Hybrid Rule-Based AI Agents in Autonomous Procurement Negotiations

MSc Thesis Proposal

Author: Kristians Davidsons (592162)

Thesis Supervisor: Klaus Garrido Tenorio

Date of Submission: 01/03/2025

Table of contents:

1. Introduction:

With rapid advancements in computing power and methods used to develop Conversational Artificial Intelligence (AI) Agents, the private and commercial adoption of intelligent dialogue systems for various tasks has increased drastically (Kulkarni et al., 2019). Consequently, these systems have evolved from speculative prototypes to revolutionary technology that if leveraged correctly provides companies with sustainable competitive advantage (Y. Cui et al., 2024). One of such areas in which the application of AI can create a competitive edge is procurement, with generative AI completely transforming core procurement functions, like category management, supplier management, risk mitigation and spend optimization (Mittal et al., 2024). AI creates value within this domain through automation and smartness (R. Cui et al., 2022). It can automate simple, repetitive, low value tasks allowing them to be completed faster and cheaper, while continuously learning and making decisions to achieve positive business outcomes (R. Cui et al., 2022). Therefore, it is essential for organizations to understand how this technology works, when and how to apply it and pinpoint where it delivers the greatest value, allowing them to remain competitive or even create a strategic advantage.

Negotiations in particular have been regarded as a time-consuming procurement process which could benefit greatly from efficiency improvements (X. Xue et al., 2007). Traditionally, negotiations have been conducted between humans. However, for large companies that have thousands of different suppliers this is simply not possible (Van Hoek et al., 2022). As a result, a situation is created where suppliers do not get to negotiate with the buying organization and are forced to accept predetermined terms instead (Van Hoek et al., 2022). Thus, many companies have already applied AI-powered bots to negotiate with “Tail-end” suppliers (Bajaj et al., 2023). This development raises an interesting scenario: as buyers increasingly employ AI to automate their negotiations, suppliers may seek to automate their side of the process, producing a situation where deals are negotiated without any human intervention on either side. Hence, exploring the dynamics and outcomes of these fully automated bot-to-bot interactions is essential for understanding the shortcomings of state-of-the-art AI-powered agents and for real-world application of this technology (Davidson et al., 2024).

Previous research has analyzed various AI agent negotiation-based games. However, many of these studies have exhibited significant drawbacks, such as AI agents relying on static strategies or assuming perfect information, limiting their applicability to real-world scenarios, as highlighted by Lopes et al. (2008). More recently, Davidson et al. (2024) suggests that these flaws of standardized inputs and outputs can be overcome with the advances in natural language understanding and processing. Indeed, several studies over the past years have used Language Model (LM) based autonomous agents in negotiation game settings to increase agents’ flexibility and create more human-like interactions (Fu et al., 2023; Brookins & DeBacker, 2023; Guo, 2023).

Despite the advantages of LM-powered agents over static models, they also face significant hurdles in recreating human behavior, most notably AI hallucinations. Hallucinations in the context of LMs describe situations where the model returns incorrect or false information, deviating from factual knowledge and potentially producing responses that may not be grounded in the model's training data (Perković et al., 2024). These hallucinations can have detrimental effects on business outcomes. For instance, an LM-powered chatbot could spread misinformation, generate flawed insights that lead to suboptimal decision-making or fail to comply with legal standards (Taplin, 2024). Such errors could be particularly damaging for businesses that use LMs in negotiations, as unfavorable deals could be made by LM agents with deal-making authority. Therefore, mitigating hallucinations is essential for automated negotiations to be implemented in practice.

To minimize hallucinations, previous studies suggest several approaches, including adjusting and fine-tuning prompts to avoid ambiguities, employing Retrieval-Augmented Generation (RAG) methods, and altering hyperparameters (Perković et al., 2024). Research shows that similar mitigation approaches can only reduce hallucinations but also improve LM performance. For example, Wei et al. (2022) found that dynamic prompting can significantly improve LM performance compared to static prompting. Further, S. Yao et al. (2022) found that with dynamically adjusted examples LMs can outperform comparable models with an absolute success rate of 10%. Dynamic prompting is often combined with robust role-based systems that, based on the conversation state, determine the appropriate prompt modifications for the LM model (Dong et al., 2024; Vertsel et al., 2024). Such hybrid LMs significantly outperform traditional role-based and LM systems (Vertsel et al., 2024). Consequently, examining hybrid LM architectures is highly relevant for applying autonomous agents to practical scenarios.

Thus far literature has mostly focused on investigating hybrid bots in isolated reasoning tasks or human-to-bot interactions. However, considering that in future bot-to-bot negotiations will become more prominent (Zhuge et al., 2023), examining performance and efficiency of such interactions is becoming increasingly more relevant. Hence, the following research objective is proposed:

- Research objective: Investigate whether a hybrid, rule-based negotiation agent demonstrates a superior negotiation performance compared to agents that rely solely on rule-based or generative AI methods, when accessed on efficiency in reaching Pareto-optimal agreements.

From this research objective the research question, with the corresponding sub-questions are posed:

- Main research question: Is the negotiation performance of a hybrid negotiating agent (rule-based + generative AI) superior to that of both rule-based and generative AI agents, when assessed based on efficiency in reaching Pareto-optimal agreements?
    - How do the negotiation outcomes (negotiated price and quality) differ between hybrid LLM agents and statically prompted LLM agents?

- How does AI agent architecture affect hallucinations within the conversations?

2. Literature Review:

This section provides an extensive overview of the literature related to this study. It begins with a thorough description of the search and selection methods used to compile the necessary research. The section is then divided into three primary segments: AI in Procurement and Negotiations (2.1), LLM Bot-versus-Bot Interactions (2.2) and LLM Hallucinations (2.3). Each section describes, analyzes and contrasts the relevant academic literature, ultimately leading to a summary of the findings in section 2.4. These sections are structured in a narrative format, with each building up to the next section and highlighting the relevance of this study.

In conducting this literature review, the author adopted a systematic strategy in order to collect all relevant literature and position this work within the existing body of literature. Firstly, the preliminary research was conducted to define and refine the research objectives and build up the information pool. To start with, information search was performed using "Google Scholar" to identify the appropriate search terms. After that, several filters were applied in "Elsevier" and "Web of Science" databases, to find relevant studies. However, considering that this study examines a novel topic, for some search terms, these databases did not have enough matches, as the relevant literature has not been published yet. Thus, the search was expanded to "Google Scholar". This was particularly the case for sections 2.2 and 2.3. Considering that much of the relevant academic literature has not gone through the review process, special care was exercised to ensure that work is consistent with the scientific standard.

After compiling the relevant findings, the main conclusions and methods used by each paper were extracted. Then, every article was meticulously evaluated for potential inclusion in the review. Subsequently, if no matching, high-quality research was found, the search terms were made more general and refined to find more applicable, high-quality research.

2.1 AI in Procurement and Negotiations:

Prior research has shown that artificial intelligence (AI) systems create significant value within the procurement domain. R. Cui et al. (2022) categorizes this value into two main aspects: automation and smartness. In terms of automation, AI can streamline repetitive time-consuming tasks, completing them more efficiently and cost-effectively (R. Cui et al., 2022). Beyond automation, AI can enhance decision-making by continuously learning, analyzing data and optimizing actions to achieve business goals (R. Cui et al., 2022). Interestingly, Boute & Van Mieghem (2020) proposes that automation implemented without smartness can backfire. This is also confirmed by findings of R. Cui et al. (2022). The author of this study concludes that most recent previous studies focus on AI creating value through both of the aforementioned aspects. As a result, AI can be applied to strategic, tactical, and operational procurement decisions.

Obinna & Kess-Momoh (2024) found that nearly 80% of procurement professionals incorporate AI in the procurement process, with the most common applications being in the contract management spend analysis and supplier selection. In the contract management field, several studies have successfully applied AI to increase procurement performance. For example, Yap et al. (2024) found that a pre-trained Natural Language Processing (NLP) model can be used to extract information from supply chain contracts, correctly identifying incoming inspection, payment terms and warranty period. Likewise, J. Zeng et al. (2024) demonstrated how AI-driven software can be used to reliably review contracts, increasing the efficiency and reducing mistakes when managing supplier contracts. Although in both studies the results were promising, they still faced some limitations, like unreliability in identifying specific contract sections or the lack of personalized outputs. Some studies have shown that by using deep learning approaches, such as Convolutional Recurrent Neural Networks, a fully automatic facilitating contract management platform can be created, reducing manpower needed and human errors (J.-H. Chen et al., 2021). Collectively, these findings imply that while AI has the potential to reduce human error and improve cost efficiency, further research and more flexible algorithms may be necessary to implement these approaches in practice.

Most AI-driven supplier selection literature leverages data from supplier spend analysis often combining AI techniques like fuzzy logic with multi-criteria decision-making (MCDM) to support supplier selection decisions (Sharma et al., 2022). For instance, Kumar et al. (2003) used a fuzzy, mixed integer, goal programming method to vendor selection problem (VSP), with multiple objectives of minimizing cost, rejections and late deliveries, allowing to solve VSPs in a supply chain where goals are not clearly stated. Other studies have similarly applied variations of fuzzy set methodology combined with MCDM to address supplier selection challenges (Boran et al., 2009; Büyüközkan & Çifçi, 2011; C. T. Chen et al., 2006; Memon et al., 2015). Using a different approach, Kuo et al. (2010) developed a hybrid neural network (NN) and a multi-attribute decision analysis (MADA) model to select green suppliers, finding that this hybrid model has greater power of discrimination and noise-insensitivity in determining their performance. Similarly, Aksoy & Öztürk (2011) found that NNs can perform better than traditional algorithms for selecting just-in-time suppliers.

The application of AI has an impact on interactions with suppliers, as concluded by R. Cui et al. (2022). In their study R. Cui et al. (2022) found that smart, automated AI requests for quotation systems achieved a better price quote than the human counterpart. This naturally extends to negotiations, where AI driven tools are increasingly being used to automate negotiations, optimize strategies and enhance decision-making in real time.

Studies that have investigated autonomous agent negotiations have been around since the 1990s (Lopes et al., 2008). Traditionally, AI negotiation research can be divided into four key groups: preliminaries, that address the fundamental nature and inherent conflicts of negotiations; pre-negotiations, which involve preparation and planning of the negotiation process; actual negotiations during which offers, counteroffers, and feedback are exchanged; and renegotiations that entail analyzing and refining the final agreement (Lopes et al., 2008). For the purposes of this study, only the negotiation category will be discussed. In earlier literature, researchers have explored various methods for autonomous agent negotiations. For example, Sycara (1993) developed an argument-based negotiation support system for resolving conflicts in the labor relation field, by generating proposals, modifying rejected proposals by the counterparty and creating persuasive argumentation. Further, Parsons et al. (1998) proposed a framework that enables autonomous agents to negotiate by exchanging proposals, critiques, counter-proposals and explanations, building arguments to justify their positions. Additionally, an algorithm that would allow autonomous agents to make trade-offs within multi-dimensional context, using fuzzy logic and hill-climbing technique was developed by Faratin et al. (2002). This algorithm was extended by Coehoorn & Jennings (2004) incorporating kernel density estimation for learning opponent preferences, resulting in improved negotiation outcomes. D. Zeng & Sycara (1998) created a sequential decision-making negotiation model that uses Bayesian learning to update its beliefs about the other party based on offers exchanged. The experimental results from this study show that, when both agents learn from each other they achieve better agreements (D. Zeng & Sycara, 1998). However, these studies had significant flaws and findings are not as relevant for contemporary settings. Earlier AI research in automated negotiation settings often relied on overly simplified human behavior and rationality, leading to experiments in which hard-coded agents operated in isolation, limiting their flexibility and real-world applicability (Baarslag et al., 2017). Moreover, these studies often overlooked the dynamic nature of negotiations, creating static bots that failed to account for evolving user preferences and the context (Baarslag et al., 2017). Davidson et al. (2024) suggests that with recent advances in LMs, there has been increased attention in the literature to explore deep learning in negotiations, allowing for dynamic and context-sensitive interactions to be made.

Sunder et al. (2018), used a reinforcement learning based agents to negotiate contract terms with human counterparts. In this study, the bots were trained by negotiating with each other and afterwards with humans, finding that purely selfish bots tended to outscore their negotiators, while prosocial bots were typically outperformed by human negotiators (Sunder et al., 2018). Similarly, in a different study, deep Q-networks have been used to create an agent that learns the optimal bidding strategies, outperforming human counterparts in terms of average utility (S. Chen & Su, 2022). Many studies have used large language model (LLM) powered bots in negotiations to investigate various negotiation outcomes, depending on the bot set up (Kwon et al., 2024). However, most recent research focuses on single-issue bot-to-bot negotiations with no human involvement (Davidson et al., 2024). This is similar to what will be done in this study.

2.2 LLM Bot-versus-Bot Interactions:

Large language models (LLMs) are a type of artificial intelligence trained on vast amounts of text data to understand, generate, and recognize patterns in human language (Hadi et al., 2025). These models excel at a wide range of natural language processing (NLP) tasks, that with fine-tuning can achieve high performance in areas such as text generation, question answering, and language translation (Hadi et al., 2025). LLMs operate by assigning probabilities to sequences of words using statistical and probabilistic methods and predicting the next word by maximizing the likelihood of the observed sequence (Singh & Mahmood, 2021). Most modern LLMs rely on transformer architecture, where raw text is tokenized into numerical embeddings with positional encoding and processed through many neural network layers (Raiaan et al., 2024). This architecture enables the models to capture long-term dependencies, allowing them to generate context relevant outputs (Raiaan et al., 2024). Within this study LLaMA language model, developed by Meta AI will be used. This model was chosen over alternatives because it is a powerful system trained explicitly on publicly available data and is open source, thereby avoiding any licensing costs (Touvron et al., 2023). Additionally, recent state-of-the-art studies have employed the LLaMA model in bot-to-bot negotiation settings (Abdelnabi et al., 2023; Davidson et al., 2024; Moghimifar et al., 2024).

In general, due to increased LLM capabilities and the relatively low technological requirements for running LLM models, there has been a surge in academic literature exploring various autonomous LLM-based interactions, where automated bots serve as all participants. For instance, LLM driven generative agents can simulate believable human behavior in a small sandbox town, with each agent having their own memories and ability to interact with other LLM agents occupying the town, in order to explore emerging social dynamics (Park et al., 2023). Further, G. Li et al. (2023) presents a cooperative role-playing framework that leverages LLMs and inception prompting techniques to enable AI agents to collaborate on complex tasks, with minimal human intervention. While these articles, similarly to this thesis, examine autonomous LLM powered bot-versus-bot interactions, they focus on generic social behavior. Within this study, AI agent performance will be evaluated on objective outcomes - the negotiated price.

Many contemporary studies have used objective metrics to evaluate performance of autonomous AI interactions. Examples include economic games, where agents' outcomes are compared to theoretically optimal strategies or human behavior (Akata et al., 2023; Brookins & DeBacker, 2023; Fontana et al., 2024; F. Guo, 2023; S. Guo et al., 2024), board games, where agents are evaluated based on win rates and strategic decision-making (Light et al., 2023; Martinenghi et al., 2024; Xu et al., 2023) and negotiation tasks, where the performance is measured through negotiated outcome. The latter is most relevant for this study and hence will be explored further.

In their study Fu et al. (2023), examined if different LLMs can improve their negotiation outcomes autonomously, by receiving feedback from the third, independent LLM. They find that only a few of the

LLM models considered in the paper improve their performance, with weaker models either not understanding the negotiation rule or not including feedback provided (Fu et al., 2023). Moreover, within their study, Bianchi et al. (2024) introduced multi-term negotiation scenarios to compare performance between different LLMs and negotiation strategies, finding that autonomous LLM negotiation outcomes are significantly dependent on behavioral strategies applied by LLMs. In particular, bots pretending to be desperate or using insults can improve their payout by 20% compared to the base model (Bianchi et al., 2024). However, these results should be interpreted with caution, as the prompts were specifically designed for the Claude LLM and have not been equally adapted for other models, potentially creating biases in the outcomes (Bianchi et al., 2024). Interestingly, another study finds that assigning Big Five personality traits to LLM agents significantly influences negotiation outcomes (Noh & Chang, 2024). Specifically, this study concludes that agreeableness affects negotiations the most, with bots that have high agreeableness being best at concluding a deal, while bots with low agreeableness get maximum value compared to other bots when a deal is made (Noh & Chang, 2024). These findings are confirmed by Huang & Hadfi (2024), who show that agreeableness has the strongest impact on negotiation outcomes, especially if the bot takes the seller's role. In a different setting, Abdelnabi et al. (2023) examine how several LLMs, like GPT-4, Llama3 70b Chat and Gemini Pro, perform in complex multi-party negotiations with agents having non-identical goals and agreement needs to be reached on multiple issues. They find that all open-source models are less successful than GPT-4, which still underperforms as the negotiations increase in complexity (Abdelnabi et al., 2023). This underperformance in increasingly complex negotiation settings underscores the importance of optimizing LLM outputs which has not been the focus of the aforementioned studies.

For example, although Fu et al. (2023) enhances the mediator performance through prompting and negotiator performance through mediator feedback, it does not incorporate a mechanism to verify the negotiators' outputs. Further, returning previous conversation history to the negotiators significantly degrades LLM performance (Hosseini et al., 2024). Likewise, none of the studies mentioned above have systematically dissected and analyzed AI agent outputs for inconsistencies, which may compromise the quality of LLM responses and limit the practical application of autonomous bots. Given that LLMs are black-box models, the output always needs to be checked if LLM did not make a mistake (Williams & Huckle, 2024). Additionally, issues with output are compounded by the length of the text given as prompt, which decreases the response quality (Williams & Huckle, 2024). Therefore, it is important to curate LLM responses so that prompts are not too long and are specific, while the outputs are checked for their quality. This is what the author aims to achieve with this research.

2.3 LLM Hallucinations:

In literature, inconsistencies or illogicalities in the generated text, failure to follow user prompts, or contradictions with previously returned responses are known as LLM hallucinations (Zhang et al., 2023). Zhang et al. (2023) identifies several reasons for LLM hallucinations, namely, reliance on massive, unfiltered training data, challenges of evaluating performance given the versatility of LLMs, and the generation of false but plausible information. Considering the exponential increase in practical applications and reliance on LLMs in recent years, it is crucial to ensure that the response generated by a LLM is accurate and free from hallucinations (Chui et al., 2023). Naturally, many studies have examined how to mitigate this phenomenon. Zhang et al. (2023) proposes a number of alleviation strategies, like hallucination mitigation during training or supervised fine tuning, but these approaches are less relevant for this study, as a pre-trained model will be used. Instead, solutions outlined by Perković et al. (2024), including Retrieval-Augmented Generation (RAG), model configuration, and prompt modification, are more appropriate.

RAG method improves LLM performance by retrieving relevant documents from an external database to provide context and enhance the model's response (J. Li et al., 2024). This method has proven effective across various fields, particularly those requiring specialized knowledge. For example, within the electronic design automation industry, an assistant powered by LLM and enhanced with RAG, demonstrated an overall improvement of recall by over 40% compared to its non-RAG counterpart (S. Yao et al., 2022). Moreover, Wang et al. (2024) developed the BioRAG model which integrates RAG with LLM to enhance biological question reasoning, finding that performance of BioRAG is superior when dealing with complex biological queries. However, using RAG to enhance LLM performance has drawbacks. Firstly, RAG systems have high training costs (Melz, 2023). Secondly, RAG models using external corpora introduce vulnerabilities to LLMs, as poisoning the retrieval databases (i.e., modifying them with malicious intent) can lead to inaccurate and even harmful answers by LLMs (J. Xue et al., 2024). Thirdly, with traditional RAG, LLMs can become overly reliant on the external retriever, causing LLMs to underutilize their own knowledge, potentially reducing performance and responsiveness of the model (C. Yao & Fujita, 2024). Fourthly, RAG techniques may introduce low quality or irrelevant information, increase LLMs noise, and consequently decrease the quality and efficiency of the generated content (C. Yao & Fujita, 2024). Lastly, and most importantly, this study does not require specialized knowledge nor very accurate responses from LLMs. Hence, RAG techniques are not the most appropriate to minimize hallucinations within this context.

Adjusting the model's configuration involves modifying parameters such as temperature, frequency and presence penalty (Perković et al., 2024). Sampling temperature hyperparameter controls the randomness of the model's output (Perković et al., 2024). Lower temperatures yield more deterministic results by returning more likely predictions, whereas higher temperatures introduce randomness, allowing less likely predictions to be generated (Renze & Guven, 2024). In other words, low temperature makes the model produce repetitive, focused and less diverse text, while high temperature creates unconventional and less probable text (Renze & Guven, 2024). Although, Renze & Guven (2024) find that sampling temperature does not produce statistically different results in answering multiple choice questions, Lee (2023) concludes that creativity can lead to more hallucinated output. Since there is no practical benefit of increasing creativity of the bots, within this study the temperature of bots will remain low.

A handful of studies have investigated prompt modification to improve LLM performance and reduce hallucinations. For example, Wei et al. (2022) demonstrated that by augmenting prompts, through chain-of-thought (CoT) reasoning - providing a series of intermediate reasoning steps - drastically improves the LLM's ability to solve complex tasks. Further, Fu et al. (2022) improved upon the CoT reasoning by proposing a complexity-based prompting, which uses only the most complex reasoning chains, rather than treating all chains equally, which substantially increases the LLM performance. For a comprehensive summary of LLM prompting refer to Qiao et al. (2023).

Prompt modification has been used in various autonomous negotiation studies. Chatterjee et al. (2024) introduced a framework called AgreeMate, that trains LLMs to negotiate prices using natural language by combining prompt engineering, fine-tuning and chain of thought prompting, to enhance the model performance. They find that their framework allows models to achieve agreement rates, fairer and less biased outcomes and more efficient negotiations (Chatterjee et al., 2024). Davidson et al. (2024) investigated various language models in multiple negotiation-style games, finding that cooperative bargaining games were the most difficult for the models, with model performance varying according to the specific game. The study employed static prompting, where a specific prompt would be used to enhance the message sent from one bot to another (Davidson et al., 2024). This was done to ensure that LLM bots understood their assignment and avoided illogical messages being generated (Davidson et al., 2024). Although these studies use prompt modifications to reduce hallucinations in LLM-powered agents, they also have notable limitations. For example, they use static prompting, where pre-defined prompts are used to specify the agent's task. This limits the flexibility of the bot, meaning that is unable to adapt to changing circumstances that may decrease their practical ability. Hua et al. (2024) addressed this limitation by introducing a third, separate remediator agent into their study, which intervened in the negotiation process to ensure that the language used by the negotiating agents adhered to social norms. In case social norm violating message was generated by one of the negotiating agents, remediator would interfere and modify the message so that it conforms to the norms (Hua et al., 2024). Dynamic in context examples were provided for remediator on how to refine these messages, and this approach proved effective in remedying social norm violations (Hua et al., 2024). However, within this study, the mediator did not aim to improve offer or check if the offers generated are profitable for the agents, rather focusing on controlling for social norm violations. Additionally, for remediators entire conversation history was returned which could significantly decrease the remediator's quality of response (Hosseini et al., 2024). Regardless, this approach can be adapted for reducing hallucinations in offers sent and received.

None of the studies that aim to address hallucinations through dynamic prompting incorporate a mechanism to verify bot responses after text generation. This means that these agents may produce false information despite specific, clarifying prompts, creating a gap in current literature. Within this study the author aims to address these limitations by creating and testing rule-based, dynamically prompted autonomous negotiation agents.

2.4 Conclusions from Literature Review:

In conclusion, the literature review underscores the growing academic and practical interest in leveraging artificial intelligence (AI) to streamline business functions through enhanced automation and outsourced intelligence. This trend is particularly evident in procurement, where AI has been widely applied in contract management, spend analysis, supplier selection and negotiations. Researchers have shown considerable interest in AI negotiations, which have evolved from static, rule-based agents to complex and flexible large language model (LLM) powered autonomous negotiators. These contemporary agents have shown a lot of promise in experimental settings, often even outperforming their human counterparts. However, practical deployment of autonomous negotiation agents still remains far, as these models often hallucinate, which makes granting them decision-making power risky. Therefore, recent studies have focused on various strategies to mitigate these hallucinations, often applying static methods. Moving forward, the author will propose a rule-based large language model that aims to minimize the agent hallucinations, while leveraging the flexibility of LLMs.

3. Research Objective and Research Questions:

The goal of this study is to develop a rule-based, dynamically prompted Large Language Model (LLM) driven negotiation agent that can perform in simulated negotiation settings and compare the performance of this agent with benchmarks - static, rule-based bot and inflexibly prompted LLM agent. This performance will be examined through two dimensions - negotiation efficiency and the agreement reached. The evaluation will focus on different negotiation scenarios, where static, rule-based agents, static LLM agents and rule-based, dynamically prompted LLM agents negotiate with one another and their own type, allowing for a comprehensive comparison of their performance. Within all the negotiation scenarios each bot will take both the buyer and the seller roles. Sub-goals of the study are to analyze the negotiation outcomes of each interaction and to examine dialogues of the LLM bots to see if any hallucinations within their conversations occur. All of the research questions can be found at the end of section 1.

The scope of this study is to examine the interactions of autonomous negotiating agents with different architectures. Besides the three bot architectures mentioned previously, other agent configurations will not be introduced. The study will focus on negotiations within procurement setting, where one bot will take a wholesaler's perspective (supplier's) and the other will take retailer's (buyer's) role. The performance of these architectures within different settings will not be examined. Further, within negotiations, both parties have to agree on price and quality of 10kg bag of wood pallets, with there being several possible Pareto efficient offers that maximize gain for both parties. Rule-based, static bot and rule-based, dynamically prompted LLM agents will have information regarding Pareto efficiency, while statically prompted LLM agents will not have access to this information. Both of these bots are designed to make and accept only Pareto efficient offers, by first determining the counterparties constraint. Other experimental set-ups are not considered in this study. Illustration of Pareto efficient offers for a rule-based buyer, who has the market retail price of 9 is shown in figure 1.

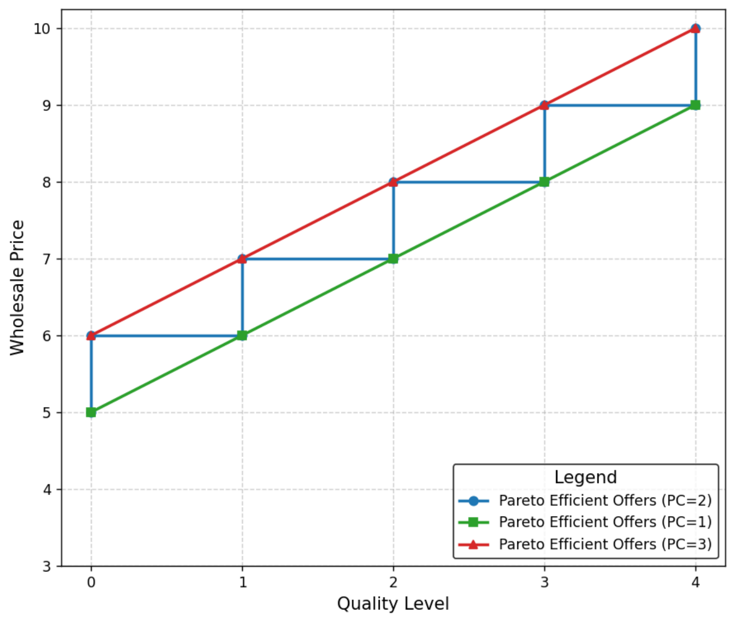

Figure 1. Pareto Efficient Frontier for Different Production Costs (PC).

4. Research Strategies and Data:

To answer the research questions, a simulation research strategy will be used. In particular, bot-to-bot negotiations will be simulated where bots with varying architectures and distinct roles engage in negotiations. The simulation setting in more detail is described in section 3. Quantitative primary data will be collected through these simulations and subsequently analyzed using statistical methods to assess the effects of different conditions.

Within the research several types of data will be collected. Firstly, a unique identifier will be recorded at the end of every negotiation. For each conversation two entries within the database will be made consisting of standard information like negotiated deal or interaction history and the data gathered about their counterparty. Therefore, unique identifiers ensure that interaction data is effectively linked and traceable.

Information regarding the negotiation setup will also be documented, including the date and time of the interaction's conclusion, the bot architecture, the bot role, and the constraints of both agents (such as production costs for sellers or retail prices for buyers). Additionally, data on negotiation outcomes, including both the negotiated price and quality, will be recorded to address the research questions and assess whether Pareto efficiency was achieved. Subsequently, agent-specific data will be collected. This includes, bot's perception of its own constraint, its belief of the counterpart's constraint, and the corresponding profits for both parties based on the negotiated terms. This is important to examine in order to identify any instances where a bot might have been hallucinating. Efficiency metrics, including the start time of the conversation, the number of offers made, and the number of messages exchanged, will also be tracked to evaluate the agents' overall efficiency. Finally, the actual conversational interactions and offers sent to different parties will be recorded. This is done for transparency and to examine potential hallucinations.

This data will be recorded only for static, rule-based agents and rule-based, dynamically prompted LLM agents, as these systems are designed to capture such information. For instance, to make a Pareto efficient offer the rule-based, dynamically prompted LLM agent would need to obtain its negotiation counterpart's offer, thus this information is naturally stored. Nevertheless, for statically prompted LLM agents, there are no prompts in place that would guide them to obtain the other counterparties constraints, and no rules integrated in the design to evaluate Pareto efficiency of the offers. Consequently, for static LLM agents, neither the estimated constraints (of either party) nor the corresponding profit estimates will be collected. Additionally, it is important to mention that in static LLM versus static LLM negotiations, data collection will be done by a third, independent LLM bot, that will determine when a deal is finalized, compile the offers made and store final negotiation terms.

5. Time Schedule:

Given the available resources and time constraints, the following project timeline is proposed. Since the project employs a simulation research strategy, there is no need for preliminary data collation, cleaning, or analysis; instead, the initial focus is on creating functional autonomous agent conditions. From January until the end of February, the statically prompted and dynamically, rule-based LLM agents were created and conversations between those two were enabled. During weeks 10 and 11 of 2025, attention will shift to the development of the rule-based bot architecture, culminating in the integration of a static rule-based agent designed to negotiate with a rule-based, dynamically prompted LLM agent. After, in the following week (week 12), rule-based agents' architecture will be integrated with a statically prompted LLM agent, as in this architecture the responses are less structured and thus may require changes in static rule-based bot architecture to account for this variability. Then, in the next week (week 13), static, rule-based architecture will be integrated to allow for negotiation with itself.

Over weeks 14, and 15, interactions between rule-based, dynamically prompted LLM agents and statically prompted LLM agents will be enabled within their respective architectures. Special attention will be given to enable agents to interact with themselves, because, in negotiations involving rule-based, dynamically prompted LLM agents, these bots initiate conversations by posing questions before actual negotiations begin. A similar initiating mechanism will need to be developed for statically prompted LLM agents and rule-based agents. Then, in the following two week (week 16) the interaction code will be optimized and prompts adjusted based on the academic literature, to improve agent performance.

Once bot-to-bot conversations are enabled, simulations will be conducted in a structured sequence. Firstly, the conversations between rule-based, dynamically prompted LLM agents and all other conditions will be simulated (week 13 - 18), with each agent taking both of the corresponding negotiation roles. Then, conversations between statically prompted LLM agents and static, rule-based agents will be simulated (week 19). Within the same week, negotiations between static rule-based agents will be conducted. After that, week 20, will be dedicated to data analysis. Finally, in weeks 23 and 24 the project will be finalized and submitted.

Throughout the timeline, work on the thesis project will progress in parallel with the experimental part of the project. The research design will be completed by the end of April 2025, and the results section will be finalized by week 20. Then the conclusions and discussion by week 23. The final project will be submitted before week 25. Table 1 illustrates the project timeline.

Table 1. Proposed MSc Thesis Project Timeline

Activity/ Months	Feb. 2025	Mar. 2025	Apr. 2025	May. 2025	Jun. 2025
Creation of Bot Architecture					
Inter-Agent Negotiation Development					
Negotiation Simulation					
Research Design					
Data Analysis					
Results Section					
Conclusion & Discussion					
Final Submission					

References:

Abdelnabi, S., Gomaa, A., Sivaprasad, S., Schönherr, L., & Fritz, M. (2023). *Cooperation, Competition, and Maliciousness: LLM-Stakeholders Interactive Negotiation*. https://arxiv.org/abs/2309.17234v2

Akata, E., Schulz, L., Coda-Forno, J., Oh, S. J., Bethge, M., & Schulz, E. (2023). *Playing repeated games with Large Language Models*. https://arxiv.org/abs/2305.16867v1

Aksoy, A., & Öztürk, N. (2011). Supplier selection and performance evaluation in just-in-time production environments. *Expert Systems with Applications*, *38*(5), 6351–6359. https://doi.org/10.1016/J.ESWA.2010.11.104

Baarslag, T., Kaisers, M., Jonker, C. M., Gerding, E. H., & Gratch, J. (2017). When will negotiation agents be able to represent us? The challenges and opportunities for autonomous negotiators. *IJCAI International Joint Conference on Artificial Intelligence*, *0*, 4684–4690. https://doi.org/10.24963/IJCAI.2017/653

Bajaj, S., Rockcliffe, S., & Raudsepp, E. (2023, March 27). *Getting recession ready: How AI chat negotiations are helping companies to improve working capital and reduce costs*. https://www.ey.com/en_ca/insights/ai/how-ai-chat-negotiations-are-helping-companies-to-improve-working-capital-and-reduce-costs

Bianchi, F., Chia, P. J., Yuksekgonul, M., Tagliabue, J., Jurafsky, D., & Zou, J. (2024). How Well Can LLMs Negotiate? NegotiationArena Platform and Analysis. *Proceedings of Machine Learning Research*, *235*, 3935–3951. https://arxiv.org/abs/2402.05863v1

Boran, F. E., Genç, S., Kurt, M., & Akay, D. (2009). A multi-criteria intuitionistic fuzzy group decision making for supplier selection with TOPSIS method. *Expert Systems with Applications*, *36*(8), 11363–11368. https://doi.org/10.1016/J.ESWA.2009.03.039

Boute, R. N., & Van Mieghem, J. A. (2020). Digital Operations: Autonomous Automation and the Smart Execution of Work. *SSRN Electronic Journal*. https://doi.org/10.2139/SSRN.3400186

Brookins, P., & DeBacker, J. (2023). Playing Games With GPT: What Can We Learn About a Large Language Model From Canonical Strategic Games? *Economics Bulletin*, *44*(1), 25–37. https://doi.org/10.2139/SSRN.4493398

Büyüközkan, G., & Çifçi, G. (2011). A novel fuzzy multi-criteria decision framework for sustainable supplier selection with incomplete information. *Computers in Industry*, *62*(2), 164–174. https://doi.org/10.1016/J.COMPIND.2010.10.009

Chatterjee, A., Miller, S., & Parepally, N. (2024). *AgreeMate: Teaching LLMs to Haggle*. https://arxiv.org/abs/2412.18690v1

Chen, C. T., Lin, C. T., & Huang, S. F. (2006). A fuzzy approach for supplier evaluation and selection in supply chain management. *International Journal of Production Economics*, *102*(2), 289–301. https://doi.org/10.1016/J.IJPE.2005.03.009

Chen, J.-H., Su, M.-C., Azzizi, V.-T., Wang, T.-K., Lin, W.-J., Azzizi, V.-T. ;, Wang, T.-K. ;, Lin, W.-J., Amorim, M., Cohen, Y., & Reis, J. (2021). Smart Project Management: Interactive Platform Using Natural Language Processing Technology. *Applied Sciences 2021, Vol. 11, Page 1597*, *11*(4), 1597. https://doi.org/10.3390/APP11041597

Chen, S., & Su, R. (2022). An autonomous agent for negotiation with multiple communication channels using parametrized deep Q-network. *Mathematical Biosciences and Engineering*, *19*(8), 7933–7951. https://doi.org/10.3934/MBE.2022371

Chui, M., Hazan, E., Roberts, R., Singla, A., Smaje, K., Sukharevsky, A., Yee, L., & Zemmel, R. (2023, June 1). *The economic potential of generative AI: The next productivity frontier*. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier#business-value

Coehoorn, R. M., & Jennings, N. R. (2004). Learning on opponent's preferences to make effective multi-issue negotiation trade-offs. *Proceedings of the 6th International Conference on Electronic Commerce - ICEC '04*, 59. https://doi.org/10.1145/1052220.1052229

Cui, R., Li, M., & Zhang, S. (2022). AI and Procurement. *Manufacturing and Service Operations Management*, *24*(2), 691–706. https://doi.org/10.1287/msom.2021.0989

Cui, Y., van Esch, P., & Phelan, S. (2024). How to build a competitive advantage for your brand using generative AI. *Business Horizons*, *67*(5), 583–594. https://doi.org/10.1016/j.bushor.2024.05.003

Davidson, T. R., Veselovsky, V., Josifoski, M., Peyrard, M., Bosselut, A., Kosinski, M., & West, R. (2024). *Evaluating Language Model Agency through Negotiations*.

Dong, Q., Li, L., Dai, D., Zheng, C., Ma, J., Li, R., Xia, H., Xu, J., Wu, Z., Liu, T., Chang, B., Sun, X., Li, L., & Sui, Z. (2024). *A Survey on In-context Learning*. https://arxiv.org/abs/2301.00234v6

Faratin, P., Sierra, C., & Jennings, N. R. (2002). Using similarity criteria to make issue trade-offs in automated negotiations. *Artificial Intelligence*, *142*(2), 205–237. https://doi.org/10.1016/S0004-3702(02)00290-4

Fontana, N., Pierri, F., & Aiello, L. M. (2024). *Nicer Than Humans: How do Large Language Models Behave in the Prisoner's Dilemma?* https://arxiv.org/abs/2406.13605v2

Fu, Y., Peng, H., Khot, T., & Lapata, M. (2023). *Improving Language Model Negotiation with Self-Play and In-Context Learning from AI Feedback*.

Fu, Y., Peng, H., Sabharwal, A., Clark, P., & Khot, T. (2022). Complexity-Based Prompting for Multi-Step Reasoning. *11th International Conference on Learning Representations, ICLR 2023*. https://arxiv.org/abs/2210.00720v2

Guo, F. (2023). *GPT in Game Theory Experiments*. https://arxiv.org/abs/2305.05516v2

Guo, S., Bu, H., Wang, H., Ren, Y., Sui, D., Shang, Y., & Lu, S. (2024). *Economics Arena for Large Language Models*. https://arxiv.org/abs/2401.01735v1

Hadi, M. U., Al tashi, qasem, Qureshi, R., Shah, A., Muneer, A., Irfan, M., Zafar, A., Shaikh, M. B., Akhtar, N., Wu, J., Mirjalili, S., & Shah, M. (2025). A Survey on Large Language Models: Applications, Challenges, Limitations, and Practical Usage. *Authorea Preprints*. https://doi.org/10.36227/TECHRXIV.23589741.V1

Hosseini, P., Castro, I., Ghinassi, I., & Purver, M. (2024). *Efficient Solutions For An Intriguing Failure of LLMs: Long Context Window Does Not Mean LLMs Can Analyze Long Sequences Flawlessly*. https://arxiv.org/abs/2408.01866v3

Hua, Y., Qu, L., & Haffari, G. (2024). Assistive Large Language Model Agents for Socially-Aware Negotiation Dialogues. *Findings of the Association for Computational Linguistics: EMNLP 2024*, 8047–8074. https://doi.org/10.18653/v1/2024.findings-emnlp.473

Huang, Y. J., & Hadfi, R. (2024). *How Personality Traits Influence Negotiation Outcomes? A Simulation based on Large Language Models*. https://arxiv.org/abs/2407.11549v2

Kulkarni, P., Mahabaleshwarkar, A., Kulkarni, M., Sirsikar, N., & Gadgil, K. (2019). Conversational AI: An Overview of Methodologies, Applications & Future Scope. *2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*, 1–7. https://doi.org/10.1109/ICCUBEA47591.2019.9129347

Kumar, M., Vrat, P., & Shankar, R. (2003). A fuzzy goal programming approach for vendor selection problem in a supply chain. *Computers & Industrial Engineering*, *46*(1), 69–85. https://doi.org/10.1016/J.CIE.2003.09.010

Kuo, R. J., Wang, Y. C., & Tien, F. C. (2010). Integration of artificial neural network and MADA methods for green supplier selection. *Journal of Cleaner Production*, *18*(12), 1161–1170. https://doi.org/10.1016/J.JCLEPRO.2010.03.020

Kwon, D., Weiss, E., Kulshrestha, T., Chawla, K., Lucas, G. M., & Gratch, J. (2024). *Are LLMs Effective Negotiators? Systematic Evaluation of the Multifaceted Capabilities of LLMs in Negotiation Dialogues*. https://arxiv.org/abs/2402.13550v2

Lee, M. (2023). A Mathematical Investigation of Hallucination and Creativity in GPT Models. *Mathematics 2023, Vol. 11, Page 2320*, *11*(10), 2320. https://doi.org/10.3390/MATH11102320

Li, G., Hammoud, H. A. A. K., Itani, H., Khizbullin, D., & Ghanem, B. (2023). *CAMEL: Communicative Agents for "Mind" Exploration of Large Language Model Society*. https://arxiv.org/abs/2303.17760v2

Li, J., Yuan, Y., & Zhang, Z. (2024). *Enhancing LLM Factual Accuracy with RAG to Counter Hallucinations: A Case Study on Domain-Specific Queries in Private Knowledge-Bases*. https://arxiv.org/abs/2403.10446v1

Light, J., Cai, M., Shen, S., & Hu, Z. (2023). *AvalonBench: Evaluating LLMs Playing the Game of Avalon*. https://arxiv.org/abs/2310.05036v3

Lopes, F., Wooldridge, M., & Novais, A. Q. (2008). Negotiation among autonomous computational agents: principles, analysis and challenges. *Artificial Intelligence Review*, *29*(1), 1–44. https://doi.org/10.1007/s10462-009-9107-8

Martinenghi, A., Donabauer, G., Amenta, S., Bursic, S., Giudici, M., Kruschwitz, U., Garzotto, F., & Ognibene, D. (2024). LLMs of Catan: Exploring Pragmatic Capabilities of Generative Chatbots Through Prediction and Classification of Dialogue Acts in Boardgames' Multi-party Dialogues. In *Proceedings of the 10th Workshop on Games and Natural Language Processing @ LREC-COLING 2024* (pp. 107–118). https://aclanthology.org/2024.games-1.12/

Melz, E. (2023). *Enhancing LLM Intelligence with ARM-RAG: Auxiliary Rationale Memory for Retrieval Augmented Generation*. https://arxiv.org/abs/2311.04177v1

Memon, M. S., Lee, Y. H., & Mari, S. I. (2015). Group multi-criteria supplier selection using combined grey systems theory and uncertainty theory. *Expert Systems with Applications*, *42*(21), 7951–7959. https://doi.org/10.1016/J.ESWA.2015.06.018

Mittal, A., Cocoual, C., Erriquez, M., & Liakopoulou, T. (2024, June 13). *Revolutionizing procurement: Leveraging data and AI for strategic advantage*. https://www.mckinsey.com/capabilities/operations/our-insights/revolutionizing-procurement-leveraging-data-and-ai-for-strategic-advantage

Moghimifar, F., Li, Y.-F., Thomson, R., & Haffari, G. (2024). *Modelling Political Coalition Negotiations Using LLM-based Agents*. https://arxiv.org/abs/2402.11712v1

Noh, S., & Chang, H.-C. H. (2024). *LLMs with Personalities in Multi-issue Negotiation Games*. https://arxiv.org/abs/2405.05248v2

Obinna, A. J., & Kess-Momoh, A. J. (2024). Systematic technical analysis: Enhancing AI deployment in procurement for optimal transparency and accountability. *Global Journal of Engineering and Technology Advances*, *19*(1), 192–206. https://doi.org/10.30574/gjeta.2024.19.1.0067

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative Agents: Interactive Simulacra of Human Behavior. *UIST 2023 - Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, *1*. https://doi.org/10.1145/3586183.3606763

Parsons, S., Sierra, C., & Jennings, N. (1998). Agents that Reason and Negotiate by Arguing. *Journal of Logic and Computation*, *8*(3), 261–292. https://doi.org/10.1093/LOGCOM/8.3.261

Perković, G., Drobnjak, A., & Botički, I. (2024). Hallucinations in LLMs: Understanding and Addressing Challenges. *2024 47th ICT and Electronics Convention, MIPRO 2024 - Proceedings*, 2084–2088. https://doi.org/10.1109/MIPRO60963.2024.10569238

Qiao, S., Ou, Y., Zhang, N., Chen, X., Yao, Y., Deng, S., Tan, C., Huang, F., & Chen, H. (2023). Reasoning with Language Model Prompting: A Survey. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, *1*, 5369–5393. https://doi.org/10.18653/v1/2023.acl-long.294

Raiaan, M. A. K., Mukta, M. S. H., Fatema, K., Fahad, N. M., Sakib, S., Mim, M. M. J., Ahmad, J., Ali, M. E., & Azam, S. (2024). A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges. *IEEE Access*, *12*, 26839–26874. https://doi.org/10.1109/ACCESS.2024.3365742

Renze, M., & Guven, E. (2024). *The Effect of Sampling Temperature on Problem Solving in Large Language Models*. https://arxiv.org/abs/2402.05201v3

Sharma, R., Shishodia, A., Gunasekaran, A., Min, H., & Munim, Z. H. (2022). The role of artificial intelligence in supply chain management: mapping the territory. *International Journal of Production Research*, *60*(24), 7527–7550. https://doi.org/10.1080/00207543.2022.2029611

Singh, S., & Mahmood, A. (2021). The NLP Cookbook: Modern Recipes for Transformer Based Deep Learning Architectures. *IEEE Access*, *9*, 68675–68702. https://doi.org/10.1109/ACCESS.2021.3077350

Sunder, V., Vig, L., Chatterjee, A., & Shroff, G. (2018). Prosocial or Selfish? Agents with different behaviors for Contract Negotiation using Reinforcement Learning. *Studies in Computational Intelligence*, *905*, 63–81. https://doi.org/10.1007/978-981-15-5869-6_5

Sycara, K. P. (1993). Machine learning for intelligent support of conflict resolution. *Decision Support Systems*, *10*(2), 121–136. https://doi.org/10.1016/0167-9236(93)90034-Z

Taplin, S. (2024, August 15). *AI Hallucinations: How Can Businesses Mitigate Their Impact?* https://www.forbes.com/councils/forbestechcouncil/2024/08/15/ai-hallucinations-how-can-businesses-mitigate-their-impact/

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). *LLaMA: Open and Efficient Foundation Language Models*. https://arxiv.org/abs/2302.13971v1

Van Hoek, R., DeWitt, M., Lacity, M., & Johnson, T. (2022, November 8). *How Walmart Automated Supplier Negotiations*. https://hbr.org/2022/11/how-walmart-automated-supplier-negotiations

Vertsel, A., Rumiantsau, M., & Bi, N. (2024). *Hybrid LLM/Rule-based Approaches to Business Insights Generation from Structured Data*. https://arxiv.org/abs/2404.15604v1

Wang, C., Long, Q., Xiao, M., Cai, X., Wu, C., Meng, Z., Wang, X., & Zhou, Y. (2024). *BioRAG: A RAG-LLM Framework for Biological Question Reasoning*. https://arxiv.org/abs/2408.01107v2

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems*, *35*. https://arxiv.org/abs/2201.11903v6

Williams, S., & Huckle, J. (2024). *Easy Problems That LLMs Get Wrong*. https://arxiv.org/abs/2405.19616v2

Xu, Z., Yu, C., Fang, F., Wang, Y., & Wu, Y. (2023). Language Agents with Reinforcement Learning for Strategic Play in the Werewolf Game. *Proceedings of Machine Learning Research*, *235*, 55434–55464. https://arxiv.org/abs/2310.18940v3

Xue, J., Zheng, M., Hu, Y., Liu, F., Chen, X., & Lou, Q. (2024). *BadRAG: Identifying Vulnerabilities in Retrieval Augmented Generation of Large Language Models*. https://arxiv.org/abs/2406.00083v2

Xue, X., Wang, Y., & Shen, Q. (2007). A hybrid method for improving agent-based negotiation efficiency in supply chain. *Proceedings of the IEEE International Conference on Automation and Logistics, ICAL 2007*, 2540–2543. https://doi.org/10.1109/ICAL.2007.4339007

Yao, C., & Fujita, S. (2024). Adaptive Control of Retrieval-Augmented Generation for Large Language Models Through Reflective Tags. *Electronics 2024, Vol. 13, Page 4643*, *13*(23), 4643. https://doi.org/10.3390/ELECTRONICS13234643

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2022). ReAct: Synergizing Reasoning and Acting in Language Models. *11th International Conference on Learning Representations, ICLR 2023*. https://arxiv.org/abs/2210.03629v3

Yap, H. P., Ong, W. L., Koh, J., Haoran, L., Tiancheng, Y., & Zihao, C. (2024). AI-based Approach to Efficient Information Extraction for Supply Chain Contracts. *Proceedings - 2024 IEEE Conference on Artificial Intelligence, CAI 2024*, 139–143. https://doi.org/10.1109/CAI59869.2024.00034

Zeng, D., & Sycara, K. (1998). Bayesian learning in negotiation. *International Journal of Human-Computer Studies*, *48*(1), 125–141. https://doi.org/10.1006/IJHC.1997.0164

Zeng, J., Chen, K., Wang, R., Li, Y., Fan, M., Wu, K., Qi, X., & Wang, L. (2024). ContractMind: Trust-calibration interaction design for AI contract review tools. *International Journal of Human-Computer Studies*, *196*, 103411. https://doi.org/10.1016/J.IJHCS.2024.103411

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., Wang, L., Tuan Luu, A., Bi, W., Shi, F., Shi, S., & lab, T. A. (2023). *Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models*. https://arxiv.org/abs/2309.01219v2

Zhuge, M., Liu, H., Faccio, F., Ashley, D. R., Csordás, R., Gopalakrishnan, A., Hamdi, A., Abed, H., Hammoud, A. K., Herrmann, V., Irie, K., Kirsch, L., Li, B., Li, G., Liu, S., Mai, J., Pi, P., Ekos, ., Ramesh, A., … Schmidhuber, J. (2023). *Mindstorms in Natural Language-Based Societies of Mind*. https://arxiv.org/abs/2305.17066v1