

SciMON /> : Scientific Inspiration Machines Optimized for Novelty

Qingyun Wang¹, Doug Downey², Heng Ji², Tom Hope^{2,3}

¹University of Illinois at Urbana-Champaign, ²Allen Institute for Artificial Intelligence (AI2)

³The Hebrew University of Jerusalem

>Background

- Millions of scientific papers are published every year.
- Human reading ability keeps almost the same over the years. MEDLINE[®] Citation Counts by

Year of Publication



> Approach Overview



...continual learning for knowledge acquisition... This approach is more efficient than exhaustive

pre-training on all existing data... Retrieved Similar Ideas Vertieval from Literature Papers 1. Continual learning (CL) aims to enable information systems to learn from a Q continuous data stream across time ... Novelty Threshold Check **1st Novelty Iteration** ...a method that combines continual learning with a dynamic knowledge distillation approach for efficient knowledge acquisition ... Retrieved Similar Ideas V Retrieval from Literature Papers . Different from previous knowledge distillation



➤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

		methods student model learns from teacher of model for incremental knowledge extraction				
	∰	Final Idea	Novelty Threshold Check			
ning:		a method that leverages memory-augmented neural networks for knowledge acquisition in a lifelong learning scenario				

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

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

Туре	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	-	-



➢ Problem Setting

 Contextualized Literature-based Discovery

Background Context 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 pretraining 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 ...

Idea Sentence

Dataset Construction



➢ Example

seed term: speech unit boundaries; context Input (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

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

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

- seed term: ARO10; context (abridged): Huangjiu is Input 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

- 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

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



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...

... an efficient monotonic segmentation Ground module ... accumulate acoustic information Truth incrementally and detect proper **speech unit** boundaries.

that their product flux may affect the synthesis of b-phenylethanol.

- We found that the key amino acid residue that + CT 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.
- Both target genes, SSA1 and **ARO10**, were deleted + KG using the CRISPR-Cas9 genome editing system.

Dataset and Code

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

