Title: LitLLMs, LLMs for Literature Review: Are We There Yet?

Speakers: Gaurav Sahu

Collection/Series: Theory + AI Symposium

Date: April 08, 2025 - 4:10 PM

URL: https://pirsa.org/25040076

Abstract:

Literature reviews are an essential component of scientific research, but they remain time-intensive and challenging to write, especially due to the recent influx of research papers. In this talk, we will explore the zero-shot abilities of recent Large Language Models (LLMs) in assisting with the writing of literature reviews based on an abstract. We will decompose the task into two components: 1. Retrieving related works given a query abstract, and 2. Writing a literature review based on the retrieved results. We will then analyze how effective LLMs are for both components. For retrieval, we will discuss a novel two-step search strategy that first uses an LLM to extract meaningful keywords from the abstract of a paper and then retrieves potentially relevant papers by querying an external knowledge base. Additionally, we will study a prompting-based re-ranking mechanism with attribution and show that re-ranking doubles the normalized recall compared to naive search methods, while providing insights into the LLM's decision-making process. We will then discuss the two-step generation phase that first outlines a plan for the review and then executes steps in the plan to generate the actual review. To evaluate different LLM-based literature review methods, we create test sets from arXiv papers using a protocol designed for rolling use with newly released LLMs to avoid test set contamination in zero-shot evaluations. We will also see a quick demo of LitLLM in action towards the end.

TMLR 2025

LitLLMs, LLMs for Literature Review: Are We There Yet?

Shubham Agarwal*, Gaurav Sahu*, Abhay Puri*
Issam H. Laradji, Krishnamurthy DJ Dvijotham, Jason Stanley,
Laurent Charlin, Christopher Pal

https://LitLLM.github.io

Affiliations:
MILA - Quebec AI Institute
University of Waterloo
ServiceNow Research

*equal contribution

The Challenge: Writing Literature Reviews

- Writing literature reviews is time-consuming
- Difficult to keep up with the trends:
  - >3000 papers/month published on arXiv just for machine learning

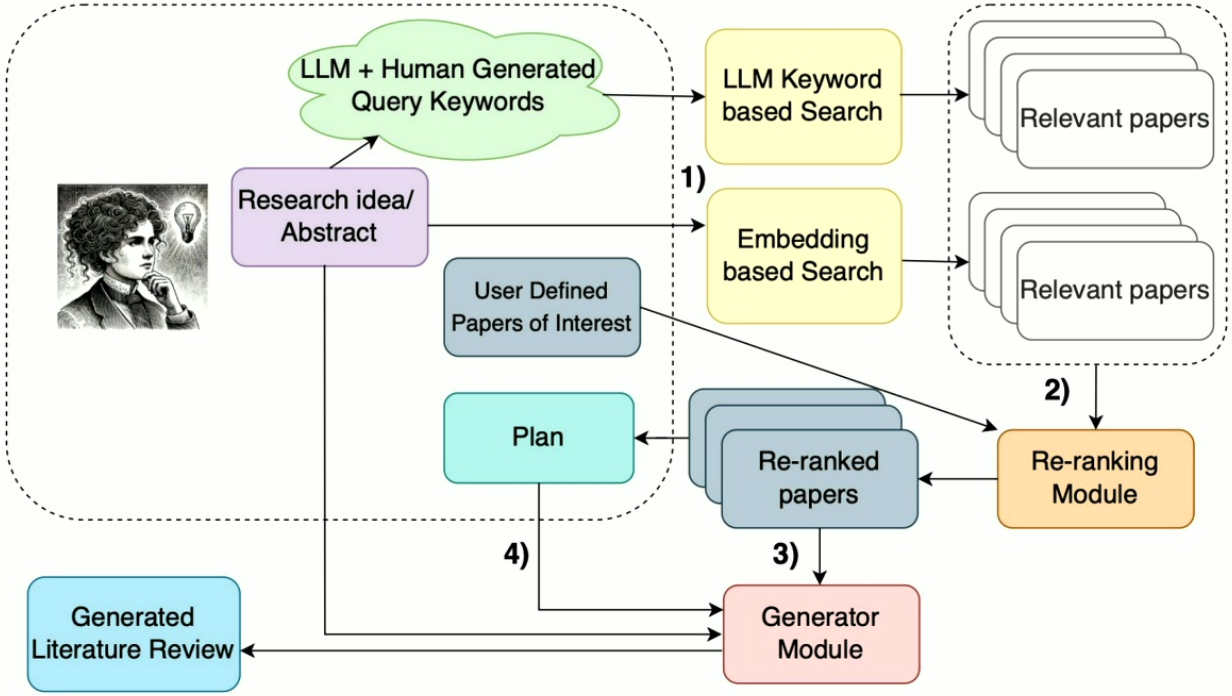
Our Focus

Writing literature review

Our Focus

- Explore the zero-shot abilities of LLMs to assist with writing literature reviews

Approach

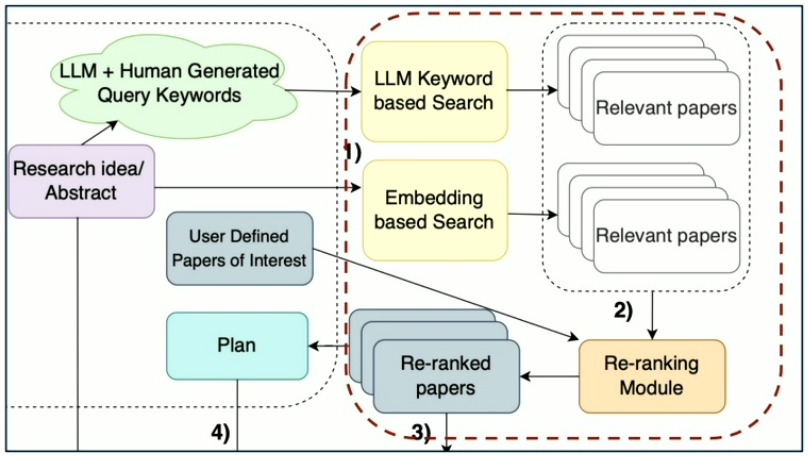

LLM + Human Generated Query Keywords
LLM Keyword based Search
Relevant papers
1)
Research idea/ Abstract
Embedding based Search
Relevant papers
User Defined Papers of Interest
2)
Plan
Re-ranked papers
Re-ranking Module
4)
3)
Generated Literature Review
Generator Module

Approach: Retrieval module

Keyword-based search
- Semantic Scholar
- SERP

Embedding-based search
- Specter2

Re-ranking:
- Specter2 embeddings
- LLM-based reranking
  - Permutation ([3] > [1] > [2])
  - Debate ranking with Attribution

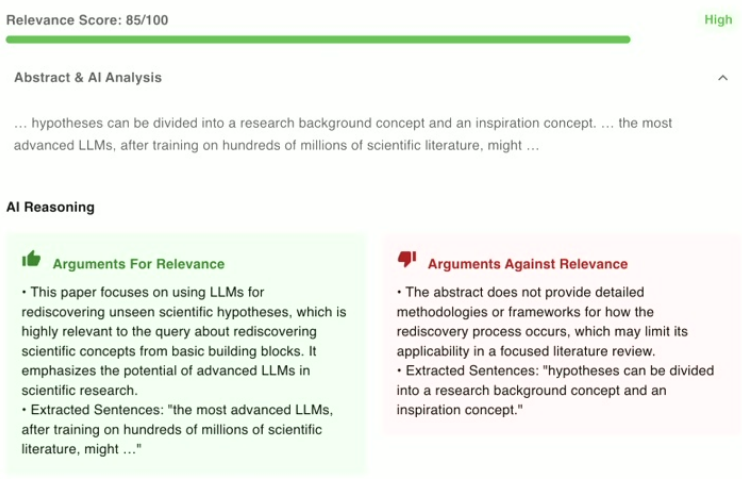


Approach: Retrieval module

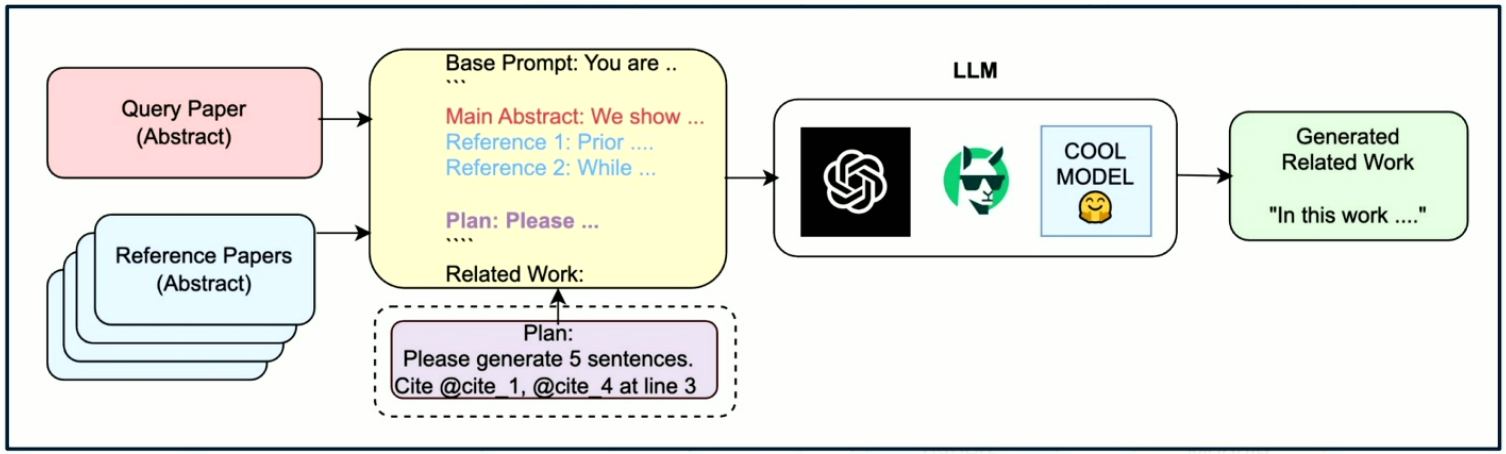
Keyword-based search
- Semantic Scholar
- SERP

Embedding-based search
- Specter2

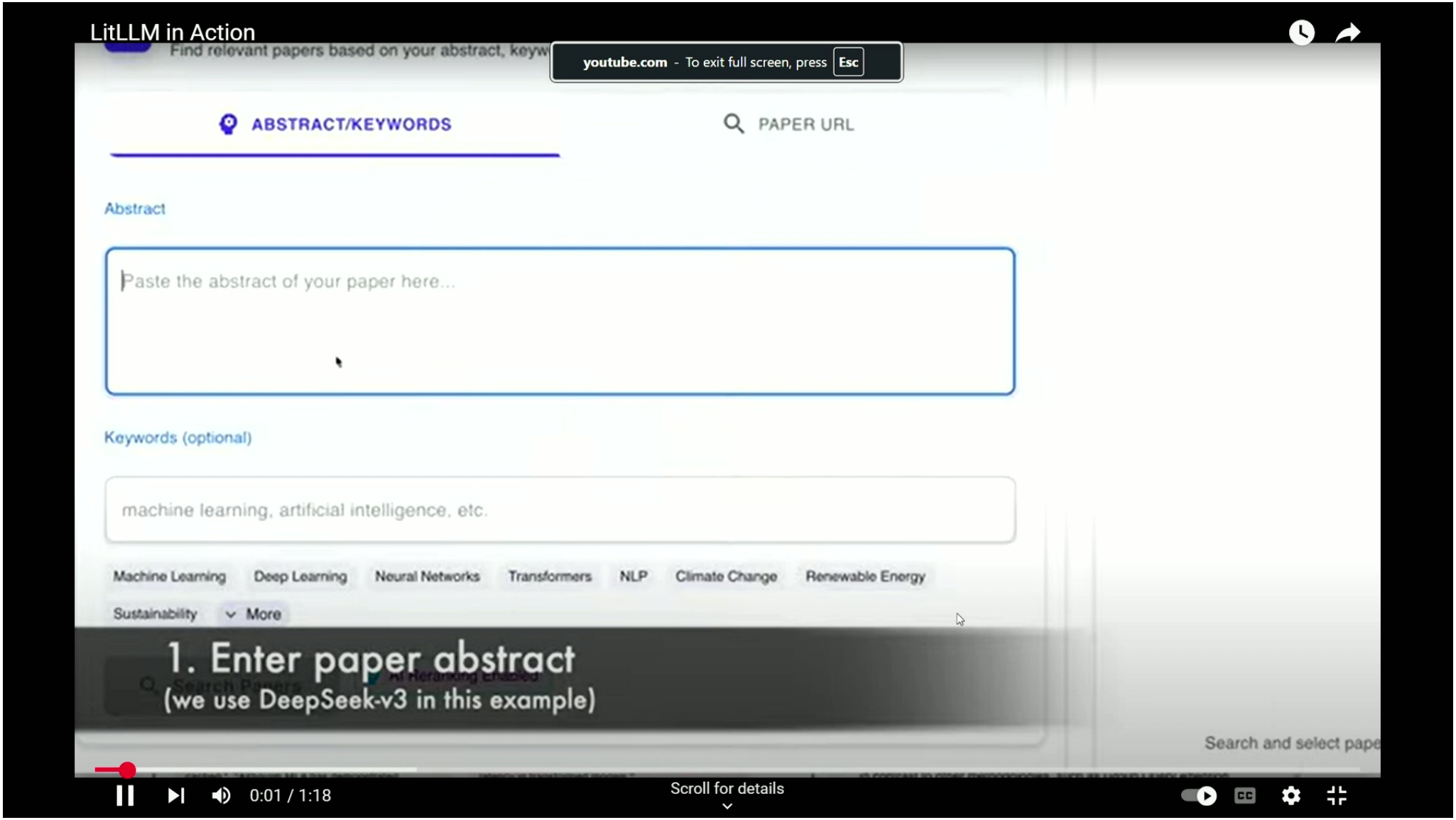
Re-ranking:
- Specter2 embeddings
- LLM-based reranking
  - Permutation ([3] > [1] > [2])
  - Debate ranking with Attribution

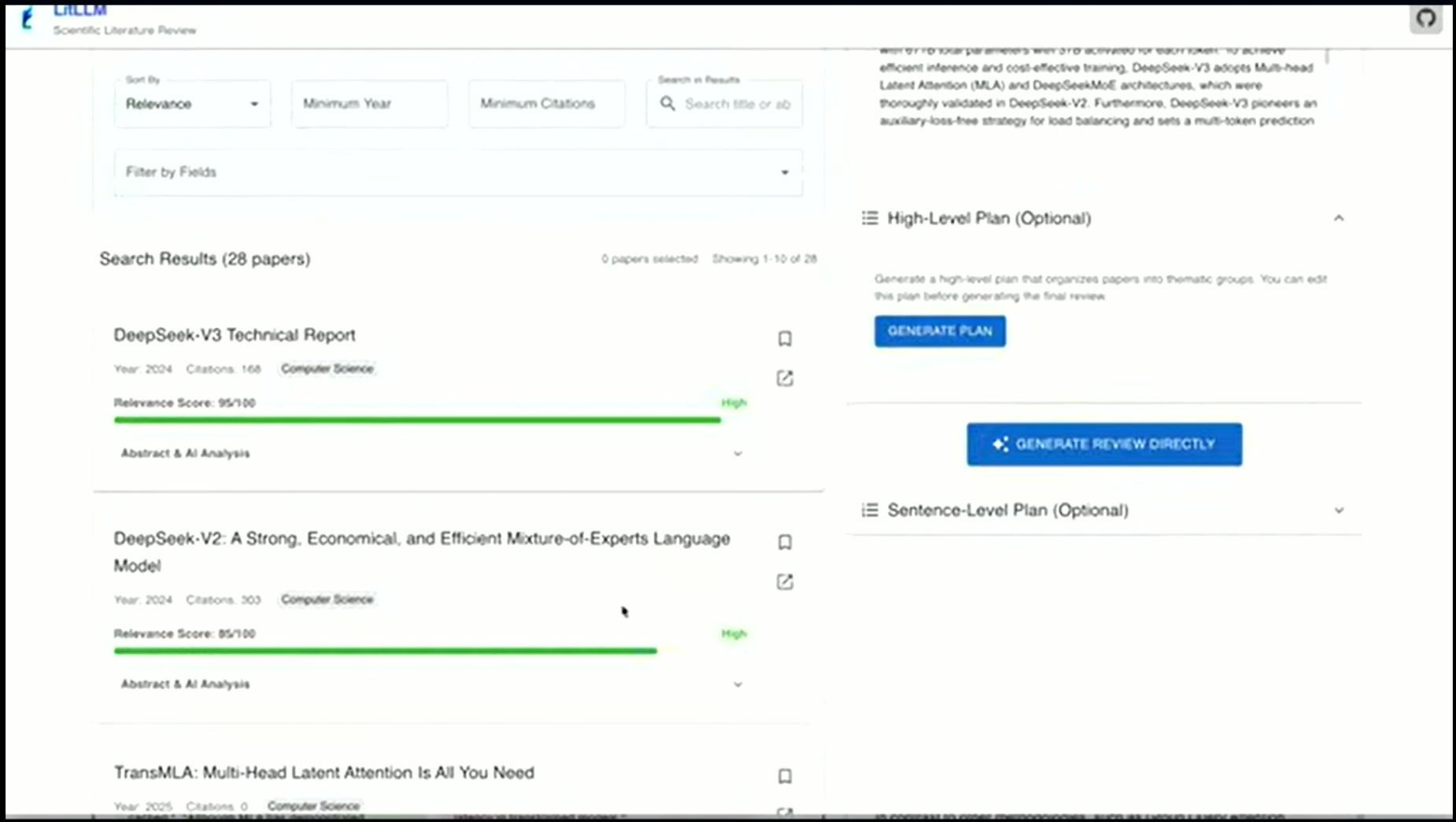


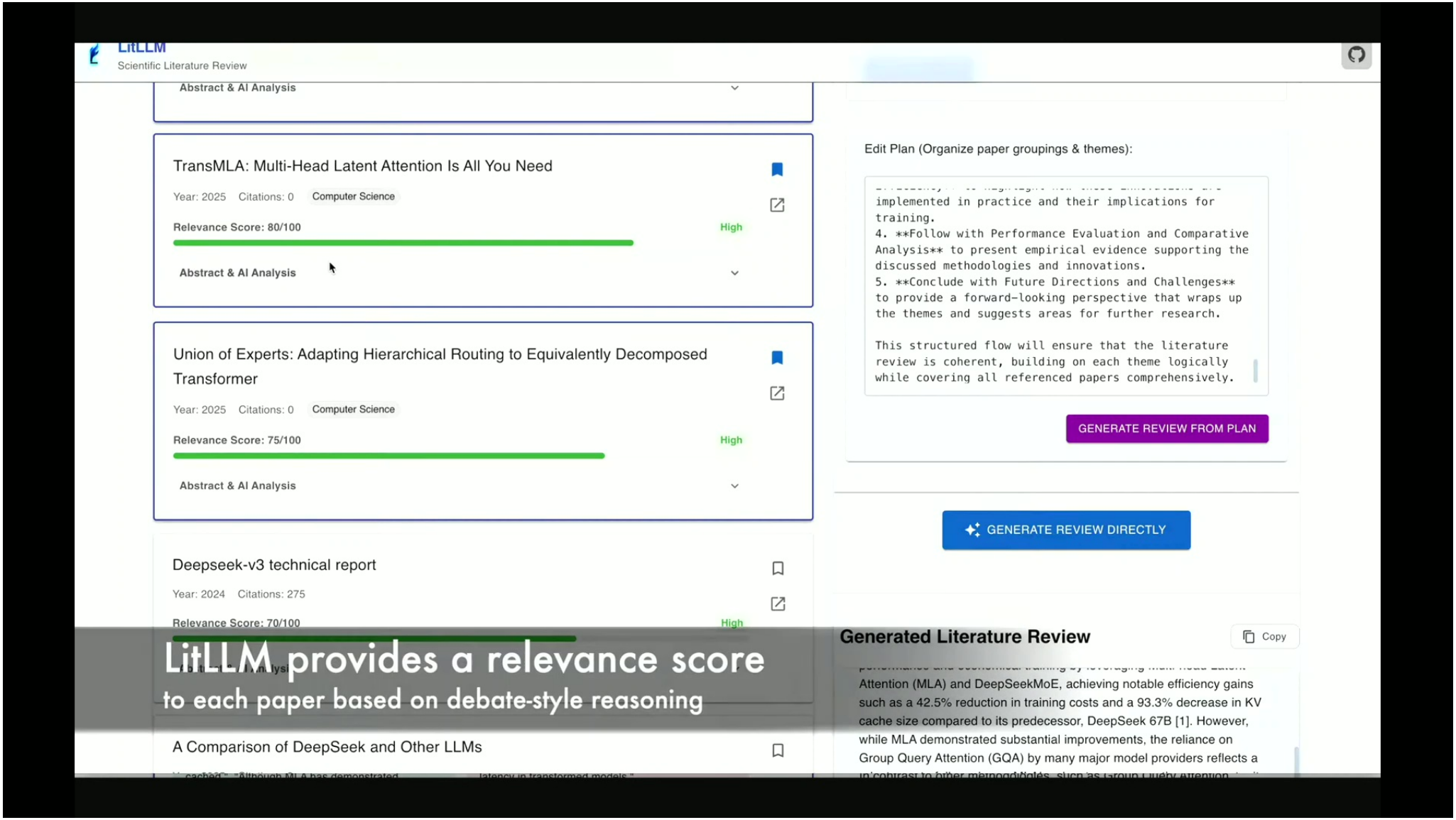
Approach: Generation Module

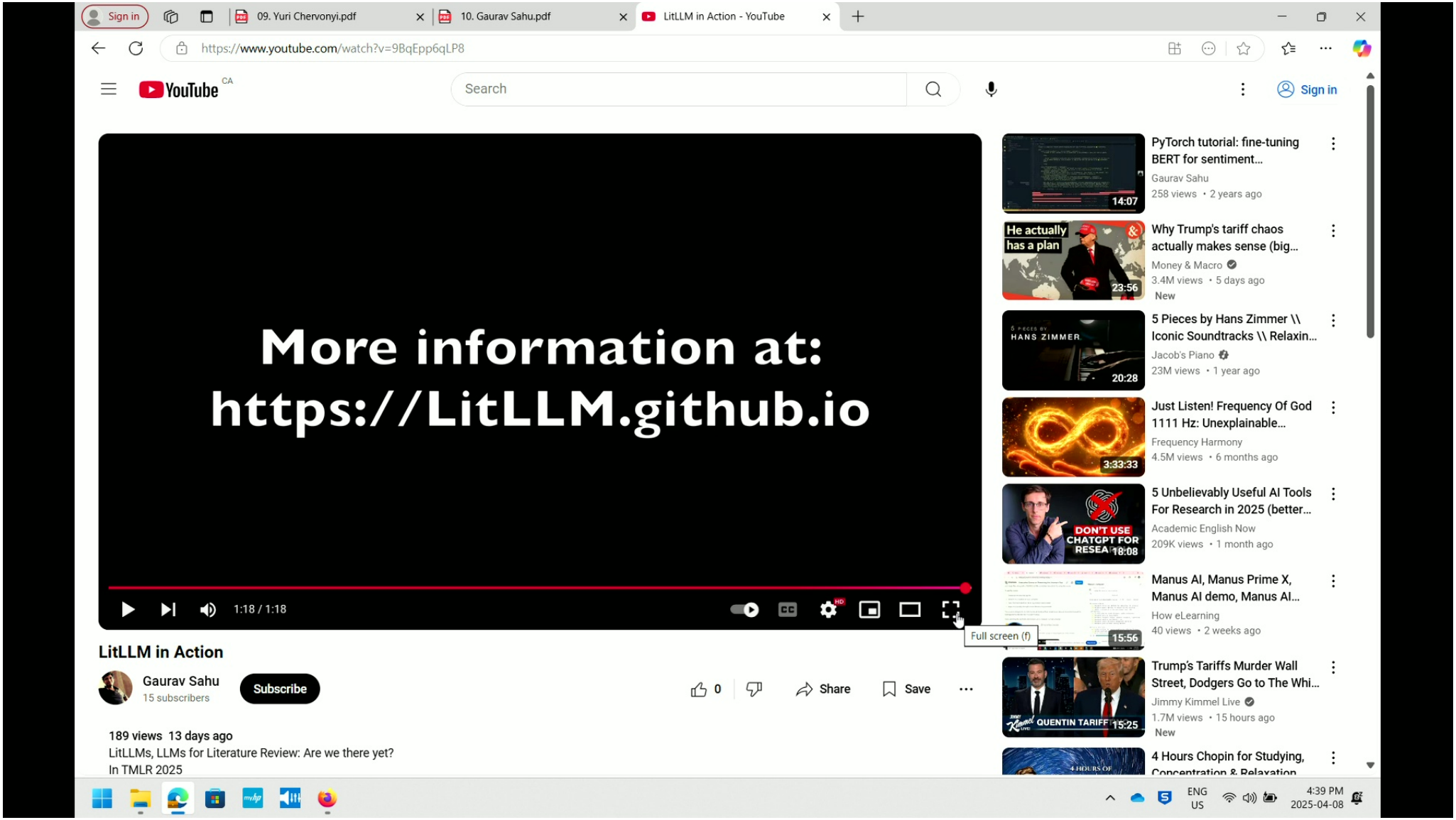


Generation module


LitLLM in Action
youtube.com - To exit full screen, press Esc
ABSTRACT/KEYWORDS
PAPER URL
Abstract
Paste the abstract of your paper here...
Keywords (optional)
machine learning, artificial intelligence, etc.
Machine Learning
Deep Learning
Neural Networks
Transformers
NLP
Climate Change
Renewable Energy
Sustainability
More
1. Enter paper abstract
(we use DeepSeek-v3 in this example)
Search and select pape
0:01 / 1:18
Scroll for details

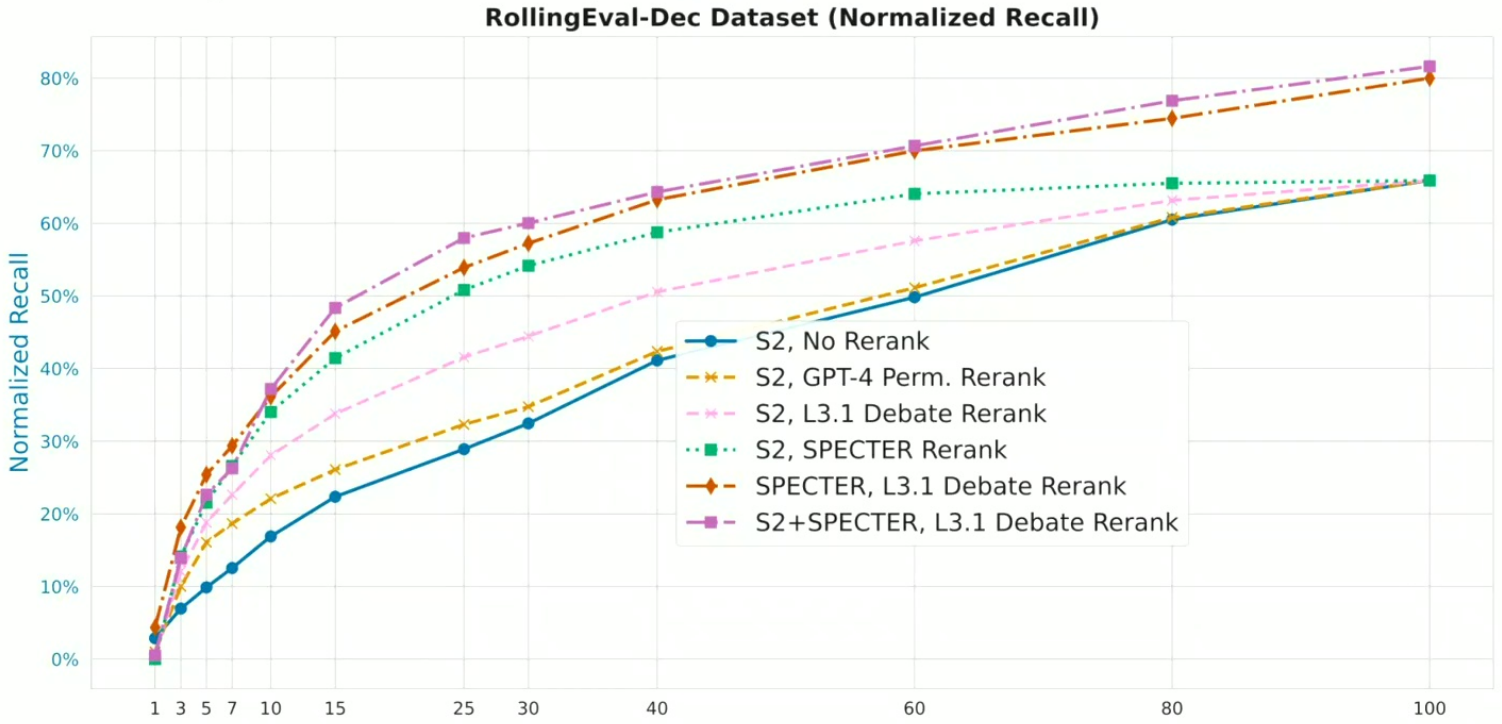

Scientific Literature Review
Sort By
Relevance
Minimum Year
Minimum Citations
Search in Results
Search title or ab
Filter by Fields
Search Results (28 papers)
0 papers selected
Showing 1-10 of 28
DeepSeek-V3 Technical Report
Year: 2024
Citations: 168
Computer Science
Relevance Score: 95/100
High
Abstract & AI Analysis
DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model
Year: 2024
Citations: 303
Computer Science
Relevance Score: 85/100
High
Abstract & AI Analysis
TransMLA: Multi-Head Latent Attention Is All You Need
Year: 2025
Citations: 0
Computer Science
efficient inference and cost-effective training, DeepSeek-V3 adopts Multi-head Latent Attention (MLA) and DeepSeekMoE architectures, which were thoroughly validated in DeepSeek-V2. Furthermore, DeepSeek-V3 pioneers an auxiliary-loss-free strategy for load balancing and sets a multi-token prediction
High-Level Plan (Optional)
Generate a high-level plan that organizes papers into thematic groups. You can edit this plan before generating the final review
GENERATE PLAN
GENERATE REVIEW DIRECTLY
Sentence-Level Plan (Optional)


LitLLM
Scientific Literature Review
Abstract & AI Analysis
TransMLA: Multi-Head Latent Attention Is All You Need
Year: 2025 Citations: 0 Computer Science
Relevance Score: 80/100 High
Abstract & AI Analysis
Union of Experts: Adapting Hierarchical Routing to Equivalently Decomposed Transformer
Year: 2025 Citations: 0 Computer Science
Relevance Score: 75/100 High
Abstract & AI Analysis
Deepseek-v3 technical report
Year: 2024 Citations: 275
Relevance Score: 70/100 High
A Comparison of DeepSeek and Other LLMs
Edit Plan (Organize paper groupings & themes):
implemented in practice and their implications for training.
4. **Follow with Performance Evaluation and Comparative Analysis** to present empirical evidence supporting the discussed methodologies and innovations.
5. **Conclude with Future Directions and Challenges** to provide a forward-looking perspective that wraps up the themes and suggests areas for further research.
This structured flow will ensure that the literature review is coherent, building on each theme logically while covering all referenced papers comprehensively.
GENERATE REVIEW FROM PLAN
GENERATE REVIEW DIRECTLY
Generated Literature Review
Copy
Attention (MLA) and DeepSeekMoE, achieving notable efficiency gains such as a 42.5% reduction in training costs and a 93.3% decrease in KV cache size compared to its predecessor, DeepSeek 67B [1]. However, while MLA demonstrated substantial improvements, the reliance on Group Query Attention (GQA) by many major model providers reflects a
LitLLM provides a relevance score
to each paper based on debate-style reasoning


Sign in
09. Yuri Chervonyi.pdf
10. Gaurav Sahu.pdf
LitLLM in Action - YouTube
https://www.youtube.com/watch?v=9BqEpp6qLP8
YouTube CA
Search
Sign in
More information at:
https://LitLLM.github.io
1:18 / 1:18
Full screen (f)
LitLLM in Action
Gaurav Sahu
15 subscribers
Subscribe
0
Share
Save
189 views 13 days ago
LitLLMs, LLMs for Literature Review: Are we there yet?
In TMLR 2025
PyTorch tutorial: fine-tuning BERT for sentiment...
Gaurav Sahu
258 views • 2 years ago
14:07
Why Trump's tariff chaos actually makes sense (big...
Money & Macro
3.4M views • 5 days ago
New
He actually has a plan
23:56
5 Pieces by Hans Zimmer \\ Iconic Soundtracks \\ Relaxin...
Jacob's Piano
23M views • 1 year ago
20:28
Just Listen! Frequency Of God 1111 Hz: Unexplainable...
Frequency Harmony
4.5M views • 6 months ago
3:33:33
5 Unbelievably Useful AI Tools For Research in 2025 (better...
Academic English Now
209K views • 1 month ago
DON'T USE CHATGPT FOR RESEA
18:08
Manus AI, Manus Prime X, Manus AI demo, Manus AI...
How eLearning
40 views • 2 weeks ago
15:56
Trump's Tariffs Murder Wall Street, Dodgers Go to The Whi...
Jimmy Kimmel Live
1.7M views • 15 hours ago
New
QUENTIN TARIFF
15:25
4 Hours Chopin for Studying, Concentration & Relaxation
ENG
US
4:39 PM
2025-04-08

Evaluation and Findings: Retrieval

$$\text{Normalized Recall@k} = \frac{|\text{Retrieved@k} \cap \text{Ground Truth}|}{|\text{Retrieved} \cap \text{Ground Truth}|}$$



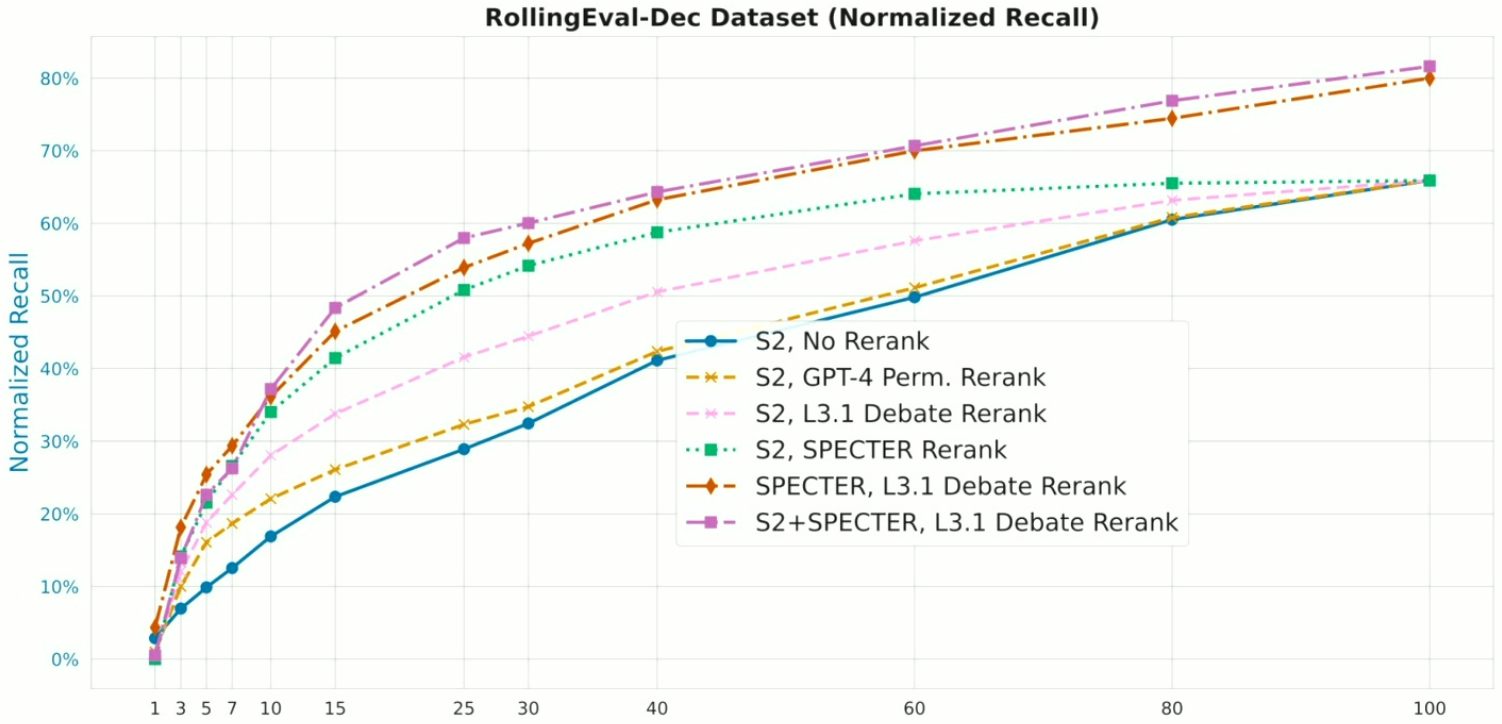
Evaluation and Findings: Retrieval

Search type	RollingEval-Aug (%)	RollingEval-Dec (%)
arXiv API (Single query)	0.65	1.41
SERP API - Google Search (Single query)	1.23	4.34
Semantic Scholar API (Single query)	**3.93**	**4.76**
arXiv API (Multiple queries)	2.75	1.92
SERP API - Google Search (Multiple queries)	**6.80**	5.04
Semantic Scholar API (Multiple queries)	6.07	**5.09**
SPECTER2	8.30	6.80
Semantic Scholar API (Multiple queries) + SPECTER2	**9.80**	**8.20**

Coverage

Evaluation and Findings: Retrieval

$$\text{Normalized Recall@k} = \frac{|\text{Retrieved@k} \cap \text{Ground Truth}|}{|\text{Retrieved} \cap \text{Ground Truth}|}$$

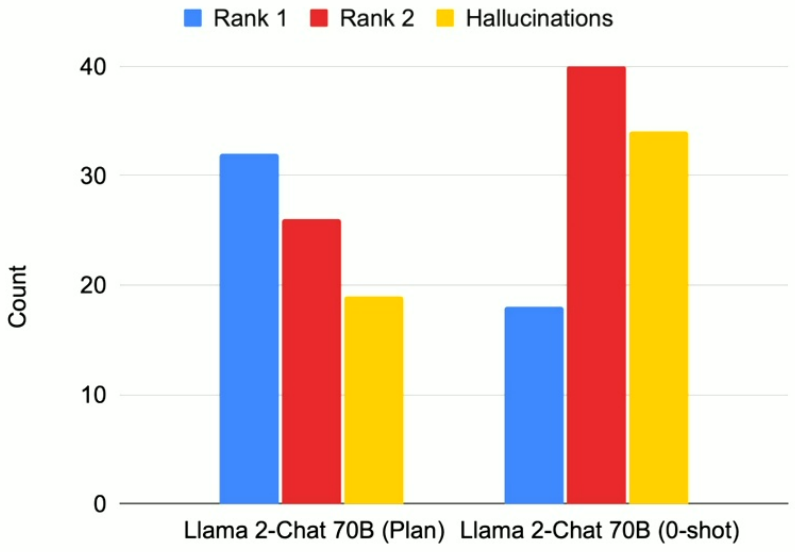


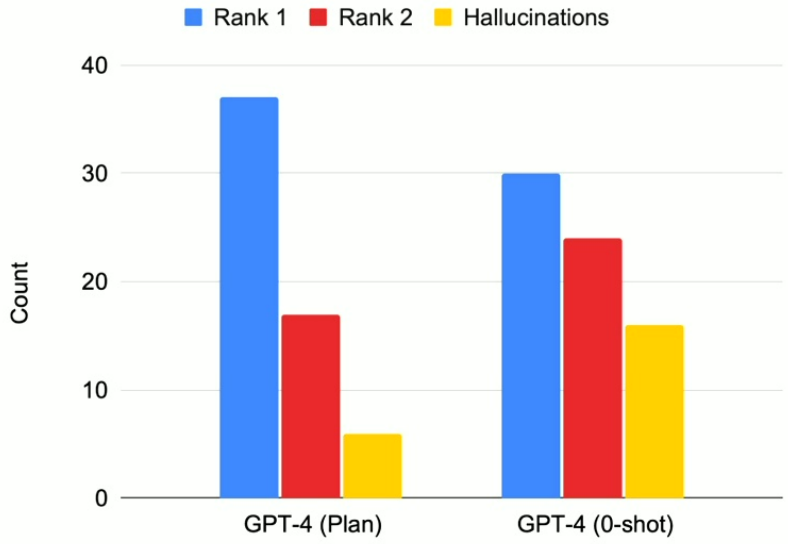
Evaluation and Findings: Generation

Model	ROUGE1 ↑	ROUGE2 ↑	ROUGEL↑	BERTScore↑	Llama-3-Eval↑
CodeLlama 34B-Instruct	22.608	5.032	12.553	82.418	66.898
CodeLlama 34B-Instruct (Plan)	27.369	5.829	14.705	83.386	67.362
Llama 2-Chat 7B	23.276	5.104	12.583	82.841	68.689
Llama 2-Chat 13B	23.998	5.472	12.923	82.855	69.237
Llama 2-Chat 70B	23.769	5.619	12.745	82.943	70.980
GPT-3.5-turbo (0-shot)	25.112	6.118	13.171	83.352	72.434
GPT-4 (0-shot)	29.289	6.479	15.048	84.208	72.951
Llama 2-Chat 70B (Plan)	30.919	7.079	15.991	84.392	71.354
GPT-3.5-turbo (Plan)	30.192	7.028	15.551	84.203	72.729
GPT-4 (Plan)	33.044	7.352	17.624	85.151	75.240

Zero-shot results on the RollingEval-Aug dataset

Evaluation and Findings: Generation





Human Evaluation Study

Key Contributions

- Novel pipeline decomposing literature review writing into LLM-assisted retrieval and plan-based generation
- *Two-step LLM-based retrieval strategy* combining keyword extraction and external search, enhanced by *re-ranking with attribution* that improves relevance
- *Plan-based generation approach* that significantly reduces hallucinations and offers more user control over the output
- New rolling evaluation protocol to avoid test set contamination

Are we there yet?

- *Not quite – but we are getting close*

Thank You!

LitLLM.onrender.com

Tutorial

X: @dem_fier
gaurav.sahu@mila.quebec
Website: LitLLM.github.io

Paper