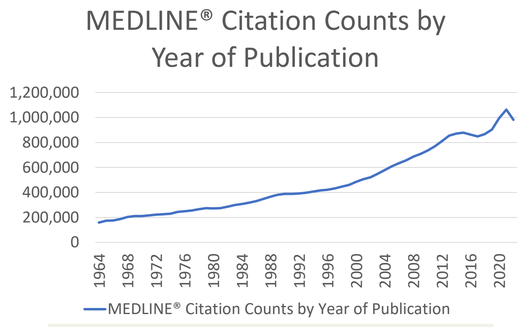
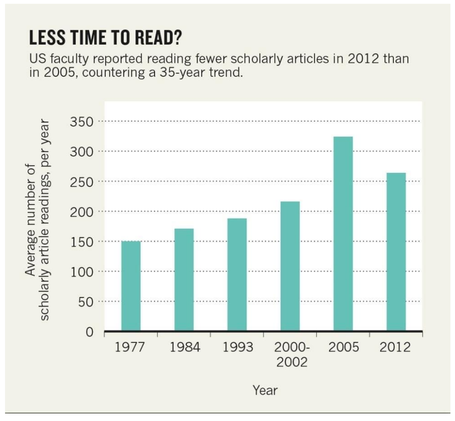
SciMON 🧪 : Scientific Inspiration Machines Optimized for Novelty

Qingyun Wang[1], Doug Downey[2], Heng Ji[2], Tom Hope[2,3]
[1]University of Illinois at Urbana-Champaign, [2]Allen Institute for Artificial Intelligence (AI2)
[3]The Hebrew University of Jerusalem

➢Background

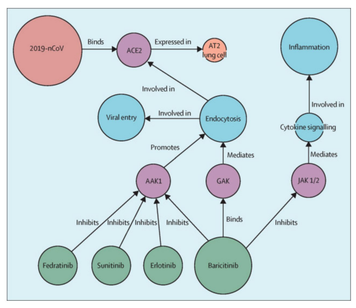
- Millions of scientific papers are published every year.
- Human reading ability keeps almost the same over the years.





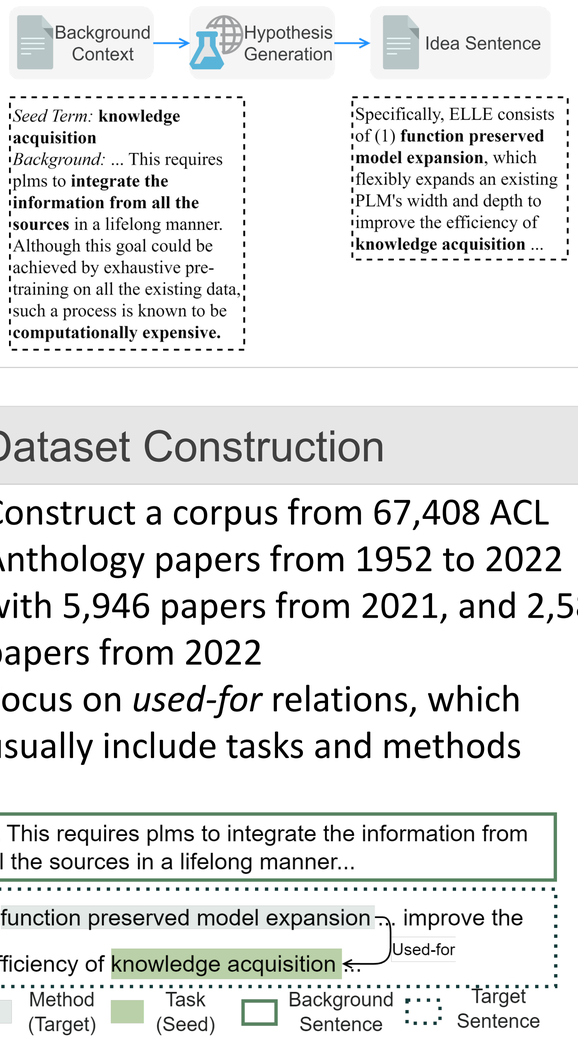
➢Challenges

- Literature-based Discovery
  - Limited to curated entities and relations
  - Limited to certain domains
  - Cannot model nuanced contexts
- LLMs for Scientific Innovation
  - Limited to code generation/experiment planning
  - Focusing on anecdotal evaluation



➢Problem Setting

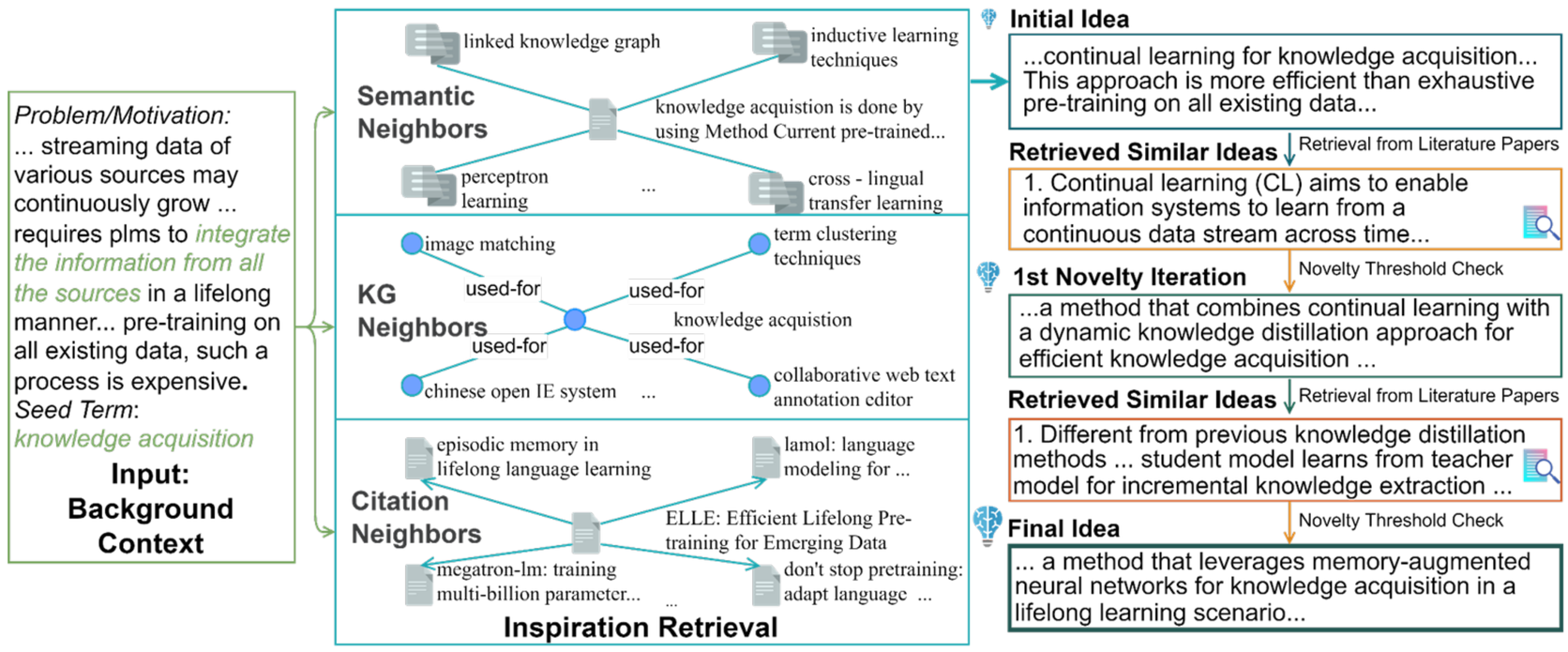
- Contextualized Literature-based Discovery



➢Dataset Construction

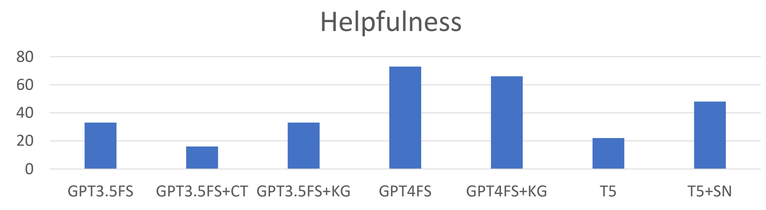
- Construct a corpus from 67,408 ACL Anthology papers from 1952 to 2022 with 5,946 papers from 2021, and 2,588 papers from 2022
- Focus on *used-for* relations, which usually include tasks and methods

➢Approach Overview



➢ Comparison across Outputs

- Comparing Outputs across Model Variants
  - GPT4FS and GPT4FS+KG outperform other models by a wide margin
  - GPT4 outputs tended to be longer, which may partially explain higher human preference
- Comparisons to Real Papers
  - 48% GPT4FS+KG shows higher technical detail
  - 45% GPT4FS+KG is more novel
  - Original ground truth ideas have significantly higher technical level and novelty in 85% of comparisons



➢Example

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**.

➢Evaluation on Iterative Novelty Boosting

- Compare the novelty-enhanced results against the previous generated ideas
- Examine the new terms added after filtering stopwords and generic words
- Ideas after novelty iterations are longer than initial ideas
- For ideas not considered more novel after applying our method, we do not observe a drop in novelty: the method either increases or maintains novelty

Type	GPT4FS	+SN	+CT	+KG
1st Novelty Δ (%)	+54.4	+**55.6**	+47.8	+46.7
2nd Novelty Δ (%)	-	+57.8	-	-
1st new terms Δ	+**23.1**	+22.8	+22.1	+21.9
2nd new terms Δ	-	+21.5	-	-

➢Case Study: Biomedical

- Collect a dataset from PubMed papers from 1988 to 2024 and construct a biomedical dataset, including 4,767 papers before 2023/02, 642 papers from 2023/02 to 2023/08, and 299 papers after 2023/08

Type	Meditron	+SN	+CT	+KG
Helpful (%)	35	80	60	50
vs. GT (%)	30	45	50	35

Input	*seed term*: **ARO10;** *context* (abridged): Huangjiu is known for its unique aroma, primarily attributed to its high concentration of b-phenylethanol (ranging from 40 to 130 mg/L).
Meditron	We then selected six key genes involved in the phenylalanine metabolism pathway and predicted that their product flux may affect the synthesis of b-phenylethanol.
+ CT	We found that the key amino acid residue that controls the activity of **Aro10p** was not conserved in wine yeast strains, which may explain the lower b-phenylethanol production in wine fermentation compared with that in Chinese huangjiu.
+ KG	Both target genes, SSA1 and **ARO10**, were deleted using the CRISPR-Cas9 genome editing system.

➢Dataset and Code

https://github.com/EagleW/Scientific-Inspiration-Machines-Optimized-for-Novelty