Experimental Protocol
Leveraging LLMs with Compositional Generalization and Agent-Based Modelling, in the scope of PubMed Clinical Prostate Cancer papers

Juan Bru
jbru@kailuasystems.com

Abstract

This study aims to explore the intersection of Large Language Models (LLMs) and Agent-Based Modeling (ABM), and their application to the clinical domain. The primary objective is to enhance the capabilities of LLMs through Compositional Generalization techniques such as Chain of Thought (CoT) and Few-Shot Learning (FSL), along with developing innovative LLM architectures using these models as foundational building blocks. A secondary aim is to apply these technological advancements to the management of medical domain documentation in prostate cancer scope. By integrating LLMs with a curated dataset of PubMed articles, this research seeks assess how to potentially support both the field of artificial intelligence and medical research and operations.

Employing a comparative analysis of LLMs, including GPT-3.5, GPT-4, and Meta's LLaMA2, against a comprehensive dataset from PubMed, this research aims to validate the proposed hypotheses and demonstrate the practical implications of enhanced LLMs and ABM in supporting clinical management and research.

1 Introduction

The rapid advancement of Artificial Intelligence (AI) in recent years has ushered in a new era of potential for processing and generating natural language, particularly with the development of Large Language Models (LLMs). These models have shown remarkable capabilities in understanding and producing text across a wide array of domains. Concurrently, the medical field, continues to generate vast amounts of data that necessitate advanced processing tools for efficient utilization at operational and research level (Batko K, 2022). In essence, the research aims to demonstrate the profound impact that advanced LLMs augmented by Compositional Generalization, and ABM, can have when applied to complex problems.

2 Objectives

2.1 Primary Objective

The primary objective of this experimental protocol is to push the boundaries of Compositional Generalization techniques (Thaddäus Wiedemer, 2023) techniques such as Chain of Thought (CoT) (Jason Wei, 2022) and Few-Shot Learning (FSL) (Yaqing Wang, 2019), coupled with the strategic use of LLMs as foundational building blocks in developing new, advanced LLM architectures. This investigation is dedicated to enhancing LLMs' performance in complex reasoning, adaptive learning tasks, and the nuanced application of these models in various domains.

2.2 Secondary Objectives

2.2.1 Secondary Objective 1

ABM (Mauricia Salgado, 2013) plays a supporting, yet significant role in this exploration, offering a method to contextualize and apply the insights generated by LLMs within simulated environments. ABM's ability to model complex systems and interactions provides a unique lens through which the potential practical applications of LLM advancements can be assessed and refined. This complementary approach enables a richer exploration of how enhanced LLMs might interact with and influence real-world systems and processes.

2.2.2 Secondary Objective 2

While the research is centered on advancing LLM and ABM technologies, medical prostate cancer serves as a practical and impactful domain for application. This specific medical field is chosen not as the primary focus but as a real-world context in which the effectiveness and versatility of the enhanced LLM capabilities, supported by ABM simulations, can be evaluated.

The use of a curated dataset of articles from PubMed (https://pubmed.ncbi.nlm.nih.gov/about/, 2024) related to prostate cancer allows for a targeted application of these technologies, showcasing their potential to revolutionize not only medical research and patient care within the prostate cancer domain but also illustrating a model for their application across a wide range of fields and challenges.

3 Justification

The integration of advanced computational techniques, specifically LLMs enhanced by compositional generalization techniques and ABM, represents a pioneering approach to addressing complex challenges across various domains, including healthcare. This experimental protocol is predicated on the assertion that the intersection of these technologies can catalyze significant advancements in the processing, understanding, and generation of complex data and scenarios. The justification for this exploration is multifaceted:

- Addressing Current Limitations of LLMs: Despite the remarkable capabilities of LLMs in language understanding and generation, there remain challenges in their application to specialized domains requiring nuanced comprehension and adaptive learning. The integration of Compositional Generalization techniques aims to surmount these hurdles, enhancing the models' ability to engage with complex, domain-specific information more effectively.
- Leveraging the Synergistic Potential of ABM: ABM offers a dynamic framework for simulating real-world systems and interactions, providing a contextual backdrop against which the enhanced capabilities of LLMs can be applied and evaluated. This symbiotic integration facilitates a deeper understanding of the potential impacts and applications of LLM advancements in practical, real-world settings.
- Practical Application in a High-Impact Domain: The choice to apply these advanced computational strategies within the context of health, in prostate cancer is strategic, offering a tangible domain where the potential benefits of this experiment can be directly observed.
- Pioneering a Model for Broad Application: While prostate cancer serves as the immediate context for this investigation, the underlying objective extends beyond a single domain. By demonstrating the effectiveness of integrating enhanced LLMs with ABM in a complex medical field, this research aims to establish a model for leveraging these technologies to tackle multifaceted challenges across a wide array of sectors.
- Contributing to the Advancement of AI and Healthcare: Ultimately, this research seeks to contribute to the broader fields of artificial intelligence and healthcare by showcasing how innovative applications of LLMs and ABM can drive forward both technological advancements and real-world impact..

In summary, the justification for this experimental protocol lies in its potential to overcome existing limitations of LLMs, harness the unique capabilities of ABM, and apply these advancements to significantly impact domains, as prostate cancer care. This research represents an ambitious step forward in the convergence of AI and practical application, with the promise of yielding insights and methodologies that can inform future innovations across diverse domains.

4 Hypothesis

4.1 Enhancing LLMs by Compositional Generalization Techniques: Chain of Thought and Few Shot Learning

Hypothesis Statement: It is hypothesized that the integration of Compositional Generalization techniques, specifically Chain of Thought (CoT) and Few-Shot Learning (FSL), can significantly enhance the capabilities of LLMs in complex reasoning and adaptive learning tasks. This enhancement is anticipated to be evident through improved accuracy in problem-solving and the model's ability to generalize from limited data inputs.

Description of Key Components

Compositional Generalization: This concept refers to the model's ability to understand and apply learned concepts and operations in novel configurations that were not explicitly encountered during training. It facilitates adaptability and creative problem-solving in changing environments.

Chain of Thought (CoT): An approach that prompts models to articulate intermediate steps or reasoning pathways when addressing complex tasks, simulating a human-like problem-solving process and potentially leading to enhanced comprehension and solutions.

Few-Shot Learning (FSL): A paradigm wherein models adapt to new tasks or assimilate new information with a minimal number of examples, leveraging their extensive pre-trained knowledge base. This method is particularly effective in extending model applicability across diverse domains where data availability is limited.

Rationale: The adoption of CoT and FSL for enhancing LLMs addresses current limitations by providing a structured framework for multi-step reasoning and enabling rapid adaptation to new tasks with minimal data. These techniques aim to foster nuanced understanding and application of knowledge, mirroring the cognitive flexibility and efficiency observed in human problem-solving.

Expected Outcomes: The integration of CoT and FSL into LLMs is expected to result in:

- Enhanced ability of LLMs to solve complex and nuanced reasoning problems with higher accuracy and efficiency.
- Improved efficiency in learning and adapting to new domains or tasks, reducing the need for extensive data for training or retraining.
- Expansion in the application of LLMs to specialized areas requiring deep, structured reasoning or where data is inherently scarce."

4.2 Advancing LLM Architecture using LLMs as building blocks

Hypothesis Statement: It is posited that leveraging LLMs as foundational building blocks within new LLM architectures can significantly advance the field by enhancing the models' complexity, adaptability, and performance across a range of linguistic and cognitive tasks. This strategy aims to harness the inherent strengths of existing LLMs to construct more sophisticated and efficient models.

Key Components

LLM Architecture: Refers to the structural and computational design of LLMs, which encompasses the underlying neural network configurations, training methodologies, and data processing mechanisms.

Building Blocks: In this context, building blocks denote pre-trained LLM components or modules that can be integrated into new LLM architectures. These components have already been trained on extensive datasets, embodying a broad understanding of language and knowledge.

Rationale: The rationale for this approach stems from the recognition that existing LLMs encapsulate valuable linguistic and world knowledge, alongside sophisticated reasoning capabilities. **By integrating these pre-trained models as components within new architectures, there is an opportunity to bypass some of the extensive resource requirements for training from scratch while also introducing novel functionalities or enhancing performance characteristics.**

Expected Outcomes

- Adopting LLMs as building blocks in the development of new LLM architectures is expected to result in:
    - o Increased efficiency in model training and development, as leveraging pre-trained components can significantly reduce the computational resources and time required.
    - o Enhanced performance and adaptability of LLMs, particularly in specialized or emerging linguistic and cognitive tasks, through the strategic combination of diverse model capabilities.
    - o The potential for innovative model architectures that transcend current limitations, offering improved understanding, generation, and interaction in natural language processing applications.

Significance: This hypothesis suggests a paradigm shift in the development of LLM architectures, advocating for a more modular and efficient approach that builds upon the successes of existing models. Such advancements could accelerate the pace of innovation in natural language processing and artificial intelligence, broadening the scope of applications and improving the effectiveness of LLMs in complex tasks. Ultimately, this could lead to more intelligent, responsive, and capable AI systems.

4.3 Improving Compositional generalization using Structured Question Taxonomies

Hypothesis Statement: The application of structured question taxonomies, such as Bloom's Taxonomy (Bloom, Engelhart, Furst, Hill, & Krathwohl, 1956), is hypothesized to significantly improve the Compositional Generalization abilities of systems engaged in generating educational content or assessments. This improvement is expected to manifest through enhanced specificity, relevance, and cognitive challenge in the generated questions.

Description of Key Components

Structured Question Taxonomies: Frameworks that categorize questions according to various cognitive levels. Bloom's Taxonomy, a widely recognized taxonomy, organizes cognitive tasks into categories ranging from basic (remembering) to complex (creating), which can guide the generation of questions targeting specific cognitive skills.

Rationale: Integrating structured question taxonomies into the process of Compositional Generalization addresses the challenge of creating educational content that is not only varied and comprehensive but also aligned with learning objectives at different cognitive levels. This approach ensures that the generated content can assess or fostering a broad spectrum of cognitive skills, from simple recall to complex analysis and creation.

Expected Outcomes: The utilization of structured question taxonomies in Compositional Generalization systems is expected to lead to:

- An increase in the diversity and cognitive range of questions or content produced, enabling a more comprehensive evaluation of learners' understanding and skills.
- Enhanced alignment of generated questions with educational goals, facilitating targeted teaching strategies and learning interventions.
- Improved ability of educational content generation systems to tailor content to different learning stages and objectives, contributing to more effective and personalized learning experiences.

4.4 Using Agent-Based Modeling for Innovative LLM Architectures

Hypothesis Statement: The integration of Agent-Based Modeling (ABM) techniques within Large Language Model (LLM) architectures is hypothesized to foster innovative developments in the structure and functionality of LLMs (Chen Gao, 2023). This approach is expected to enhance model adaptability, scalability, and interaction capabilities, by simulating complex systems and interactions at the agent level.

Key Components

Agent-Based Modeling (ABM): A computational modeling approach that simulates the actions and interactions of autonomous agents (individuals or collective entities) to assess their effects on the system. ABM is particularly useful for understanding complex, dynamic systems.

Rationale: The rationale behind integrating ABM with LLM architectures lies in the potential of ABM to simulate and analyze complex, emergent behaviors from simple agent interactions. By applying ABM principles to LLM development, it is posited that LLMs can achieve greater flexibility and efficiency in processing and generating language, especially in scenarios involving dynamic contexts or requiring nuanced understanding of interactions.

Expected Outcomes: The application of ABM to the innovation of LLM architectures is anticipated to lead to:

- Enhanced adaptability of LLMs to varying contexts and requirements, facilitated by the dynamic simulation capabilities of ABM.
- Improved scalability of models, enabling them to handle more complex tasks and larger datasets more efficiently.
- Increased sophistication in the models' interaction capabilities, allowing for more nuanced and contextually relevant language generation and comprehension.

4.5 Digitalization of Cancer Prostate Unit based in Agent-Based Modeling and LLMs

Hypothesis Statement: The hypothesis posits that the digitalization of Prostate Cancer Unit operations can be significantly improved through the integration of Agent-Based Modeling (ABM) and Large Language Models (LLMs). This approach is anticipated to refine diagnostic, treatment, and patient care processes by leveraging the detailed simulation capabilities of ABM and the sophisticated language understanding and generation features of LLMs.

Key Components

Digitalization of Prostate Cancer Unit: Involves applying digital technologies to optimize the management of patient data, treatment planning, and patient interaction within the context of prostate cancer care.

Rationale: Integrating ABM with LLMs in the digitalization of Prostate Cancer Unit operations addresses the complexity of cancer care, which involves numerous variables including patient conditions, treatment responses, and healthcare delivery systems. ABM can simulate the dynamics of the healthcare ecosystem, offering insights into optimal care pathways and potential outcomes. Concurrently, LLMs can enhance the processing and generation of medical documentation, patient communication, and the synthesis of the latest research and clinical guidelines. Together, these technologies promise to streamline operations, improve patient care, and support data-driven decision-making.

Expected Outcomes: The combined application of ABM and LLMs in prostate cancer care units is expected to yield:

- Enhanced efficiency and accuracy in the management of patient information and treatment protocols, enabled by LLMs' capacity to process and generate relevant documentation and communications.
- Improved patient care strategies, resulting from ABM's ability to simulate patient trajectories and treatment outcomes, allowing for personalized care planning.
- Greater adaptability in responding to emerging treatment modalities and patient needs, facilitated by the predictive modeling of ABM and the dynamic information processing capabilities of LLMs.

Significance

This hypothesis underscores a transformative approach to the digitalization of Prostate Cancer Unit operations, suggesting that the integration of ABM and LLMs could lead to significant improvements in how prostate cancer care is delivered. By enhancing the precision, efficiency, and adaptability of care processes, this approach has the potential to positively impact patient outcomes and healthcare provider workflows. Furthermore, it could serve as a model for the digitalization of other specialized healthcare units, contributing to broader advancements in medical care and patient management.

5 Data

5.1 Dataset Overview

The dataset comprises a collection of articles from PubMed (https://pubmed.ncbi.nlm.nih.gov/about/, 2024), a free search engine accessing primarily the MEDLINE database of references and abstracts on life sciences and biomedical topics. This specific dataset focuses on articles related to prostate cancer, encompassing a wide range of topics including diagnosis, treatment options, patient care, epidemiology, genetic factors, and advancements in medical technology pertaining to prostate cancer.

5.2 Dataset Scope

Temporal Coverage: Our dataset spans one year, a period selected due to budget and time constraints. This timeframe is ideal for, developing all the experiments pointed out previously, but can impact on the quality of results, that could require bigger datasets.

Subject Coverage: The dataset covers a broad spectrum of subjects within the realm of prostate cancer, including but not limited to: Clinical research on the efficacy and side effects of various treatment options. Studies on the genetic and environmental risk factors associated with prostate cancer. Research on diagnostic techniques and Articles on patient care, quality of life, survivorship, and psychosocial aspects of prostate cancer treatment and recovery.

5.3 Dataset Contents

Articles: Each entry in the dataset represents a scholarly article that has been peer-reviewed and published in various medical and scientific journals. The content of these articles includes research findings, clinical trial results, review articles, meta-analyses, and editorial opinions related to prostate cancer.

Metadata: For each article, the dataset includes metadata such as the title, authors, publication date, journal name, abstract, keywords, and, where available, links to the full text. The dataset may also contain citations and references to other relevant articles.

5.4 Data Sources and Accessibility

Source: All articles and metadata are sourced from PubMed, which aggregates content from MEDLINE, life science journals, and online books. PubMed provides access to millions of citations and abstracts freely available to the public, with links to full-text articles when available through PubMed Central or publisher websites.

Accessibility: The dataset can be accessed through PubMed's search interface or via programmatic interfaces such as the Entrez Programming Utilities (E-utilities) for automated searches and data retrieval. Users can perform targeted searches using specific keywords, author names, or journal titles to compile a dataset tailored to the scope of their research on prostate cancer.

5.5 Dataset Utilization

To utilize the PubMed dataset of prostate cancer articles effectively in validating the various hypotheses prepared, a structured approach will be adopted for each hypothesis. This involves extracting relevant information, conducting analyses, and applying findings directly to the hypotheses. Here's how the dataset can be employed to validate each specific hypothesis:

1. **Data Preparation:** Identify and preprocess relevant articles.
2. **Model Training:** Use the half of dataset for LLM training and ABM simulations.
3. **Evaluation:** Test models with the rest of the dataset

6 Metrics

6.1 Qualitative Evaluation

- **Consistency Check** (Lukas Fluri, 2023)**:** Evaluate the consistency of the LLM's responses across similar documents or questions. Consistent responses to similar analytical prompts indicate the model's reliability.
- **Error Typology Analysis**
- (https://www.qt21.eu/mqm-definition/issues-list-2015-12-30.html)**:** Categorize and analyze the types of errors (e.g., factual inaccuracies, misinterpretations) the model makes. This analysis can guide targeted improvements and training.
- **Comparison among LLM Models:** Compare the LLM's output against those from baseline models known for their performance in text analysis. This comparison can highlight the LLM's unique strengths or weaknesses. We plan to use LLaMA2 and GPT3.5 as base, and GPT4.0 as evaluator.

6.2 Quantitative Evaluation

- **Accuracy Measurement:** For each question posed to the model, compare the LLM's responses with a set of predefined correct answers or annotations. This comparison can be automated to some extent, especially for straightforward, fact-based questions.

- **Precision, Recall, and F1 Score** (Powers, 2007)**:** Implement automated scripts to calculate these metrics, especially for tasks like information extraction where the model's responses can be clearly categorized as true positives, false positives, and false negatives. Using LLMs to assist with the annotation of data, especially in the context of evaluating other LLM-generated answers, is an innovative approach that can significantly streamline the process, particularly for large datasets. However, it's important to approach this with caution due to potential biases or errors that the assisting LLM might introduce. Here's how you might go about it:
- **Document Coverage Analysis** (Max Schäfer, 2023)**:** Automatically analyze the percentage of documents in your dataset for which the model provides non-trivial, relevant analysis. This can be quantified by setting thresholds for response length, specificity, or relevance scores.
- **Response Time Tracking** (Ramanujan, 2023)**:** Use time stamps before and after each model query to calculate the average response time. Optimizing for faster responses may be necessary for real-time applications.
- **Automated Readability Index (ARI)** (in, 1967)**:** Utilize existing software libraries to calculate the ARI for the model's written outputs. This metric helps ensure the accessibility of the model's analyses to a broader audience.
- **Topic Coherence** (Hamed Rahimi, 2023)**:** For models that generate or identify topics, use coherence measures such as NPMI (Normalized Pointwise Mutual Information) or UMass coherence score to evaluate the quality and distinctiveness of the topics identified by the model.
- **Keyword Frequency Comparison** (James S. Adelman, 2006)**:** Use automated scripts to count keyword frequencies in the model's responses and compare these with frequencies from a manually annotated set. This requires creating a script that can accurately parse and count keywords in both the model's output and the benchmark dataset.

By combining these qualitative and quantitative evaluation strategies, you can comprehensively assess your LLM's performance in quantitative analysis of PubMed articles. This thorough evaluation will help identify areas of strength and opportunities for further.

7 Models

To thoroughly investigate the potential of Large Language Models (LLMs) acting as building blocks within LLMs Architecture and for manage in the medical domain with medical information, particularly focusing on prostate cancer research, we plan to employ a comparative analysis approach. This involves using both established and cutting-edge models as baselines and subjects of investigation. Our selection includes GPT-3.5, GPT-4 (OpenAI, s.f.), and Meta's LLaMA2 (Meta, 2024) models, chosen for their prominence, capabilities, and varying architectural innovations.

We plan to use GPT3.5 and LLaMA2 as bases for the work, while using GPT4.0 as support for evaluation.

8 General Reasoning

Focusing on the utilization of LLMs, Compositional Generalization, and primarily as tools for enhancing the quality and accessibility of datasets derived from articles for prostate cancer research, we shift away from direct clinical application towards supporting the foundational research aspects. These advanced computational techniques serve as a backbone to refine, understand, and expand the datasets that researchers rely upon.

Enhancing Datasets with Large Language Models: The role of LLMs, including GPT-3.5, GPT-4, and Meta's LLaMA2, extends into the realm of data curation and enhancement for prostate cancer research. By processing extensive arrays of research articles and clinical data, these models are adept at identifying, synthesizing, and summarizing key findings. This capability is instrumental in transforming vast quantities of unstructured data into organized, accessible formats that researchers can easily navigate. The goal is to enrich the datasets with comprehensive insights drawn from the latest studies, making it easier for researchers to draw connections, formulate hypotheses, and identify gaps in the current knowledge base.

Leveraging Compositional Generalization: Compositional Generalization plays a pivotal role in augmenting the utility of LLMs for dataset enhancement. This technique empowers models to apply learned concepts in new ways, facilitating the generation of novel insights and the identification of patterns within the data. The application of Compositional Generalization ensures that the datasets are not just repositories of information but are dynamic resources that can adapt and evolve with the emerging trends and discoveries in prostate cancer research. It enables the creation of more nuanced and comprehensive datasets that reflect the complex nature of medical research.

Agent-Based Modeling for Simulating and Digitalization: While ABM is typically associated with simulating complex systems, its application in the context of supporting prostate cancer research through dataset development is somewhat indirect yet impactful.

ABM can be utilized to simulate the effects of various research findings or treatment outcomes, generating simulated data that can fill in the gaps in existing datasets or highlight areas where additional research is needed. This approach can help in forecasting future trends, evaluating the potential impact of new therapies, and identifying under-researched areas, thus guiding researchers towards promising directions for further investigation.

By focusing on these technologies as enablers for research rather than direct clinical tools, the aim is to fortify the foundation upon which prostate cancer research is built, enhancing the capacity for discovery and innovation in the field..

8.1 Model and Data Integration

The choice of GPT-3.5, and Meta's LLaMA2 models as subjects of investigation is informed by their demonstrated capabilities in processing natural language at scale, their varying approaches to learning and generation, and their potential applicability to domain-specific challenges found in medical research. Using GPT-4, as help for evaluator closes the loop. Each model represents a point along the spectrum of current LLM technology, from GPT-3.5's foundational capabilities to GPT-4's advanced processing power and LLaMA2's efficiency in scaling. Evaluating these models against a corpus of recent, relevant medical literature on prostate cancer allows for a nuanced analysis of each model's strengths and weaknesses in interpreting and synthesizing complex medical information.

Our core hypotheses revolve around the potential of LLMs to not only understand and generate text that accurately reflects the depth and complexity of medical knowledge but also to do so in a way that is creatively enriched and contextually relevant. By employing Compositional Generalization techniques and structured question taxonomies, we hypothesize that LLMs can achieve a higher degree of precision and relevance in their outputs, thereby becoming more effective tools for medical professionals and researchers.

The use of prostate cancer research articles from PubMed as our data source is deliberate, chosen for the depth and specificity of information these articles contain. This dataset will challenge the LLMs to navigate and interpret specialized vocabulary, complex treatment protocols, and the latest research findings, offering a rigorous testing ground for our hypotheses.

8.2 Preliminary Description of Investigation Focus

Through quantitative and qualitative evaluations, including accuracy measurements, expert reviews, and comparisons with baseline models, this investigation aims to extent LLMs by Compositional Generalization Techniques and new architecture more modular. Finally, if we can contribute to the field of medical operations and research. Specifically, we will assess how well these models can:

- Incorporate and leverage Compositional Generalization techniques and structured question taxonomies to improve the generation of medical content.
- Serve as modular building blocks within LLM architectures, potentially reducing the need for developing new neural network structures from scratch.
- Be utilized within Agent-Based Modeling frameworks to develop innovative LLM architectures and apply these models in organizational digitalization, especially within the context of healthcare.

This exploration is not just about advancing our understanding of LLM capabilities but also about identifying practical, end user applications and improvements that can be made to leverage these technologies in addressing real-world challenges in prostate cancer research and treatment.

9 Summary of Progress

As our project enters a critical phase, we have made progress in setting the groundwork for a comprehensive analysis of LLMs, ABM and within the domain of medical research. Here is a summary of our progress to date.

9.1 Work done.

- Data Acquisition: We have curated a dataset from PubMed, on prostate cancer. This dataset, rich in domain-specific knowledge and up-to-date research findings, is the base for testing our hypotheses and evaluating the capabilities of LLMs in understanding and generating medically relevant content.
- Infrastructure setup: we have established virtual machines on Azure and local Spanish hosting company. The plan is to use one for core software of ABM in Python and PostgreSQL as based, while the other to runs LLaMA2. The core of the software in Python will call LLaMA2 machine and OpenAI APIs for GPT3.5 and GPT4.0.
- Model Access and Setup:
    - We have secured access to OpenAI's GPT-4 and GPT-3.5, two of the most advanced LLMs available, known for their ability to process and generate natural language at an unprecedented scale.

- Additionally, we have almost ready a machine for running Meta's LLaMA2 "on-premises", in opposite to APIs provided by other players, ensuring we have a controlled environment to evaluate its performance closely and securely, Let's point we have quite a lot of problems to make LLaMA2 work, and in fact is note ready yet.

- Development of Agent-Based Modeling basic Framework: In alignment with our hypotheses, we have been developing modules based on ABM. These modules are specifically designed to enhance Compositional Generalization techniques. This framework is based in Python and PostgreSQL to provide agent's communication and persistence for experiment's inputs and outputs.

9.2 Next Steps

- **Finalize the design of the system based on building blocks, for Chain of Thought, Few Shot Learning.**
- **Final configuration of Meta's LLaMA2 on Azure or AWS.**
- **Implement the system, on a minimum viable product approach, based on Ubuntu, Python, PostgreSQL, OpenAI GPT4.0, GPT3.5, LLaMA2.**
- **Launch Experimental Processes:** Initiate the execution of our LLMs processes against the prostate cancer dataset, applying Chain of Thought, Few Shot Learning building blocks developed.
- **Model and simulate prostate cancer unit** with previous building blocks developed based on LLMs and ABM.
- **Results Data Collection:** Systematically gather the outputs generated by GPT-4, GPT-3.5, and Meta's LLaMA2 for subsequent analysis..
- **Analysis and Evaluation:** Employ both qualitative and quantitative evaluation metrics to assess the performance of each model.
- **Writing final Paper.**

References

Batko K, Ś. A. (2022). The use of Big Data Analytics in healthcare. *J Big Data. 2022;9(1):3.*

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *(1956). Taxonomy of educational objectives: The classification of educational goals. Vol. Handbook I: Cognitive domain. .* New York: David McKay Company.

Chen Gao, X. L. (2023). Large Language Models Empowered Agent-based Modeling and Simulation: A Survey and Perspectives. *Arxiv*.

Hamed Rahimi, J. L. (2023). Contextualized Topic Coherence Metrics. *Arxiv*.

https://pubmed.ncbi.nlm.nih.gov/about/. (2024). Obtained from PubMed.

https://www.qt21.eu/mqm-definition/issues-list-2015-12-30.html. (s.f.). Obtenido de https://www.qt21.eu/mqm-definition/issues-list-2015-12-30.html.

in, A. w. (1967). *https://readable.com/readability/automated-readability-index/*. Obtained de The Automated Readability Index.

James S. Adelman, G. D. (2006). Contextual Diversity, Not Word Frequency, Determines Word-Naming and Lexical Decision Times. *Psychology and Counseling. SAGE Journals*.

Jason Wei, X. W. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Arxiv*.

Lukas Fluri, D. P. (2023). Checks, Evaluating Superhuman Models with Consistency. *Arxiv*.

Mauricia Salgado, N. G. (2013). *Agent-BAsed Modelling. In book: Handbook of Quantitative Methods for Educational Research (pp.247-265)Chapter: 12.*

Max Schäfer, S. N. (2023). An Empirical Evaluation of Using Large Language Models for Automated Unit Test Generation. *Arxiv*.

Meta. (2024). *Meta LLama*. Obtenido de Meta LLama: https://llama.meta.com/

OpenAI. (s.f.). *OpenAI*. Obtenido de https://openai.com/.

Powers, D. M. (2007). Evaluation: From Precision, Recall and F-Factor. *Technical Report SIE-07-001*.

Ramanujan, S. (2023). *How to Evaluate LLMs: A Complete Metric Framework*. Obtenido de https://www.microsoft.com/en-us/research/group/experimentation-platform-exp/articles/how-to-evaluate-llms-a-complete-metric-framework/.

Thaddäus Wiedemer, P. M. (2023). Compositional Generalization from First Principles. *Arxiv*.

Yaqing Wang, Q. Y. (2019). Generalizing from a Few Examples: A Survey on Few-Shot Learning. *Arxiv*.