Integrating Large Language Models with Agent-Based Modeling for New Architectures and Digital Transformation

Juan Bautista Bru Sanchis
Kailua Systems
jbru@kailuasystems.com

Abstract

This paper explores the potential of integrating Large Language Models (LLM) and Agent-Based Modeling (ABM) techniques, focusing on their practical application for enhancing LLM architectures and developing ABM-LLM based digitalization frameworks. The research assesses Compositional Generalization techniques, such as Few Shot Learning and Chain of Thought, to improve LLM performance in complex problem-solving scenarios tailored for real-world environments. A curated dataset of prostate cancer research articles concerning the therapeutic use of Sipuleucel-T, extracted from PubMed, serves as the base for experimentation. The study critically assesses the strengths and limitations of these methods and suggests potential avenues for further application-driven research.

1 Introduction

The application of advanced technologies such as Large Language Models (LLMs) and Agent-Based Modeling (ABM) holds transformative potential for various scopes, in our case to improve LLM architecture and models, and in the realms of Digital transformation

This research is structured around a comprehensive analysis of recent literature, an ad hoc built dataset on prostate cancer and the medicinal product Sipuleucel-T, and an innovative LLM high level architecture proposal that leverages GPT3.5.

These elements collectively enables to address complex problems through Compositional Generalization and to simulate organizational dynamics and decision-making processes.

The experiment in this scope will use as driver the following hypothesis:

- H1: Enhancing LLMs by Compositional Generalization Techniques (CG): Few Shot Learning (FS) and Chain of Thought (CoT).
- H2: Advancing LLM Architecture using LLMs as building blocks.
- H4: Application of ABM for developing Hybrid LLM Architectures
- H5: Digitalization of Cancer Prostate Unit based in ABM and LLMs

The study produced a new dataset curated from PubMed, focused on prostate cancer research articles published from 2001 onwards. This dataset, serves as a basis for both training and testing the models, ensuring that the insights are grounded in current and relevant scientific findings..

It was designed and built a new architecture upon OpenAI's GPT3.5 model (OpenAI, 2024), tailored to enhance the CG capabilities of these LLMs using FSL and CoT. Despite initially was planned to use also Meta LLAMA2 (Meta, 2024) and nanoGPT (karpathy, 2024) architecture and models, unfortunately, due to the timing of this training course were discarded, but offers promising alternatives research lines.

2 Prior Literature Review

In the quest to deepen understanding of CG within LLMs, and the burgeoning field of combining LLMs with ABM applied to organizational structures, the literature review done has been selective yet comprehensive. This point focus on the main studies that have underpinned the

experiments and informed the approach taken. The selected papers build a foundation that supports our endeavor to harness the strengths of LLMs and ABMs for the new LLM Architecture and digital transformation of organizations.

Compositional Generalization: The review covers the challenge of CG within LLMs. An approach to address the lack of CG in ML, particularly in NLU, is proposed in (Daniel Keysers, 2020)**.** R&R, a data augmentation method discussed in (Ekin Akyürek, 2021)improves Compositional Generalization by recombining training examples and resampling. Regarding Few-Shot Learning (Xin Liu, 2024) express that LLMs can capture representations of concepts useful for real-world tasks. However, language alone is limited. While existing LLMs excel at text-based inferences, health applications require that models be grounded in numerical data that is not easily or readily expressed as text in existing training corpus.

Regarding Chain of Thought (Jason Wei, 2022) explores how generating a chain of thought -- a series of intermediate reasoning steps -- significantly improves the ability of large language models to perform complex reasoning. Experiments on three LLM show that CoT prompting improves performance on a range of arithmetic, commonsense, and symbolic reasoning tasks. The empirical gains can be striking. In summary, the articles encapsulate the progress in refining LLMs to better align with human behavior taking several work lines always in the scope of Compositional Generalization.

Agent Based Modelling: The article referenced as (Christopher W. Weimer, 2016) provides a foundational introduction to ABM, scoping it as a bottom-up simulation technique that emphasizes the autonomous interactions of agents within an environment. The article listed as (Singh, 2019) discusses the escalating interest among researchers in utilizing ABM to conceptualize and analyze complex systems. It outlines the benefits of ABM over Equation-Based Modeling (EBM), particularly for understanding dynamic interactions within systems. Further delving into the intricacies of ABM, the text by G. Wade McDonald and Nathaniel D. Osgood, mentioned in (Osgood, 2023), specifically addresses the challenge of comprehending the behavior of complex systems through the lens of agent interactions and the environment. The article articulates why traditional compartmental models are often inadequate for representing the heterogeneity of agents, the complexity of their networks and spatial contexts, and the significance of individual histories within these systems.

LLMs, ABM and Organization Modelling.

The study from (Mohammad Asfour, 2023) focuses on the simulation and understanding of human responses to social engineering attacks, specifically phishing attacks, through the lens of LLMs like OpenAI's GPT-4. In (Yuan Li, 2023), the paper explores the capacity of LLMs to coordinate within task-oriented social contexts, with a particular focus on a simulated job fair environment. It introduces collaborative generative agents designed to display human-like social behavior and collaboration skills, aiming to bridge the gap in understanding LLMs' potential for complex coordination tasks. Lastly, (Mingyu Jin, 2024) jumps into the speculative domain of simulating interactions between human and extraterrestrial civilizations through LLMs, inspired by concerns raised by Stephen Hawking. The study introduces "Cosmo Agent," an AI framework designed to explore the dynamics of inter-civilizational relations across diverse ethical and moral landscapes. The study conducted by Yuan Li, Yixuan Zhang, and Lichao Sun (Yuan Li, 2023) explores the development and application of "METAAGENTS," which are collaborative generative agents powered by Large Language Models (LLMs) such as GPT-4. These agents are designed to simulate human-like behaviors and capabilities, particularly focusing on task-solving abilities and coordination within a job fair context. The researchers developed a comprehensive framework for METAAGENTS that incorporates four key modules: perception, memory, reasoning, and execution. The results of the study reveal that METAAGENTS exhibit a promising level of performance in terms of information retrieval and coordination capabilities. However, the study also uncovers limitations that affect the agents' effectiveness in more complex coordination tasks.

This literature review casts a discerning eye over seminal works that have shaped the dialogue around Compositional Generalization in Large Language Models (LLMs) and Agent-Based Modeling (ABM) in the realm of organizational

structures. Through this review, the stage for the deployment of a new LLM architecture that seeks to blend the computational finesse of LLMs with the systemic dexterity of ABM.

3 Data

Here a description the dataset, of how dataset was built for the experiments.

3.1 Data Sources and Query

As the author employment and other projects he is involved are in the scope of Health, the Health Domain was the starting point to build a dataset.

Following these, Cancer Prostate and Sipuleucel-T product scoped the dataset. PubMed (PubMed, 2024) was selected as origin of data and the list of Prostate Cancer at the National Cancer Institute (NCI) (National Cancer Institute, 2024) to focus and narrow the data as will be later explained.

The query was designed to retrieve articles related to prostate cancer published from 2001 to the present. Below can be seen the query for PubMed interface:

```
("prostatic neoplasms"[MeSH Terms] OR ("prostatic"[All Fields] AND "neoplasms"[All Fields]) OR "prostatic neoplasms"[All Fields] OR ("prostate"[All Fields] AND "cancer"[All Fields]) OR "prostate cancer"[All Fields]) AND ( (ffrft[Filter]) AND (fha[Filter]) AND ({}/1/1:{}[pdat]))
```

The time frame selected was 2001-2024. This resulted in 61720 articles. After that as this quantity of articles was too much, some actions were taken to reduce the scope. It was reviewed the list of prostate cancer products (Institute, 2024) of NCI, and based on that list an screening of the articles was done. Finally it was selected articles regarding SIPULEUCEL-T, cause they had a good distribution along years. Also in order to narrow the scope of data to make it appropriate to the work of this course, only the abstract and title of the articles, where used, discarding the full article text.

3.2 Dataset

After all that and to summarize, **the FINAL DATASET of prostate cancer articles from PubMed and focused in Sipuleucel-T was 224 article abstracts**. The last step was to split it in 112 abstracts for generate context for future work and to prepare the questions used for evaluation of the models. The dataset was extracted focusing exclusively on articles that are freely accessible. This deliberate choice was guided by the principles of open science and the aim to ensure reproducibility and inclusivity in our research.

3.3 Purpose of the dataset

The dataset is intended for experiment on Artificial Intelligence targets, as LLM Models and ABM, and Digitalization, but not to make any research in the scope of pure Health domain.

The dataset is split in two sets of 112 articles in order to use one part to prepare the models and prepare the questions for evaluations, while the second half, 112 articles more, were used to test the LLM models and architectures of the experiments and get results in order to probe the initial hypothesis.

4 Models

4.1 LLM Models

4.1.1 Baseline Model

GPT3.5 (OpenAI, 2024), developed by OpenAI, served as the baseline and foundational model for the work. This model is part of the Generative Pre-trained Transformer family, noted for its effectiveness in natural language understanding and generation.

Other models were initially intended to be used as Meta-LLAMA2 (Meta, 2024) and nanoGPT (karpathy, 2024) but unfortunately, time needed for deployments overcame the scope of this work. It would be outlined as future work.

4.1.2 LLM Hybrid Model Architecture

A new LLM model architecture named as GPTH as developed, intended to enhance the capabilities of basic LLM models as GPT3.5 through the usage of Compositional Generalization (CG) techniques, as Few Shot Learning (FSL) and Chain of Thought (CoT). To implement this new architecture, instead of modifying the architecture at inner level, agent-based modeling software techniques where applied, that means that it was built the new model architecture named GPTH by integrating three building blocks: FSL block and CoT block on top of GPT3.5 block. The primary aim of GPTH is to augment the existing GPT3.5 model without modifications to its internal architecture, focusing

instead on input pre-processing strategies to improve the input query.

Below a description of three modules:

1. **Basic GPT3.5 Module:** This module simply gets a question and sends to OpenAI API selecting GPT3.5 Model.

2. **Few-Shot Learning Module:** This module preprocess input questions to simulate a few-shot learning scenario. It formats the inputs by incorporating contextually relevant examples, thereby priming the GPT-3.5 model to adapt its output to the query specifics efficiently and with accuracy. After that, the new input question is sent to the Basic GPT3.5 module.

3. **Chain of Thought Module**: This module takes a question input and preprocesses also but adding instruction to process applying Chain of Thought. Then, as in module 2, the new question is sent to the Module 1.

4.2 Digitalization Architecture

Additionally and as part of our experiments a basic Agent Based Modelling oriented framework model was implemented, leveraging the synergy between ABM and LLMs, in concrete the advanced language understanding capabilities of GPT-4, creating a dynamic interaction system between digital agents representing patients and doctors. This architecture model, encapsulating roles within the urology domain, signifies a substantial leap forward in simulating and streamlining patient-doctor interactions.

The architecture is composed of two primary agents: a **Urology Patient** and a **Urology Doctor**. Each agent is equipped with a profile, a defined set of actions, and the capability to interact with other agents within the system.

The `Urology Patient` can perform actions such as requesting appointments and providing feedback, mirroring the behavior and needs of real-world patients. Conversely, the `Urology Doctor` focuses on tasks aligned with the medical professional's responsibilities, including diagnosing, treating, and conducting follow-ups.

The integration with GPT-4 allows each agent to process natural language input and generate contextually relevant responses, facilitating a conversation flow that closely resembles human interaction. This seamless communication is facilitated by the `Messaging` system, a backend service that enables the exchange of messages and the coordination of tasks between agents.

5 Methods

In this section, the methodologies used to implement are presented. It was evaluated and compared the GPT-3.5 base model with the new enhanced GPTH hybrid model, and also the possibilities of combining LLMs and ABM for data digitalization.

The methods includes dataset preparation implementation of the model architectures in Python, PostgreSQL, OpenAI GPT3.5 API, execution of tests and saving of results. This process ensures the reproducibility of results. The method is split into two main points, one for LLM architectures and second for LLM/ABM digitalization.

5.1 Methods for LLM Architectures

5.1.1 Data download and Dataset preparation

Question preparation.

To assess the comparative effectiveness of the baseline GPT-3.5 model and the enhanced GPT3.5Hybrid model, the study started by generating a set of questions. This generation was performed using GPT-4, leveraging its advanced linguistic capabilities to produce two distinct types of questions: simple and complex. The use of GPT-4 for generating these questions ensured a high standard of linguistic quality and relevance.

5.1.2 GPTH Architecture Implementation

IT was implemented the GPTH Architecture model through a modular ABM approach designed to enhance the functionality of the basic GPT-3.5 model provided by OpenAI. This system integrates several components to enhance interaction and processing capabilities, leveraging Python's ecosystem APIs, database interactions, and object-oriented programming. Here's a breakdown of the design details:

Basic GPT3.5 Module: Serving as the core component that interfaces with the OpenAI API. It handles questions from FSL and CoT Modules, send to the GPT-3.5 engine and receiving the responses.

Few-Shot Learning and Chain of Thought Modules: Both modules preprocess the input user questions to enhance understanding and response

quality. The Few-Shot Learning Module adds contextually relevant examples to the input, while the Chain of Thought Module instructs the model to apply step-by-step reasoning. These modules modify the input queries by appending necessary details or instructions before passing them back to the Basic GPT3.5 Module.

Messaging Communication Module: A custom Python class manages the asynchronous communication between different components of the system, facilitating a decoupled architecture where modules operate independently yet interact seamlessly.

PostgreSQL Database: This is used for persistent storage of questions and their corresponding responses, allowing for robust data management and retrieval. The system interacts with PostgreSQL using Python's `psycopg2` (Gregorio, s.f.)library, which provides comprehensive functionalities for database operations.

The Python-based implementation ensured components can be independently developed, tested, and optimized. The use of PostgreSQL for storing interactions enabled data analysis and traceability of model performance over time. These features created a simple but versatile system.

5.1.3 Testing

The experimental setup involved executing the generated questions across different configurations of the language models to observe and compare their performance in handling varying complexities of queries.

- **Simple Questions Execution:** Simple questions were tested against GPT3.5, and against GPTH with application of FSL.
- **Complex Questions Execution**: Complex questions were executed through the GPT-3.5 base model and through GPTH with CoT Module.

Responses from all four cases were stored in PostgreSQL for later evaluation and analysis.

5.1.4 Metrics

To evaluate the quality and effectiveness of the model's answers, a set of five metrics where selected, described below:

- **Accuracy:** Factual correctness of the responses, ensuring that the information provided is accurate and reliable.
- **Relevance:** Measures the alignment of questions with the specifics of the queries posed, addressing the intended topic without deviation.
- **Comprehensiveness:** Measures the model's ability to provide detailed and complete answers, covering all aspects of the query without leaving critical questions.
- **Clarity:** Measures if model's responses are easy to understand, using straightforward, concise, and unambiguous language.
- **Coherence:** Measures the logical structure and flow of the response, ensuring that the model produces answers logically organized and progress naturally from start to finish.

5.1.5 Evaluation

The evaluation method was designed to systematically assess the performance of the GPT-3.5 base model and the GPTH Model against a series of metrics. The responses from both models were rated on five key metrics proposed. This rating process was automated using Python, PostgreSQL and OpenAI GPT4 API was used to apply GPT4 to do the rates. This allowed:

- Consistency: The rating process with GPT-4 should ensure a consistent application of evaluation criteria across all responses, eliminating human bias and variability in rating.
- Efficiency and Scalability: Using an automated process allows for the rapid processing of large volumes of data. And makes feasible to extend the testing framework to include more data or additional models.

Each response generated by the models was sent to GPT-4, which was tasked with applying these metrics. The system was configured to evaluate each response and provide a score for each metric, ranging from 0 to 100, where 0 indicates poor performance and 100 indicates exceptional performance.

Data Compilation and Analysis: The scores for each response were compiled into a dataset in PostgreSQL for later analysis.

5.2 Methods for Digitalization Architecture

This point presents the method for the second experiment run, that is, ABM and LLM combination for organization Modelling.

5.2.1 Implementation

The Agent-Based Digitalization Model was operationalized through the development of two digital agents: **`AgentUrologyPatient`** and **`AgentUrologyDoctor`**. These agents were implemented in Python applying ABM techniques and utilizing GPT-4 for natural language processing.

The patient agent was programmed to simulate patient behaviors such as requesting appointments and providing feedback on treatments, while the doctor agent was designed to respond with diagnoses, treatment plans, and conduct follow-ups. Each agent's behavior was defined by a series of potential actions encoded within their profile.

A message-passing persistent infrastructure was created to facilitate communication between agents, The architecture's modularity, facilitated by ABM, enabled rapid prototyping and iterative development.

5.2.2 Testing and Issues.

This experiment was limited due to time restrictions as first basic test shown that needed much more work, to provide capabilities as mainly Memory and controlling actions to avoid the agents enter in no ended loops.

6 Results

6.1 LLM Architectures

This chapter presents the results obtained from the evaluation of the GPT-3.5 base model and the GPTH model. The performance of those models was assessed based on five key metrics pointed out: Accuracy, Relevance, Comprehensiveness, Clarity, and Coherence. Each response was rated on a scale from 0 to 100. The table below summarizes mean and median for the answer for the corresponding metrics, for each model architecture.

		Simple Questions		Complex Questions	
		GPT3.5	GPTH FSL	GPT3.5	GPTH CoT
Mean	Accuracy	93,949	92,124	82,548	84,706
	Relevance	94,574	93,216	85,715	86,735
	Compr	88,239	84,769	78,084	81,737
	Clarity	88,001	86,659	83,074	83,695
	Coherence	89,905	89,705	82,349	83,787
Median	Accuracy	100,000	100,000	85,000	85,000
	Relevance	100,000	100,000	85,000	85,000
	Compr	90,000	95,000	85,000	85,000
	Clarity	85,000	85,000	85,000	85,000
	Coherence	95,000	95,000	85,000	85,000

6.2 Digitalization Architecture

In the development and deployment of the digitalization model involving Agent-Based Modeling (ABM) and Large Language Models (LLMs) to simulate interactions within the healthcare context, particularly within the Cancer Prostate Unit, the project faced significant constraints with respect to time. This limitation impacted the chances to gather a comprehensive set of results for in-depth analysis. This lack of extensive data collection and subsequent analysis leaded to the decision to not to present preliminary results as part of this project's outcomes

7 Analysis

This section delves into the comparative performance analysis of two distinct configurations of Large Language Models (LLMs): the standard GPT3.5 model and the advanced GPTH model, which incorporates Few Shot Learning (FSL) and Chain of Thought (CoT) enhancements. The evaluation is focused on their effectiveness in addressing simple and complex queries, reflecting on key metrics such as accuracy, relevance, comprehensiveness, clarity, and coherence.

Lastly, the role of Agent-Based Modeling (ABM) in enhancing the LLM architecture is considered. This analysis highlights how ABM contributes to the rapid and flexible integration of new functionalities like FSL and CoT into the existing GPT3.5 framework, enhancing the model's adaptability and performance without extensive modifications to its core architecture.

7.1 LLM Architectures

Simple Question to GPT3.5

Mean Scores: The model is performing well, with the lowest mean score being for Clarity.

Median Scores: Perfect median scores for Accuracy and Relevance indicate that at least half the responses are scoring the maximum, suggesting consistent high performance for these metrics.

Simple Questions to GPTH with FSL

Mean Scores: Slightly lower than the Simple Questions without FSL, for Comprehensiveness and Clarity. This might indicate that while FSL adds context, it may also introduce complexity to the answers which affects clarity and the completeness of responses.

Median Scores: Medians remain high, with perfect scores for Accuracy and Relevance, and an improvement in Comprehensiveness compared to the non-FSL condition, indicating a strong tendency for high performance, despite a slight drop in the average scores.

Complex Questions to GPT3.5

Mean Scores: Scores drop across all metrics compared to the Simple Questions, which is to be expected as the questions are more challenging. Comprehensiveness sees the largest decrease, indicating that providing complete answers to complex questions is particularly challenging.

Median Scores: Median scores are notably lower than the perfect scores seen in the Simple Question categories, but are still high (85,000 across all metrics), indicating that a significant portion of responses meet a high standard even not perfect.

Complex Questions to GPTH with CoT

Mean Scores: Introducing the Chain of Thought appears to improve performance across all metrics when compared to the Complex Questions without this module. This suggests that structuring responses to encourage a step-by-step approach is beneficial in tackling complex queries.

Median Scores: All median scores are the same as in the Complex Questions without Chain of Thought, suggesting that while the average performance is improved, the central tendency of high performance remains consistent.

Overall Interpretation: Across all scenarios, the models demonstrate a strong capacity to deliver accurate and relevant answers, as shown by the high median and mean scores. The addition of Few-Shot Learning appears to have a slightly negative effect on the mean scores for simpler questions, possibly due to increased complexity in the responses. However, the use of Chain of Thought with complex questions leads to improvements in the means, indicating its effectiveness for improving the handling of complexity.

7.2 Agent Based Modelling, for LLM Architecture

The integration of Agent-Based Modeling (ABM) techniques in the development of GPTH has emerged as a significant factor contributing to the model's enhanced performance. ABM provided a flexible and efficient framework to rapidly iterate on the GPT-3.5 architecture, facilitating the swift incorporation of enhancements without the need to delve into the underlying neural network's complexities. Some advantages of Agent-Based Modeling Approach:

Flexibility: ABM allowed for the modular addition of functionalities to the existing GPT-3.5 model. By treating Few-Shot Learning (FSL) and Chain of Thought as separate 'agents' operating on top of the base model, ABM avoided the intricacies involved in altering the model's internal architecture.

Rapid Development and Testing: The agent-based approach significantly reduced the development time. It allowed for parallel development and testing of enhancements, streamlining the process of determining the most effective strategies for different types of questions.

Adaptability: Using ABM made the system highly adaptable, as new agents can be added or existing ones modified to respond to evolving requirements or to incorporate new insights into model performance.

In summary, the agent-based development of the GPTH model not only facilitated enhancements that improved performance but also showcased a scalable and flexible approach to AI development. The improvements in handling complex questions, brought about by the Chain of Thought module, are indicative of the potential that ABM holds in advancing the field of AI. This methodology, while effective in the current context, also lays the groundwork for future explorations into creating even more nuanced and sophisticated AI models.

8 Conclusions

In the culmination of this work into the enhanced application of LLMs through Compositional Generalization techniques, the use of LLMs as modular building blocks, and the implementation of Agent-Based Modeling (ABM) to develop hybrid architectures, the research has produced insights in favor and expand upon the initial hypotheses. Here, conclusions that directly connect the research outcomes to the initial hypotheses are presented. .

H1: Enhancing LLMs by Compositional Generalization Techniques: Chain of Thought and Few Shot Learning: The findings are in favor that the integration of Compositional

Generalization techniques enhances the processing capabilities of GPT3.5..

H2: Advancing LLM Architecture Using LLMs as Building Blocks: The successful implementation of a hybrid model using GPT3.5 as a foundational component enhanced by additional modules for CoT and FSL illustrates the practicality and efficacy of using existing LLMs as building blocks without the need for extensive retraining or restructuring.

H4: Application of Agent-Based Modelling for Developing Hybrid LLM Architectures: The application of ABM in structuring our GPTH architecture has proven instrumental. By treating the enhancement techniques as independent yet intercommunicating agents, enabled to encapsulate specific functionalities within an overarching system effectively.

H5: Digitalization of Cancer Prostate Unit Based in Agent-Based Modelling and LLMs: The digitalization of the Cancer Prostate Unit using an ABM framework integrated with LLMs showcases the transformative potential of these technologies in a healthcare setting. Although preliminary, these results encourage further exploration and refinement of ABM and LLM integration in healthcare and other sectors.

Overall Conclusion

The research conducted provides evidence supporting initial hypotheses, demonstrating that LLMs enhanced with Compositional Generalization techniques can achieve higher levels of linguistic and cognitive performance. Additionally, the use of LLMs as building blocks within an ABM framework offers a robust method for developing sophisticated, adaptable, and efficient digital architectures. These findings support the theoretical propositions but also open avenues for practical applications in digital transformation.

Known Project Limitations

While our study has yielded promising results several limitations must be acknowledged.

Scope of Compositional Generalization Techniques: The enhancements introduced by the CoT and FSL techniques have shown improvements in model performance; however, their effectiveness is contingent upon the nature and complexity of the input queries. CoT requires well-structured and articulated prompts to guide the reasoning process effectively. Furthermore, the training data for FSL was limited to a specific domain (prostate cancer), which may affect the generalizability of the model to other domains or broader applications.

Integration of Agent-Based Modeling: While ABM provided a flexible and scalable approach to developing hybrid LLM architectures, the integration complexity can introduce new challenges. Coordination and communication between agents can become a bottleneck.

Data Dependency: The effectiveness of the models heavily relies on the quality and relevance of the training data. Given that our dataset was exclusively sourced from publicly available PubMed articles focusing on a specific medical treatment, the models' applicability to other dataset might be limited.

Ethical and Privacy Considerations: Deploying LLMs and ABM in real-world settings like healthcare, raises ethical and privacy concerns. The use of patient data must be controlled to comply with regulations such as GDPR and HIPAA.

Experimental Constraints: The experimental design, while robust, was limited by time and the scope of this study. Certain planned components, such as the MetaLLAMA2 and nanoGPT architectures, were not explored due to these constraints. The insights are therefore based on a subset of the potential model configurations.

Authorship statement

The author is currently serving as an independent AI Advisor specializing within the Health Domain. His role involves overseeing the strategic integration and management of AI technologies to enhance health systems. This project is conducted, independently and has no relation with his current work. **External Collaboration:** This research was independently designed and conducted with no external help or collaboration a part of the support provided from the course instructors. **Rationale for Single Authorship:** Given the unique position and expertise in the field, single authorship was deemed appropriate for this research. However, the author remains open to collaborative opportunities with the aim to enrich the research and its applicability in real-world settings.

References

Christopher W. Weimer, J. O. (2016). Agent Based Modelling: An Introduction. *Proceedings of the 2016 Winter Simulation Conference.*

Daniel Keysers, N. S. (2020). Measuring Compositional Generalization: A Comprehensive Method on Realistic Data. *arxiv > Computer Science > Machine Learning*.

Ekin Akyürek, A. F. (2021). Learning to Recombine and Resample Data for Compositional Generalization. *arxiv > Computer Science > Computation and Language*.

Giabbanelli, P. J. (2023). GPT-Based Models Meet Simulation: How to Efficiently Use Large-Scale Pre-Trained Language Models Across Simulation Tasks. *Arxiv > Computer Science > Human-Computer Interaction*.

Institute, N. C. (2024). *National Cancer Institute - List of prostate cancer medicianl products*. Obtenido de https://www.cancer.gov/about-cancer/treatment/drugs/prostate

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

karpathy, A. (30 de 03 de 2024). *Github - nanoGPT*. Obtenido de https://github.com/karpathy/nanoGPT

Lake, B. M. (2019). Compositional generalization through meta sequence-to-sequence learning. *https://proceedings.neurips.cc/*.

Meta. (30 de 03 de 2024). *Meta LLAMA*. Obtenido de Meta LLAMA: https://llama.meta.com/

Mingyu Jin, B. W. (2024). What if LLMs Have Different World Views: Simulating Alien Civilizations with LLM-based Agents. *Arxiv > Computer Science > Computation and Language*.

Mohammad Asfour, J. C. (2023). Harnessing Large Language Models to Simulate Realistic Human Responses to Social Engineering Attacks: A Case Study. *International Journal of Cybersecurity Intelligence & Cybercrime*.

Najoung Kim, T. L. (2020). COGS: A Compositional Generalization Challenge Based on Semantic Interpretation. *Arxiv > Computer Science > Computation and Language*.

National Cancer Institute. (30 de 03 de 2024). Obtenido de https://www.cancer.gov/

OpenAI. (30 de 03 de 2024). *Open AI*. Obtenido de Open AI: https://platform.openai.com/docs/models/overview

Osgood, G. W. (2023). Agent-Based Modeling and its Tradeoffs: An Introduction & Examples. *Arxiv > Computer Science > Multiagent Systems*.

PubMed. (03 de 03 de 2024). Obtenido de https://pubmed.ncbi.nlm.nih.gov/

Singh, K. (2019). *Introduction to Agent-Based Modeling.* Obtenido de https://dzone.com/articles/introduction-to-agent-based-modelling

Thaddäus Wiedemer, P. M. (2023). Compositional Generalization from First Principles. *Advances in Neural Information Processing Systems 36* .

Xin Liu, D. M.-L.-Z. (2024). Large Language Models are Few-Shot Health Learners. *Arxiv*.

Yuan Li, Y. Z. (2023). MetaAgents: Simulating Interactions of Human Behaviors for LLM-based Task-oriented Coordination via Collaborative Generative Agents. *Arxiv > Computer Science > Artificial Intelligence*.

A Supplementary Material

The code developed for the experiments is available in GitHub: https://github.com/jbruks/NLU_XCS224U/tree/main/implementation