experiences.	Novel
	Truthful reviews would acknowledge the reviewer's personal biases or preferences.	Novel
	Deceptive ones may present the reviewer's opinion as objective facts.	Novel
	Truthful reviews may contain reviewers' past experiences or future travel plans.	Novel

Hypothesis Generation

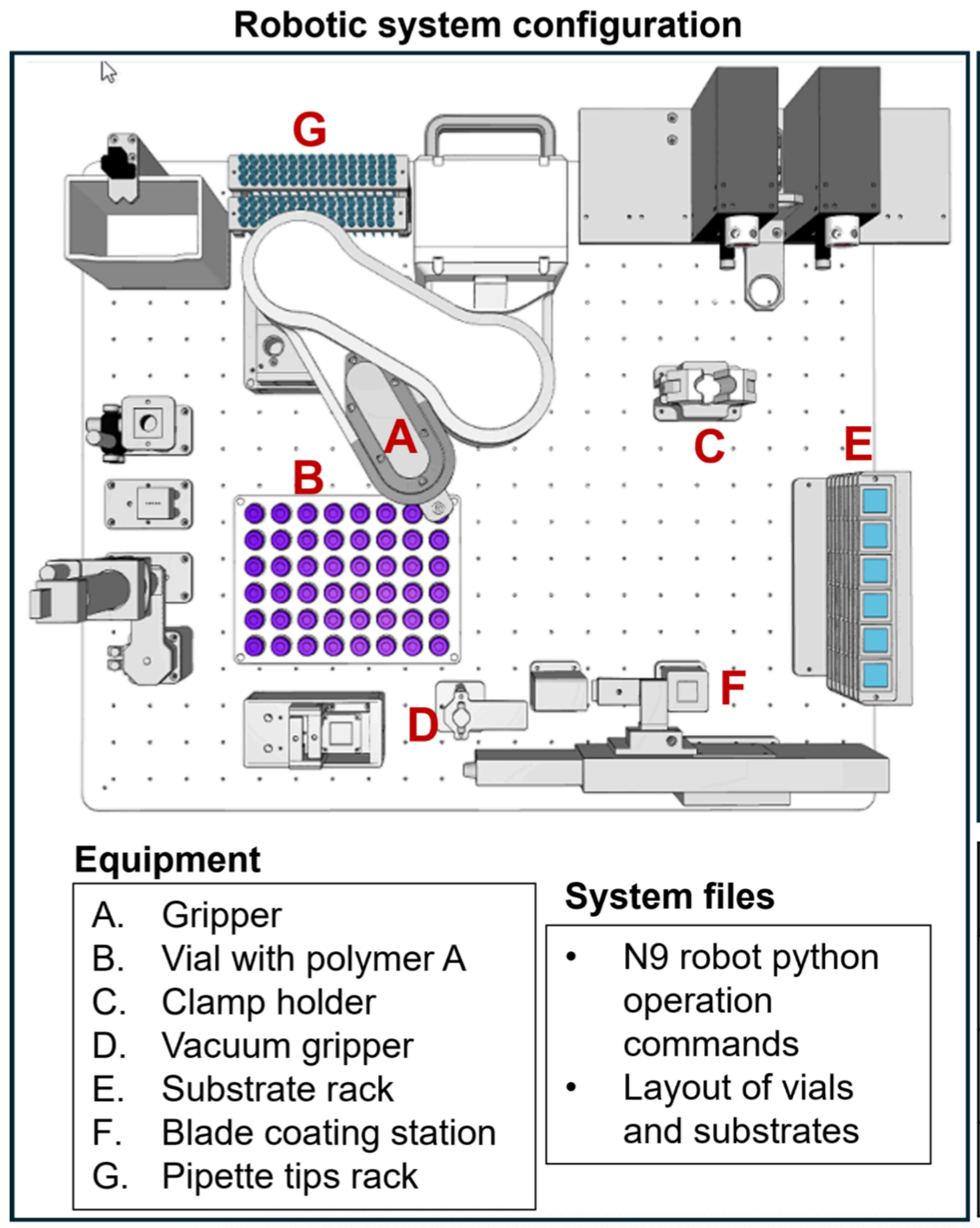
- Just as in business applications, the problem of **hallucinations** remains a great danger to the quality of the generations
- [6] integrates *knowledge graphs* (KG) to constrain and guide hypothesis generation toward **factually grounded** directions [6]
  - First, the relevant relations are retrieved from the KG, and relevant literature fragments are retrieved from a curated database
  - Based on the retrieved data, the initial query is enriched with relevant keywords
  - Augmented with the relations and the literature, the LLM generates the hypothesis in a CoT manner
  - A dedicated module checks separate reasoning thoughts, retrieves entities and checks if they actually are related somehow in the KG => assigns overall answer confidence
- Consistent and significant performance improvements on a biological relations dataset (3-fold classification task) on both closed-, open-source, small and larger models
- Ablation studies showed that all components / techniques (relations, literature excerpts, query enrichment, CoT prompting) turned out to be contributing to the quality of the generation

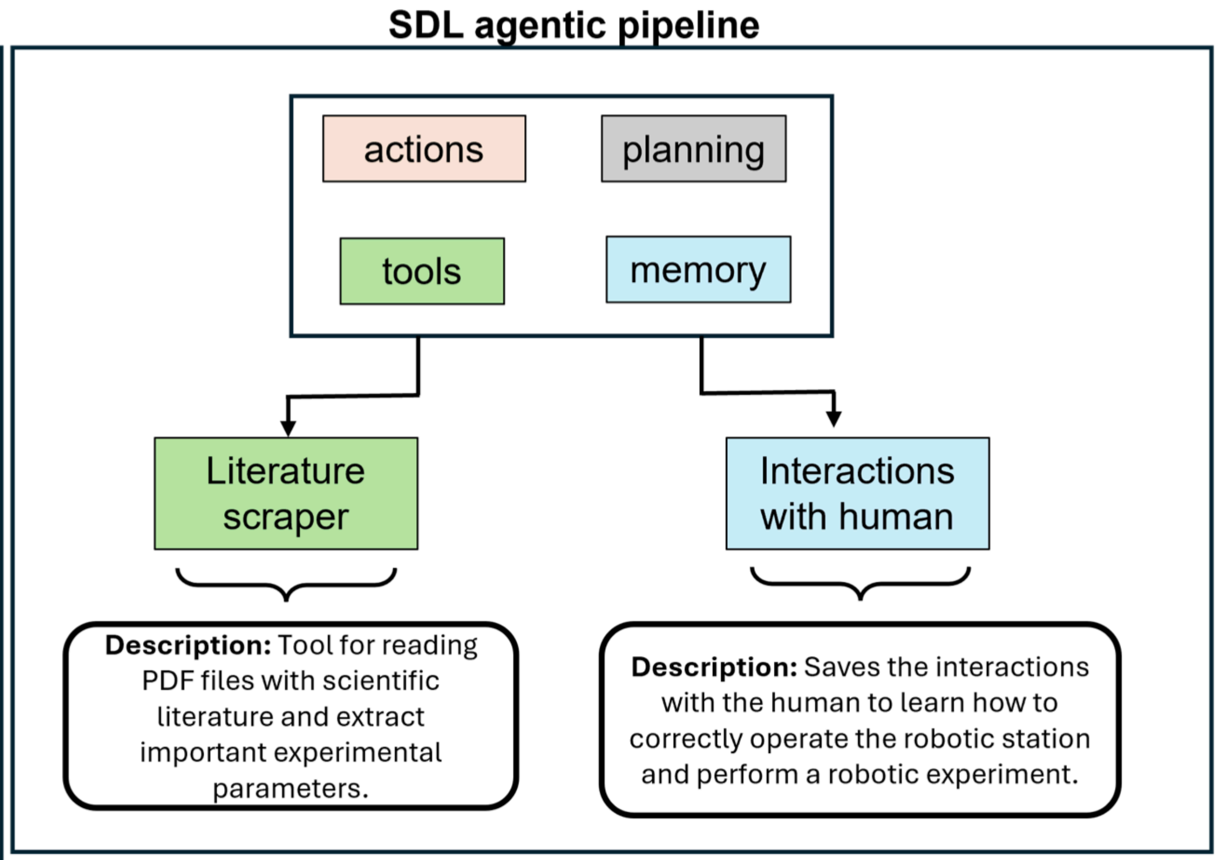
End-to-end Solutions

- LLMs can translate natural language queries to experimental *protocols* converting high-level research objectives into detailed, executable experimental procedures with specific parameters: researchers can specify experimental goals using conversational interaction: "Test compound X against cell line Y with controls Z" [2, 9]
- Systems like [2, 9] are aware of the real-world equipment they have as well as provided the documentation to invoke it programmatically based on the protocols generated from the researcher's input
- Learning mechanisms allow robotic systems to improve performance over time by analyzing experimental outcomes and refining procedural efficiency like in [9]
- Finally, system like AI Scientist [10] perform literature review, hypothesis generation, experimental design, data analysis, and manuscript writing without human intervention

End-to-end Solutions

- [9] was built on Autogen with 6 specialized AI agents (code writer, critic, administrator, paper scraper, manager, teachability) controlling an N9 robot
- It incorporated modules for literature and feedback storage





	Tasks	Complexity	No of sequential steps
Task 1	Pick up the vial with PEDOT:PSS and move it to the clamp.	Easy	8
Task 2	Pick up a substrate and move it to the coating station.	Medium	10
Task 3	Create a polymer film on a substrate using PEDOT:PSS.	Hard	25

- The system could successfully plan sequences of actions for complex operations involving invoking the very robot to perform operations in a real lab
- The final product of the pipeline is code for the robot; out of several metrics for assessing the code quality (correctness, execution, reproducibility etc.), all reached 100% for bigger models with human guidance (but were much worse on more complex tasks without it)

End-to-end Solutions

Example from [2]

F.

You have a source plate with three samples in unknown positions (yellow, blue, and red), each potentially located in one of the wells (A1, B1, and C1). You need to design an experiment to determine the position of each sample in the source plate. <setup description>

User prompt

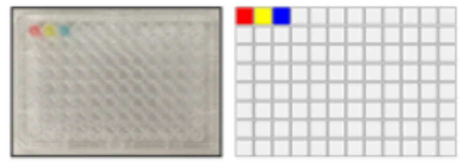
First, I will transfer a small amount of each sample from the source plate to separate wells in the target plate. Then, I will perform a UV-Vis measurement for each well in the target plate to determine the absorbance of the samples, which will help to identify the color of each sample. Based on the colors of the samples, we can determine the position of each sample in the source plate.

```
CLOUD
from opentrons import protocol_api
...

# Perform UV-Vis measurement on the target plate
UVVIS plate 1
```

In the code the Agent moves liquids from A1, B1, and C1 to A1, A2, and A3. Then it dilutes them.

Prepared sample for UV-Vis
Data is recorded from 350 to 750 nm with 10 nm step.



Planner's output

UV-Vis data from plate 1 is ready. You can access it using PYTHON, it will be in the file named 'plate1.npy'. You will have NumPy installed. The file contains NumPy array of shape (8, 12, 41). There are 41 wavelengths from 350 to 750 with 10 nm step. **Think about how would different colors absorb first.**

Automated answer
+ guiding prompt

Different colors absorb light at different wavelengths. Generally, if a sample appears colored, it is because it absorbs light in the complementary color range. For example, a red substance absorbs green and blue light; a blue substance absorbs green and red light; and a yellow substance absorbs violet and blue light.

First, I will load the UV-Vis data from the file 'plate1.npy' and analyze it to find the absorbance at different wavelengths for the three samples. Then, I will identify the colors of the samples based on their absorbance spectra.

```
PYTHON
import numpy as np
...
print(samples_max_wavelength_nm)
```

[530, 430, 630]

Automated answer

The maximum absorbance wavelengths for the three samples are 530 nm, 430 nm, and 630 nm. Based on the absorbance spectra, we can deduce their colors:

1. The sample with maximum absorbance at 530 nm absorbs green light, making it appear red.
2. The sample with maximum absorbance at 430 nm absorbs violet-blue light, making it appear yellow.
3. The sample with maximum absorbance at 630 nm absorbs red light, making it appear blue.

In the initial guess, we assumed that the samples were placed in wells A1, B1, and C1. However, based on the absorbance data analysis, we find:

1. The red sample is located in well A1.
2. The yellow sample is located in well B1.
3. The blue sample is located in well C1.

Verification and Validation

- LLM hallucinations pose significant risks in scientific contexts where accuracy and reliability are necessary [6]
- [8] provides evidence that not only plain text but also code generation may sometimes be unreliable for scientific research (to be fair, so is it for other applications)
- We have already seen a few ways to mitigate this problem: knowledge graph and CoT reasoning steps evaluation [6], human-AI collaboration as in [9]
- LLM-based pipelines for predicting peer reviews can indirectly be used for quality assessment as they might generate factual criticism [3]
- Benchmark development establishes standardized evaluation metrics for LLM research assistance tools, including accuracy and reproducibility scores [3]; those are basically mostly usual benchmarks (QA and similar) on scientific material

References

[1] Hypothesis Generation with Large Language Models, University of Chicago & Toyota Technological Institute at Chicago

[2] Emergent autonomous scientific research capabilities of large language models, Carnegie Mellon University

[3] LLM4SR: A Survey on Large Language Models for Scientific Research, University of Texas at Dallas & Nanyang Technological University

[4] SciLitLLM: How to Adapt LLMs for Scientific Literature Understanding, University of Science and Technology of China & DP Technology

[5] Paper Copilot: A Self-Evolving and Efficient LLM System for Personalized Academic Assistance, University of Illinois at Urbana-Champaign, Carnegie Mellon University & Carleton College

[6] Improving Scientific Hypothesis Generation with Knowledge Grounded Large Language Models, University of Virginia

[7] Mapping the Increasing Use of LLMs in Scientific Papers, Stanford

[8] LLMs for Science: Usage for Code Generation and Data Analysis, TUM

[9] Operating Robotic Laboratories with Large Language Models and Teachable Agents, Argonne National Laboratory

[10] The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery, Sakana AI et al.