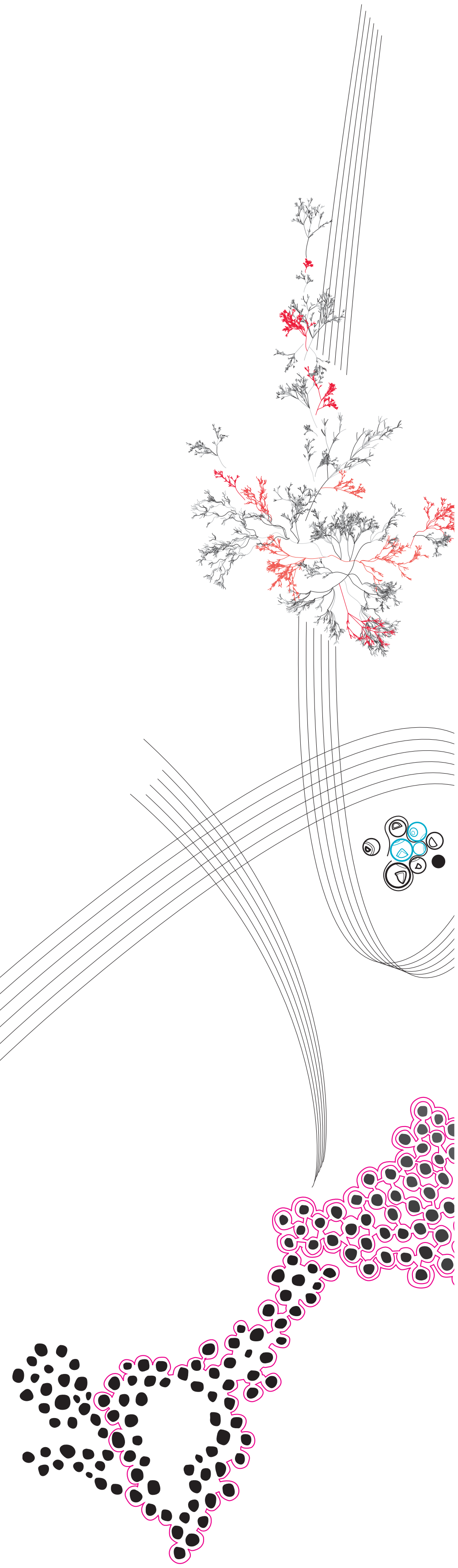
MSc Computer Science
Final Project



Facilitating Industrial B2B e-Auctions through Multi-Agent and Retrieval Augmented Large Language Models

Noor Mansour

Supervisor:
Dr. F.A. Bukhsh (Faiza)
Prof. Dr. H. Schiele (Holger)
Dr. M. Daneva (Maya)

April, 2024

Department of Computer Science
Faculty of Electrical Engineering,
Mathematics and Computer Science,
University of Twente

UNIVERSITY OF TWENTE.

Contents

Abstract

Purchasing professionals often face challenges in the auction design process due to the time-intensive and complex nature of the task. The intricacies of auction dynamics, and the negative impact of subpar auction design choices on realized savings, highlight the need for advanced tools that can facilitate more informed decision-making, optimize auction strategies, and automate the design process. This research project addresses these challenges by designing, implementing, and evaluating a multi-agent and retrieval augmented conversational chatbot system, based on Large Language Models (LLMs) to assist in the auction design process.

Leveraging a proprietary dataset from a case company of executed reverse e-auctions, the study first conducts an empirical analysis of theoretical auction design models and examines the impact of several key factors like the number of bidders, price dispersion, and the risk aversion of bidders on the auction design choice and realized savings. A new empirically optimized recommendation model that demonstrates the potential for achieving higher expected savings compared to existing models is proposed. Notably, the study defines specific operationalizations for the practical use of the recommendation models and finds that the empirical evidence supports the theoretical recommendations in a small majority of cases.

The insights derived from this empirical analysis are integrated into the chatbot system's knowledge base alongside the dataset of e-auctions, allowing it to generate customized auction design recommendations. The effectiveness of this system is evaluated using the AuctionEval dataset, created specifically to assess the performance of the system on common use cases around auction design. The experiments explore how information volume, type, and various prompt engineering techniques impact the chatbot's performance. The findings reveal that smaller models significantly benefit from tailored, focused corpora, while larger models gain from a more diverse corpus. Additionally, the implementation of chain of thought prompting has markedly improved the relevance, quality of answers, and reasoning accuracy of the system.

The study is limited by the analysis of only single-phase auctions and a limited range of auction types in the dataset, which restricts the applicability of the validation of theoretical models. Furthermore, the multi-agent conversational chatbot has limitations regarding performance in data retrieval tasks, and the evaluation procedure can benefit from a larger experimental setup and evaluation dataset. These limitations highlight the need for future research to encompass multi-phase auctions and a larger experimental setup and evaluation of the chatbot application.

The practical and scientific implications of this research are twofold: first, filling the research gap of missing empirical evidence on auction design recommendation models; second, delivering an innovative, adaptable, and robust LLM system that simplifies the auction design process for purchasing professionals and highlights the reasoning capabilities of LLMs for currently under-explored game theoretical use cases. Through the large adaptability of the system to new data and tasks, the architecture promises to be beneficial for a range of tasks in the purchasing domain beyond auction design.

Future work should focus on improving the accuracy of data retrieval tasks and expanding the overall system architecture as well as experimental configurations. Further research could also expand the empirical analysis to multi-phase auctions and extend the system's capabilities to cover more complex auction scenarios, and a more varied range of tasks, thereby enhancing its practical applicability and effectiveness in real-world settings.

Keywords: large language models, auction theory, multi-agent systems, game theory, empirical analysis, purchasing

Chapter 1

Introduction

The following section outlines the research problem and the questions that will be addressed in this study. Section 1.1 gives an overview of the problem situation by introducing the concept of electronic auctions and their use in the purchasing field. Section 1.2 outlines problems faced by purchasing professionals relating to the design of auctions. Section 1.3 then formulates the research goal and related research questions to be answered.

1.1 Initial Situation: Bounded Rationality Wastes Potential Cost Reductions

Auctions are prevalent selling and buying mechanisms across industries and product types. Art, toys, collectibles, and public contracts, as well as electric power or radio spectrums, are allocated through auction mechanisms [37, 43, 21]. Business-to-business (B2B) purchasing activities also utilize auctions as an integral competitive mechanism to reduce purchasing costs. Especially online reverse auctions, also known as reverse e-auctions, are incorporated as an important competitive procurement method in the purchasing process by most large companies and government entities [86, 43, 21].

Purchasing professionals often face the challenge of designing reverse auctions that optimize a certain (set of) goal(s). Designing auctions is a time-intensive and complex task to perform [85, 67, 21]. In combination with limitations on the available information, cognitive ability, and time to design auctions, most purchasers design auctions to achieve satisfactory but not optimal performance [85]. At the same time, it has been shown that auctions, that follow mechanism design principles for a specific scenario, yield lower bids and achieve significant cost reductions, while some auction designs can adversely impact the final price and buyer-supplier relationships [85, 11, 72]. Therefore, potential cost reductions are left untouched and possible adverse impacts are tolerated due to bounded rationality [17] and limited knowledge on the topic of mechanism design and game theory.

A more detailed explanation of several concepts relating to auctions can be found in Chapter 3.1.

1.2 Complication: Lack of Guidance on Auction Design

An auction designer needs to decide on many different factors like the auction format (e.g. Dutch or English format) and parameter settings (e.g. reservation prices and minimum bid step size), which have an interdependent effect on the final auction outcome [37, 67, 86]. The amount of possibilities is not only overwhelming for a human but can even provide a computationally intractable solution space, especially for combinatorial auctions [91].

Purchasers who turn to scientific articles to search for decision-making models that would support them in choosing the most appropriate auction design will find an extensive body of theoretical research. The focus of that research lies predominantly on the optimal auction design for experimental scenarios that are built upon sets of assumptions like the seminal VCG auction mechanisms [64, 93]. Empirical research on the optimal auction design in complex real scenarios with different auction formats and parameter settings is scarcely found in the literature [44, 21, 7].

Another burden purchasers face is that many of the theoretical assumptions made in these cases are not met in practical scenarios and bidder behavior may differ significantly from what the theory predicts [80]. For instance, a frequently made assumption is a competitive market without implicit or explicit collusion [6, 42, 44]. However, in real-world auction scenarios, collusion among bidders poses a realistic concern for auction designers [80] and a range of example cases has been captured in the literature [74, 4, 75, 92]. In addition, the assumption on perfect rationality of participants or identical valuation of the auctioned item underpins many of the foundational concepts in auction theory such as the Revenue Equivalence Theorem [65, 66], neglecting the possibility of bounded rationality of bidders [84] due to differences in information [35] or emotional control [102]. For instance, despite its advantageous game theoretical properties, the VCG mechanism is rarely employed in practical scenarios [14, 80].

Much of the existing empirical research in purchasing auctions is based on publicly available data such as public governmental procurement auctions [28, 50, 55] and laboratory/experimental setups with professionals or students [2, 15, 30, 87]. Due to the confidential nature of private auction datasets, they rarely appear in the literature [31, 62]. The few papers that utilize a relationship with an industrial partner to conduct an empirical analysis of auction performances and design parameters, have different focuses for the analysis [31, 36, 47, 60, 62], which makes the impact of the auction design and the applicability of recommendation models in practice elusive.

Existing research for empirical decision-making models specific to auction-type recommendations such as [6, 44, 20, 84] often define selection criteria and give recommendations of auction formats based on game theoretical reasoning or mathematical approaches like order statistics. Most of the authors remain consistent on the choice of important selection criteria for the auction format decision, but are inconsistent in the specific recommendations of an auction format, giving an indecisive signal to practitioners. This indecisive signal is also shown by the conflicting observations and recommendations of the theoretical and empirical research [14].

Although large organizations provide internal learning material for their purchasers, many are overwhelmed by the number of theoretical frameworks and suffer from limited time and ability to use that information to design auctions [85]. Learning on the job is therefore more prevalent in the design of auctions within an industrial B2B context, which encourages purchasers to remain with standard auction design templates that have been used in the past. Therefore, a research-practitioner gap can be observed in the field of reverse e-auction design, in which practitioners only take limited value out of the work of research.

A more elaborate discussion of the complications and problem situation can be found in Chapter 2.3.

1.3 Research Goal: Maximizing Buyer Surplus through LLMs

In a World Cafe survey [85], practitioners and Artificial Intelligence (AI) experts expressed their belief in the potential of AI to facilitate the creation of mechanism design-based auctions. AI is seen as a major player in enabling practitioners to design auctions faster and with increased achievement of desired objectives such as revenue maximization. The AI experts participating in the study judged this idea with limited technical feasibility as of the year 2020. Since then, large language models have dominated news headlines and businesses are rushing to implement this new technology, due to its strong ability to understand language, engage in reasoning, and process multi-modal inputs, acting as a one-shot learner for many different tasks with an intuitive user experience [34]. To address the research-practitioner gap in reverse e-auction design and to explore the potential of utilizing large language models (LLMs) to help purchasers design auctions, the following research question is defined:

RQ *How can we facilitate the design of reverse e-auctions that maximize buyer surplus in practice by using large language models?*

The following sub-questions help in answering the main research question:

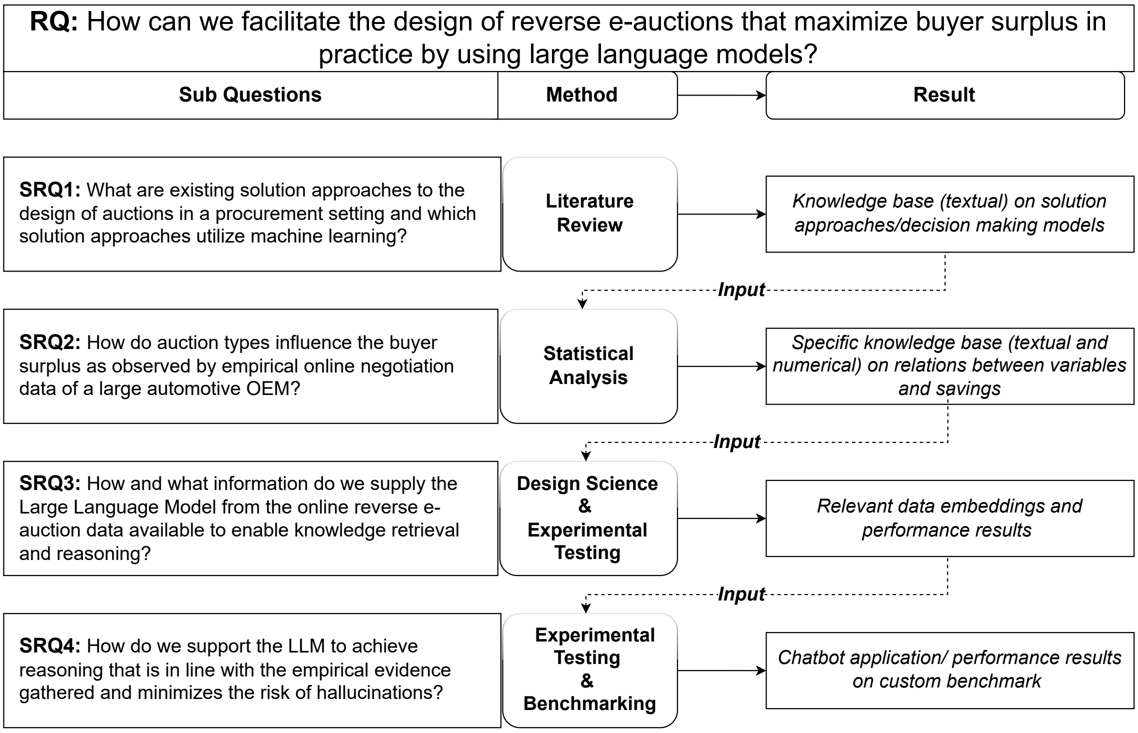
SRQ1 What are existing solution approaches to the design of auctions in a procurement setting and which solution approaches utilize machine learning?

SRQ2 How do auction types influence the buyer surplus as observed by empirical online negotiation data of a large automotive firm?

SRQ3 How and what information do we supply the LLM from the online reverse e-auction data available to enable knowledge retrieval and reasoning?

SRQ4 How do we support the LLM to achieve reasoning that is in line with the empirical evidence gathered and minimizes the risk of hallucinations?

Figure 1.1 highlights the main and sub-questions and their relation graphically. The sub-questions were designed to divide the bigger main research question into smaller parts that utilize the results of each other to build towards answering the main research question. SRQ1 builds the theoretical foundation of existing solution approaches and establishes the direction of the research, while also creating a knowledge base of basic information and decision-making models related to the optimal auction design. Additionally, it yields the hypothesis tested on the empirical data for SRQ2. The detailed literature review for SRQ1 is included in Appendix A. In combination with SRQ2, which specifically analyses the data from e-auctions conducted at the case company based on the hypothesis formulated from SRQ1, they both build the knowledge base for the LLM about relevant solution approaches, decision-making models, and relations of variables in textual and numerical form. This knowledge base provides context to the LLM about relationships of variables, specific auction scenarios, and mechanism design principles to design auctions based on the up-to-date research and specific data available. It is also utilized to check if the output



RQ: How can we facilitate the design of reverse e-auctions that maximize buyer surplus in practice by using large language models?

Sub Questions
Method
Result

SRQ1: What are existing solution approaches to the design of auctions in a procurement setting and which solution approaches utilize machine learning?

Literature Review
Knowledge base (textual) on solution approaches/decision making models
Input

SRQ2: How do auction types influence the buyer surplus as observed by empirical online negotiation data of a large automotive OEM?

Statistical Analysis
Specific knowledge base (textual and numerical) on relations between variables and savings
Input

SRQ3: How and what information do we supply the Large Language Model from the online reverse e-auction data available to enable knowledge retrieval and reasoning?

Design Science & Experimental Testing
Relevant data embeddings and performance results
Input

SRQ4: How do we support the LLM to achieve reasoning that is in line with the empirical evidence gathered and minimizes the risk of hallucinations?

Experimental Testing & Benchmarking
Chatbot application/ performance results on custom benchmark

Figure 1.1: Visualization of the relation between sub-questions and how which results are used as inputs to the subsequent questions and activities

of the LLM suffers from hallucinations and false facts.

To answer the research question, the design science methodology as proposed by [90] is followed to create an artifact that fills the research-practitioner gap and yields novel scientific insights into the optimal auction design in practice and the application of LLMs in auction mechanism design. The following Chapter 2.1 details the research method followed to solve the research questions and relates each chapter to the steps of the research methodology.

Chapter 2

Design Science Methodology

This chapter aims to provide an overview of the methodological framework guiding the research and structure of the thesis. It situates the problem within a broader research context, offers detailed clarification of the problem, and specifies the properties and objectives the solution should encompass to effectively address the issue.

Section 2.1 delves into the methodological framework guiding the research, specifically employing the Design Science Methodology specific to the creation of artifacts in the purchasing science field. It outlines the seven-step process and aligns each step with the corresponding chapters within the thesis. Section 2.2 contextualizes the thesis as an interdisciplinary research field at the intersection of mechanism design, autonomous agents, and LLM's. It explores the theoretical foundations and practical implications of these research areas, situating the thesis at the broader challenge among this interdisciplinary framework. Section 2.3 defines the specific problems within auction design in theory and practice, highlighting challenges such as bounded rationality, lack of decision-making support, and variability in auction scenarios. Section 2.4 concludes the chapter by outlining solution objectives, emphasizing the need for a versatile, transparent, and context-aware smart assistant to facilitate optimal auction design.

2.1 Overview Research Methodology

The research is structured according to the design science methodology by [90], as the authors propose the methodology specifically for the purchasing field with the aim to support the creation of artifacts suitable to solve complex problems faced in theory as well as practice. It enables the research to be structured and reproducible while giving a robust iterative framework for the design and evaluation of the LLM-based chat-bot application. Figure 2.1 presents the methodology in its seven steps and their corresponding chapters:

1. identification & clarification of the problem
2. exploration of the solution space
3. selection of development approach
4. designing the solution concept
5. applying the solution
6. evaluating the solution

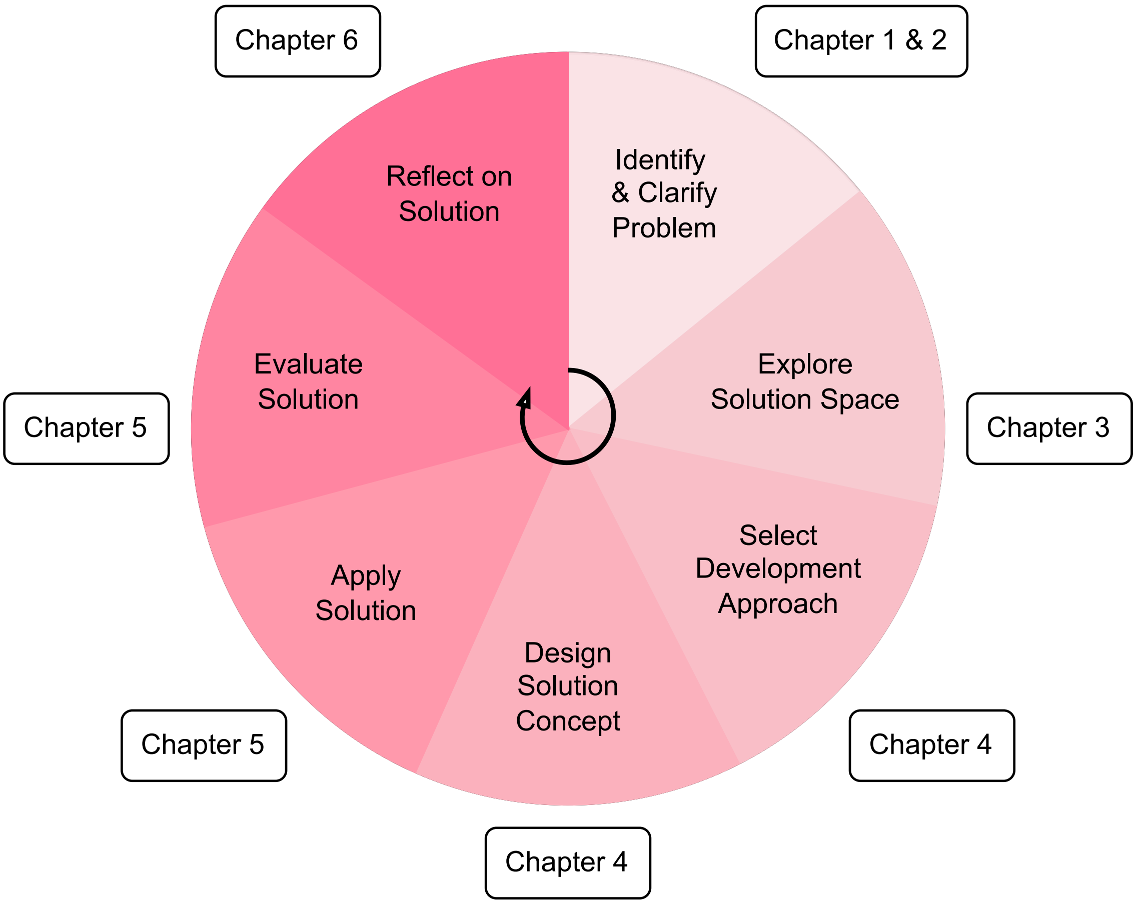


Figure 2.1: Design science research cycle according to [90] and corresponding chapters in this thesis

7. reflecting upon the performance of the solution and the research

Chapter 1 serves as the introduction and initial identification of the problem. Chapter 2 additionally goes into more detail about the problem definition by placing it in the broader research context, explaining the problem in more detail, and defining specific solution objectives. Both chapters correspond to the first step of the design science approach. Chapter 3 explores the extant solutions to the problem by performing a literature review and identifying research gaps, highlighting relevant literature, as well as defining a hypothesis for SRQ2 about the influences of auction design types on buyer surplus. It also explains the basic concepts related to auction and game theory. Chapter 4 discusses the methodology and development approach used for both as well as the statistical analysis and creation of the LLM based chatbot application, that aims to facilitate the creation of auctions in practice, along with the concept design and evaluation procedure. Therefore, Chapter 4 covers the selection of the development approach and the design of the solution concept of the design science cycle. Chapter 5 presents the results of the statistical analysis and the design as well as the evaluation results of the LLM based system. It hence incorporates in addition with the relevant appendices the two phases of application and evaluation of the solution. Chapter 6 concludes with a discussion of the results and its limitations, summarizing the main findings, reflecting upon the results and highlighting steps for future work.

2.2 Broader Challenge: Designing Mechanisms with a Smart Assistant

Our research lies at the intersection of (empirical-)mechanism design, autonomous agents, and large language models. The general class of problem aims to optimize auctions, with the special restrictions that the designed system not only bases its decision on empirical data but also provides greater reasoning and serves as a possible stepping stone for an autonomous negotiation agent. Figure 2.2 highlights the position of the research as the intersection between these fields as shown by the white dot.

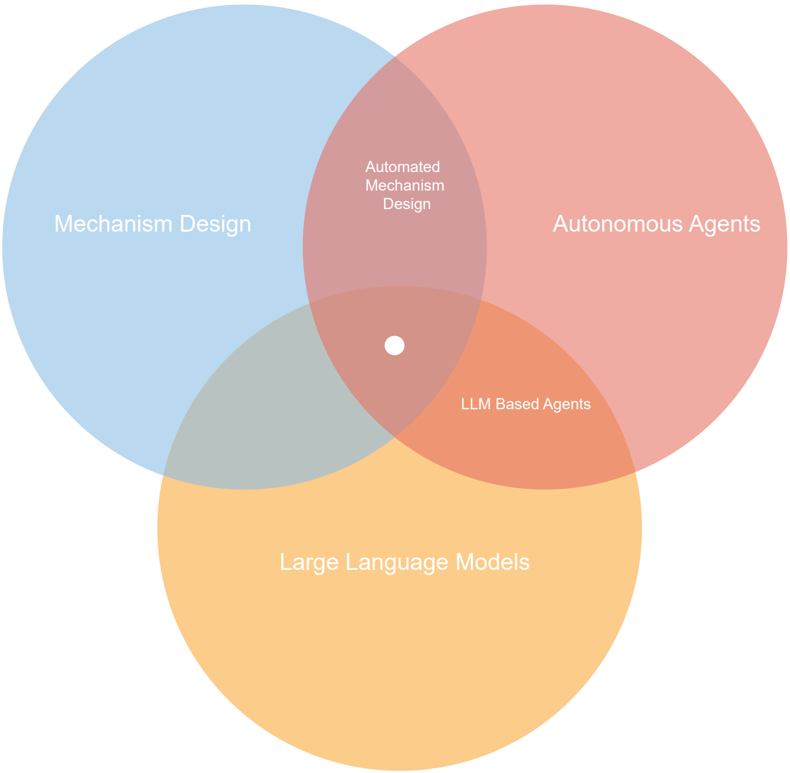


FIGURE 2.2: The three research areas. Our research is placed at the intersection between these three on the white dot.

Mechanism Design The main research streams for auctions revolve around theoretical mechanism design research and its empirical counterpart. Mechanism design is the counterpart of game theory. Given a defined game or scenario, game theory tries to study the impact of different behaviors and choices on the game and objectives, while trying to optimize them. Mechanism Design has the reverse view, as it tries to create the games or mechanisms to achieve certain objectives from the game designer's point of view. While the theoretical research defines auctions mathematically and tries to find formal analytical solutions for the optimal design, the empirical counterpart uses simulations, case studies, expert interviews, or real data to come to the optimal auction design. Mechanism Design is the research field that deals with the design of mechanisms, auctions being a certain type of economic mechanisms, that fulfill a set of objectives. Results in the theoretical mechanism design area are only of limited value to practitioners, as often the assumptions necessary to describe and solve the mathematical problems do not hold in practice [93, 43, 80]. Additionally, most of the auction problem definitions that have closed form analytical solutions are simpler auction types, that do not encompass the multi-attribute, multi-item, and combinatorial auction scenario with different bidder and bidder types that purchasers face in practice [19, 89].

Empirical Mechanism Design

Empirical Mechanism Design started to emerge as a response to the limitations of practical use for the theoretical mechanism design results and this research field is also known as automated mechanism design [96]. While the theoretical mechanism design defines auctions mathematically and tries to find formal analytical solutions for the optimal design, the empirical counterpart uses simulations, case studies, expert interviews, or real data to come to the create mechanisms. Empirical mechanism design tries to use data from practice to inform the optimal auction design for insights on general recommendations or for specific auction scenarios. Automated mechanism design is therefore aiming to provide models that potentially could be used universally in different scenarios to design the optimal auction. There is increasing research interest in the utilization of machine learning models to the research field of automated mechanism design [19, 104], as deep learning approaches have shown strong capabilities in creating possibly optimal auction designs even for complex and numerically intractable situations. The most prominent models are detailed in chapter 3.2.4.

Autonomous Agents

Autonomous Agents research aims to create systems that can act with purpose and independently in complex and dynamically changing environments, without human intervention [97, 100]. Autonomous agents perform various tasks like information retrieval, dialogue, planning and decision making. Many autonomous agents are powered through deep learning and large amounts of training data to complete their assigned task. Challenges faced by the research area are creating the knowledge representations of the model and reasoning capabilities, as well as letting the agents create novel and creative solutions or paths [22]. Reinforcement Learning is often utilized in order to drive decision making and exploration in autonomous agents, but often the limited representation of specific scenarios leaves the research field with the challenge of adeptness to new situations and interacting with humans through several interfaces [97].

Large Language Models

Large Language Models have not only been dominating news headlines, but have also received considerable attention in research [61]. It ranges from the technical side of developing new architectures, new training procedures, or prompt engineering to the application of LLM's to a vast array of tasks from a multitude of domains. Especially the field of Autonomous Agents experiences heightened attention due to the capabilities of large language models to understand and reason about natural language as well for its creativity and adaption skills [97].

The challenge lies in creating an agent that can interact with purchasers in natural language and retrieve relevant information, while also being able to reason with both specific (tabular-)data about the auction situation as well as general (textual-)knowledge on mechanism design. From an empirical mechanism design point of view, the research contributes by utilizing a novel method namely large language models on the task of general auction design, including complex scenarios. From an autonomous agent point of view, the research contributes by creating an autonomous agent to a novel application task and with a still novel technology of Large language models driving the development. Lastly, from a LLM research point of view the novel application area, but also the exploration of different kinds of reasoning on mechanism design challenges combined with the empirical nature of the project contribute to the body of knowledge. Therefore it combines these three research areas at an intersection that to the best of our knowledge is sparsely explored in the current literature.

2.3 Problem: Difficulty of Designing Cost Optimal Reverse E-Auctions

Designing reverse e-auctions with the objective of maximizing savings is a complex task. There are a multiple interrelated issues that contribute to the difficulty of the task that are outlined in this section.

Large Space of Possible Combinations

A purchaser is faced with a large space of possible auction design parameters to decide upon. The impact on the buyer surplus is not only affected by the interrelated decisions on the auction design but also on the parameters of the specific negotiation situation. Table 2.1 shows the specific variables to be decided upon by the purchase in our specific problem situation at the case company. More specifically we have around 22 variables on which decisions can be made on, with many being categorical variables, including some continuous variables. As every decision on these variables can have significant impact on the auction outcome and considering the interrelated effects the variables have on each other, bounded rationality hinders the creation of optimal auction designs.

Bounded Rationality

Purchasers are typically not specialized in the field of mechanism design for negotiations in general and auctions in specific [85]. In combination with the limitations of an individual's rational decision-making by the finite amount of time available to reach decisions, the amount of information available, and their cognitive ability, purchasers tend to develop satisfying rather than optimized solutions. This phenomenon is also known as bounded rationality [17]. The issue of bounded rationality and the missing specialization of purchasers present a major part of the problem in designing optimal auctions, also given the large possible solutions space and difficult game theoretical reasoning process necessary to design auctions [5].

Lack of Decision Making Support

While there exists a set of guidelines and internal recommendation material on how to choose some of the possible auction design parameters highlighted in Table 2.1, most of these internal documents have no underlying research-based or practical source. There also exist models in the literature that aim to help purchasers in deciding on the correct design of auctions, who take up to three parameters into account like the number of suppliers, the initial bid price spread, or the strategic importance of the business to the supplier [6, 44, 43]. Many of these decision models are based on either (game-)theoretical reasoning or mathematical reasoning but do not take into account the individuality of each negotiation situation and the possible large space of auction design parameters to decide upon. Additionally, the recommendations on which specific auction design to choose lack a clear consensus in the literature. Adding that empirical research sends indecisive signals, purchasers can be overwhelmed by the literature and the already existing frameworks for designing auctions. Hence in practice using standard templates and on the job learning is prevalent, rather than relying on theoretical or empirical research.

Trustworthiness and Reasoning

As many purchasers are not classically trained in mechanism design or game theory and mathematical optimizations, much of the research remains elusive to them [85, 67]. The lack of understanding and therefore the implied lack of trust decreases the willingness to adopt the recommendations of research to guide the creation of auctions, leading to the application of standard auction designs from previous negotiations. This approach has been shown to potentially miss increasing profit gains in auctions, and even induce strong negative consequences [72].

Highly Variable Auction Scenarios

Different from ad auctions,which are repeated frequently every minute, the negotiation scenarios at the case company are highly variable and individual. The terms and criteria to be negotiated, the suppliers participating in the auction but also the behaviour, and the strategic negotiation situation can be different for every single item to be sourced. Given this landscape of highly individualized auctions with not many repetitions, it is difficult to propose a common decision-making framework that will be able to apply to all situations and satisfy the necessary conditions. Furthermore, quantitative data is not necessarily available for some important considerations to be made in the decision of the optimal negotiation such as whether the supplier is an incumbent, if a monopoly situation exist or how the attractivity of the auction is to the supplier.

Table 2.1: List of changeable auction design parameters

Variable	Values	Description
General Settings		
Auction Type	[English, Dutch, Japanese]	Describes the general auction type, either a classical English auction or a ticker auction.
Direction	[Forward, Reverse]	Describes the bid directions. Bids going up (Forward) or down (Reverse) on the criteria level.
Duration	[0-200] minutes	The regular duration of the auction in minutes.
Prolongation	[0-200] minutes	Time the auction could be prolonged according to the prolongation strategy.
Prolongation Strategy	[Standard, Limited End Date, Limited Times, Best Bid, Improved Bid]	Standard - Extend Auction Time by a specified time; Limited End Date: Allow prolongations until a specified end date; Limited Times: Allow a specified amount of prolongations; Improved Bid: Only prolong if a supplier improves his rank; Best Bid: Only prolong if there is a new best bid.
Prolongation Max	Date	The maximum to which an auction could be prolonged to.
Feedback Mechanism		
Feedback mechanism	[Best, Rank, Tier]	Best Bid - Show the best value to all participants; Rank - Show individual rank to each participant; Tier - Show individual tier green/yellow/red to each participant. Not mutually exclusive.
tiergreen	[1 - 10]	Defines up to which rank the supplier sees the color green given RANK mechanism.
green blur	[0-100]	If a blur of 1 is defined (1%), the best supplier will see the green color in the yellow area if he has less than 1% distance to the worst supplier in the green area.
tieryellow	[1 - 10]	Defines up to which rank the supplier sees the color yellow given RANK mechanism.
yellow blur	[0-100]	If a blur of 1 is defined (1%), the best supplier in the red area will see the yellow color if he has less than 1% distance to the worst supplier in the yellow area.
Allow equal ranks	[0,1]	Suppliers can share the same rank.
Hide before start	[0,1]	The criterion will only be visible after the start of the negotiation.
hide without initial value	[0,1]	The criterion will be visible only if there is an initial value.
initial value is a global start	[0,1]	Bidding below or above the given initial value is not allowed.
show equal values	[0,1]	Show that a supplier has offered the same value as at least one other.

Continued on next page

Table 2.1 – Continued

Variable	Values	Description
prevent equal values	[0,1]	Equal supplier offers are prevented.
	Target Value	
Target Value	Num	The target value used to compare supplier offers.
target value transparency	[Never, Reached: Status, Reached: Value and Status, Always: Status, Always: Value and Status]	Once a set target value is reached, it can be communicated to the bidders via this option. The "reached" options share either the status in the form of text that the target value is reached or both the value and that it has been reached, only when the value has been reached. The "always" options always show the status, whether the price is reached or not and possibly the corresponding value.
threshold value	num	A value used as a threshold for the supplier to show him feedback on his offers.
threshold value mode	[normal, only red, target traffic light]	Normal - normal traffic light or no traffic light until the threshold is reached; only red - only red until the threshold is reached; target traffic light - Only yellow & red until the threshold is reached.
threshold value transparency	[Never, Reached: Status, Reached: Value and Status, Always: Status, Always: Value and Status]	Same as in target value transparency.
	Bid Increment	
step required	[0,1]	The value of the offer for a criterion must be changed in every offer.
minimum step increment	[absolute/own, percent/own, absolute/overall, percent/overall]	Absolute/own - Bid increments need to be better than own best bid by a specified absolute value; Percent/own - Bid increments need to be better than own best bid by a specified percentage; Absolute/overall - Bid increments need to be better than the overall best bid by a specified absolute value; Percent/overall - Bid increments need to be better than the overall best bid by a specified percentage.
minimum step	absolute value or percentage	The minimum increment value as a percentage or absolute value depending on the former setting.
maximum step increment	[absolute/own, percent/own, absolute/overall, percent/overall]	Same as in minimum step increment.

2.4 Solution Objectives: Versatile, Transparent, Context-Aware

Given the detailed problem discussion of 2.3 and after some discussions with a group of former and still active purchasers at the case company, the system should ideally incorporate the following functional requirements to achieve the goal of facilitating the e-auction design:

- The system provides the user with past historical data and the negotiation parameters
- The system analyses past historical data in a descriptive manner
- The system gives auction design recommendations based on historical data and auction theory knowledge
- The system is flexible in its response and the possible inputs from the user
- The system is able to reason and explain its auction design recommendations, referring to relevant past data and materials like handouts, or research articles.
- The system is able to help the user not only with general recommendations but also takes into account the specific context provided to him
- The system is able to process new information outside of the existing databases in textual as well as numerical form.

To solve the issue, the creation of a smart assistant that can act as a knowledge discovery tool but also provides reasoning to recommend auction designs based on past data and game theoretical knowledge is proposed. The goal is to facilitate the creation of auction designs by letting the tool provide all necessary information and also process it to go beyond the limits of bounded rationality and fill the research practitioner gap. The requirements for the project are that the main input of the system is the available data on the conducted online reverse e-auctions at the case company, and is able to reason with the buyer about the optimal design of auctions given the specific individualized auction scenario. It helps by tackling the problem of bounded rationality, giving the buyer access to knowledge, be it the specific data or material on the theory of auction design, but also giving recommendations and reason upon the optimal design to take. Specific focus was laid on the explainability and interpretability of the model's output. In industrial auctions, multiple millions of euros are at stake and therefore trust in the model's output is of paramount importance for its use in practice.

Chapter 3

Background & Literature

This chapter aims to provide background knowledge and compile the relevant literature in the topics of auctions, machine learning in auction design and LLM's. The contents of the chapter are the result of a literature review as found in Appendix A, conducted to answer the first sub-question SRQ1. First, section 3.1 explores Game Theory and its counterpart Mechanism design as the foundation of auction design, in addition to introducing and specifying the auction types to be analyzed in this research. It also explores recommendation models in auction theory and important influences on the auction design parameters on savings based on contemporary empirical research. Section 3.2 explores the solution space on how machine learning is utilized in auction design problems, introducing and analyzing several architectures on their functionality and benefits/weaknesses. Section 3.3 specifically approaches LLM's, including the techniques and concepts of retrieval augmented generation and LLM reasoning, which are used in the creation of the chatbot application. Section 3.4 highlights the identified gaps in contemporary research and state of the art solutions, while section 3.5 reflects upon the studied literature and the solutions gap, establishing the opportunity for contributions to the theory.

3.1 Reverse Auctions, Mechanism Design, and Recommendation Models

The subsequent subsections aim to provide a comprehensive background on several key components relevant to this thesis: theoretical frameworks on the design of optimal auctions, the design and functionality of specific auction types, and contemporary research on recommendation models concerning the optimal auction design. Moreover, the impact of auction design parameters on final savings through empirical research is analyzed.

Section 3.1 details the framework of Game Theory and Mechanism Design, offering an in-depth exploration of the mathematical framework underpinning mechanisms and auctions as specialized economic mechanism. This section facilitates a more nuanced understanding of how auction mechanisms are formulated and enables discourse on their design and expected outcomes.

Section 3.1.2 introduces specific examples of auction designs central to this thesis, illustrating the application of theoretical frameworks in practice. The discussed general auction design types are well known and in use by the case company: English auction, Dutch auction, and the Japanese auction. By applying the theoretical frameworks from the previous subsection and by elucidating the theoretical assumptions and consequences of various design decisions, this section lays the groundwork for subsequent analysis.

In Section 3.1.3, attention shifts to a detailed examination of selected auction recommendation models. This section not only highlights the variance and complexity in recommended auction designs across different scenarios but also underscores disparities in conclusions regarding optimal auction design within the research landscape. Furthermore, hypotheses synthesized therein are subsequently subjected to statistical analysis in Chapter 5 and it is a part of the knowledge base built for the chat bot application.

Lastly, Section 3.1.4 offers a conclusion to the background section by providing an overview of several variables and parameters impacting auction savings potential, as evidenced by empirical research in the purchasing domain. By contextualizing the relationship between these variables and negotiating scenarios, this section enhances understanding of factors crucial to determining the optimal auction design, and also serves as part of the knowledge base for the chat bot application.

3.1.1 Game Theory and Mechanism Design: Foundations and Applications in Auction Design

The essential framework to understand the behaviours within auctions, anticipate its expected outcome and asses the influences of design choices is rooted in the domain of Game Theory [81, 5, 65]. On the other hand, Mechanism Design is concerned with creating games or mechanisms to achieve predefined objectives, in alignment with assumptions regarding the participating players and the environment [85, 84, 67].

Game Theory

Game theory is a research area within mathematics and economics, focusing on strategic interactions and decision-making among rational agents in a game [66]. A game is any interaction between agents in which their decisions influence each other. It provides a structured approach to modeling situations where each agent's utility is contingent on the collective choices of agents and establishes an understanding of how these decisions interrelate. At its core, game theory states that players strategize to maximize their utilities, taking into account the potential decisions of others [66]. Game theory's utility in auction theory lies in its analysis of bidder strategies for winning at desirable prices and auctioneer design strategies for revenue or efficiency maximization. The design of an auction can be considered to be part of mechanism design, the counterpart of game theory. While game theory is primarily concerned with the identification of optimal strategies and behaviors for the agents in a game, mechanism design theory adopts the perspective of the game designer, aiming to identify how a game or mechanism should be designed to achieve specific objectives considering the anticipated agent behaviors [79].

Mechanism Design

Mechanism design is the research field that deals with the design of mechanisms, auctions being a certain type of economic mechanisms, that fulfill a set of objectives. It has the reverse view of Game Theory, as it tries to create the games or mechanisms to achieve certain objectives from the game designer's point of view. From the lens of mechanism design, many seminal papers have graced the body of research like the VCG mechanism by

Myerson [64] or the Radio Allocation Spectrum by Mcmilan et al. [56]. Robert Wilson and Paul Milgrom even won the Nobel Prize in 2020 for their contribution to auction theory as well as their involvement in the US Radio Allocation Spectrum Auctions [68]. The concept of a dominant strategy exemplifies the application of game theory and mechanism design to auctions. A strategy qualifies as dominant if it represents the optimal course of action for an agent, independent of the decisions or strategies adopted by other agents [65]. In auction theory, this concept is utilized to construct an auction mechanism where it is a dominant strategy for all players to truthfully reveal their reservation prices [95]. The objective is to create a bidding environment in which, for every auction, participants are naturally incentivized to truthfully disclose their genuine valuation for the auctioned item, resulting in an efficient and transparent market outcome. An example of this principle was provided in one of the seminal works on auction theory by Vickrey [95]. The Vickrey Auction [95] is a sealed-bid auction in which the best bidder wins the auction but pays the price of the second-best bidder. In this auction design, the dominant strategy for all players is to bid their true valuations, because the final price paid by the winning bidder is not determined by his submitted bid but by the second-best bid, making the attempt to bid above their reservation price disadvantageous for the bidder. The design of this second-price auction maximizes that bidder's payoff by bidding one's true valuation regardless of how others bid.

Another significant game theoretical concept for auction design is the Revenue Equivalence Theorem (RET), also introduced by Vickrey [95]. It states that, under specific conditions, all standard auction formats yield identical expected revenue for the auctioneer. This theorem forms one of the fundamental concepts of auction theory, on which the relaxation of the assumptions made reveals the superior performance of different auction formats compared to others. For example, when questioning the assumptions of risk neutrality and identical valuations among buyers, first-price and Dutch auctions tend to offer greater cost savings in scenarios where bidders have risk aversion or when asymmetries exist in the bidders' valuations [54]. Conversely, modifying the assumption that bidders have independent private valuations of the item's worth towards a scenario where bidders' valuations are influenced by the valuations of their competitors, the English auction format becomes more advantageous [59].

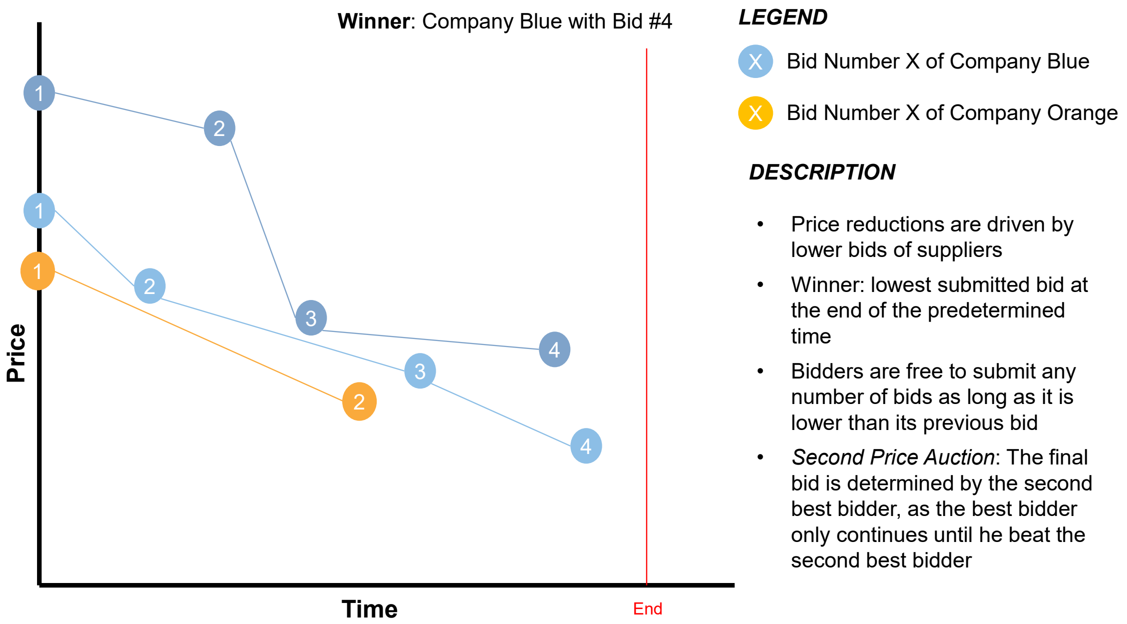
It is important to note that although theoretical auction formats exist that inhibit the truth revealing characteristic, such as the Vickrey auction, they are still rarely deployed in practice [80]. Several factors contribute to the limited application, including the issue of preventing collusion among bidders and the in-applicability of the assumptions made underlying the design of the auction [93, 43, 80]. In most practical industrial negotiation scenarios, established standard auction formats, such as an English, Dutch, or Japanese auction format, are applied [6, 44, 80]. Additionally, most of the auction problem definitions that have closed form analytical solutions are simpler auction types, that do not encompass the multi-attribute, multi-item, and combinatorial auction scenario with different bidder and bidder types that purchasers face in practice [19, 89].

3.1.2 Reverse Auction Types: Dynamic English, Dutch Ticker, and Japanese First Price

The design of an auction encompasses a range of decisions, including the degree of bidding freedom given to participants, the trajectory of the price evolution, the manner in which bid position information is shared, and constraints imposed on time and bid increment/decrement sizes [42]. Some example design considerations are elucidated in Table 2.1.

In most practical industrial negotiation scenarios, established standard auction types, such as an English or ticker auction formats, are applied [6, 44, 80]. The English auction, Dutch auction, and Japanese auction formats are commonly observed in diverse settings such as e-commerce platforms, art auctions, and fish markets [43]. These are also the auction types available in our empirical dataset and in use by the case company. The following sections present a detailed explanation of the mechanism for each of them.

English Auction



- Price reductions are driven by lower bids of suppliers
- Winner: lowest submitted bid at the end of the predetermined time
- Bidders are free to submit any number of bids as long as it is lower than its previous bid
- *Second Price Auction*: The final bid is determined by the second best bidder, as the best bidder only continues until he beat the second best bidder

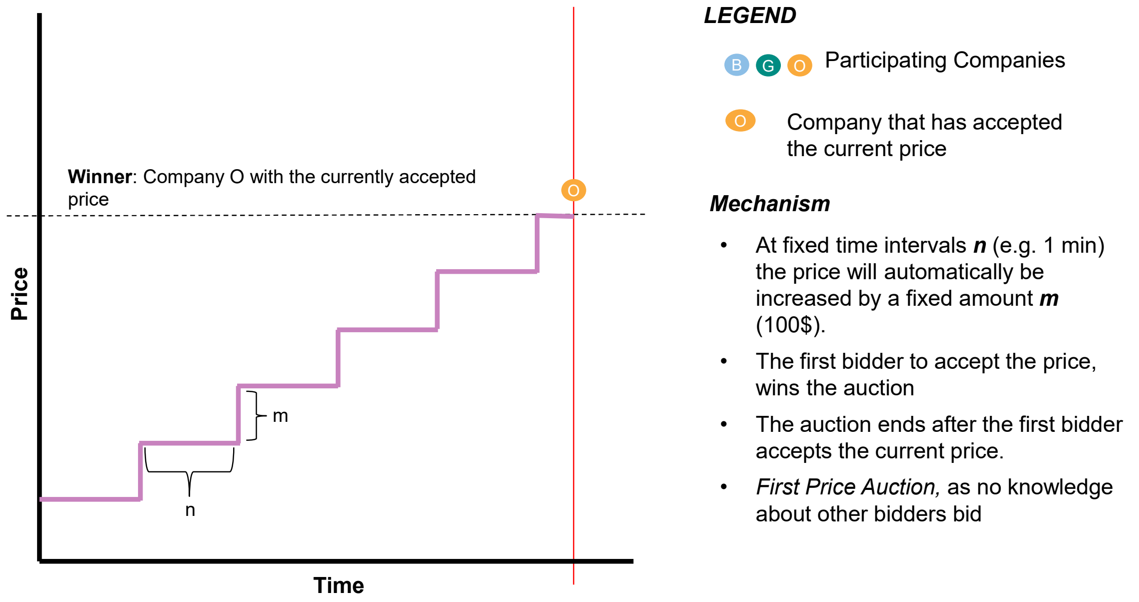
FIGURE 3.1: Visualization of a English auction.

English auctions are the most commonly used auction type and are also known as Open Descending Bid Auctions [62]. It is 'open' because information about other bidders bids is shared among all bidders anonymously. Furthermore, it is based on a descending principle, where bidders need to place bids that are lower than any bid they have previously made. Different from a ticker auction, bidders are given the freedom to choose both the timing and the value of their bid, provided it is lower than their last bid. There is no limitation on the number of bids that can be submitted, as long as the bids are successively lower than the last. The auction ends after a defined period of time, and the auction is won by the bidder who has placed the lowest bid by the conclusion of the process at the price of the submitted bid. This non ticker variant of the English auction is also called the dynamic English auction [6].

Auction types can be adapted through simple modifications to the information shared among all bidders. There are three primary modifications for information exchange: best bid, rank, and traffic light. Best bid reveals the value of the leading bid to all participants anonymously, giving bidders a concrete benchmark on their position and distance from the leading bidder. Still, the submitted bids of each bidder only need to be lower than their last submitted bid, not necessarily lower than the current best bid. A rank modification shares the standing of a bidder relative to other bidders without revealing either bid values or the current leading bid. Traffic light modifications provide even less transparency, sorting bids into only three categories represented by different colors (green, yellow, and red), without revealing the number of other bids in each category. Achieving 'green' status indicates a leading position among an unknown number of bidders, but no more information is provided on a bidder's relative standing or its distance to the leading bid.

From a game-theoretical perspective, the English auction is expected to behave similar to a second-price auction, wherein the final auction price will be close to the second-lowest bid. In an English auction, a bidder reduces their bid incrementally until they reach close to their reservation price/valuation, at which point they quit the auction to avoid bidding below this indifference price. Consequently, the auction is typically won by the bidder with the lowest reservation price, but the winning bid is set just above the second-lowest bidder's reservation price. This is because the leading bidder needs only to outbid the second-lowest bid by a marginal amount to win the auction, rather than approaching their own reservation price further. Moreover, the existing information flow in English auctions allows bidders to adjust their valuations based on others' perceived values of the item, introducing a more dynamic bidding environment than a sealed-bid auction, impacting the auction outcome [6, 84].

Dutch Ticker Auction



Winner: Company O with the currently accepted price
O
m
n
Price
Time
LEGEND
B G O Participating Companies
O Company that has accepted the current price

Mechanism

- At fixed time intervals ***n*** (e.g. 1 min) the price will automatically be increased by a fixed amount ***m*** (100$).
- The first bidder to accept the price, wins the auction
- The auction ends after the first bidder accepts the current price.
- *First Price Auction,* as no knowledge about other bidders bid

Figure 3.2: Visualization of a dutch ticker auction.

A Dutch auction, or ascending ticker bid, starts with a low price that steadily increases by a fixed value at predetermined time intervals. Bidders have the choice to accept the current price or await the subsequent increase. The first bidder to accept the current price wins the auction at the proposed price. If no bidder accepts the price, the auction continues with the increased price, repeating the process until one bidder accepts and ends the auction. Essential to the design of the Dutch auction is the nonexistent information about the number of other bidders or their bids.

Different from the English auction, the Dutch auction is categorized as a first-price auction (Vickrey, 1961), because the first bidder to accept the price wins the auction at that price level. Therefore, the final price of a Dutch auction is expected to be the lowest bid, consisting of a strategic margin added on top of the reservation price of each bidder. The auction design incentivizes bidders to bid close to their reservation value, as the risk of losing the auction by not bidding close to their best possible price is high, and more pressure is applied as no information about other bidders is available.

Japanese Ticker Auction

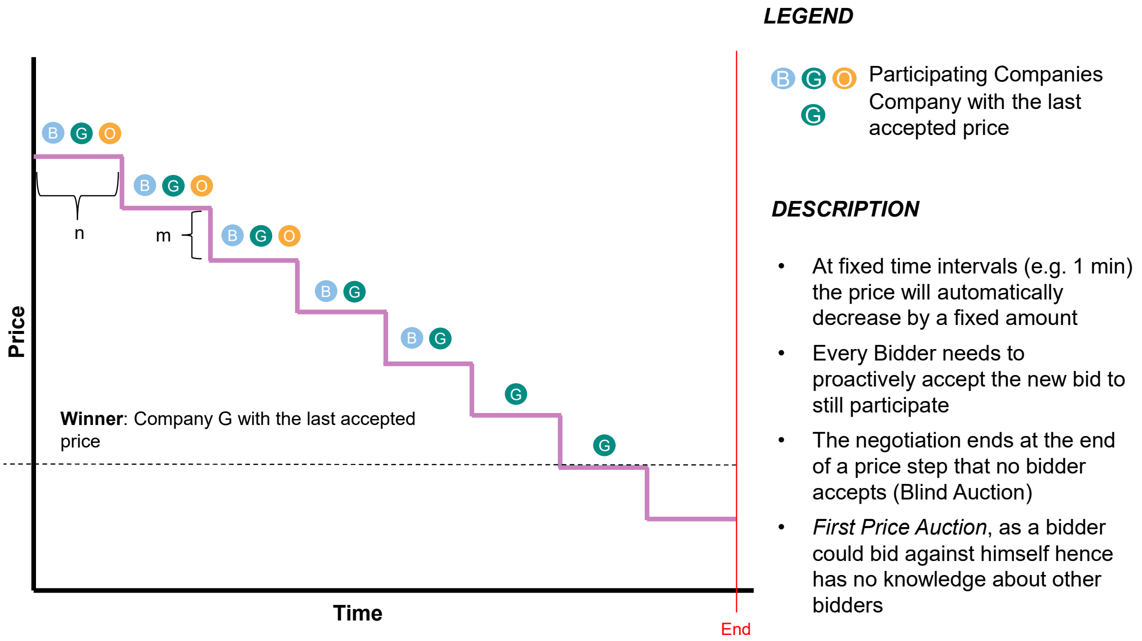

LEGEND
B G O Participating Companies
G Company with the last accepted price
DESCRIPTION
• At fixed time intervals (e.g. 1 min) the price will automatically decrease by a fixed amount
• Every Bidder needs to proactively accept the new bid to still participate
• The negotiation ends at the end of a price step that no bidder accepts (Blind Auction)
• First Price Auction, as a bidder could bid against himself hence has no knowledge about other bidders
Price
Time
End
n
m
B G O
B G O
B G O
B G
B G
G
G
Winner: Company G with the last accepted price

FIGURE 3.3: Visualization of a Japanese auction.

In a Japanese auction, or a descending ticker bid, the price starts high and steadily decreases by a fixed value at predetermined time intervals. Bidders need to proactively confirm their continued participation in the auction at every time step. Once a time step passes and the bidder does not confirm the price, they are excluded from the auction. In this version of a Japanese first-price auction, the auction ends only after a price step passes without any bidder confirming this price. The auction winner is the bidder with the last accepted price step. The winner is informed about winning the auction only after the last not accepted price step ends.

In this specific Japanese auction type, bidders have no information about the bidders remaining or their accepted bids. This allows a single bidder to keep accepting lower price steps, while no other bidder is still in the auction. Because bidders do not have any information about the remaining competitors and because a single bidder can continue to accept lower time steps even if no other bidders are still accepting, the Japanese auction can be classified similar to the Dutch auction as a first-price auction. Therefore, this version of a Japanese auction is also known as a Japanese first-price auction. In contrast to that is the Japanese second-price auction, which ends as soon as the second-best bidder leaves and only one bidder accepts the price step. At the end of this price step, which was only accepted by one bidder, the auction ends. In this case, similar to an English auction, the expected price of the auction is determined by the second-best bidder, implying that it is categorized as a second-price auction. The Japanese second-price auction is also known as an English ticker auction, as it inhibits the same second-price properties but instead of having dynamic bids, it proceeds with falling ticker steps [6, 84].

Similar to the Japanese auction, a Hong Kong auction also has descending ticker steps, but different from the Japanese second-price auction, the bidder who quits the auction and triggers the end of the auction by leaving the last bidder on the price step, also wins the auction. In the Hong Kong auction the exact remaining number of participants is not shared. The Hong Kong auction cannot easily be classified as a second-price auction due to the fact that the bidder who leaves the auction also wins the auction, and therefore it is assumed that the bidders prepare a strategic margin in case they are the worst winner to avoid winning at their indifference price [6]. The studied auction types in the empirical observations are only three: the dynamic English auction, the Dutch auction, and the first-price Japanese auction. Some of the other described auction types are mentioned in the recommendations of the decision-making models that ought to be tested with empirical data, and therefore not every recommendation can be directly comparable.

3.1.3 Auction Type Recommendation Models

For the analysis of the empirical data, three auction type recommendation models are outlined in detail in the following paragraphs. Most recommendation models regarding the optimal auction design in contemporary research focus on the definition of scenarios based upon one or more variables that describe a situation or a game, in which the purchasing professional is required to set up the auction [36, 30] (refer to Appendix A). Optimal auction designs are then proposed for each of these defined scenarios, based upon the characteristic of the scenario and the potential impact of the auction type on the savings potential. A summary of the recommendations and the related hypotheses for all three models is shown in Table 3.1. For the first model, the hypotheses are detailed specifically, while for the other two models, the hypotheses are discussed but only outlined in Table 3.1 to provide a better overview.

Eichstädt (2007) Recommendation Model on Two Dimensions: Number of Bidders and Price Dispersion

Eichstädt [20] bases his auction type recommendation on the evaluation of an empirical survey of 113 companies and the theoretical formulation of the Revenue Equivalence Theorem [95]. The essential assumption of the model is that bidders incur bidding costs that increase with time spent bidding and the disclosure of information about their willingness to accept prices. This implies that bidders stop bidding in auctions, if the probability of winning the auction is perceived to be small. The reduction in active bidders is considered to impact the savings negatively, and therefore the model is framed such that bidders are kept from quitting early.

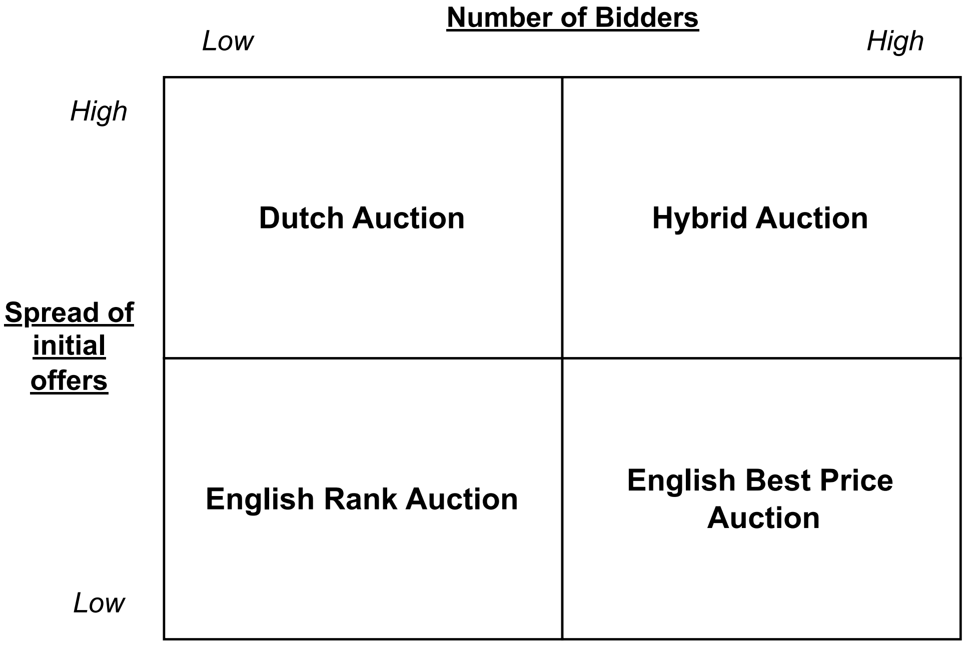
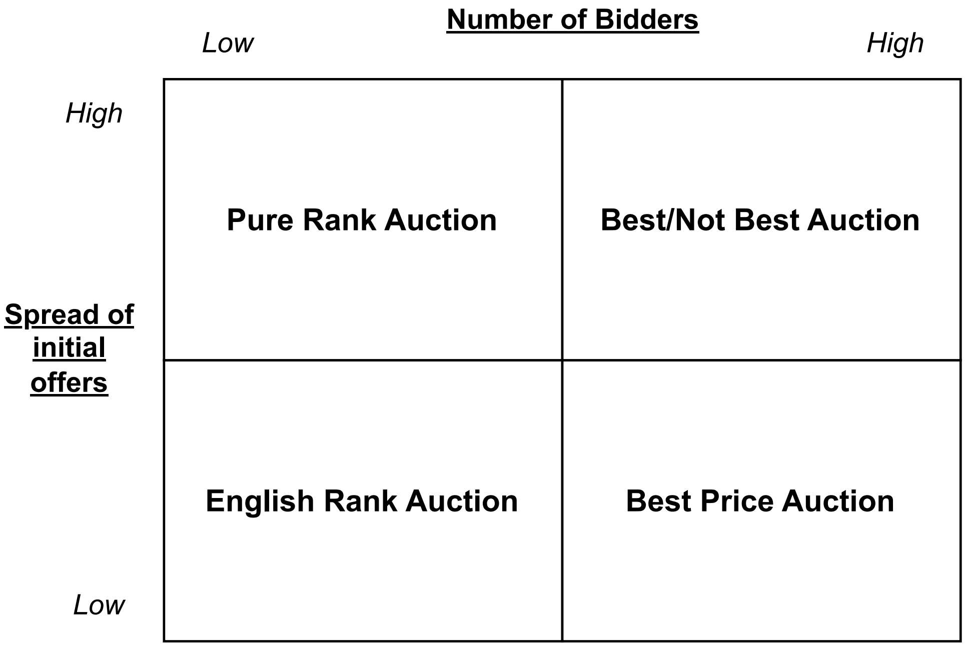


FIGURE 3.4: Summarized visualization of Eichstädt [20] auction type recommendation model

Two dimensions, the number of bidders and the spread of the initial bids, or the price dispersion, are defined as decision criteria in the recommendation model. The spread of the initial bids and the number of bidders are categorized into two levels: 'low' and 'high'. Both criteria have no specific threshold defined to describe the classification of 'low' or 'high', and the specific computation of the initial bid spread is not defined. Based on these four defined quadrants of scenarios, Eichstädt [20] proposes one auction type for each quadrant. Figure 3.4 and 3.5 present a summary of the recommended auction types based on the four described quadrants.

English auctions are deemed to perform superiorly in cases where the bidder's valuations are dependent on each other. This should imply that in competitive markets, the English auction should perform better than a first-price auction. Hence, if the initial bid spread is low and indicates a competitive market, the English auction achieves higher cost savings. Another important factor to consider in scenarios with close bid prices is the assumption that bidders, believing they have a low chance of winning, may quit the auction or cease to bid actively. This withdrawal negatively affects the outcome by decreasing the number of bidders who are actively participating. Hence, if the bid spread is low and the number of bidders is high, showcasing only the best price in an English auction format is expected to achieve higher cost savings compared to showcasing the rank of the bidder.

Revealing the rank of a bidder when the number of bidders is high might disincentivize bidders who have a low rank to bid further. The best auction format therefore hides the rank of the bidders but highlights the short distance to the leading bid, encouraging suppliers to assume that their probability of winning is high. If the number of bidders is low, revealing information about a low rank and a close best bid encourages bidders to perceive their probability of winning as high. Therefore, the following hypotheses are formulated:



Spread of initial offers / Number of Bidders	Low	High
High	Pure Rank Auction	Best/Not Best Auction
Low	English Rank Auction	Best Price Auction

FIGURE 3.5: Summarized visualization of Eichstädt [20] information policy recommendation model

- H1E: With a low number of bidders and low price dispersion, an English rank and best bid auction achieves the highest cost savings
- H3E: With a high number of bidders and low price dispersion, an English best bid auction achieves the highest cost savings

Eichstädt [20] furthermore argues that first-price and Dutch auctions achieve superior cost savings in the case of risk averse bidders or when asymmetries exist in bidder valuations, in line with the Revenue Equivalence Theorem [95](Vickrey, 1961). This is also due to the fact that first-price auctions hide the amount of other bidders participating, increasing the insecurity of the participants. Therefore, the first-price auctions are expected to achieve higher cost savings when the spread of the initial offers is high, which indicates higher asymmetries amongst bidders. In case a large number of bidder coincides with a large bid spread, a hybrid auction format, first an English auction, followed by a Dutch ticker auction is recommended. Thereby, the English auction utilizes the positive impact of the open information exchange on the adaptation of bidders prices based on each other's valuations, and then a Dutch ticker auction follows up to pressure the possible dominant bidder. This hybrid auction format is also known as the Klemperer auction [42]. This argumentation yields the following hypothesis:

- H2E: With a low number of bidders and a high price dispersion, a Dutch auction achieves the highest cost savings
- H4E: With a high number of bidders and a high price dispersion, a hybrid auction consisting of an English, followed by a Dutch auction achieves the highest cost savings

Moreover, a recommendation is made on variations of the English auctions that differ only by the type of information shared with bidders, such as a ranking or only the highest bid. This recommendation follows the similar assumption that auction designs should increase the perceived probability of bidders winning the auction so as to mitigate the negative impacts of bidders exiting the auction. If the number of bidders is low, communicating the rank to bidders makes use of the perceived closeness to the best rank. Adding a high bid spread, a pure rank auction is recommended without showing the distance to the best bid, mitigating the adverse effects of the larger distance to the best bid. But, if the bids are closer together, a rank and the best bid should be communicated, encouraging the bidders to bid more due to their close rank and close distance to the best bid.

- H5E: With a high price dispersion and a low number of bidders, an English pure rank auction achieves the highest cost savings among English auctions

It is argued that if the number of bidders is high, the goal is to mitigate the discouraging effect of the bidder perceiving a large number of potential competitors in the auction. Combined with low spreads, a best price auction is recommended to profit from the encouragement of bidders due to the small distance to the best bid. On the other hand, if the bid spread is high, only the best bidder should be informed about his position, while all other participants do not receive any information (blind auction).

- H6E: With a high price dispersion and a high number of bidders, an English Best bid Blind auction achieves the highest cost savings among English auctions.

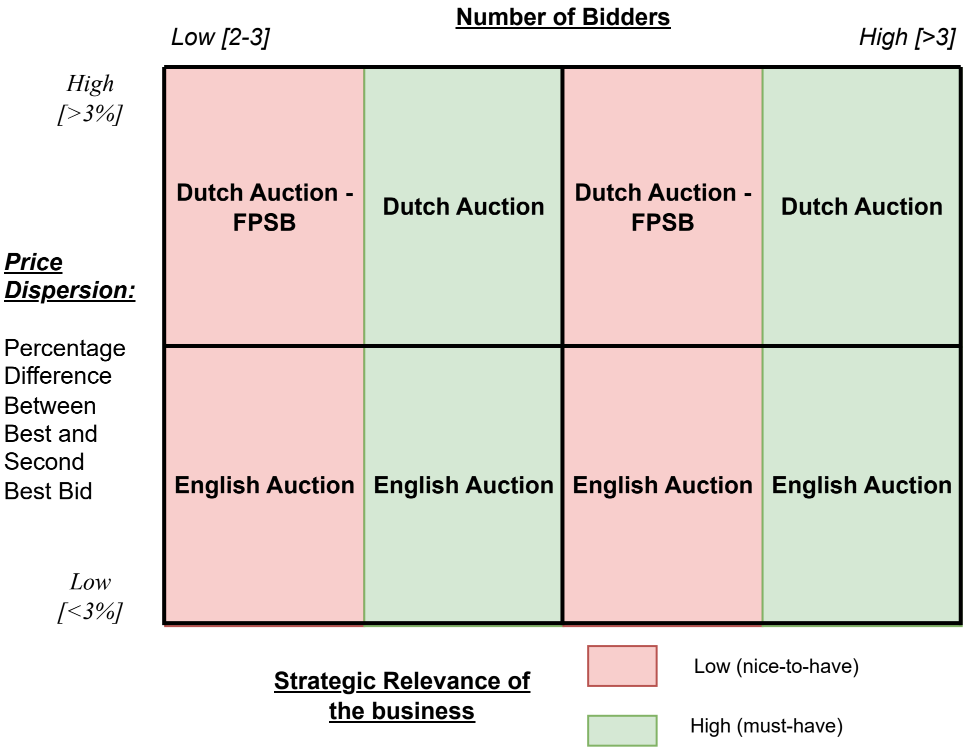
As opposed to Eichstädt (2007), the following model of Schulze-Horn et al. (2018) introduces a third dimension to the model and chose their recommendation based on empirical insights gained at a case company.

Schulze-Horn et al. (2018) Mechanism Design based Recommendation Model on Three Dimensions: Number of Bidders, Price Dispersion and Bidder's Risk Aversion

Schulze-Horn et al. [84] propose a decision-making model based on the an additional dimension of the bidders risk aversion. The elicitation of the recommendations comes from empirical insights gained through an action research methodology, combining data from documents, participant observations, and workshops, as well as participation in the execution of mechanism design-based negotiations at a case company. Differing from the other models, this paper has a broader focus to not only include auctions but encompass broader mechanism designs, such as exclusive offers.

Different from the other discussed works, Schulze-Horn et al. [84] operationalize the initial bid spread specifically. The spread of the initial bids is defined by the difference between the best and second-best initial offers, and a threshold of 3% is given for the classification of a 'high' or 'low' difference. Still, the evaluation of the attractiveness of the suppliers needs to be estimated by the purchaser based on market knowledge and relationship insights. Figure 3.6 showcases the recommendations of Schulze-Horn et al. in a summarized visualization. The third dimension, the attractivity or also known as the strategic relevance of the business is depicted by the red and green colours of the quadrants.

Distinct in this work is the inclusion of a sequential Dutch auction-type variation. Different from the parallel case, where all bidders get the same quote at the same time, in the sequential version, a previous ranking establishes the order in which bidders receive the first quotation. The first rank can then decide whether to accept the given price, and if the price is declined, the same price is offered to the second rank until one participant accepts.



	Low [2-3] 🟥	Low [2-3] 🟩	High [>3] 🟥	High [>3] 🟩
High [>3%]	Dutch Auction - FPSB	Dutch Auction	Dutch Auction - FPSB	Dutch Auction
Low [<3%]	English Auction	English Auction	English Auction	English Auction

🟥 Low (nice-to-have)
🟩 High (must-have)

FIGURE 3.6: Summarized visualization of Schulze-Horn et al. [84] auction type recommendation model

The classical and previously defined Dutch auction type is thereby named a parallel Dutch auction, in which each participant receives the same quotation at the same time. Translating the different auction types as stipulated by Schulze-Horn et al. [84] into our definition, a proposed English ticker auction is synonymous with a Japanese second price auction and falls into the category of second-price mechanisms such as the open bid English auction. To facilitate the comparison between the multiple-phase mechanisms, all initial requotes of the initial bids are omitted, as are all subsequent mechanisms, such as exclusive offers, from the comparison. The considered auction type from Schulze-Horn et al. [84] is shown in Figure 3.6 Table 3.1, along with the named hypotheses to be tested on the empirical data.

If the bid spread is low, Schulze-Horn et al. [84] always recommend a Japanese 2nd price auction (English ticker auction). This is because a Japanese 2nd price auction is favorable if there are at least two similar strong bidders in the auction that can challenge each other towards their reservation price. This follows the same line of argument for the general second-price auction types, including the English auction, as proposed by Eichstädt [20]. If there is an indication of asymmetries in bidders valuation by a higher bid spread, in line with the other authors, a Dutch auction format is recommended. Here a distinction is made: if the attractivity of the lot is high, the sequential version of the Dutch auction is preferred to encourage bidders from the lower ranks to seize the opportunity that the first ranks passed by adjusting their reservation price. A FPSB is always following a parallel Dutch auction in cases where the attractivity of the business is 'low' for the supplier.

The following model by Berz et al. [6] propose the same three dimensions as Schulze-Horn et al. [84], but define the operationalization of the attractivty dimension differently and elicit the recommendations based on a game theoretical framework rather than empirical knowledge.

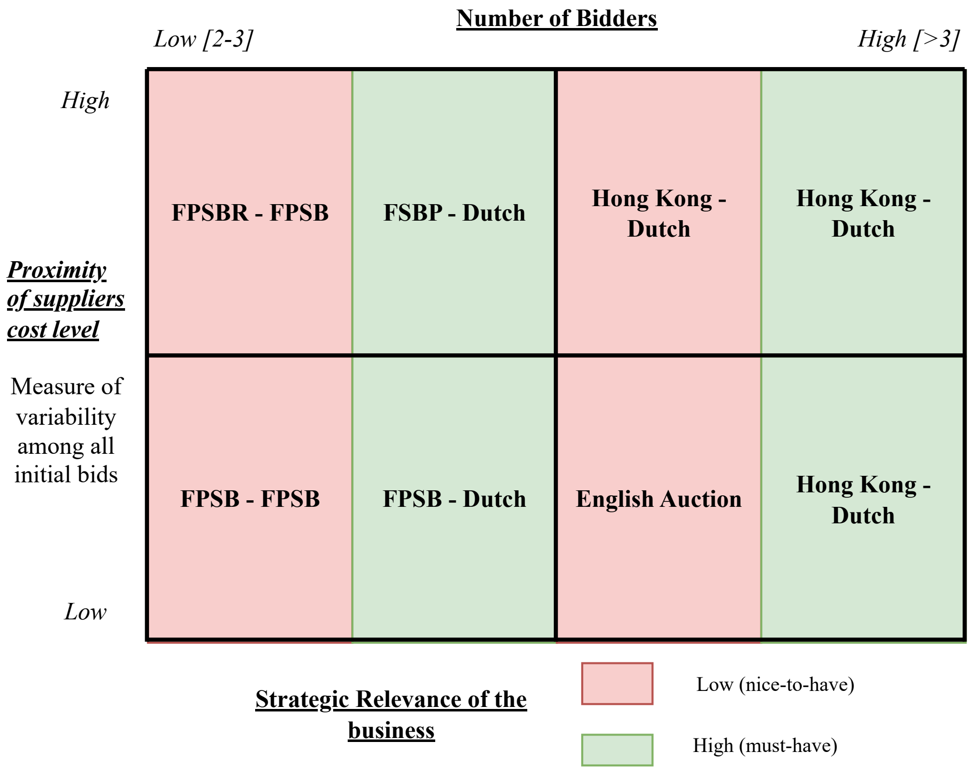
Berz et al. (2021) Game Theoretical Recommendation Model on Three Dimensions: Number of Bidders, Price Dispersion and Bidder's Risk Aversion

Berz et al. [6] study the challenge of the optimal auction design for single-lot negotiations by theoretically approaching the problem through order statistics. Defining the expected results of first-price and second-price auctions, the authors find a criterion that defines the threshold for the choice between first- and second-price auctions. By challenging assumptions made by the Revenue Equivalence Theorem of Milgrom [59] as described previously and utilizing order statistics, Berz et al. [6] compare first and second-price auctions under different scenarios. The authors conclude a recommendation model based on three decision criteria, coined the Auction Cube. For the auction cube to be applicable, three preconditions are defined:

1. Competitive market with no implicit or explicit collusion and at least two bidders or the perception of at least two bidders
2. Same level of risk aversion to not winning the auction for all bidders and maximal price sensitivity of the buyer/auctioneer with regards to the bidders offers.
3. Only a single object of negotiation, compared to multiple lots that may be won by multiple different bidders.

Berz et al. [6] extends the model of Eichstädt [20] with a third dimension, the strategic relevance of the business to the supplier. Similar to Eichstädt [20], the three decision criteria—the number of bidders, the bid spread of initial offers, and the attractiveness of the business to the buyer—are categorized into two levels: 'low' and 'high'. For the number of bidders, more than three participating bidders are considered to be 'high'. The spread of the initial bid or the proximity of the bidders is approximated in practice by the sample standard deviation of the initial bids submitted, but no specific threshold for the categorization is defined. The sample standard deviation is a statistical measure that quantifies the amount of variation or dispersion within a dataset. It calculates the average distance between each data point and the mean of the dataset, highlighting the extent to which individual data points deviate from the sample mean. This third dimension is considered to be an approximation of the possible risk aversion of a supplier, which again is inversely related to the strategic margin that a supplier adds towards his reservation price.

First, Berz et al. [6] define the expected results of a second and first-price auction mathematically utilizing orders statistics. In a second-price auction, the participant with the lowest indifference price wins the auction, but at the value of the indifference price of the second lowest valuation [6, 44]. The expected result of a second-price auction is therefore defined as the expected value of the second best valuation, formally described by the value of the second orders statistics. On the other hand, in a first price auction, the participant with the lowest price wins at this specific price, prompting participants to add a strategic margin "M" to their indifference price to avoid a profit of zero [6, 44]. Therefore, the expected result is the lowest indifference price, or first-price, with an addes strategic Margin 'M'. By then introducing the computation of a threshold 'G' that defines up to which value of the strategic margin 'M' a first price auction is preferable to a second price auction, Berz et al. [6] find a specific way on how to theoretically categorize the attractiveness of the business to the supplier via order statistics. They assume the


Number of Bidders
Low [2-3]
High [>3]
High
Proximity of suppliers cost level
Measure of variability among all initial bids
Low
Strategic Relevance of the business
🟥 Low (nice-to-have)
🟩 High (must-have)

Proximity of suppliers cost level	Low [2-3] 🟥	Low [2-3] 🟩	High [>3] 🟥	High [>3] 🟩
High	FPSBR - FPSB	FSBP - Dutch	Hong Kong - Dutch	Hong Kong - Dutch
Low	FPSB - FPSB	FPSB - Dutch	English Auction	Hong Kong - Dutch

FIGURE 3.7: Summarized visualization of Berz et al. [6] auction type recommendation model

bidders indifference prices to be normally distributed and strategic margins to be symmetric, implying all bidders need to have the same strategic margin (which is likely to be an unrealistic assumption). Furthermore, the utilization of the threshold G requires the estimation of the strategic margin of suppliers, which in most practical cases necessitates a specific cost structure analysis for every negotiation situation and bidder. Therefore, the practical operationalization of the attractiveness of the business to the supplier remains a challenge.

Based on these three decision criteria and the two levels for each, the Auction Cube presents eight different scenarios and recommendations for auction types, highlighted in Table 3.1. The hypotheses to be studied are also introduced in Table 3.1, labeled by square brackets and an identifier. The basis for the recommendations is the design of the Klemperer auction [42]. The Klemperer auction, also known as the Anglo-Dutch auction, is a hybrid auction with English auction as its first phase, followed by a Dutch auction, as previously recommended in Eichstädt [20]. Instead of an English auction that aims to influence the indifference prices of the bidders through price transparency, a Hong Kong auction with two winners is proposed that then goes into the Dutch auction due to its established result in industrial practice [6, 5].

In line with Eichstädt [20], in the case of a high number of bidders and a low bid spread, an English auction is recommended, but only if the attractivity of the business to the bidders is low and a lower risk aversion to winning the lot is to be expected. If the risk aversion is high, it is recommended to perform a hybrid auction consisting of the Hong

Kong-Dutch auction.

Generally, if there is an indication of a high attractivity of the business to the bidders, implying that the strategic margin is small and the risk aversion of winning the auction is high, the Auction Cube recommends conducting a first-price auction, namely a Dutch auction, in the second phase of the two-phase auction design. In cases where the number of bidders allows for it, a Hong Kong auction preempts the Dutch auction. Otherwise, a first-price sealed bid is recommended to be used in the first round.

Both authors also agree to perform a Dutch auction if the number of bidders is low and the bid spread is high, but Berz et al. (2021) argue that a First Price Sealed Bid auction (FPSB) needs to preempt the Dutch auction. Given a low-risk aversion, even if the number of bidders is low and the spread is high, the auction cube speaks against a Dutch auction in favor of a FPSB with two winners, followed by another FPSB.

Following a similar line of reasoning as Eichstädt [20], if bid dispersion is high and risk aversion is low among a few bidders, an FPSBR-FPSB auction is recommended. The first round is a first-price sealed bid auction that informs the bidders at the end only of their rank rather than the best price. This design aims to maintain competitiveness without discouraging the higher bidder by only revealing rank, not exact bid amounts.

Concluding, both models agree that in cases of a large number of bidders and a large bid spread, a hybrid auction that ends in a first-price auction is recommended. The difference is that Eichstädt [20] recommends an English auction in the first round to enable competitors to update their evaluations based on competitor bids, while Berz et al. [6] recommend a Hong Kong auction with two winners. Table 3.1 summarizes the recommendations made and classifies the hypotheses to be tested with the empirical analysis.

Given the three models and their differences in recommendations, it is valuable to empirically investigate which recommendation model is more accurate and achieves higher savings.

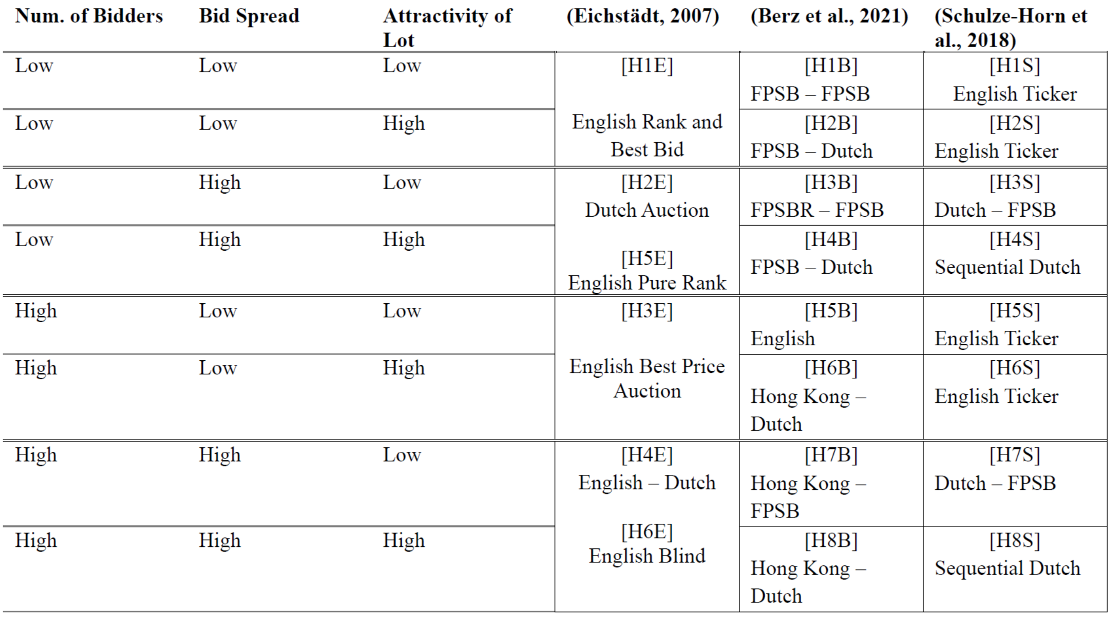
3.1.4 Empirical Influence of Auction Design Parameters on Savings

A collection of general statements made from the research papers can be found in Table 3.2. Table 3.2 summarizes the key auction design parameters and their expected impact on the auction savings/buyer surplus, given the existence of certain conditions outlined and their respective source citation. The collection of statements from the different papers that studied the influence of auction design parameters on the buyer surplus will be the basis for textual knowledge database of the LLM based system and also serves the evaluation of the LLM models output. In what follows, the key parameters indicated in Table 3.2, will be explained in more detail.

Amount of Bidders and Bidding Activity

Many authors analyze the key parameter of the number of bidders and number of bids on the buyer surplus [62, 36, 23, 2, 70]. In line with the theoretical literature, the authors observe that the number of bidders and the number of bids have a positive influence on the buyer surplus. A higher number of bidders implies increased competition among the

Table 3.1: Auction type recommendations per defined scenario and summary of hypotheses named in square brackets



Num. of Bidders	Bid Spread	Attractivity of Lot	(Eichstädt, 2007)	(Berz et al., 2021)	(Schulze-Horn et al., 2018)
Low	Low	Low	[H1E] English Rank and Best Bid	[H1B] FPSB – FPSB	[H1S] English Ticker
Low	Low	High		[H2B] FPSB – Dutch	[H2S] English Ticker
Low	High	Low	[H2E] Dutch Auction [H5E] English Pure Rank	[H3B] FPSBR – FPSB	[H3S] Dutch – FPSB
Low	High	High		[H4B] FPSB – Dutch	[H4S] Sequential Dutch
High	Low	Low	[H3E] English Best Price Auction	[H5B] English	[H5S] English Ticker
High	Low	High		[H6B] Hong Kong – Dutch	[H6S] English Ticker
High	High	Low	[H4E] English – Dutch [H6E] English Blind	[H7B] Hong Kong – FPSB	[H7S] Dutch – FPSB
High	High	High		[H8B] Hong Kong – Dutch	[H8S] Sequential Dutch

supplier base for the auction. The number of bids is considered an even stronger indication of competition in an auction, also resulting in an even higher influence and relevance than the number of bidders in a couple of works [62, 2, 70]. It can be observed that some bidders are not participating in the auction due to a variety of possible reasons, among them the strategy to not participate in the auction but acquire information about the cost structure of competitors [70]. Park et al. [70] also analyze the influence of the number of repeated bidders in auctions and find that repeated bidders contribute the majority of bids in auctions. They can also become inactive if their competitive bids do not yield a winning bid, discouraging them from participating actively the next time around. Similar to that is the percent of bidding participants, analyzed by Millet et al. [60], which was observed to have the most significant influence on the buyer surplus among all the tested variables in their research. Summarizing, we formulate the following statements:

s1a The higher the number of bidders in an auction, the higher the buyer surplus

s1b The higher the number of bids in an auction, the higher the buyer surplus

s1c The higher the number of repeated bidders participating in an auction, the higher the buyer surplus

s1d The higher the percentage of bidders that participate in an auction, the higher the buyer surplus

Information Exchange Mechanism

Another widely studied parameter is the Information Revelation Mechanism. In the design of an auction, the auctioneer can decide how much information each bidder receives from the other participating bidders and his position relative to the competition. In a sealed bid auction, no information about any other bids is shared with the bidder, only information

Key Parameter	Impact on Buyer Surplus	Conditions	Citation
# of bidders	positive	[2]: third bidder needs to be a low cost type	[62], [36], [23], [2], [70]
# of repeated bidders	positive		[70]
# of bids	positive		[62], [23]
Bid Decrement	no effect		[62]
Setting a Reserve Price	positive		[62]
Lower Reserve Price	positive		[62]
Auction Duration	no effect		[62]
Auction Duration	positive	exogenous bidder entry	[47]
Higher Information Transparency	negative		[62], [99], [103], [30], [15]
Higher Information Transparency	positive	lower cost certainty;	[87]
Auction Type (English vs Dutch)	postive (English); negative (dutch)		[36], [51]
Well Informed Bidder Type	positive	well informed = has a precise definition of his cost estimates and the value of the item	[99]
Percent Bidding	positive		[60]
Allowing overtime	positive		[60]
Allowing overtime	negative	Only with Best Bid and Rank Feedback	[60]
Two phase negotiations	positive	Complex Item project including non-price attributes	[57, 6, 5]

Table 3.2: Summarized influences of studied key auction design parameters on the buyer surplus. The key parameters are listed along with the expected impact on the savings of an auction mechanism. The third column highlights some conditions posed by the papers for this relation to hold and the fourth column references the citations that observe these relations between key parameters and their impact on the auction savings.

about the amount of his own bid is known to him. Then the next level comprises sharing the relative position of the bidder by means of a traffic light. In the traffic light system, once a bidder submits a bid he gets a green, yellow, or red light as feedback. For example, a red light could indicate that his bids are among the lower range of submitted bids, a yellow light that he is in the middle range, and a green light that his bid is among the best. Furthermore, rank information could be shown to the bidder, for example, a bidder receives information that his bid is on Rank 3. Additionally, the current best bid could be shown to all bidders. These different mechanisms have different anticipated effects on bidders' behavior [36, 70]. The mechanisms mainly differentiate themselves through the level of information revelation and transparency they provide to the bidder. Many authors observed through their analysis that a higher level of information transparency influences the buyer surplus negatively in most auction scenarios [62, 99, 103, 30, 15]. But Setia et al. [87] find that higher information transparency influences the buyer surplus positively in case the cost structure certainty of each participating bidder is low, hence increased information transparency about other bids might promote bidders to get more confident estimations of possible realistic bids. Apart from the statistical analysis in the literature, there are some decision-making models that advise auctioneers to choose different information revelation mechanisms based on a couple of key parameters and situations. The number of bidders and the spread between the lowest and highest initial bid across the participating bidders are considered the two key parameters that should inform the decision on which information mechanism to choose [5, 6]. From a collection of internal training material and decision-making models from the literature [6, 103], the following statements are formulated:

s2a The less information transparent an auction is the higher the buyer surplus

s2b Rank and best bid information exchange mechanisms have higher average buyer surplus than only rank or only best bid information exchange options.

s2c If the number of bidders is high and the initial bid spread is low, the best bid information exchange has a higher average buyer surplus *s2d* If the number of bidders is low, and the initial bid spread is low, the rank and best bid option has the highest buyer surplus

s2e If the number of bidders is low and the initial bid spread is high, the rank option has the highest buyer surplus.

s2f If the number of bidders is high and the initial bid spread is high a traffic light option has the highest buyer surplus on average.

Reserve Price

According to Mithas and Jones [62] setting a reserve price in general and having this reserve price be a value close to or even lower than historically achieved prices is of significant positive influence on the buyer surplus. Reserve prices act as the maximum possible price a buyer would be willing to pay for the item or service, hence all submitted bids need to be lower than the set reserve price. We formulate the following statements:

s3a Setting a reserve price yields a higher average buyer surplus than not setting a reserve price

s3b Setting a relatively low reserve price based on historical prices yields a higher average buyer surplus than setting a relatively high reserve price

Auction Duration and Overtime

Different conclusions can be drawn by different authors on the impact of the auction duration on buyer surplus. Liang et al. [47] conclude that longer auction duration has a positive influence on the buyer surplus, given that the auction is dependent on exogenous bidder entry. Exogenous bidder entry means that the fact that bidders join is dependent on the auction design. For example, very small auction duration could impact the decision of bidders to participate. Mithas and Jones [62] on the other hand do not observe any impact of the auction duration on the buyer surplus, based on their dataset from a large automotive OEM.

Millet et al. [60] also study the effect of allowing overtime in an auction and observe different impacts based on the information exchange mechanism of the auction. They find that overtime has a positive impact on buyer surplus on all auction types except for best-bid-and-rank auctions, for which its impact is negative. Therefore we define the following statements:

s4a Given any information exchange type other than best-bid-and-rank, allowing overtime in an auction yields a higher average buyer surplus

s4b Given a best-bid-and-rank information exchange type, allowing overtime yields a lower average buyer surplus

Apart from the compiled statements, the literature review also yields some additional insights which are discussed in the following sections in more detail.

Inconsistent Conclusions

Empirical research also comes to conclusions inconsistent with theory and theoretical assumptions. Aloysius et al. [2] found in their experimental investigation that bidders tend to change their bidding behavior in a manner that is inconsistent with theory when an additional bidder joins the auction. Chen-Ritzo et al. [15] conclude from their experimental analysis of a proposed auction mechanism that practical bidder behavior is not close to theoretical predictions even when experienced bidders participate in the auction. Mithas and Jones [62] also observe a difference between theoretical predictions and practice on the influence of reverse prices on the buyer surplus. While theoretical approaches predict that setting a reserve price increases buyer surplus, empirical research observes that an increase in the reserve price is associated with a decreased probability of selling the good and decreasing buyer surplus, but higher revenues if the goods are sold. Mithas and Jones [62] also compile the predictions of theoretical research with their empirical results and find several discrepancies in conclusions related to the auction duration and bid decrement, which had no effects on the buyer surplus according to their analysis. Furthermore, Wooten et al. [99] identify a reversal of a theoretical prediction compared to practice on the hypothesis that first-price auctions increase buyer surplus compared to second-price auctions in a private value setting. Similar to Mithas and Jones [62], Aloysius et al. [2], Block et al. [7], Matthäus et al. [55], Wooten et al. [99] come to the conclusion that bidder behavior in practice is difficult to predict even for private value settings, yielding inconsistencies between the theoretical literature and empirical practice, especially considering common value settings in which the revenue equivalence theorem does not hold [16].

Majority of Research on Publicly Available Data

Another observation is that empirical research is mostly based on publicly available data such as public governmental procurement auctions [51, 55, 28] and laboratory/experimental setups with professionals or students [15, 87, 30, 2]. Due to the confidential nature of private auction datasets, they rarely appear in the literature [31], although there are a couple of papers that utilize a relationship with an industrial partner like [62, 36, 47, 60].

3.2 Machine Learning in Auction Design: Surpassing Theoretical Boundaries

In the following section, the existing state-of-the-art machine learning models that address the task of designing optimal auctions are introduced and elucidated. They can be categorized according to the class of machine learning model architectures they are based upon. Initial solutions in this research became prominent on the basis of the Multi-Layer Perceptron (MLP) architecture [76]. Foremost the seminal MenueNet [89], and RegretNet [19], emerge as the most prominent contemporary solutions to the optimal auction design problem in machine learning research, and are detailed in Subsection 3.2.1. Another type of machine learning architecture, the DeepSet architecture [104], addresses certain limitations of the MLP-based architectures, such as potential data scarcity, model sensitivity to data order, and variations in data input sizes, as described in Subsection 3.2.2. Subsection 3.2.3 introduces reinforcement learning approaches [10] to auction mechanism design, which exhibit greater adaptability to diverse auction scenarios, albeit encountering convergence issues. Section 3.2.4 concludes with an analysis and comparison of the differing strengths and weaknesses of these machine learning solutions in the context of designing the optimal auction based on five factors defined on the basis of the solutions objectives from Section 2.4.

3.2.1 Multi-Layer Perceptron Architectures

Researchers have utilized the capability of deep learning approaches for the optimization of auction designs, trying to solve the existing bottlenecks of theoretical frameworks and numerical intractability [104]. One of the earliest works at the intersection of deep learning and automated mechanism design is the use of Multi Layer Perceptrons (MLP) [76] with the aim to find theoretical optimal mechanisms, titled the MenueNet [89]. The MenuNet is followed by its sucessor, the RegretNet [19], which in turn is the foundation for the PreferenceNet [73] and other extensions that address a subset of assumptions or limitations of their predecessors.

MenueNet

This MenueNet is split into two networks, the Mechanism and the Buyer Network. The Mechanism Network receives a simple one-dimensional constant as input and subsequently generates two outputs: an allocation matrix and a payment vector. The output of the mechanism network are then used as inputs to the buyer network, which computes the buyers strategy for the supplied mechanism and its associated utility. The MenueNets optimization procedure is guided by the Incentive Compatability (IC) and Individual Rationality (IR) Constraints. Incentive Compatability [86, 43] tries to ensure that the best strategy for all participants in an auction is to bid according to their true valuation. The revelation of the true valuation of bidders implies more efficient allocations and a reduction of the likelihood of strategic bidding or manipulation. Individual Rationality ensures participants in an auction will perceive to only benefit from participating in the auction. The architecture is kept simple for efficient optimization and training procedures of the MLP, and the authors have been able to verify that the MenueNet can compute theoretically known optimal mechanisms for the single buyer case [89]. The aim of the MenueNet architecture is to find theoretical mechanisms that achieve optimal efficiency in the allocation of goods and provide mechanismsm that satisfy the IC and IR constraints. This is also represetented by the fact that the MenueNet does not use any practical auction data to train its model. Other Limitations of the MenueNet are on the complexity and variety of possible mechanisms that it can represent, additionally to the deterministic nature of the assumed buyer valuations for the items [104].

RegretNet

Building upon the MenuNet, the RegretNet represents a more complex and flexible approach to automated mechanism design [19]. Simmilar to MenueNet, RegretNet has an allocation network and payment network. Different to MenueNet, RegretNet applies additional constraints on the objective function during the optimization process, with the aim to deter participants or buyers from hiding their true valuations, making the auctions Incentive Compatible and minimizing regret. Thereby, regret is defined as the loss incurred by a buyer for not bidding his true valuation in the auction. Also, RegretNet is trained using a uniformly drawn sample of allocations as inputs. In General, RegretNet is able to find optimal solutions for a wider range of different scenarios, including more complex mutli-item and higher variety in buyer valuations. It is still limited by the fact that it assumes additive and unit demand valuations, while also needing to be retrained with more extensive computational requirements than MenueNet for every single auction scenario. Additive and unit demand valuations impose that the total value assigned to multiple items is the sum of the individual valuations for each item, ignoring possible synergy values between bundles of items and restricting the buyer only to be interested in a single item respectively.

PreferenceNet

Peri et al. [73] introduced PreferenceNet, one of the extensions of RegretNet that integrates human preferences into auction designs; this architecture combines the RegretNet with a 3-layer MLP, trained using expectation maximization to refine predictions based on observed bidder behaviors and preferences for items. The new 3 layer MLP acts as an adaptable allocation predictor, helping to train the RegretNet with the added preference information of bidder valuations. This addition enables to apply the model in a more dynamic and responsive manner. Feng et al. [25] created an additional neural network using RegretNet's structure. This model integrated budget constraints and tackled both Bayesian Incentive Compatible (BIC) as well as conditional IC restrictions. The augmentation of budget constraints and BIC amplifies the capacity of RegretNet to manage more intricate forms of IC, thereby enhancing its flexibility in diverse auction environments [104].

3.2.2 DeepSets Architecture

DeepSets [104] are neural networks specifically designed to process set-based data, differing from traditional structures like Multi-Layer Perceptrons. Unlike these usually fixed-size and order-sensitive models, DeepSets exhibit permutation invariance. The capacity for permutation invariance is given by its abillity to independently address each element within the input set and consolidating information through pooling operations (sums, means or maximums). Hence, DeepSets are suitable for auction data, where input data lacks a natural order or varies in size. Auctions often involve multiple items with each bid combining into a set of valuations for these items. MLP based models suffer from the restrictions on the fixed size and order sensitivity of inputs, however DeepSets can processes the input data regardless of their number or order.

Liu et al. [49] propose the Deep Neural Auction model based on the DeepSet architecture and tailor its application on the auctions of online ad spaces. The model comprises of three parts: The Set Encoder, the Rank Score Network and the Differentiabel Sorting Engine. First, the encoder takes the input set and creates set embedings, a compact representation of all input features, enabling the model to handle any number of input bids/ads. Then the Rank Score Network applies a score for every ad included in the set following constraints like non-negativity that represent the value of each ad/bid to the seller. Then the Sorting engine based on the NeuralSort architecture that is subject to machine learning optimization methods, sorts the ads/bids according to their calculated value. In this dynamic approach, the model is able to continuously learn and optimize its Ranks Score and Sorting Engine Networks with a wide applicability to many auction scenarios, while being limited by its computational complexity, and intensive requirement for a extensive amount of high dimensional feature data.

3.2.3 Reinforcement Learning

Apart from MLP approaches, reinforcement learning approaches have been used to navigate the variety, complexities, and dynamic nature of auctions and mechanisms. In reinforcement learning (RL), an agent actively engages with its environment, making decisions and taking actions to optimize cumulative reward [97]. Unlike traditional auction methods that operate on fixed regulations and predefined strategies, RL enables the auction mechanism to perpetually change and adapt based on participants' behaviors and interactions within the auction. For example, Cai et al. [10] apply RL and the notion of no regret from the RegretNet architecture to design cost-optimal mechanisms using historical data on the behavior of participants and the evolution of the auction. While designed to be very adaptable to different scenarios and constantly changing to adapt its mechanisms to a range of bidder behaviours, reinforcement learning methods often struggle to converge for more complex scenarios with a number of bidders as the size of the state space to explore scales exponentially. Moreover, these model needs extensive amounts of historical data as well as a modeled auction environment to be able to learn via exploration and exploitation [10].

3.2.4 Strengths and Weaknesses

The strengths and weaknesses of each model related to five main factors, the flexibility, the data dependency, the computational efficiency, the interpretability and the constraints imposed on the model are highlighted in Table 3.3. While all of these methods try to solve the greater problem of designing revenue optimal mechanisms, their main drawback is the lack of interpretability and explainability, yielding a low basis of trust from the practical purchasers perspective. Another dimension they lack in is that most of these models aim to present the optimal mechanism based on multiple assumptions like the additive and unit demand constraints, limited rationality or incentive compatibility constraints that often do not hold in practice [44, 19]. Methods that do not have these constraints such as Deep Neural Auction and Reinforcement learning approaches, often need a extensive amount of high dimensional and varied data in order to converge to an reasonable performance, also often coupled with resource intensive training procedures [97, 49, 10]. In our specific case, we do not have the amount of repeated historical auctions, that can be sourced in the often applied to scenario of online ad auctions, but rather a great range of very diverse auctions scenarios with different bidding behaviour as well as bidder dynamics that only repeat every year or quarter. Given the importance for interpretability in the sensitive and impactful design of real industrial auctions, the in comparison sparse data available, and the often varying auction scenarios as well as input features, none of these approaches seem to fit the requirements fully. There seems to be a gap in the research on more practical models with high interpretability, while at the same time having low data quantity demands and still being able to handle the dynamic nature of auctions in practice.

3.3 Large Language Models: Knowledge Retrieval, Recommendation, and Reasoning

3.3.1 Large Language Models and Transformers

Large Language Models (LLMs), such as GPT (Generative Pre-trained Transformer) [9] and BERT (Bidirectional Encoder Representations from Transformers) [18], present a no-

TABLE 3.3: Comparison of Strengths and Weakness of Machine Learning Models for Auction Design

Model	Flexibility	Data Dependency	Computational Efficiency	Interpretability	Constraints
MenuNet [89]	**Middle**; Single Buyer Scenario	**High**; Relies on the optimization procedure	**High**; Simple Architecture	**Low**; MLP Structure	**Middle**; IC, IR, Order, Size
RegretNet [19]	**Middle**; Multi-Item, Multi-Buyer Scenario	**High**; Uniform sampling from Bids	**Middle**; Complicated Optimization	**Low**; More complex structure and optimization	**High**; IC, IR, Regret, Order, Size
PreferenceNet [73] / BudgetNet [25]	**Middle**; Includes bid preferences / Budget Constraints	**High**; Detailed Data on bidder preferences and bidding history	**Middle**; Increased complexity with preference layers and budget constraints	**Low**; Added three layer MLP	**High**; IC, IR, Order, Size
Deep Neural Auction [49]	**High**; Variable-sized, unordered inputs	**High**; Diverse set of feature-rich data	**Low**; Multiple Networks to optimize with large datasets	**Middle**; Rank Scores and Sorting Engine provide insight into Ranking	**Low**; Permutation invariant, auction type invariant
RL [10]	**High**; Dynamically changing environments	**High**; Many runs/historical bids training the valuation network	**Low**; Training resource-intensive and convergence difficult in large state space	**Low**; Deep RL model valuations and decision making unclear	**Low**; State space and convergence constraints

table progression in the field of natural language processing (NLP) and machine learning in general. These models have a vast number of parameters, and undergo training on extensive corpora of text data, giving them the capacity to generate, comprehend, and manipulate text that resembles human-like proficiency. They are able to have high proficiency in a range of NLP tasks, including but not limited to text completion, translation, question-answering, and summarization. Primarily based on the transformer architecture [94], it facilitates the handling of dependencies in input data: this enables the effective capture of contextual information by the model. LLMs have showcased the ability to understand syntax, semantics, and common knowledge, generating coherent text that remains contextually relevant. LLMs have exceeded initial expectations and are also able to perform complex tasks such as coding, university exams, or identifying patterns in data, demonstrating their versatility and advanced reasoning capabilities across languages [69].

3.3.2 Retrieval Augmented Generation

Through extensive training on large text corpora, LLMs acquire a large range of different knowledge. OpenAI's GPT 3 was trained on 43 terabytes of compressed text data sourced from the internet, with its newer model GPT 3.5 expected to be trained on even more data [1]. However, the knowledge encoded in the parameters of the model remains static and restricted to the period up until the last training update. The well-known GPT 3.5 Model of OpenAI has a diverse range of knowledge about many topics and events only until September 2021, limiting its knowledge to a certain timeframe.

There exist techniques such as Retrieval Augmented Generation (RAG) [46] that supply such LLMs with new external knowledge through the model's input processing, augmenting the statically encoded knowledge [58]. These techniques have been shown to enhance the model output, enabling it to remain factually accurate and contextually as well as time-relevant, boosting its reasoning capabilities compared to models without RAG [58]. It helps mitigate a common limitation of LLM models called Hallucination, in which the model produces outputs that are factually incorrect or logically incoherent.

In general, most RAG systems follow the same pattern: The user input prompt to a large language model is first used to query a database and extract relevant documents. Then the resulting query search result is provided as an added context to the LLM, providing it with information that is relevant to the prompt. Figure 3.8 provides an example of a retrieval augmented generation architecture. There are different decisions that can be made in the design of the architecture. It is for example possible to query based on precise keyword matching or through computing embeddings and finding semantic similarity with an existing database. Simple keyword matching might work well with simple exact phrase retrieval, but the data needed to answer a more complex question might not have similar syntactic similarity but rather need semantic searching to perform well.

A couple of versions of this technique with the common idea of extending the knowledge of LLMs through external non-parametric memory to augment its capabilities exist in the literature like RAG [46], RETRO [8], or REALM [29]. Fine-tuning or retraining LLMs is associated with a large overhead of computing power, time, and costs. Specifically RAG with a focus on external data retrieval only during the inference face and not during the pertaining procedure relieves the burden of needing to fine-tune the model to make it effective in solving domain-specific tasks as we set out to solve.

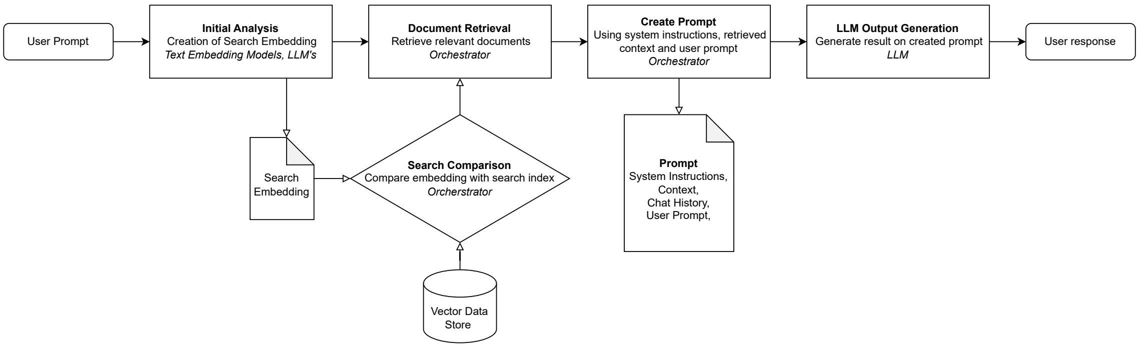


Figure 3.8: Retrieval Augmented Generation (RAG) Architecture

3.3.3 Covariance and Logic-Based Reasoning

Apart from the ability of LLMs to access and retrieve external knowledge through RAG, its reasoning capabilities are suitable to enable the recommendation of auction mechanisms based on game theoretical criteria, statistically observed relationships, and logical reasoning. Reasoning is the ability to make inferences using evidence and logic. Reasoning often happens in multiple steps in so-called inference chains, which involve deductions from multiple steps. Complex reasoning problems can be defined by the fact that they need to be divided into multiple sub-problems and these sub-problems need to be brought back together.

To understand how LLMs can reason, it is important to make a distinction in the type of causality [39, 34]. So-called covariance-based causal analysis is mainly used in fields like statistics and econometrics by using statistical approaches to discover and estimate causal relationships from data. For example, evaluating drug efficacy, optimizing business decisions, or analyzing the impact of auction design parameters on the buyer surplus are all covariance-based causality. On the other hand, we have logic-based causality [39] which aims to reason about casual relationships in systems by utilizing logic and domain knowledge. For example, in law, the notion of legal liability includes the creation of estimated causes for an effect based on counterfactuals and plausible scenarios. Counterfactual is a what-if analysis in a mental simulation of a scenario, yielding alternative possibilities for past or future events.

Another dimension of causality is the distinction between type (general) and actual (specific) causality. Type causality focuses on the relationship between variables and the impact that one variable can have on another variable. Actual causality focuses on the degree to which specific events cause other events. While type (general) causality focuses on variables and their average effects, actual causality also known as specific causality focuses on specific events and their causes.

The technique of prompting an LLM is one of the most computationally efficient, simple, and effective ways to elicit more advanced reasoning capabilities from LLMs Prompting can be broadly classified in two ways: Zero-shot and few-shot prompting [39, 34]. In few-shot prompting, a few examples of a task's input and output expectations are prompted along with the actual question prompt, whereas in zero-shot prompting, the prompt is compromised out of only the actual question prompt, without any prior examples. Few-shot prompting in-context learning or few-shot learning has been shown to enhance the accuracy on several tasks and benchmarks [39, 97, 34, 1, 69].

Auction design is a process that can contain both, covariance based causality as well as logical causality. Currently, most practitioners are using logical causality to develop auction design scenarios. For example, if we have invited 3 suppliers into an auction for bolts, and the initial bid spreads are high, a logical causality could be:

> A wide bid spread can be a sign of asymmetries between bidders. If there is an asymmetry between bidders, there could be one dominant bidder. If there is one dominant bidder a Dutch ticker auction could challenge the dominant bidder as the position of direct competition is unknown, and time pressure is applying. Therefore a Dutch auction would be the best auction scenario since the initial bid spread is high and we have a small amount of bidders.

If data from the created knowledge base is to be used to understand a general or average effect of a specific auction design on the buyer surplus, we would need covariance-based causal reasoning. For example, given a history of auction designs and their buyer surplus, the model could reason that based on the past auctions on bolts, the application of an two-phase auction has yielded higher buyer surplus compared to only one phase auctions when this set of suppliers was invited. In an more advanced scenario, logic and covariance based reasoning could be combined to create counterfactual scenarios that are started on the logic based reasoning but also tested or verified by covariance based reasoning, founded upon already existing coefficients of (linear)models that predict the impact on the buyer surplus. More common in current research on Large Language Models is the focus on causal tasks of effect inference, explanation, and decision-making on the general causality level with logic-based general causality, rather than the covariance based specific causality [34].

3.4 Research and Solutions Gap: Missing Flexibility, Logic-Based Reasoning, and Trust

Inconsistent Conclusions and Missing Standardized Approach

The analysed research papers based on the literature review conducted in Appendix A and detailed in Section 3.1.4 and Section 3.1.3 highlighted the sometimes contradictory conclusions on the relation of auction design parameters and the buyer surplus, influencing the recommendation given to practitioners. As shown by the wide range of auction scenarios studied from different papers, recommendations made are difficult to generalize due to the widely situational negotiation scenario in which the auction is placed and its dynamic external parameters [44]. Although there exist some common ground on specific parameters of the auction and its expected impact, and decision making models that try to close the research practitioner gap, the known risk of the impact of sub-optimally designed auctions and the individuality of the auction negotiation scenario, always warrants researchers to highlight that in practice, every auction scenario should be considered carefully and individually. Therefore, even though there already exists statistical analysis on auction design parameters, conducting statistical analysis for a case specific online auction history and comparing them to the existing hypothesis is necessary in order to gather trustworthy proposals and relations.

Adaptable, Interpretable and Practice Oriented Models

Machine Learning approaches such as RegretNet [19] or MenueNet [89] have a low flexibility in the auction designs/scenarios they can represent, while being rather effective in terms of demands on data and computational requirements. The models that inhibit high flexibility desirable for real life application like the Deep Neural Auction [49] or Reinforcement Learning [10] approaches have high data and computational requirements and are complicated to train to converge towards reasonable performance. All of the models share a lack of interpretability that is essential in the eyes of practitioners for the recommendation of auction designs to which possibly large monetary gains or losses are associated with [85]. Hence current research misses a model that inhibits a high interpretability, high flexibility and low data and computational demands. There is a gap in developing models that are directly applicable to industrial auction settings that are not repeated often like online ad auctions and account the more diverse and dynamic nature of industrial procurement auctions.

LLM Reasoning on Game Theoretical Tasks

Large Language Models (LLMs) show potential in various domains, but their application in auction theory and mechanism design remains unexplored [97, 34]. There is a gap in leveraging LLMs for advanced reasoning and decision-making based on game-theoretical principles for auction design. Combining LLMs with the field of empirical auction mechanism design could provide new insights and solutions, especially in complex reasoning tasks and interpretation of auction dynamics. This integration can fill the current gap in understanding and applying game-theoretical reasoning in practical auction scenarios.

3.5 Reflection on Literature and Solution Exploration

Concluding, in the field of automated auction mechanism design through deep learning, there is a noticeable gap in solutions characterized by deficiencies in flexibility, interpretability, and practical applicability. The current methodologies especially lack a comprehensible understanding of the recommended auction design and its underlying reasoning, a critical aspect of the high-stakes use case at hand. The missing possibility to extract logic-based reasoning along with the numerical optimizations provided hinders the translation of theoretical research into real-world purchasing practice. Current methodologies are characterized by a noticeable inflexibility, as they are often rigidly designed for specific scenarios and lack the ability to adapt or customize various auction types. This limitation hampers their applicability in a wide range of auction environments and contexts, thereby limiting their usefulness. Solutions with an emphasis on practical applications are also lacking as most works center around theoretical constructs, offering limited value to pragmatically solve the task in real-world scenarios. Combined with the not easily generalizable recommendations and statistical analysis of existing literature, the reasons for the research-practitioner gap can be observed. Despite the recent spread of LLMs in many different domains, their utilization in the field of auction theory and mechanism design presents an interesting research gap, especially considering the insights into advanced reasoning capabilities and flexibility of these models.

Chapter 4

Analysis and Evaluation

This chapter aims to explain the analysis procedure of the empirical data, as well as the design and evaluation of the auction chatbot. Section 4.1 gives an overview of the different research activities spanning from the initial data collection to the evaluation of the multi-agent system. Furthermore, it links the following sections of the chapter to the Flowchart of the overall research process. Following this, section 4.2 explains how the secondary data analysis of the auction type performance based on the empirical data is conducted. This section breaks down the process into three key components: data collection, operationalization of dimensions, and the statistical analysis. Section 4.3 shifts the focus to the development of the conversational chatbot, detailing the components of a multi-agent LLM system and the methodologies employed for retrieval augmented generation and prompt engineering. Lastly, section 4.4 presents the evaluation and experimental setup for the auction chatbot, including assessments of context retrieval, data retrieval, reasoning capabilities, and the overall experimental framework.

4.1 Process Overview

Figure 4.1 provides an overview of some detailed activities and process steps regarding the analysis of the research sub-questions (SRQ) 2, 3, and 4 which include the following high level phases:

1. Data Collection & Statistical Analysis
2. Model Development
3. Model Evaluation

The flowchart in Figure 4.1 is split up into 4 main boxes that represent the high level phases of collecting the tabular data for the statistical analysis and conducting it (see Section 4.2), the development process of the multi-agent retrieval augmented model (see Section 4.3), and the creation of an evaluation set as well as the evaluation process itself (see Section 4.4). White rectangle boxes represent activities, while red parallelograms define tabular data, green documents represent textual results and yellow documents are the research results in tabular and textual form marked with the corresponding research sub-question.

The data cleaning process in the upper left corner, represents a standard approach to data collection and cleaning [33], involving the special focus on the three data quality attributes of data relevancy to the research questions and formulated hypothesis, the data accuracy to the settings and results of the conducted auctions, and the reliability of the data stored in the system. The statistical analysis describes the operationalization of decision dimensions based upon the theory and empirical data available, the grouping of the auctions in their respective scenarios according to the operationalization of the three dimension decisions, the iterative validation of auction type diversity and size in their respective groups, and the statistical analysis with its check on assumptions. This process is repeated for each auction recommendation model as discussed in Section 3.1.3.

The model development is driven by an iterative design approach and Figure 4.1 highlight the distinction in the design of the multi-agent system architecture and the single retrieval augmented agents. The multi-agent system architecture process is defined by the specification of the general agents and their roles, their communication architecture and the selection and set up of prompt engineering techniques (see Section 4.3.1).
The model evaluation process is preceded by the creation of the evaluation dataset, found in the right hand corner. The process consists of mainly the brainstorming of relevant question categories and creation of ground truth labels to these questions (see Section 4.4). The evaluation procedure itself then consists of the initialization of a Judge LLM and the gathering of the systems responses to the evaluation dataset, including the evaluation of these answers by the Judge LLM. The computed evaluation metrics then represent the analyses to SRQ2 & 4 for each experimental setting proposed in Section 4.4.

4.2 Secondary Data Analysis of Auction Type Performance

4.2.1 Data Collection: Online e-Auction Platform

To analyze SRQ2, an empirical secondary data analysis on a large sample of 4731 conducted online reverse e-auctions of a large European automotive OEM from 2020 to 2023 is executed. The online reverse e-auctions are sourced from an online auction platform used by the purchasing department of the case company. The collected data entails the conduct of initial single-phase auctions, including the standard auction types of English auction, Dutch auction, and Japanese auction as defined previously. Before any online auction is conducted, suppliers are asked to submit an initial offer for the specified sourcing. Then the suppliers are invited to participate in the online auction, and the final bids are recorded. The main target variable is the savings achieved through the auction, and they are calculated as the percentage difference between the minimal initial offer and the minimal final offer achieved through the auction.

$$\text{Savings} = \frac{(\text{min initial bid - min final bid})}{\text{min initial bid}} \tag{4.1}$$

The auctions in the dataset deal with only one price position, and only single-phase auctions are subject to analysis. Out of the 4731 auctions in the dataset, 3845 are English, 705 are Dutch, and 186 are Japanese, as defined in the Theory section. The awarding commitment of the buyer is not necessarily given to the supplier with the smallest or winning price but is decided by a sourcing committee at a later stage.

4.2.2 Operationalization of Dimensions: Estimating Decision Criteria from Empirical Data

In order to group the available data into categories as defined by the decision-making models discussed in the theory section, an operationalization of the three decision criteria

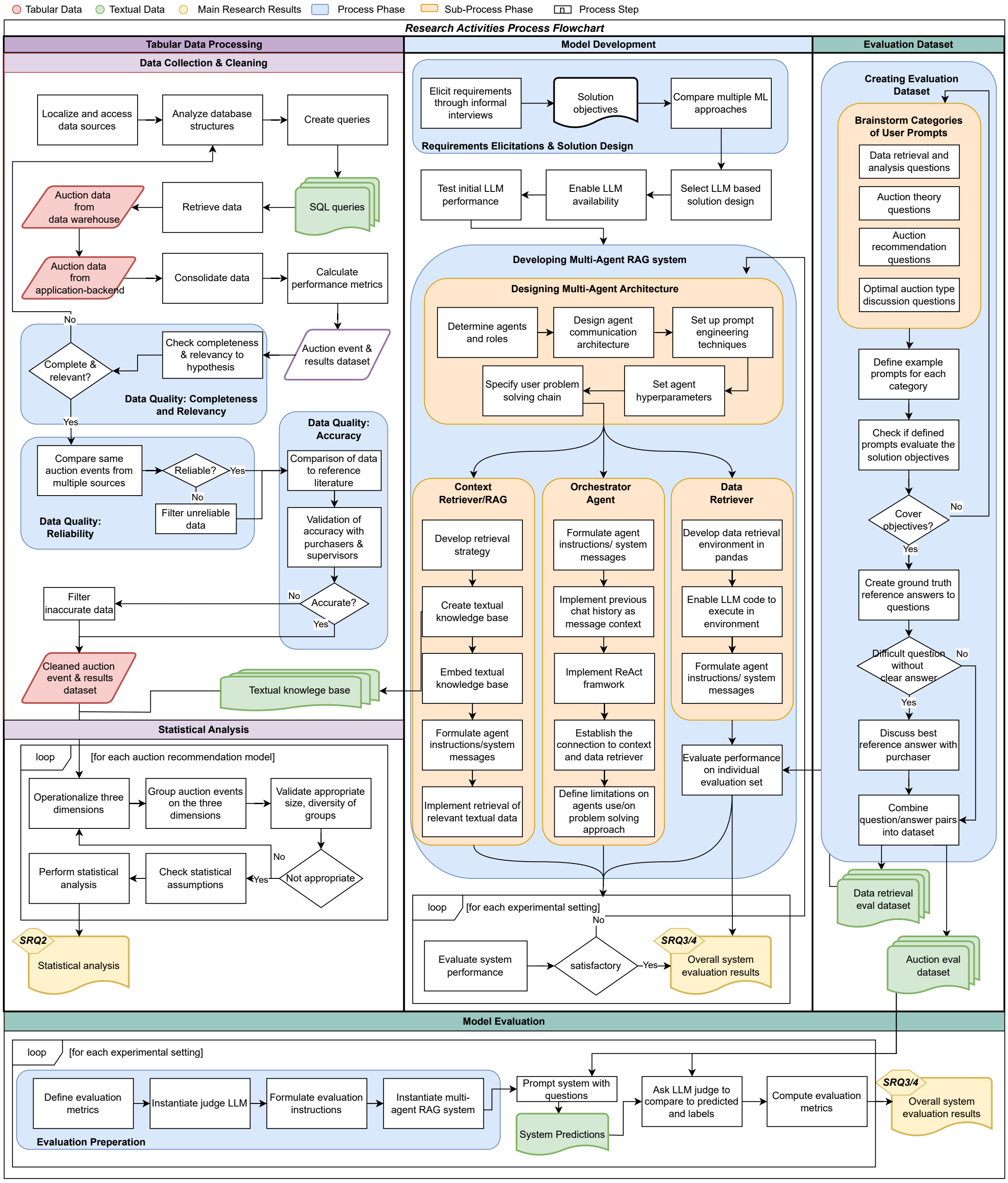

Tabular Data
Textual Data
Main Research Results
Process Phase
Sub-Process Phase
Process Step
Research Activities Process Flowchart
Tabular Data Processing
Data Collection & Cleaning
Localize and access data sources
Analyze database structures
Create queries
Auction data from data warehouse
Retrieve data
SQL queries
Auction data from application-backend
Consolidate data
Calculate performance metrics
No
Complete & relevant?
Check completeness & relevancy to hypothesis
Auction event & results dataset
Data Quality: Completeness and Relevancy
Yes
Data Quality: Accuracy
Compare same auction events from multiple sources
Reliable?
Yes
No
Filter unreliable data
Comparison of data to reference literature
Data Quality: Reliability
Validation of accuracy with purchasers & supervisors
Filter inaccurate data
No
Accurate?
Yes
Cleaned auction event & results dataset
Textual knowlege base
Statistical Analysis
loop
[for each auction recommendation model]
Operationalize three dimensions
Group auction events on the three dimensions
Validate appropriate size, diversity of groups
No
Perform statistical analysis
Check statistical assumptions
Yes
Not appropriate
SRQ2
Statistical analysis
Model Development
Elicit requirements through informal interviews
Solution objectives
Compare multiple ML approaches
Requirements Elicitations & Solution Design
Test initial LLM performance
Enable LLM availability
Select LLM based solution design
Developing Multi-Agent RAG system
Designing Multi-Agent Architecture
Determine agents and roles
Design agent communication architecture
Set up prompt engineering techniques
Specify user problem solving chain
Set agent hyperparameters
Context Retriever/RAG
Orchestrator Agent
Data Retriever
Develop retrieval strategy
Create textual knowledge base
Embed textual knowledge base
Formulate agent instructions/system messages
Implement retrieval of relevant textual data
Formulate agent instructions/ system messages
Implement previous chat history as message context
Implement ReAct framwork
Establish the connection to context and data retriever
Define limitations on agents use/on problem solving approach
Develop data retrieval environment in pandas
Enable LLM code to execute in environment
Formulate agent instructions/ system messages
Evaluate performance on individual evaluation set
loop
[for each experimental setting]
No
Evaluate system performance
satisfactory
Yes
SRQ3/4
Overall system evaluation results
Evaluation Dataset
Creating Evaluation Dataset
Brainstorm Categories of User Prompts
Data retrieval and analysis questions
Auction theory questions
Auction recommendation questions
Optimal auction type discussion questions
Define example prompts for each category
Check if defined prompts evaluate the solution objectives
Cover objectives?
No
Yes
Create ground truth reference answers to questions
Difficult question without clear answer
No
Yes
Discuss best reference answer with purchaser
Combine question/answer pairs into dataset
Data retrieval eval dataset
Auction eval dataset
Model Evaluation
loop
[for each experimental setting]
Define evaluation metrics
Instantiate judge LLM
Formulate evaluation instructions
Instantiate multi-agent RAG system
Evaluation Preperation
Prompt system with questions
System Predictions
Ask LLM judge to compare to predicted and labels
Compute evaluation metrics
SRQ3/4
Overall system evaluation results

Figure 4.1: Research activity flowchart that describes the data processing, statistical analysis, model development, and model evaluation phases and some of their related steps and outputs

is necessary. While the number of bidders is a straight-forward computation, the initial bid spread and especially the risk aversion or attractiveness of the business to the supplier require more complex considerations.

Number of bidders

The number of bidders is the number of bidders that have accepted and are participating in the auction, which does not coincide necessarily with the number of active participants in English auctions. The threshold defined by Berz et al. (2021) and Schulze-Horn et al. (2018) for the 'high' number of bidders classification of more than three bidders is adopted for the analysis. For any auction, at least two bidders need to be participating.

Initial bid spread

The initial bid spread is directly specified by Schulze-Horn et al. (2018) as the difference between the best bid and the second best bid as a percentage (Equation 2), and a threshold of 3% is given to categorize a 'low' initial bid spread. On the other hand, Berz et al. (2021) consider the initial bid spread not only on the basis of the two best bidders and their differences, but more generally through the sample standard deviation of all initial bids. Instead of directly taking the absolute sample standard deviation as a possible operationalization, the coefficient of variation is used as a measure to ensure comparability between different auctions (Equation 3). But no specific threshold for the two levels is defined by the literature. Therefore, the threshold is defined by quantile binning. In quantile binning, the groups are divided into 'low' and 'high' based on the median value for the initial bid spread to ensure that both groups have the same number of observations.

$$\text{Spread (Schulze-Horn)} = \frac{\text{(second lowest initial bid - min initial bid)}}{\text{second lowest initial bid}} \tag{4.2}$$

$$\text{Spread(Berz)} = \text{Coefficient of Variation} = \frac{\sigma}{\mu} \tag{4.3}$$

with $\mu = \frac{\sum(x_i)}{n}$ and $\sigma = \sqrt{\frac{\sum(x_i-\mu)^2}{n}}$, given that we observe a set of initial bids X with size n and x_i being the ith initial bid.

Strategic Relevance for Suppliers

The strategic relevance or attractiveness of the business for suppliers is considered to be a complex function that needs to be evaluated by the purchaser on the basis of market knowledge, relationship insights, and the competitive situation [84]. Berz et al. [6] use the strategic margin 'M' of the suppliers in their theoretical model, which are defined as the difference between the indifference prices and the submitted price in a first-price auction. Computing the strategic margin 'M' in practice would only be available after a thorough cost analysis and represent an approximation at best. Then the threshold G acts as a reference point upon which the decision on a first-price or second-price auction is made. Given that no information about the possible margins is available for each auction in our dataset, the strategic relevance of the business in this analysis is approximated by the purchasing volume. The hypothesis is that a larger purchasing volume auction will, on average, be more strategically relevant to most participating suppliers than a lower purchasing volume auction. Again, since no specific thresholds are given for the

Variable	Mode	Count	Min	Max	Mean	Median	Std
Number of Bidders	Dutch	705	2	9	2.77	2	1.21
	Japanese	186	2	11	3.08	3	1.38
	English	3845	2	26	6.62	6	3.59
	All	4731	2	26	5.91	5	3.60
Initial Spread Best & Second (%)	Dutch	705	0.00	41.55	8.17	5.17	8.37
	Japanese	186	0.00	40.49	6.91	4.06	8.22
	English	3845	0.00	41.95	11.62	8.71	9.84
	All	4731	0	41.95	10.92	7.84	9.69
Coefficient of Variation	Dutch	705	0.00	0.22	0.04	0.03	0.05
	Japanese	186	0.00	0.21	0.04	0.03	0.05
	English	3845	0.00	0.26	0.09	0.09	0.04
	All	4731	0.00	0.26	0.08	0.08	0.05
Initial Minimum Bid & Auction Volume	Dutch	705	112	261,523,019	3,624,878	1,137,400	12,371,909
	Japanese	186	141	137,500,000	3,327,854	529,650	13,140,358
	English	3845	806	434,480,850	12,698,297	1,845,827	36,127,814
	All	4731	112	434,480,850	10,979,618	1,573,115	33,194,398
Savings (%)	Dutch	705	0.00	17.40	3.70	3.09	3.07
	Japanese	186	0.00	12.22	3.37	2.79	2.51
	English	3845	0.00	18.87	3.78	2.54	4.01
	All	4731	0.00	18.87	3.75	2.65	3.84

Table 4.1: Descriptive statistics per auction mode

categorization into levels, the quantile binning method based on the separation of the data on the median line is applied.

4.2.3 Statistical Analysis

Table 4.1 highlights descriptive statistics for the main variables grouped by the auction mode and for the whole dataset. On the request of the case company, the absolute values have been symmetrically changed, retaining the important data relations.

The English auction achieved the highest cost savings on average, closely followed by the Dutch and Japanese auctions. The median is almost always lower than the mean, indicating positive skewness in the data. The English auction seems to be favored in situations with a higher number of bidders, as the mean and median are substantially higher than for the Japanese auction and Dutch auction. The Japanese auction, in turn, tends to be used by a marginally higher number of bidders than the Dutch auction. With the spread between the best and second-best initial offer, the distribution for each auction mode seems to be comparable, with the mean of the Japanese being the lowest, followed by the Dutch and English auctions. Also, for the coefficient of variation, the mean seems to be much higher for the English auction than for the first-price auctions. The English auction is also used in scenarios that have a higher auction volume, with the mean and median being multiple times larger than those of the Dutch or Japanese auction. The Japanese auction is more often used in smaller auction volumes.

We test the hypotheses with an inferential statistical test. In preparation for the decision on the most suitable statistical test, which establishes whether a statistically significant difference between the population means of the distributions of savings for the auction types under the different predefined scenarios exists, three assumptions have to be checked on the data. First, the assumption of independent groups. As the observed auctions are only single-phase auctions, which do not influence the results in any of the other auctions or between groups, the assumption that the groups are independent is warranted. Second, there is the assumption of the normality of the data. To test the assumption of normality of the data distribution, a Shapiro-Wilk test (Shapiro & Wilk, 1965) is applied to the divided groups. Third is the common assumption of the homogeneity of variances. For this assumption, the Levenes test (Levene, 1960) is applied to the equal variances of two groups.

Figure 4.2 showcases the decision tree regarding choice of the appropriate statistical test based on the assumptions. In case either the normality of the two groups or the homogeneity of variances assumption fails, different statistical tests are necessary. If the sample size of the groups is small (n<15) [71] the recommendation would be to apply a non-parametric test, more specifically the Mann-Whitney U-test [52]. But due to the large sample size for each group ($n > 15$), although normality of the data is not given in most cases, the application of a parametric test is still possible based on the central limit theorem [83]. Depending on the homogeneity of the variances, a suitable test would be either the independent two-samples t-test, given the homogeneity of the variances, or Welch's t-test otherwise [71]. All tests are two-tailed tests conducted at a significance level of five percent ($\alpha = 0.05$).

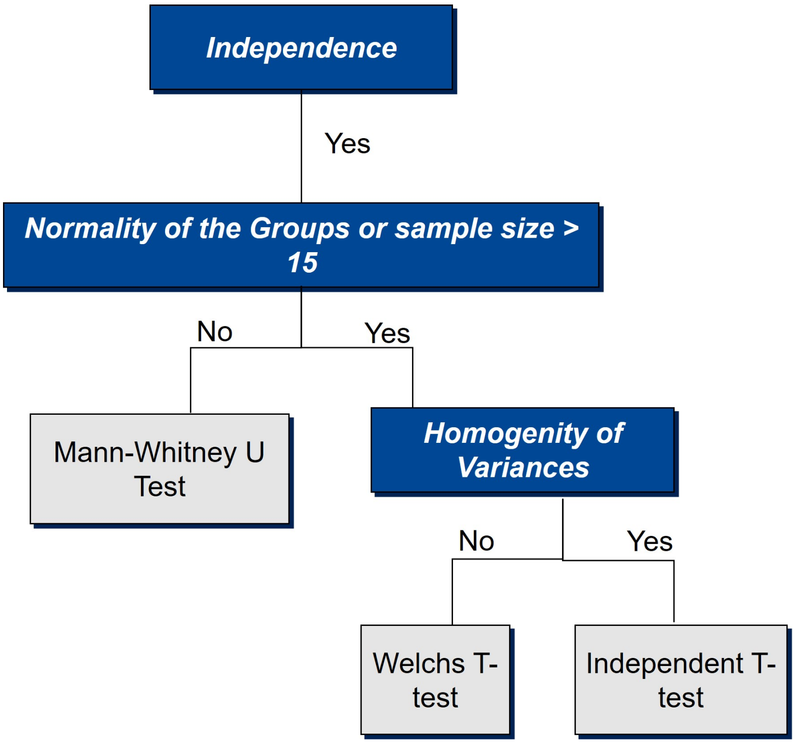
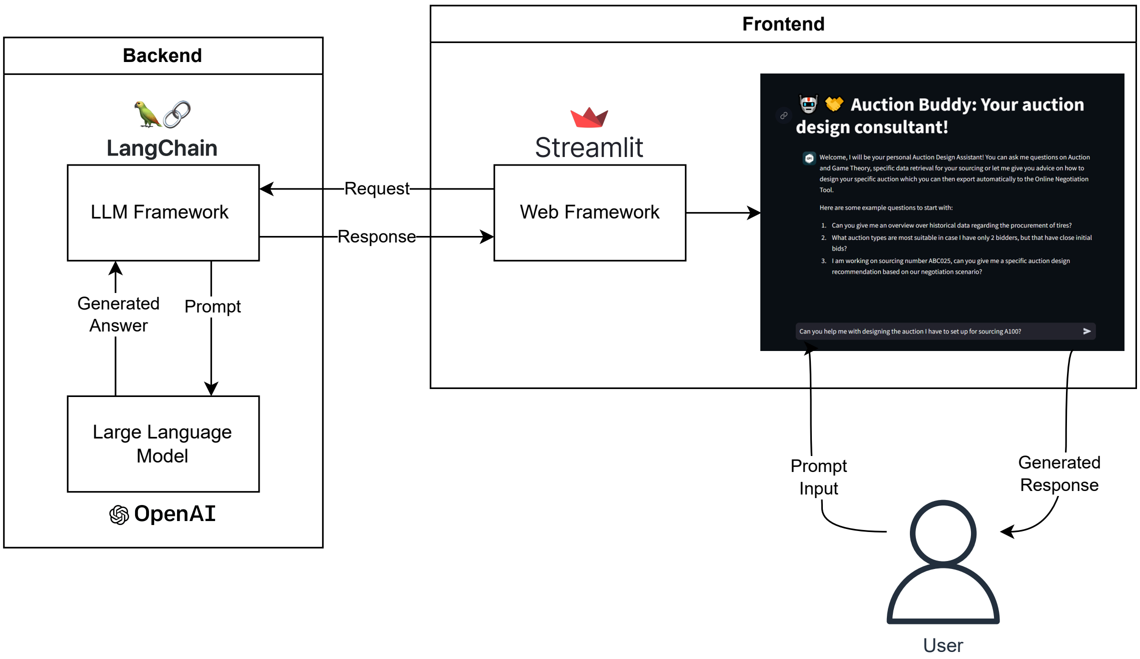


FIGURE 4.2: Decision Tree for the appropriate selection of statistical tests

4.3 Auction ChatBot Development Process

The system presented in the following sections is a multi-agent conversational LLM-powered chatbot with memory and step-by-step problem-solving capabilities. Based on the requirements outlined in section 2.4, the system was designed to facilitate the creation of auction designs for purchasers, utilizing knowledge of multi-modal data of online reverse auctions and auction design. An simple high level overview of the application is given in Figure 4.3 and example screenshots of the user interface and interactions with the system can be found in Appendix.



Auction Buddy: Your auction design consultant!

Welcome, I will be your personal Auction Design Assistant! You can ask me questions on Auction and Game Theory, specific data retrieval for your sourcing or let me give you advice on how to design your specific auction which you can then export automatically to the Online Negotiation Tool.

Here are some example questions to start with:

1. Can you give me an overview over historical data regarding the procurement of tires?
2. What auction types are most suitable in case I have only 2 bidders, but that have close initial bids?
3. I am working on sourcing number ABC025, can you give me a specific auction design recommendation based on our negotiation scenario?

Can you help me with designing the auction I have to set up for sourcing A100?

FIGURE 4.3: High Level Overview of the application structure

The chatbot application is powered by two main Frameworks in the Python programming language: LangChain and Streamlit. Streamlit is a Web Framwork with which it is possible to create a web application and user interface for an User to interact with. The LangChain library is essential in establishing a framework for the LLM powered multi-agent system behind the chatbot. The Large Language Models are either run locally on a computer or exposed via an API to the application. The application has access to specific and relevant textual and numerical data about online e-auctions. On the one hand, structured tabular data about the historical auction design (see Table 2.1) and results are available. On the other hand, internal guidelines and possible recommendations on the auction design are available as textual data. Additionally, from the result of research question 1, textual data that sets the relationship between different variables and best practice recommendations based on the preceding statistical analysis of the data can be incorporated.

The following sections dive deeper into the mechanics and concepts in the Backend of the application. Section 4.3.1 explains the multi-agent framework in place that includes the orchestrator, context retrieval, and data retrieval agents and their collaboration architecture. Section 4.3.2 and section 4.3.3 explain the context retrieval and data retrieval agent mechanisms in more technical detail respectively. The concluding section 4.3.4 highlights some prompt engineering techniques deployed to enhance the reasoning and error handling of the whole system.

4.3.1 Multi-Agent LLM System: Orchestrator, Retriever, and Calculator

Building upon the cognitive abilities demonstrated by single-agent LLM systems [27], the complexities of combining textual and numerical knowledge necessitate a more intricate architecture. The proposed multi-agent framework tackles the challenge of not only handling textual information retrieval but also incorporating numerical data processing seamlessly into the general Retrieval-Augmented Generation (RAG) pipeline. The core idea of the architecture is the centralized communication structure, inspired by the improved tackling of complex tasks including the planning and solving of subgoals [27]. This centralized approach is detailed in Figure 4.4.

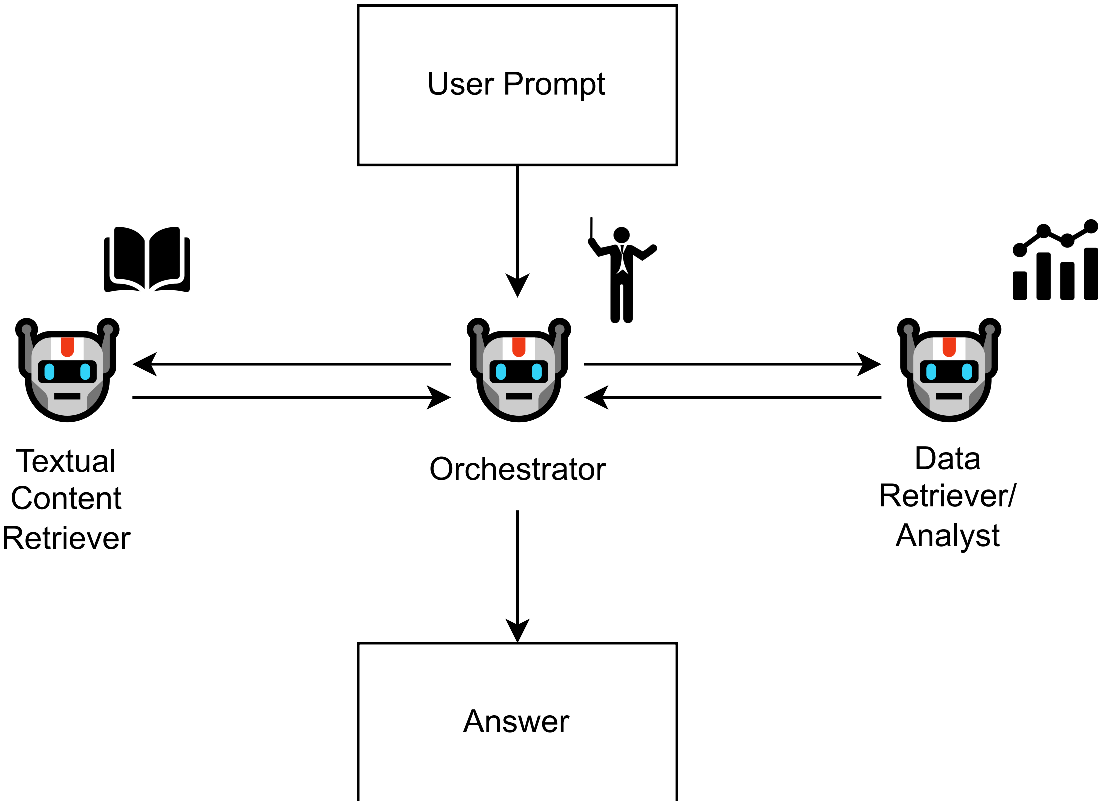


FIGURE 4.4: The agent communication architecture: Centralized and Cooperative

The **Orchestrator** is responsible for the planning, aggregation and answering of the user input prompt. The orchestrator receives the user prompt and plans how to reach an answer to this prompt. This plan includes the choice of agents to make and the orchestrator can choose whether an agent is needed to answer this question, and which agent is most suitable. He can invoke the agents individually up to 3 times in total each and is not restricted in the order of which agent to call. Therefore, for any given prompt the agent may call no agents, only the textual context retriever, only the data retriever, or both, even multiple times after each other. The orchestrator creates the prompts for the agents based on the assessed information needed to answer the question and evaluates the received answers from the agent. After receiving the required context from the agents, the orchestrator generates the final answer to the user prompt. Figure 4.5 showcases a simple example of the information and execution flow between the three actors and the user.

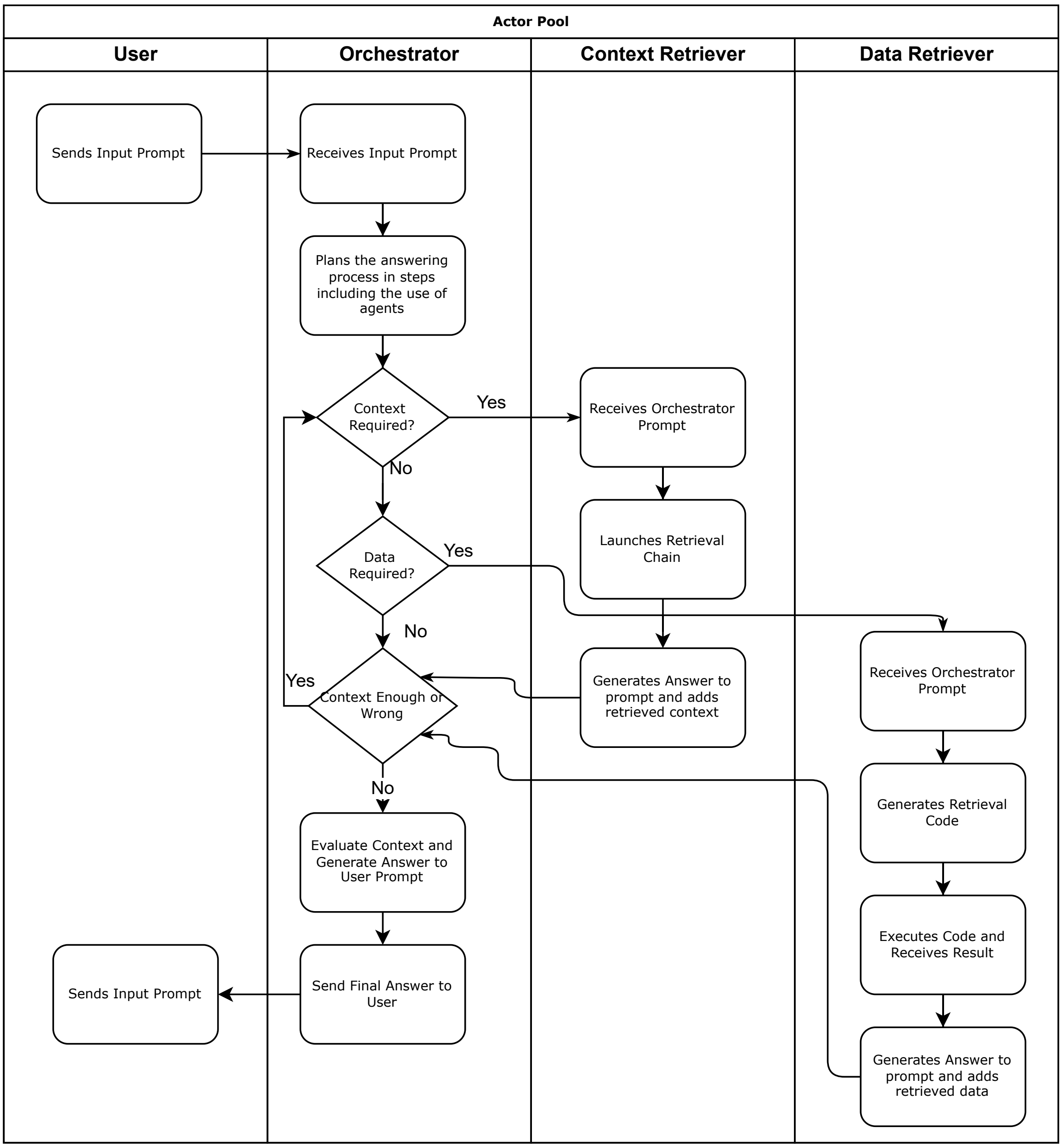

Actor Pool
User
Orchestrator
Context Retriever
Data Retriever
Sends Input Prompt
Receives Input Prompt
Plans the answering process in steps including the use of agents
Context Required?
Yes
No
Receives Orchestrator Prompt
Launches Retrieval Chain
Data Required?
Yes
No
Generates Answer to prompt and adds retrieved context
Receives Orchestrator Prompt
Yes
Context Enough or Wrong
No
Generates Retrieval Code
Evaluate Context and Generate Answer to User Prompt
Executes Code and Receives Result
Sends Input Prompt
Send Final Answer to User
Generates Answer to prompt and adds retrieved data

Figure 4.5: Flow diagram of the agent communication and executions

The **Context Retriever** specializes in answering prompts based on context from text corpora supplied to him via text documents. This agent can retrieve relevant text chunks from the text corpora and incorporate them into the creation of an answer to the prompt from the orchestrator. More detailed explanations of the technical implementation are in section 4.3.2.

The **data retriever/analyst** can retrieve data from numerical data saved in tables via access to a Python environment and the panda's library. The agent can not only retrieve data but also aggregate and compute metrics based on the input prompt. The main responsibility of the data retriever is to retrieve data and computations from the relevant table and return an answer to the prompt of the orchestrator. More technical descriptions of the implementation are detailed in section 4.3.3.

The communication structure leverages a centralized architecture with a cooperative communication paradigm [27]. The Orchestrator agent acts as the central node, facilitating communication between itself and the two specialized agents: the Textual Context Retriever and the Data Analyst. This centralized approach ensures coordinated information flow and enables the Orchestrator to address potential errors or inconsistencies in the answers received by the agents. To deliver the most accurate and informative response to the user's query, the agents work together to achieve a common goal by sharing information and executing specific actions appropriate to their specialization. Although this cooperative communication paradigm is chosen, the system still incorporates a feedback loop within this cooperative structure for the Orchestrator. The Orchestrator assesses the retrieved information from the other agents and strategically reuses or refines prompts based on this evaluation. For instance, if the Textual Context Retriever retrieves irrelevant information or the LLM Calculator encounters unexpected or irrelevant numerical formats, the Orchestrator can adjust its prompts or request clarifications to steer the information retrieval and response generation processes towards a more accurate outcome.

4.3.2 Retrieval Augmented Generation: Context Retriever

The retrieval augmented generation pipeline for the context retriever is split into two essential phases: Indexing and Retrieval. Before runtime, the text documents need to go through the indexing process, highlighted in Figure 4.6. The indexing stage is the foundation for efficient retrieval within the RAG framework and it transforms raw text data into an efficiently and semantically searchable format. This process typically occurs offline, allowing for storage of the information before real-time interaction with user queries.

First, the text data stored in pdf or Word documents are loaded as strings. Given the potentially large size of the documents, splitting them into smaller, manageable chunks becomes important for efficient retrieval. Here the chunk strategy, comprising of the method of text splitting and the chunk size, can be chosen in different ways. The chosen recursive text splitting method, splits the text based on paragraphs and other newline characters, as they are assumed to form semantic relevant groups. It recursively splits those chunks until the specified chunk size is reached and the chunks are small enough. The trade-off in selecting the size of the chunks occurs in the retrieval efficiency and the retrieval accuracy. Larger chunks facilitate a faster retrieval process but have a smaller retrieval accuracy, while smaller chunks have a higher retrieval accuracy but a slower retrieval process. Also, the finite context window of LLMs limits the possible chunk size.

Following text splitting, the processed chunks require an embedding for an effective retrieval process. Embedding models transform the textual splits into a more suitable representation for indexing. This transformation involves mapping the textual data into a high-dimensional vector space, where similar items are positioned close together. This facilitates the retrieval process by allowing the system to identify splits containing information relevant to the user's query based on their proximity within the vector space. The all-MiniLM-L6-v2 embedding model [78] from Huggingface embeds the text chunks into vector space, and then stores them in a vector database.

The retrieval process starts with the orchestrator prompt that is based on the ini-

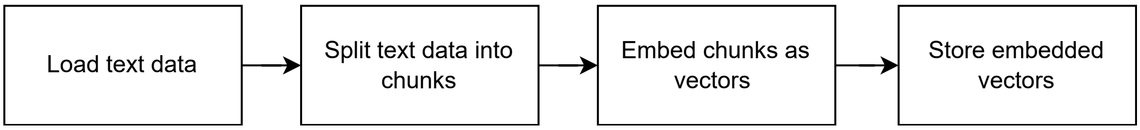


FIGURE 4.6: The indexing procedure

tial user prompt. The retriever agent utilizes the previously constructed index via vector storage to identify the most relevant splits from the stored data. The query is embeded in the same way as the vector index and the retrieval process leverages the vector representations of both the query and the indexed splits. By calculating the similarity between these vectors, the Retriever can efficiently identify splits containing information that closely matches the query's semantic meaning. This similarity calculation often involves techniques like cosine similarity, which measures the angle between the query and split vectors in the high-dimensional space. For this specific implementation, the Maximal Marginal Relevance calculation [53] ranks the vector chunks based on their relevance with higher similarity scores deemed more relevant to the user's query and are prioritized for retrieval.

The Retriever selects a predefined number of the most relevant splits. This selection process ensures that the LLM in the generation stage has access to relevant context while considering the limitations of context windows. The number of retrieved splits again represents a trade-off between comprehensiveness and computational efficiency. A larger number of retrieved splits provides the LLM with a more comprehensive understanding but requires more processing power during the generation stage.

4.3.3 Retrieval Augmented Generation: Data Retriever

Mirroring the textual context retriever is the data retriever that focuses on retrieving relevant data from structured sources such as CSV files. The agent is composed of an LLM-powered agent who is triggered by the invocation through the orchestrator. Specific to this agent is the ability to execute Python code in the environment and being able to access a directory, in which the desired structured context data is stored. The process of the data retriever starts with the prompt of the orchestrator which is interpreted and transformed into code by the underlying LLM agent based on a set of prompt instructions. In the first run, the LLM agent loads all available CSV files in the directory to which the user uploads the required files into. After that, the generated code is executed in the Python environment and can run code for the popular Pandas Dataframe library specifically. This enables the agent to use all the capabilities of the Pandas library, including the following important functionalities:

- **Simple Data Retrieval:** The agent can execute queries instructed by the orchestrator. These queries typically specify the desired information to be retrieved from the DataFrame. The queries might involve filtering criteria based on specific data points (e.g., "Give me an overview over all conducted auction that are available in the uploaded file for the item with itemnumber: ITEM00001") or more complex conditional statements encompassing multiple attributes.

- **Data Aggregation and Calculation:** The agent can compute metrics like the average savings over all previous auctions for specific items or give descriptive statistics about distributions.

- **Metadata analysis:** The analysis of metadata of the table/dataframe is possible, such as returning the number of rows or columns or checking if there are any missing values in the dataset.

The functionality enables the orchestrator to combine general theoretical knowledge about auction design, with specific information about auctions and negotiation situations. The orchestrator is therefore able to engage in logic-based reasoning on the textual context and in covariance-based reasoning on the numerical data provided, as well as combine both of these sources to give specific recommendations on for example the optimal auction design.

4.3.4 Prompt Engineering & Reasoning: ReAct Framework

The ReAct Framework [101] is a prompt engineering approach designed to improve the reasoning and action-planning capabilities of LLMs. It is the framework that enables the LLMs underpinning the agents, especially the orchestrator agent, to require factual accuracy, reason about the user prompt, and interact with the two other agents. The ReAct framework encourages the LLM to explicitly state its reasoning process in a step-by-step fashion, combining the chain of thought prompting technique. It also specifies that the LLM takes specific actions from a set of predefined possible actions with descriptions of these actions and their usefulness in different situations. Additionally, an observation step follows the reasoning and acting steps to enhance the accuracy and handle possible errors that occur.

4.4 Auction Chat-bot Evaluation & Experiments

The evaluation of outputs from an LLM is an ongoing complex challenge, as academia works on creating established standardized metrics that encompass the varied nature of LLMs [34, 105]. Existing traditional evaluation metrics such as ROUGE [48] compute a measure of similarity between outputs and reference answers, but fall short at tasks that require advanced reasoning such as in auction design. Simmilarity-based metrics such as ROUGE do not suit for the evaluation of auction design recommendations and fact-checking the contents of the retrieved documents and data, because of the absence of a singular correct answer. The benchmarks commonly used in evaluating LLM often depend on large datasets or multiple-choice formats that may not correspond to the specialized reasoning needed in domain-specific applications [12, 63, 105]. Hence, a tailored evaluation approach is more suitable considering the specialized nature of the auction and mechanism design. Although human evaluation is considered as the benchmark for assessing the quality of LLM outputs, it is limited by the scalability and practicality of extensive experimentation and its associated cost and time investments. Recently, instead of humans, LLMs have been used to evaluate the outputs of RAG systems and other more complex LLM-powered systems [105]. Zheng et al. [105] found that strong LLM judges like GPT-4 can match the evaluation of human evaluators.

As the system is composed of different components, it is necessary to evaluate the components in the different phases of the processing pipeline: The context retrieval, the data retrieval, and the reasoning in the final answer.

4.4.1 Evaluation Context Retrieval

On the evaluation procedure of the context retrieval of textual documents, the RAGAS [24] Evaluation Framework is adopted. RAGAS is based upon the LLM-as-a-Judge principle [105] and therefore utilizes an evaluator LLM to score the accuracy measures proposed. RAGAS entails the assessment of the context relevance, the answer relevance, and the faithfulness to the source documents.

The context relevance metric represents the focus and containment of irrelevant information in the retrieved context. To compute that it extracts all sentences from the context $c(q)$ given a question/prompt q that are considered to be relevant sentences by the evaluator LLM based on the initial question. The subset of relevant sentences SR from the set of all Sentences SE that could answer the question from the provided context, and sets it about the total number of sentences in the context.

$$\text{Context Relevance} = \frac{\text{Number of extracted relevant sentences}}{\text{Total Number of sentences}} = \frac{|SR|}{|SE|} \tag{4.4}$$

The answer relevance metric is computed by calculating the cosine similarity of the original question q given by the user and a newly generated question q_i. The evaluation LLM generates this new question q_i based on the answer $as(q)$ given by the agent LLM to the original question q. The cosine similarity of the original question and the newly formed question based on the given answer is calculated as follows:

$$\text{Answer Relevance} = \frac{1}{n}\sum_{i=1}^{n} \text{sim}\left(q, q_i\right) \tag{4.5}$$

The faithfulness metric assesses how well the generated answer, $as(q)$, aligns with the factual information presented in the context, $c(q)$. Essentially, it measures whether the claims made in the answer can be logically derived from the context. An LLM is employed to extract a set of statements, $S(as(q))$, from the generated answer. For each statement s_i, the LLM evaluates whether it can be inferred from the context $c(q)$ or not. $|V|$ is the number of statements that were supported by the context and $|S|$ is the total number of statements. The Faithfulness score is then computed as:

$$\text{Faithfulness} = \frac{|V|}{|S|} \tag{4.6}$$

4.4.2 Evaluating Data Retrieval

Different from the evaluation of context retrieval, the evaluation procedure for the data retrieval is centered around a custom evaluation dataset of 20 questions and correct reference answer pairs, shown in Appendix D.1. The questions are formulated with the goal to test the ability of the data retrieval agent to do the following actions:

- Basic Data Retrieval
- Simple Calculations
- Handling of different data types (text, numerical, dates, lists of text)
- Join information from two tables

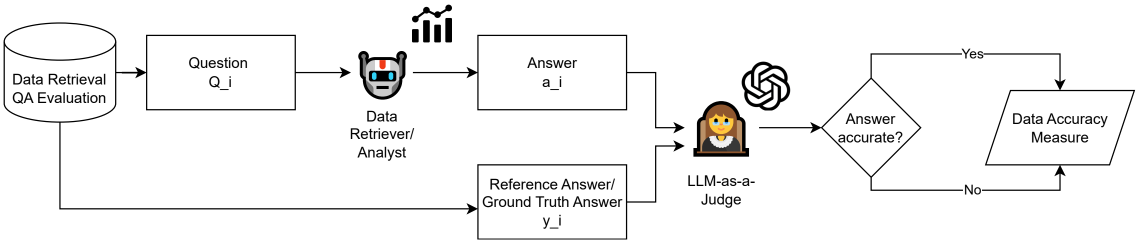


FIGURE 4.7: Evaluation Procedure for the Data Retrieval Agent

In order to perform the evaluation, two data tables containing auction data, on which the data retriever will perform the evaluation on, are uploaded to the system. Appendix D.1 shows an example of the data tables uploaded, one table representing the auction event information, and the other, information about the participating companies.

Figure 4.7 explains the simple evaluation procedure. The dataset consists of a set of 20 questions q_i, and corresponding reference or ground truth answers y_i. A question is given to the data retrieval agent and the response of the data retrieval agent $a_i = answer(q_i)$ is compared to the ground truth reference values by an evaluator LLM, following the LLM-as-a-Judge principle [105]. The Judge LLM compares both inputs and returns a of 0 or 1 for a wrong or correct answer respectively, as well as a reasoning on why this score was given. This is repeated over all question answer pairs in the dataset and the resulting scores and reasoning are saved. $|G|$ represents the number of correct answers and $|Q|$ the number of total questions. The accuracy for the data retrieval is called the Data Accuracy and is then computed as the fraction of correct answers as follows:

$$\text{Data Accuracy} = \frac{|G|}{|Q|} \tag{4.7}$$

4.4.3 Evaluating Reasoning: AuctionEval

The evaluation of the final answer of the whole system, especially related to the task of recommending the optimal auction design, given general theoretical recommendations about abstract scenarios as well as recommendations on specific scenarios is more intricate. Not only is there no clear and single correct answer to specifically more complex questions, but also the quality of the reasoning provided by the answer should be evaluated. Therefore, a custom evaluation dataset named AuctionEval is created, containing 30 question input prompts and ground truth reference answer pairs, as found in Appendix D.2.

The AuctionEval dataset (Table D.4) consist of 30 questions with varying degrees of difficulty ranging from simple general questions regarding theoretical recommendations up to specific questions aiming to test the covariance-based and high-level logic based reasoning (see Section 3.3.3) with expected counterfactual scenarios. The question answer pairs were created with the aim to cover a wide range of common use cases for purchasers utilizing the system and explore the capabilities of multi-agent retrieval systems to solve more complex reasoning tasks including the use of textual as well as numerical context. Some question do not have a clear single correct recommendation, and some also contain multiple possible correct reasoning traces. he creation of the reference or ground truth answers of the AuctionEval dataset was supported by the evaluation of two humans, one of which has 5+ years of professional experience as a purchaser, to ensure reliable reference

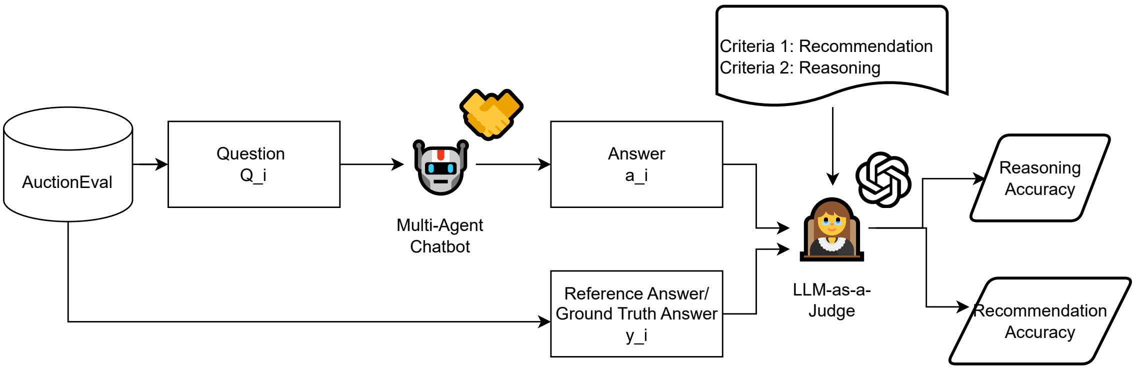


FIGURE 4.8: Evaluation Procedure for the whole multi-agents system

answers to the created questions. This lead to the decision to avoid the usage of a multiple choice based evaluation [77], but rather again use the LLM-as-a-Judge methodology [105] due to its flexibility.

The LLM-as-a-Judge [105] concept, as the current standard for evaluation procedures, serves as the basis for the evaluation of the recommendation and reasoning of the system. As shown in Figure 4.8, the evaluator LLM is given two criteria, the correct auction design recommendation and the reasoning process, on which to compare the system output with the ground truth reference answers. For each criterion, a description of the criterion and how it should be evaluated is given to the evaluator LLM. The LLM then provides a score of 0 or 1 and a reasoning for his score for each criterion for each question answer pair. In the special case of multiple possible reasoning traces, the evaluator LLM is specifically told to consider the comparison between both possible solutions.

Similar to the accuracy measures before, the percentage of correct answers over all answers in the dataset describes the performance of the system on the reasoning and correct recommendation criteria. Let $|R|$ be the number of correct reasoning outputs, $|CR|$ the number of correct recommendations and $|Q|$ the total number of questions in the dataset. Then the accuracy measures are defined as follows:

$$\text{Reasoning Accuracy:} = \frac{|R|}{|Q|} \tag{4.8}$$

$$\text{Recommendation Accuracy:} = \frac{|CR|}{|Q|} \tag{4.9}$$

4.4.4 Experimental Setup

The experimental setup aims to answer SRQ3 and 4 by evaluating the multi-agent system under different configurations of Large Language Model choice, available documents for the context retrieval and the application of prompt engineering techniques, more specifically chain of thought prompting. Table 4.2 summarizes the 12 configurations of the system based on the LLM in use for the agents, the supply of information to the system, and the use of prompt engineering techniques.

Experimental Setup		
LLM	**Document Set**	**Chain of Thought**
gpt-3.5-turbo	Small_R	No
gpt-3.5-turbo	Small_R	Yes
gpt-3.5-turbo	Large_RET	No
gpt-3.5-turbo	Large_RET	Yes
Nous-Hermes2	Small_R	No
Nous-Hermes2	Small_R	Yes
Nous-Hermes2	Large_RET	No
Nous-Hermes2	Large_RET	Yes
gpt4-turbo	Small_R	No
gpt4-turbo	Small_R	Yes
gpt4-turbo	Large_RET	No
gpt4-turbo	Large_RET	Yes

Table 4.2: Summary of the 12 configuration for the experimental setup

Each Agent is underpinned by an LLM. The choice of the LLM in use has therefore major implications regarding multiple parameters including the context window, the latency, and the output quality. As the focus lies on the evaluation of reasoning capabilities, three LLM are chosen that each represent a category of LLMs:

- Nous-Hermes2: A LLM model based on Llama2-13B and fine-tuned on research discussions and reasoning tasks. This is a model that can run locally on the laptop due to its relatively small size and the quantization of its parameters [34]. It represents the low parameter and locally runnable class of LLM models.

- GPT3.5-Turbo: Still one of the biggest, fastest and among the highest performing LLM model across benchmarks [12, 69, 3]. As one of the most commonly used models that also powers the popular ChatGPT application, this model represents the class of models with a high parameter count, which cannot be run locally, but are also not state-of-the-art anymore.

- GPT4-Turbo: The flagship model of OpenAI and leader in most benchmarks especially in reasoning and complex problem solving tasks [69]. The newest model represents the current state-of-the-art in LLM models that have a very high parameter count and cannot be run locally, while being expected to perform among the best across all available models.

To answer SRQ3, the supply of information with focus on the textual data needs to be experimented upon. The exploration of the trade-off between size of the text corpora and the relevance of the retrieved context and efficiency of the system is of major interest and concern in most RAG implementations [58, 34]. For that reason, different document sets with mainly size differences are defined that will be made available to the context retriever of the multi-agent system:

- Small_R: The small document set only includes the three research papers by Berz et al. [6], Eichstädt [20], and Schulze-Horn et al. [84] as well as the recommendations and analysis given by the results of RQ1. It focuses purely on the recommendation models that are highly relevant to the task of designing auctions, and therefore is abbreviated as Small_R.

- Large_RET: This bigger document set contains 12 research paper about the topic of auction design from the Literature Review outlined in Appendix A. They include the recommendation model documents from Small_R but also add documents about empirical research into auction design through regression models, or mathematical analysis from a wide range of industries like healthcare or public procurement. Therefore this document set presents a much wider view and a key part to explore the trade offs and differences between small and concentrated corpora against a larger corpora with more variety.

Lastly, the impact of prompt engineering techniques are to be explored for SRQ4, especially regarding its influence on the reasoning ability. To test the influence of the prompt engineering, the choice to use or not use chain of through prompting in the initial system messages as well as partly in the ReAct Framework and in the orchestrator prompts to the other agents will be one of the experimental variables. All in all, there are 12 configurations defined and summarized in Table 4.2, and these 12 system configuration will undergo the evaluation procedure as described in Section 4.4. The results of the evaluation can be found in the following results chapter, specifically in Section 5.2.

Chapter 5

Results

The following chapter presents the results of the statistical analysis and the experimental evaluations, structured based on the stipulated research question in chapter 1.3. section 5.1 details the statistical analysis of the three recommendations model discussed in chapter 3.1.3 with the approach described in chapter 4.2. Thereby, the auction type recommendations proposed by the models are compared to the empirical performance from the data, followed with an empirically optimized auction type recommendation model in subsection 5.1.4, and a computation of the expected savings achieved per recommendation model in subsection 5.1.5. This section concludes with the answers to SRQ2. section 5.2 shows the results of the performance evaluation of the multi-agent retrieval system as described in chapter 4, and concludes with the answers to SRQ3 and SRQ4.

5.1 SRQ2: Design of an Empirically Optimized Auction Type Recommendation Model

The following section includes a more complex visualization of the savings distribuitions per auction type, consisting of 4 boxplot visualizations, and the results of the statistical analysis for every recommendation model. The operationalized recommendation models of Eichstädt [20], Schulze-Horn et al. [84], and Berz et al. [6] are applied on the empirical data to see if the recommendation of the best auction type per scenario are in line with the empirical observations in Figure 5.2, Figure 5.5, and Figure 5.4, respectively. Additionally, the results of the statistical analysis of the difference in population means between the auction types inside a quadrant are shown in Table 5.1, Table 5.4, and Table 5.3, respectively. In the next paragaphs, the visualizations and the statistical result tables are explained in more detail.

Boxplot Savings Distribution Visualization

Figure 5.1 highlights the important elements on how to read a boxplot. The median is denoted by the horizontal line, whereas the average is represented by the red dot for each boxplot. To understand how to read the visualization, Figure 5.2 will yield as an example. The figure is composed out of 4 single boxplot visualization combined into a bigger figure, mimicing the recommendation model visualization in the likes of Figure 3.4, and 3.7. The y-axis, comparable to the recommendation model visualizations like Figure 3.4, categorizes the single boxplots by the price dispersion. The x-axis analogously represent the categorization of the number of bidders. This implies that the boxplot in the upper

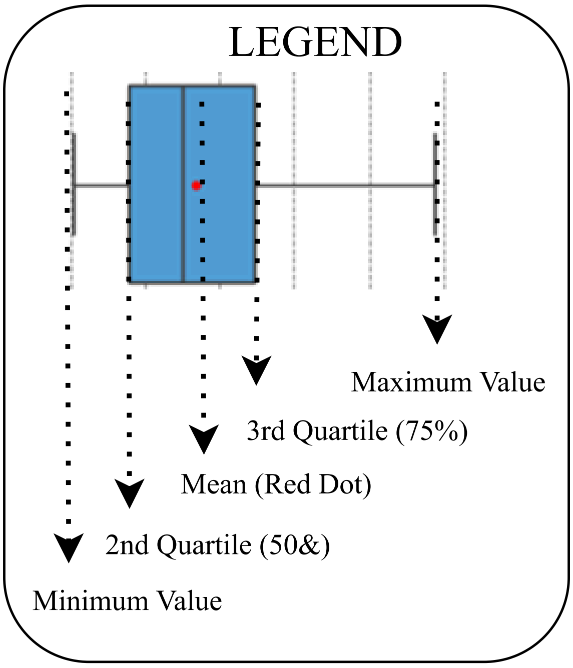


FIGURE 5.1: Legend of Boxplot Visualization

right hand corner is the savings distribution of the different auction types, for all auctions that have a high number of bidders and a high price dispersion. Similarly, the lower left hand corner shows the savings distribution of the different auction types for all auctions that have a low number of bidders and a low price dispersion. To represent the models with three dimensions as well, the auction volume is denoted by a color hue applied to the boxplot and the legend inside the figures explain the specific threshold values as well as the color allocation. For both Figure 5.5 and Figure 5.4, the red colored versions of the boxplot for every auction type in each quadrant are the auctions with low volume, while green are the auctions with higher volumes.

Tables of Statistical Test Result

The tables representing the results of the conducted statistical tests, like Table 5.1, contain the information about the whole statistical analysis, including the statistical test performed, the degrees of freedom (DF), the value of the test statistics with its corresponding p-value at a 5% significance level, a summary of the auction type distributions and their means in the group, as well as the result of the Shapiro-Wilk test[88] and Levene's test [45] for the assumption of the homogeneity of variances. For the former to last column, a p-value less than 5% means that we assume normality based on the Shapiro-Wilk test [88]. For the last column, a p-value less than 5%, we assume the homogeneity of the variances between the groups to be significantly different, based on Levene's test [45]. The table additionally summarizes which hypothesis was tested, which scenario applied and whether the hypothesis is supported by the statistical test.

Combining the information form the visual insights with the results of the statistical test, enables the answering of the SRQ2 and the analysis of the hypothesis stipulated by the recommendation models. The following sections discuss the results for each model tested in more detail.

5.1.1 Eichstädt (2007) Hypotheses

Figure 5.2 draws the savings distributions for each quadrant and each auction mode in every one of the four defined scenarios of the recommendation model. As mentioned in the section 4.2, for Eichstädt's [20] model, the price dispersion shown on the outer y-axis is considered to be the spread between the best and second-best bid, and the cutoff between high and low is set at 3%, as defined by Schulze-Horn et al. [84]. The outer x-axis is the number of bidders, with the commonly agreed-upon cutoff of three bidders. The y-axis for each boxplot shows the savings in percent for each auction.

The empirical findings support Eichstädt's (2007) hypotheses across all instances where a direct comparison is feasible, as outlined in Table 5.1. When there are few bidders and the initial bid spread is narrow, the English auction is shown to yield higher average savings than both the Japanese and Dutch auctions, with this difference being statistically significant. The average savings from the Japanese and Dutch auctions do not differ significantly from one another. A similar trend is observed when the number of bidders increases, with the mean savings from the English auction significantly outpacing those of the first-price auctions, namely the Dutch and Japanese auctions. Both observations tend to support the argument that the English auction is able to utilize the competition between bidders in these scenarios to drive prices down.

In scenarios where the gap between the highest and second-highest initial bids is large but the number of bidders is small, the first-price auctions demonstrate statistically significant higher average savings compared to the English auction. However, this distinction disappears when both the bid spread and the number of bidders are large, as no significant difference in mean savings is observed in any auction mode.

No statistical difference is observable between the Dutch and Japanese auctions in any scenario, highlighting that the first-price auctions seem to tend towards the same result on average. Generally, the first-price auctions inhibit a smaller variance compared to the English auction across all scenarios. Especially if the bid spread is large, many of the savings values for the English auction in the distribution are at the lower tail of the first quantile and close to 0% saving, indicating low bidder activity.

To summarize the results related to the auction type recommendation model of Eichstädt [20], the hypotheses H1E, H2E, and H3E, as defined in Table 5.1, are supported by our empirical test. While no inference can be made from the available data about Hypothesis H4E as it does not contain hybrid auctions, the similar mean across all three modes might indicate that choosing the English auction as the first phase in this scenario might not be the only viable possibility.

Moreover, the hypotheses related to the information exchange mechanisms of the English auction are mostly in line with the empirical findings, as shown in Figure 5.3 and Table 5.1. In the scenarios of low spread and high number of bidders, as well as high spread and low number of bidders, the differences between the information exchanges are statistically significant. As stipulated by hypothesis H3E, once the bid spread is low, the empirical insights indicate that English auctions, which show the best bid to all bidders, seem to have higher means with statistical significance at a 10% significance level. This

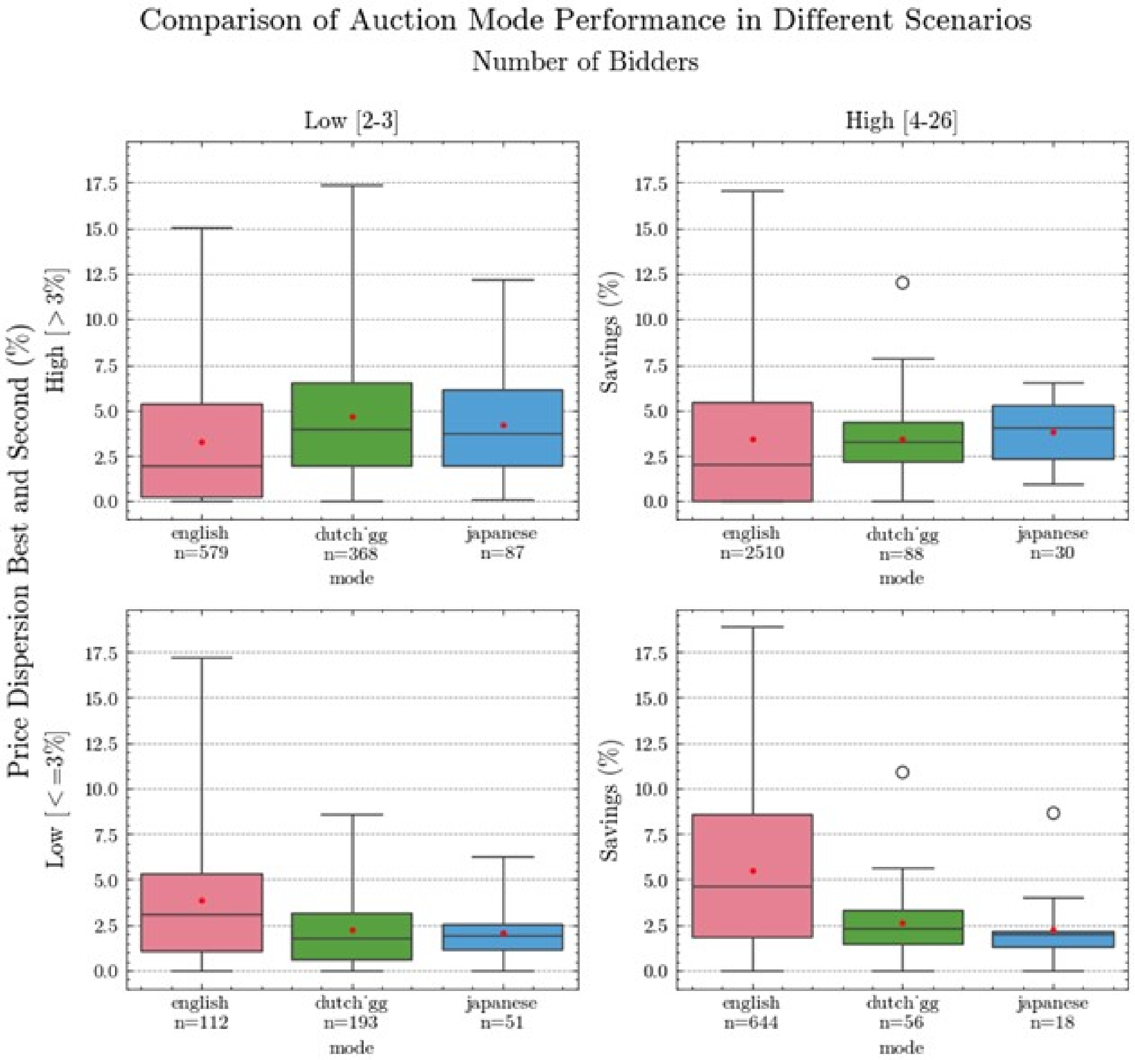
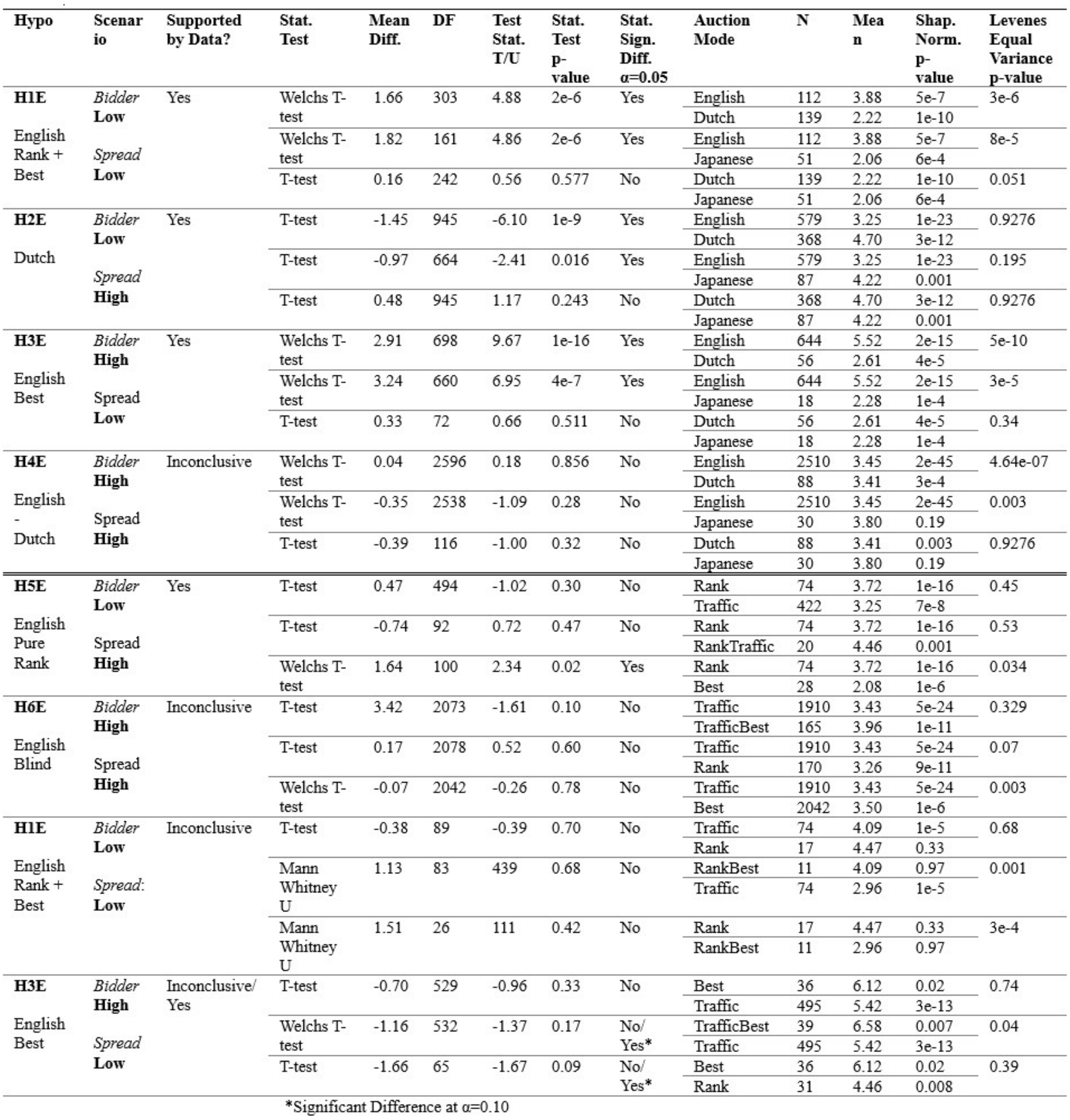


FIGURE 5.2: Empirical Analysis of Eichstädt, [20] Model with the savings distributions plotted against the number of bidders and the initial spread of best and second best bid for each auction mode.

Table 5.1: Results of statistical analysis of auction modes and information exchanges for different scenarios by means of two tailed statistical tests. Comparison of scenarios and auction type recommendations made by Eichstädt [20] on two dimensions



Hypo	**Scenario**	**Supported by Data?**	**Stat. Test**	**Mean Diff.**	**DF**	**Test Stat. T/U**	**Stat. Test p-value**	**Stat. Sign. Diff. α=0.05**	**Auction Mode**	**N**	**Mean**	**Shap. Norm. p-value**	**Levenes Equal Variance p-value**
H1E English Rank + Best	*Bidder* **Low** *Spread* **Low**	Yes	Welchs T-test	1.66	303	4.88	2e-6	Yes	English	112	3.88	5e-7	3e-6
									Dutch	139	2.22	1e-10	
			Welchs T-test	1.82	161	4.86	2e-6	Yes	English	112	3.88	5e-7	8e-5
									Japanese	51	2.06	6e-4	
			T-test	0.16	242	0.56	0.577	No	Dutch	139	2.22	1e-10	0.051
									Japanese	51	2.06	6e-4	
H2E Dutch	*Bidder* **Low** *Spread* **High**	Yes	T-test	-1.45	945	-6.10	1e-9	Yes	English	579	3.25	1e-23	0.9276
									Dutch	368	4.70	3e-12	
			T-test	-0.97	664	-2.41	0.016	Yes	English	579	3.25	1e-23	0.195
									Japanese	87	4.22	0.001	
			T-test	0.48	945	1.17	0.243	No	Dutch	368	4.70	3e-12	0.9276
									Japanese	87	4.22	0.001	
H3E English Best	*Bidder* **High** Spread **Low**	Yes	Welchs T-test	2.91	698	9.67	1e-16	Yes	English	644	5.52	2e-15	5e-10
									Dutch	56	2.61	4e-5	
			Welchs T-test	3.24	660	6.95	4e-7	Yes	English	644	5.52	2e-15	3e-5
									Japanese	18	2.28	1e-4	
			T-test	0.33	72	0.66	0.511	No	Dutch	56	2.61	4e-5	0.34
									Japanese	18	2.28	1e-4	
H4E English - Dutch	*Bidder* **High** Spread **High**	Inconclusive	Welchs T-test	0.04	2596	0.18	0.856	No	English	2510	3.45	2e-45	4.64e-07
									Dutch	88	3.41	3e-4	
			Welchs T-test	-0.35	2538	-1.09	0.28	No	English	2510	3.45	2e-45	0.003
									Japanese	30	3.80	0.19	
			T-test	-0.39	116	-1.00	0.32	No	Dutch	88	3.41	0.003	0.9276
									Japanese	30	3.80	0.19	
H5E English Pure Rank	*Bidder* **Low** Spread **High**	Yes	T-test	0.47	494	-1.02	0.30	No	Rank	74	3.72	1e-16	0.45
									Traffic	422	3.25	7e-8	
			T-test	-0.74	92	0.72	0.47	No	Rank	74	3.72	1e-16	0.53
									RankTraffic	20	4.46	0.001	
			Welchs T-test	1.64	100	2.34	0.02	Yes	Rank	74	3.72	1e-16	0.034
									Best	28	2.08	1e-6	
H6E English Blind	*Bidder* **High** Spread **High**	Inconclusive	T-test	3.42	2073	-1.61	0.10	No	Traffic	1910	3.43	5e-24	0.329
									TrafficBest	165	3.96	1e-11	
			T-test	0.17	2078	0.52	0.60	No	Traffic	1910	3.43	5e-24	0.07
									Rank	170	3.26	9e-11	
			Welchs T-test	-0.07	2042	-0.26	0.78	No	Traffic	1910	3.43	5e-24	0.003
									Best	2042	3.50	1e-6	
H1E English Rank + Best	*Bidder* **Low** *Spread*: **Low**	Inconclusive	T-test	-0.38	89	-0.39	0.70	No	Traffic	74	4.09	1e-5	0.68
									Rank	17	4.47	0.33	
			Mann Whitney U	1.13	83	439	0.68	No	RankBest	11	4.09	0.97	0.001
									Traffic	74	2.96	1e-5	
			Mann Whitney U	1.51	26	111	0.42	No	Rank	17	4.47	0.33	3e-4
									RankBest	11	2.96	0.97	
H3E English Best	*Bidder* **High** *Spread* **Low**	Inconclusive/ Yes	T-test	-0.70	529	-0.96	0.33	No	Best	36	6.12	0.02	0.74
									Traffic	495	5.42	3e-13	
			Welchs T-test	-1.16	532	-1.37	0.17	No/ Yes*	TrafficBest	39	6.58	0.007	0.04
									Traffic	495	5.42	3e-13	
			T-test	-1.66	65	-1.67	0.09	No/ Yes*	Best	36	6.12	0.02	0.39
									Rank	31	4.46	0.008	

*Significant Difference at α=0.10

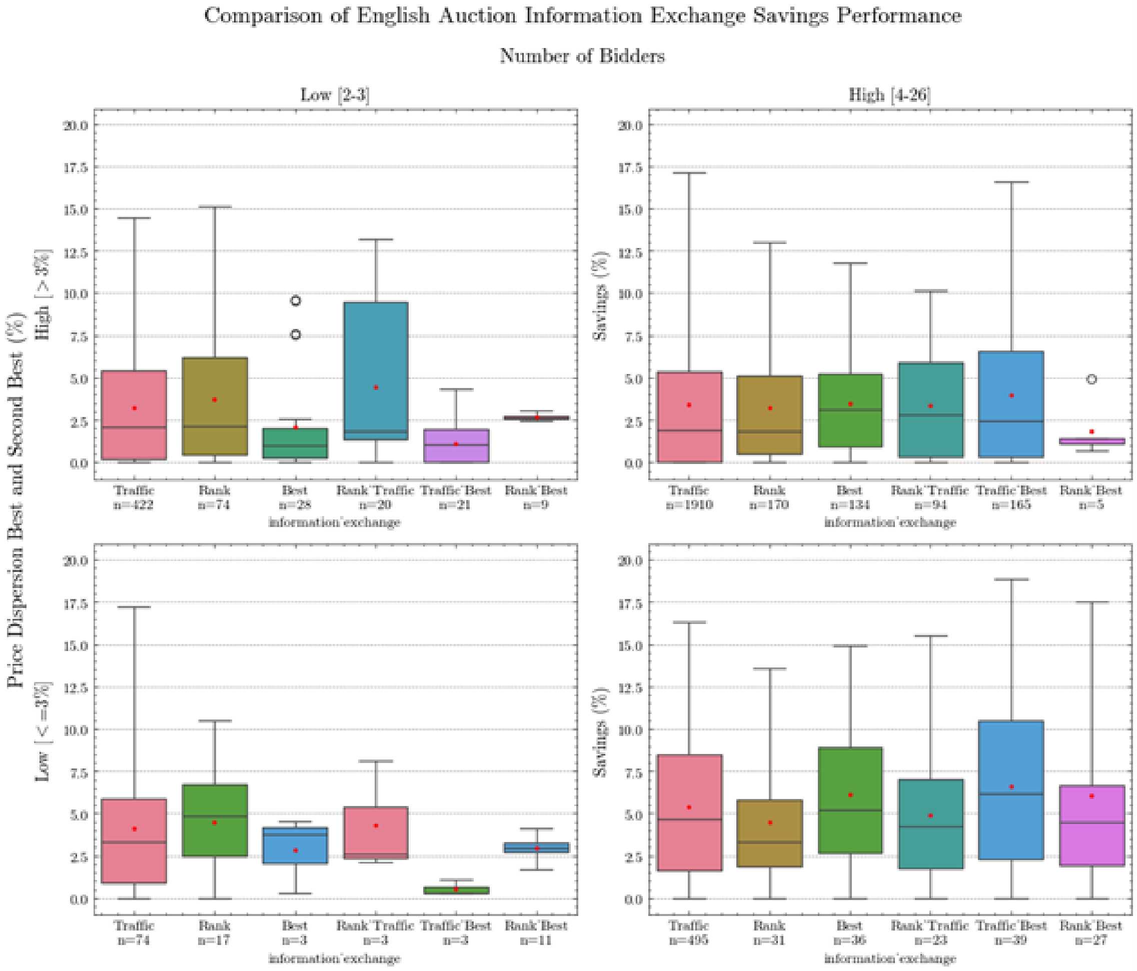


FIGURE 5.3: Empirical Analysis of Eichstädt, [20] Model with the savings distributions plotted against the number of bidders and the initial spread of best and second best bid for each information policy for the english auction type.

also holds for hypothesis H5E, for which the high bid spread and low number of bidders propose a rank auction that hides the best bid. The data indicate that the theoretical line of argumentation to pay special attention to the bidder's perceived probability of winning coincides with practical results. For both the high number of bidders and high price dispersion, as well as for the low number of bidders with low price dispersion scenarios, the analysis does not yield any conclusive results of significant differences between the information exchange modes.

5.1.2 Schulze-Horn et al. (2018) Hypotheses

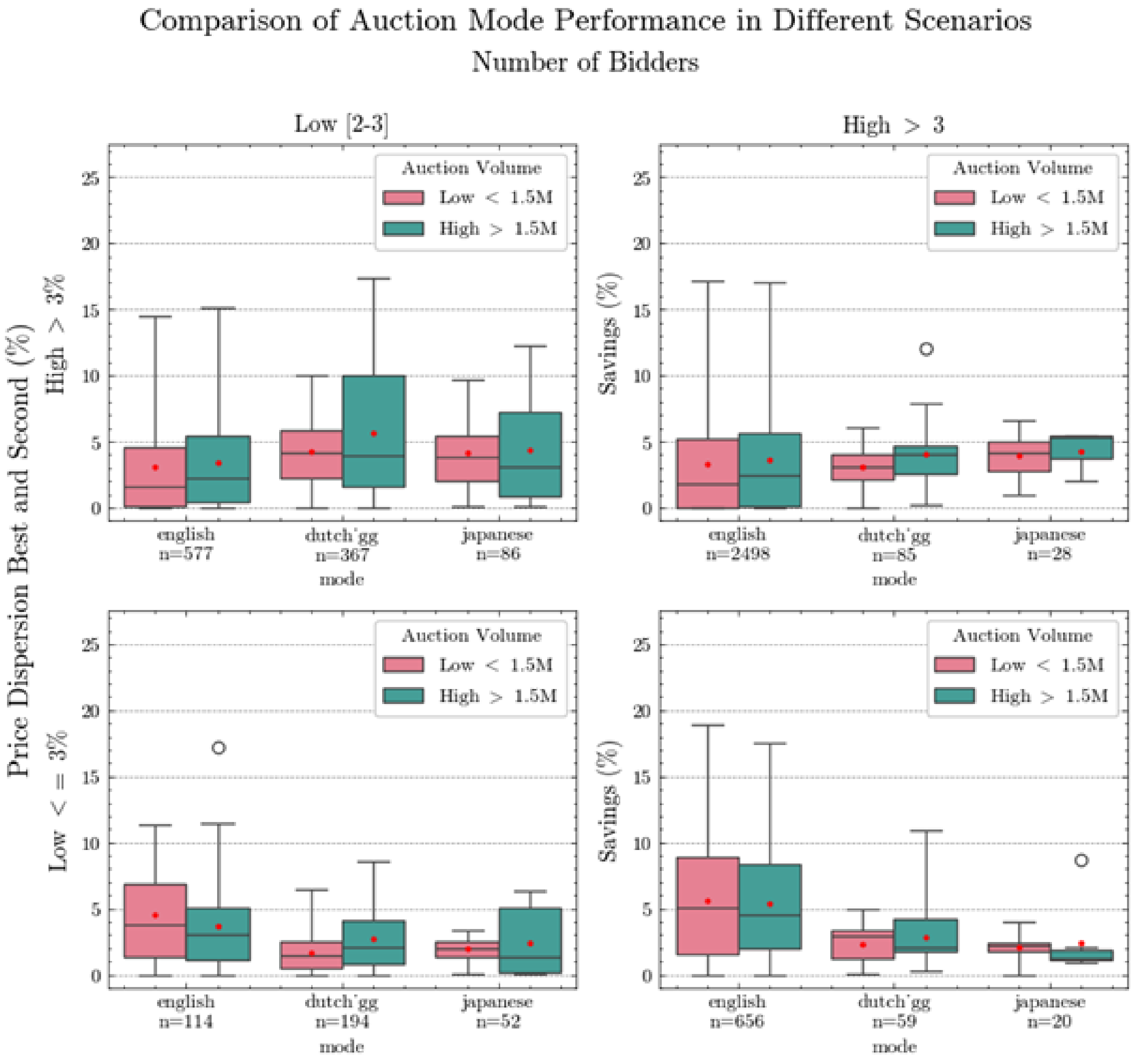


Figure 5.4: Empirical Analysis of Schulze-Horn et al. [84] model with the savings distributions plotted against the number of bidders, the auction volume, and the initial bid spread as defined by the spread of best and second best bid

For the model represented in Figure 5.4, the price dispersion is defined as the difference between the best and second-best bid, as described by Schulze-Horn et al. [84]. In contrast to Berz et al. [6] and in alignment with the empirical insights, they recommend using a second-price ticker auction, also known as an English ticker auction or Japanese 2nd price auction, in the case that a low number of bidders has close bid spreads, indicating equally strong bidders in Hypothesis H1S. The attractivity does not influence the recommendation, and the empirical insights show that although the difference between the first-price and second-price auctions is getting smaller, the English auction is still significantly different from the Dutch auction, aligning with the hypothesis H2S.

Also, in cases of high price dispersion with a low number of bidders, H3S is supported, as the Dutch auction achieves significantly higher performance than the other auction types, and while H4S stipulates the sequential Dutch auction, which is not present in the data, it is still clear that the first-price auction mechanisms are outperforming.

Again, although not directly comparable, Schulze-Horn et al. [84] hypothesize that H5S and H6S fit the observations from the data, in which second-price auctions outperform first-price auctions in the scenario that a high number of bidders are met with a close bid spread between them. Independent of the attractivity, the recommendation of the auction type remains the same.

For both the high number of bidders and the high bid spread, the Dutch auction is stipulated to be the best recommendation, but there is statistically significant evidence that the Japanese auction achieves higher savings on average than either the Dutch or the English auction if the auction volume is low. This represents the first scenario in which the Japanese auction distinguishes itself from the Dutch auction significantly.

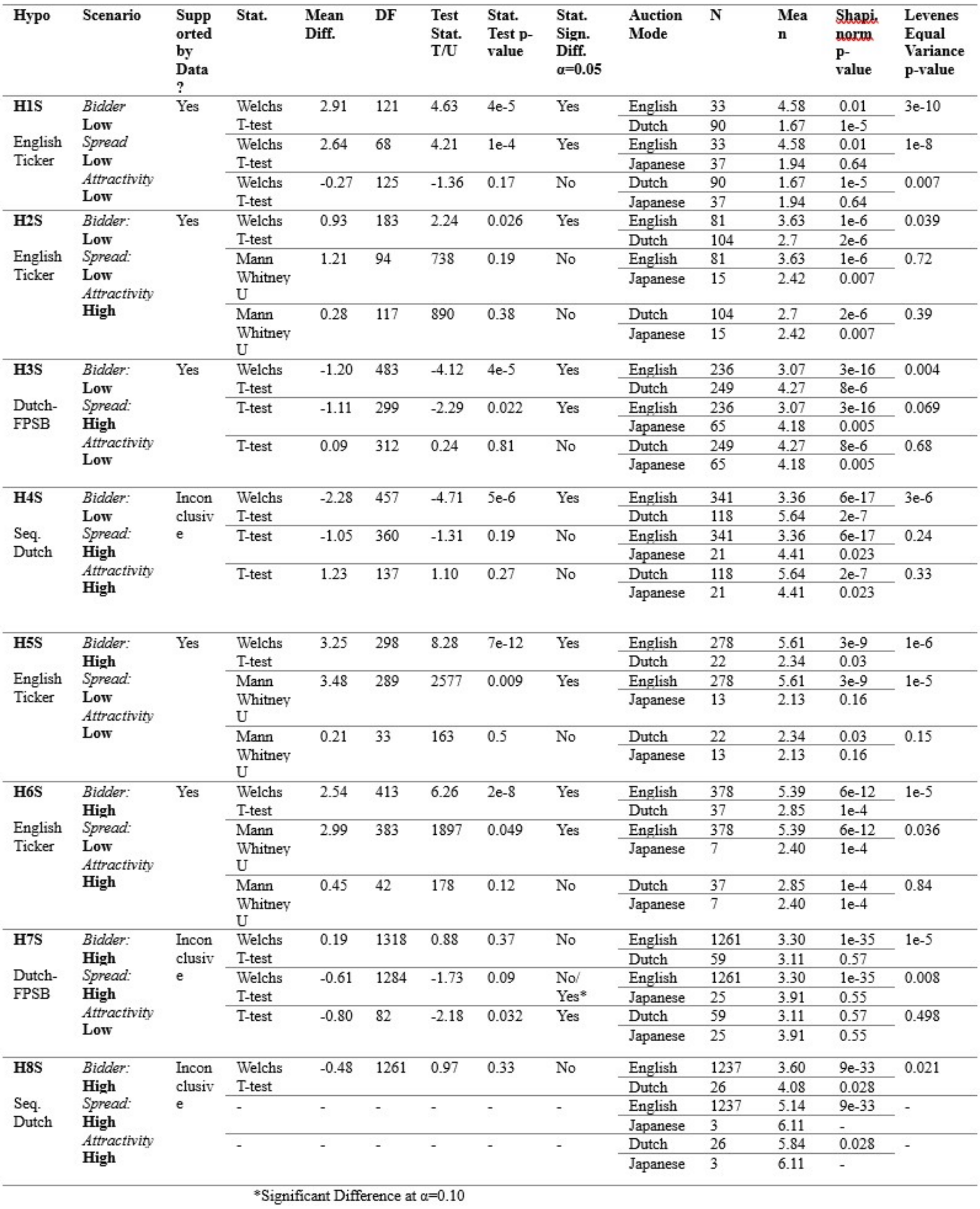
Possible explanations for the higher performance of the Japanese auctions, although it is a first-price auction, might lead to the more active participation of bidders in the process by always confirming every ticker step until the end. It might alleviate some of the pressures a Dutch auction could induce while at the same time capitalizing on the first-price nature to get close to the bidder's reservation prices. Interestingly, if the price dispersion is computed along all initial bids, as in the case of Table 5.3 and Figure 5.5, the higher cost savings of the Japanese to the Dutch auction is less pronounced. This could hint at the fact that if the best bidder is much more dominant, a Japanese auction might show higher cost savings than a Dutch auction.

5.1.3 Berz et al. (2021) Hypotheses

Figure 5.5 adds the third dimension to the analysis by introducing the distinction in the auction volume, coded by the red color for low auction volumes and green for high auction volumes of the same auction type distribution. As can be seen by the general trend in Figure 5.5, it is observable that if the auction volume is high, the first-price auctions seem to yield higher cost savings compared to their low-volume counterparts. This supports the hypothesis of Berz et al. [6] that first-price auctions are more effective when the risk aversion of the bidders is high. This is especially pronounced if the initial price dispersion is high and there is a low number of bidders for the Dutch auction.

Contrary to hypothesis H1B, where they propose to conduct two subsequent FPSB auctions in case of a low number of bidders and low price dispersion, the first-price auctions appear to yield lower savings than the second-priced English auction. Although we cannot directly compare this recommendation due to the missing FPSB auction in the data, the observation might indicate that the decision to apply an English auction only in the case with a higher number of bidders participating might not be reflected in the

Table 5.2: Results of statistical analysis of auction modes for different scenarios by means of two tailed statistical tests. Comparison of scenarios and auction type recommendations made by Schulze-Horn et al. [84] along three dimensions



Hypo	Scenario	Supported by Data ?	Stat.	Mean Diff.	DF	Test Stat. T/U	Stat. Test p-value	Stat. Sign. Diff. α=0.05	Auction Mode	N	Mean	Shapi. norm p-value	Levenes Equal Variance p-value
H1S English Ticker	*Bidder* **Low** *Spread* **Low** *Attractivity* **Low**	Yes	Welchs T-test	2.91	121	4.63	4e-5	Yes	English	33	4.58	0.01	3e-10
									Dutch	90	1.67	1e-5	
			Welchs T-test	2.64	68	4.21	1e-4	Yes	English	33	4.58	0.01	1e-8
									Japanese	37	1.94	0.64	
			Welchs T-test	-0.27	125	-1.36	0.17	No	Dutch	90	1.67	1e-5	0.007
									Japanese	37	1.94	0.64	
H2S English Ticker	*Bidder:* **Low** *Spread:* **Low** *Attractivity* **High**	Yes	Welchs T-test	0.93	183	2.24	0.026	Yes	English	81	3.63	1e-6	0.039
									Dutch	104	2.7	2e-6	
			Mann Whitney U	1.21	94	738	0.19	No	English	81	3.63	1e-6	0.72
									Japanese	15	2.42	0.007	
			Mann Whitney U	0.28	117	890	0.38	No	Dutch	104	2.7	2e-6	0.39
									Japanese	15	2.42	0.007	
H3S Dutch-FPSB	*Bidder:* **Low** *Spread:* **High** *Attractivity* **Low**	Yes	Welchs T-test	-1.20	483	-4.12	4e-5	Yes	English	236	3.07	3e-16	0.004
									Dutch	249	4.27	8e-6	
			T-test	-1.11	299	-2.29	0.022	Yes	English	236	3.07	3e-16	0.069
									Japanese	65	4.18	0.005	
			T-test	0.09	312	0.24	0.81	No	Dutch	249	4.27	8e-6	0.68
									Japanese	65	4.18	0.005	
H4S Seq. Dutch	*Bidder:* **Low** *Spread:* **High** *Attractivity* **High**	Inconclusive	Welchs T-test	-2.28	457	-4.71	5e-6	Yes	English	341	3.36	6e-17	3e-6
									Dutch	118	5.64	2e-7	
			T-test	-1.05	360	-1.31	0.19	No	English	341	3.36	6e-17	0.24
									Japanese	21	4.41	0.023	
			T-test	1.23	137	1.10	0.27	No	Dutch	118	5.64	2e-7	0.33
									Japanese	21	4.41	0.023	
H5S English Ticker	*Bidder:* **High** *Spread:* **Low** *Attractivity* **Low**	Yes	Welchs T-test	3.25	298	8.28	7e-12	Yes	English	278	5.61	3e-9	1e-6
									Dutch	22	2.34	0.03	
			Mann Whitney U	3.48	289	2577	0.009	Yes	English	278	5.61	3e-9	1e-5
									Japanese	13	2.13	0.16	
			Mann Whitney U	0.21	33	163	0.5	No	Dutch	22	2.34	0.03	0.15
									Japanese	13	2.13	0.16	
H6S English Ticker	*Bidder:* **High** *Spread:* **Low** *Attractivity* **High**	Yes	Welchs T-test	2.54	413	6.26	2e-8	Yes	English	378	5.39	6e-12	1e-5
									Dutch	37	2.85	1e-4	
			Mann Whitney U	2.99	383	1897	0.049	Yes	English	378	5.39	6e-12	0.036
									Japanese	7	2.40	1e-4	
			Mann Whitney U	0.45	42	178	0.12	No	Dutch	37	2.85	1e-4	0.84
									Japanese	7	2.40	1e-4	
H7S Dutch-FPSB	*Bidder:* **High** *Spread:* **High** *Attractivity* **Low**	Inconclusive	Welchs T-test	0.19	1318	0.88	0.37	No	English	1261	3.30	1e-35	1e-5
									Dutch	59	3.11	0.57	
			Welchs T-test	-0.61	1284	-1.73	0.09	No/ Yes*	English	1261	3.30	1e-35	0.008
									Japanese	25	3.91	0.55	
			T-test	-0.80	82	-2.18	0.032	Yes	Dutch	59	3.11	0.57	0.498
									Japanese	25	3.91	0.55	
H8S Seq. Dutch	*Bidder:* **High** *Spread:* **High** *Attractivity* **High**	Inconclusive	Welchs T-test	-0.48	1261	0.97	0.33	No	English	1237	3.60	9e-33	0.021
									Dutch	26	4.08	0.028	
			-	-	-	-	-	-	English	1237	5.14	9e-33	-
									Japanese	3	6.11	-	
			-	-	-	-	-	-	Dutch	26	5.84	0.028	-
									Japanese	3	6.11	-	

*Significant Difference at α=0.10

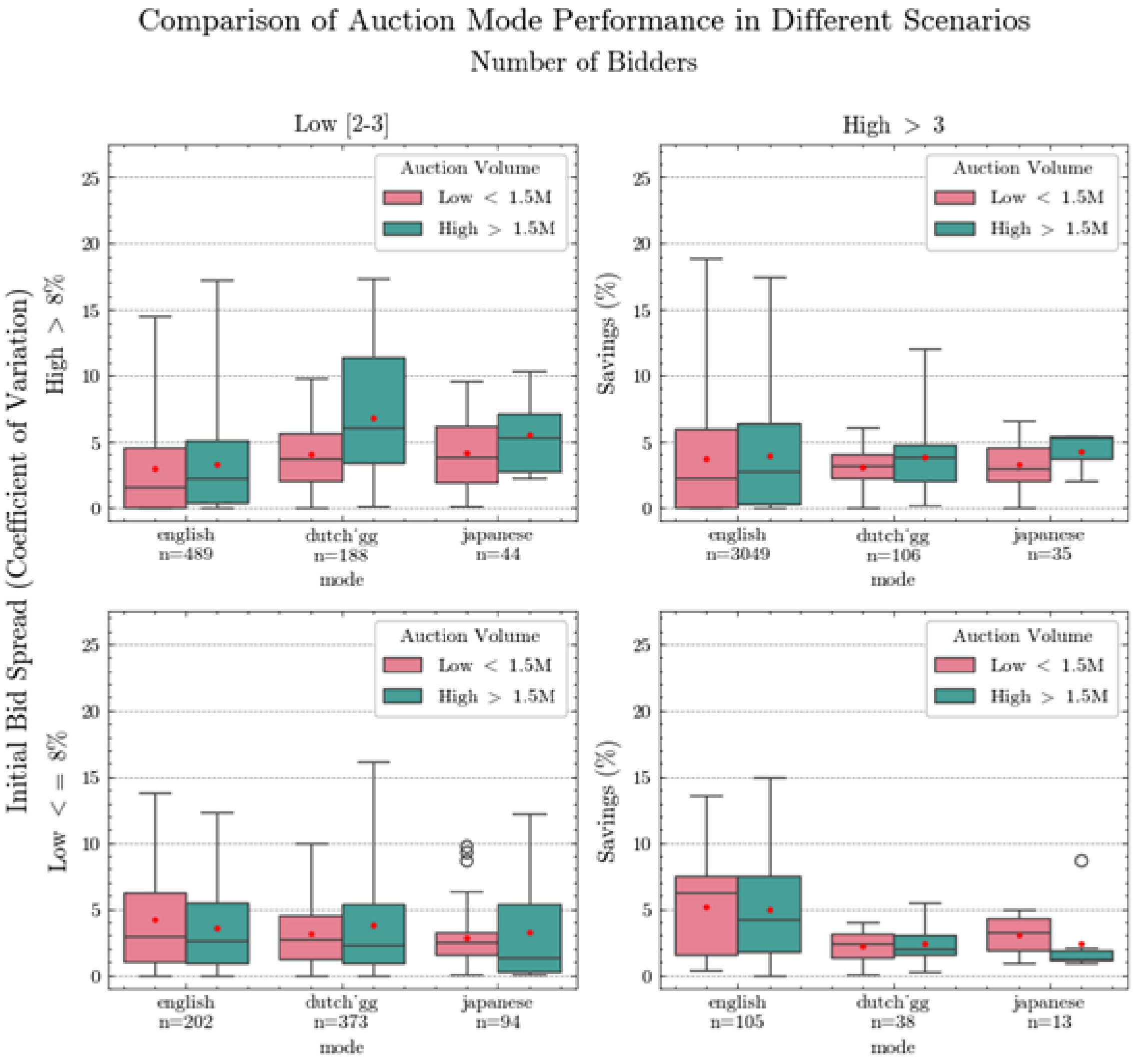


FIGURE 5.5: Empirical analysis of Berz et al. [6] model with the savings distributions plotted against the number of bidders, the initial bid spread as defined by the coefficient of variation and the auction volume

practical results. The difference between the English auction and the Dutch auction is not statistically significant in this case.

In a situation with a low number of bidders and a close bid spread combined with a high auction volume, the Dutch auction has the highest recorded mean. Although the differences are not statistically significant, it still highlights that as the auction volume and therefore the possible attractivity of the business grow, first-price auctions seem to perform better. Remaining with the low number of bidders but recording a higher spread of all initial bids, a clear favorite, namely the Dutch auction, appears, especially if the auction volume is high. This observation aligns with the hypotheses H4B and H3B, considering that we use the first-price auctions in Dutch and Japanese as substitutes for the missing FPSB.

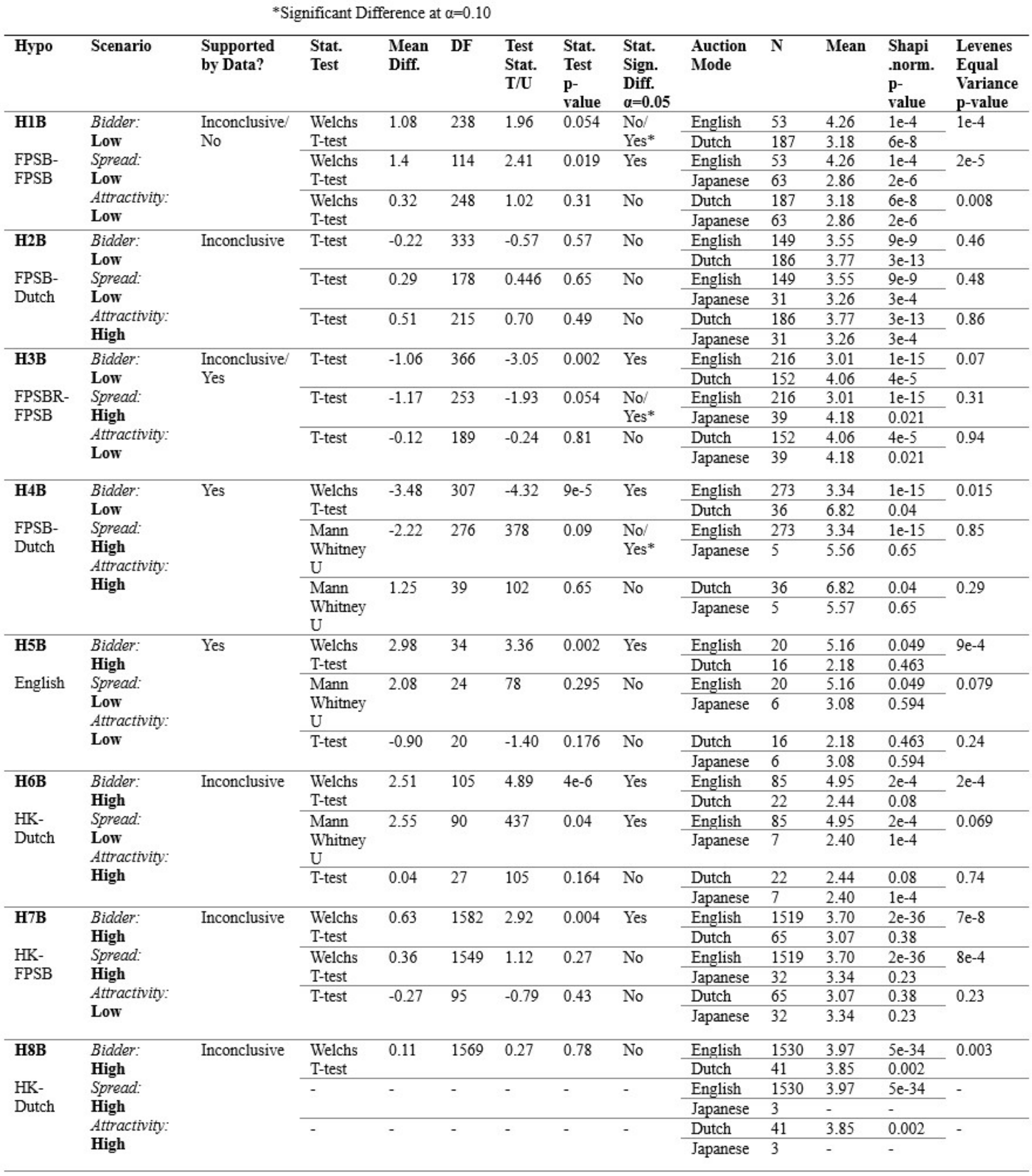
In cases where the number of bidders is high and their initial bids have low variation, independent of the auction volume, the English auction significantly outperforms the Dutch and Japanese auctions, aligning with H5B of Berz et al. [6]. In cases where the spread between the initial bids increases, the English auction seems to have the upper hand compared to the Dutch auction, but no significant difference in means can be observed between the English and Japanese auctions. Nevertheless, first-price auctions have increased savings if the attractivity is higher, indicating the alignment of the theoretical hypotheses regarding the relation between possible risk aversion and the use of first-price auctions, although the empirical data cannot test the hypothesis H8B directly.

5.1.4 Optimized Recommendation Model based on Empirical Data

To summarize, most hypotheses that can be directly compared were found to be supported by the empirical data for all three models analyzed. In some cases, no conclusive support for the hypotheses could be found, as only the general class of first- or second-price auction mode could be compared to the existing auction types in the data, in addition to the missing wholistic evaluation of the effect of multiple phases. In general, the data shows that the small bid spread, either defined by the coefficient of variation or the difference between the best and second-best bid, tends to favor the first-price auctions, especially if the number of bidders is low. The Dutch and Japanese auctions achieve comparable performances in most scenarios. Except in the case of a high price dispersion between the best two bidders and if the number of bidders is low, the Dutch auction outperforms the Japanese auction, but if the number of bidders is high, the Japanese auction outperforms. The second price auction, specifically the English auction is favored when the bid spread is low and in case the number of bidders is high, independent of the bid spread. Based on the analysis of the models and the insights from the empirical data, a new recommendation framework is proposed and presented in Figure 5.6.

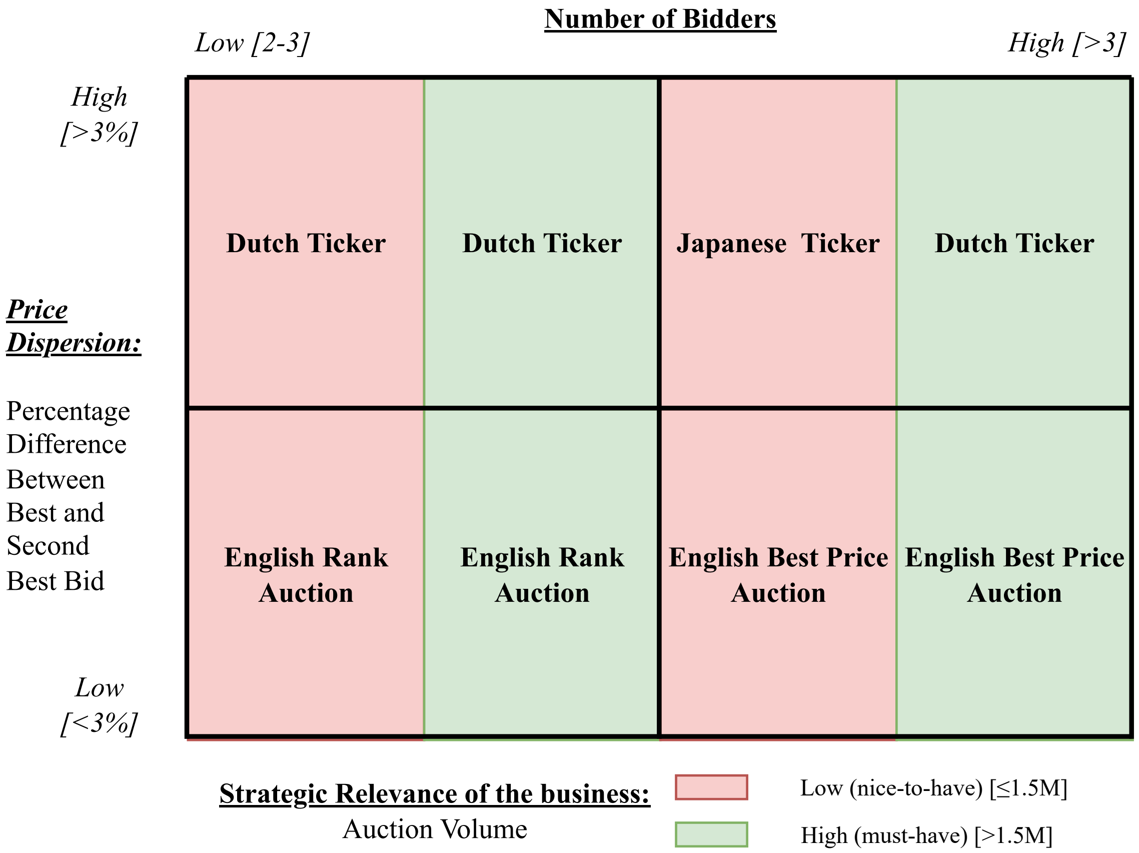
The three dimensions represent the three dimensions of the auction cube, which specifically considers the price dispersion between the two best initial bids rather than the price dispersion among all submitted bids. The former shows slightly higher differences between the performances of the first-price auctions compared to the second-price auctions, and the operationalization of Schulze-Horn et al. [84] seems to also work for the specific empirical data at hand, indicating a possible generalizability of the chosen threshold values. For the strategic relevance of the business dimension, which is approximated in the proposed model by the auction volume, it is recommended to utilize the quantile binning technique to obtain the necessary practical thresholds based on the median of specific material groups.

Table 5.3: Results of statistical analysis of auction modes for different scenarios by means of two tailed statistical tests. Comparison of auction type recommendations made by Berz et al. [6] along three dimensions



*Significant Difference at α=0.10

Hypo	Scenario	Supported by Data?	Stat. Test	Mean Diff.	DF	Test Stat. T/U	Stat. Test p-value	Stat. Sign. Diff. α=0.05	Auction Mode	N	Mean	Shapi .norm. p-value	Levenes Equal Variance p-value
H1B FPSB-FPSB	*Bidder:* **Low**	Inconclusive/ No	Welchs T-test	1.08	238	1.96	0.054	No/ Yes*	English	53	4.26	1e-4	1e-4
									Dutch	187	3.18	6e-8	
	Spread: **Low**		Welchs T-test	1.4	114	2.41	0.019	Yes	English	53	4.26	1e-4	2e-5
									Japanese	63	2.86	2e-6	
	Attractivity: **Low**		Welchs T-test	0.32	248	1.02	0.31	No	Dutch	187	3.18	6e-8	0.008
									Japanese	63	2.86	2e-6	
H2B FPSB-Dutch	*Bidder:* **Low**	Inconclusive	T-test	-0.22	333	-0.57	0.57	No	English	149	3.55	9e-9	0.46
									Dutch	186	3.77	3e-13	
	Spread: **Low**		T-test	0.29	178	0.446	0.65	No	English	149	3.55	9e-9	0.48
									Japanese	31	3.26	3e-4	
	Attractivity: **High**		T-test	0.51	215	0.70	0.49	No	Dutch	186	3.77	3e-13	0.86
									Japanese	31	3.26	3e-4	
H3B FPSBR-FPSB	*Bidder:* **Low**	Inconclusive/ Yes	T-test	-1.06	366	-3.05	0.002	Yes	English	216	3.01	1e-15	0.07
									Dutch	152	4.06	4e-5	
	Spread: **High**		T-test	-1.17	253	-1.93	0.054	No/ Yes*	English	216	3.01	1e-15	0.31
									Japanese	39	4.18	0.021	
	Attractivity: **Low**		T-test	-0.12	189	-0.24	0.81	No	Dutch	152	4.06	4e-5	0.94
									Japanese	39	4.18	0.021	
H4B FPSB-Dutch	*Bidder:* **Low**	Yes	Welchs T-test	-3.48	307	-4.32	9e-5	Yes	English	273	3.34	1e-15	0.015
									Dutch	36	6.82	0.04	
	Spread: **High**		Mann Whitney U	-2.22	276	378	0.09	No/ Yes*	English	273	3.34	1e-15	0.85
									Japanese	5	5.56	0.65	
	Attractivity: **High**		Mann Whitney U	1.25	39	102	0.65	No	Dutch	36	6.82	0.04	0.29
									Japanese	5	5.57	0.65	
H5B English	*Bidder:* **High**	Yes	Welchs T-test	2.98	34	3.36	0.002	Yes	English	20	5.16	0.049	9e-4
									Dutch	16	2.18	0.463	
	Spread: **Low**		Mann Whitney U	2.08	24	78	0.295	No	English	20	5.16	0.049	0.079
									Japanese	6	3.08	0.594	
	Attractivity: **Low**		T-test	-0.90	20	-1.40	0.176	No	Dutch	16	2.18	0.463	0.24
									Japanese	6	3.08	0.594	
H6B HK-Dutch	*Bidder:* **High**	Inconclusive	Welchs T-test	2.51	105	4.89	4e-6	Yes	English	85	4.95	2e-4	2e-4
									Dutch	22	2.44	0.08	
	Spread: **Low**		Mann Whitney U	2.55	90	437	0.04	Yes	English	85	4.95	2e-4	0.069
									Japanese	7	2.40	1e-4	
	Attractivity: **High**		T-test	0.04	27	105	0.164	No	Dutch	22	2.44	0.08	0.74
									Japanese	7	2.40	1e-4	
H7B HK-FPSB	*Bidder:* **High**	Inconclusive	Welchs T-test	0.63	1582	2.92	0.004	Yes	English	1519	3.70	2e-36	7e-8
									Dutch	65	3.07	0.38	
	Spread: **High**		Welchs T-test	0.36	1549	1.12	0.27	No	English	1519	3.70	2e-36	8e-4
									Japanese	32	3.34	0.23	
	Attractivity: **Low**		T-test	-0.27	95	-0.79	0.43	No	Dutch	65	3.07	0.38	0.23
									Japanese	32	3.34	0.23	
H8B HK-Dutch	*Bidder:* **High**	Inconclusive	Welchs T-test	0.11	1569	0.27	0.78	No	English	1530	3.97	5e-34	0.003
									Dutch	41	3.85	0.002	
	Spread: **High**		-	-	-	-	-	-	English	1530	3.97	5e-34	-
									Japanese	3	-	-	
	Attractivity: **High**		-	-	-	-	-	-	Dutch	41	3.85	0.002	-
									Japanese	3	-	-	



	Number of Bidders			
	Low [2-3]			*High [>3]*
High [>3%]	🟥 **Dutch Ticker**	🟩 **Dutch Ticker**	🟥 **Japanese Ticker**	🟩 **Dutch Ticker**
Low [<3%]	🟥 **English Rank Auction**	🟩 **English Rank Auction**	🟥 **English Best Price Auction**	🟩 **English Best Price Auction**

Price Dispersion:

Percentage Difference Between Best and Second Best Bid

Strategic Relevance of the business:
Auction Volume

🟥 Low (nice-to-have) [≤1.5M]

🟩 High (must-have) [>1.5M]

FIGURE 5.6: Auction type recommendation model based upon the analysis of the empirical data and its statistical analysis along three dimensions

Ideally, the attractiveness of the business to the supplier should be evaluated based on a multitude of criteria and measures, not only the auction volume. In line with the common theoretical consensus, the threshold for the high number of bidder categories is defined as more than three participating bidders.

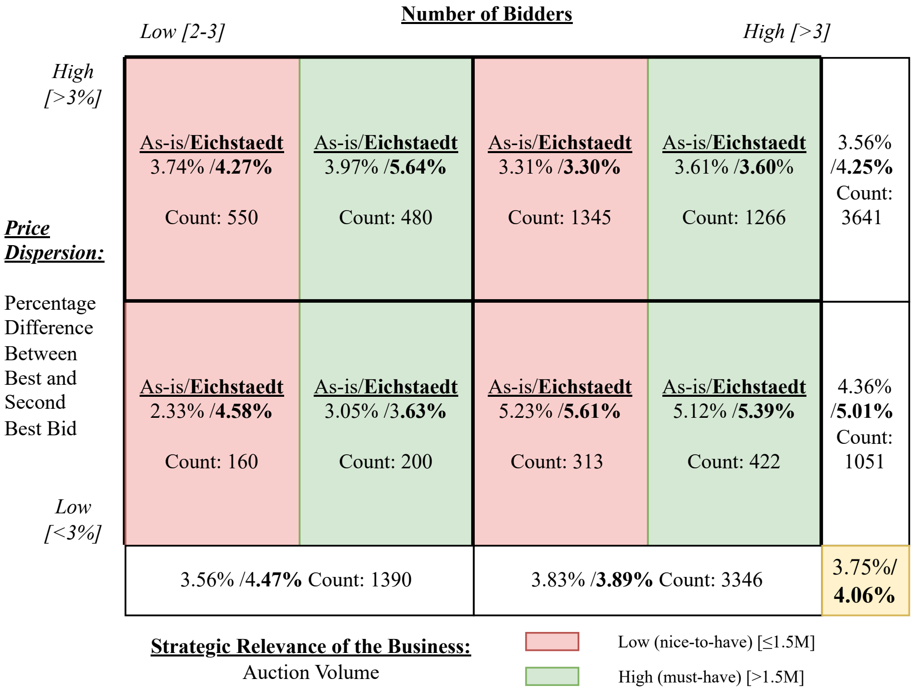
The recommendations are based on the observed empirical data and the statistical significance test. For three of the four quadrants defined by the number of bidders and the price dispersion, a different version of the dynamic English auction is proposed. When a few bidders with close prices meet, an English rank auction gives the bidders the impression of a low rank, encouraging more active bidding to finish along the top ranks. This recommendation is independent of the assumed attractivity of the business of the supplier as estimated by the auction volume in the case of the price dispersion being defined by the best two bids. If the price dispersion considers all initial bids, then a Dutch auction is recommended if the attractivity of the business is high. When many bidders with close price dispersion meet, showing a short distance to the leading bid establishes the groundwork for active participation to reach the best price. On the other hand, if the price dispersion is high, the empirical insights show the superior performance of a first-price auction, given this sign of a possible dominant buyer. The risk of losing the auction plays into the reduction of the strategic margin towards the reservation price to minimize the cost incurred due to losing the auction. The Dutch auction seems to excel, especially in cases where the strategic relevance of the business is large, while the Japanese auction performs well if the attractivity of the auction and the implied risk aversion to losing the auction is lower.

5.1.5 Expected Savings per Recommendation Model

By calculating the expected savings of the different auction type recommendation models in case they would have been applied compared to the As-is situation in the empirical data, the evaluation of the expected performances of the recommendation models is facilitated. The methodology for the calculation will be described based on Figure 5.7, which presents the calculation of the expected savings for the model of Eichstädt [20] compared to the as-is situation.

Similar to previous figures the quadrants describe the specific classification of the negotiation scenario based on the three dimensions. Inside each of these groups, the total count of auctions conducted in the empirical data is shown, alongside the average savings of the auctions conducted in that group, denoted as the as-is expected saving in percent. Beside that is the expected savings percentage if all auctions in that group would have been the auction type that Eichstädt [20] recommends, namely a Dutch auction as can be seen from Figure 3.4. The assumption made is that the calculated average savings of the Dutch auctions in the as-is situation for this group, is also the expected average saving for the whole group if all auctions would have been Dutch auctions. In that manner, there are two expected savings per group, the as-is savings percentage and the assumed recommended model percentage.

Additionally, the row and column marginals are computed as the weighted average of the auction count per group multiplied by the expected savings and summed up over the row or columns. These marginals are then again weighted by multiplying with the count of auctions and then summed to arrive at the total expected savings for the as-is situation and for the application of the recommendation model. In this way, the final achieved savings are made comparable and a closer look at the savings per group or margin is made possible.



Number of Bidders

Price Dispersion: Percentage Difference Between Best and Second Best Bid

	Low [2-3]		*High [>3]*		
High [>3%]	As-is/**Eichstaedt** 3.74% /**4.27%** Count: 550	As-is/**Eichstaedt** 3.97% /**5.64%** Count: 480	As-is/**Eichstaedt** 3.31% /**3.30%** Count: 1345	As-is/**Eichstaedt** 3.61% /**3.60**% Count: 1266	3.56% /4.**25%** Count: 3641
Low [<3%]	As-is/**Eichstaedt** 2.33% /**4.58%** Count: 160	As-is/**Eichstaedt** 3.05% /3**.63%** Count: 200	As-is/**Eichstaedt** 5.23% /**5.61%** Count: 313	As-is/**Eichstaedt** 5.12% /**5.39%** Count: 422	4.36% /**5.01%** Count: 1051
	3.56% /4**.47%** Count: 1390		3.83% /**3.89%** Count: 3346		3.75%/ **4.06%**

Strategic Relevance of the Business:
Auction Volume

🟥 Low (nice-to-have) [≤1.5M]

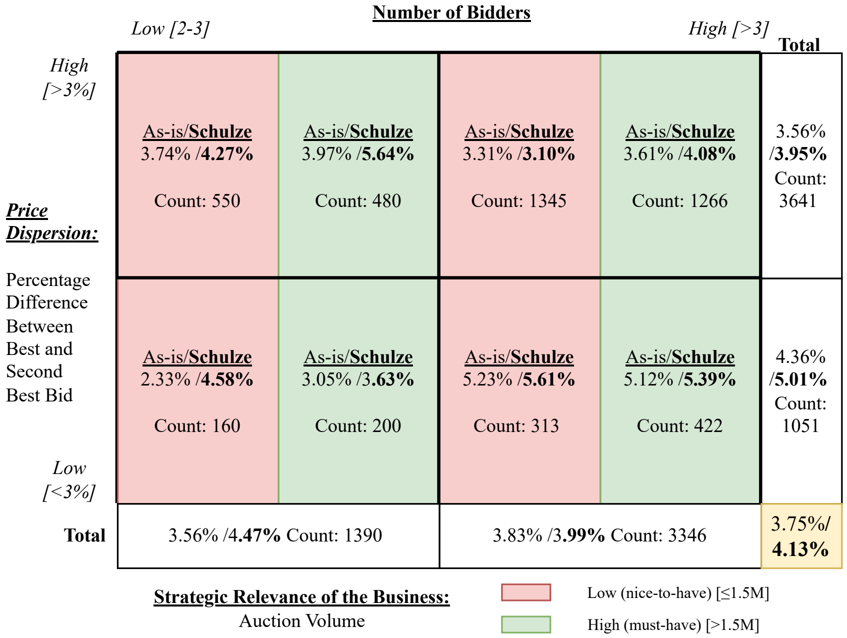
🟩 High (must-have) [>1.5M]

FIGURE 5.7: Expected Savings calculation based on the application of Eichstädt [20] Model. The as-is situation based on the empirical data is compared to the savings that would have been achieved if the recommendation model would have been applied.

Figure 5.7 highlights that the recommendations for each group would increase the expected savings potential. This is especially prominent in the case with a low number of bidders, where the recommendations increase the savings significantly, sometimes even double, for example in the case of low number of bidders, low price dispersion and low attractivity. In total the recommendations of Eichstädt [20] Model perform a 0.31% increase on the total expected savings.

Figure 5.8 shows the Schulze-Horn et al. [84] model performance. While in the case of a low number of bidders, the savings are identical to Eichstädt [20], the savings for a high price dispersion and high number of bidders differ. Schulze-Horn et al. [84] recommendations have minimally lower expected savings for the scenario with low attractivity and a significantly higher expected savings in comparison to both as-is and Eichstädt [20]. In total the recommendations improve the as-is situation by 0.38% and achieve 0.07% higher savings than Eichstädt [20] recommendations.

Figure 5.9 applies the model of Berz et al. [6] and it shows that for the row of high variability in the initial bids, Berz et al. [6] is able to significantly increase the expected savings except in the case of a high number of bidders and a low attractivity, for which the as-is situation has a higher average saving than the recommendation. This is also the case for the low attractivity scenario when there are a low amount of bidders and they have close bids, as well as for the high attractivity case when there are many bidders and close bid spreads. In total the recommendation improve the as-is by 0.21% and are lower than both Eichstädt [20] and Schulze-Horn et al. [84] by 0.10% and 0.17% respectively.



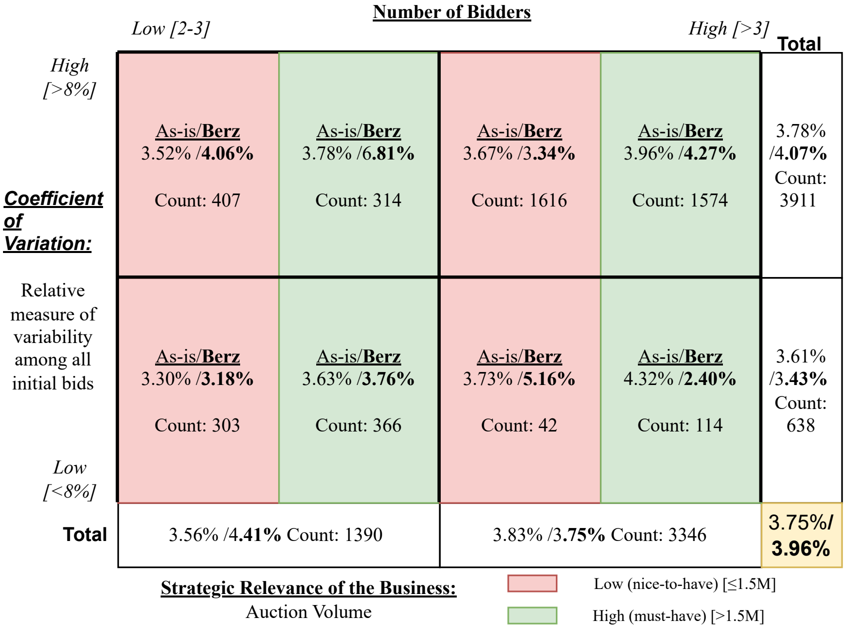
	Number of Bidders				
Price Dispersion: Percentage Difference Between Best and Second Best Bid	*Low [2-3]*		*High [>3]*		**Total**
	🟥	🟩	🟥	🟩	
High [>3%]	As-is/**Schulze** 3.74% /**4.27%** Count: 550	As-is/**Schulze** 3.97% /**5.64%** Count: 480	As-is/**Schulze** 3.31% /**3.10%** Count: 1345	As-is/**Schulze** 3.61% /**4.08%** Count: 1266	3.56% /**3.95%** Count: 3641
Low [<3%]	As-is/**Schulze** 2.33% /**4.58%** Count: 160	As-is/**Schulze** 3.05% /**3.63%** Count: 200	As-is/**Schulze** 5.23% /**5.61%** Count: 313	As-is/**Schulze** 5.12% /**5.39%** Count: 422	4.36% /**5.01%** Count: 1051
Total	3.56% /**4.47%** Count: 1390		3.83% /**3.99%** Count: 3346		3.75%/ **4.13%**

Strategic Relevance of the Business:
Auction Volume

🟥 Low (nice-to-have) [≤1.5M]

🟩 High (must-have) [>1.5M]

FIGURE 5.8: Expected Savings calculation based on the application of Schulze-Horn et al. [84] Model. The as-is situation based on the empirical data is compared to the savings that would have been achieved if the recommendation model would have been applied.



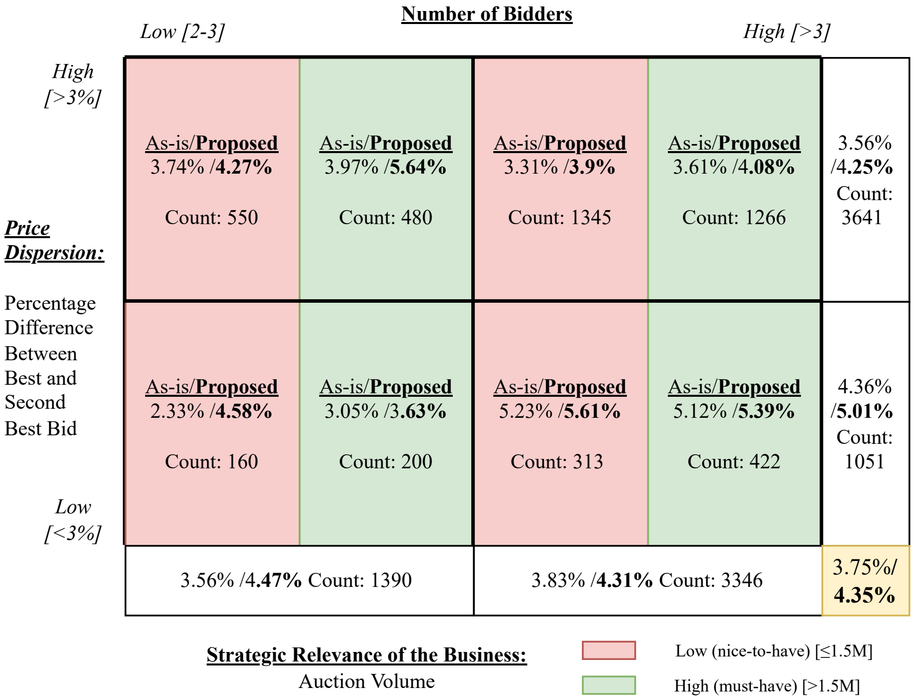
	Number of Bidders				
Coefficient of Variation: Relative measure of variability among all initial bids	*Low [2-3]*		*High [>3]*		**Total**
	🟥	🟩	🟥	🟩	
High [>8%]	As-is/**Berz** 3.52% /**4.06%** Count: 407	As-is/**Berz** 3.78% /**6.81%** Count: 314	As-is/**Berz** 3.67% /**3.34%** Count: 1616	As-is/**Berz** 3.96% /**4.27%** Count: 1574	3.78% /**4.07%** Count: 3911
Low [<8%]	As-is/**Berz** 3.30% /**3.18%** Count: 303	As-is/**Berz** 3.63% /**3.76%** Count: 366	As-is/**Berz** 3.73% /**5.16%** Count: 42	As-is/**Berz** 4.32% /**2.40%** Count: 114	3.61% /**3.43%** Count: 638
Total	3.56% /**4.41%** Count: 1390		3.83% /**3.75%** Count: 3346		3.75%/ **3.96%**

Strategic Relevance of the Business:
Auction Volume

🟥 Low (nice-to-have) [≤1.5M]

🟩 High (must-have) [>1.5M]

FIGURE 5.9: Expected Savings calculation based on the application of Berz et al. [6] Model. The as-is situation based on the empirical data is compared to the savings that would have been achieved if the recommendation model would have been applied.



Number of Bidders

Price Dispersion: Percentage Difference Between Best and Second Best Bid	*Low [2-3]*		*High [>3]*		
High [>3%]	As-is/**Proposed** 3.74% /**4.27%** Count: 550	As-is/**Proposed** 3.97% /**5.64%** Count: 480	As-is/**Proposed** 3.31% /**3.9%** Count: 1345	As-is/**Proposed** 3.61% /**4.08%** Count: 1266	3.56% /**4.25%** Count: 3641
Low [<3%]	As-is/**Proposed** 2.33% /**4.58%** Count: 160	As-is/**Proposed** 3.05% /**3.63%** Count: 200	As-is/**Proposed** 5.23% /**5.61%** Count: 313	As-is/**Proposed** 5.12% /**5.39%** Count: 422	4.36% /**5.01%** Count: 1051
	3.56% /**4.47%** Count: 1390		3.83% /**4.31%** Count: 3346		3.75%/ **4.35%**

Strategic Relevance of the Business:
Auction Volume

🟥 Low (nice-to-have) [≤1.5M]

🟩 High (must-have) [>1.5M]

FIGURE 5.10: Expected Savings calculation based on the application of our proposed model **??** Model. The as-is situation based on the empirical data is compared to the savings that would have been achieved if the recommendation model would have been applied.

Based on our proposed model in Figure 5.10, it is observed that in all cases the expected savings are higher than the as-is situation and while the low number of bidder cases are again identical to Schulze-Horn et al. [84], proposing the Japanese auction type in the case of a high number of bidders, a large price dispersion and a low attractivity, yields higher expected savings by 0.80% in this group, impacting the overall expected result to improve upon the as-is situation by 0.60% and the best model so far from Schulze-Horn et al. [84] by 0.22%.

5.1.6 Answers to SRQ2 and Conclusions

The empirical analysis and test of theoretical hypotheses conclude that the three dimensions of number of bidders, price dispersion, and the attractivity of the business to the supplier are relevant determinants for the auction design choice, particularly between the choice of a first-price auction type and a second-price auction type, as is prominent in Figure 5.4. Additionally, the auction volume appears to be an indicator of the attractivity of the business to the supplier, with which it is possible to observe differences in the expected savings of an auction following theoretical expectations. Still, it is a limited indicator for the attractivity of the business, as shown by the in most situations only marginal difference observed between the ‘low’ and ‘high’ categorization, for example in Figure 5.5. Considering a measure that accounts for the individual calculation of the strategic relevance of the business, including information about the possible utilization of production capacities, current market position, and market-specific competitive situation for every supplier, could be relevant improvements for future work.

For the price dispersion, both the difference between the two leading initial bids (see Figure 5.4)as well as the price dispersion across all initial bids (see Figure 5.5) are suitable indicators of the dispersion and yield comparable results. The price dispersion between the two best initial bids highlights the greater performance of first-price auctions for higher dispersion more compared to the price dispersion through the coefficient of variation. This observation might hint that a difference between the best two bids could be a more relevant decision criteria for the choice of a first-price or second price auction rather than the spread of all initial bids.

The differences of the resulting expected average savings between the auction types are mostly in line with the proposed theoretical recommendations. Different from the models based on theoretical analysis, the empirical observations, as shown in Table 5.1, also see English auctions in a setting with a low number of bidders as the best auction choice if the price dispersion is low, assuming that the bidders participating in the English auction have no knowledge about how many other bidders are participating.

Additionally, much of the game's theoretical reasoning on the effects and behavior of bidders in first-price auctions depends on the strict commitment of the buyer to allocate the business to the winning bidder. Although this assumption is not fully met by the empirical data, the trends and expected main benefits of first-price auctions can still be observed, for example in Figure 5.5. This might indicate that even a less strong commitment to the allocation of the business to the winner of the auction results in similar patterns of behavior. On the other hand, potentially establishing this strong commitment could yield higher savings for the first-price auctions than the currently observed performance. Another limitation is that only single-phase auctions and only three auction modes were analyzed, which could not be directly compared with some multi-phase auction recommendations of the theoretical models. A more holistic analysis of multiple-phase auctions and the specific bidding behavior in online auctions can further enhance the comparability and understanding of game theoretical recommendations in industrial practice.

The evaluation of the recommendations models represented in Figure 5.7, 5.8, 5.9, and 5.6 show that expected savings increase, if the models would have been applied compared to the as-is situation. The model of Schulze-Horn et al. [84] achieves the highest savings and provides the most specific operationalization of dimensions for practical use. Only our model achieves higher savings increases of 0.60% from the as-is situation by differentiating the case of a large amount of participants and a large spread to recommend the Japanese Ticker auction if the attractivity of the business is lower. All in all, the evaluation stresses the importance of the auction type decision on the expected savings, as following the recommendation models can yield significant increases from 0.21% to 0.60% in overall average savings.

The evaluation is limited by the assumption that the expected savings achieved by conducting all auctions in a group according to one type would approach expected savings of this subgroup based on the empirical analysis. Moreover, as some of the recommended auction types, such as the Hong Kong auction by Berz et al. [6]), are not available in the empirical data, they have been assumed to behave similar to one of the closest available auction types, while not being identical in their mechanism. This limits the generalizability and conclusion of the results to the performance of the specific auction recommendations in these model, but nevertheless it gives an indication of the suitability of first or second-price auction types.

5.2 SRQ3 & 4: Evaluation of the Online Negotiation Auction Bot

Table 5.4 shows the results of the performance evaluation for the experiments outlined in section 4.4. In total twelve different configurations of LLM, the available document set and the use of chain of thought prompting combinations have been evaluated. The best performance for each performance measure is highlighted in bold.

The results for the **context relevance** indicate that the performance decreases significantly across all LLMs if the documentation set is large, compared to the small documentation set. This observation is expected, as a larger pool of possible contexts increases the chance to retrieve context that are adjacent but not fully relevant to the question at hand. Interestingly, the use of chain of thought prompting increases the context relevance significantly for the more capable GPT3.5 and GPT4 models and less significant for the smaller Nous-Hermes2 model. The possible increase in relevant and better formulated question as a result of the chain of thought prompting seems to impact the relevance of the retrieved content towards the original question.The difference in performance between the different models seem to be minimal, especially between GPT3.5 and GPT4.

The **answer relevance** shows no clear tendency except for an increase in performance if the Chain of Thought prompt engineering technique is utilized. The different document set seem to have limited impact on the answer relevance, but the GPT4 model outperforms both others with a wide margin in the configuration with the large document set and the chain of thought prompting.

The **Faithfulness** metric shows the general trend that it is more difficult for the generated answer to remain faithful to the context, if the context is wider and more varied. Additionally, similar trends between the performances of the LLM models can be observed as before, as GPT4 tops the other models, followed closely by GPT3.5 and with some more distance Nous-Hermes2. Again chain of though prompting also shows its positive impact on the performance on the faithfulness of the answer to the context.

Data Accuracy performance measures are low across the board with the best model configuration only solving 65% of all tasks in the evaluation test set correctly. While GPT4 and GPT3.5 are again close competitors, the Nous-Hermes2 model only achieves a maximum performance of almost half the best with 35%. Chain of Thought prompting seems to be a large influence on the performance of the data retrieval tasks for all models but especially strong for the high parameter count GPT models. The available textual context has almost no influence on the data accuracy metric.

The **reasoning accuracy** paints a stark divide between the GPT models and the small local Nous-Hermes2 model. The dominance of the larger parameter models in the reasoning process shows even between GPT3.5 and GPT4. Moreover, chain of thought prompting is a major influence on the reasoning capabilities for all models and independent of the document set available. The available document set does not seem to have a major influence

Experimental Setup			Performance Measures					
LLM	Document Set	Chain of Thought	Context Relevance	Answer Relevance	Faithfulness	Data Accuracy	Reasoning Accuracy	Recommendation Accuracy
gpt-3.5-turbo	Small_R	No	35.87	79.44	80.82	40	63.33	70
gpt-3.5-turbo	Small_R	Yes	45.09	81.65	84.77	50	73.33	76.66
gpt-3.5-turbo	Large_RET	No	15.34	73.56	73.60	40	56.66	66.66
gpt-3.5-turbo	Large_RET	Yes	19.68	81.90	76.68	50	73.33	73.33
Nous-Hermes2	Small_R	No	30.54	70.88	70.43	30	26.66	40
Nous-Hermes2	Small_R	Yes	32.66	73.10	72.34	35	33.33	50
Nous-Hermes2	Large_RET	No	12.22	67.22	66.60	25	30	33.33
Nous-Hermes2	Large_RET	Yes	14.05	69.56	68.23	30	33.33	40
gpt4-turbo	Small_R	No	42.01	83.60	83.45	55	73.33	83.33
gpt4-turbo	Small_R	Yes	**47.94**	88.08	**86.54**	**65**	80	86.66
gpt4-turbo	Large_RET	No	14.67	83.44	71.60	55	73.33	83.33
gpt4-turbo	Large_RET	Yes	23.33	**89.50**	77.08	**65**	**83.33**	**90**

Table 5.4: Results of the experiments and evaluation metrics for different configurations regarding the LLM, the document set and whether chain of thought prompting is used.

if chain of thought prompting is missing. The best performance is achieved by combining the largest model with the largest document set and the chain of thought prompting. The **recommendation accuracy** follows similar trends compared to the reasoning accuracy, but it has higher values in every configuration, indicating that there are some correctly predicted auction design recommendations that have a unsatisfactory line of reasoning. At the end the best performer also in most other metrics is the configuration with GPT4 as the LLM, a large document set with a higher variety of information and chain of thought prompting active.

5.2.1 Answers to SRQ3 & SRQ4 and Conclusions

All in all, the experiments and evaluations showcased that the kind and volume of information supplied to the large language model impacts the retrieval performance and the quality of the final output specific to the task of auction design recommendations. A smaller corpus that is concentrated around one specific topic related to the specific task at hand seems to yield better performances in the relevance of the retrieved context, and the faithfulness of the generated answer to the context. In case of a very large model like GPT4, a larger volume of documents and more variety in the discussed topics can yield to higher performance gains especially with regards to the relevance of the answer, but also on the correctness of the final auction recommendation and the the accuracy of the reasoning process. It seems that the utilization of a larger volume of documents depends on the capability of the LLM model to meaningfully use the context.

On the other hand, chain of thought prompting is an essential part to any RAG system, as it has shown to impact the performances for all performance metric in a major positive way. Chain of thought prompting helps in increasing the relevance of the retrieved context by asking more precise search questions, increases the answer relevance by decomposing complex problems and approaching sub problems first, and elevates the faithfulness of the generated answer to the context. In combination with an strong increase in the accuracy of recommending the correct auction type and providing a logical and correct reasoning, it is a fundamental part and highlights the strong impact of prompt engineering on the final results.

Chapter 6

Conclusion & Limitations

In this last chapter, the results of the thesis are concluded in section 6.1, while section 6.2 highlights the contributions of the research to science and practice. Finally, the limitations of the research and next steps for future work are discussed in section 6.3.

6.1 Conclusion

This research designed, implemented and evaluated a multi-agent conversational chatbot system powered by Large Language Models to facilitate the creation of auctions for purchasing professionals and explore the ability of such a system to generate meaningful recommendations alongside logical reasoning paths.

As a first step, an **empirical statistical analysis** on online e-reverse auction data partly confirmed existing theoretical models. The study confirms that the number of bidders, price dispersion, and supplier interest in the auction are crucial factors for choosing between first-price and second-price auctions. First-price auctions perform best when the price dispersion is high. The study finds that both measures of price dispersion are useful indicators, with the difference between the top bids potentially being a more relevant decision factor than overall spread. English auctions perform well in most cases and can ,in contrast to the theoretical hypothesis, be effective for few bidders and low dispersion (assuming blind bidding). Clear practical operationalization of the three dimensions are outlined and while auction volume can indicate supplier interest, it is a limited measure. More comprehensive metrics that consider individual supplier calculations are recommended for future research. Finally, the study focused on single-phase auctions and limited auction types, restricting comparisons with multi-phase auction recommendations from theory. A broader analysis incorporating multi-phase auctions and online bidding behavior could improve the real-world applicability of these theoretical recommendations.

Insights from the first step flowed into the knowledge provided to the **multi-agent retrieval augemented chatbot** system along with other textual and numerical data. The created system is able to plan, take action and execute steps necessary to come to a final verdict for a range of questions with regards to auction design with a specific focus on optimal auction design recommendation. The context retriever, and data retriever collaboratively work together, with the orchestrator as the central coordinating unit. In combination with prompt engineering frameworks such as ReAct, the system is capable of holding conversations with a user and enacting upon the retrieval agents available.

The system is evaluated based on a custom evaluation dataset **AuctionEval** to asses the influence of the amount and type of information supplied. Additionally, the impact of prompt engineering techniques on the systems performance is also evaluated using the same custom dataset. The experiments and evaluations indicated that the type and quantity of information provided to a large language model significantly influence its performance in retrieving relevant information and generating the auction design recommendations. Smaller, focused text corpora centered around the specific task tend to result in better performance, showing improved relevance of retrieved context and faithfulness of generated answers. However, for very large state-of-the-art models like GPT-4, a larger corpus with diverse topics can lead to even greater performance gains, particularly in answer relevance, recommendation accuracy, and reasoning precision. The effectiveness of utilizing a larger volume of documents depends on the model's ability to extract meaningful context.

Additionally, the incorporation of chain of thought prompting is crucial for the system, as it consistently improves performance across various metrics. Chain of thought prompting enhances relevance by asking precise search questions, improves answer quality by breaking down complex problems into manageable sub-problems, and ensures faithfulness of the answer to the context. In combination with an increase in the accuracy of recommended auction types and the ability to provide logical reasoning traces, prompt engineering emerges as a fundamental aspect, underscoring its significant impact on final results.

6.2 Contribution to Science & Practice

The empirically tested auction design recommendation models, along with the newly proposed model, fills the gap in the body of auction literature on missing validations of models and the few existing research with robust empirical evidence. The development and evaluation of the multi-agent conversational system investigated the reasoning capabilities of LLM's in the under-explored area of game theory and mechanism design, for the specific use case of purchasing auction design. It highlighted the ability of larger models to perform well on game theoretical reasoning tasks including the use of general and specific covariance based logic, while still being limited in more complex and specific data retrieval tasks. The agent LLM architecture including retrieval augmentation and multiple agents with specialized tasks, highlights the increased capability and robustness of multi-agent LLM systems to plan and act in steps to solve complex problems across domains.

From a practical point of view, the analysis and proposal of an empirically backed recommendation model provides purchasing professionals with a practice-relevant recommendation and guidance on the specific operationalization and use of such models in real-world negotiations. The conversational auction chat-bot promises to be a flexible, robust and performing application for the automation and facilitation of the auction design procedure in real-world commercial environments. The adaptability of the conversational chatbot with its innovative use of multi-agents LLM's make it a versatile application for various tasks in daily purchasing practice. Quick adaptability to new information and guidelines, the flexibility to solve a range of tasks and not be limited, and the ability to reason and discuss recommendations made, make the system stand out from traditional approaches. Additionally, the thesis outlines an example evaluation procedure, which, through the AuctionEval dataset, provides a scalable method for assessing the effectiveness of possible configurations and variations of the system. It provides the ground for future expansions on the evaluation of LLM systems on practical use cases.

6.3 Limitations & Future Work

The research conducted faces several limitations. For the **statistical analysis**, only one phase auctions were available in the empirical data, limiting the ability to asses the impact of two-phase auction designs. Additionally, the not exact matching of auction designs in the empirical data compared to the proposed designs in theoretical models limits the validation of the expected achieved savings. Future work could try to expand the analysis to multi-phase auction designs and involve richer information about the auctions and especially bidder behaviour. The assumptions made by the recommendations model and auction theory on the expected behaviour could be investigated, yielding valuable contributions to the literature.

Furthermore, the **comprehensiveness of the data** and other data quality issues limited the extension of the analysis to include other factors such as the impact of overtime or reserve prices decisions on the auction outcome. Therefore, also the approximation of the attractivity of an auction or risk aversion of a bidder by the auction volume is limited, as the attractivity of an auction to a business incorporates more consideration than only the size of an auction. Future work can establish a more complex metric for the attractivity of the auction to a participant, by for example considering the current production capacity of the participant or incorporating relevant market data into the metric.

The **multi-agent conversational chatbot** faces limitations especially in data retrieval tasks, even in relatively simple use cases. On the other hand, the system showed great capabilities of performing well on the auction design recommendation task with a focus on the reasoning behind the recommendation, although at the cost of computation time. The three agent design including the ability to correct for mistakes and reuse the agents, enables a robust system that can handle missing data or more complex questions, but requires a higher computation time for solving the questions. Furthermore, the evaluation of the multi-agent retrieval system is limited by the amount and variety of LLM's and prompt engineering techniques tested. Moreover, the small size of the dataset of 50 total questions limits the generalizability and robustness of the evaluation. Therefore, future work needs to expand the scope of the evaluation including the experimental configurations, and the size and variety of questions in the dataset to increase the robustness of the evaluation and the level of insight. The addition of more open-sourced LLM models or pretrained LLM's with the evaluation on different prompt techniques like few shot prompting or tree of thought prompting enable a more sophisticated analysis. Adapting a new data retrieval strategy or even agent architecture to increase the performance while maintaining or decreasing the computation time is a future step.

Bibliography

[1] Gpt-3: Ai game-changer or environmental disaster? *Communications of the ACM*, 2020. Accessed: 2023-09-25. URL: https://cacm.acm.org/news/246586-gpt-3-ai-game-changer-or-environmental-disaster/fulltext#:~:text=The%20GPT,3.

[2] J Aloysius, C Deck, L Hao, and R French. An experimental investigation of procurement auctions with asymmetric sellers. *PRODUCTION AND OPERATIONS MANAGEMENT*, 25:1763–1777, 2016. doi:10.1111/poms.12576.

[3] Yuvanesh Anand, Zach Nussbaum, Brandon Duderstadt, Benjamin Schmidt, and Andriy Mulyar. Gpt4all: Training an assistant-style chatbot with large scale data distillation from gpt-3.5-turbo. https://github.com/nomic-ai/gpt4all, 2023.

[4] Laura H Baldwin, Robert C Marshall, and Jean-Francois Richard. Bidder collusion at forest service timber sales. *Journal of Political Economy*, 105(4):657–699, 1997.

[5] Gregor Berz. *Game theory bargaining and auction strategies: Practical examples from internet auctions to investment banking, second edition*. 2016. doi:10.1057/9781137475428.

[6] Gregor Berz, Florian Rupp, and Brian Sieben. How the 'auction cube' supports the selection of auction designs in industrial procurement. 2021.

[7] Carsten Block and Dirk Neumann. *A Decision Support System for Choosing Market Mechanisms in e-Procurement*. 2008. doi:10.1007/978-3-540-77554-6_3.

[8] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR, 2022.

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[10] Qingpeng Cai, Aris Filos-Ratsikas, Pingzhong Tang, and Yiwei Zhang. Reinforcement mechanism design for e-commerce. In *Proceedings of the 2018 World Wide Web Conference*, pages 1339–1348, 2018.

[11] Craig R. Carter and Cynthia Kay Stevens. Electronic reverse auction configuration and its impact on buyer price and supplier perceptions of opportunism: A laboratory experiment. *Journal of Operations Management*, 25(5):1035–1054, 2007. URL: https://onlinelibrary.wiley.com/doi/abs/10.1016/j.jom.2006.10.005, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1016/j.jom.2006.10.005, doi:10.1016/j.jom.2006.10.005.

[12] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*, 2023.

[13] Qian Chen, Xuan Wang, Zoe Lin Jiang, Yulin Wu, Huale Li, Lei Cui, and Xiaozhen Sun. Breaking the traditional: a survey of algorithmic mechanism design applied to economic and complex environments. *Neural Computing and Applications*, pages 1–30, 2023.

[14] C.-H. Chen-Ritzo, T P Harrison, A M Kwasnica, and D J Thomas. Better, faster, cheaper: An experimental analysis of a multiattribute reverse auction mechanism with restricted information feedback. *Manage Sci*, 51:1753–1762, 2005. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-29144524547&doi=10.1287%2fmnsc.1050.0433&partnerID=40&md5=238ff59f0699c3e1e9ade7a6cf0403e6, doi:10.1287/mnsc.1050.0433.

[15] C.-H. Chen-Ritzo, T P Harrison, A M Kwasnica, and D J Thomas. Better, faster, cheaper: An experimental analysis of a multiattribute reverse auction mechanism with restricted information feedback. *Manage Sci*, 51:1753–1762, 2005. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-29144524547&doi=10.1287%2fmnsc.1050.0433&partnerID=40&md5=238ff59f0699c3e1e9ade7a6cf0403e6, doi:10.1287/mnsc.1050.0433.

[16] Harrison H. Cheng and Guofu Tan. Asymmetric common-value auctions with applications to private-value auctions with resale. *Economic Theory*, 45(1/2):253–290, 2010. URL: http://www.jstor.org/stable/40864845.

[17] Oliver Curry. Bounded rationality: The adaptive toolbox edited by gerd gigerenzer and reinhard selten. (2002). 07 2023.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[19] Paul Dütting, Zhe Feng, Harikrishna Narasimhan, David C. Parkes, and Sai Srivatsa Ravindranath. Optimal auctions through deep learning: Advances in differentiable economics, 2022. arXiv:1706.03459.

[20] Tilman Eichstädt. *Einsatz von Auktionen im Beschaffungsmanagement: Erfahrungen aus der Einkaufspraxis und die Verbreitung auktionstheoretischer Konzepte*. 01 2008. doi:10.1007/978-3-8349-9744-9.

[21] W Elmaghraby. Auctions within e-sourcing events. *PRODUCTION AND OPERATIONS MANAGEMENT*, 16:409–422, 2007.

[22] Ulle Endriss, Ann Nowé, Maria Gini, Victor Lesser, Michael Luck, Ana Paiva, and Jaime Sichman. Autonomous agents and multiagent systems: perspectives on 20 years of aamas. *AI Matters*, 7(3):29–37, jan 2022. URL: https://doi-org.ezproxy2.utwente.nl/10.1145/3511322.3511329, doi:10.1145/3511322.3511329.

[23] R Engelbrecht-Wiggans, E Haruvy, and E Katok. A comparison of buyer-determined and price-based multiattribute mechanisms. *MARKETING SCIENCE*, 26:629–641, 2007. doi:10.1287/mksc.1070.0281.

[24] Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. Ragas: Automated evaluation of retrieval augmented generation, 2023. arXiv:2309.15217.

[25] Zhe Feng, Harikrishna Narasimhan, and David C. Parkes. Deep learning for revenue-optimal auctions with budgets. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '18, page 354–362, Richland, SC, 2018. International Foundation for Autonomous Agents and Multiagent Systems.

[26] Dinesh Garg, Yadati Narahari, and Sujit Gujar. Foundations of mechanism design: A tutorial part 1-key concepts and classical results. *Sadhana*, 33:83–130, 04 2008. doi:10.1007/s12046-008-0008-3.

[27] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges, 2024. arXiv:2402.01680.

[28] A Gupta, S T Parente, and P Sanyal. Competitive bidding for health insurance contracts: lessons from the online hmo auctions. *INTERNATIONAL JOURNAL OF HEALTH CARE FINANCE ECONOMICS*, 12:303–322, 2012. doi:10.1007/s10754-012-9118-x.

[29] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR, 2020.

[30] E Haruvy and E Katok. Increasing revenue by decreasing information in procurement auctions. *Prod. Oper. Manage.*, 22:19–35, 2013. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84872713617&doi=10.1111%2fj.1937-5956.2012.01356.x&partnerID=40&md5=b73582ec2fb51c7a59dcc08db4fd075d, doi:10.1111/j.1937-5956.2012.01356.x.

[31] E Haruvy, PTLP Leszczyc, O Carare, J C Cox, E A Greenleaf, W Jank, S Jap, Y H Park, and M H Rothkopf. Competition between auctions. *MARKETING LETTERS*, 19:431–448, 2008. doi:10.1007/s11002-008-9037-2.

[32] Anne-Wil Harzing and Satu Alakangas. Google scholar, scopus and the web of science: a longitudinal and cross-disciplinary comparison. *Scientometrics*, 106:787–804, 2016.

[33] Fang Huang. *Data Cleansing*, pages 1–4. Springer International Publishing, Cham, 2019. doi:10.1007/978-3-319-32001-4_300-1.

[34] Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. Challenges and applications of large language models. *arXiv preprint arXiv:2307.10169*, 2023.

[35] Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*, pages 99–127. World Scientific, 2013.

[36] O Karabag and B Tan. An empirical analysis of the main drivers affecting the buyer surplus in e-auctions. *INTERNATIONAL JOURNAL OF PRODUCTION RESEARCH*, 57:3435–3465, 2019. doi:10.1080/00207543.2018.1536835.

[37] Michał P. Karpowicz. Designing auctions: A historical perspective. *Journal of telecommunications and information technology*, pages 114–122, 2011.

[38] John Kennes. Competitive auctions: Theory and application. *Contributions to Economic Analysis*, 275:145–168, 2006.

[39] Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050*, 2023.

[40] Barbara Kitchenham. *Kitchenham, B.: Guidelines for performing Systematic Literature Reviews in software engineering. EBSE Technical Report EBSE-2007-01*. 01 2007.

[41] Paul Klemperer. Auction theory: A guide to the literature. *Journal of economic surveys*, 13(3):227–286, 1999.

[42] Paul Klemperer. Auctions: theory and practice. 2018.

[43] Vijay Krishna. *Auction theory*. Academic press, 2009.

[44] Jacob Gorm Larsen. *A Practical Guide to E-auctions for Procurement: How to Maximize Impact with E-sourcing and E-negotiation*. Kogan Page Publishers, 2021.

[45] Howard Levene. Robust tests for equality of variances. *Contributions to probability and statistics*, pages 278–292, 1960.

[46] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[47] C Liang, Y Hong, P.-Y. Chen, and B B M Shao. The screening role of design parameters for service procurement auctions in online service outsourcing platforms. *Inf. Syst. Res.*, 33:1324–1343, 2022. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85150024671&doi=10.1287%2fisre.2022.1168&partnerID=40&md5=21dfff4f210c5746b8b1b34a3eeb134d, doi:10.1287/isre.2022.1168.

[48] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL: https://aclanthology.org/W04-1013.

[49] Xiangyu Liu, Chuan Yu, Zhilin Zhang, Zhenzhe Zheng, Yu Rong, Hongtao Lv, Da Huo, Yiqing Wang, Dagui Chen, Jian Xu, et al. Neural auction: End-to-end learning of auction mechanisms for e-commerce advertising. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 3354–3364, 2021.

[50] Sofia Lundberg. *Auction formats and award rules in Swedish procurement auctions*. CERUM, Umeå, 2005.

[51] Sofia Lundberg. Auction formats and award rules in swedish procurement auctions. *Rivista di Politica Economica*, 96:91–114, 01 2006.

[52] Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60, 1947.

[53] Yuning Mao, Yanru Qu, Yiqing Xie, Xiang Ren, and Jiawei Han. Multi-document summarization with maximal marginal relevance-guided reinforcement learning, 2020. arXiv:2010.00117.

[54] Eric Maskin and John Riley. Asymmetric auctions. *The review of economic studies*, 67(3):413–438, 2000.

[55] D Matthaus. Designing effective auctions for renewable energy support. *ENERGY POLICY*, 142:–, 2020. doi:10.1016/j.enpol.2020.111462.

[56] John McMillan. Selling spectrum rights. *Journal of Economic Perspectives*, 8(3):145–162, September 1994. URL: https://www.aeaweb.org/articles?id=10.1257/jep.8.3.145, doi:10.1257/jep.8.3.145.

[57] M Mediavilla, K Mendibil, and C Bernardos. Making the most of game theory in the supplier selection process for complex items. *PRODUCTION PLANNING CONTROL*, 32:845–860, 2021. doi:10.1080/09537287.2020.1773560.

[58] Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, et al. Augmented language models: a survey. *arXiv preprint arXiv:2302.07842*, 2023.

[59] Paul R Milgrom and Robert J Weber. A theory of auctions and competitive bidding. *Econometrica: Journal of the Econometric Society*, pages 1089–1122, 1982.

[60] Ido Millet, Diane H. Parente, John L. Fizel, and Ray R. Venkataraman. Metrics for managing online procurement auctions. *Interfaces*, 34(3):171 – 179, 2004. URL: http://ezproxy2.utwente.nl/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=pbh&AN=13643072&site=ehost-live.

[61] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey, 2024. arXiv:2402.06196.

[62] Sunil Mithas and Joni L. Jones. Do auction parameters affect buyer surplus in e-auctions for procurement? *Production and Operations Management*, 16:455–470, 2009.

[63] Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*, 2022.

[64] Roger B. Myerson. Optimal auction design. *Mathematics of Operations Research*, 6(1):58–73, 1981. arXiv:https://doi.org/10.1287/moor.6.1.58, doi:10.1287/moor.6.1.58.

[65] Roger B. Myerson. Game theory - analysis of conflict. 1991. URL: https://api.semanticscholar.org/CorpusID:27263747.

[66] Roger B Myerson. *Game theory*. Harvard university press, 2013.

[67] Noam Nisan. *Introduction to Mechanism Design (for Computer Scientists)*, page 209–242. Cambridge University Press, 2007. doi:10.1017/CBO9780511800481.011.

[68] NobelPrize.org. Press release: The prize in economic sciences 2020, 2020. Accessed on 2023-09-27. URL: https://www.nobelprize.org/prizes/economic-sciences/2020/press-release/.

[69] OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023. URL: https://api.semanticscholar.org/CorpusID:257532815.

[70] Jong Han Park, Jae Kyu Lee, and Hoong Chuin Lau. Bidder behaviors in repeated b2b procurement auctions. pages 145–152. Association for Computing Machinery, 2012. doi:10.1145/2346536.2346563.

[71] Cecilia Maria Patino and Juliana Carvalho Ferreira. Meeting the assumptions of statistical tests: an important and often forgotten step to reporting valid results. *Jornal Brasileiro de Pneumologia*, 44:353–353, 2018.

[72] Dawn H. Pearcy and Larry C. Giunipero. The impact of electronic reverse auctions on purchase price reduction and governance structure: an empirical investigation. *Int. J. Serv. Technol. Manag.*, 7:215–236, 2006.

[73] Neehar Peri, Michael Curry, Samuel Dooley, and John Dickerson. Preferencenet: Encoding human preferences in auction design with deep learning. *Advances in Neural Information Processing Systems*, 34:17532–17542, 2021.

[74] Martin Pesendorfer. A study of collusion in first-price auctions. *The Review of Economic Studies*, 67(3):381–411, 2000.

[75] Owen R Phillips, Dale J Menkhaus, and Kalyn T Coatney. Collusive practices in repeated english auctions: Experimental evidence on bidding rings. *American Economic Review*, 93(3):965–979, 2003.

[76] Marius-Constantin Popescu, Valentina Balas, Liliana Perescu-Popescu, and Nikos Mastorakis. Multilayer perceptron and neural networks. *WSEAS Transactions on Circuits and Systems*, 8, 07 2009.

[77] Zafaryab Rasool, Stefanus Kurniawan, Sherwin Balugo, Scott Barnett, Rajesh Vasa, Courtney Chesser, Benjamin M. Hampstead, Sylvie Belleville, Kon Mouzakis, and Alex Bahar-Fuchs. Evaluating llms on document-based qa: Exact answer selection and numerical extraction using cogtale dataset, 2024. arXiv:2311.07878.

[78] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL: https://arxiv.org/abs/1908.10084.

[79] Alvin E Roth. The economist as engineer: Game theory, experimentation, and computation as tools for design economics. *Econometrica*, 70(4):1341–1378, 2002.

[80] Michael H. Rothkopf, Thomas J. Teisberg, and Edward P. Kahn. Why are vickrey auctions rare? *Journal of Political Economy*, 98(1):94–109, 1990. URL: http://www.jstor.org/stable/2937643.

[81] Tim Roughgarden. Algorithmic game theory. *Commun. ACM*, 53(7):78–86, jul 2010. doi:10.1145/1785414.1785439.

[82] Tim Roughgarden and Inbal Talgam-Cohen. Approximately optimal mechanism design. *Annual Review of Economics*, 11:355 – 381, 2019. Cited by: 9; All Open Access, Green Open Access. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85071615535&doi=10.1146%2fannurev-economics-080218-025607&partnerID=40&md5=a765b0b64489b4cbba3a7bbd35a709cb, doi:10.1146/annurev-economics-080218-025607.

[83] Graeme D Ruxton. The unequal variance t-test is an underused alternative to student's t-test and the mann–whitney u test. *Behavioral Ecology*, 17(4):688–690, 2006.

[84] Ines Schulze Horn. Mechanism design theory in buyer-supplier negotiations: Designing incentive systems to achieve price reductions. 2019.

[85] Ines Schulze-Horn, Sabrina Hueren, Paul Scheffler, and Holger Schiele. Artificial intelligence in purchasing: Facilitating mechanism design-based negotiations. *Applied Artificial Intelligence*, 34(8):618–642, 2020. arXiv:https://doi.org/10.1080/08839514.2020.1749337, doi:10.1080/08839514.2020.1749337.

[86] Loay M. Sehwail, Ricki G. Ingalls, and David B. Pratt. Business-to-business online reverse auctions: A literature review and a call for research. *International Journal of Services and Operations Management*, 4(4):498 – 520, 2008. Cited by: 8. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-40749136635&doi=10.1504%2fIJSOM.2008.017432&partnerID=40&md5=27844b6dbf2b01d439dcafb1ce1563a3, doi:10.1504/IJSOM.2008.017432.

[87] P Setia and C Speier-Pero. Reverse auctions to innovate procurement processes: Effects of bid information presentation design on a supplier's bidding outcome. *Decis. Sci.*, 46:333–366, 2015. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84928399727&doi=10.1111%2fdeci.12127&partnerID=40&md5=00308a244385780c5565c1cc56e763df, doi:10.1111/deci.12127.

[88] Samuel Sanford Shapiro and Martin B Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965.

[89] Weiran Shen, Pingzhong Tang, and Song Zuo. Computer-aided mechanism design: designing revenue-optimal mechanisms via neural networks. *CoRR*, abs/1805.03382, 2018. URL: http://arxiv.org/abs/1805.03382, arXiv:1805.03382.

[90] Raphael Stange, Holger Schiele, and Jörg Henseler. Advancing purchasing as a design science: Publication guidelines to shift towards more relevant purchasing research. *Journal of Purchasing and Supply Management*, 28(1):100750, 2022. URL: https://www.sciencedirect.com/science/article/pii/S147840922200005X, doi:10.1016/j.pursup.2022.100750.

[91] Moshe Tennenholtz. Tractable combinatorial auctions and b-matching. *Artificial Intelligence*, 140(1):231–243, 2002. URL: https://www.sciencedirect.com/science/article/pii/S0004370202002291, doi:10.1016/S0004-3702(02)00229-1.

[92] CS Timothy. Preventing collusion among firms in auctions. In *Auctioning Public Assets: Analysis and Alternatives*, pages 80–107. Cambridge University Press Cambridge, 2004.

[93] Hal R. Varian and Christopher Harris. The VCG Auction in Theory and Practice. *American Economic Review*, 104(5):442–445, May 2014. URL: https://ideas.repec.org/a/aea/aecrev/v104y2014i5p442-45.html.

[94] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[95] William Vickrey. Counterspeculation, auctions, and competitive sealed tenders. *The Journal of Finance*, 16(1):8–37, 1961. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.1961.tb02789.x, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-6261.1961.tb02789.x, doi:10.1111/j.1540-6261.1961.tb02789.x.

[96] Enrique Areyan Viqueira, Cyrus Cousins, Yasser Mohammad, and Amy Greenwald. Empirical mechanism design: Designing mechanisms from data. In Ryan P. Adams and Vibhav Gogate, editors, *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pages 1094–1104. PMLR, 22–25 Jul 2020. URL: https://proceedings.mlr.press/v115/viqueira20a.html.

[97] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Ji-Rong Wen. A survey on large language model based autonomous agents, 2023. arXiv:2308.11432.

[98] Elmar Wolfstetter. Auctions: an introduction. *Journal of economic surveys*, 10(4):367–420, 1996.

[99] J O Wooten, J M Donohue, T D Fry, and K M Whitcomb. To thine own self be true: Asymmetric information in procurement auctions. *Prod. Oper. Manage.*, 29:1679–1701, 2020. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85085102830&doi=10.1111%2fpoms.13174&partnerID=40&md5=12bedba2527a751b25fb0eadd7a99cc3, doi:10.1111/poms.13174.

[100] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang

Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wensen Cheng, Qi Zhang, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huang, and Tao Gui. The rise and potential of large language model based agents: A survey, 2023. arXiv:2309.07864.

[101] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models, 2023. arXiv:2210.03629.

[102] Marcel Zeelenberg. Anticipated regret, expected feedback and behavioral decision making. *Journal of behavioral decision making*, 12(2):93–106, 1999.

[103] J Zhang, J Tian, and X Gong. Study on the effect of information disclosure policy in multi-attribute reverse auctions. volume 2, pages 504–508, 2014. doi:10.1109/ISCID.2014.201.

[104] Zhanhao Zhang. A survey of online auction mechanism design using deep learning approaches, 2021. arXiv:2110.06880.

[105] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023.

Appendix A

Literature Review

A.1 Background

Auction and Mechanism Design Theory enjoys a long and extensive history of research, four of the seminal researchers in the field have even been awarded the Nobel Prize in Economics [13]. Hence also many literature reviews exist in the field of auction theory [41, 38, 98], mechanism design [26, 82], and some on the intersecting area of computer science and mechanism design, namely Algorithmic Game Theory and Algorithmic/Automated Mechanism Design [81, 13]. In the Algorithmic Game Theory field, a number of different categories of solution approaches exist due to the diverse backgrounds of researchers in this interdisciplinary field [89]. One of these categories uses AI approaches to search for the optimal mechanism and has seen a great increase in research interest [19] due to its adeptness to new situations and ability to approach optimal mechanisms even for numerically intractable situations like combinatorial auctions.

For the special intersection of AI and mechanism design, a review regarding optimal auction design through deep learning techniques is available by Zhang et al. [104]. They discuss the recent AI techniques used for determining the optimal auction mechanism design. Still, most approaches found use sampled data from an assumed distribution of valuations of bidders and are generally aiming to solve theoretical auction scenarios, rather than applying them to empirical real auction data. Research that uses Large Language Models to optimize auctions is neither mentioned in the review nor to be found in the literature to the best of our knowledge. But there exist many comprehensive and recent literature reviews on the general topic of large language models and its current application challenges [34], the discussion about their reasoning capabilities [39], their application as autonomous agents [97] and the open research challenge on their evaluation [12].

The proposed research lies at he intersection of mechanism design, the research field of autonomous agents, and the Machine Learning sub-discipline on Large Language Models. We have found extensive literature reviews on the topic of mechanism design, the intersection between autonomous agents and mechanism design, but no reviews on the intersection between all three topics with a focus on the practical application of such models. This literature review adds to the existing reviews by focusing on methods and solution approaches that take empirical data into account in the automated design of mechanisms via machine learning or statistical analysis. The following literature review aims to accomplish three main goals:

1 Identify gaps in the research on the design of reverse online auctions for procurement.

2 Explore the influences of the auction design scenario on the buyer's surplus in reverse e-auctions, to generate hypotheses that will be tested on the empirical data available.

3 Find approaches to solve the auction design problem for practice.

A.2 Research Questions

In order to achieve the aforementioned goal the following research questions for the literature review are defined:

LRQ1 How do auction design parameters influence the buyer surplus as observed by empirical online negotiation data?

LRQ2 What solution approaches to the design of optimal auctions in a procurement setting exist that utilize machine learning?

LRQ3 How could Large Language Models be used to generate auction design recommendations that are individualized and interpretable?

While LRQ1 aims to summarize existing knowledge of the empirical evidence of different auction design parameters on the buyer surplus, LRQ2 aims to highlight research gaps that exist in the soulution space for the optimal auction design in procurement. The following literature research aims to primarily solve these two research question. The third research question focuses on the possible use of large language models and open research gaps in the application of them on the mechanism design and auction design tasks. Due to the existence of many extensive and recent literature reviews about large language models, this question will be answered mainly using the existing reviews and surveys [34, 97, 39, 12].

In order to sharpen our scope, the specific auction design scenario we are facing needs to be explained in more detail. As the research interest lies in online reverse e-auctions in an industrial B2B setting, optimal auctions are not characterized by a social welfare function, but rather simply by the generated revenue of the auction. The higher the revenue for the auctioneer the more optimal the auction design. This higher revenue is also called buyer surplus and is the difference between the total price from the initial bids before the auction and the total price after the auction.

A.3 Search Strategy

The defined research question guides the search strategy and therefore informs the definition of the search string, inclusion and exclusion criteria, and on the assessment of the literature, as inspired by the methodology of Kitchenam et al. [40]. Table A.1 presents the chosen Scientific Databases for extracting the relevant articles and short reasoning for each. Scopus and Web of Science contain publications from a range of academic fields, major journals, and conferences. Both are considered to be the most comprehensive scientific databases [32] and ensure that the search covers a wide range of relevant literature from different disciplines. ACM and IEEE Explore give the search its special focus on the computer science field, specifically the search for machine learning optimization techniques and large language models. They are used to ensure that specific technical papers are not missed. Lastly, Business Source Elite is one of the most important databases for the field of Business Administration, giving the search coverage on articles in the purchasing and management field.

Source	Reasoning
Scopus	All major publications; covers wide range of literature
Web of Science	All major publications; covers wide range of literature
ACM	Computer Science and Computation Focus; retrieve technical implementation papers
IEEE Explore	Computer Science and Computation Focus; retrieve technical implementation papers
Business Source Elite	Business Administration Focus; coverage of purchasing and management papers

Table A.1: Scientific Databases used for the Literature Review

The search in these libraries is done on August 2, 2023 and found the following specified search string in the article title, abstract or keywords:

"reverse auction" AND "design" AND "procurement"

The search in these digital libraries itself did not contain any restrictions by publication year, source (i.e. journal or conferences), or the research domain (computer science, Business, or Engineering), to ensure that a wide variety of literature is captured and no important articles are missed. To guarantee that relevant synonyms/keywords were found and that the search result would not be too narrow nor too broad, an iterative learning process consistent of several experiments with different keywords and keyword combinations preceded the final selection of keywords.

A.4 Inclusion and Exclusion Criteria

The first step after the removal of duplicates is to screen the Title and Abstract. The following Inclusion criteria (IC) are the initial main pillars for the screening:

IC1 The paper directly relates to the topic of our review. This means, we include papers that explicitly analyze the effect of independent variables on the reverse auction outcome, especially regarding the final price/buyer surplus, or propose a decision-making model/machine learning model in which auction designs are recommended based on different negotiation scenarios.

IC2 The paper uses empirical research methods or laboratory experiments.

IC3 The paper addresses the research question

IC4 The paper is peer-reviewed.

IC5 The paper is available for download.

The second main pillar for the screening are the exclusion criteria (EC) and they ensure that only the truly relevant articles for our research question are filtered:

EC1 The paper deals with in-person auctions.

EC2 The paper talks about the procurement process as a whole and excludes reverse auctions.

EC3 The paper only tackles theoretical mechanism design problems and solves them analytically.

EC4 The paper deals with an application area that has little application to the procurement field, like crypto or energy markets.

EC5 The paper studies auction settings that are not relevant due to their simplicity.

EC6 The paper is not peer-reviewed.

As the literature review distinguishes itself by the focus on models and analysis that use empirical data, IC1, IC2, and EC3 are essential to retrieve the relevant papers. Also, we include laboratory experiments in IC2, as only a fraction of the papers use empirical data and they provide insights that are relevant to practice.

A.5 Snowballing

Looking at the citations for possible other research papers, no additional papers have been found that could be of relevance. Therefore no additional results through snowballing are reported. This could indicate that the search strategy was successful in capturing most of the relevant research papers.

A.6 Overview

Figure A.1 shows the extraction process in detail. First, the outlined databases have been queried with the search string and duplicates have been removed. Second, a Title and Abstract Screening filtered out or included papers that satisfy the mentioned Inclusion and Exclusion Criteria. Third, The papers that were left undergo a Full-text review for further inspection of the inclusion and exclusion criteria, as well as a closer look at the relevance to the research question.

After the whole process, 32 papers were left. Figure A.2 shows the distribuition of the focus of the papers from the literature review. Most papers, around 56 %, explore the impact of different auction design parameters on the buyer surplus for auctions, using statistical analysis to prove significant correlations. 25 % of the retrieved papers utilize machine learning techniques in the application of designing auctions automatically that achieve optimality, often with the aim to overcome computational and analytical boundaries of classical approaches. Papers that propose decision making models for the selection of the best practical auction design with the aim to guide praciticoners on which designs to choose among different scenarios, make up 24 % of the retrieved papers.

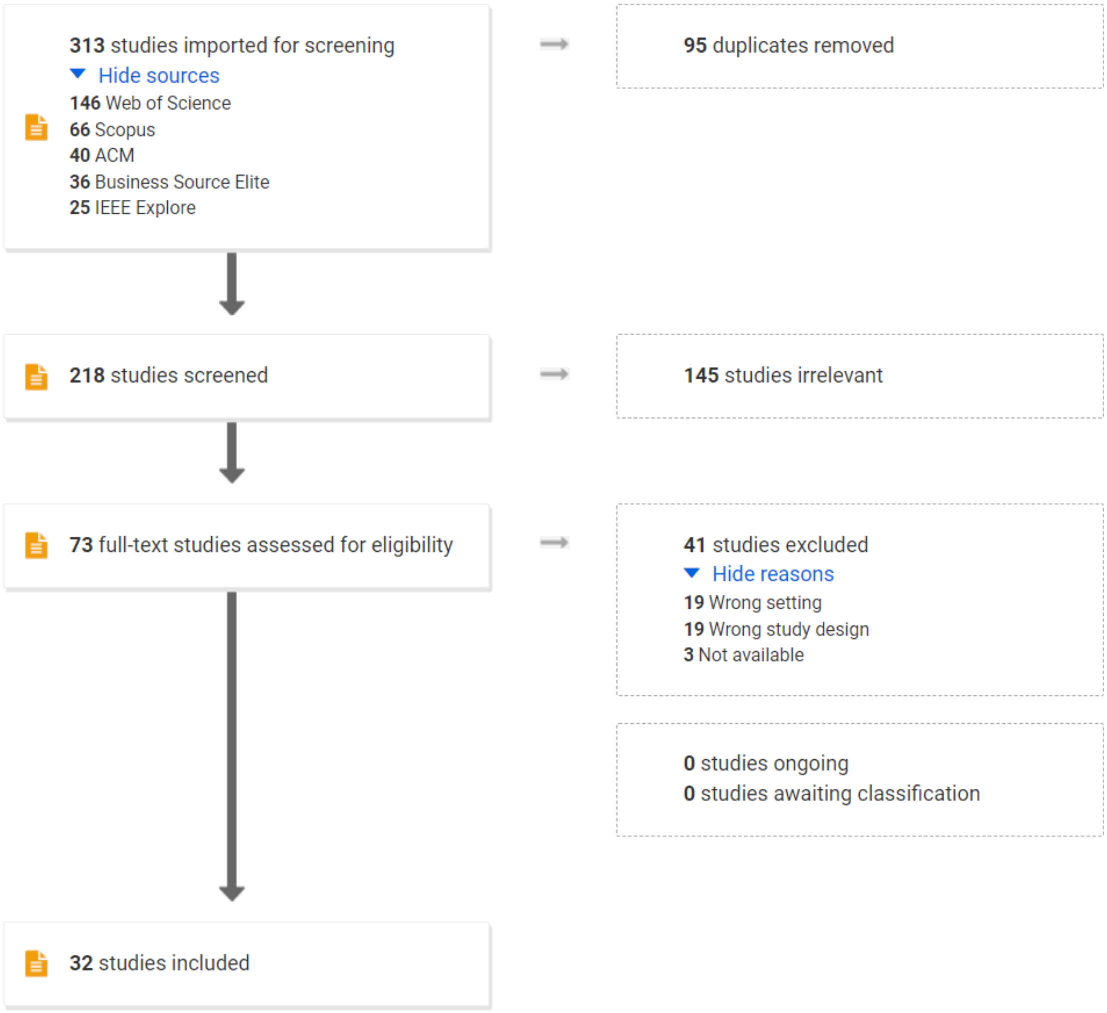


Figure A.1: Literature Review PRISMA Process Chart

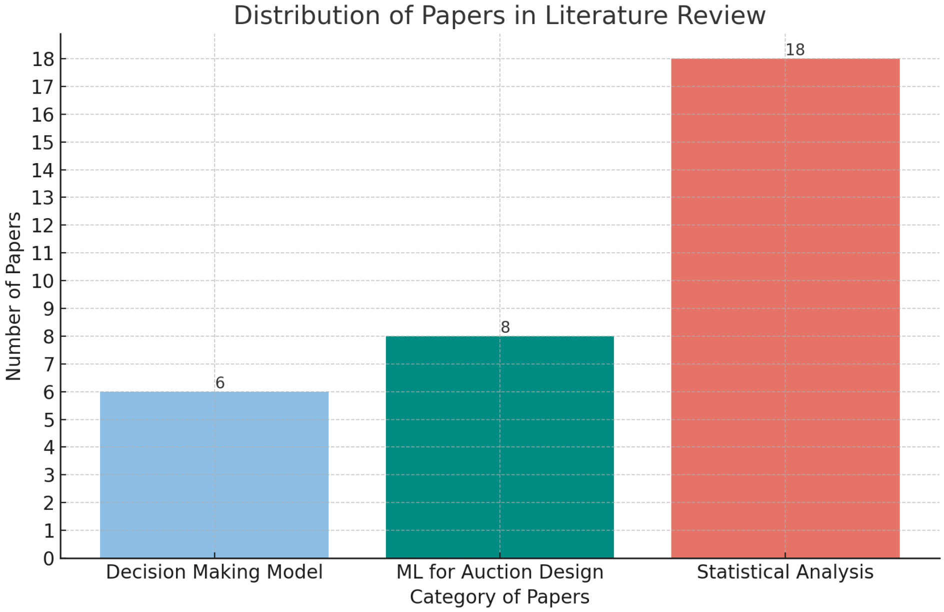
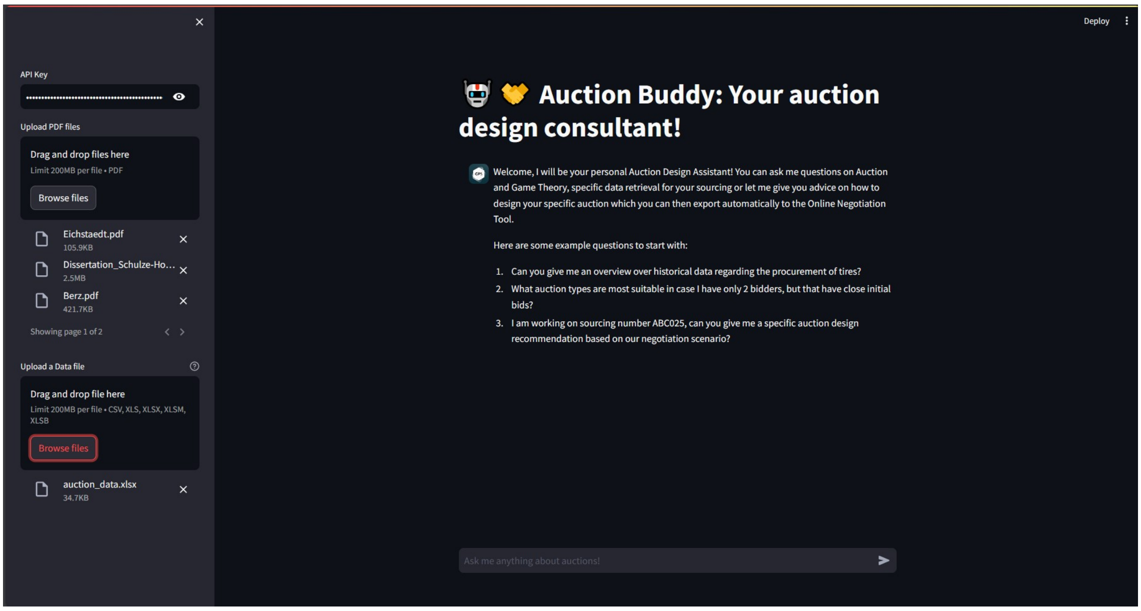


Figure A.2: Distribution of Papers dealing with decision making models, machine learning in auction design and the statistical analysis of auction design parameters on the buyer surplus

Appendix B

AuctionBuddy: Chatbot Application Interface and Use Cases



Auction Buddy: Your auction design consultant!

Welcome, I will be your personal Auction Design Assistant! You can ask me questions on Auction and Game Theory, specific data retrieval for your sourcing or let me give you advice on how to design your specific auction which you can then export automatically to the Online Negotiation Tool.

Here are some example questions to start with:

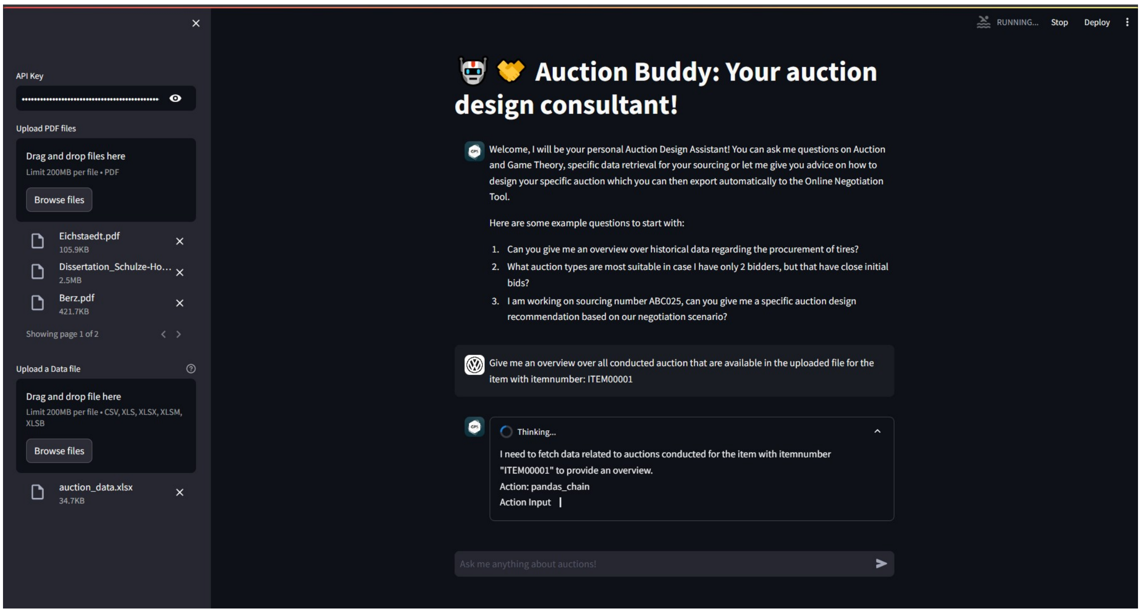
1. Can you give me an overview over historical data regarding the procurement of tires?
2. What auction types are most suitable in case I have only 2 bidders, but that have close initial bids?
3. I am working on sourcing number ABC025, can you give me a specific auction design recommendation based on our negotiation scenario?

Figure B.1: The User Interface of the AuctionBot

Figure B.1 showcases the user interface of the application. It represents a well known chatbot, with a messaging box at the bottom of the page in which the user can prompt the model and the the chat style question and answer visualization that encourages a conversational chain with the system. In the side panel, the user can upload text documents that he wants the system to be able to access, as well as numerical documents such as .csv files. In case a model with an API is used, the key can be inserted in this password field.

Different from well known ChatGPT, the AuctionBot shows you what he thinks and the steps he takes to solve your question. Figure B.2 shows the AuctionBot revealing his thinking appraoch and the steps he wants to take. The other agents also report back to this main orchestrator and provide their reasoning and outputs as well. Figure B.4 highlights the chain of actions and planning moments taken, that can be collapsed by clicking on it to view the content of that particular step along the action trajectory.

Figure B.4 presents the answer to the question asked in Figure B.2. As is observable, the system retrieves the correct data by invoking the data retrieval agent and prints the



Auction Buddy: Your auction design consultant!

Welcome, I will be your personal Auction Design Assistant! You can ask me questions on Auction and Game Theory, specific data retrieval for your sourcing or let me give you advice on how to design your specific auction which you can then export automatically to the Online Negotiation Tool.

Here are some example questions to start with:

1. Can you give me an overview over historical data regarding the procurement of tires?
2. What auction types are most suitable in case I have only 2 bidders, but that have close initial bids?
3. I am working on sourcing number ABC025, can you give me a specific auction design recommendation based on our negotiation scenario?

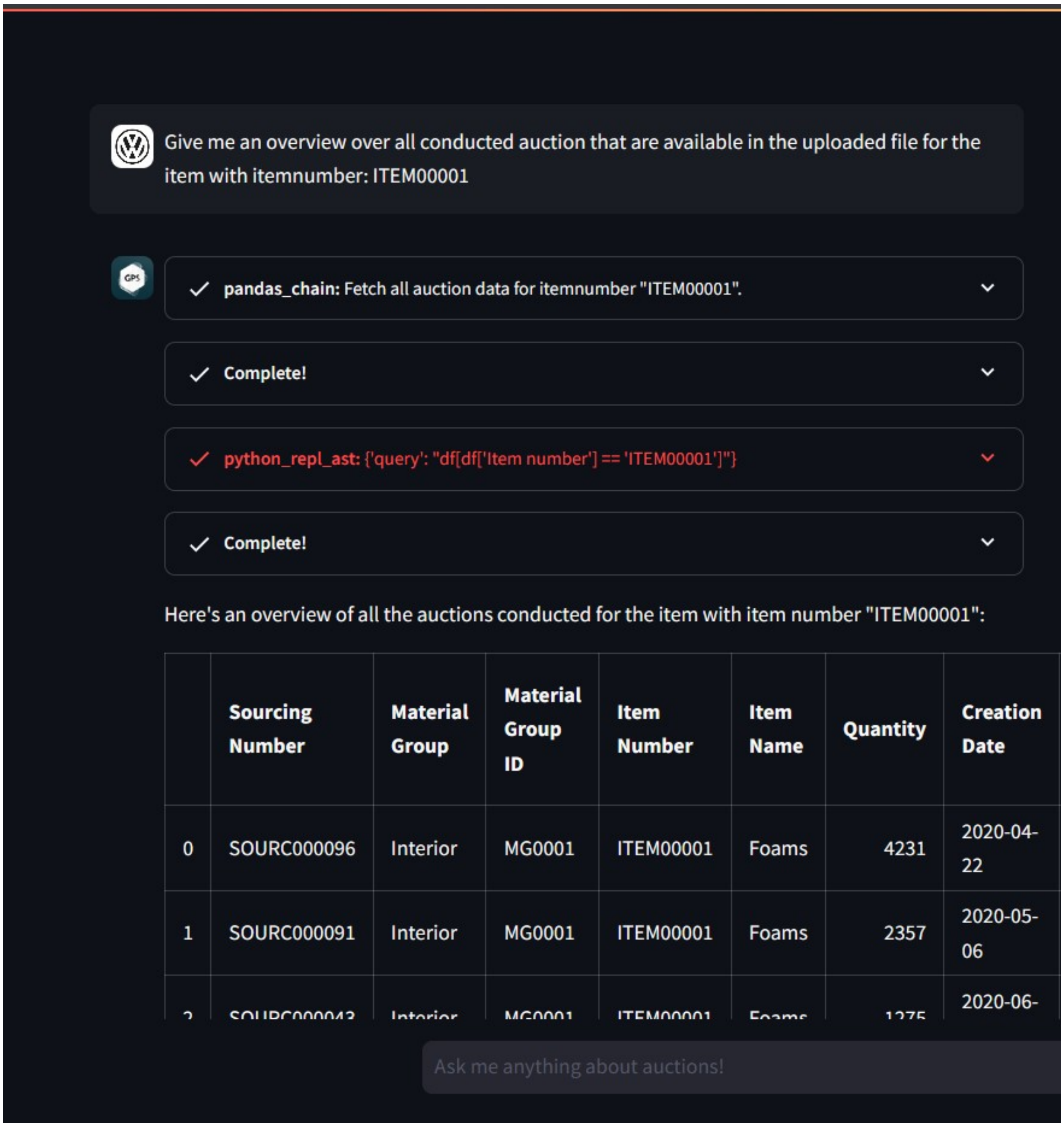
Give me an overview over all conducted auction that are available in the uploaded file for the item with itemnumber: ITEM00001

Thinking...

I need to fetch data related to auctions conducted for the item with itemnumber "ITEM00001" to provide an overview.
Action: pandas_chain
Action Input

Figure B.2: The User Interface of the AuctionBot

results to the web application for the user to see. Now the user can continue to converse with the AuctionBot.



Give me an overview over all conducted auction that are available in the uploaded file for the item with itemnumber: ITEM00001

✓ **pandas_chain:** Fetch all auction data for itemnumber "ITEM00001".

✓ **Complete!**

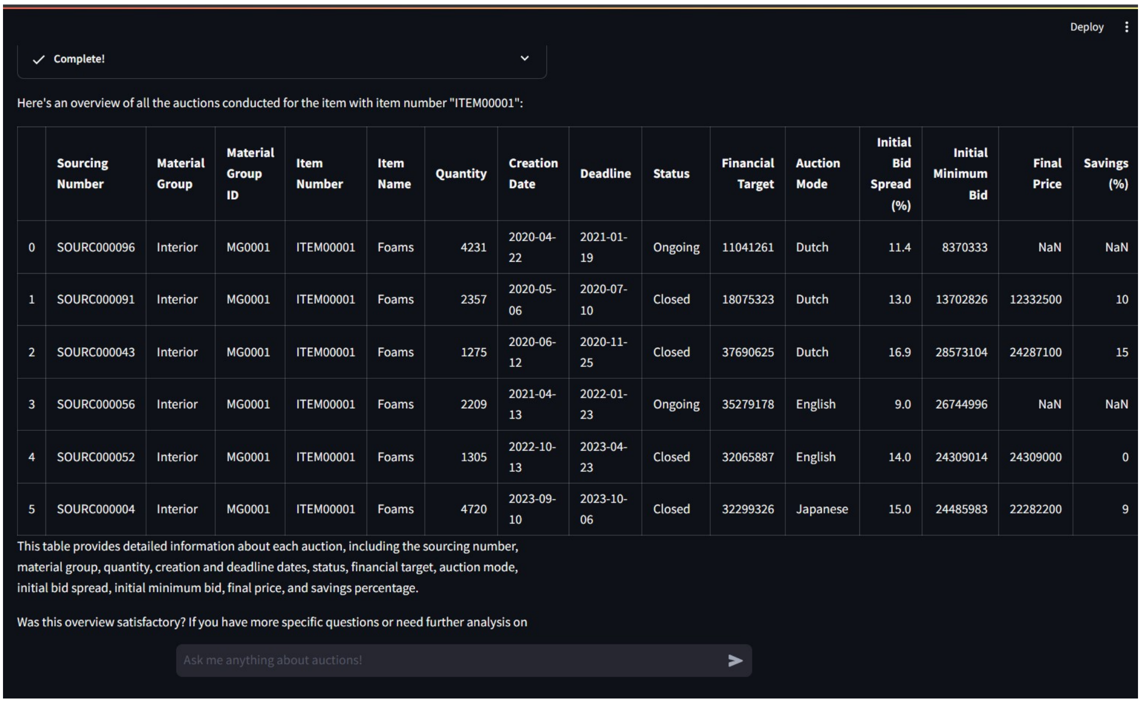
✓ **python_repl_ast:** {'query': "df[df['Item number'] == 'ITEM00001']"}

✓ **Complete!**

Here's an overview of all the auctions conducted for the item with item number "ITEM00001":

	Sourcing Number	Material Group	Material Group ID	Item Number	Item Name	Quantity	Creation Date
0	SOURC000096	Interior	MG0001	ITEM00001	Foams	4231	2020-04-22
1	SOURC000091	Interior	MG0001	ITEM00001	Foams	2357	2020-05-06
2	SOURC000043	Interior	MG0001	ITEM00001	Foams	1275	2020-06-

Figure B.3: The User Interface of the AuctionBot



Here's an overview of all the auctions conducted for the item with item number "ITEM00001":

	Sourcing Number	Material Group	Material Group ID	Item Number	Item Name	Quantity	Creation Date	Deadline	Status	Financial Target	Auction Mode	Initial Bid Spread (%)	Initial Minimum Bid	Final Price	Savings (%)
0	SOURC000096	Interior	MG0001	ITEM00001	Foams	4231	2020-04-22	2021-01-19	Ongoing	11041261	Dutch	11.4	8370333	NaN	NaN
1	SOURC000091	Interior	MG0001	ITEM00001	Foams	2357	2020-05-06	2020-07-10	Closed	18075323	Dutch	13.0	13702826	12332500	10
2	SOURC000043	Interior	MG0001	ITEM00001	Foams	1275	2020-06-12	2020-11-25	Closed	37690625	Dutch	16.9	28573104	24287100	15
3	SOURC000056	Interior	MG0001	ITEM00001	Foams	2209	2021-04-13	2022-01-23	Ongoing	35279178	English	9.0	26744996	NaN	NaN
4	SOURC000052	Interior	MG0001	ITEM00001	Foams	1305	2022-10-13	2023-04-23	Closed	32065887	English	14.0	24309014	24309000	0
5	SOURC000004	Interior	MG0001	ITEM00001	Foams	4720	2023-09-10	2023-10-06	Closed	32299326	Japanese	15.0	24485983	22282200	9

This table provides detailed information about each auction, including the sourcing number, material group, quantity, creation and deadline dates, status, financial target, auction mode, initial bid spread, initial minimum bid, final price, and savings percentage.

Was this overview satisfactory? If you have more specific questions or need further analysis on

Figure B.4: The User Interface of the AuctionBot

Appendix C

Data Cleaning

C.1 Number of Bids, Participants, Active and repeated bidders

The SQL Query below Listing C.1 is used to extract the number of participants, the amount of active participants and the number of bids from the production Database of the Online Negotiation Tool. For the repeated number of bidders, a separate SQL Query is done and the actual number of repeated bidders is calculated in Python outlined in Listing C.2and Listing ??

```sql
With TotalParticipants AS (
SELECT
    p.eventid as eventid,
    Count(DISTINCT p.companyid) as numberparticipants
FROM events_event_participants p
GROUP BY eventid

),
Active AS (
SELECT
    e.id               AS eventid,
    e.name             AS eventname,
    a.id               AS auctionid,
    COUNT(DISTINCT b.companyid) AS activeparticipants,
    COUNT(b.id)      AS numberofbids
FROM
         events_events e
    INNER JOIN events_auctions     a ON a.eventid = e.id
    INNER JOIN events_auction_bids b ON b.auctionid = a.id
WHERE b.isvalid = 1
GROUP BY
    e.id,
    e.name,
    a.id
)

SELECT
    tp.eventid as eventid,
    ap.eventname,
    ap.auctionid,
    tp.numberparticipants,
    ap.activeparticipants,
    ap.numberofbids
FROM TotalParticipants tp
```

```
LEFT JOIN Active ap ON tp.eventid = ap.eventid
```

Listing C.1: Participants Query

```
SELECT
    si.id AS sourcingid,
    s.id AS sourcingitemid,
    e.id AS eventid,
    e.dateofentry as dateofentry,
    p.companyid as companyid
FROM
         events_events e
    INNER JOIN events_event_sourcingitems s ON s.eventid = e.id
    INNER JOIN si_si                      si ON si.id = s.sourcingitemid
    INNER JOIN events_event_participants p ON p.eventid = e.id
```

Listing C.2: Repeated Participants per sourcing item Query

```
# Step 1: Remove rows where COMPANYID is empty
# df_analysis_corrected = df_csv_corrected[df_csv_corrected['COMPANYID'] != ""]
sourc.dropna(axis=0).info()

# Step 2: Sort the dataframe by DATEOFENTRY to ensure we process events in chronological order
# df_analysis_corrected.sort_values(by='DATEOFENTRY', inplace=True)
sourc.sort_values(by='dateofentry', inplace=True)

# Step 3: We will create a dictionary to keep track of the COMPANYIDs that have appeared in any EVENTID
company_history = {}

# Step 4: We will iterate through each event, and for each company in the event, check if it's in the history
# If it is, we increment a counter; if not, we add it to the history
repeated_counts= {}

for eventid, group in sourc.groupby('eventid'):
    repeated_counts[eventid] = 0
    for companyid in group['companyid'].unique():
        if companyid in company_history:
            repeated_counts[eventid] += 1
        company_history[companyid] = True

# Step 5: Convert the repeated_counts dictionary to a dataframe
repeated_counts_df = pd.DataFrame(list(repeated_counts.items()), columns=['eventid', 'repeatedparticipants'])

repeated_counts_df
repeated_counts_df['repeatedparticipants'].unique()
```

Listing C.3: Python Computation Repeated Bidders

C.2 TTO SQL Statements

The SQL Query below is used to extract XXX from the production Database of the Online Negotiation Tool

```
WITH latestsnapshot AS (
    SELECT
```

```
        eab.auctionid,
        eab.companyid,
        MAX(eas.numbr) AS latestnumbr
    FROM
             events_auction_bids eab
        INNER JOIN events_auction_snapshots eas ON eab.auctionid = eas.auctionid
    GROUP BY
        eab.auctionid,
        eab.companyid
)
SELECT
    ee.id                                    AS eventid,
    ee.name                                  AS eventname,
    ls.auctionid,
    easn.name,
    easn.direction,
    easn.itemid,
    STATS_MODE(eas.id),
    STATS_MODE(easn.itemlabel)               AS itemlabel,
    MAX(easn.targetvaluetier)                AS targetvalue,
    MAX(easn.reservation)                    AS reservation,
    MIN(easn.initialvalue)                   AS initialttomax,
    MAX(easn.initialvalue)                   AS initialttomin,
    CASE
        WHEN MIN(easn.initialvalue) = 0 THEN
        CASE
            WHEN easn.direction = 0 THEN MAX(CASE WHEN easn.value != 0 THEN easn.value END)
            WHEN easn.direction = 1 THEN MIN(CASE WHEN easn.value != 0 THEN easn.value END)
        END
        WHEN MIN(easn.initialvalue) <> 0 THEN MIN(easn.initialvalue)
    END AS trueinitialvalue,
    DECODE(MIN(easn.initialvalue),MAX(easn.initialvalue), 'Y', 'N') AS compareinit,
    MIN(CASE WHEN easn.value != 0 THEN easn.value END) AS ttomin,
    MAX(easn.value)                          AS ttomax,
    DECODE(MIN(easn.value),MAX(easn.value), 'Y', 'N') AS compareto,
    CASE
        WHEN easn.direction = 0 THEN MIN(easn.initialvalue) - MIN(easn.value)
        WHEN easn.direction = 1 THEN MAX(easn.value) - MAX(easn.initialvalue)
    END AS calculated_difference
FROM
         latestsnapshot ls
    INNER JOIN gpsovq.events_auction_snapshots eas ON ls.auctionid = eas.auctionid
                                                      AND ls.latestnumbr = eas.numbr
    INNER JOIN gpsovq.events_auction_snpshtent easn ON eas.id = easn.snapshotsavepointid
    INNER JOIN events_auctions                 ea ON ea.id = ls.auctionid
    INNER JOIN events_events                   ee ON ee.id = ea.eventid
WHERE
    ee.istraining = 0
    AND ee.isonline = 1
GROUP BY
    ee.id,
    ee.name,
    ls.auctionid,
    easn.name,
    easn.direction,
    easn.itemid
```

Listing C.4: TTO Query

C.3 Information Exchange Mechanism

The SQL Query below is used to extract the information exchange mechanism for each auction from the production Database of the Online Negotiation Tool.

```
WITH BidsPerAuction AS (
    -- Calculate the number of bids and bidders per auction
    SELECT
        eab.AUCTIONID,
        COUNT(DISTINCT eab.COMPANYID) AS NumberOfBidders,
        COUNT(eab.ID) AS NumberOfBids
    FROM
        events_auction_bids eab
    WHERE
        eab.isvalid = 1
    GROUP BY
        eab.AUCTIONID
),

InitialBidSpread AS (
    -- Calculate the spread of initial bids for each auction and each criteria
    SELECT
        eab.AUCTIONID,
        easn.name,
        MAX(CASE WHEN easn.value != 0 THEN easn.value END) AS maxbid,
        MIN(CASE WHEN easn.value != 0 THEN easn.value END) AS minbid,
        MAX(CASE WHEN easn.value != 0 THEN easn.value END) - MIN(CASE WHEN easn.value != 0 THEN easn.value END) AS SpreadOfInitialBids
    FROM
        events_auction_bids eab
    JOIN
        events_auction_snpshtent easn ON eab.ID = easn.RANKABLEID
    WHERE
        eab.ISINITIAL = 1 AND eab.isvalid = 1
    GROUP BY
        eab.AUCTIONID,
        easn.name
),

GeneralBidSpread AS (
    -- Calculate the spread of general bids (excluding initial bids) for each auction
    SELECT
        eab.AUCTIONID,
        easn.name,
        MAX(CASE WHEN easn.value != 0 THEN easn.value END) AS maxbid,
        MIN(CASE WHEN easn.value != 0 THEN easn.value END) AS minbid,
        MAX(CASE WHEN easn.value != 0 THEN easn.value END) - MIN(CASE WHEN easn.value != 0 THEN easn.value END) AS SpreadOfGeneralBids
    FROM
        events_auction_bids eab
    JOIN
        events_auction_snpshtent easn ON eab.ID = easn.RANKABLEID
    WHERE
        eab.ISINITIAL = 0 AND eab.isvalid = 1
    GROUP BY
        eab.AUCTIONID,
        easn.name
)


SELECT DISTINCT
    ea.ID AS auctionid,
    bpa.NumberOfBidders,
    bpa.NumberOfBids,
    eac.id AS criteriaid,
    eac.version,
    eac.num,
    eac.name,
    eac.description,
    ibs.minbid AS initminbid,
    ibs.maxbid AS initmaxbid,
    ibs.SpreadOfInitialBids,
    gbs.minbid AS genminbid,
    gbs.maxbid AS genmaxbid,
    gbs.SpreadOfGeneralBids,
    eac.scpe,
    eac.direction,
    eac.ismonetary,
    eac.formula,
    eac.hidebeforestart,
    eac.initialvalue,
    eac.hidewoinitialvalue,
    eac.initialvalueisglobalstart,
    eac.steprequired,
    eac.minimumstep,
    eac.maximumstep,
    eac.maximumstepin,
    eac.minimumstepin,
    eac.pattern,
    eac.showbest,
    eac.showrank,
    eac.showtier,
    eac.tiergreen,
    eac.tieryellow,
    eac.steprequired,
    eac.showequalvalues,
    eac.preventequalvalues,
    eac.allowequalranks,
    eac.allowifallgranted,
    eac.allowifanygranted,
    eac.allowifnotgranted,
    eac.blurtiergreen,
    eac.blurtieryellow,
    eac.targetvalue,
    eac.targetvaluetier,
    eac.targetvaluetiermode,
    eac.targetvaluetranspar,
    eac.targetvaluetiertranspar
FROM
    events_auctions ea
JOIN
```

```
    events_events ee ON ea.EVENTID = ee.ID AND ee.ISTRAINING = 0
LEFT JOIN
    BidsPerAuction bpa ON ea.ID = bpa.AUCTIONID
LEFT JOIN
    events_auction_criteria eac ON ea.ID = eac.AUCTIONID
LEFT JOIN
    InitialBidSpread ibs ON ea.ID = ibs.AUCTIONID
LEFT JOIN
    GeneralBidSpread gbs ON ea.ID = gbs.AUCTIONID
WHERE
    ibs.minbid is not NULL
ORDER BY
    ibs.spreadofinitialbids DESC
```

Listing C.5: Information Exchange Mechanism

```
WITH BidsPerAuction AS (
    -- Calculate the number of bids and bidders per auction
    SELECT
        eab.AUCTIONID,
        COUNT(DISTINCT eab.COMPANYID) AS NumberOfBidders,
        COUNT(eab.ID) AS NumberOfBids
    FROM
        events_auction_bids eab
    WHERE
        eab.isvalid = 1
    GROUP BY
        eab.AUCTIONID
),

InitialBidSpread AS (
    -- Calculate the spread of initial bids for each auction and each criteria
    SELECT
        eab.AUCTIONID,
        easn.name,
        MAX(CASE WHEN easn.value != 0 THEN easn.value END) AS maxbid,
        MIN(CASE WHEN easn.value != 0 THEN easn.value END) AS minbid,
        MAX(CASE WHEN easn.value != 0 THEN easn.value END) - MIN(CASE WHEN easn.value != 0 THEN easn.value END) AS SpreadOfInitialBids
    FROM
        events_auction_bids eab
    JOIN
        events_auction_snpshtent easn ON eab.ID = easn.RANKABLEID
    WHERE
        eab.ISINITIAL = 1 AND eab.isvalid = 1
    GROUP BY
        eab.AUCTIONID,
        easn.name
),

GeneralBidSpread AS (
    -- Calculate the spread of general bids (excluding initial bids) for each auction
    SELECT
        eab.AUCTIONID,
        easn.name,
        MAX(CASE WHEN easn.value != 0 THEN easn.value END) AS maxbid,
        MIN(CASE WHEN easn.value != 0 THEN easn.value END) AS minbid,
        MAX(CASE WHEN easn.value != 0 THEN easn.value END) - MIN(CASE WHEN easn.value != 0 THEN easn.value END) AS SpreadOfGeneralBids
```

```sql
    FROM
        events_auction_bids eab
    JOIN
        events_auction_snpshtent easn ON eab.ID = easn.RANKABLEID
    WHERE
        eab.ISINITIAL = 0 AND eab.isvalid = 1
    GROUP BY
        eab.AUCTIONID,
        easn.name
)


SELECT DISTINCT
    ea.ID AS auctionid,
    bpa.NumberOfBidders,
    bpa.NumberOfBids,
    eac.id AS criteriaid,
    eac.version,
    eac.num,
    eac.name,
    eac.description,
    ibs.minbid AS initminbid,
    ibs.maxbid AS initmaxbid,
    ibs.SpreadOfInitialBids,
    gbs.minbid AS genminbid,
    gbs.maxbid AS genmaxbid,
    gbs.SpreadOfGeneralBids,
    eac.scpe,
    eac.direction,
    eac.ismonetary,
    eac.formula,
    eac.hidebeforestart,
    eac.initialvalue,
    eac.hidewoinitialvalue,
    eac.initialvalueisglobalstart,
    eac.steprequired,
    eac.minimumstep,
    eac.maximumstep,
    eac.maximumstepin,
    eac.minimumstepin,
    eac.pattern,
    eac.showbest,
    eac.showrank,
    eac.showtier,
    eac.tiergreen,
    eac.tieryellow,
    eac.steprequired,
    eac.showequalvalues,
    eac.preventequalvalues,
    eac.allowequalranks,
    eac.allowifallgranted,
    eac.allowifanygranted,
    eac.allowifnotgranted,
    eac.blurtiergreen,
    eac.blurtieryellow,
    eac.targetvalue,
    eac.targetvaluetier,
    eac.targetvaluetiermode,
    eac.targetvaluetranspar,
    eac.targetvaluetiertranspar
FROM
```

```
    events_auctions ea
JOIN
    events_events ee ON ea.EVENTID = ee.ID AND ee.ISTRAINING = 0
LEFT JOIN
    BidsPerAuction bpa ON ea.ID = bpa.AUCTIONID
LEFT JOIN
    events_auction_criteria eac ON ea.ID = eac.AUCTIONID
LEFT JOIN
    InitialBidSpread ibs ON ea.ID = ibs.AUCTIONID
LEFT JOIN
    GeneralBidSpread gbs ON ea.ID = gbs.AUCTIONID
WHERE
    ibs.minbid is not NULL
ORDER BY
    ibs.spreadofinitialbids DESC
```

Listing C.6: Decoding the Information Exchange Mechanism

C.4 Auction Type

The SQL Query below is used to extract the buyer surplus for each ticker auction from the production Database of the Online Negotiation Tool. Combined with the buyer surplus of the English auctions, it is used to test the hypothesis on the auction type.

```

SELECT
    e.id                            AS eventid,
    t.id                            AS tickerid,
    i.id                            AS itemid,
    STATS_MODE(i.name)              AS itemname,
    MIN(t.japaneseminbidders)       AS "Min_japaneseminbidders",
    STATS_MODE(t.tickermode)        AS "Stats_Mode_tickermode",
    STATS_MODE(t.dutchmode)         AS "Stats_Mode_dutchmode",
    STATS_MODE(t.tickermode_backup) AS "Stats_Mode_tickermode_backup",
    STATS_MODE(s.version)           AS versionstep,
    MIN(s.startdate)                AS min_start,
    MAX(s.startdate)                AS max_start,
    MIN(s.enddate)                  AS min_end,
    MAX(s.enddate)                  AS max_end,
    MIN(p.price)                  AS min_price,
    MAX(p.price)                  AS max_price,
    DECODE(MIN(p.price), MAX(p.price), 'Y', 'N')       AS compare,
    (MAX(p.price) - MIN(p.price)) AS diff,
    (MAX(p.price) - MIN(p.price))/MAX(p.price) AS result
FROM
         events_events e
    INNER JOIN events_ticker_auctions t ON t.eventid = e.id
    INNER JOIN events_tauction_items i ON i.tickerauctionid = t.id
    INNER JOIN events_tauction_steps  s ON s.tickerauctionid = t.id
    INNER JOIN events_tauction_prices p ON p.tickerstepid = s.id AND p.
    tickeritemid = i.id
    INNER JOIN events_tauction_bids b ON b.tickerpriceid = p.id
GROUP BY
    e.id,
    t.id,
    i.id
HAVING MAX(p.price) <> 0
```

Listing C.7: Auction Type

C.5 Auction Duration

The SQL Query below is used to extract the buyer surplus for each ticker auction from the production Database of the Online Negotiation Tool. Combined with the buyer surplus of the English auctions, it is used to test the hypothesis on the auction type.

```sql
SELECT
    ea.eventid,
    ea.strt,
    ea.stp,
    ea.stp - ea.strt as duration,
    ea.initialstop,
    ea.prolongation,
    ea.prolongationstrategy,
    ea.prolongationcount
FROM events_auctions ea
```

Listing C.8: Auction Duration and Overtime/Prolongation

Appendix D

AuctionEval: The Evaluation Dataset

This section devolves in detail about the process and datasets used for the evaluation of the data retrieval agent in Section D.1 and the overall system in Section D.2. Accompanied by descriptions, Table D.3 represents the 20 question answer pairs used to evaluate the data retriever, while Table D.4 shows the 30 question used to evaluate the overall system.

D.1 Data Retrieval Evaluation

Two data tables are uploaded and made available to the data retrieval agent. The data tables are simpler representations of the actual database underlying the online negotiation and serve as the ground for evaluating the ability of the data retrieval agent to do the following actions:

- Basic Data Retrieval
- Simple Calculations
- Handling of different data types (text, numerical, dates, lists of text)
- Join information from two tables

The first data table includes the majority of information including the process sourcing number, the material group, the quantity to be purchased, several date data regarding an auction, the auction mode, the financial targets and also information related to the auction outcome. An example of the first data table is shown in Table D.2. The second data table consist of simple information of the participants that are linked to the auction in the last column of Table D.2. Table D.1 showcases the supplier information table, which can be linked to the auction information table in different ways

Table D.3 showcases the custom question answer pairs utilized for the evaluation of the data retriever. It consists of 20 questions that range from basic data retrieval, simple calculations to joining the two tables, merging information from more complex queries, and handling no returns or missing data. Since data sources are dynamic and can change over time, the evaluation was designed to work with dynamic data through the use of the dynamic evaluator.

Table D.1: Exemplary dataset representing a simplification of the real dataset of participant information at the case company.

Supplier ID	Supplier Name	Risk Aversion	Market Dominance
SUP79040	MotoCraft	High	Low
SUP73160	ElectroMotive	Low	Low
SUP12441	TireTech	Low	Low
SUP92890	DriveLine	High	High
SUP39837	GearHead	Low	Low
SUP77315	GearHead	High	High
SUP80226	Michelin	High	Low

This dynamic evaluator takes the reference code created to being able to answer the question and executes it, and places the retrieved values in the reference answer. For example in the second question of Table D.3, the dynamic evaluator would execute the reference code, and have the instruction to construct the reference answer to fill out the date with the 'creation_date' variable. In this way, the evaluators reference answer is "The creation date of sourcing number SOURC000091 is creation_date.", where the 'creation_date' variable would be filled with the value retrieved by the reference code. For the sake of showing the reference answers in their final form, example values for these are showcased in Table D.3, and the reference codes are highlighted in the last column.

Table D.3: The individual evaluation dataset for the data retrieval agent. They represent 20 question and reference answer pairs based on the tabular structures of the previously shown tabular data made available to the agent.

Questions	Reference Answer	Reference Code
Give me an overview over all conducted auctions that are available for the item with itemnumber: ITEM01	The conducted auctions available for ITEM01 are: [Table showing relevant information about auctions for ITEM01]	item_auctions = auction_df[auction_df['Item number'] == 'ITEM01']
What is the creation date of sourcing number SOURC000091?	The creation date of sourcing number SOURC000091 is 06/05/20.	creation_date = item_auctions[item_auctions ['Sourcingnumber'] == 'SOURC000091']['Creation Date']. values[0]
How many items were auctioned in sourcing number SOURC000043?	The number of items auctioned in sourcing number SOURC000043 is 1275.	items_auctioned = item_auctions[item_auctions ['Sourcingnumber'] == 'SOURC000043'] ['Quantity'].values[0]
What was the lowest initial minimum bid for sourcing number SOURC000096?	The lowest initial minimum bid for sourcing number SOURC000096 is $8370333.	lowest_initial_bid_SOURC000096 = item_auctions[item_auctions ['Sourcingnumber'] == 'SOURC000096'] ['Initial Minimum Bid'] .min()
Which sourcings are still ongoing with the material group MG001?	There are no ongoing sourcings with the material group MG001.	ongoing_MG001_sourcings= auction_df [(auction_df['Material Group'] == 'MG001') & (auction_df['Status'] == 'Open')]['Sourcingnumber'].tolist()

Table D.3: The individual evaluation dataset for the data retrieval agent. They represent 20 question and reference answer pairs based on the tabular structures of the previously shown tabular data made available to the agent.

Questions	Reference Answer	Reference Code
What is the most common auction mode used in the Material Group MG001?	The most common auction mode used in the Material Group MG001 is Dutch.	most_common_auction_MG001 = auction_df[auction_df[’Material Group’] == ’MG001’][’Auction Mode’].mode()[0]
When is the deadline for the item ITEM03 in this 2023?	The deadline for the item ITEM03 in 2023 is [Deadline Date].	deadline_ITEM03_2023 = item_df[(item_df[’Item number’] == ’ITEM03’) &(item_df[’Deadline’]. dt.year == 2023)] [’Deadline’].values[0]
What is the highest achieved saving for ITEM01? What auction type did that sourcing have?	The highest achieved saving for ITEM01 is 15%, and the auction type for that sourcing was Dutch.	highest_saving_ITEM01 = item_auctions[item_auctions [’Item number’] == ’ITEM01’] [’Savings (%)’].max() auction_type_highest_saving_ ITEM01 = item_auctions[item_auctions [’Savings (%)’] == highest_saving_ITEM01] [’Auction Mode’]. values[0]
Given ITEM01, calculate the average quantity of items to be purchased.	The average quantity of items to be purchased for ITEM01 is 2690.	average_quantity_ITEM01 = item_auctions[item_auctions [’Item number’] == ’ITEM01’] [’Quantity’].mean()
Calculate the average savings percentage across all auctions of the Interior Material Group.	The average savings percentage across all auctions of the Interior Material Group is 12.5%.	interior_auctions = auction_df[auction_df [’Material Group’] == ’Interior’] average_savings_interior = interior_auctions[’Savings (%)’].mean()
Calculate the desired savings percentage If the financial target is to be met for sourcing SOURC000011	If the financial target is to be met for sourcing SOURC000011, the desired savings percentage would be 20%.	financial_target_SOURC000011 = item_auctions[item_auctions[’Sourcingnumber’] == ’SOURC000011’] [’Financial Target’].values[0] initial_minimum_bid_SOURC000011 = item_auctions[item_auctions[’Sourcingnumber’] == ’SOURC000011’] [’Initial Minimum Bid’] .values[0] desired_savings_percentage = ((initial_minimum_bid_SOURC000011 - financial_target_SOURC000011) / initial_minimum_bid_SOURC000011) * 100

Table D.3: The individual evaluation dataset for the data retrieval agent. They represent 20 question and reference answer pairs based on the tabular structures of the previously shown tabular data made available to the agent.

Questions	Reference Answer	Reference Code

What is the current saving for sourcing SOURC000056 and what should have been the savings if the financial target was met?	The current savings for sourcing SOURC000056 is 0%. If the financial target was met, the expected savings percentage would have been 2%.	current_savings = auction_df.loc[auction_df[’Sourcingnumber’] == ’SOURC000056’, ’Savings (%)’].values[0]
		financial_target = auction_df.loc[auction_df[’Sourcingnumber’] == ’SOURC000056’, ’Financial Target’].values[0]
		initial_minimum_bid = auction_df.loc[auction_df[’Sourcingnumber’] == ’SOURC000056’, ’Initial Minimum Bid’].values[0]
		expected_savings = ((initial_minimum_bid - financial_target) / initial_minimum_bid) * 100

Calculate the per unit price for sourcing SOURC000056?	The per unit price for sourcing SOURC000056 is $12.11.	final_price = auction_df.loc[auction_df[’Sourcingnumber’] == ’SOURC000056’, ’Final Price’].values[0]
		quantity = auction_df.loc[auction_df[’Sourcingnumber’] == ’SOURC000056’, ’Quantity’].values[0]
		per_unit_price = final_price / quantity

How many participants participated in Sourcing SOURC000043?	There were 2 participants in sourcing SOURC000043.	participants_list = auction_df.loc[auction_df[’Sourcingnumber’] == ’SOURC000043’, ’Participants’].values[0]
		num_participants = len(participants_list.strip(’[]’).split(’,’))

Who are the recurring participants for the auction that needs to take place until the 19th of January 2021?	The recurring participants for the auction until the 19th of January 2021 are suppliers SUP12441 and SUP92890.	auction_before_date = auction_df[auction_df[’Creation Date’] <’19/01/2021’]
		recurring_participants = auction_before_date[’Participants’].explode().value_counts()[auction_before_date[’Participants’].explode().value_counts() >1].index.tolist()

How many and which participants are market dominant in the still ongoing auction for the Alloy wheels in the exterior material group?	There are 0 market dominant participants in the ongoing auction for the Alloy wheels in the exterior material group.	ongoing_auction = auction_df[(auction_df[’Material Group’] == ’Exterior’) &(auction_df[’Status’] == ’Open’)]
		dominant_participants = participant_df[participant_df[’Market Dominance’] == ’High’][’Supplier ID’].tolist()
		num_dominant_participants = len(set(ongoing_auction[’Participants’].explode()).intersection(dominant_participants))

What initial bid spread does item ITEM05 have for each auction mode? Which auction mode has the highest value?	For ITEM05, the initial bid spread for the Dutch auction mode is 105% and for the English auction mode is 9%. The Dutch auction mode has the highest initial bid spread.	bid_spread_by_mode = auction_df[auction_df[’Item name’] == ’Alloy Wheels’].groupby(’Auction Mode’)[’Initial Bid Spread (%)’].mean()
		highest_bid_spread_mode = bid_spread_by_mode.idxmax()

Table D.3: The individual evaluation dataset for the data retrieval agent. They represent 20 question and reference answer pairs based on the tabular structures of the previously shown tabular data made available to the agent.

Questions	Reference Answer	Reference Code
What are the characteristics of the participants of sourcing SOURC000043?	The characteristics of the participants of sourcing SOURC000043 are: SUP79040: High risk aversion, Low market dominance SUP92890: High risk aversion, High market dominance	participant_characteristics = participant_df[participant_df ['Supplier ID'].isin(auction_df. loc[auction_df['Sourcingnumber'] == 'SOURC000043', 'Participants'] .values[0].strip('[]').split(','))]
List the sourcing numbers for which the final price is below the financial target	The sourcing numbers for which the final price is below the financial target are: SOURC000056.	below_target_sourc_numbers = auction_df[auction_df['Final Price'] <auction_df['Financial Target']] ['Sourcingnumber'].tolist()
Identify the sourcing numbers where the auction mode was Dutch, a dominant bidder participated, and the savings percentage was above 10%	There are no sourcing numbers meeting all the specified criteria.	dutch_dom_savings_above_10 = auction_df[(auction_df ['Auction Mode'] == 'Dutch') & (auction_df['Participants']. str.contains(','.join (dominant_participants))) & (auction_df['Savings (%)'] >10)] ['Sourcingnumber'].tolist()
What auction mode was the best performing for ITEM01 in the last 6 months?	The best performing auction mode for ITEM01 in the last 6 months was the Dutch auction mode, with an average saving of 25%	six_months_ago = pd.to_datetime ('today') - pd.DateOffset(months=6) best_performing_mode = auction_df [(auction_df['Item name'] == 'Foams') & (pd.to_datetime(auction_df' ['Creation Date'], format='%d/%m/%y') >six_months_ago)]. groupby('Auction Mode')['Final Price'] .mean().idxmax()

D.2 Overall System Evaluation: AuctionEval Dataset

For the evaluation of the whole systems performance, the custom made dataset AuctionEval is created. As shown in Table D.4, AuctionEval consists of question answer pairs representing probable prompts of purchasers to the system to guide the auction design recommendation. The questions are divided into three categories

- Auction Theory Questions
- Auction Recommendation Questions
- Auction Theory & Recommendation Discussions

The three categories relate to the reasoning tasks to be studied as described in section 3.3.3, namely logic based and covariance based type (general) and specific causality. Thereby the first category of questions covers the logic based reasoning on a general level, asking about theoretical concepts, recommendation models and advice taken from the provided textual documents and research papers. The second category then goes into the common use case of recommending an auction type for a specific sourcing process to which the purchaser needs to decide and conduct the auction. Compared to the data retrieval evaluation, this is a test of the whole system, including the orchestrate, the context retriever and the data retriever, as the data from the specific sourcing needs to be retrieved, and then relevant context needs to be combined with it to make a sound recommendation. The last category deals with more complex questions and critical discussion about auctions and their theory. Some questions also include a check for possible hallucinations by referring to empty data fields or stipulating wrong assumptions.

Table D.4: AuctionEval Dataset

Questions	**Reference Answers**
Auction Theory Questions (Type Causality)	

Table D.4: AuctionEval Dataset

Questions	**Reference Answers**
Given a low number of bidders and a close bid spread, what auction design would you recommend?	Recommendation: English Rank Auction Reasoning: A low price dispersion indicates higher competitivness between bidders. This higher competitiveness can yield to more bids in an English Auction, which encourages bidder through its information policy and open design, to increse their bidding activity based on the activity and valutions of other bidders. Because the valuations for the bidders are close, it is expected that the competition in the English Auction will get us beyond the strategic margin of a possible First-price auction (Dutch or Japanese). As the number of bidders is low, it is advised to use the Rank information policy to give bidders a sense of closeness to the best position.
Explain why an English auction is preferable over a Dutch auction in a scenario with a low number of bidders and a close bid spread.	Recommendation: English Auction Counterfactual Reasoning: If an English auction is used in the case of low number of bidders and close bid spread, the auction is able to utilize competitivness of the bidders to drive the bids down, given that bidders perceive to be close to winning. If a dutch auction is used in the case of low number of bidders and close bid spread, the competitivness of the bidder would not be utilized as the bidders would only add a strategic margin to their respective bids indpendent of the bids of others.
Given a low number of bidders and a high bid spread, what auction design would you recommend?	Recommendation: Dutch Auction Reasoning: For a scenario with a low number of bidders and a high price dispersion, the Dutch auction is recommended as it achieves higher cost savings due to the strategic margin that a possible dominant bidder adds to their bid in a first-price auction setting, to secure the auction. A dutch auction also gives the bidder no information about his compeition, which is utilized to let the dominant buyer not know about his dominant position.
Given a low number of bidders and a low bid spread, what auction design would you recommend?	Recommendation: English Rank Auction Reasoning: When the number of bidders is low and because it uses the competitive environment to induce bidders to decrease their bids, taking advantage of the perceived closeness to the best bid to enhance the competitiveness of the auction.

Table D.4: AuctionEval Dataset

Questions	**Reference Answers**
Given a high number of bidders and a high bid spread, what auction design would you recommend?	Recommendation: Hybrid Auction (English followed by Dutch Auction) or First-price Auction Reasoning: With a high number of bidders and high bid spread, a hybrid auction is recommended. The initial phase of an English auction allows for an open exchange of information which encourages price adjustments based on bidders' valuations, followed by a Dutch auction to exert pressure on dominant bidders. In case a single phase auction is desired, a first-price auction is advised. The decision on which specific auction type depends on the attractivity of the auction, for low attractity it is the japanese auction, for high attractivity the dutch auction.
Given a high number of bidders, a high bid spread, and a high auction vole/attractity of the auction what auction design would you recommend?	Recommendation: Dutch Auction Reasoning: High attractivity coupled with a high bid spread indicates that bidders are very interested and the auction is competitive. A Dutch auction encourages a the bidders to set a low strategic margin close to their indifference price, in cases where the attractivity is high, due to the high risk aversion of possible dominant buyers , whose existence is indicated by the large bid spread.
With a high number of bidders, a high bid spread, and a low auction vole/attractity of the auction what auction design would you recommend?	Recommendation: First-price Auction or Japanese Auction Reasoning: With high bidder numbers and bid spread, but low attractivity, a first-price auction format, such as a Japanese auction, is recommended as it encourages the dominant bidder, as indicated by the high bid spread, to bid closer to his reservation price to secure the auction. Given the low attractivity of the auction, a japanese auction encourages bidders to step by step approach their reservation price without applying pressure as the dutch auction would.
Explain why in the case of a low number of bidders, a high bid spread, and a high auction volume/attractivity of the auction, a dutch auction is recommended?	Recommendation: Dutch Auction Reasoning: In cases of low bidder numbers and high bid spread with high attractivity, a Dutch auction tends to perform better as it benefits from bidders' high interest and encourages them to bid close to their true valuations.
For which auction types should I allow overtime?	Recommendation: Allow overtime for english auctions except with best-bid and rank information exchange Reasoning: Overtime should be allowed in all English Auction Types, except for the best-bid and rank information exchange, because overtime in this case allows the bidders to only bid towards the very end of the auction because of the high information transparency about their position and the best bid.
Should I invite as many particpants as possible or should I only include the former participants to the next auction?	Recommendation: Invite as many as possible including the repeated participants Reasoning: The empirical research has shown that the higher number of bidders, the higher the savings. This is due to the fact, depending on the auction format, that more competition yields to more bids, which reduces the final price, especially if the information exhcange policy enables the valuations of the bidders to be impacted by other bids.

Table D.4: AuctionEval Dataset

Questions	**Reference Answers**
How would suspected bidder collusion affect the outcome of an English auction for an item with a high initial bid spread? What auction design adaptations can we make to mitigate the effects of collusion?	Recommendation: Auction would yield no significant savings. Set reserve price for maximum acceptable price based on historical data. Change information policy to avoid sharing how many other bidders are in the auction. Reasoning: Suspected bidder collusion in an English auction with a high initial bid spread could depress the final sale price. Colluders may agree not to outbid each other, minimizing the incremental bidding and potentially ending the auction at a higher price than if they had competed as independent, non-collusive bidders. The collusive behaviour in the special case of an english auction with high initial spread can be masked or mistaken as inactive bidding, because the high difference in initial price might seem to discourage bidders from bidding. To mitigate the risk of collusion, setting a reserve price close to the historical price and a change in the information policy of the english is advised. The change in information policy of the auction avoids informing the colluding bidders about how many other bidders are there or the current best price, increasing the insecurity of bidders to be the only ones in the auction, hence potentially missing out on the auction.
How does a rank and best-bid english auction work?	Explanation: In an English auction, bidders need to place bids that are lower than any bid they have previously made. Different from a ticker auction, bidders are given the freedom to choose both the timing and the value of their bid, provided it is lower than their last bid. There is no limitation on the number of bids that can be submitted, as long as the bids are successively lower than the last. The auction ends after a defined period of time, and the auction is won by the bidder who has placed the lowest bid by the conclusion of the process at the price of the submitted bid. A rank and best-bid English auction works by providing bidders with two pieces of information: their current rank in the bidding process and the current best bid. This format encourages competitive behavior among the bidders.
What is the difference between a Japanese first-price auction and a dutch auction?	The primary difference between a Japanese first-price auction and a Dutch auction lies in their operational mechanisms. In a Dutch auction, the auctioneer begins with a low asking price which increases until one of the bidders accepts the price. The first bidder to accept the price wins and ends the auction, without revealing any information about the other bidders prices. A Japanese first-price auction, on the other hand, requires bidders to stay in the auction by actively confirming their willingness to pay the current price as the price starts from a high asking price and continues to decrease. Bidders drop out as the price decreases and they're unwilling to pay the price. The auction ends when only one bidder remains, willing to accept the current price, thus being a less stressfull experience for the bidders, while still retaining the first-price auction mechanisms and benefits.
Auction Design Recommendation (Specific)	

Table D.4: AuctionEval Dataset

Questions	**Reference Answers**
Recommend an auction design for the ongoing auction with sourcing number SOURC000056 for ITEM01, considering the specifics of the auction context provided in the data	Data Context: - Number of bidders: 4 (High, since >3 bidders) - Initial Bid Spread: 9% (High, since >3% spread) - Initial Minimum Bid: $26,744(Highsince$ >25,000) Recommendation: Dutch Ticker Auction The high number of bidders suggests that there is considerable interest in ITEM01, which could lead to increased competition in the auction. A high initial bid spread indicates that there is a significant difference between the minimum bid and the potential maximum value perceived by the bidders, hinting at a possible dominant buyer. High attractivity implies that risk aversion of bidder to losing the auction is high. Therefore, a dutch auction is recommended that challenges dominant bidders, and uses the pressure of the attractivity and the risk aversion to reduce the strategic margins of bidders.
Which auction type is the most suitable for the ongoing auction with sourcing number SOURC000096 in the year 2020?	Data Context: - Number of bidders: 2 (Low, since <3 bidders) - Initial Bid Spread: 11,4% (High, since >3% spread) - Initial Minimum Bid: $83703(Highsince$ >25,000) Recommendation: Dutch Ticker Auction A high initial bid spread and a low number of bidder indicates that there is a significant difference between the minimum bid and the potential maximum value perceived by the bidders, hinting at a possible dominant buyer. High attractivity implies that risk aversion of bidder to losing the auction is high. Therefore, a dutch auction is recommended that challenges dominant bidders, and uses the pressure of the attractivity and the risk aversion to reduce the strategic margins of bidders.
I have conducted the sourcing with number SOURC000056, in retrospective would it have been better to use a dutch auction due to the large bid spread?	Data Context: - Number of bidders: 4 (High, since >3 bidders) - Initial Bid Spread: 9% (High, since >3% spread) - Initial Minimum Bid: $26744(Highsince$ >25,000) Recommendation: Yes, a dutch ticker auction would be better as apparent by the 0% savings achieved. Looking at the participants information, there is a dominant bidder. Alongside the high initial bid spread, an English auction risks no bidding activity as the non-dominant bidders are discouraged by the large differences in bids. Therefore, a dutch auction is recommended that challenges dominant bidders, and uses the pressure of the attractivity and the risk aversion to reduce the strategic margins of bidders, while hiding information of other bidders.

Table D.4: AuctionEval Dataset

Questions	**Reference Answers**
Which auction type is the most suitable for the ongoing auction with sourcing number SOURC000131 with item ITEM87?	Data Context: - Number of bidders: 5 (High, since >3 bidders) - Initial Bid Spread: 1,8% (High, since <3% spread) - Initial Minimum Bid: 12000($Lowsince >$25,000) Recommendation: English Rank or English Best bid Auction When the number of bidders is high and the bid spread is close, an English auction is advised because it uses the competitive environment to induce bidders to decrease their bids, taking advantage of the perceived closeness to the best bid to enhance the competitiveness of the auction. The Rank information policy can be used if there is the desire to not diclose the actual best bid for compliancy or competitive reasons. Otherwise a best bid English auction gives the impression to the bidders that everyone is close to the best bid. Adding both information policies increases the competition as bidders also see their relative position while they see the distance to the best price, possibly encouraging bidders on lower ranks to experience the competitiveness and bid more agressively
Which auction type is the most suitable for the ongoing auction with sourcing number SOURC000033 for Material Group MG0005?	Data Context: - Number of bidders: 4 (High, since >3 bidders) - Initial Bid Spread: 8,6% (High, since >3% spread) - Initial Minimum Bid: 10000($Highsince >$25,000) Recommendation: Japanese Auction With high bidder numbers and bid spread, but low attractivity, a first-price auction format, such as a Japanese auction, is recommended as it encourages the dominant bidder, as indicated by the high bid spread, to bid closer to his reservation price to secure the auction. Given the low attractivity of the auction, a japanese auction encourages bidders to step by step approach their reservation price without applying pressure as the dutch auction would.
It seems that for SOURC000666 there are two dominant buyers and one risk averse buyer? Can you verify that claim and recommend the optimal auction type in that case?	Data Context: - 1 Risk averse buyer - 2 dominant buyers Recommendation: Yes the claim can be verified and the optimal auction is an English Rank Auction. Although the existence of a dominant buyer might imply the possibility to reduce their strategic margin by a first-price auction like the dutch auction, the existence of two dominant buyers prompts the utilization of the competitive environment through an English auction at the possible expense that the risk-averse participant is initimidated by the large distance to the leading bids. This intimidation can be minimized by using the rank information policy.

Table D.4: AuctionEval Dataset

Questions	**Reference Answers**
What is the optimal auction type for the SOURC000767, also considering that it only consists of participants that are repeated bidders?	Data Context: - Number of bidders: 3 (Low, since <=3 bidders) - Initial Bid Spread: 5.6% (High, since >3% spread) - Initial Minimum Bid: 120000($Highsince$ >25,000) Recommendation: Dutch Ticker Auction The fact that only repeated participants are in the auction does not influence the auction recommendation, compared to the risk aversion type or the dominance of a participant in the market. A high initial bid spread and a low number of bidder indicates that there is a significant difference between the minimum bid and the potential maximum value perceived by the bidders, hinting at a possible dominant buyer. High attractivity implies that risk aversion of bidder to losing the auction is high. Therefore, a dutch auction is recommended that challenges dominant bidders, and uses the pressure of the attractivity and the risk aversion to reduce the strategic margins of bidders.
What would be the ideal auction format for sourcing number SOURC000289 that deals with a highly valued and item?	Data Context: -Number of bidders: 6 (High, since >3 bidders) -Initial Bid Spread: 12% (High, since >3% spread) -Initial Minimum Bid: 4000($Lowsince$ >25000) -Dominant Bidder: 3 Dominant Bidder Recommendation: Japanese Auction or Dutch Auction "Given the high interest reflected by the number of bidders, the significant initial bid spread for ITEM52, and the large number of dominant supplier participating, a First-price auction either the Japanese Auction or the dutch auction is recommended. These factors suggesting that bidders will be willing to bid competitively. Although the initial minimum bid is low, the other factors and the description of the as a highly valued item yield the conclusion that the attractivity can be defined as high and therfore the dutch auction is an alternative.
Auction Theory and Recommendation Discussion	
I would like to conduct a two-phase auction, what combination of auctions would you recommend?	Recommendation: First English Auction, Second Dutch Auction Reasoning: For the first phase, an english auction utilizes the competition to decrease the auction price and ends in mutiple winning bidders being invited into a dutch auction. This dutch auction then pressures the remaining bidders to decrease their strategic margin in order to win the auction.

Table D.4: AuctionEval Dataset

Questions	**Reference Answers**
Why did the auction in sourcing SOURC000234 yield no savings?	Data Context: - Savings (%): 0 - Auction Type: English - Information Exchange: Best Bid -Number of bidders: 2 (Low, since <3 bidders) -Initial Bid Spread: 15% (High, since >3% spread) -Initial Minimum Bid: 4000($Lowsince >$25000) Answer: The auction type was chosen incorrectly, instead of an English auction, a First-Price auction like a Japanese Auction would have yielded a higher saving. Reasoning: The existence of only two bidders with a very high spread, indicates that the english auction did not have any bidding activity as the dominant bidder was not challenged by the weaker bidder and the weaker bidder might have been discouraged by the large distance to the best bid.
For previous auctions of the Item ITEM09, what auction designs achieved the highest savings and why?	For most auctions for Item ITEM09, the english rank auction yielded the highest savings. The reason for that is the always close initial bid spread that auctions with this item have, indicating a competitive market situation. As also only 3 bidders are participating in the auction, most of them repeated participants, the rank information policy gives the bidders the impression that they are close to the best bid without revealing the best price to the competitors.
In Material Group MG089, the auctions conducted from 2018-2020 seem to have higher achieved savings, compared to 2020 until 2023, although the auction type has not been changed? How can that be?	The number of participants has changed in comparison between these two time periods. In the time period of 2018-2020, the average number of bidders is 5, while in the 2020-2023 time period, the averge number of bidders is 3. In combination that the english auction type is always chosen, one reason for the decline in savings can be the reduced amount of bidders present in the auctions.
With what reasoning does Berz propose the use of a Honk Kong auction in a two phase auction design rather than a English auction such as Schulze-Horn proposes?	Berz proposes the use of the Hong Kong auction due to its interesting properties as a so called "1,5-price" auction. The Hong Kong auction is not a clear first-price auction nor a clear second-price auction, as it shares some level of information to the bidders, elimintating a full first-price characteristic, but also is not limited by the second-price logic found in the English auction, as there are multiple winner to the auction and the winner only realize that they won once they have left the auction. Hence he proposes that the Hong Kong auction is able to reduce the number of bidders alongside a price discovery, which is more beneficial for the subsequent first-price auction, than a First-price Sealed bid or an english auction.
What overtime rule would you recommend me to set up in the dutch auction?	The dutch auction cannot have an overtime rule, as the end time and the beginning time are clearly and strictly specified.
Why does Schulze-Horn recommend an exclusive offer at the end of almost all of the english auction designs?	Schulze-Horn recommends an exclusive offer to be reasonable if there is a preference to award a certain supplier. Furthermore, the exclusive offer represents a cooperative negotiation element between the participant and the auctioneer, which can yield better auction outcomes in situations where the competition among bidders is low comapred to competitive elmenets such as an auction.

Table D.4: AuctionEval Dataset

Questions	**Reference Answers**
When should I choose to do a multi-phase auction and when only a single phase auction?	Multi-phase auctions can be chosen in case there is a large number of bidders and there is a possibility to maximize the savings by encouraging a price disovery among multiple steps. This possibility is given in case there exist a high risk aversion to losing the auction and therfore a high attractivity.
Can I combine a Japanese Auction with multiple winners followed by a dutch auctions as a mult-phase auction design? What would be the implications of that design for the sourcing SOURC0852?	Data Context: -Number of bidders: 2 (Low, since <3 bidders) -Initial Bid Spread: 5% (High, since >3% spread) -Initial Minimum Bid: 13456($Lowsince$ >25000) Answer: Yes combining is possible in general. For sourcing SOURC0852, the Japanese-Dutch auction design is not suitable. Reasoning: In general a Japanese Auction with multiple winners in the first-price auction variant supports the price discovery among bidders and utilizes the competition to reduce strategic margins at the same time. A dutch auction could then follow up to pressure the strategic margin of the bidders further, yielding higher achieved savings. For SOURC0852, the low amount of bidders indicates that a multi-phase auction design is not suitable, as only 2 available bidders is only enough for a single phase auction.

TABLE D.2: Exemplary dataset representing a simplification of the real dataset of auction information of the case company. The data tables are used to test the data retrieval agent. Such a dataset would be uploaded to the agent's environment, and the agent would access it through the Python pandas library by executing pandas code.

Sourcingnumber	Material Group	Material Group ID	Item number	Item name	Quantity	Creation Date	Status	Financial Target	Auction Mode	Initial Bid Spread (%)	Initial Minimum Bid	Final Price	Savings (%)	Participants
SOURC000096	Interior	MG0001	ITEM01	Foams	4231	22/04/2020	Open	66000	Dutch	11,4	83703			[SUP79040, SUP73160]
SOURC000091	Interior	MG0001	ITEM01	Foams	2357	06/05/20	Closed	11121	Dutch	13	13702	12332	10	[SUP12441, SUP92890, SUP73160]
SOURC000043	Interior	MG0001	ITEM01	Foams	1275	12/06/20	Closed	25000	Dutch	16,9	28573	24287	15	[SUP79040, SUP92890]
SUP79040, SUP92890]														
SOURC000056	Interior	MG0001	ITEM01	Foams	2209	13/04/21	Open	26000	English	9	26744	26744		[SUP13901, SUP42236, SUP96891, SUP45907]
SOURC000011	Exterior	MG0003	ITEM02	Alloy Wheels	4702	13/09/20	Closed	2300	Japanese	1026	2536	1923,29	24,19	[SUP19515, SUP45907]
SOURC000084	Tires	MG0001	ITEM05	Alloy Wheels	2258	25/11/21	Closed	9000	English	105	13146	13146	0	