

Leverage GenAI and Causal Inference to Disrupt Innovation

Yong Zhang, Kelly Anderson

P&G-CF-R&D-DIP-DT-RTIC/DSAI

April 12, 2024



D&IP

Discovery & Innovation Platforms

2024 Bayesialab Spring Conference, Cincinnati

Outline

- GenAI Powered AI-BBN for Causal Inference
- GenAI Powered ScienceSage for Causal Inference

What is AI-BBN?

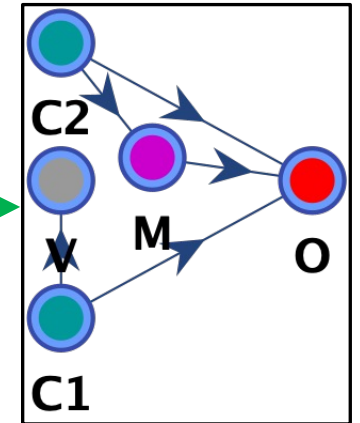
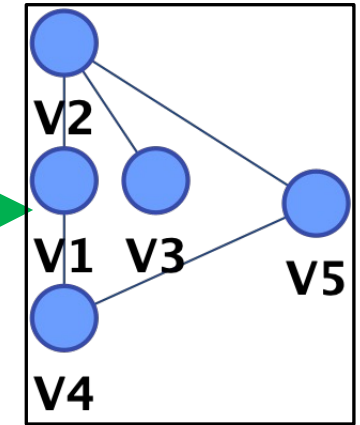
Observational Data



AI-BBN GenAI Powered Causal Inference



Regular BBN



Causal BBN



Domain Experts

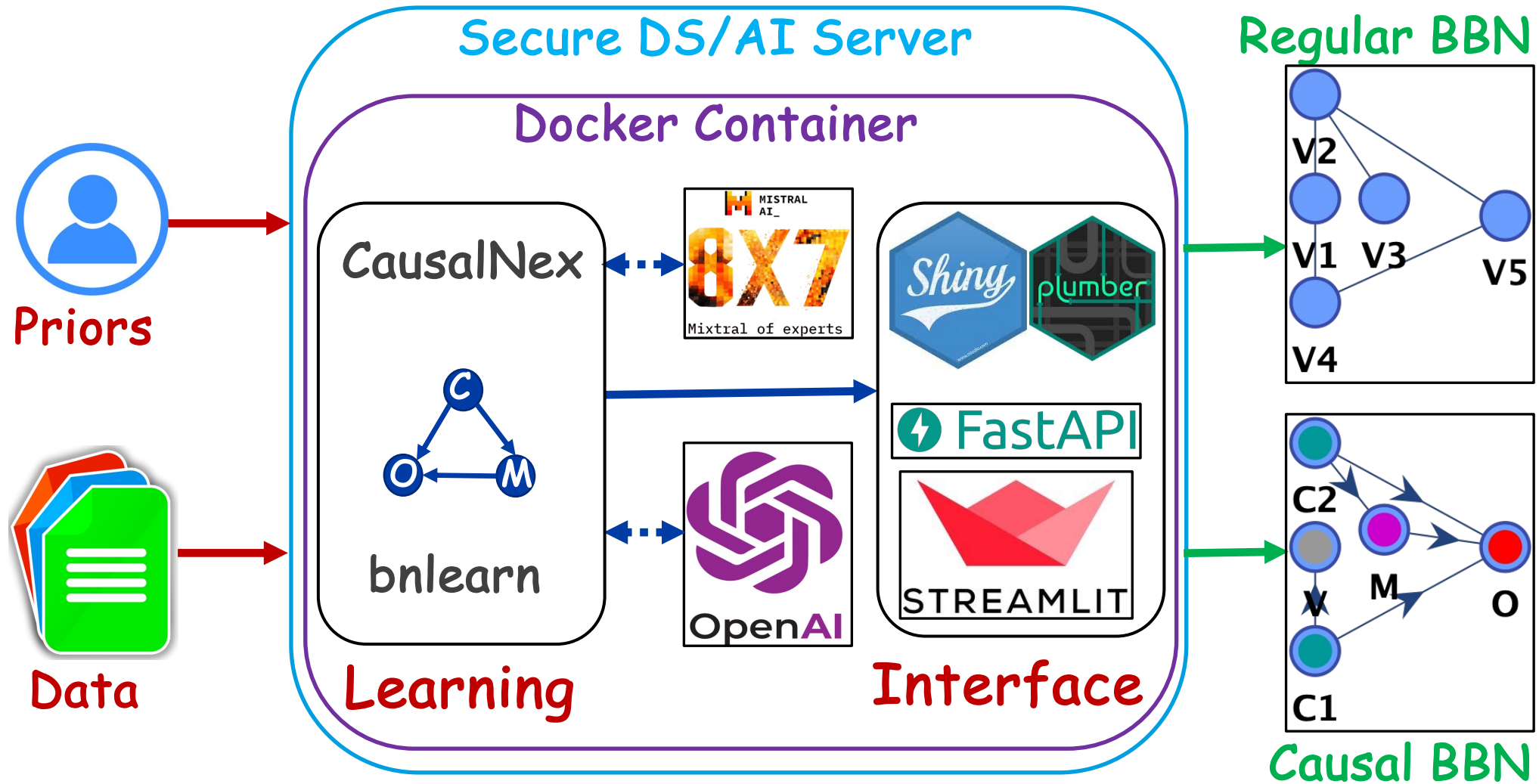
Disrupt Speed & Economics

- 100X faster

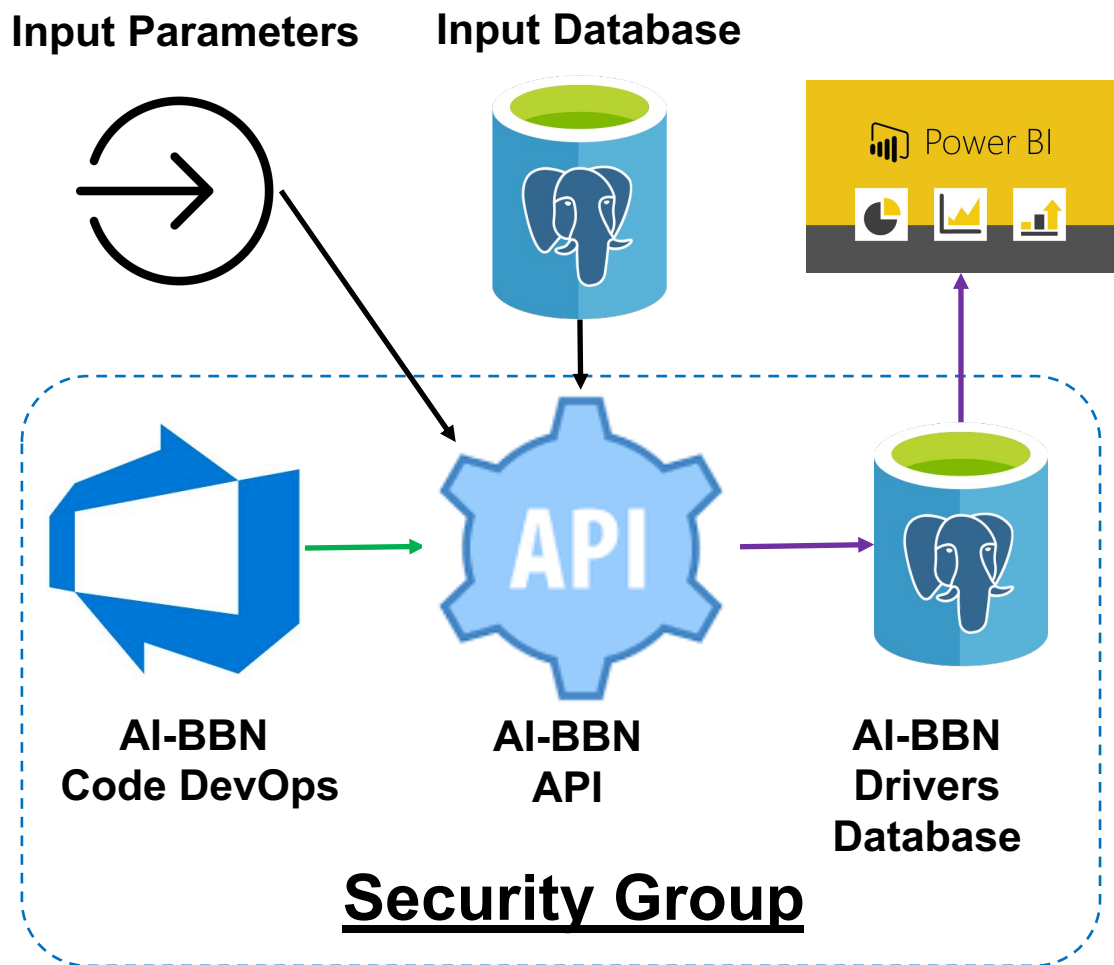
Features & Advantages

- Use LLM as an 'expert'
- Iteratively refine the learned BBN
- Leverage tabular and text data
- Output regular & Causal BBN

Architecture of AI-BBN API/APP



Example: API Interface



AI-BBN API With Database Connection 1.0.0 OAS3

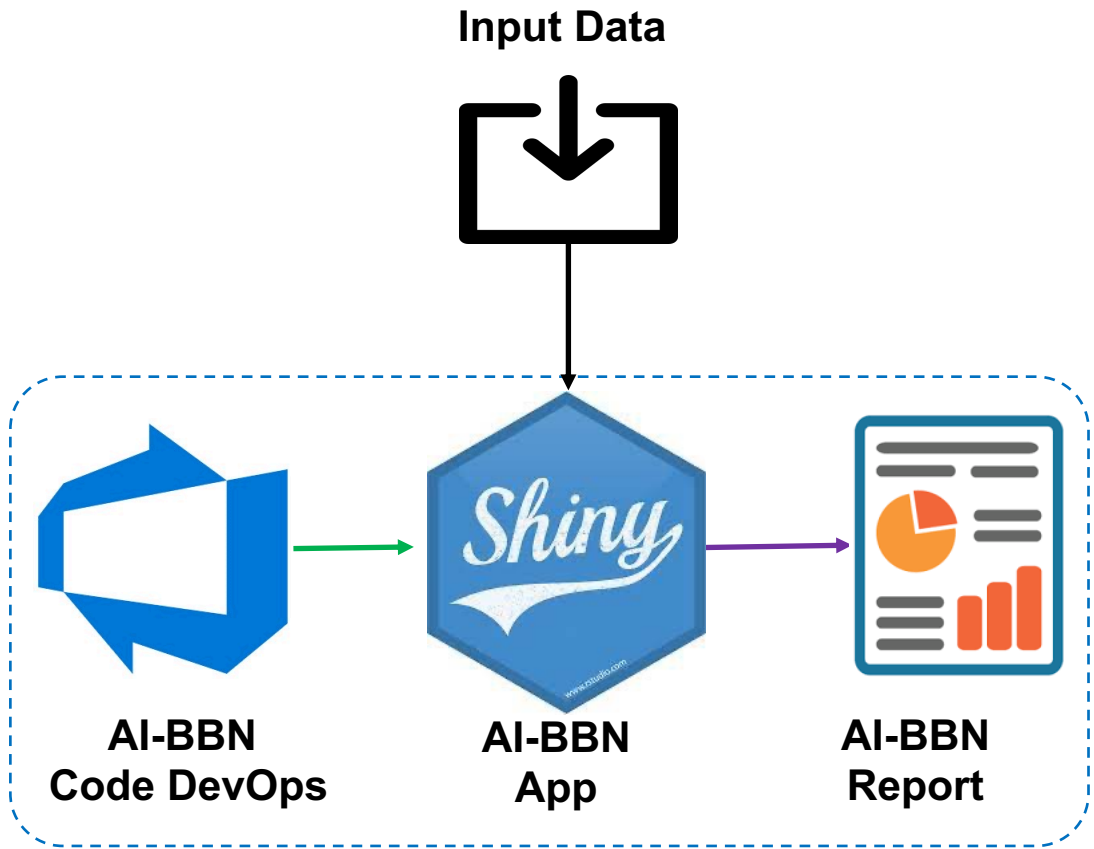
default

GET `/aibbn` Grab data from STC and build AI-BBN

Parameters

Name	Description
Region * required string (query)	Region/Market: e.g. NORTH AMERICA Region - Region/Market: e.g. NORTH AMERI
Segment * required string (query)	Product segment: e.g. Shampoo Segment - Product segment: e.g. Shampoo
Target * required string (query)	Name of Target: e.g. eRR_ReviewRating Target - Name of Target: e.g. eRR_ReviewRc
topicLevel * required string (query)	Level of Topic: e.g. Vector topicLevel - Level of Topic: e.g. Vector
BeginDate * required string (query)	Beginning date: e.g. 2020-01-01 BeginDate - Beginning date: e.g. 2020-01-01
EndDate * required string (query)	Ending date: e.g. 2021-12-31 EndDate - Ending date: e.g. 2021-12-31

Example: APP Interface



AI-BBN Web APP

QuickStart

Data Exploration

BBN Model & Drivers

Landscape & Profile Analyses

Influence/Impact Analysis

AI-BBN App is used to build BBN model and conduct driver and influence/impact analyses automatically, at large scale on large volume of data. This web App is to help users to build a BBN model and conducts drivers in a product category. It then calculate influences/impacts of product attributes/benefits on the target for an individual product. It also provides landscape analysis to understand what position each product is at and relative to each other and market average. You can also compare two products on multiple benefits using profile analysis. If you are interested in leveraging AI-BBN capability in your platform (e.g., App and Web Service), we also provide API services (See details below). The web App is consist of 4 different tabs as described below.

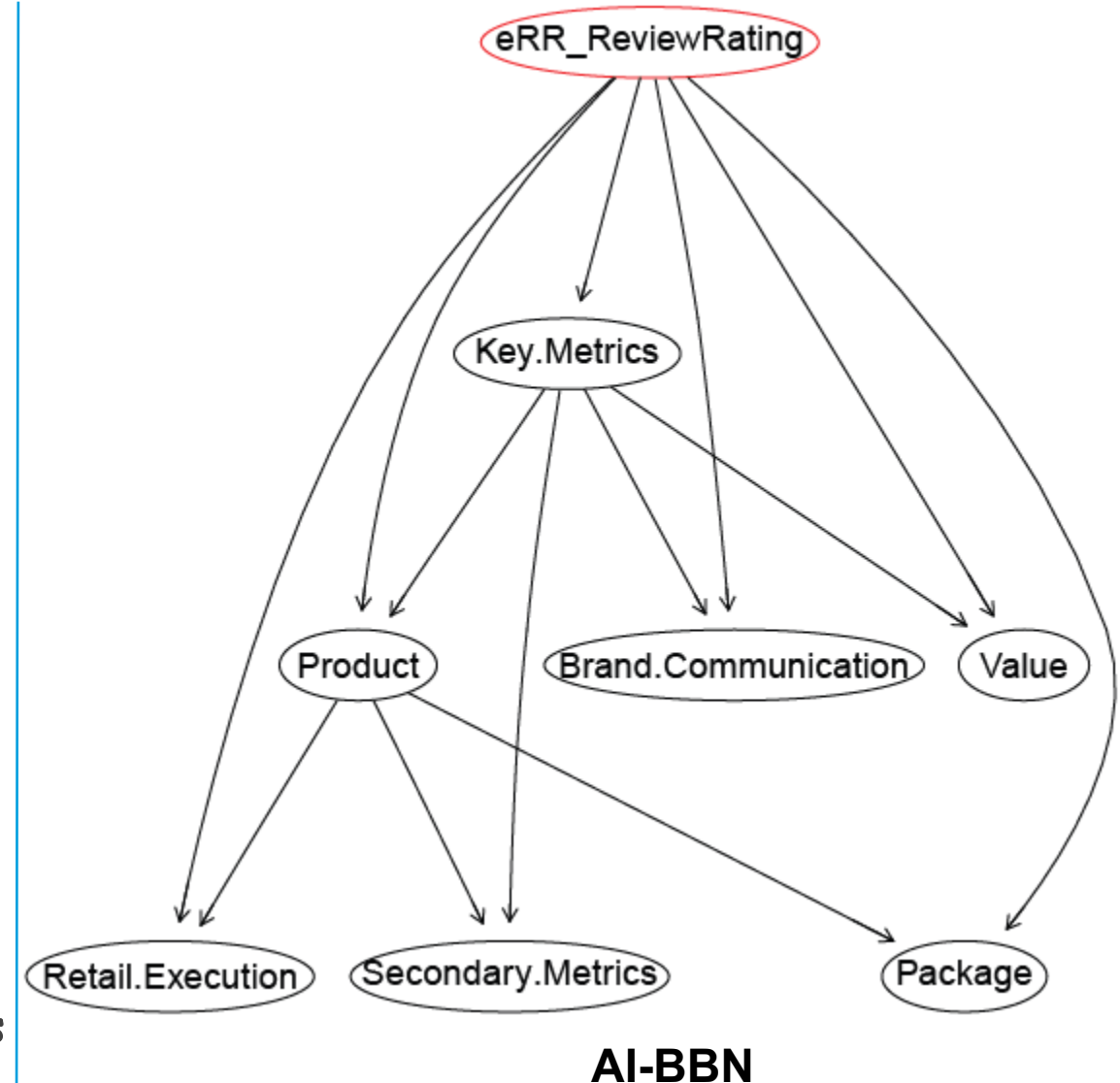
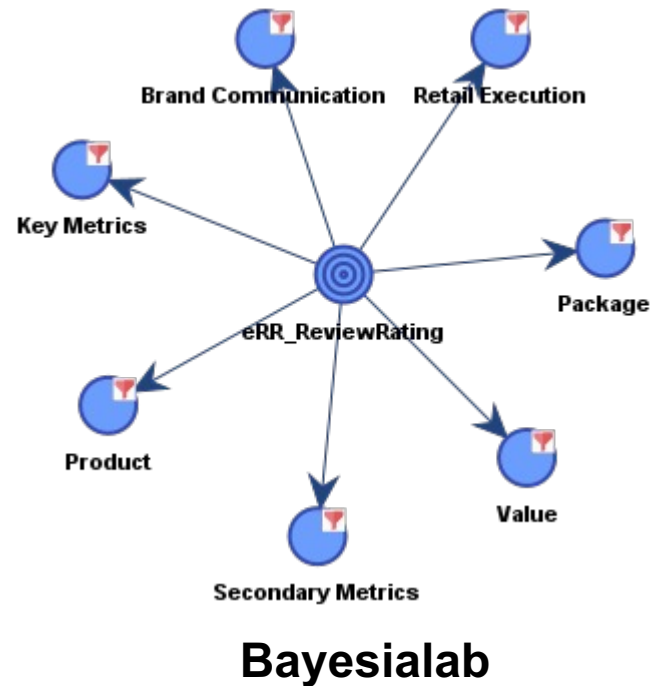
Data Exploration

This tab explores the input data and put the data into a format for conducting influence/impact, landscape and profile analyses.

BBN Models & Drivers

This tab builds BBN model and conduct driver analyses for the whole product category in the input data.

AI-BBN Performs Well on Data with **High Sparsity**



- Automatic driver analysis at scale on large volume of data (100x faster)
- Benchmark on multiple datasets shows that the AI-BBN is correct, consistent, and robust.
- AI-BBN also tends to identify and reveal fine structures

LLM as 'Expert' to Build Causal BBN Model

Iterative LLM Supervised Causal Structure Learning (ILS-CSL)

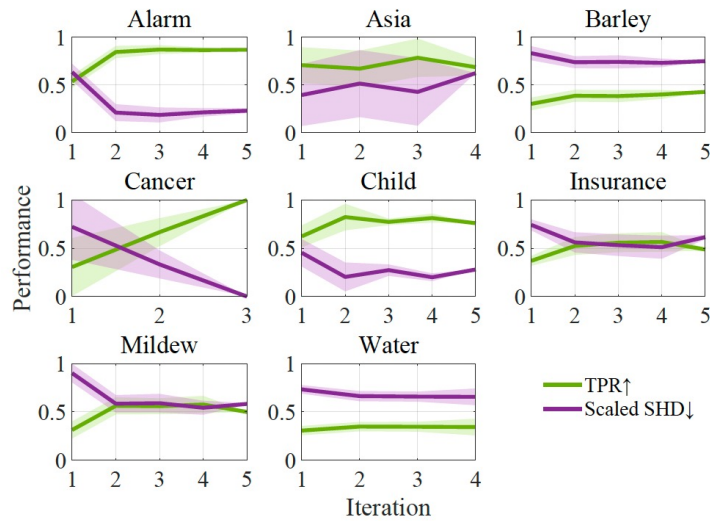


Fig. 3: Trend of TPR↑ (green line) and scaled SHD↓ (purple line) of HC+BIC+ILS-CSL-hard on various datasets.

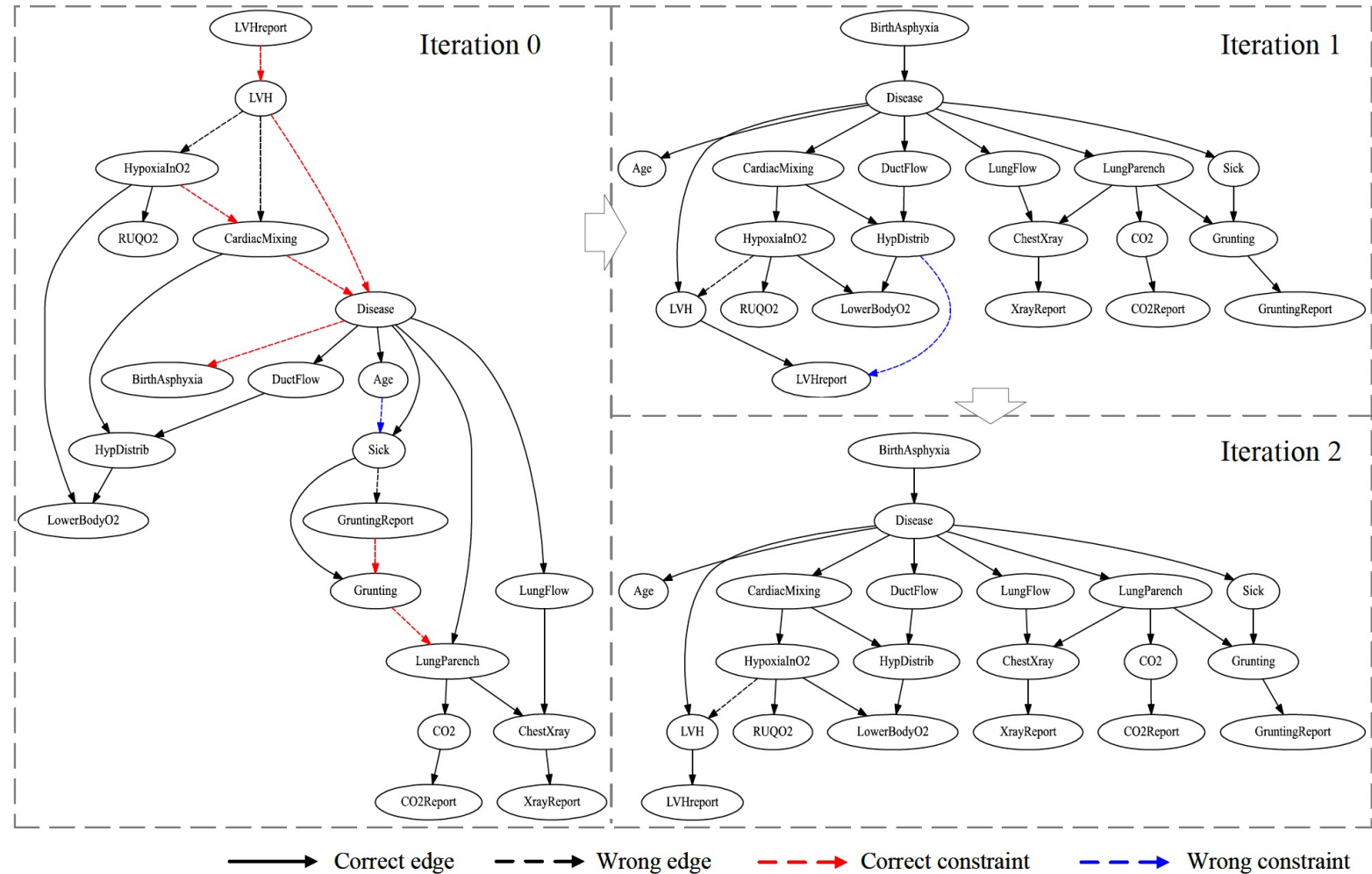


Fig. 4: Visualized process of HC-BDeu+ILS-CSL-hard on a set of observed data of *Child*, 2000 samples. The SHD of iterations are: 12 for Iteration 0, 3 for Iterations 1 and 2.

LLM Substantially Improves Causal BBN Model Business Use

TABLE VII: Scaled SHD↓ enhancement on data-based CSL with different scores, search algorithms and approaches to apply prior constraints, by the proposed framework.

Dataset N	Cancer		Asia		Child		Insurance	
	250	1000	250	1000	500	2000	500	2000
HC-BDeu	0.58±0.13	0.33±0.26	0.56±0.27	0.23±0.17	0.57±0.12	0.49±0.18	0.69±0.06	0.68±0.09
+ILS-CSL-hard	0.50±0.22 ^{-14%}	0.29±0.29 ^{-12%}	0.46±0.33 ^{-18%}	0.15±0.15 ^{-35%}	0.24±0.07 ^{-58%}	0.10±0.02 ^{-80%}	0.45±0.06 ^{-35%}	0.34±0.04 ^{-50%}
+ILS-CSL-soft	0.50±0.22 ^{-14%}	0.29±0.29 ^{-12%}	0.44±0.30 ^{-21%}	0.15±0.15 ^{-35%}	0.26±0.06 ^{-54%}	0.11±0.03 ^{-78%}	0.50±0.08 ^{-28%}	0.35±0.04 ^{-49%}
MINOBSx-BDeu	0.75±0.22	0.46±0.29	0.52±0.32	0.31±0.07	0.38±0.08	0.21±0.04	0.46±0.05	0.29±0.02
+ILS-CSL-hard	0.50±0.22 ^{-33%}	0.29±0.29 ^{-37%}	0.42±0.37 ^{-19%}	0.15±0.15 ^{-52%}	0.25±0.06 ^{-34%}	0.07±0.03 ^{-67%}	0.42±0.03 ^{-9%}	0.28±0.06 ^{-3%}
+ILS-CSL-soft	0.50±0.22 ^{-33%}	0.29±0.29 ^{-37%}	0.42±0.37 ^{-19%}	0.15±0.15 ^{-52%}	0.25±0.04 ^{-34%}	0.08±0.04 ^{-62%}	0.41±0.03 ^{-11%}	0.26±0.04 ^{-10%}
HC-BIC	0.92±0.29	0.62±0.34	0.48±0.36	0.31±0.29	0.53±0.07	0.38±0.16	0.76±0.05	0.72±0.06
+ILS-CSL-hard	0.92±0.29 ^{+0%}	0.42±0.34 ^{-32%}	0.33±0.25 ^{-31%}	0.19±0.17 ^{-39%}	0.26±0.07 ^{-51%}	0.07±0.03 ^{-82%}	0.60±0.03 ^{-21%}	0.41±0.03 ^{-43%}
+ILS-CSL-soft	0.92±0.29 ^{+0%}	0.42±0.34 ^{-32%}	0.35±0.26 ^{-27%}	0.21±0.19 ^{-32%}	0.27±0.08 ^{-49%}	0.07±0.05 ^{-82%}	0.62±0.06 ^{-18%}	0.42±0.03 ^{-42%}
MINOBSx-BIC	1.00±0.25	0.62±0.21	0.46±0.23	0.27±0.05	0.34±0.06	0.18±0.04	0.62±0.05	0.55±0.05
+ILS-CSL-hard	0.92±0.29 ^{-8%}	0.38±0.26 ^{-39%}	0.42±0.40 ^{-9%}	0.12±0.08 ^{-56%}	0.24±0.08 ^{-29%}	0.06±0.02 ^{-67%}	0.55±0.03 ^{-11%}	0.39±0.08 ^{-29%}
+ILS-CSL-soft	0.92±0.29 ^{-8%}	0.38±0.26 ^{-39%}	0.35±0.26 ^{-24%}	0.15±0.12 ^{-44%}	0.25±0.05 ^{-26%}	0.06±0.02 ^{-67%}	0.55±0.03 ^{-11%}	0.41±0.09 ^{-25%}

Dataset N	Alarm		Mildew		Water		Barley	
	1000	4000	8000	32000	1000	4000	2000	8000
HC-BDeu	0.65±0.12	0.64±0.09	0.79±0.11	0.99±0.07	0.76±0.07	0.64±0.08	0.80±0.06	0.65±0.06
+ILS-CSL-hard	0.12±0.02 ^{-82%}	0.08±0.01 ^{-88%}	0.46±0.01 ^{-42%}	0.22±0.02 ^{-78%}	0.64±0.02 ^{-16%}	0.55±0.03 ^{-14%}	0.69±0.06 ^{-14%}	0.57±0.06 ^{-12%}
+ILS-CSL-soft	0.30±0.05 ^{-54%}	0.25±0.06 ^{-61%}	0.43±0.00 ^{-46%}	0.47±0.04 ^{-53%}	0.64±0.01 ^{-16%}	0.56±0.03 ^{-12%}	0.76±0.04 ^{-5%}	0.62±0.03 ^{-5%}
MINOBSx-BDeu	0.21±0.06	0.14±0.04	0.50±0.02	0.46±0.05	0.77±0.07	0.61±0.04	0.56±0.04	0.40±0.03
+ILS-CSL-hard	0.09±0.03 ^{-57%}	0.08±0.02 ^{-43%}	0.43±0.00 ^{-14%}	0.33±0.18 ^{-28%}	0.68±0.05 ^{-12%}	0.56±0.02 ^{-8%}	0.54±0.02 ^{-4%}	0.38±0.02 ^{-5%}
+ILS-CSL-soft	0.09±0.02 ^{-57%}	0.07±0.01 ^{-50%}	0.47±0.01 ^{-6%}	0.37±0.02 ^{-20%}	0.68±0.04 ^{-12%}	0.56±0.02 ^{-8%}	0.55±0.03 ^{-2%}	0.38±0.02 ^{-5%}
HC-BIC	0.68±0.05	0.59±0.10	0.90±0.06	0.91±0.13	0.76±0.04	0.70±0.03	0.87±0.05	0.80±0.08
+ILS-CSL-hard	0.22±0.04 ^{-68%}	0.12±0.04 ^{-80%}	0.58±0.01 ^{-36%}	0.46±0.04 ^{-49%}	0.69±0.02 ^{-9%}	0.61±0.03 ^{-13%}	0.76±0.02 ^{-13%}	0.69±0.06 ^{-14%}
+ILS-CSL-soft	0.41±0.04 ^{-40%}	0.35±0.11 ^{-41%}	0.71±0.01 ^{-21%}	0.57±0.02 ^{-37%}	0.69±0.02 ^{-9%}	0.61±0.03 ^{-13%}	0.82±0.04 ^{-6%}	0.74±0.09 ^{-8%}
MINOBSx-BIC	0.32±0.08	0.15±0.04	0.74±0.01	0.73±0.09	0.82±0.03	0.77±0.03	0.79±0.04	0.58±0.03
+ILS-CSL-hard	0.16±0.07 ^{-50%}	0.09±0.03 ^{-40%}	0.58±0.01 ^{-22%}	0.45±0.03 ^{-38%}	0.69±0.03 ^{-16%}	0.62±0.01 ^{-19%}	0.73±0.03 ^{-8%}	0.55±0.03 ^{-5%}
+ILS-CSL-soft	0.19±0.06 ^{-41%}	0.10±0.01 ^{-33%}	0.73±0.01 ^{-1%}	0.64±0.04 ^{-12%}	0.70±0.02 ^{-15%}	0.64±0.02 ^{-17%}	0.76±0.02 ^{-4%}	0.56±0.03 ^{-3%}

AI-BBN Summary

- **AI-BBN is not meant to replace Bayesialab!** It helps conduct automatic driver analysis at scale on large volume of data with **high sparsity**.
- **LLM can substantially improve quality of Causal BBN** in just few iterations. It opens a new frontier of causal inference using GenAI and PGM.

What is ScienceSage?

P&G Docs & Question

P&G Answers

ScienceSage
GenAI Powered
Research Assistant



Research Report



Research Question

Disrupt Speed & Economics

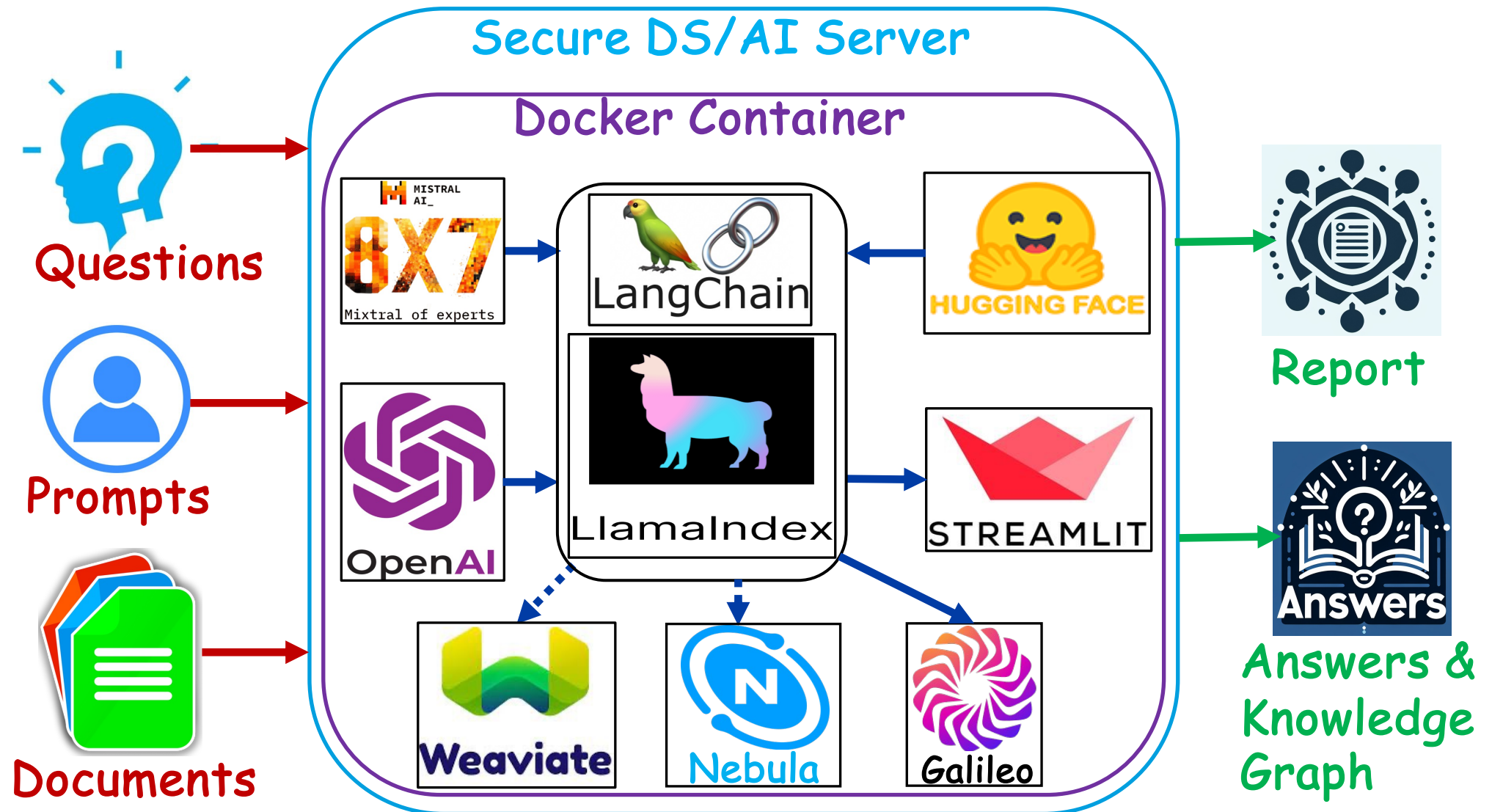
- 100X faster

Features & Advantages

- Latest information
- Comprehensive report
- Data security
- Avoid Microsoft filtering
- Chat with documents
- Support multiple files & format
- Support multiple LLMs

Architecture of ScienceSage Web App

Business Use



ScienceSage User Interface

Business Use



ScienceSage - GenAI Powered Research Assistant to Extract Insight from Voice of Science

Use GenAI to Generate a Research Report

Please type in your research question:

What is the impact of climate change on airborne allergens?

Use Next Three Template Boxes from Generic Research Report to Customize Your Instructions

Scroll down as needed inside each box to see and edit the whole prompt. Do not change the variables in curly bracket, namely {text}, {question}, {research_summary}; please also keep the split/division lines

☐ Quick Start Menu

☒ Generate Research Report

☐ Chat With Your Documents

Select Search Method:

searchArxiv

Number of Queries to Decompose Your Question:

3

26

Top N Records Per Query:

3

220

Select a Report Template:

Generic Research Rep...

Select a LLM:

AzureChatGPT

generateReport



Use GenAI to Chat with Your Documents

Upload Your Files

Choose input files



Drag and drop files here

Limit 200MB per file • CSV, TXT, DOCX, XLSX, PDF

Browse files



Zhang-etal_SLR-AI-Summarizer_WebApp_Llama2Update_Nov2023.pdf 363.0KB



Zhang-etal_SLR-AI-BBN_Influence-LandScape-Impact_12122022.docx 0.8MB



Ask your question



Select a Tab

- ☐ Quick Start Menu
- ☐ Generate Research Report
- ☒ Chat With Your Documents

Select a LLM:

Mixtral_8X7B

Select a Usage Scenario:

Use Vector

ScienceSage Outputs a Report & References

Question: What is the MoE architecture in GenAI models?

Select a Tab

☐ Quick Start Menu
 ☒ Generate Research Report
 ☐ Chat With Your Documents

Select Search Method:

searchArxiv

Number of Queries to Decompose Your Question:

3

26

Top N Records Per Query:

3

220

Select a Report Template:

Generic Research Report

Select a LLM:

AzureChatGPT

generateReport

Mixture of Experts Architecture in Generative AI Models: An In-depth Analysis

Generative Artificial Intelligence (GenAI) models have been at the forefront of AI research, with their ability to generate new data that resembles the input data. These models have found applications in various fields, from natural language processing to image synthesis. However, the architecture of these models plays a crucial role in their performance and scalability. One such architecture that has gained attention is the Mixture of Experts (MoE) architecture. This report aims to provide an in-depth analysis of the MoE architecture in GenAI models, its benefits, drawbacks, and its role in advancing the field of AI.

Understanding MoE Architecture

The MoE architecture is a machine learning model that consists of multiple 'expert' models, each specializing in different parts of the input space. Each expert is responsible for a subset of the data, and the final output is a weighted combination of the outputs of all experts. This architecture allows the model to scale and handle complex tasks by dividing them into smaller, manageable parts.

Despite their success, MoEs have several issues such as training instability, token dropping, inability to scale the number of experts, and ineffective finetuning. To address these challenges, a new architecture called Soft MoE, a fully-differentiable sparse Transformer, has been proposed. Soft MoE performs an implicit soft assignment by passing different weighted combinations of all input tokens to each expert. It only processes a subset of the combined tokens, enabling larger model capacity at lower inference cost. In visual recognition, Soft MoE outperforms standard Transformers (ViTs) and popular MoE variants. For instance, Soft MoE-Base/16 requires 10.5x lower inference cost (5.7x lower wall-clock time) than ViT-Huge/14 while matching its performance after similar training. Soft MoE Huge/14 with 128 experts in 16 MoE layers has over 40x more parameters than ViT Huge/14, while inference time cost grows by only 2%, and it performs substantially better.

MoE in Large Language Models

MoE models are a popular choice for large language models due to their scalability and sublinear computational complexity for training and inference. However, the existing MoE models have two main drawbacks. Firstly, they introduce a significant amount of inner-node and inter-node communication overhead due to all-to-all dispatching and gathering. Secondly, they have limited scalability for the backbone due to the bound data parallel and expert parallel to scale in the expert dimension.

To address these issues, a novel MoE architecture called Pipeline MoE (PPMoE) has been proposed. This new architecture incorporates expert parallel with tensor parallel and replaces the communication-intensive all-to-all dispatching and gathering with a simple tensor index slicing and inner-node all-reduce. PPMoE can also easily integrate pipeline parallel to further scale the backbone due to its flexible parallel architecture.

The authors conducted extensive experiments and found that PPMoE not only achieves a speed up of more than 1.75 times compared to existing MoE architectures, but also reaches 90% throughput of its corresponding backbone model that is 20 times smaller.

MoE in GenAI: Opportunities and Challenges

While the MoE architecture has shown promise in large language models and visual recognition tasks, its application in GenAI models is still an open area of research. The construction industry, for instance, has been slow in adopting AI, and there is a lack of studies investigating the opportunities and challenges of implementing GenAI with MoE architecture in this sector.

The potential of large language models (LLM) like OpenAI's GPT, Google's PaLM, and Meta's Llama has been highlighted, and the prospects and complexities of GenAI integration in various industries are being explored. However, the specific benefits and drawbacks of MoE architecture in GenAI models are yet to be fully understood and documented.


Conclusion


In conclusion, the MoE architecture offers a promising approach to handle the scalability and complexity issues in AI models, including GenAI models. The introduction of Soft MoE and PPMoE architectures has shown significant improvements in performance and inference cost. However, more research is needed to fully understand the potential of MoE architecture in GenAI models and to address the existing challenges. As the field of AI continues to evolve, the MoE architecture is likely to play a crucial role in shaping the future of GenAI models.

References

- From Sparse to Soft Mixtures of Experts. Retrieved from <http://arxiv.org/abs/2304.11414v1>
- Pipeline MoE: A Flexible MoE Implementation with Pipeline Parallelism. Retrieved from <http://arxiv.org/abs/2304.11414v1>
- Generative AI in the Construction Industry: Opportunities & Challenges. Retrieved from <http://arxiv.org/abs/2304.11414v1>

ScienceSage Outputs a Knowledge Graph to Provide Causal Inference

 what is NLP?



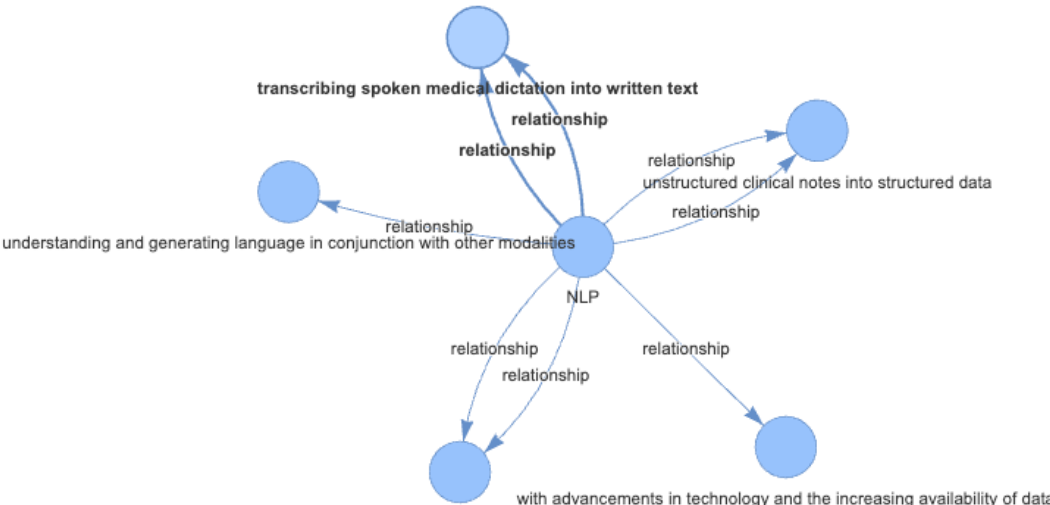
NLP, or Natural Language Processing, is a field that involves several functions. It automates the process of transcribing spoken medical dictation into written text and helps convert unstructured clinical notes into structured data. NLP also enables businesses to automate customer support processes using chatbots. Additionally, it involves understanding and generating language in conjunction with other modalities. NLP continues to grow and evolve with advancements in technology and the increasing availability of data.

Query based entity-relationship-entity

	Subject.entity.name	Predicate.relationship	Object.entity.name
0	NLP	continue to grow and evolve	with advancements in technology and the increasing availability of data
1	NLP	involves	understanding and generating language in conjunction with other modalities
2	NLP	helps in converting	unstructured clinical notes into structured data
3	NLP	enables	businesses to automate customer support processes using chatbots
4	NLP	automates	transcribing spoken medical dictation into written text

[Download Subject-Predicate-Object table based on the query in CSV format](#)

Rendered Graph



ScienceSage Summary

- ScienceSage can generate an answer with a **Knowledge Graph** to assist causal inferences.
- ScienceSage can search internet or scientific paper database based on a question. It can then generate a research **report with references** based on **latest relevant information**.

Questions