A comparative study between humans and GPT for TradeSpace analysis

by

Aathira Anil Kumar

(Under the Direction of Beshoy Morkos)

Abstract

With large language models becoming more and more popular and their usage expanding into several different domains, this study explores how effective a Large Language Model like GPT4 would be in analyzing requirement specification documents and using relevant information from those to create a tradespace matrix. The research compares GPT-4's performance with 30 human participants divided into three groups: engineering background, non-technical background, and computer science/artificial intelligence (CS/AI) background. Each participant and LLM completed a survey based on the provided materials. The analysis included within-group heatmaps, across-group comparisons, and human vs. GPT-4 heatmap evaluations. Results showed the CS/AI group had the closest responses to GPT-4. The study showcases how LLMs can be used to augment trade-space exploration and can be used as an alternative in cases where a team of human experts from different backgrounds is not feasible.

Index words: LLM, GPT, ReAct, RAG, Multi-Agent, Tradespace Analysis, Tradespace Exploration, Requirements Engineering

A comparative study between humans and GPT for TradeSpace analysis

by

Aathira Anil Kumar

B.Tech., Rajagiri School of Engineering and Technology, India, 2019

A Thesis Submitted to the Graduate Faculty of the

University of Georgia in Partial Fulfillment of the Requirements for the Degree.

Master of Science

Athens, Georgia

2024

A comparative study between humans and GPT for TradeSpace analysis

by

Aathira Anil Kumar

Major Professor: Beshoy Morkos

Committee: Ismailcem Budak Arpinar

Lefteris Jason Anastasopoulos

Electronic Version Approved:

Ron Walcott

Dean of the Graduate School

The University of Georgia

August 2024

Dedication

This work is dedicated to my father, who dreamed of this degree before I did, to my mother, who hates it when I leave but always encourages me to spread my wings, to my baby sister, who, one might say, was quite brutal about it but still made sure I never quit, and to my roommate, who spent many an hour listening to me while I ranted about life.

It is also dedicated to those loved ones who made me laugh and touched my heart.

Acknowledgments

Special thanks go to my thesis committee: Dr. Beshoy Morkos, Dr. Ismailcem Budak Arpinar and Dr. Lefteris Jason Anastasopoulos who accepted and encouraged me to try out all that I wanted to during my study and guiding me through this research when it felt like I was adrift. I would also like to thank all the members of the Institute of Artificial Intelligence, UGA that pushed me towards the finish line. And of course, to God, for giving me the life that I have.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

A tradespace can be thought of as a space that contains all the design variables related to a problem (Ross & Hastings, 2005). The primary benefit of tradespace analysis lies in helping stakeholders understand the relationship between different options or utilities and evaluating their respective benefits and costs. This calculation is essential for a project as one would want to select the best combination of attributes that satisfy some predefined criteria. Traditionally, tradespace analysis requires extensive manual effort, expert knowledge, and time-consuming calculations. Analysts must sift through vast amounts of data, often dealing with complex, multi-dimensional trade-offs to identify the best solutions. However, as technology continues to advance, new methodologies have emerged that can be used to enhance and assist the process. The use of Large Language Models (LLMs) is one such promising avenue.

1.1 Motivation For The Study

LLMs, like OpenAI's GPT-4, are powerful models that have been trained extensively on diverse datasets, leaving them with remarkable capabilities when it comes to understanding and generating textual information. This capability makes them suitable for a wide variety of tasks, of which this thesis is interested in their ability to perform data analysis and decision support. The use of LLMs for tradespace analysis appears to have a lot of potential. To begin with, when compared to traditional methods, LLMs can process

and analyze large volumes of information at a faster rate and with less effort. Second, LLMs might identify patterns within data that might be overlooked by human analysts, leading to more accurate and insightful analyses. Furthermore, like all other Artificial Intelligence (AI) algorithms, LLMs can continuously learn and improve , enhancing their performance over time to adapt to evolving requirements and constraints.

1.2 Research Objective

The primary objective of this research is to explore the potential of Large Language Models (LLMs) in supporting tradespace analysis, particularly by addressing the multidisciplinary background that is typically required for effective decision-making. Tradespace analysis often demands input from specialists across various domains to evaluate complex trade-offs and identify optimal solutions. However, not all organizations have access to a diverse team with different backgrounds, which can limit the scope and depth of their analyses. This research aims to investigate how LLMs can bridge this gap by simulating the knowledge and insights of a multidisciplinary team. Specifically, the objectives include:

1. Evaluating the capability of LLMs in Tradespace Analysis: Determine if LLMs can effectively conduct tradespace analysis using project specification documents and if they can justify the trade-offs made.

2. Enhancing Performance with Retrieval-Augmented Generation (RAG): Investigate how RAG can enhance the performance of LLMs in tradespace analysis and if the model can retrieve and access relevant information, thus improving the accuracy and relevance of their analyses.

3. Implementing a ReAct (Reasoning and Acting) approach: Explore whether the ReAct method enhances LLM decision-making in tradespace analysis. This would involve evaluating how well the LLM can reason through complex trade-offs and take appropriate actions based on its analysis.

4. Incorporating Multi-Agent Systems: To mimic a multidisciplinary team's collaborative effort, multiple agents with different roles can be created and then evaluated to see how the different agents can work together to perform a comprehensive analysis.

By achieving these objectives, this research seeks to demonstrate the potential of LLMs as a valuable tool in tradespace analysis, capable of enhancing the decision-making processes through their advanced data processing capabilities.

1.3 Research Question

The research questions that this study attempts to resolve are along the lines of the objectives that were listed above and are:

- Are Large Language Models (LLMs) capable of performing tradespace analysis when presented with a project specification document(s)?
- If they can, how does it compare to how humans from different domains approach the same task?
- Can the ability to perform data analysis be enhanced or improved using techniques such as Retrieval Augmented Generation, ReAct, or Mutli-Agent systems?

1.4 Structuring of Thesis

The rest of the paper is organized as follows: Chapter 2 contains the literature review pertaining to relevant topics, and Chapter 3 talks about the software environments and libraries that were used and presents some background information about the different approaches. In Chapter 4, the methodology for the experiments is detailed, and Chapter 5 focuses on the discussion regarding the outcomes of the experiments. Finally, in Chapter 6, the conclusion is presented, as well as what the future works could entail.

Chapter 2

Literature Review

Research work that relates particularly to the use of Large Language Models for tradespace analysis was not found despite an exhaustive search. Papers relating to tradespace analysis, interdisciplinary design, and requirements engineering (RE) were explored instead to provide the base for this research.

2.1 Tradespace Analysis

As defined by Spero et al., 2014, a tradespace can be thought of as a "multidimensional solution space" in which various design options and their associated performance metrics are presented. It includes all information in a detailed manner so that how each utility relates to another and the extent to which it does can be understood. This enables stakeholders to analyze the trade-offs between different design options. Tradespace analysis can then be understood as the process of systematically exploring, understanding and evaluating the tradespace in order to perform an informed decision-making. As shown in the paper (Spero et al., 2014), this involves generating a number of design alternatives early in the process and then assessing them based on certain criteria before selecting the best one. The main goal is to provide a comprehensive understanding of the trade-offs that are present and what the best combination of the utilities is under a given condition. Some of the benefits of tradespace analysis or exploration, as can be seen in the paper

"Designing the Design Space: Evaluating Best Practices in Tradespace Exploration, Analysis and Decision-Making" (Daniels et al., 2022), include:

1. Informed decision-making. The decision-makers in a project are presented with all the information regarding the risks and benefits of the different utilities and their relations with each other. The potential trade-offs and compromises are presented to the stakeholders after performing several optimizations and considering the design variables in different manners. Since all the possible solutions are considered and evaluated, the stakeholders can be assured that the best one is chosen.

2. Risk assessment. Performing tradespace analysis in the early stages of a project ensures that the given plan can be implemented, has minimal risks, and can present the solution that is required by the project. This early assessment can differentiate between project success and failure by addressing potential issues before they escalate.

3. Collaborative development. Since tradespace exploration would involve multiple stakeholders, it means that different perspectives and requirements would have to be considered. This would ensure that the final solution would be satisfactory for all the parties and would be more likely to be relevant, and thus accepted.

4. Factors in complexity and uncertainty. Tradespace analysis can be manipulated to factor in different conditions that could arise during the implementation of a large project and can be used to simulate any potential problem that could be faced and what the best plan of action, in that case, could be. In addition to this, as changes happen, the tradespace can be updated to factor in the latest developments, and the analysis can be performed again so that the best plan is continually chosen.

Tradespace exploration presents several challenges, particularly when applied to complex systems. Considering an industry such as the construction machine industry, some of the challenges, as identified by Machchhar et al., 2024 are:

1. Integrating new technologies. As new ideas and novel approaches and utilities emerge, it becomes hard to incorporate these into the traditional workspace because one might not be aware of the effects they would have on the existing design variables.

2. Quantifying changeability. When performing tradespace analysis, it is important to understand the potential impact of each design variable on the system and to evaluate its value with regard to the solution and with regards to the other variables. This impact then has to be presented in a quantified format for the stakeholders to make a decision. In order to do this, a detailed understanding of the system is required. For large projects, this would be a very complex process.

3. Factoring in uncertainty. At times, multiple tradespaces would have to be developed to simulate potential situations in the future, and the best strategies for those would also have to be developed. And even when the situation is as predicted, and the design variables remain unchanged, plans might have to be adapted to consider factors such as human interactions, legislation, and others. Whilse these factors do not relate to the dependencies within the system, they can still influence the outcome of a project.

Tradespace analysis is often performed in tandem with a cost model (Xu et al., 2023). This allows designers to compare the estimated expenses of the various design options with the project's budget. This ensemble approach helps to identify the best utility design that achieves a good performance and is cost-effective as well.

2.2 Interdisciplinary Research

Globalization has greatly advanced the scope of interdisciplinary research by removing geographical and communication obstacles, as detailed in the paper "Getting Closer or Drifting Apart?" (Rosenblat and Mobius, 2004). The availability of cheaper communication technologies like the Internet and email meant that people were no longer forced to pay steep prices in order to collaborate with peers from different fields and different locations. This, in turn, increased the ability to share data and insights, collaborate

on projects, and to collectively research and implement new, interdisciplinary methodologies and, by extension, interdisciplinary projects.

Dalton et al., 2022 in their paper "Interdisciplinary Research as a Complicated System" describe interdisciplinary research as a "complex system consisting of researchers from different disciplines that have undergone a pseudomorphosis". This essentially means that it is a system where people from different fields come together to work on a particular task. "Pseudomorphosis" is used to show how the researchers trade in a part of their individual freedom in order to work more effectively as a team. A large part of an interdisciplinary research's success depends on how clearly the research problem is defined and how each researcher from a different domain can contribute towards the common goal. Some of the benefits of interdisciplinary research include:

1. Working with researchers from different backgrounds exposes everyone to different viewpoints and methodologies, which can broaden their understanding and help them think outside the box.

2. Researchers also pickup the skills on how to communicate more effectively and clearly. As their colleagues would come from different domains, complex ideas or theories would have to be broken down into simpler terms. In the long run, this means that the findings of the research can also be better explained to or understood by a larger audience.

3. Interdisciplinary research facilitates resource sharing, including equipment and data. This, in turn, offers the researchers a more flexible approach to their methods and presents them with new approaches that could be explored.

4. Real-world issues are often multifaceted and complex and would require input from different disciplines to be addressed. Working with people from other fields also means that new insights could be generated based on how each field approaches the problem, giving a more robust solution.

Despite the benefits of interdisciplinary research, it is not without challenges. Some of the challenges that were noted as part of a workshop and released by Daniel et al., 2022 include:

1. Communication discrepancies. The language and terminologies that are used by different disciplines would differ. Each discipline would have its own "jargon" that people in that domain would understand, but it would not be universally understood. This necessitates additional time and effort to clarify terms and ensure effective communication among team members.

2. Conflicting approaches. Interdisciplinary research often involves having to combine or choose between different problem-solving approaches that are inherent to each discipline. Researchers need to navigate through these conflicting paradigms, which requires a deep understanding of the various methodologies. More importantly, the collaborators must be willing to adapt and integrate diverse research standards.

3. An interdisciplinary researcher might find themselves feeling somewhat lost as they struggle to identify themselves as a central part of an interdisciplinary group. However, at the same time, they also feel isolated within their home department. This problem is enhanced when it comes to publishing their research as well. A researcher might find it difficult to publish his or her work in journals outside their primary discipline, which often leads to challenges in gaining recognition and advancing their career. Even after modifying their literature to reflect approaches from different domains, the question arises about where the journal should be published.

4. Securing funding and resources. Funding is an essential requirement for research. However, proposals are typically evaluated using discipline-specific standards, making it difficult for interdisciplinary research projects to receive support.

A recent paper released by Hu et al., 2024 shows how valuable interdisciplinary research is when tackling challenges that span different areas, such as the scientific, societal, and policy domains, by performing a study that shows a positive correlation between research and policy attention in the COVID-19 context. Having individuals from different domains often enhances the value of the research and drives technological innovation in a collaborative and social manner.

2.3 Requirements Engineering

Requirements Engineering (RE) is a foundational process in software and system development, encompassing the activities of gathering, analyzing, specifying, validating, and managing requirements. It serves as a critical phase in ensuring that the developed systems meet the needs and constraints of stakeholders and are aligned with business goals. Requirements remain as a major component of engineering design. To assist in their management, there are tools that exist to support requirements elicitation, management, and verification (McLellan et al., 2010; Morkos et al., 2010a, 2010b). However, requirements nonetheless experience requirement changes that have to be managed. Hein et al., 2015, 2021, 2022; Htet Hein et al., 2017; Morkos and Summers, 2010; Morkos et al., 2012; Shankar et al., 2012; Summers et al., 2014). These changes may occur due to trade space exploration where changes to a requirement to accommodate a trade space will necessitate subsequent downstream propagation. Newer tools to support this include the use of Natural Language Processing methods and Transformers to (Chen, 2022; Chen, Carroll, and Morkos, 2023; Chen and Morkos, 2023; Chen, Wei, and Morkos, 2023; Chen et al., 2021; Mullis et al., 2023).

We have seen requirements also play a role in manufacturing of systems (Olajoyegbe and Morkos, 2022; Piazza et al., 2017). Silva et al., 2019 and Gong et al., 2021 define requirements as a key process in manufacturing design, highlighting the stages of eliciting, crowdsourcing, modeling, and validation or verification, which rely on a formal approach initiated at the beginning of the process. Elneel et al., 2022 describe RE as a set of activities essential for extracting accurate requirements, emphasizing its significance in the software development life cycle and the success of software projects. Additional challenges in requirements engineering as understood by Silva et al., 2019 and Elneel et al., 2022 include:

- Requirements Elicitation:
  - Understanding the objectives and the purpose for the project can be difficult, especially when dealing with complex requirements.

  - As stakeholders come from different backgrounds and have different manners of understanding and communication, gathering the accurate requirements can be challenging.

- Ambiguity and Inconsistency:
  - Requirements are often specified using natural language, which can lead to misunderstandings, ambiguities, and inconsistencies in the documented requirements.
  - As different stakeholders and components are involved in requirements documentation, it can become difficult to maintain consistent reports across the system.

- Complexity in Large-Scale Systems:
  - It is difficult to maintain a shared understanding of the requirements when working together with multiple teams on large-scale projects.
  - For complex projects, it is difficult to manage and map requirements as they change, and to make sure that all changes are noted in a manner that can be traced across the teams.

Despite the challenges, effective RE provides numerous benefits that are crucial for the success of engineering and system development projects:

- Enhanced Quality and Consistency:
  - Requirement documentations help ensure that the developed system meets the needs and expectations that were set by the stakeholders.

- Improved Communication and Understanding:
  - Helps reduce misunderstandings between the developers and the project owners by aligning the objectives with the system design.
  - Allows for a better understanding of the requirements, which is important in order to accurately implement the functionalities.

- Risk Mitigation:
  - It identifies potential risks early in the development, which allows them to be addressed beforehand, thereby preventing costly mistakes and delays.
  - Serves as a placeholder that helps to manage all the changes in the requirements, and makes sure that the changes are aligned with the stakeholders' end-goals.

Chapter 3

Environments and Background Theory

This section details the experimental setup, including the libraries, frameworks, and software environments utilized. It also provides background information on large language models (LLMs) and describes the various approaches employed in the experiments.

3.1 Environment and Software

3.1.1 Jupyter Notebook and Google Colab

For the practical components of this thesis, which involves the use API calls to the GPT models and utilizing Python for visualizing and analyzing the results, it was advantageous to use Jupyter Notebooks [1] that were hosted on Google Colab [2].

[1] https://jupyter.org/
[2] https://colab.research.google.com/

Jupyter Notebook

Project Jupyter ([3]) is a spin-off project that is committed to creating open-source software. Jupyter notebooks, in particular, are helpful in integrating live code with visualizations, making it easier to analyze and present data. The notebook consists of "code" cells that execute the code in the order of your choosing, thus promoting a modular approach towards coding. This approach was highly beneficial for the purposes of this thesis. API calls were made in one code cell, cleaned in another, and then fed into a function for visualization. Each cell executed a specific task, which simplified the debugging process and allowed code blocks to be organized into distinct sections. This organization made modifications more effortless and efficient, contributing to a smoother workflow and more effective data analysis.

Google Colab

The primary motivation for using Colab was to gain access to its powerful computational resources, including CPUs, GPUs, and TPUs. For visualization purposes, these computational resources were invaluable as there was a lot of graphical data that needed to be generated and analyzed. In addition to this, as Google Colab primarily extends the functionalities of Jupyter Notebook into the cloud, it was easier to integrate it with a Google Drive containing all the data collected as part of this experiment, process it as required, and save the results in the same drive. This approach mitigated the risk of data loss and made it easier to share the results that were gathered. Finally, as Colab comes with several pre-installed libraries that could be used for handling the data and creating the graphs, the environment was easy to set up and use.

[3] https://en.wikipedia.org/wiki/Project_Jupyter

3.1.2 LangChain and LangGraph

LangChain

LangChain [4] is a framework that enhances the capabilities of Large Language Models by enabling their use in a chained manner. Whereas traditionally, connecting a LLM to an external data source would take some effort, LangChain abstracts and simplifies the process for the user. In the context of this thesis, LangChain has been used for the RAG approach and the ReAct agent implementation.

LangGraph

LangGraph [5] extends the LangChain framework to build multi-agent applications. While LangChain can create chains between the LLMs and other components, LangGraph introduces the cyclic capabilities that allow you to control the interactions between the different agents in a more controlled manner without repetitive code. As was said in their blog post, "LangGraph is a module built on top of LangChain to better enable creation of cyclical graphs, often needed for agent runtimes." (https://blog.langchain.dev/langgraph/).

3.1.3 Vector Database

A vector database[6] is a data storage system that is designed to manage and store data in a vector format. Unlike most databases that deal with structured data, vector databases are capable of handling unstructured data, including text, images, and audio. The information is stored in a manner that captures its semantic meaning, which makes vector databases a suitable option for use cases such as similarity search and clustering. For the purpose of this research, ChromaDB was used to handle the storage and retrieval of the information when implementing the RAG, ReAct and Multi-Agent approach. Chroma [7] is open-

[4] https://www.langchain.com/
[5] https://langchain-ai.github.io/langgraph/
[6] https://en.wikipedia.org/wiki/Vector_database
[7] https://www.trychroma.com/

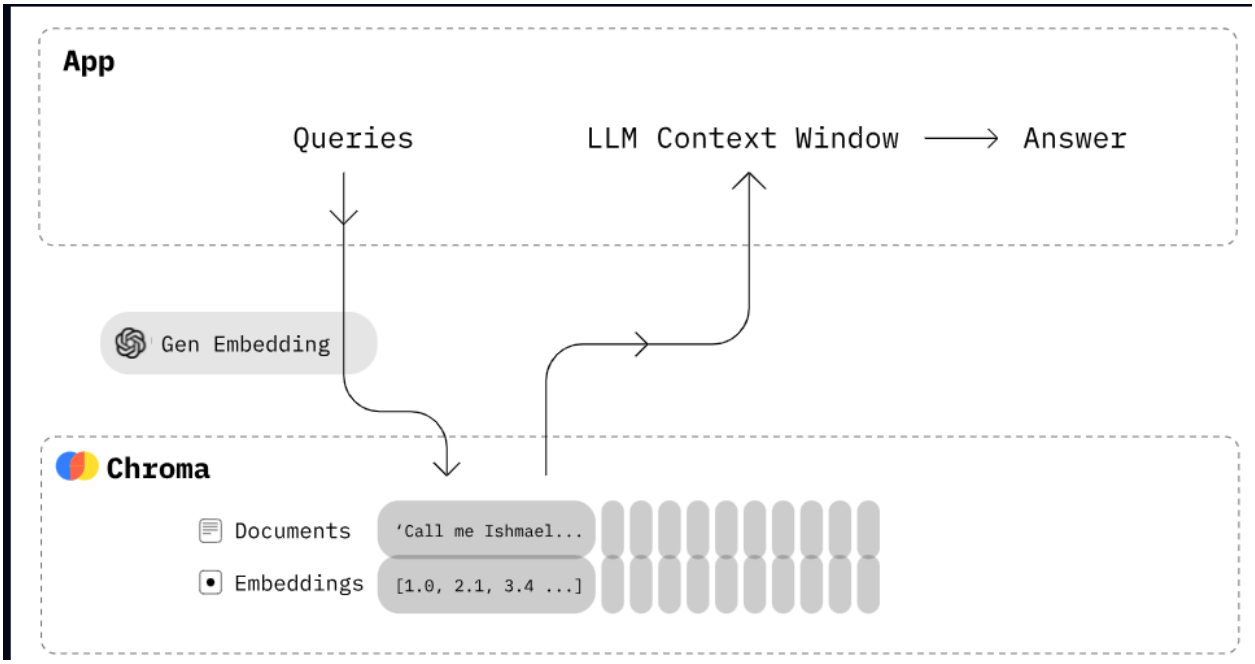


Figure 3.1: Chroma (https://docs.trychroma.com/)

source, has a self-hosted option, and, most importantly, worked in the Google Colab environment. Figure 3.1 shows how Chroma usually works with a LLM:

- Documents are chunked and stored as embeddings in the database.
- When the user sends a query to the LLM, it is first converted to embeddings and then sent to the database.
- A semantic search is done to retrieve the most similar or relevant chunks.
- The retrieved information is added to the query and then sent to the LLM
- The LLM uses the retrieved information present in the context to form its response to the query.

3.2 Background Theory

3.2.1 Transformers

The transformer architecture was introduced in the paper "Attention Is All You Need" (Vaswani et al., 2017). The architecture was primarily introduced for the purpose of using a self-attention mechanism alone to model the dependencies between the input and output sequences in place of several recurrent neural networks (RNNs) and convolutional neural networks (CNNs), which were traditionally used in sequence transduction models. The Transformer model (Figure 3.2) employs an encoder-decoder structure, where both the encoder and decoder consist of several identical layers. Each encoder layer includes a multi-head self-attention mechanism followed by a position-wise feed-forward network. The decoder layers are similar but incorporate an additional layer for attention over the encoder's output, allowing the model to attend to all positions in the input sequence while generating the output. The model has an encoder-decoder architecture, as can be seen from the figure:

- Encoder: Made up of identical layers, each with a multi-head self-attention mechanism followed by a feed-forward network. The self-attention mechanism allows the model to weigh the importance of the different words with respect to each other, and the feed-forward network processes this information.

- Decoder: The decoder is similar to the encoder's structure but has an additional multi-head attention layer that processes the encoder's output. The self-attention mechanism is also changed so that while the decoder can attend to all positions in the input sequence, during a prediction, it can only use the words that are present prior to the input and not those that follow it (Vaswani et al., 2017).

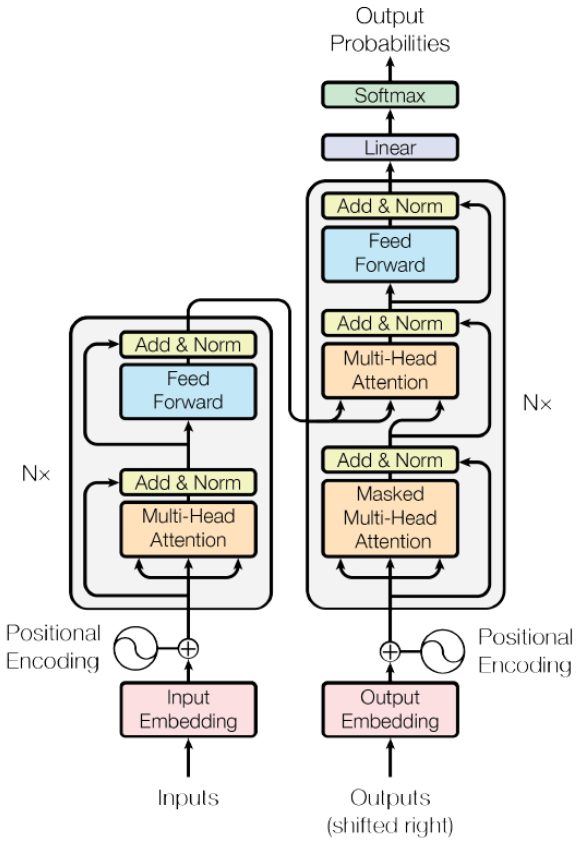


Figure 3.2: Transformer - model architecture (Vaswani et al., 2017)

3.2.2 ReAct

"ReAct: Synergizing Reasoning and Acting in Language Models" (Yao et al., 2023) shows a new approach, named ReAct, to improve the performance and trustworthiness of Large Language Models by combining reasoning traces with task-specific actions. The ReAct approach builds on the Chain-of-Thought (CoT) (Wei et al., 2023) approach, where a LLM "thinks" through a given prompt, and the "act" approach which allows LLMs to interact with external environments. Quite notably, the ReAct approach was not susceptible to hallucinations, which, along with error propagation, was one of the significant drawbacks of the CoT approach. A closer look at Figure 3.3 explains the Reason Only" and "Act Only" approach as

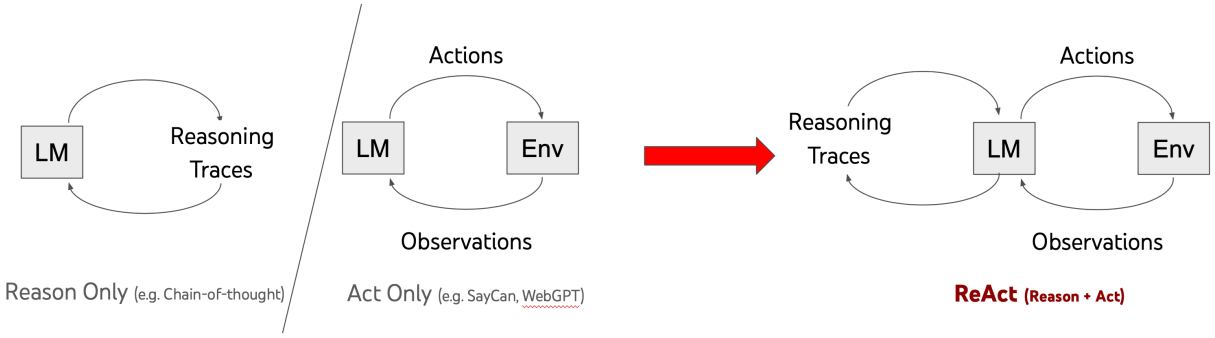


Figure 3.3: ReAct: Synergizing Reasoning and Acting in Language Models (https://react-lm.github.io/)

below:

- Reason Only: Here, "reasoning traces" are generated by the LLM and then fed back into them to produce a step-by-step approach for a given prompt. While this does produce good results, the LLM does not have the ability to interact with other external sources, which is why the issues mentioned previously (hallucination, error propagation) are likely to occur.

- Act Only: Conversely, in papers like Nakano et al., 2022 we can see that language models can be used to interact with tools and APIs, and using the output from these interactions, the LLM can improve their answers. While this approach can provide a model with updated information, it does not explicitly follow a "reasoning" approach.

3.2.3 Multi-Agents

As Large Language Models continued to reach new levels of success in different tasks, it was only a matter of time before an iterative workflow where these agents could communicate with each other to achieve a goal was designed. In a multi-agent framework like AutoGen(Wu et al., 2023), each agent could be a single LLM that is directed by a prompt to have a specific capability or is specialized in a particular domain and can take advantage of a certain set of tools or functions, and it works in tandem with other such diverse agents. As seen in Guo et al., 2024, multi-agent systems have been used in various domains ranging from software programming to policy making. In every multi-agent system, the vital decisions are:

- How the agent interacts with its environment. For an agent to make informed decisions or perform specific actions, it must be aware of the current state of its environment (in a physical or virtual context). Additionally, understanding the impact of its actions on this environment is essential. This is then used to guide the agent on what should be done next.

- The role of the agent(s). It is important to understand the problem that the system is being designed for and to determine what kind of traits or skills must be given to each agent. Each agent in the network would contribute towards a specific task, and would be given the necessary functions or tools that could be used to improve the output.

- How the agents communicate with each other. The multi-agent system is based on the models communicating with one another. What is communicated, how it is communicated, and the order in which information is communicated plays an important role. The messages could be instructions, information, or maybe the output of a certain action that one agent took. The order could be a layered one where one LLM takes a supervisory role, or it could be one where the LLMs act as peers.

Whilst multi-agent systems have a huge potential, factors such as the cost of using several agents using models like GPT-4 and the increasing complexity when it comes to determining what role each agent plays must be considered and optimized.

3.2.4 RAG

Retrieval Augmented Generation (RAG) is a technique that enhances Large Language Models by allowing them to interact with external sources of information when responding to a query (Gao et al., 2024). This approach helps address some of the critical limitations of LLMs, such as hallucination and not having access to current or updated information or to information pertaining to a particular domain. RAG combines the extensive, pre-existing knowledge within LLMs with up-to-date information from any external data source. RAG operates through a structured process consisting of three phases: retrieval,

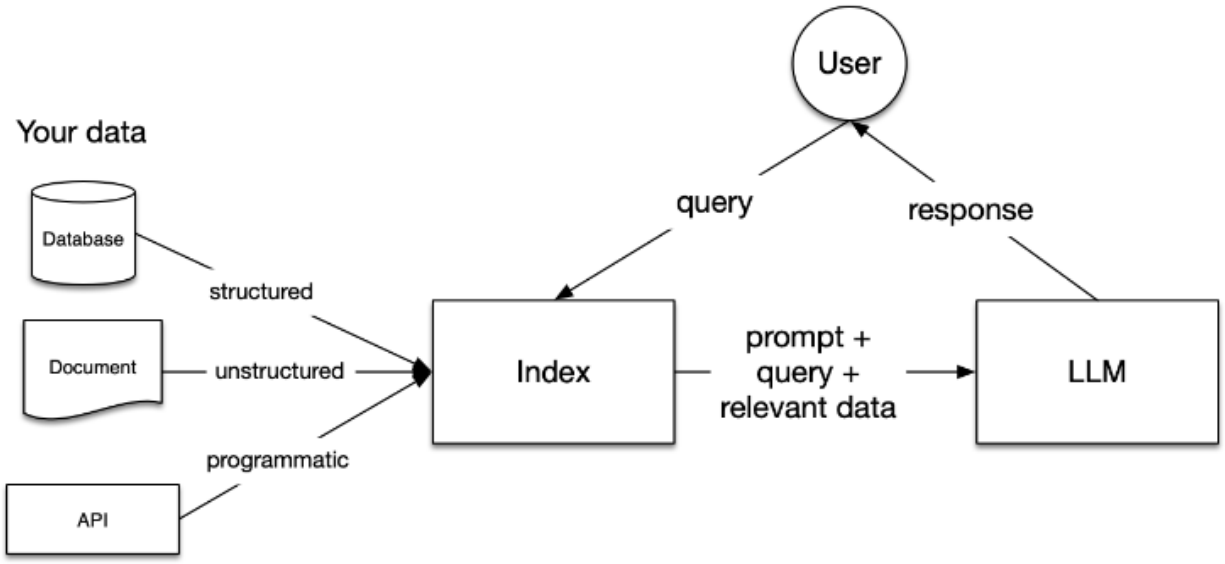


Figure 3.4: RAG https://docs.llamaindex.ai/en/stable/getting_started/concepts/

generation and augmentation. As shown in seen in Figure 3.4, the process usually follows the steps below:

- Indexing: Information from different sources (databases, PDFs, HTML pages etc.) is collected, cleaned, and then segmented into smaller chunks. Using an embeddings model, the chunks are converted into vectors to store into a vector database.

- Retrieval: When the user submits a query, the system encodes the query into a vector as well, and then a semantic or similarity search is performed so that the document chunks with the best similarity score with respect to the query are retrieved.

- Generation: The retrieved document chunk(s) are then combined with the user's query to create a comprehensive prompt, which is fed into an LLM. With the augmented information, the LLM is more likely to come up with an answer that is contextually relevant and more accurate.

As RAG became popular, it has evolved to become more accurate using techniques such as metadata attachment (better indexing), improved querying processing (makes the query clearer), re-ranking (retrieves the most relevant context from the retrieved information), context compression (large contexts can cause the LLM to focus on the wrong information) and the introduction of specialized modules and patterns (Gao et al., 2024).

Chapter 4

Methodology

The primary objective of this study was to systematically evaluate and compare the performance of an LLM like GPT against human participants in trade-space analysis. The study was conducted in two phases. In the first phase, the trade-space analysis was performed using a ReAct agent and then a multi-agent system. Both the agent and the system used GPT-4 as the driving model. These responses were then processed for the second phase, where a comparative analysis of GPT-4's responses and those that were collected from the survey was done.

4.1 RAG, ReAct and Multi-Agent System

4.1.1 ReAct

As outlined in Chapter 2, the ReAct agent is one that thinks (or reasons) through a solution and then follows an action that was decided during this process. The design and workflow of the agent that was implemented for the purpose of this thesis are as follows:

- Determining the LLM that should be used. Selecting the appropriate Large Language Model (LLM) was critical to ensure a fair evaluation and optimal results. For this purpose, the latest model, GPT-4-O [1], was used.

- Determining the tools that the LLM would need to perform the trade-space analysis. To determine the tools necessary for the LLM to perform the trade-space analysis, several resources were identified and provided. The LLM required access to the SKA1 system document to evaluate trade-offs effectively and to enable retrieval-augmented generation for this purpose, the Chroma vector store was supplied as a retrieval tool for the agent. Additionally, to facilitate the creation of a matrix, a tool allowing the agent to execute Python code was also provided.

- Configuring the prompts. Configuring the prompts involved designing and inputting appropriate descriptions for the tools into the tool executor, ensuring the agent could determine which tool was optimal for each specific action. The prompt template for the agent itself also had to be designed to reflect its capabilities and skills. This is different from the query that we pass to the agent during the execution. The prompts used for each object can be seen in Table 4.1

- Creating the agent executor and passing the query. During the experiment, the entire matrix was initially provided to the agent for completion. However, to achieve a more nuanced analysis, each pairing was subsequently passed to the agent individually to obtain the trade-off score for each specific comparison. Table 4.2 shows the complete query that was passed and one instance of the individual pairing that was passed to the agent.

4.1.2 Multi-Agent System

One of the more intricate tests conducted in this thesis involved the same question posing to a multi-agent system and observing the resulting scores. Implementing the multi-agent approach was a bit more complex because it required determining the number of agents, assigning roles, and deciding how to connect them.

[1] https://platform.openai.com/docs/models/gpt-4o

Table 4.1: Objects and Prompts Used

Object	**Prompt Used**
Tool - Python code execution	Use this to execute python code. The output should be made visible to the user. Use the print statement `print()` to display the result.
Tool - ChromaDB retriever tool	Search and return information of the SKA1_System_Baseline_Design_Document about how each utility works and how they relate to the other utilities.
Agent - ReAct	You are an expert trade space analyst with over 20 years of experience and expertise when it comes to handling requirement analysis. You need to perform a tradespace analysis for the 2 utilities based on the information. Retrieve the information using the tools you have. Give a score from -2, -1, 0, 1, 2 for the tradespace relation between the utilities, with -2 being a negative impact and 2 being a positive one. You have access to the following tools: {tools} Use the following format: Question: the trade space question you must answer Thought: you should always think about what to do Action: the action to take, should be one of [{tool_names}] Action Input: the input to the action Observation: the result of the action ... (this Thought/Action/Action Input/Observation can repeat N times) Thought: I now know the final answer Final Answer: the final answer to the original input question Begin! Question: {input} Thought:{agent_scratchpad}

Knowing the necessary tools that the agents required access to was equally important. The workflow for this approach can be seen in Figure 4.1, and the steps taken to implement it are as follows:

- Determining the LLM that should be used. In order to ensure a fair evaluation, the same LLM that was used for the ReAct approach was chosen for the multi-agent system as well (GPT-4-0).

Table 4.2: User Queries

Query type	**Prompt Used**
Complete	As an expert trade space analyst with over 20 years of experience and expertise when it comes to handling requirement analysis, perform the trade off for the below matrix on a score from -2,-1,0,1,2 based on the SKA1_System_Baseline_Design_Document. Go with the best score based on your experience and knowledge and the information that is present in the document. The columns are: Frequency Range, Sensitivity, Bandwidth, Data Processing Capability, Polarisation Capability, Distribution of Collecting Area, Maximum Baseline, Site Locations, Systematic Error Control, Integration Time. The rows are: Technological Complexity, Scientific Output Quality, Resolution, Error and Noise Management, Operational Efficiency, Sensitivity, Data Volume and Processing, Cost Implications, Geographical Constraints. Then present the trade-off score and the utilities in a matrix. Once you code it up, finish.
Individual	What is the trade-off between Technological Complexity and Sensitivity according to SKA1_System_Baseline_Design_Document?

- Determining the tools that should be used. Again, as the same problem was being presented to the system, in order to ensure fairness, consistency was maintained. Consequently, the agents within the network were given access to the same set of tools: a retriever tool and a Python code executor tool.

- Deciding on the agents and what their roles would be. Several discussions were made regarding how many agents should be used and how they would interact with each other. Finally, after contemplating the issue with different people from different areas, the problem was posed to GPT to come up with a team that was designed to perform a tradespace analysis when a specification document was presented (Table 4.3). From the given response, some of the agents were chosen and after making some modifications to their roles, the workflow, as seen in Figure 4.1, was implemented.

- Choosing a prompt for the system. Prompting plays a significant role when it comes to a multi-agent system. The generic prompt that is used for each agent makes them aware of each other and of the tools that they have access to. The prompt chosen in this experiment resembles what is used by LangGraph in their example notebooks, and can be seen in Figure 4.2.

Overall, the experiments were designed to perform a fair and accurate evaluation, and those factors which could be controlled were kept the same throughout. By using GPT-4-o for both the ReAct approach and the multi-agent system, the study ensures that the comparison remained unbiased. Equipping the agents with the same retriever and Python code executor tools further reinforced this consistency.

Table 4.3: Prompt to get the agents and roles

Prompt for the LLM
I want to create a team to perform tradespace analysis for a specification document to figure the tradeoffs between utilities. What would the team members be? Output just the names and their prompt description and nothing else as comma separated list

4.2 Participants Selection and Categorization

In this study, I attempted to consider individuals from varying backgrounds and how their different levels of experience with requirements and whether they are from a technical or non-technical domain could influence their approach in analyzing and filling out the matrix given the same set of information.

4.2.1 Selection Criteria

The selection of participants was designed to ensure a diverse representation of experiences and backgrounds in relation to requirements engineering. These criteria included educational background, professional experience, and familiarity with requirements engineering concepts (find detailed descriptions below). The primary aim was to include individuals who could provide a wide range of perspectives on the task, thus widening the scope of the survey to get more meaningful results.

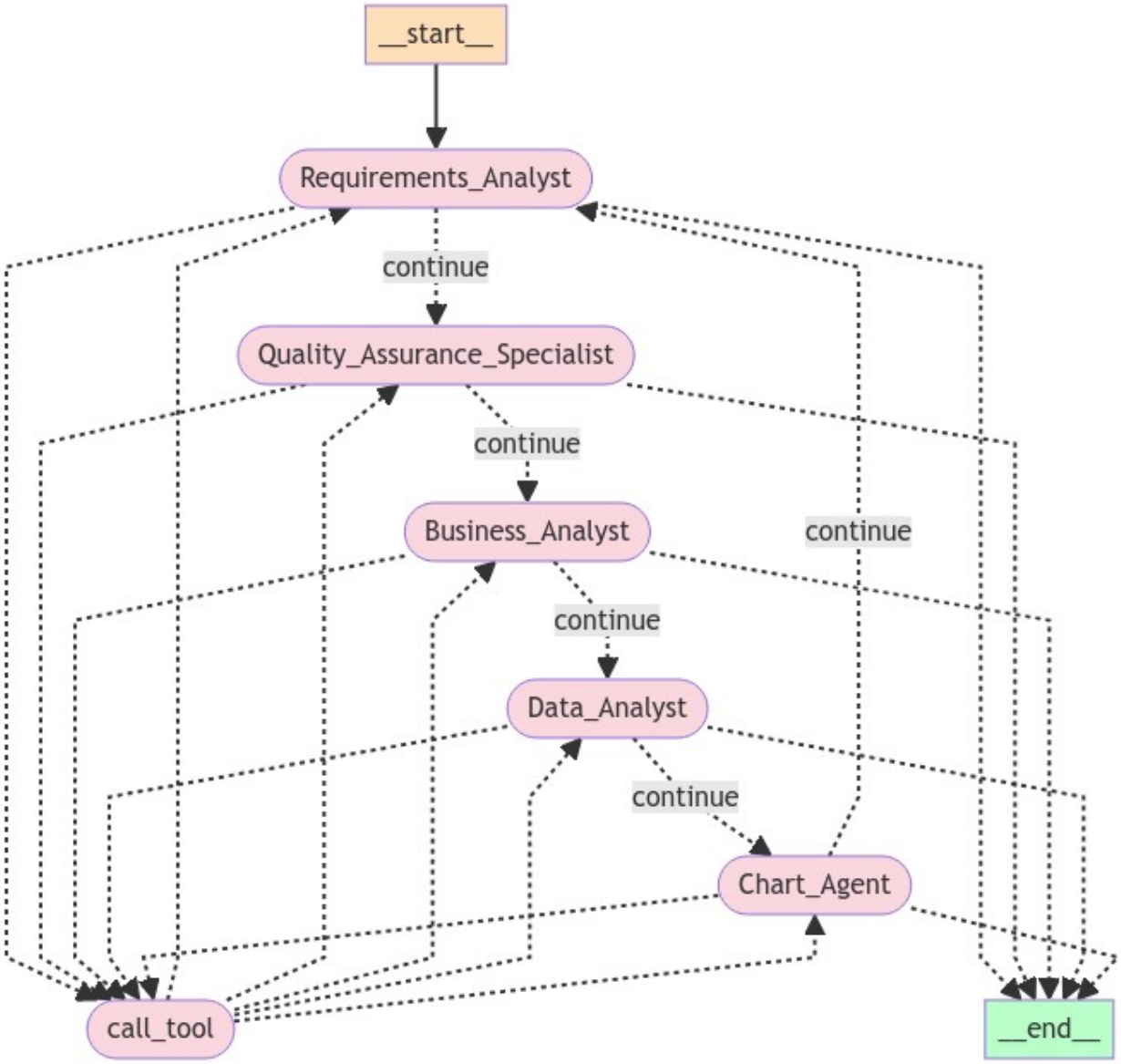


Figure 4.1: The Multi-Agent System

```
"system",
"You are a helpful AI assistant, collaborating with other assistants."
" Use the provided tools to progress towards answering the question."
" If you are unable to fully answer, that's OK, another assistant with different tools "
" will help where you left off. Execute what you can to make progress."
" If you or any of the other assistants have the final answer or deliverable,"
" prefix your response with FINAL ANSWER so the team knows to stop."
" You have access to the following tools: {tool_names}.\n{system_message}",
```

Figure 4.2: Generic Agent Prompt for The Multi-Agent workflow (https://github.com/langchain-ai/langgraph/blob/main/examples/multi_agent/multi-agent-collaboration.ipynb)

- Educational Background: Participants must have completed or be currently enrolled in a bachelor's or master's degree program in engineering, computer science, artificial intelligence, or a non-technical field.
- Professional Experience: For those not currently enrolled in an academic program, professional experience related to their field of study, particularly experiences that involve requirements engineering, software development, or project management, is considered.
- Familiarity with Requirements Engineering: Participants should have varying degrees of familiarity with requirements engineering, from theoretical knowledge acquired through coursework to practical experience gained through professional engagement.

4.2.2 Grouping Methodology

Around 30 participants were chosen to take part in the survey. They were then categorized into three distinct groups as follows:

- Group 1: Engineering Background: This group included individuals that were pursuing or had completed either a bachelor's or master's degree in an engineering domain like manufacturing. It would have individuals who would have a solid foundation in designing systems, and would be

more familiar with requirements documentation, and have a more nuanced understanding of the technical terms, which are critical when it comes to filling out the matrix.

- Group 2: Computer Science / Artificial Intelligence background: Participants in this group would include individuals who are associated with the software field in either a professional or educational capacity. This group would be more familiar with LLMs like ChatGPT and would also be able to apply the trade space that is usually done as part of the software development life-cycle to a different domain.

- Group 3: Non-Technical Background: The third and final group consisted of individuals who are from a non-technical domain, such as humanities, who may not have a formal education in engineering or computer science but would have interacted with Generative AI technologies in some manner in their day-to-day activities. The primary purpose for including this group is to reflect on how the model would fair against individuals who would approach this task without a technical foundation but rather use their intuitive understanding and general problem-solving skills to complete the survey.

4.2.3 Demographic and Background Controls

Acknowledging that demographic and background variables could influence the results, I attempted to have an inclusive group with the aim of isolating the effects of the primary variables under investigation: the educational background and experience level in requirements engineering. The study collected demographic information from the participants, including age, geographic location, educational level, research interest, and familiarity with Generative AI tools. This data was used to assess the diversity of the participants pool and to ensure a broad representation. It also helped account for and mitigate potential biases that may arise from demographic skewness. For example, age-related differences in familiarity with digital tools and professional experiences in technical fields can influence outcomes.

To control for these demographic and background factors, I tried to, where possible, match participants

across the primary groups based on key demographic and background variables, aiming to reduce variability not related to the study's main focus.

4.3 Task Design for Comparative Analysis

This section details the design and execution of the comparative part of the study, outlining the nature of the task and the method by which the responses were collected. The goal is to ensure a comprehensive understanding of the comparative analysis's methodology and to provide a clear basis for evaluating the outcomes.

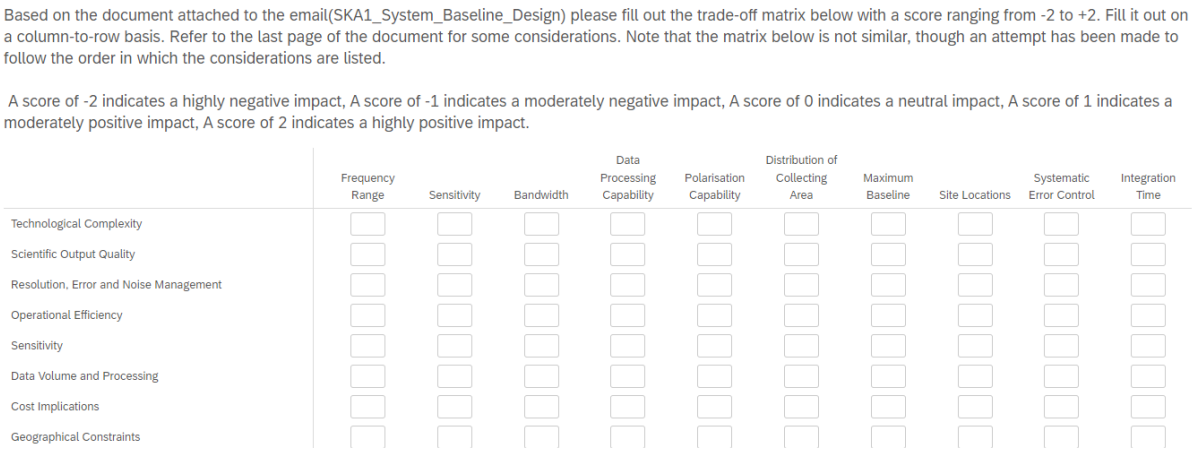
4.3.1 Task Description

Participants in this study were first given a concise 2-page document that presents key information regarding the SKA1 System Baseline Design. This document was designed to impart a foundational understanding of what the SKA1 system would need and what the most important factors for this trade-space analysis would be. Additionally, participants have the option to view two supplementary videos (3mins [2] and 8mins [3] long), which offer further insights into what the SKA1 observatory would do and how it functions.

Following this preparatory phase, the participants were asked to fill in an 8x10 matrix with different factors, using the knowledge acquired from the document and, optionally, the videos. Keeping in mind the different backgrounds from which those filling in the matrix came, the document included a section titled 'Trade-off Considerations', which explained some of the factors that were present and how they could influence the others.

This task was designed to test the various levels of understanding between the three buckets of human

[2] https://www.youtube.com/watch?v=nXcsFWw2qd8
[3] https://www.youtube.com/watch?v=Dc7xRjRvN3Y



Based on the document attached to the email(SKA1_System_Baseline_Design) please fill out the trade-off matrix below with a score ranging from -2 to +2. Fill it out on a column-to-row basis. Refer to the last page of the document for some considerations. Note that the matrix below is not similar, though an attempt has been made to follow the order in which the considerations are listed.

A score of -2 indicates a highly negative impact, A score of -1 indicates a moderately negative impact, A score of 0 indicates a neutral impact, A score of 1 indicates a moderately positive impact, A score of 2 indicates a highly positive impact.

	Frequency Range	Sensitivity	Bandwidth	Data Processing Capability	Polarisation Capability	Distribution of Collecting Area	Maximum Baseline	Site Locations	Systematic Error Control	Integration Time
Technological Complexity										
Scientific Output Quality										
Resolution, Error and Noise Management										
Operational Efficiency										
Sensitivity										
Data Volume and Processing										
Cost Implications										
Geographical Constraints										

Figure 4.3: The matrix that the human participants were asked to fill out

participants and how multiple factors, such as experience as a requirements engineer or experience with tools such as Generative AI, might affect the exercise.

4.3.2 Response Collection Procedures

As this study involved both humans and a GPT model, the procedures for capturing the responses were designed differently, as detailed below:

- Human Participants: After reading the comprehensive summary and watching the videos, the participants were asked to fill out and submit a matrix through a Qualtrics survey. In order to learn more about the survey taker, other information such as their age, interest area, and familiarity with Generative AI was also collected. Although no time constraints were imposed, the amount of time that was taken by each individual was also noted. I believe that this would reflect the ease with which each group approached the matrix.

- Large Language Model: In contrast to the human participants who were given a summarized document, I uploaded a 69-page document detailing the SKA1 System Baseline Design to ChatGPT to perform the trade space analysis. This disparity in document length was intentional and was

designed to test both the model's ability to process and synthesize information from a significantly larger dataset and to measure how rapid its analysis could be. The same matrix that was given in the Qualtrics survey was given as a query to the LLM to fill out. In order to ensure fairness, I also later uploaded the same 2-page document that was given to the other participants and asked ChatGPT to fill in the matrix again.

4.4 Data Analysis Methods

4.4.1 Comparative Methods

In order to evaluate the responses between individuals within a group, across different groups, and in comparison to GPT-4, heatmaps were used to visually represent the information.

- Within-group analysis: Heatmaps were generated for each individual's response within a group, and this was then used to identify the differences in opinions within a homogeneous group.
- Across-group analysis: The responses for each trade-off within a group was averaged and then compared to the other groups in order to visualize how the patterns across groups varied.
- Groups v/s GPT-4: The average heatmaps that were generated for the groups were then compared to the heatmap that was generated for GPT-4's responses. This was done to identify which group(s) had the closest response to GPT-4 and the areas where GPT-4's responses significantly differed from the human responses.

4.4.2 Statistical Methods

- Mean: Calculating the mean of the responses for each group played a pivotal role as this served as the main manner of comparing the responses between the human participants and GPT-4.

$$Mean, \overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

where:

- n is the number of observations,
- x_i is each individual observation

- Standard Deviation : Standard deviation is used to measure the amount of variation of a dataset from the mean. A low standard deviation means that the data points are close to the mean, which implies a similarity in responses within a group. On the other hand, a high standard deviation would mean a wider spread of responses, highlighting diversity in how individuals or GPT-4 approach the tasks. In the context of the study, standard deviation was used to help the consistency of performance within each group and between groups. This is also necessary as having a high or low standard deviation determines if the average is representative of the group and whether it is a reliable measure when I use it to compare the results between the groups and GPT-4.

$$StandardDeviation, \sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

where:

- N is the number of observations,
- x_i is each individual observation,
- $\overline{x}$ is the sample mean.

- ANOVA: Analysis of Variance (ANOVA) is a statistical method that analyzes the difference between the means of three or more groups. In this study, ANOVA was used to calculate the variance between the three human groups compared to that of the GPT model. When running the test, we get two values: F-statistic and p-value. The F-statistic value shows the ratio of the variance, with a high value suggesting that the group means are not equal, while a low value indicates that they are. The p-value, on the other hand, tells us whether the difference is significant or occurred by chance.

Usually, a p-value that is less than 0.05 means that the difference occurred as a result of the values in the data and is relevant. Once the ANOVA results were achieved, a post-hoc analysis was also required as the ANOVA just tells us that one group is different from the others, but not which one. A post-hoc test like Dunnett's test can be used to compare several groups to one particular group, which aligns with the goal of this study.

4.4.3 Interpretation of Results

A large part of understanding the results focused on the heatmaps that were generated for each group and for the model. The visual data representations made it easier to find the trade-offs that were approached similarly, the ones where the groups and/or GPT-4 had differing opinions and helped highlight the overall distribution of the responses.

4.5 Ethical Considerations, Limitations and Assumptions

4.5.1 Human Participants, Ethics, and Fairness

When conducting the research, informed consent was obtained from all the participants, making them fully aware of the nature of the study and what role their responses played in it. Privacy and confidentiality was strictly implemented, and the participants' data was kept anonymous.

4.5.2 Study Limitations

While the aim of this study was to provide a fair comparison between human participants from different backgrounds and LLMs when performing trade space analysis, I fully acknowledge that there are several limitations present.

- Variability in participants: Despite efforts having been taken to categorize participants based on their educational backgrounds and experiences, individual differences in cognitive abilities and prior

knowledge could have affected the results. As the survey was not taken in a controlled environment, motivation and attention to detail might also have influenced the outcome.

- Information provided: A 69-page document was fed to ChatGPT, while a 2-page summary was provided to the human participants, along with the option to watch 2 supplementary videos. This discrepancy in information could potentially bias the outcome, especially considering that the two page document summary was also generated by ChatGPT. To compensate for this, ChatGPT was also asked to fill in the matrix based on the summarized document. Even so, this remains a significant limitation in terms of the information that was accessible to both the human participants and ChatGPT.

- Generalization: This study was performed using the SKA Baseline system design document alone. Using a simpler document, a more complex system or incorporating more documents into the study could bring about different results. The document chosen directly influences the complexity of the task and the participant's ability to comprehend and apply information, which in turn influences the analysis' outcomes.

4.5.3 Assumptions

- The participants within each group are similar in terms of their relevant skills and background. This assumption does not fully account for the individual differences within each group.

- The study assumes that the document and/or the videos provide adequate information to the participants and should be enough for them to fill out the matrix.

Chapter 5

Results and Discussion

The Results and Discussion section of this thesis aims to comprehensively analyze the performance of Large Language Models (LLMs) in comparison to human participants from various backgrounds when it comes to performing tradespace analysis in requirements engineering.

The section is organized to first present the within-group analysis, where individual responses within each homogeneous group are examined to identify variations in opinions. Heatmaps are utilized to visually represent these differences. Next, the across-group analysis compares the averaged responses for each trade-off among the different human groups to understand how patterns of responses vary across groups. Subsequently, the responses from the human groups are compared against those from GPT-4. Average heatmaps generated for each group are compared to GPT-4's heatmap to determine which human group's responses most closely align with those of GPT-4 and to highlight areas of significant divergence. Additionally, statistical analyses, including the calculation of average responses and standard deviations, are presented. A radar graph illustrates the extent of variation in the average responses for each trade-off among the groups and GPT-4.

Through these analyses, this section seeks to uncover insights into the capabilities and limitations of LLMs in performing tasks typically associated with human requirements engineers, and to explore the potential for integrating LLMs into the requirements engineering process.

5.1 Within-Group Analysis

5.1.1 Non-technical Background Group

The trends observed in 10 of the participants are as seen below:

1. Participant 1:

    - Positive Correlations: Notable in Sensitivity, Data Processing Capability, and Site Locations, indicating a strong alignment of these factors with overall project goals.
    - Negative Correlations: Observed in Bandwidth, Maximum Baseline, and Systematic Error Control, particularly under Technological Complexity and Integration Time, suggesting potential conflicts that need to be addressed.

2. Participant 2:

    - Strong Positive Correlations: Found in Technological Complexity and Scientific Output Quality, reflecting a robust link between these critical areas.
    - Neutral/Positive Correlations: Predominant in Resolution, Error and Noise Management, and Cost Implications, showing a balanced trade-off scenario.

3. Participant 3:

    - Mixed Correlations: Positive values in Sensitivity and Data Processing Capability indicate beneficial interactions, whereas higher negative correlations in Cost Implications and Geographical Constraints highlight areas of concern..

4. Participant 4:

    - Positive Correlations: Systematic Error Control and Integration Time show strong positive correlations.

- Negative Correlations: Mixed correlations in other factors with some noticeable red areas indicating conflicts.

5. Participant 5:
    - Positive Correlations: Observed in Bandwidth, Data Processing Capability, and Geographical Constraints, suggesting these areas support project objectives.
    - Negative Correlations: Noted in Sensitivity and Maximum Baseline, requiring further investigation.
6. Participant 6:
    - Higher Positive Correlations: Frequency Range and Integration Time exhibit strong positive correlations.
    - Mixed Correlations: Observed in Scientific Output Quality and Operational Efficiency.
7. Participant 7:
    - Positive Correlations: Strong in Scientific Output Quality and Integration Time, indicating these factors are well-aligned.
    - Generally Low Correlations: Across other factors, suggesting a balanced or neutral dataset.
8. Participant 8:
    - Positive Correlations: In Sensitivity, Data Processing Capability, and Site Locations.
    - Negative Correlations: In Integration Time and Geographical Constraints.
    - Mixed Correlations: In Scientific Output Quality and Technological Complexity.
9. Participant 9:
    - Strong Positive Correlations: In Technological Complexity, Scientific Output Quality, and Systematic Error Control.

- Negative Correlations: In Integration Time and Data Volume and Processing.
- Mixed Correlations: In Operational Efficiency and Sensitivity.

10. Participant 10:

    - Positive Correlations: In Data Processing Capability and Sensitivity.
    - Negative Correlations: Significant in Geographical Constraints and some technological complexity areas.

Observations

The significant patterns and trends withing this group are:

- Concerns About Technological Complexity: There is a mix of positive and negative responses, but the trend shows a general discomfort or concern about handling complex technologies.
- Cost Implications: Continues to be a significant concern, with many participants showing negative responses. This indicates that cost is a critical trade-off factor for this group.
- Geographical Constraints: Generally negative responses, indicating concerns about the practicality of geographical constraints.
- Positive Focus on Operational Efficiency and Error Management: Many participants showed positive responses towards operational efficiency and error management, highlighting these as areas of importance and confidence.
- Varied Opinions on Data Processing Capability and Sensitivity: There are mixed responses regarding data processing capability and sensitivity, indicating varied levels of confidence or priority placed on these aspects within the group.
- Confidence in Scientific Output Quality and Systematic Error Control: Positive responses indicate that participants believe in their ability to maintain high scientific output quality and control systematic errors effectively.

Overall, the within-group analysis of the non-technical background group reveals significant variation in opinions on technological complexity and cost implications. These areas are sources of concern for many participants. Conversely, there is a notable confidence in operational efficiency, error management, and scientific output quality. The mixed responses on data processing capability and sensitivity reflect the diverse perspectives and priorities within the group.

5.1.2 Engineering Background Group

1. Participant 1
    - Positive Correlations: In Bandwidth, Distribution of Collecting Area, and Integration Time.
    - Negative Correlations: Geographical Constraints.
    - Mixed Correlations: In Frequency Range, Sensitivity, Data Processing Capability, Polarisation Capability, and Systematic Error Control.
2. Participant 2
    - Positive Correlations: In Data Processing Capability and Distribution of Collecting Area.
    - Negative Correlations: In Sensitivity, Site Locations, and Systematic Error Control.
    - Mixed Correlations: In Frequency Range, Scientific Output Quality, Resolution, Error and Noise Management, and Cost Implications.
3. Participant 3
    - Positive Correlations: In Maximum Baseline, Site Locations, and Systematic Error Control.
    - Negative Correlations: In Sensitivity, Cost Implications, and Geographical Constraints.
    - Mixed Correlations: In Frequency Range, Bandwidth, and Data Volume and Processing.
4. Participant 4

- Positive Correlations: In Data Processing Capability, Polarisation Capability, and Integration Time.
- Negative Correlations: In Resolution, Error and Noise Management, and Sensitivity.
- Mixed Correlations: In Frequency Range, Scientific Output Quality, Operational Efficiency, and Cost Implications.

5. Participant 5
    - Positive Correlations: In Scientific Output Quality, Sensitivity, and Systematic Error Control.
    - Negative Correlations: In Technological Complexity, Resolution, Error and Noise Management, and Geographical Constraints.
    - Mixed Correlations: In Frequency Range, Bandwidth, and Data Volume and Processing.
6. Participant 6
    - Positive Correlations: In Frequency Range, Scientific Output Quality, and Integration Time.
    - Negative Correlations: In Cost Implications.
    - Mixed Correlations: In Sensitivity, Data Processing Capability, and Polarisation Capability.
7. Participant 7
    - Positive Correlations: In Frequency Range, Scientific Output Quality, and Operational Efficiency.
    - Negative Correlations: Geographical Constraints.
    - Mixed Correlations: In Sensitivity, Data Volume and Processing, and Cost Implications.
8. Participant 8
    - Positive Correlations: In Data Processing Capability, Distribution of Collecting Area, and Systematic Error Control.

- Negative Correlations: In Technological Complexity, and Cost Implications.
- Mixed Correlations: In Frequency Range, Scientific Output Quality, and Sensitivity.

9. Participant 9

    - Positive Correlations: In Technological Complexity, Frequency Range, and Polarisation Capability.
    - Negative Correlations: In Sensitivity, and Data Volume and Processing.
    - Mixed Correlations: In Scientific Output Quality, and Cost Implications.

Observations

The significant patterns within the group are:

- Emphasis on Scientific Output Quality: Many participants see positive correlations with Scientific Output Quality, underlining its critical role in the project's success.
- Geographical Constraints as a Negative Factor: Multiple participants view Geographical Constraints negatively, indicating it is a common challenge.
- Varied Opinions on Frequency Range: Mixed correlations suggest that Frequency Range is seen differently by participants, showing it is a contentious factor.
- Cost Implications as a Concern: The negative correlation with Cost Implications highlights the financial constraints and considerations important to the participants.
- Data Processing Capability's Positive Impact: Recognized positively by several participants, it suggests its importance in achieving project goals.

The overall trends in this group shows that Scientific Output Quality is widely seen as crucial, while Geographical Constraints and Cost Implications are major concerns. Data Processing Capability is positively viewed for its role in technological success. Opinions on Frequency Range are mixed, reflecting diverse priorities.

5.1.3 CS/AI Background Group

The trends observed in 9 of the participants are as seen below:

1. Participant 1:
    - Positive Correlations: In Frequency Range, Integration Time, and Systematic Error Control.
    - Negative Correlations: In Scientific Output Quality, Sensitivity, and Geographical Constraints.
    - Mixed Correlations: In Operational Efficiency and Data Volume and Processing.
2. Participant 2:
    - Positive Correlations: In Data Processing Capability, Distribution of Collecting Area, and Site Locations.
    - Negative Correlations: In Sensitivity, Cost Implications, and Integration Time.
    - Mixed Correlations: In Technological Complexity and Resolution, Error and Noise Management.
3. Participant 3:
    - Positive Correlations: In Technological Complexity, Systematic Error Control, and Integration Time.
    - Negative Correlations: In Sensitivity, Geographical Constraints, and Operational Efficiency.
    - Mixed Correlations: In Scientific Output Quality and Data Volume and Processing.
4. Participant 4:
    - Positive Correlations: In Technological Complexity, Integration Time, and Polarisation Capability.

- Negative Correlations: In Data Volume and Processing, Geographical Constraints, and Site Locations.
- Mixed Correlations: In Scientific Output Quality and Sensitivity.

5. Participant 5:
   - Positive Correlations: In Technological Complexity, Data Processing Capability, and Systematic Error Control.
   - Negative Correlations: In Cost Implications, Geographical Constraints, and Resolution, Error and Noise Management.
   - Mixed Correlations: In Operational Efficiency and Scientific Output Quality.
6. Participant 6:
   - Positive Correlations: In Technological Complexity, Frequency Range, and Systematic Error Control.
   - Negative Correlations: In Cost Implications, Sensitivity, and Geographical Constraints.
   - Mixed Correlations: In Operational Efficiency and Scientific Output Quality.
7. Participant 7:
   - Positive Correlations: In Sensitivity, Maximum Baseline, and Distribution of Collecting Area.
   - Negative Correlations: In Cost Implications, Geographical Constraints, and Integration Time.
   - Mixed Correlations: In Scientific Output Quality and Technological Complexity.
8. Participant 8:
   - Positive Correlations: In Sensitivity, Data Processing Capability, and Site Locations.
   - Negative Correlations: In Integration Time and Geographical Constraints.

- Mixed Correlations: In Scientific Output Quality and Technological Complexity.

9. Participant 9:

    - Positive Correlations: In Technological Complexity, Systematic Error Control, and Distribution of Collecting Area.
    - Negative Correlations: In Geographical Constraints, Cost Implications, and Scientific Output Quality.
    - Mixed Correlations: In Operational Efficiency and Sensitivity.

Observations

The significant patterns and trends within this group are:

- There is significant variation in opinions on technological complexity. Some participants view it as a highly positive factor, while others see mixed or neutral impacts.
- Cost implications are another area with notable variation. Some participants highlight it as a significant negative factor, while others show mixed or less pronounced impacts.
- There is notable confidence in systematic error control across most participants. It is consistently seen as a positive factor in ensuring the quality and reliability of the project.
- Data processing capability is generally viewed positively, reflecting its importance in handling the large volumes of data associated with the project.
- Confidence in site locations is also high among participants, indicating a consensus on the strategic importance of site selection for the project's success.
- The responses on scientific output quality are mixed, with some participants showing strong positive correlations, while others display more neutral or mixed views.

- Sensitivity shows mixed responses, reflecting diverse perspectives on its impact. While some see it as a positive factor), others show neutral or varied impacts.
- Operational efficiency is generally viewed positively, but with some variation in the extent of its impact.
- Geographical constraints are seen as a negative factor by several participants, indicating concerns about the challenges associated with site selection and infrastructure).
- Integration time shows both positive and negative correlations, indicating varying impacts on different aspects of the project. Some participants highlight its importance in operational efficiency , while others see it as a constraint .

Overall, the CS/AI group displays significant variation in opinions on technological complexity and cost implications, while showing confidence in systematic error control, data processing capability, and site locations. The mixed responses on scientific output quality and sensitivity reflect the diverse perspectives and priorities within the group, with additional concerns around geographical constraints and integration time.

5.2 Across-Group Analysis - Average

Figure 5.1 shows the average of the responses for the 3 groups (engineering background, CS/AI background and non-technical background).

5.2.1 Similarities

1. High consistency in some areas: All three groups show consistently high values in certain trade-offs, indicating shared importance or preference areas. For instance:

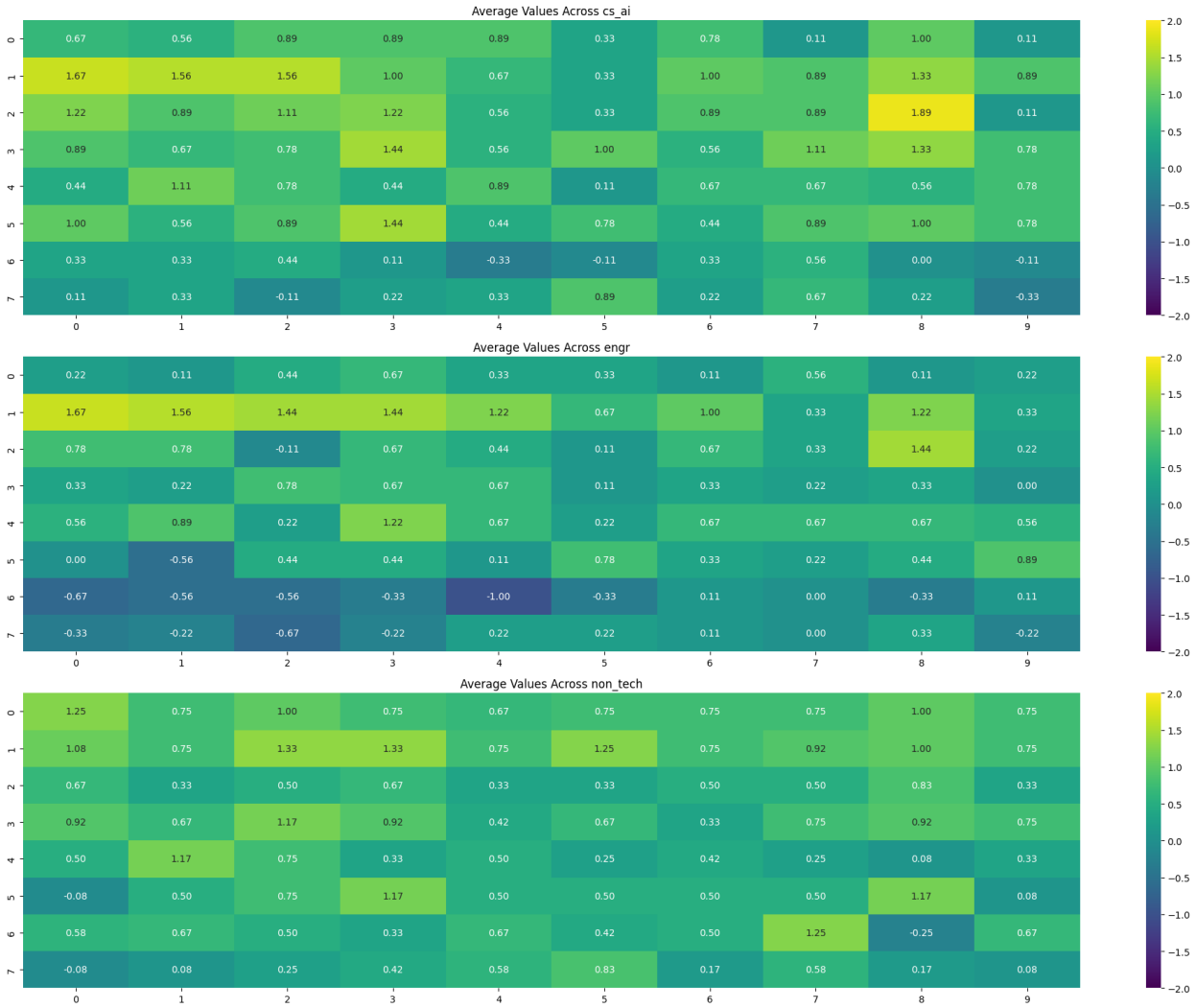


Figure 5.1: Average across the 3 human groups

- All three groups have high values for the second column (corresponding to (Sensitivity) especially in the first two rows (Technological Complexity) and (Scientific Output Quality), showing that they all believe that this trade-off has a positive relation.

2. The negative values and moderate score are spread across different columns for the different groups, and the only similarity that can be drawn is that each group believes that not all the factors can or should be prioritized. However it can be seen that the last row seems to have consistently lower values across all the groups.

5.2.2 Differences

1. High variation in Columns 8 (Systematic Error Control) and 9 (Integration Time). The CS/AI groups shows significantly higher values in those two columns when compared to the other two groups.

2. The non-technical background groups has more consistently positive values across most columns when compared to the groups with a CS/AI or Engineering background.

5.2.3 Reasoning

The heatmap analysis reveals distinct trade-offs between the CS/AI, Engineering, and non-technical groups. CS/AI emphasizes high technological complexity and sensitivity, and would prefer to investment in a better data processing capability, leading to higher costs. In contrast, Engineering adopts a balanced approach, focusing on operational efficiency and practical error management, with moderate costs. The non-technical group focuses on practical applications, preferring to have lower sensitivity and data volume handling if the implementation costs are also lower. These differences reflect each domain's approach to the implementation and how their focuses can differ.

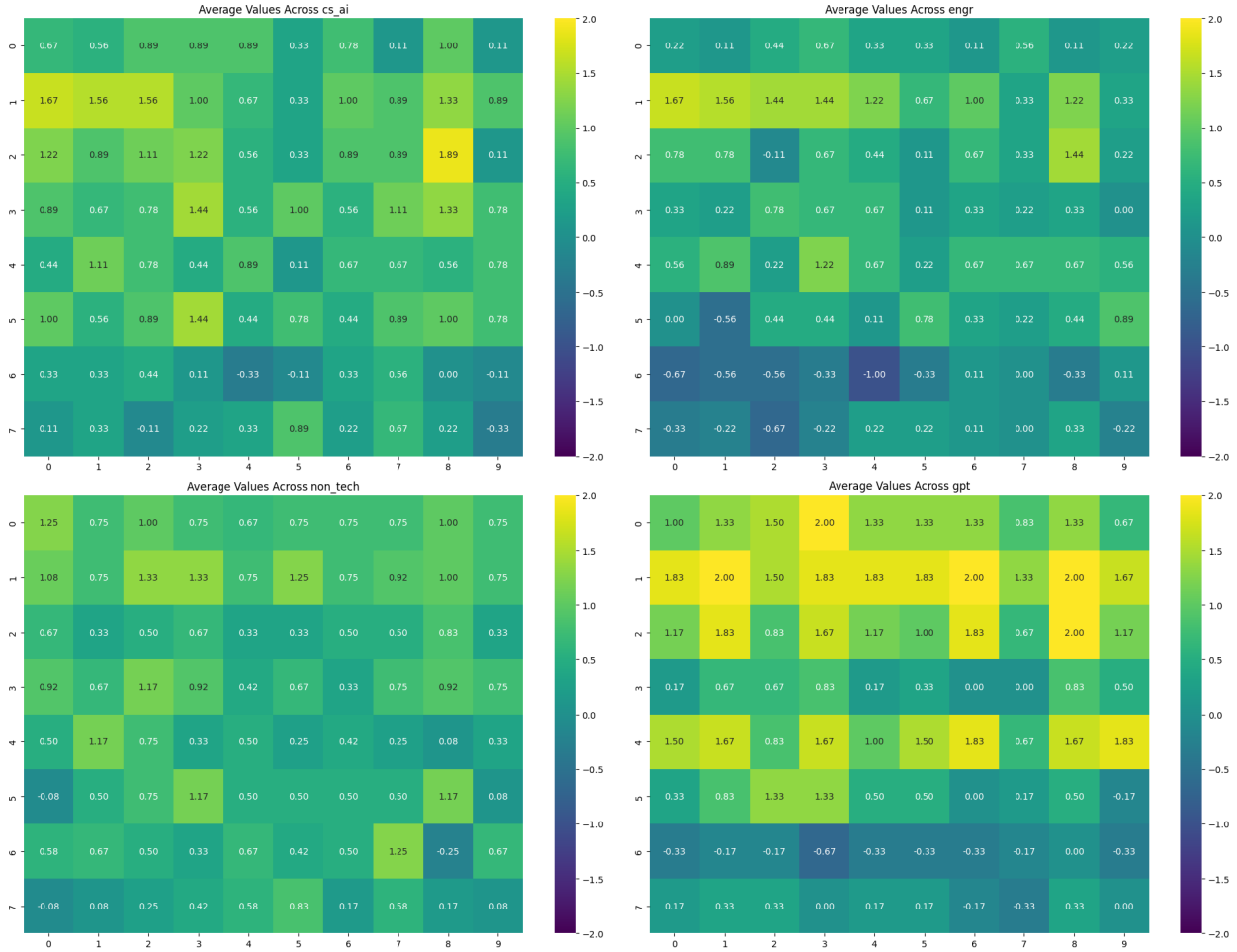


Figure 5.2: Average across human groups and GPT

5.3 Human Groups vs. GPT-4 Analysis

Figure 5.2 shows the average of the responses collection from all the participants and that of the ones that was obtained from the pipelines implemented with GPT.

5.3.1 Comparison of Average Heatmaps

Analyzing the data showed that the group with the CS/AI background had the closest responses to GPT. This was determined by averaging the differences between all the responses. The average difference for CS/AI was 0.547, compared to 0.599 for the engineering group and 0.620 for the non-technical group. This could be because those with a computer science or artificial intelligence background would take a more similar approach to interpreting and responding to the survey like GPT would due to their familiarity with technology and logical processing.

5.3.2 Significant Differences Between Human Groups and GPT-4

Using the ANOVA method, the areas with the largest differences were identified for each group and can be seen in Table 5.1.

Table 5.1: Top 5 Differences for Each Background

Background	Metric Comparison	stat
CS/AI Background	Sensitivity - Distribution of Collecting Area	4.5354
	Scientific Output Quality - Polarisation Capability	3.6303
	Scientific Output Quality - Distribution of Collecting Area	3.4526
	Scientific Output Quality - Maximum Baseline	3.4205
	Sensitivity - Maximum Baseline	3.0532
Engineering Background	Resolution, Error and Noise Management - Maximum Baseline	3.6303
Non-Tech Background	Resolution, Error and Noise Management - Maximum Baseline	4.4164
	Sensitivity - Integration Time	4.3818
	Sensitivity - Maximum Baseline	3.6296
	Sensitivity - Distribution of Collecting Area	3.5921
	Scientific Output Quality - Maximum Baseline	3.4816

Table 5.1 highlights the areas were the model's responses and the human's responses were different.
For the CS/AI group, significant discrepancies were noted for trade-offs like "Sensitivity - Distribution of Collecting Area" and "Scientific Output Quality-Distribution of Collecting Area". This could be because the group has a deeper understanding of the technical nuances and practical implications of these

utilities, especially when considering that "Scientific Output Quality" is considered multiple times. This in turn allows them to consider factors regarding the practical application of these variables, which GPT, despite its advanced capabilities, might not consider. However, there were several metrics where the CS/AI group and GPT were similar, such as "Cost Implications-Polarisation Capability" and "Resolution, Error and Noise Management-Systematic Error Control". These similarities could be reflecting the ability of both groups to rely on logical reasoning and structured data analysis, implying that when deep technical expertise is not required, the results are comparable.

The engineering group showed the most notable differences can be seen in "Resolution, Error and Noise Management - Maximum Baseline" (3.6303). This could be a reflection of a more hands-on perspective, with a focus on the more practical aspects of data processing and noise management. The tendency of an engineer to focus on the practical feasibility and efficiency of technological implementation would be different from GPT's theoretical and generalized approach which would not fully consider the constraints that are present in a real-world scenario.

The non-technical background group seems to be more focused on trade-offs that include "Sensitivity". Individuals in this group had varying backgrounds and experiences which probably influenced them to consider factors that went beyond the technical specifications. Their diverse perspectives and that their difference with GPT (which was prompted to think from a technical view point) were numerically higher, showing the importance of how non-technical contributors can play a significant role when evaluation solutions as they bring with them a new view point.

5.4 Statistical Analysis

To look at the responses of the three distinct groups—Engineers, Non-technical, and CS/AI backgrounds—alongside GPT-4 from a statistical perspective, I consider the standard deviation in each group. Using radar graphs, the variations in their responses across different parameters such as Sensitivity, Data Volume and Processing, and Geographical Constraints are observed. I attempt to identify which group GPT shows the most similarity with.

5.4.1 Radar Graph Analysis

To show case the difference observed, we highlight three utilities - Technological Complexity, Sensitivity and Operational Efficiency.

1. From Figure 5.3 which shows the trade-off between "Technological Complexity" with the other design variables, the following trends can be observed:

    - Engineering Background: The group shows a greater difference in standard deviation compared to the GPT model. Especially when it comes to "Bandwidth" and "Sensitivity". This could be an indication that engineers from different backgrounds (but in the same field) would have more variable and specialized responses as post to a general assessment that GPT would do.
    - Non-technical Background: This group has the least difference with GPT. This indicates that the variability in GPT's responses is similar to that of the people that come from varying background.

2. Figure 5.4 shows a trend where the GPT model resembles the variability in the non-technical group, followed by the CS/AI group and the least with the engineering group. This in essence means that GPT can effectively be used to represent the non-technical group and maybe even the CS/AI group, but not the engineering group.

3. A change in the trend is seen in Figure 5.5 where the CS/AI group shows the highest similarity to the model. In this case the CS/AI group can be effectively represented by the mode, but the other two groups do not show the same degree of variance as the model.

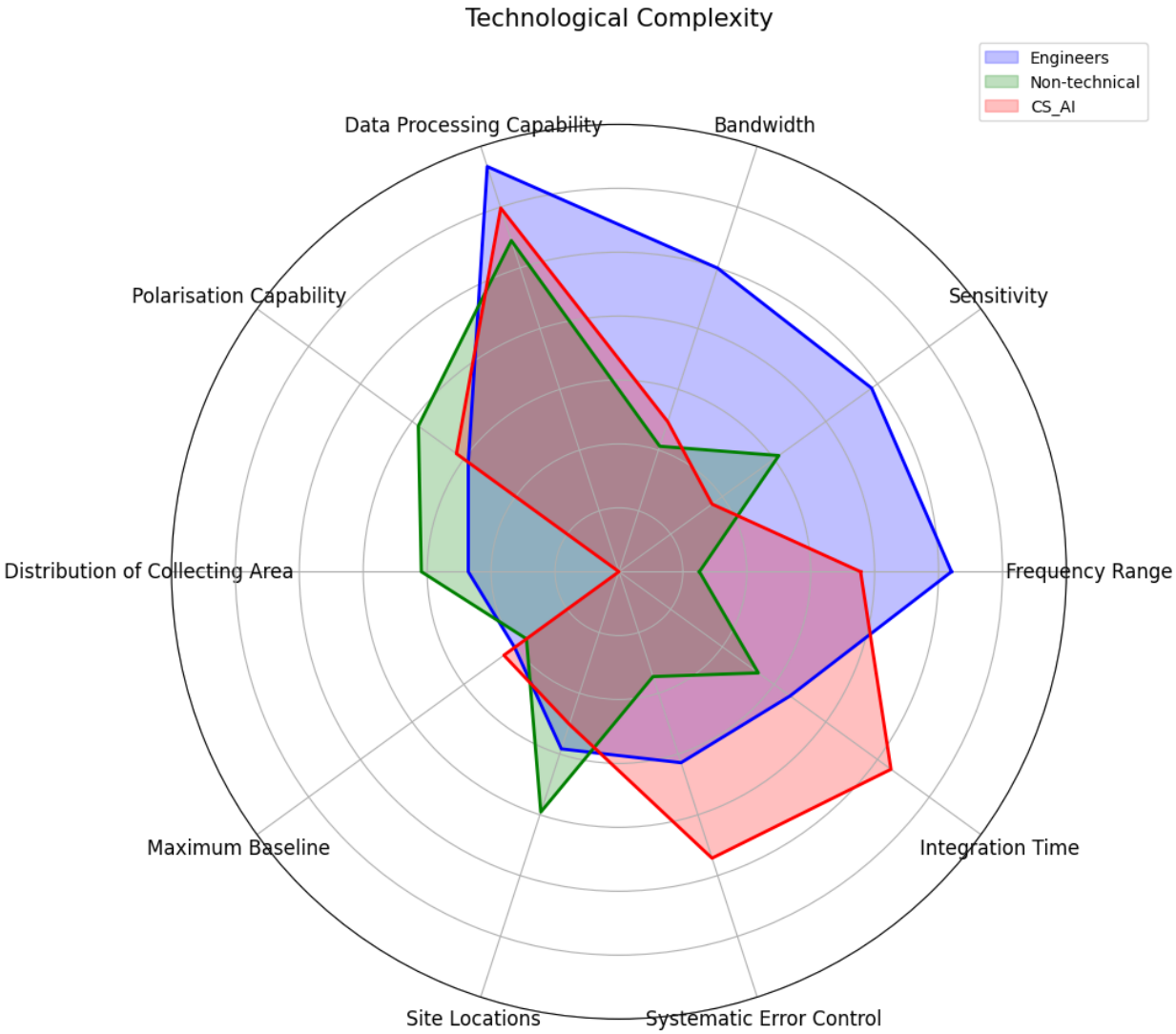


Figure 5.3: Technological Complexity

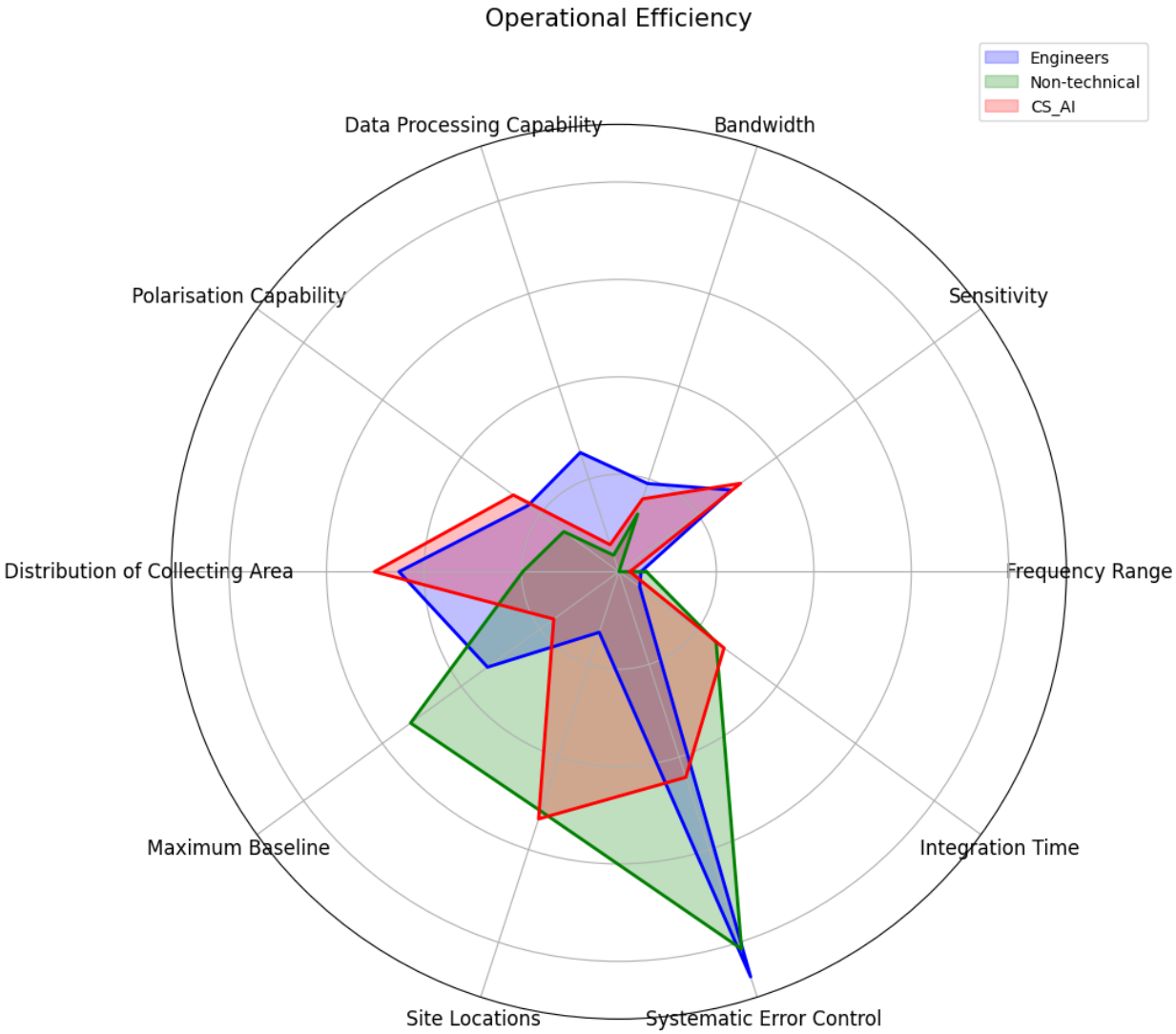


Figure 5.4: Operational Efficiency

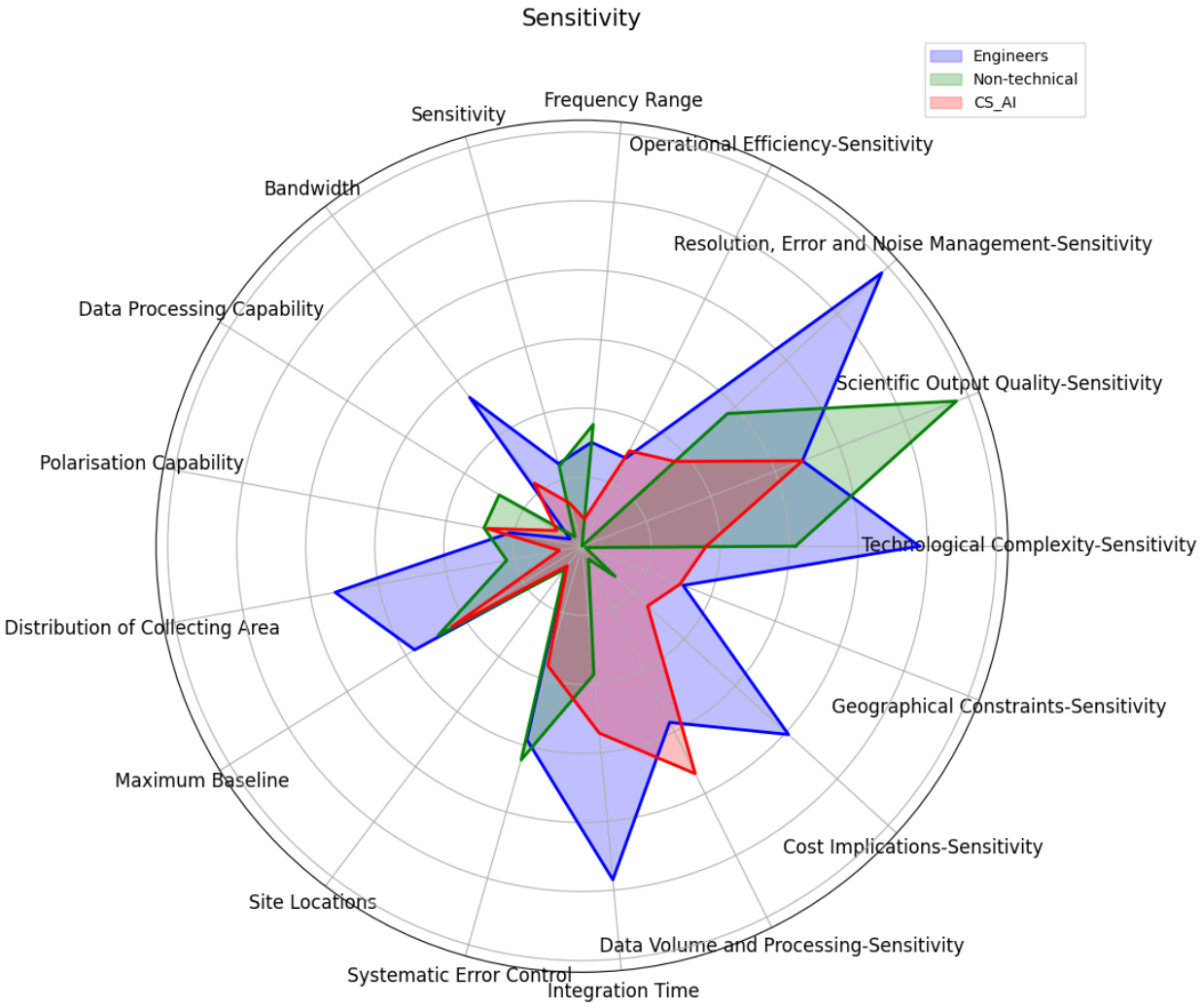


Figure 5.5: Sensitivity

Chapter 6

Conclusion and Future Work

6.1 Conclusion

Trade space analysis is a crucial step in requirements engineering, helping stakeholders to easily visualize and compare the different design alternatives based on multiple criteria such as cost, performance, and risk. By visualizing trade-offs, they can better understand the implications of their choices, leading to optimal solutions. This method is the focus of this study, where I assessed the performance of human experts from various backgrounds compared to Large Language Models (LLMs) like GPT-4 in understanding and evaluating requirements.

The study highlights the importance of interdisciplinary teams in requirements engineering. The diverse backgrounds of the selected participants, engineering, non-technical and CS/AI, provided unique perspectives and insights, contributing to a more comprehensive understanding of the requirements. Engineers might focus on technical feasibility, non-technical experts might emphasize user experience and practicality, while CS/AI specialists might bring in the latest technological advancements. This diversity ensures that various aspects of the problem are considered and no critical factors are overlooked.

However not every organization would have the resources to hire or organize such an interdisciplinary team. A limited budget might cause an organization to struggle to maintain the team for the duration of the project as well. In addition to this, ensuring coordination between individuals from diverse back-

grounds can be complex and time consuming.

Large Language Models (LLMs) like GPT-4 offer a promising solution to these challenges. Through this study, I observed that GPT-4's responses aligned most closely with individuals having a CS/AI background. The study shows how LLMs can be used to augment human expertise, especially when operating in a limited environment. LLMs can provide cost-effect expertise, and can be used to generate the initial analysis and recommendation, which can then be, if required, refined by human experts.

The findings of the study show the need for a balanced approach that integrates both AI and human expertise. It also highlights the importance of incorporating diverse human perspectives in the evaluation of AI systems to ensure comprehensive and balanced decision-making processes. Understanding where AI responses diverge from human judgment can guide the refinement of AI models to better align with human expectations and improve their applicability in real-world scenarios. By acknowledging these differences, we can better appreciate the value of human expertise and judgment in areas where AI models may have limitations. This understanding is crucial for developing AI systems that complement human decision-making, leading to more effective and reliable outcomes in requirements engineering and beyond.

6.2 Future Work

The integration of LLMs in requirements engineering for tradespace analysis represents a significant advancement, but it is just the beginning. Future research can focus on:

- Improving AI Interpretability: If the reasoning behind the scores that the model(s) assigned to the trade-off was known, it would be easier to understand and trust the decisions that were taken.
- Refining AI-Human collaboration: Maybe implementing a Human-in-the-loop approach to the system so that the model periodically takes feedback from the human could help improve the scores.
- Expanding Use Cases: Exploring the applicability of LLMs in other domains of engineering and decision-making to further validate their utility.

Appendix A

Responses for CS/AI Group

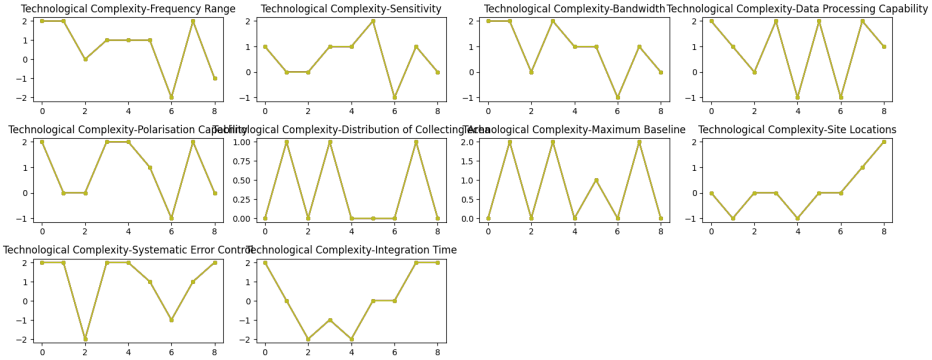


Figure A.1: CS/AI Comparison for: Technological Complexity

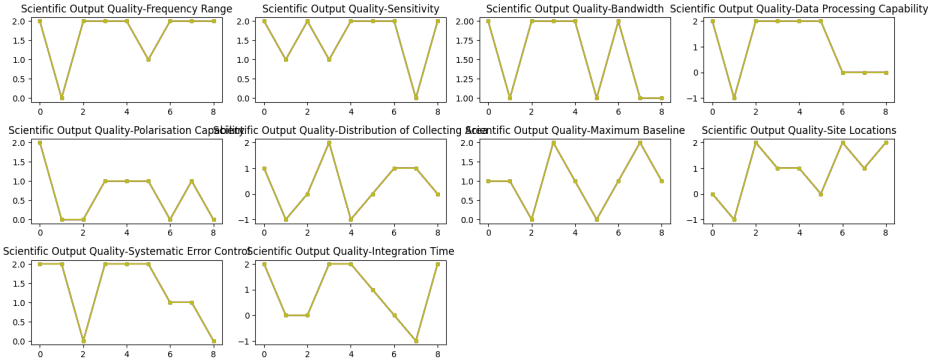


Figure A.2: CS/AI Comparison for: Scientific Output Quality

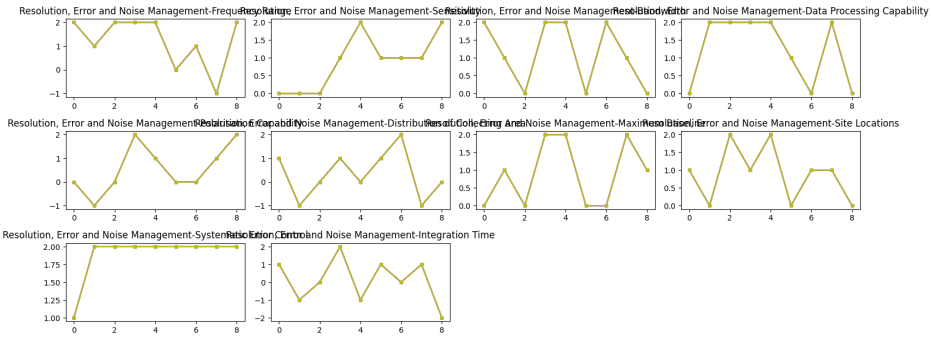


Figure A.3: CS/AI Comparison for: Resolution, Error and Noise Management

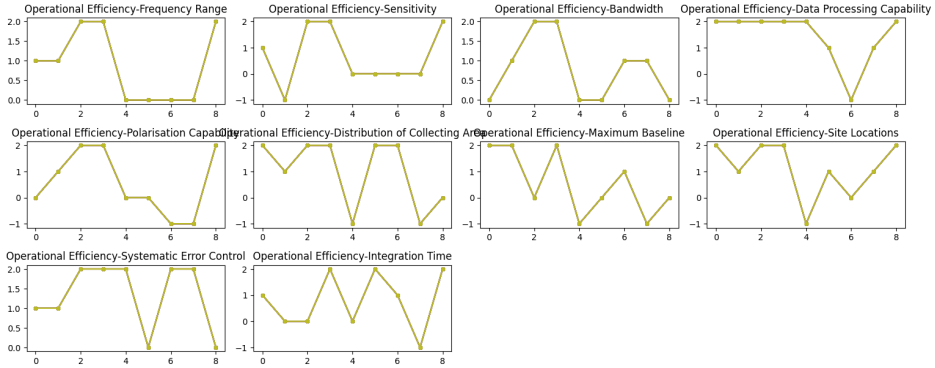


Figure A.4: CS/AI Comparison for: Operational Efficiency

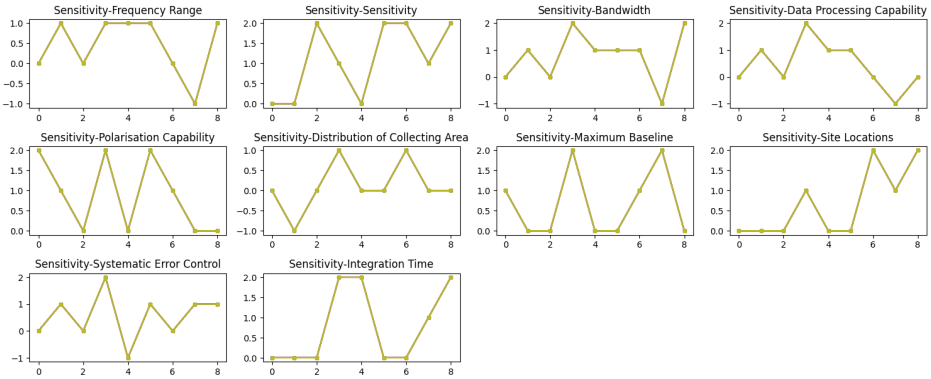


Figure A.5: CS/AI Comparison for: Sensitivity

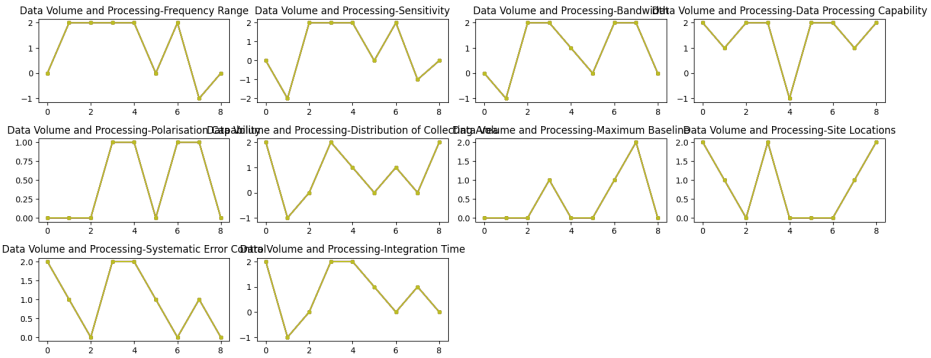


Figure A.6: CS/AI Comparison for: Data Volume and Processing

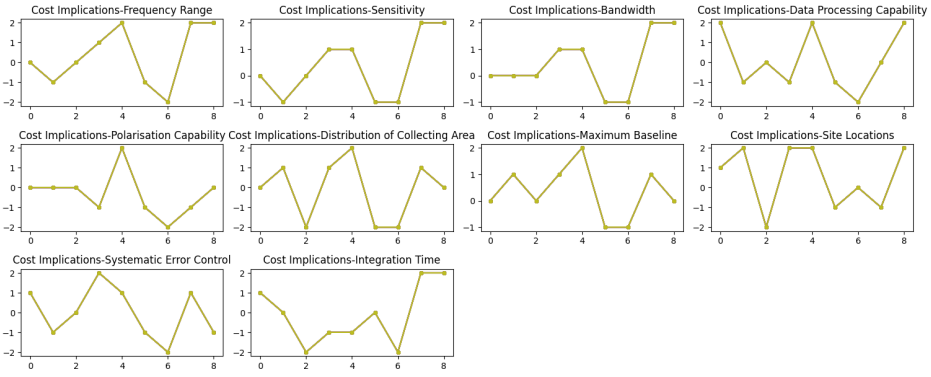


Figure A.7: CS/AI Comparison for: Cost Implications

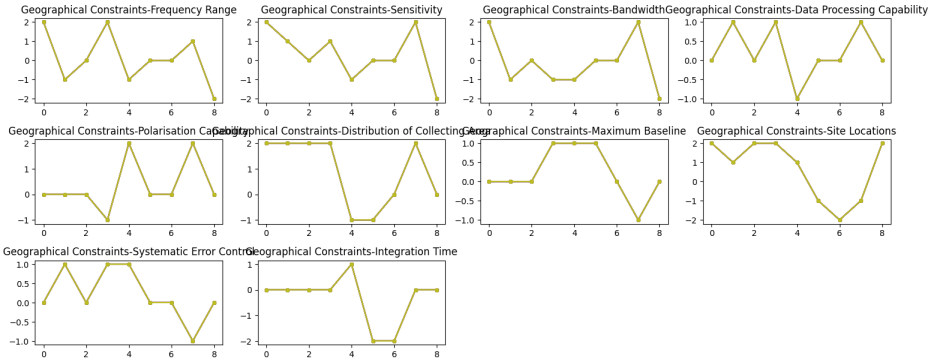


Figure A.8: CS/AI Comparison for: Geographical Constraints

Appendix B

Responses for Engineering Group

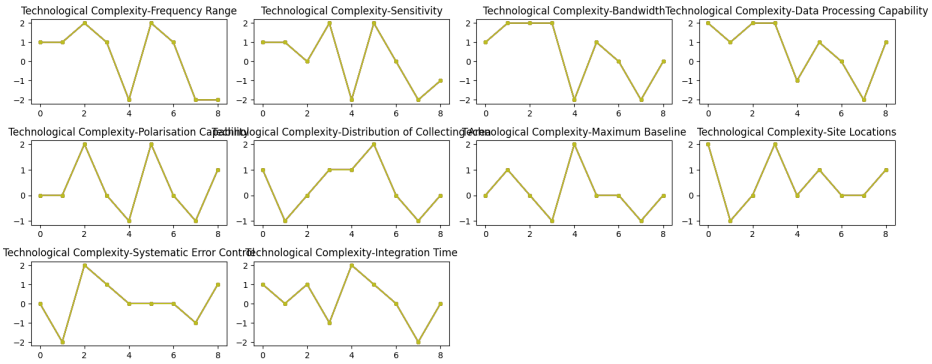


Figure B.1: Engineering Comparison for: Technological Complexity

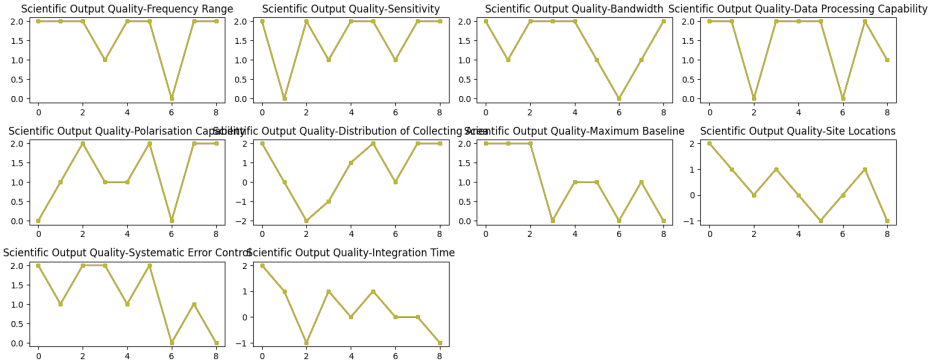


Figure B.2: Engineering Comparison for: Scientific Output Quality

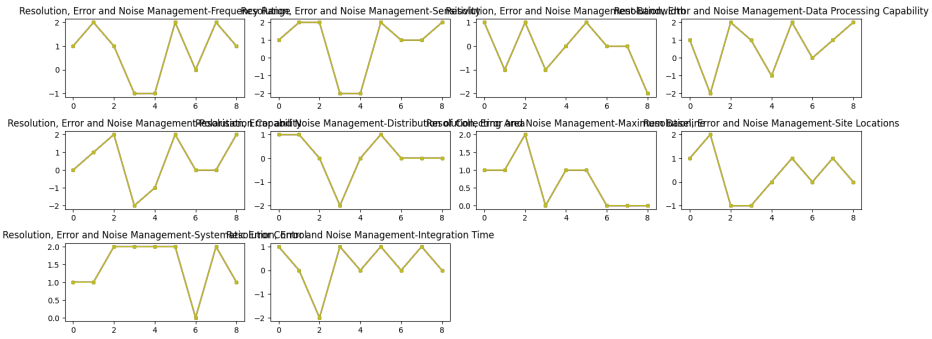


Figure B.3: Engineering Comparison for: Resolution, Error and Noise Management

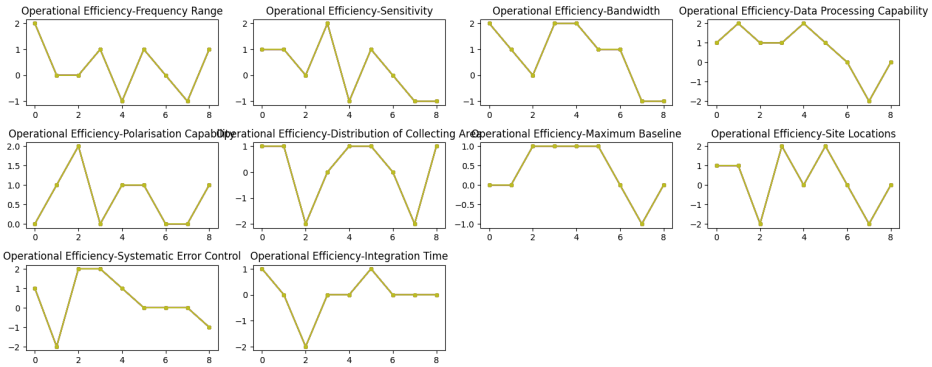


Figure B.4: Engineering Comparison for: Operational Efficiency

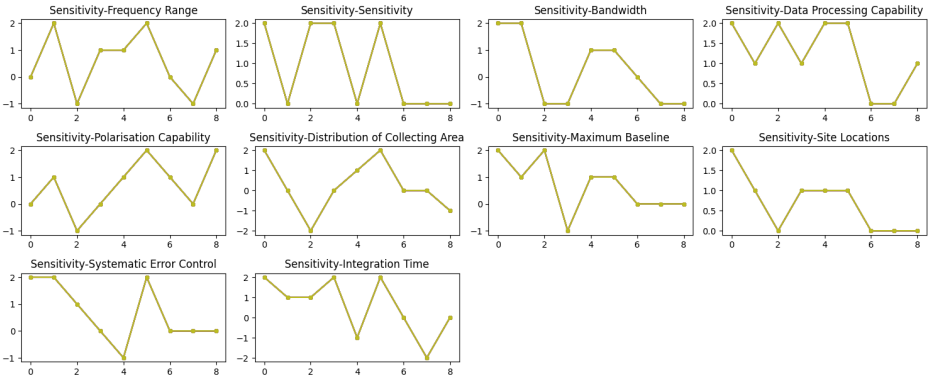


Figure B.5: Engineering Comparison for: Sensitivity

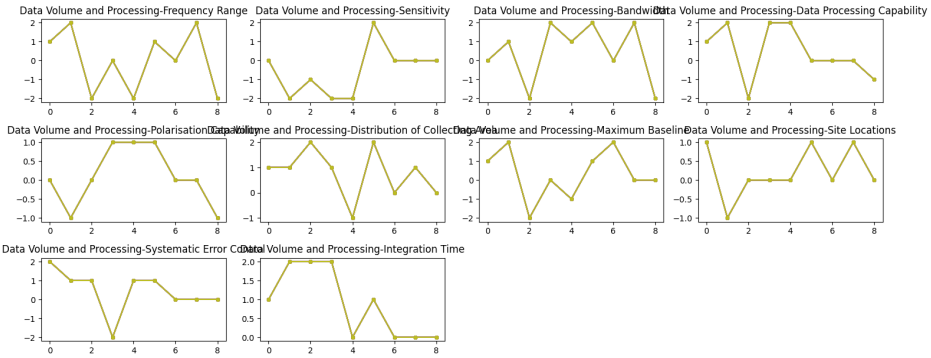


Figure B.6: Engineering Comparison for: Data Volume and Processing

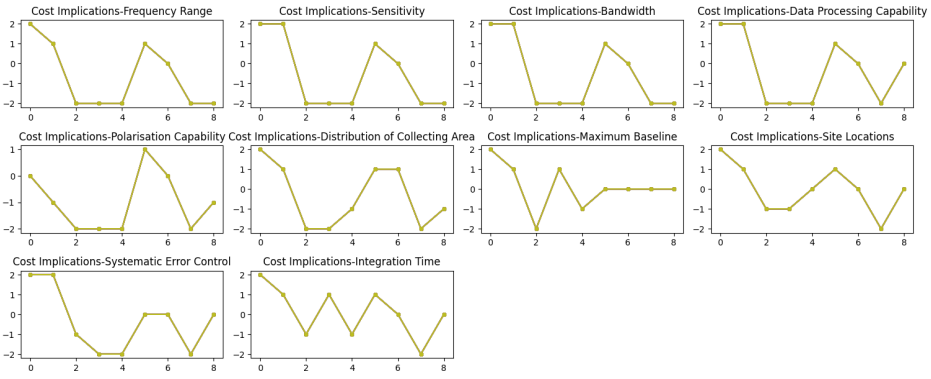


Figure B.7: Engineering Comparison for: Cost Implications

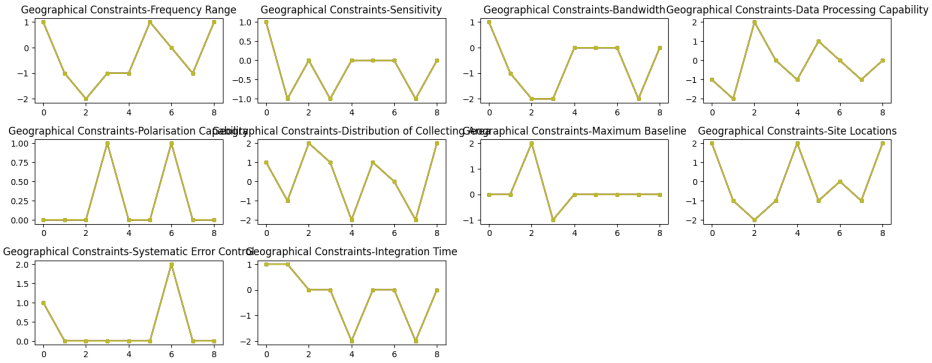


Figure B.8: Engineering Comparison for: Geographical Constraints

Appendix C

Responses for Non Technical Group

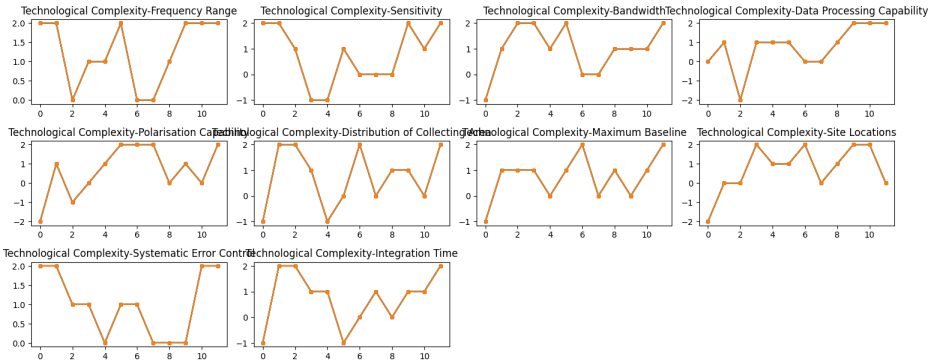


Figure C.1: Non Technical Comparison for: Technological Complexity

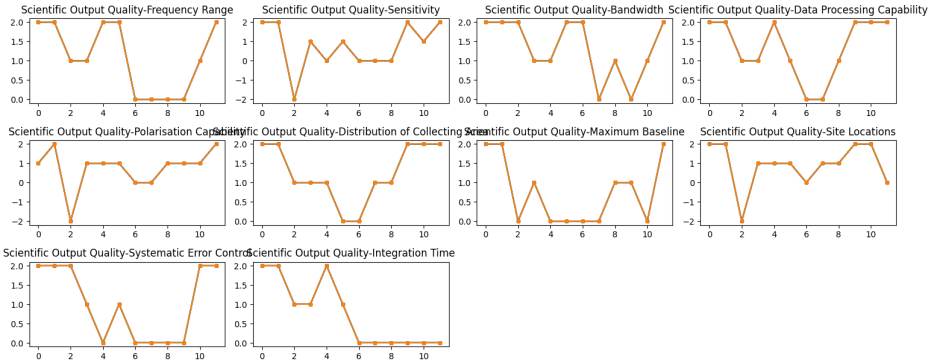


Figure C.2: Non Technical Comparison for: Scientific Output Quality

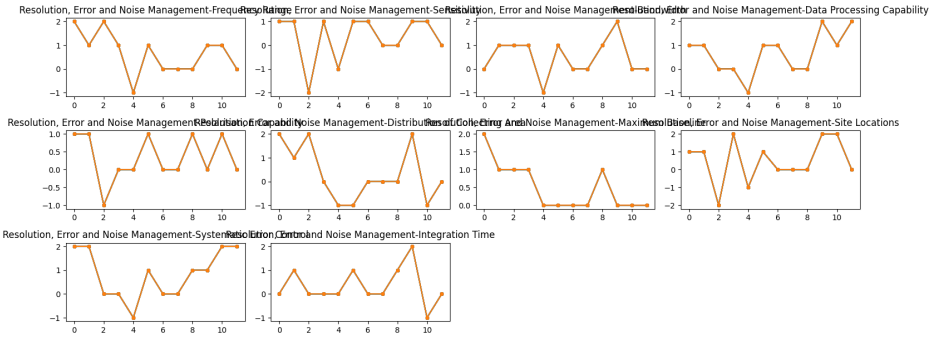


Figure C.3: Non Technical Comparison for: Resolution, Error and Noise Management

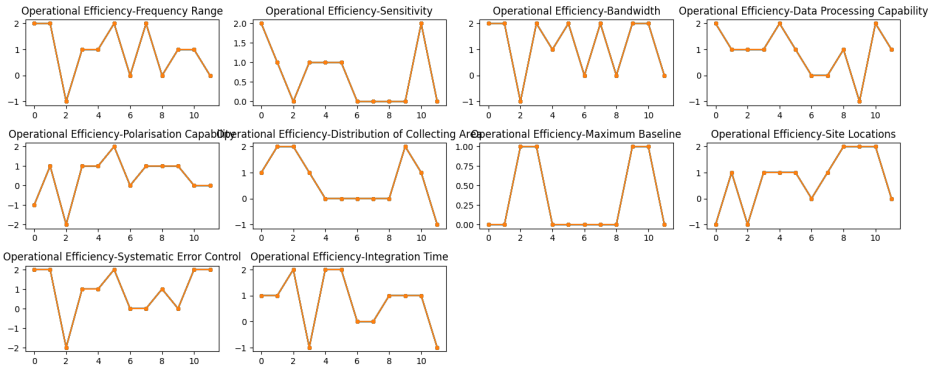


Figure C.4: Non Technical Comparison for: Operational Efficiency

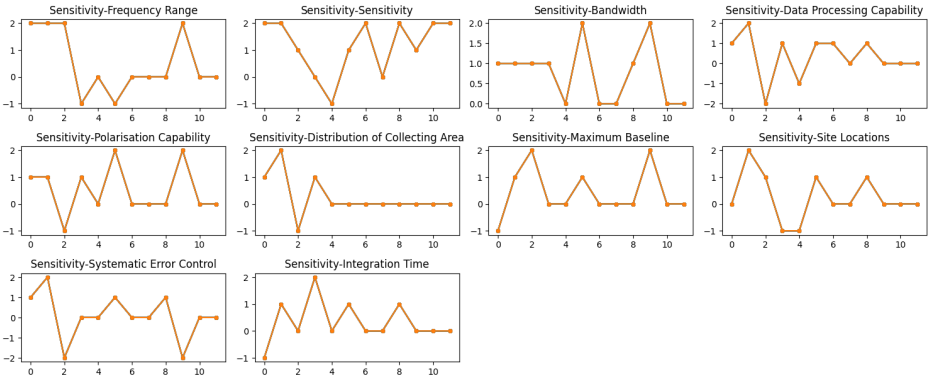


Figure C.5: Non Technical Comparison for: Sensitivity

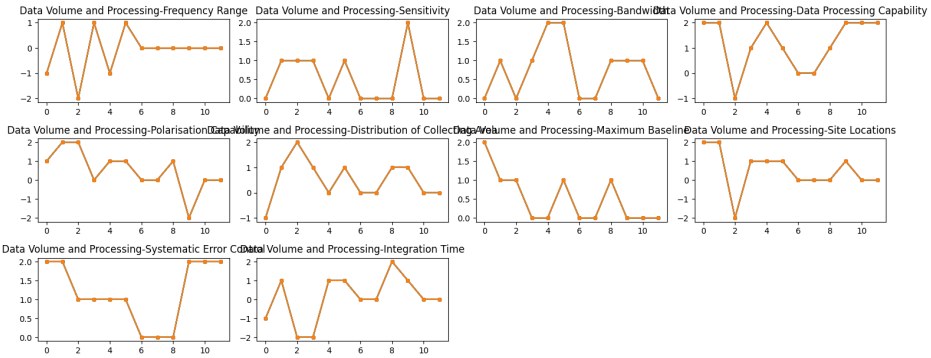


Figure C.6: Non Technical Comparison for: Data Volume and Processing

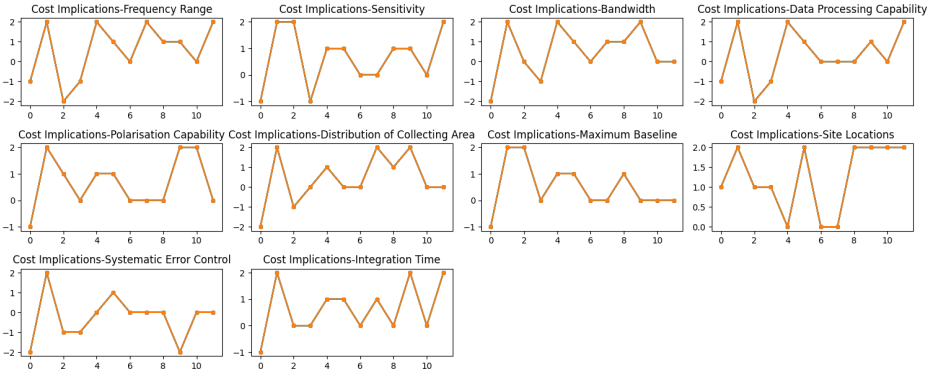


Figure C.7: Non Technical Comparison for: Cost Implications

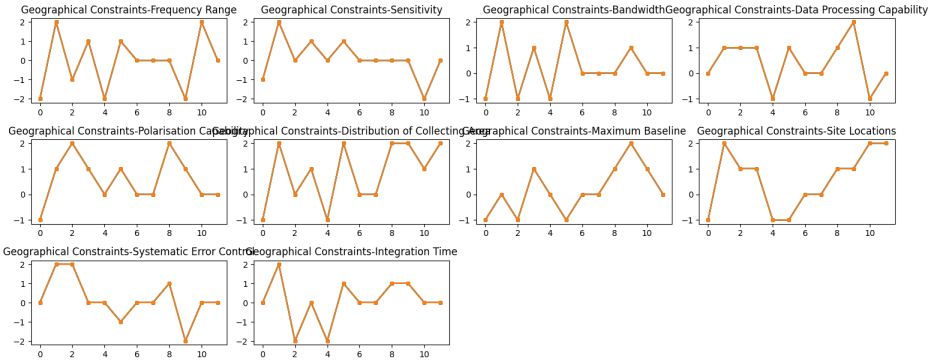


Figure C.8: Non Technical Comparison for: Geographical Constraints

Appendix D

Responses for GPT

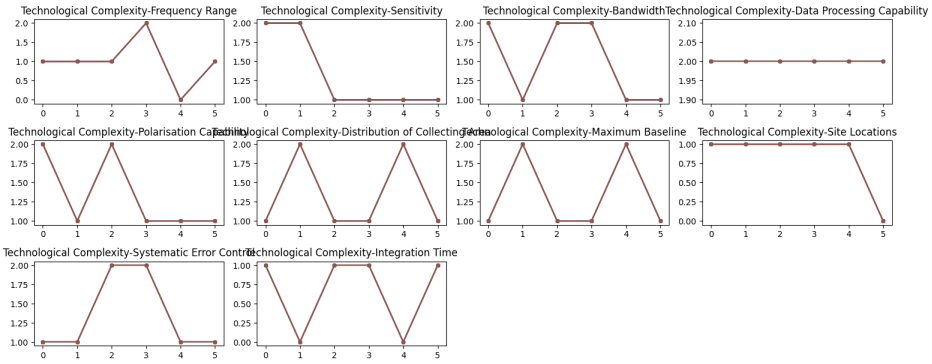


Figure D.1: GPT Comparison for: Technological Complexity

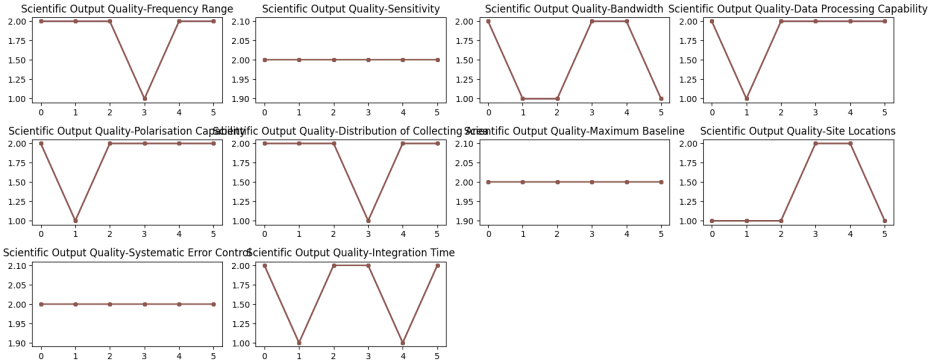


Figure D.2: GPT Comparison for: Scientific Output Quality

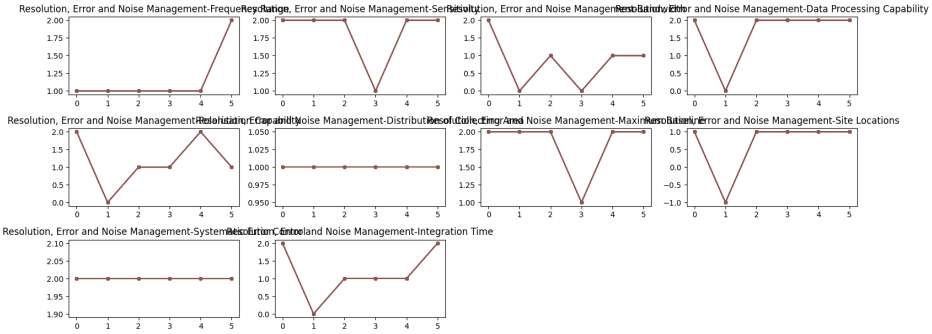


Figure D.3: GPT Comparison for: Resolution, Error and Noise Management

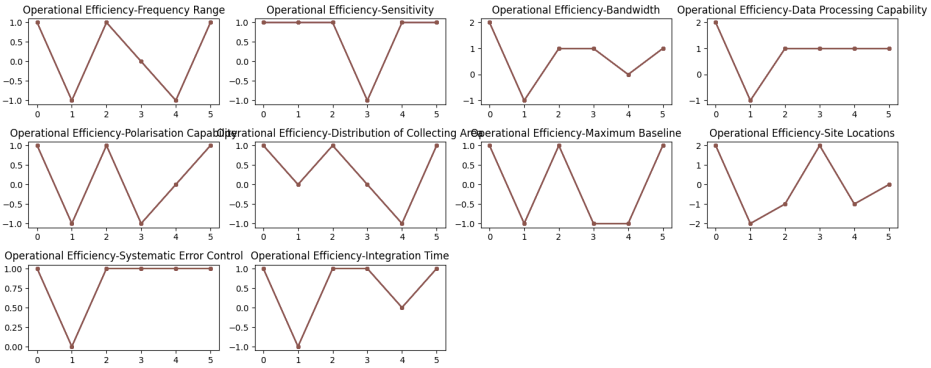


Figure D.4: GPT Comparison for: Operational Efficiency

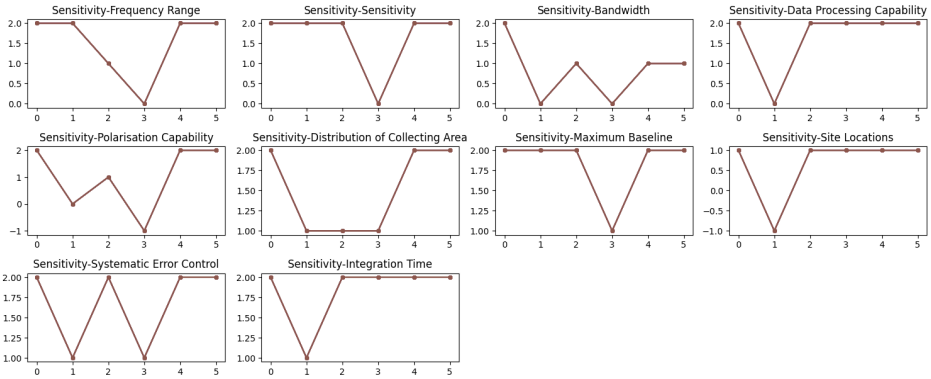


Figure D.5: GPT Comparison for: Sensitivity

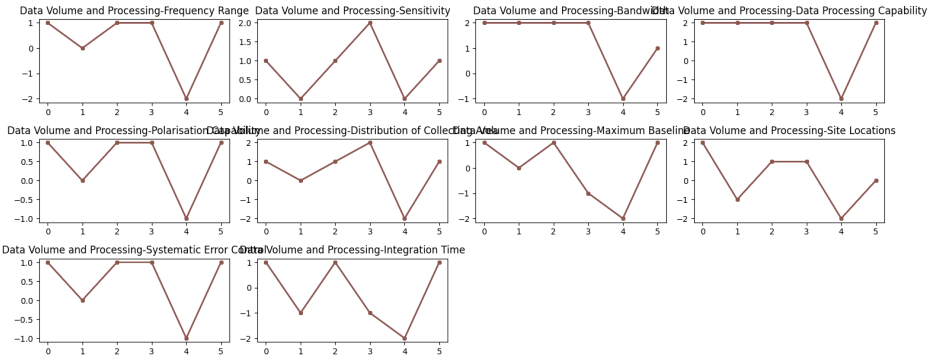


Figure D.6: GPT Comparison for: Data Volume and Processing

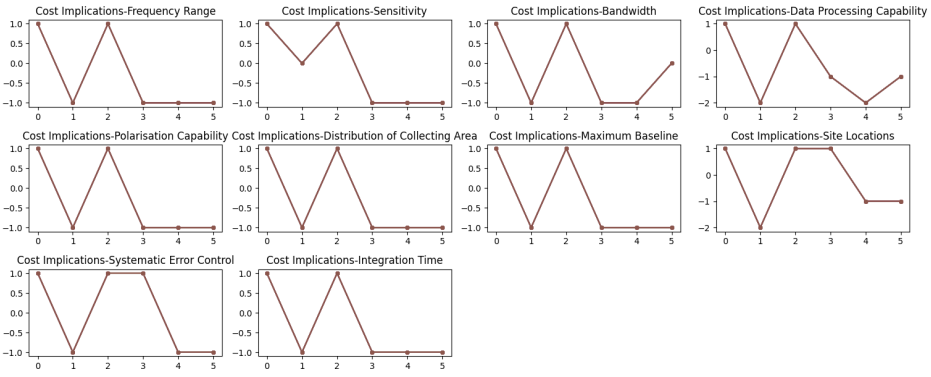


Figure D.7: GPT Comparison for: Cost Implications

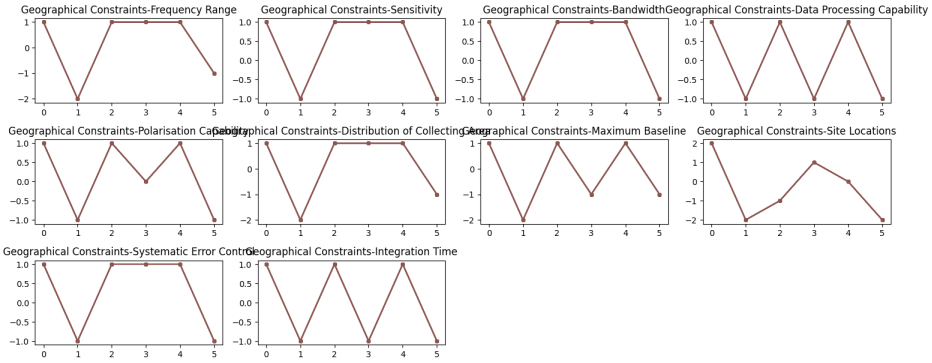


Figure D.8: GPT Comparison for: Geographical Constraints

Bibliography

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *CoRR*, *abs/1706.03762*. http://arxiv.org/abs/1706.03762

Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., Awadallah, A. H., White, R. W., Burger, D., & Wang, C. (2023). Autogen: Enabling next-gen llm applications via multi-agent conversation.

Chen, C. (2022). *Realization of inter-model connections: Linking requirements and computer-aided design* [Doctoral dissertation, University of Georgia].

Chen, C., Carroll, C., & Morkos, B. (2023). From text to images: Linking system requirements to images using joint embedding. *Proceedings of the Design Society*, *3*, 1985–1994.

Chen, C., & Morkos, B. (2023). Exploring topic modelling for generalising design requirements in complex design. *Journal of Engineering Design*, *34*(11), 922–940.

Chen, C., Mullis, J., & Morkos, B. (2021). A topic modeling approach to study design requirements. *International design engineering technical conferences and computers and information in engineering conference*, *85383*, V03AT03A021.

Chen, C., Wei, S., & Morkos, B. (2023). Bridging the knowledge gap between design requirements and cad-a joint embedding approach. *2023 ASEE Annual Conference & Exposition*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023). Chain-of-thought prompting elicits reasoning in large language models.

Gong, X., Jiao, R., Jariwala, A., & Morkos, B. (2021). Crowdsourced manufacturing cyber platform and intelligent cognitive assistants for delivery of manufacturing as a service: Fundamental issues and outlook. *The International Journal of Advanced Manufacturing Technology*, *117*(5), 1997–2007.

Hein, P. H., Kames, E., Chen, C., & Morkos, B. (2021). Employing machine learning techniques to assess requirement change volatility. *Research in engineering design*, *32*, 245–269.

Hein, P. H., Kames, E., Chen, C., & Morkos, B. (2022). Reasoning support for predicting requirement change volatility using complex network metrics. *Journal of Engineering Design*, *33*(11), 811–837.

Hein, P. H., Menon, V., & Morkos, B. (2015). Exploring requirement change propagation through the physical and functional domain. *International design engineering technical conferences and computers and information in engineering conference*, *57052*, V01BT02A051.

Htet Hein, P., Morkos, B., & Sen, C. (2017). Utilizing node interference method and complex network centrality metrics to explore requirement change propagation. *International design engineering technical conferences and computers and information in engineering conference*, *58110*, V001T02A081.

Dalton, A., Wolff, K., & Bekker, B. (2022). Interdisciplinary research as a complicated system. https://doi.org/10.1177/16094069221100397

Rosenblat, T. S., & Mobius, M. M. (2004, August). Getting Closer or Drifting Apart?*. https://doi.org/10.1162/0033553041502199

Daniel, K. L., McConnell, M., Schuchardt, A., & Peffer, M. E. (2022). Challenges facing interdisciplinary researchers: Findings from a professional development workshop. *PloS one*, *17*(4 April). https://doi.org/10.1371/journal.pone.0267234

Hu, L., Huang, W.-b., & Bu, Y. (2024). Interdisciplinary research attracts greater attention from policy documents: Evidence from covid-19. *Humanities and Social Sciences Communications*, *11*. https://doi.org/10.1057/s41599-024-02915-8

McLellan, J. M., Morkos, B., Mocko, G. G., & Summers, J. D. (2010). Requirement modeling systems for mechanical design: A systematic method for evaluating requirement management tools and lan-

guages. *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, *44113*, 1247–1257.

Morkos, B., Joshi, S., & Summers, J. D. (2012). Representation: Formal development and computational recognition of localized requirement change types. *International design engineering technical conferences and computers and information in engineering conference*, *45028*, 111–122.

Morkos, B., Joshi, S., Summers, J. D., & Mocko, G. G. (2010a). Evaluation of requirements and data content within industry in-house developed data management system. *International Design Engineering Technical Conference*.

Morkos, B., Joshi, S., Summers, J. D., & Mocko, G. G. (2010b). Requirements and data content evaluation of industry in-house data management system. *International design engineering technical conferences and computers and information in engineering conference*, *44113*, 493–503.

Morkos, B., & Summers, J. D. (2010). Requirement change propagation prediction approach: Results from an industry case study. *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, *44090*, 111–121.

Mullis, J., Chen, C., Morkos, B., & Ferguson, S. (2023). Deep Neural Networks in Natural Language Processing for Classifying Requirements by Origin and Functionality: An Application of BERT in System Requirements. *Journal of Mechanical Design*, *146*(4), 041401. https://doi.org/10.1115/1.4063764

Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N. V., Wiest, O., & Zhang, X. (2024). Large language model based multi-agents: A survey of progress and challenges.

Olajoyegbe, T. O., & Morkos, B. (2022). Utilizing bayesian inference to optimization manufacturing facility configuration and task sequencing in product remanufacturing. *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, *86250*, V005T05A012.

Piazza, A., Bielanos, K., & Morkos, B. (2017). Exploration of various methods for cost considerations in additive manufacturing. *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, *58165*, V004T05A016.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2024). Retrieval-augmented generation for large language models: A survey.

Silva, J. M., Javales, R., & Silva, J. R. (2019). A new requirements engineering approach for manufacturing based on petri nets [13th IFAC Workshop on Intelligent Manufacturing Systems IMS 2019]. *IFAC-PapersOnLine*, *52*(10), 97–102. https://doi.org/https://doi.org/10.1016/j.ifacol.2019.10.006

Elneel, D. A., Fakharudin, A. S., Ahmed, E. M., Kahtan, H., & Abdullateef, M. (2022). Stakeholder identification overview and challenges in requirements engineering prospective, 314–319. https://doi.org/10.1109/ICCIT52419.2022.9711653

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). React: Synergizing reasoning and acting in language models.

Ross, A. M., & Hastings, D. E. (2005). 11.4.3 the tradespace exploration paradigm. *INCOSE International Symposium*, *15*(1), 1706–1718. https://doi.org/https://doi.org/10.1002/j.2334-5837.2005.tb00783.x

Shankar, P., Morkos, B., & Summers, J. D. (2012). Reasons for change propagation: A case study in an automotive oem. *Research in Engineering Design*, *23*, 291–303.

Summers, J. D., Joshi, S., & Morkos, B. (2014). Requirements evolution: Relating functional and non-functional requirement change on student project success. *International design engineering technical conferences and computers and information in engineering conference*, *46346*, V003T04A002.

Spero, E., Bloebaum, C. L., German, B. J., Pyster, A., & Ross, A. M. (2014). A research agenda for tradespace exploration and analysis of engineered resilient systems [2014 Conference on Systems Engineering Research]. https://doi.org/https://doi.org/10.1016/j.procs.2014.03.091

Daniels, J., Turner, C., Wagner, J., Masoudi, N., Agyemang, M., Hartman, G., Rizzo, D., Gorsich, D., Skowronska, A., & Agusti, R. (2022, March). Designing the design space: Evaluating best prac-

tices in tradespace exploration, analysis and decision-making. https://doi.org/10.4271/2022-01-0354

Machchhar, R. J., Toller Melén, C. N. K., & Bertoni, A. (2024). A tradespace exploration approach for changeability assessment from a system-of-systems perspective: Application from the construction machinery industry. *Proceedings of the Design Society*, *4*, 2655–2664. https://doi.org/10.1017/pds.2024.268

Xu, Y., Turner, C., & Wagner, J. (2023). Multi-Objective Optimization and Tradespace Analysis of a Mechanical Clock Movement Design. *ASME Open Journal of Engineering*, *2*, 021029. https://doi.org/10.1115/1.4062410

Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., Jiang, X., Cobbe, K., Eloundou, T., Krueger, G., Button, K., Knight, M., Chess, B., & Schulman, J. (2022). Webgpt: Browser-assisted question-answering with human feedback.