## Define Topics:

### Tech Topics:

RAG, GPT4, OpenAI.

### Domain Topics:

Manufacturing System Optimization.

### Background Related Works (Complimentary knowledge):

**Retrieval-Augmented Generation (RAG) for Knowledge-Intensive NLP Tasks:**

* **RAG models, which combine pre-trained parametric memory with non-parametric memory (e.g., a dense vector index of Wikipedia), have shown state-of-the-art performance on open domain QA tasks, demonstrating more specific, diverse, and factual language generation compared to parametric-only models** [**(Lewis et al., 2020)**](https://consensus.app/papers/retrievalaugmented-generation-knowledgeintensive-lewis/fbc9d8d6f6de501987cc8c3afa034696/?utm_source=chatgpt)**.**

**REALM: Retrieval-Augmented Language Model Pre-Training:**

* **REALM models enhance performance in knowledge-intensive tasks by integrating a latent knowledge retriever, outperforming traditional LLMs in open-domain QA tasks by a significant margin** [**(Guu et al., 2020)**](https://consensus.app/papers/realm-retrievalaugmented-language-model-pretraining-guu/f82ec2e5e9675040a5a416f2a051a031/?utm_source=chatgpt)**.**

**Minimizing Factual Inconsistency and Hallucination in Large Language Models:**

* **This research addresses the critical challenge of ensuring the factual accuracy and reliability of LLM-generated responses. In the context of manufacturing QA, maintaining high accuracy and minimizing errors is crucial, making this work highly relevant** [**(Muneeswaran et al., 2023)**](https://consensus.app/papers/minimizing-factual-inconsistency-hallucination-large-muneeswaran/cb770953cfd65ba2b0b9dcacf1bef23a/?utm_source=chatgpt)**.**

**The Program Testing Ability of Large Language Models for Code**:

* Evaluating the capability of LLMs in program testing can provide insights into how these models can be used to test and verify manufacturing processes and systems. This paper's findings on improving code pass rates and leveraging human feedback are directly relevant to manufacturing QA applications [(Xiong et al., 2023)](https://consensus.app/papers/program-testing-ability-large-language-models-code-xiong/c9a0462513115833a03487e911dc4afa/?utm_source=chatgpt).

#### NSF Proposal

**https://new.nsf.gov/funding/opportunities/manufacturing-systems-integration-msi**

**The Manufacturing Systems Integration (MSI) Program:**

* addressing the opportunities and challenges that digital technologies present for the next industrial revolution, with particular emphasis on the digital integration of design and manufacturing within the larger life cycle ecosystem.
* address underlying principles and advances that are generalizable for globally competitive and world-leading industries.

Fundamental generalizable research for manufacturing systems integration:

* Digital representation, protocols, and/or processes for integration and collaboration in manufacturing systems (machines and/or humans)
* Intelligent self-organizing production systems
* Ease of use, interoperability and seamless integration of technologies, machines, and humans
* Service-oriented architectures and systems
* Datasets that are compatible and usable across platforms
* Reliable and secure communications within and across the manufacturing value chain
* Integration of distributed manufacturing systems across time and space, including incorporating both legacy and leading-edge equipment and technologies
* Methods for assessing the impact and value of externalities throughout the life cycle within the digital environment

### Correlated works(Competitor)

**Generative Retrieval-Augmented Ontologic Graph for Materials Design**:

* Using LLMs with retrieval-augmented ontological knowledge graphs aids in engineering analysis, materials design, and hypothesis generation, providing an interpretable structure for complex relationships and improving problem-solving strategies in material sciences [(Buehler, 2023)](https://consensus.app/papers/generative-retrievalaugmented-multiagent-strategies-buehler/936db9a659aa58c097329bdb2518e802/?utm_source=chatgpt).

**MechGPT for Mechanics and Materials Modeling**:

* The MechGPT model, fine-tuned for mechanics of materials, effectively retrieves and integrates domain-specific knowledge, demonstrating the potential of LLMs in bridging disparate areas of knowledge and providing interpretable insights through ontological knowledge graphs [(Buehler, 2023)](https://consensus.app/papers/mechgpt-languagebased-strategy-mechanics-materials-buehler/7926d56af9235d5aa5b3e3fac0a02647/?utm_source=chatgpt).

**ChatENT: Augmented Large Language Models for Expert Knowledge Retrieval in Otolaryngology**:

* This research presents a domain-specific fine-tuned model, ChatENT, which improves performance in medical examinations and education by integrating RAG techniques [(Long et al., 2023)](https://consensus.app/papers/chatent-augmented-large-language-models-expert-knowledge-long/249fd4c935b65ae0bc2ffdad8f610dc7/?utm_source=chatgpt).

**Transforming Healthcare Education: Harnessing Large Language Models for Frontline Health Worker Capacity Building using Retrieval-Augmented Generation**:

* This case study discusses the application of RAG models in enhancing healthcare education for community health workers, highlighting the potential of LLMs in educational improvements [(Al Ghadban et al., 2023)](https://consensus.app/papers/transforming-healthcare-education-harnessing-large-ghadban/a30de7ca31685a6d9f46daa702d11139/?utm_source=chatgpt).

**\*\* Generative Retrieval-Augmented Ontologic Graph and Multi-Agent Strategies for Interpretive Large Language Model-Based Materials Design**:

* The research explores the use of LLMs in materials design and manufacturing, demonstrating the benefits of fine-tuning with domain knowledge and utilizing retrieval-augmented ontological knowledge graphs [(Buehler, 2023)](https://consensus.app/papers/generative-retrievalaugmented-multiagent-strategies-buehler/936db9a659aa58c097329bdb2518e802/?utm_source=chatgpt).

**Injecting Domain Knowledge in Language Models for Task-Oriented Dialogue Systems**:

* This paper showcases the advantages of injecting domain-specific knowledge into pre-trained language models to improve task-oriented dialogue systems, particularly in response selection and generation tasks [(Emelin et al., 2022)](https://consensus.app/papers/injecting-domain-knowledge-language-models-taskoriented-emelin/90cfb54ef62857c1a1e5084ddef6e305/?utm_source=chatgpt).

**\*\* How Can Large Language Models Help Humans in Design and Manufacturing?**:

* This paper investigates the application of LLMs in various stages of the design and manufacturing workflow, from converting text prompts into design specifications to producing manufacturing instructions. This broad exploration is crucial for understanding the potential benefits and limitations of LLMs in manufacturing QA contexts [(Makatura et al., 2023)](https://consensus.app/papers/language-models-help-humans-design-manufacturing-makatura/69ae91fda8155d6dbb2f41635bba0ef4/?utm_source=chatgpt).

**\*\*** **Program Synthesis with Large Language Models**:

* This paper explores the capability of LLMs to synthesize code from natural language descriptions, which can be directly applicable in creating automated QA systems for manufacturing processes. The ability to generate and refine code based on human feedback is particularly relevant for developing robust QA tools [(Austin et al., 2021)](https://consensus.app/papers/program-synthesis-large-language-models-austin/7ef9943556d65cd78073d0ae4775f076/?utm_source=chatgpt).

**RAG vs Fine-tuning: Pipelines, Tradeoffs, and a Case Study on Agriculture**

* This paper explores two approaches for incorporating proprietary and domain-specific data into LLM applications: Retrieval-Augmented Generation (RAG) and fine-tuning. It proposes a multi-stage pipeline for both methods and assesses their performance using metrics on an agricultural dataset. The study shows that fine-tuning can increase accuracy by over 6 percentage points, with RAG adding a further 5 percentage points. The findings highlight how LLMs can be adapted for industry-specific applications, offering significant improvements in accuracy and knowledge integration. [(Balaguer2024)](https://ui.adsabs.harvard.edu/abs/2024arXiv240108406B/abstract)

## Story Build-up:

**Background in Manufacturing System:**

1. **The Importance of Manufacturing Systems for the Economy:** 
   * Manufacturing systems are crucial for economic vitality, contributing significantly to GDP, job creation, and technological advancement.
2. **Digital Integration of Design and Manufacturing**:
   * Advanced digital tools like CAD/CAM, digital twins, and simulation technologies allow for seamless integration between design and manufacturing stages. This integration enables real-time adjustments and optimizations, reducing time-to-market and enhancing the capacity to customize products to individual customer needs.
3. **Enhanced Data Analytics**:
   * Leveraging big data analytics and AI, industries can gain unprecedented insights into the manufacturing process, predicting maintenance needs, optimizing resource allocation, and improving product quality.

**LLM and RAG:**

1. **Emergence of LLMs:**
   * Large Language Models (LLMs) have been around for some time.
   * Applications of LLMs in various domains are becoming increasingly common.
2. **Benefits of LLMs:**
   * LLMs have demonstrated the potential to improve efficiency.
   * They can also help reduce costs.
3. **Challenges with LLMs:**
   * Issues such as hallucinations and incorrect answers can be confusing.
4. **Introduction of RAG:**
   * Scientists introduced Retrieval-Augmented Generation (RAG) to address these issues.
   * RAG can provide citations from knowledge bases, offering stronger reasoning.
5. **Introduction of Agents:**
   * Agents enhance reasoning for knowledge that cannot be directly extracted from articles.
   * They use tools like Python and user proxies to execute code, making answers more reliable.

**Material Manufacturing Domain:**

1. **Challenges for Manufacturing Students:**
   * Students face difficulties in understanding complex manufacturing processes.
   * Integrating theoretical knowledge with practical applications is challenging.
2. **Potential of LLMs in Manufacturing Education:**
   * LLMs can assist with various aspects of education and training.
   * Fine-tuned knowledge bases can help students gain a better understanding and grasp knowledge more effectively.
3. **Problems with Previous Works:**
   * **Problem 1:** Lack of domain-specific accuracy leading to irrelevant or incorrect information.
   * **Problem 2:** Inability to provide practical, hands-on examples and applications.
   * **Problem 3:** Difficulty in integrating theoretical knowledge with practical tools and processes.
4. **Our Proposed Solution:**
   * **Problem 1:**
     + Implementing a domain-specific fine-tuning process to improve the accuracy and relevance of the information provided by LLMs.
   * **Problem 2:**
     + Utilizing RAG to integrate practical examples and applications into the learning material, making it more relevant and useful for students.
   * **Problem 3:**
     + Developing Agent-based systems that can execute code and simulate real-world manufacturing scenarios, helping students bridge the gap between theory and practice.

## Branch – Different Rags:

Ways: {

Dense Retrieval,

Sparse Retrieval,

Hybrid Retrieval,

Retrieval-Enhanced Transformers,

Memory-Augmented Models,

Retriever-Reader Architecture,

Indexing Techniques,

Knowledge-Augmented Retrieval,

Meta-Learning for Retrieval,

}

**RAGAS: Automated Evaluation of Retrieval Augmented Generation**:

* Authors: ES Shahul, Jithin James, Luis Espinosa Anke, S. Schockaert
* Abstract: Introduces RAGAs, a framework for evaluating RAG pipelines without relying on ground truth human annotations.
* [Read more](https://consensus.app/papers/ragas-automated-evaluation-retrieval-augmented-shahul/1e214d3a38e2558b8d6595f964842bca/?utm_source=chatgpt)

**Retrieval-augmented Generation across Heterogeneous Knowledge**:

* Author: W. Yu
* Abstract: Discusses the challenges and solutions for retrieving knowledge from a heterogeneous corpus for RAG methods.
* [Read more](https://consensus.app/papers/retrievalaugmented-generation-across-heterogeneous-yu/1884e4026419520f9be8409a2402a340/?utm_source=chatgpt)

**Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks**:

* Authors: Patrick Lewis, Ethan Perez, Aleksandara Piktus, et al.
* Abstract: Explores a fine-tuning recipe for RAG models combining pre-trained seq2seq models and dense vector indexes.
* [Read more](https://consensus.app/papers/retrievalaugmented-generation-knowledgeintensive-lewis/fbc9d8d6f6de501987cc8c3afa034696/?utm_source=chatgpt)

**Benchmarking Large Language Models in Retrieval-Augmented Generation**:

* Authors: Jiawei Chen, Hongyu Lin, Xianpei Han, Le Sun
* Abstract: Systematically investigates the impact of RAG on large language models and proposes a new benchmark.
* [Read more](https://consensus.app/papers/benchmarking-language-models-retrievalaugmented-chen/7e1513ce958c5b7b8b96901750b26f3f/?utm_source=chatgpt)

**Using retrieval-augmented generation to elevate low-code developer skills**:

* Author: Nakhod O
* Abstract: Applies RAG to improve low-code developer skills by augmenting LLMs with domain-specific knowledge.
* [Read more](https://consensus.app/papers/using-retrievalaugmented-generation-elevate-lowcode-o/6784651a164f5711a6f4fa6affdafa3b/?utm_source=chatgpt)

**GAR-meets-RAG Paradigm for Zero-Shot Information Retrieval**:

* Authors: Daman Arora, Anush Kini, Sayak Ray Chowdhury, et al.
* Abstract: Proposes a novel GAR-meets-RAG recurrence formulation for zero-shot information retrieval.
* [Read more](https://consensus.app/papers/garmeetsrag-paradigm-zeroshot-information-retrieval-arora/38c91730674f521a864864b2b9d5a807/?utm_source=chatgpt)

**Enhancing LLM Intelligence with ARM-RAG: Auxiliary Rationale Memory for Retrieval Augmented Generation**:

* Author: Eric Melz
* Abstract: Explores the use of ARM-RAG to improve problem-solving performance in LLMs.
* [Read more](https://consensus.app/papers/enhancing-intelligence-armrag-auxiliary-rationale-melz/67809c740f7e5135bd930c6f2bcca5b8/?utm_source=chatgpt)

**IAG: Induction-Augmented Generation Framework for Answering Reasoning Questions**:

* Authors: Zhebin Zhang, Xinyu Zhang, Yuanhang Ren, et al.
* Abstract: Proposes IAG framework utilizing inductive knowledge for implicit reasoning in QA tasks.
* [Read more](https://consensus.app/papers/inductionaugmented-generation-framework-answering-zhang/d454855ba2c2564094a027a7d9cbc101/?utm_source=chatgpt)

**Retrieval Augmented Generation and Representative Vector Summarization for large unstructured textual data in Medical Education**:

* Authors: S. S. Manathunga, Y. A. Illangasekara
* Abstract: Discusses the application of RAG in medical education using representative vector summarization.
* [Read more](https://consensus.app/papers/retrieval-augmented-generation-representative-vector-manathunga/81b3d5e132545a4798f18bfeab41fe8d/?utm_source=chatgpt)

**Improving the Domain Adaptation of Retrieval Augmented Generation (RAG) Models for Open Domain Question Answering**:

* Authors: Shamane Siriwardhana, Rivindu Weerasekera, Elliott Wen, et al.
* Abstract: Evaluates joint training of retriever and generator components of RAG for domain adaptation in ODQA.
* [Read more](https://consensus.app/papers/improving-domain-adaptation-retrieval-augmented-siriwardhana/e7ce39892716576d921937ae79fd172e/?utm_source=chatgpt)

## 20240712

#### Purpose:

1. contribute manufacturing RAG dataset

2. 尝试提升performance of manufacturing Information retrieval using fine-tuning or graph rag (it’s a tool)

3. based on 2加点自己的创新

4.考虑解决多选题和计算题的RAG，这部分看123的完成度再考虑加不加

5.

#### TODO This week, due next week:

Z:

1. Better background study about how to generate RAG answer files and the format.
2. Experiment on rag files: [GrooverSolution/ch18.pdf] to rag file format.
3. Find out how to calculate the RAG loss.
4. Better understanding of different rag methods. (Later in the experiment for comparison) 先做openai 的作为baseline，来看其他的在manufacturing中的表现怎么样

Connie:

1. Reorganize the files.
2. Project direction leading. Now: More Engineering project
3. How to evaluate the quality of RAG dataset

Z:

Background:

* LlamaIndex and Langchain are similar

Pre: ReadPDF

Mission 1: Chunk\_size and Chunk\_Overlap ration around: 10:1 a reasonable amount. Chunk\_size = [300,500]

Mission 2: Embedding, normally OpenAI Embedding

Mission 3, Save the Embedded Vector to Database: VectorDB

储存的文件是一个project file， 路径是：CHROMA\_PATH

db\_chroma = Chroma(persist\_directory=CHROMA\_PATH)

Ref： [Link](https://medium.com/@drjulija/what-is-retrieval-augmented-generation-rag-938e4f6e03d1)

Discussion about：​​我觉得与其去研究funetune 不finetune， 我觉得我们可以着重看一下不同的rag方式的表现差异， 做实验对比， 我们选择最好的（contrb 1）然后人工做一个benchmark的数据集作为标准答案， 方便对比模型生成的答案是不是好的（contrb 2）， 在大部分解决了单选题的基础上， 我们同时可以提供graphgrag， 可以非常有助于学生/提问的人 对于问题和解决思路的直观理解， 这个我觉得在很多领域都还没有广泛运用起来, 可以多少算点创新（contrib 3） 在manufacturing system 中， 肯定会有计算的问题， 我们能否把计算的方法也放进去， 让模型可以更好的解决计算的问题， 从而提高准确度， 解决calculation intensive的问题（contrib4， 有时间再搞）最后就是搞个前端， 部署在hugging face 上之类的， 让public accessible， 让科研可以产生实际的价值（contrib5）