COPPE
UFRJ

**Instituto Alberto Luiz Coimbra de
Pós-Graduação e Pesquisa de Engenharia**

# COMPARATIVE ANALYSIS OF SINGLE AND MULTI-AGENT LARGE LANGUAGE MODEL ARCHITECTURES FOR DOMAIN-SPECIFIC TASKS IN WELL CONSTRUCTION

Vitor Brandão Sabbagh

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia de Sistemas e Computação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia de Sistemas e Computação.

Orientador: Geraldo Bonorino Xexéo

Rio de Janeiro
Julho de 2025

COMPARATIVE ANALYSIS OF SINGLE AND MULTI-AGENT LARGE
LANGUAGE MODEL ARCHITECTURES FOR DOMAIN-SPECIFIC TASKS IN
WELL CONSTRUCTION

Vitor Brandão Sabbagh

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO
ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE
ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO
GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E
COMPUTAÇÃO.

Orientador: Geraldo Bonorino Xexéo

Aprovada por: Prof. Geraldo Bonorino Xexéo
      Prof. Jano Moreira de Souza
      Prof. Arnaldo Cândido Júnior

RIO DE JANEIRO, RJ – BRASIL
JULHO DE 2025

*To Carolina, my life partner.*

# Acknowledgements

To my daughter, Marina, who came into the world just two months ago, bringing a new light and a new purpose to my life. I dedicate every page of this work to you, with the hope of building a bright future for you.

To my parents, Vera and Nicolau, for all the love, unconditional support, and for always believing in me. Your faith in my abilities was the foundation for this achievement.

To my beloved wife, Carolina, my gratitude for all the patience, understanding, and love, especially during the most challenging moments of this journey. Without your support, this work would not have been possible.

To my stepson, Filipe, thank you for the moments of joy and relaxation that helped me maintain balance, especially during our Minecraft adventures. May our friendship continue to grow.

I express my deep gratitude to my mentor, Claudio, for his unwavering support and trust since the beginning of my career in digital transformation. His mentorship was fundamental to my professional development.

To my advisor, Xexéo, thank you for the wise guidance, academic rigor, and patience throughout this entire process. Your teachings were crucial to the quality of this work.

I extend my gratitude to the well construction engineering experts, Marcelo Grimberg, Rafael Peralta, and Lorenzo Simonassi, whose expertise and dedication significantly contributed to this research.

I also want to thank Ashish Vaswani. His work on "Attention Is All You Need" paved the way for the Large Language Models that were not only the subject of this dissertation but also an invaluable tool that helped me put ideas into words.

Finally, a special thanks to my colleagues from Petrobras and the Tecgraf Institute. Our daily discussions about Gen-AI were an inexhaustible source of inspiration and knowledge, immensely enriching this dissertation.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

COMPARATIVE ANALYSIS OF SINGLE AND MULTI-AGENT LARGE LANGUAGE MODEL ARCHITECTURES FOR DOMAIN-SPECIFIC TASKS IN WELL CONSTRUCTION

Vitor Brandão Sabbagh

Julho/2025

Orientador: Geraldo Bonorino Xexéo

Programa: Engenharia de Sistemas e Computação

Esta dissertação apresenta a aplicação de modelos de linguagem (LLM) no setor de petróleo e gás, especificamente em tarefas de construção e manutenção de poços. O estudo avalia o desempenho de uma arquitetura baseada em LLM de agente único e de múltiplos agentes no processamento de diferentes tarefas, oferecendo uma perspectiva comparativa sobre sua precisão e as implicações de custo de sua implementação. Os resultados indicam que sistemas multiagentes oferecem desempenho melhorado em tarefas de perguntas e respostas, com uma medida de veracidade 28% maior do que os sistemas de agente único, mas a um custo financeiro mais alto. Especificamente, a arquitetura multiagente incorre em custos que são, em média, 3,7 vezes maiores do que os da configuração de agente único, devido ao aumento do número de tokens processados. Por outro lado, os sistemas de agente único se destacam em tarefas de texto para SQL (Linguagem de Consulta Estruturada), especialmente ao usar o Transformador Pré-Treinado Generativo 4 (GPT-4), alcançando uma pontuação 15% maior em comparação com as configurações multiagentes, sugerindo que arquiteturas mais simples podem, às vezes, superar a complexidade. A novidade deste trabalho reside em seu exame original dos desafios específicos apresentados pelos dados complexos, técnicos e não estruturados inerentes às operações de construção de poços, contribuindo para o planejamento estratégico da adoção de aplicações de IA generativa, fornecendo uma base para otimizar soluções contra parâmetros econômicos e tecnológicos.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

COMPARATIVE ANALYSIS OF SINGLE AND MULTI-AGENT LARGE LANGUAGE MODEL ARCHITECTURES FOR DOMAIN-SPECIFIC TASKS IN WELL CONSTRUCTION

Vitor Brandão Sabbagh

July/2025

Advisor: Geraldo Bonorino Xexéo

Department: Systems Engineering and Computer Science

This article explores the application of large language models (LLM) in the oil and gas sector, specifically within well construction and maintenance tasks. The study evaluates the performances of a single-agent and a multi-agent LLM-based architecture in processing different tasks, offering a comparative perspective on their accuracy and the cost implications of their implementation. The results indicate that multi-agent systems offer improved performance in question and answer tasks, with a truthfulness measure 28% higher than single-agent systems, but at a higher financial cost. Specifically, the multi-agent architecture incurs costs that are, on average, 3.7 times higher than those of the single-agent setup due to the increased number of tokens processed. Conversely, single-agent systems excel in text-to-SQL (Structured Query Language) tasks, particularly when using Generative Pre-Trained Transformer 4 (GPT-4), achieving a 15% higher score compared to multi-agent configurations, suggesting that simpler architectures can sometimes outpace complexity. The novelty of this work lies in its original examination of the specific challenges presented by the complex, technical, unstructured data inherent in well construction operations, contributing to strategic planning for adopting generative AI applications, providing a basis for optimizing solutions against economic and technological parameters.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

In the dynamic and ever-changing oil and gas (O&G) industry, ⚡[1] ⚡[2] digital transformation has emerged as a key element to achieve operational efficiency, sustainability, and competitiveness. At the forefront of this transformation are Large Language Models (LLMs), which have the potential to process unstructured queries, map out alternatives, and advise users on possible actions (KAR and VARSHA, 2023). We also note the advantage of increased engagement, cooperation, accessibility, and ultimately profitability. These models redefine paradigms in knowledge management and information retrieval and impact a variety of other areas (ECKROTH and GIPSON, 2023), making it crucial to adopt these technologies to remain competitive.

A study conducted by DELLACQUA *et al.* (2023), in collaboration with the Boston Consulting Group, shows that in knowledge-intensive tasks, consultants equipped with access to LLMs such as GPT-4 not only completed tasks more efficiently (25.1% more quickly on average) but also with substantially higher quality, achieving results more than 40% better compared to those without AI assistance (DELLACQUA *et al.*, 2023). Increase in productivity of knowledge workers was 12% on average. A major oil company spent in 2023 $2.8B with employee compensation (PETROBRAS, 2024). A potential increase of 12% in knowledge workers productivity, given they represent 60% of all employee, could represent $204M annual savings in this scenario. ⚡[3] ⚡[4]

Broader economic indicators predict significant transformations due to generative AI (Gen-AI) across various industries. A report from Goldman Sachs (HATZIUS

---

[1]X:eu acho que aqui falta alto. TEm um pouco há ver com o uso do "in the dynamic changing of the" não ser realmente muito adequado a "oil and gas industry", falta algo, um qualificador a mais (mercado? . Minha IA sugeriu mudar para "In the dynamic and ever-changing oil and gas (O&G) industry", o que eu achei muito melhor

[2]V:feito

[3]X:Aqui está um problema clássico. Você quer fornecer um dado mas provavelmente não quer dizer que é a Petrobras. Mas provavelmente esse dado está em algum relatório público. Não consegue ele de algum lugar e aí pode citar?

[4]V:feito

*et al.*, 2023) highlights that Gen-AI is poised to increase global GDP by nearly 7%, increasing productivity growth by 1.5 percentage points over the next decade. This economic uplift is expected due to AI's ability to automate complex workflows and create new business opportunities, significantly impacting employment and productivity sectors worldwide.

Expanding on the broader discussion on data utilization within organizations, an important issue is the challenge of extracting relevant information from extensive databases (SINGH *et al.*, 2023). Initially, the challenge of knowing, finding, and accessing data poses a significant obstacle to decision-making processes. Collaborators at O&G companies often face the intensive task of manually searching large data repositories to find useful information.

Focusing specifically on the activities of drilling and completion of offshore and onshore wells, a major challenge lies in the inherently complex and technical nature of the data involved, which can be from various types: operations, projects, technologies, supply chains, and others. Inefficiency in leveraging large volumes of unstructured data worsens these challenges, as observed by SINGH *et al.* (2023). A significant amount of the data generated and collected in this sector is unstructured, ranging from text reports and emails to images and videos of exploration and production activities. Examples include hundreds of daily operational reports from drilling rigs, well execution projects, nonproductive time (NPT) reports, and operational lessons learned documents, as illustrated in Figure 1.1. As a result, valuable information can remain untapped, and the potential to find insights, informed decision-making, and innovation is significantly compromised. SINGH *et al.* (2023) showcases the capabilities and potential of Generative AI-enabled chatbots for the O&G sector, particularly in enhancing drilling and production analytics to achieve better business results. The author concludes that companies that adopt these technologies in the coming years will see clear advantages.

However, the deployment of such technologies presents limitations and introduces challenges, including biased data, hallucinations, lack of explainability, and logical reasoning errors, among others (HADI *et al.*, 2023), which require a balanced approach to harness their potential in a responsible manner. Although previous research has focused mainly on the broader applications of AI in industry, the novelty of our research lies in its original examination of the specific challenges and solutions presented by the complex, technical and unstructured data inherent in O&G operations. By comparing single- and multi-agent systems, this study fills a knowledge gap, providing empirical insights into the effectiveness of different Gen-AI architectures in a domain where such studies are scarce.

The adoption of these technologies by a major oil company underscores their potential to revolutionize data analysis and management, presenting an opportunity

Figure 1.1: Sample of drilling & completion learned lesson partial document. (translated from Portuguese)

for deeper exploration and application.

## 1.1 Business Scope Delimitation

To contextualize the scope of this study, it is necessary to understand the life cycle of an oil field, which begins with Exploration and progresses to the Development of Production, followed by effective production, and culminates in decommissioning (BADIRU and OSISANYA, 2016). Gen-AI has the potential to impact each of these phases, but the focus of this work lies in the operations of the development and maintenance stages.

Well construction is a highly specialized activity that involves drilling and completion of wells for hydrocarbon extraction (THOMAS, 2004). In this context, Gen-AI can be applied in various ways. For example, a chatbot could manage knowledge by answering queries about operations and well projects by retrieving information from the organization's databases. Additionally, LLM-based agents could be used in executive project review to ensure that drilling or completion operations comply with the organization's standards and adhere to best operational practices. Moreover, Gen-AI could perform inference in unstructured databases to extract specific information from text reports and obtain structured data. This business scope emphasizes the importance of Gen-AI in the construction and maintenance of wells.

### 1.1.1 Key Information Sources in Well Engineering

To fully appreciate the challenges in this domain, it is important to understand the primary data sources that specialists interact with daily. The following sources, used in this research's experiments, exemplify the complex information landscape of well engineering:

**Operational Learned Lessons** During drilling, completion, and workover interventions, documents called Knowledge Items are written by specialists, as depicted in Fig 1.1. These can be of four types: Technical Alert, Learned Lesson, Good Practice, and Well Observation. This system serves as a critical tool for knowledge management, considering the large number and variety of specialists involved and well operations performed.

**Operational NPTs (Non-Productive Time)** This data source contains structured records of anomalies that occurred during well interventions, detailing the title, description, location, operation type, responsible sector, rig involved, time lost, and event dates. These data are critical for the industry, as NPTs represent periods when operations are interrupted. The identification and analysis of these events are essential for continuous process improvement, cost reduction, and increased operational efficiency.

**Collaborator Finder** The third data source is a collaborator finder, an important internal tool for consulting and managing employee data. This system allows for the quick identification of employees through information such as name, workplace, and role. The importance of this tool lies in the ability to cross-reference employee data with operational events, enabling a more complete analysis by an intelligent agent.

## 1.2 Objectives

This research directly addresses the challenges facing major oil companies. By investigating the comparative advantages and limitations of various Gen-AI architectures, including single and multi-agent systems, for Q&A and Text-to-SQL tasks, this study aims to identify the most efficient and cost-effective solutions.

The specific objectives of this research are to assess the suitability and effectiveness of multi-agent systems based on LLMs for complex, domain-specific tasks in well engineering, aiming to streamline information access and decision-making.

The study will compare single-agent and multi-agent AI systems in terms of their ability to address well engineering queries. Finally, it will map the potential

obstacles and limitations associated with deploying Gen-AI applications.

The insights gained from this research will directly contribute to O&G companies strategic goals by improving access to well engineering information and automated data analysis tasks. A comprehensive understanding of the challenges and limitations associated with Gen-AI will enable informed decisions about its adoption, maximizing the return on investment.

To achieve these objectives, this research was conducted through two distinct experimental phases. The first, carried out in 2024, focused on a foundational comparison between single and multi-agent architectures, which revealed that, although the multi-agent architecture achieved 28% higher truthfulness in Q&A tasks, its cost was on average 3.7 times higher. Furthermore, the single-agent architecture proved to be surprisingly more effective in Text-to-SQL tasks. The rapid evolution of generative AI frameworks and models prompted a second, more advanced experiment in 2025. This second phase built upon the initial findings, also employing non-agentic workflows as baseline and a more rigorous, quantitative evaluation methodology to address the challenges identified in the first experiment and automated evaluation based on the concept commonly reffered to as "LLM-as-judge" ((GU *et al.*, 2025)).

✐[5] ✐[6]                                                                          X:5

V:6

«PENDENTE»»

FAZER ALINHAMENTO FINAL (QUESTÃO DE PESQUISA - EXPERIMENTO - CONCLUSÕES)

DETALHAMENTO LLM: "BOTAR DEPOIS DAS QUESTÕES DE PESQUISA" E "SEMPRE ALINHANDO TUDO" ESTA É A "LINHA DE OURO" (GOLDEN THREAD) DE UMA DISSERTAÇÃO. TUDO PRECISA ESTAR CONECTADO. A ESTRUTURA QUE ELE SUGERE CRIA UM FLUXO MUITO LÓGICO PARA O LEITOR:

OBJETIVOS: O QUE VOCÊ QUER ALCANÇAR (VISÃO GERAL).

QUESTÕES DE PESQUISA (A SEREM ADICIONADAS): AS PERGUNTAS ESPECÍFICAS E FOCADAS QUE SUA PESQUISA VAI RESPONDER.

PARÁGRAFO COM "SPOILER" (O QUE ESTAMOS DISCUTINDO): UM RESUMO DE COMO VOCÊ RESPONDEU A ESSAS PERGUNTAS E O QUE ENCONTROU.

CONCLUSÕES (NO FINAL DA DISSERTAÇÃO): ONDE VOCÊ RESPONDE FORMALMENTE ÀS QUESTÕES DE PESQUISA, USANDO OS RESULTADOS DETALHADOS DOS EXPERIMENTOS.

AO COLOCAR O PARÁGRAFO DE "SPOILER" DEPOIS DAS QUESTÕES

---

[5]X:Acho que pode aumentar um pouco isso, já falar dos resultados (dissertação tem spoiler), motivar o segundo a partir dos resutlados do primeiro e botar depois das questões de pesquisa (leia o todo), sempre alinhando tudo (questão de pesquisa - experimento - conclusões

[6]V:Incluído spoiler e ponte do 1o justificando o 2o experimento.

DE PESQUISA, VOCÊ CRIA UMA CONEXÃO DIRETA: "PARA RESPON-
DER A ESTAS PERGUNTAS (QUESTÃO 1, QUESTÃO 2...), EU CONDUZI
ESTES EXPERIMENTOS (EXPERIMENTO 1, EXPERIMENTO 2), QUE ME
LEVARAM A ESTAS DESCOBERTAS PRINCIPAIS (RESULTADO A, RESUL-
TADO B)."

ESSA ESTRUTURA AMARRA TODA A SUA DISSERTAÇÃO, TORNANDO-
A COESA, LÓGICA E FÁCIL DE ACOMPANHAR. ««FIM»»

Ok, está bom, porém seria melhor para dissertação se agora você definisse questões de pesquisa. Essas questões de pesquisa serão respondidas na conclusão, a partir do que você fez. Eu só li até aqui, então não tenho sugestões fortes agora, mas as questões podem ser coisas coisa: Qual a eficácia e eficiência de LLMs para extrair dados de bases .... Como sistemas single e multi agentes se comparam... Elas podem ser bem melhores e bem mais objetivas e ao longo do texto, se eu detectar alguma , escrevo. A questão é perguntar aqui no fim da introdução e responder na conclusão, caracterizando a colaboração

SERÁ FEITO NO FIM

## 1.3 Research Methodology

This research follows the Design Science Research (DSR) methodology, a framework particularly suited for studies that develop and evaluate technological artifacts to address specific organizational problems. DSR provides a structured approach for creating innovative solutions while maintaining scientific rigor through empirical validation (HEVNER, 2007).
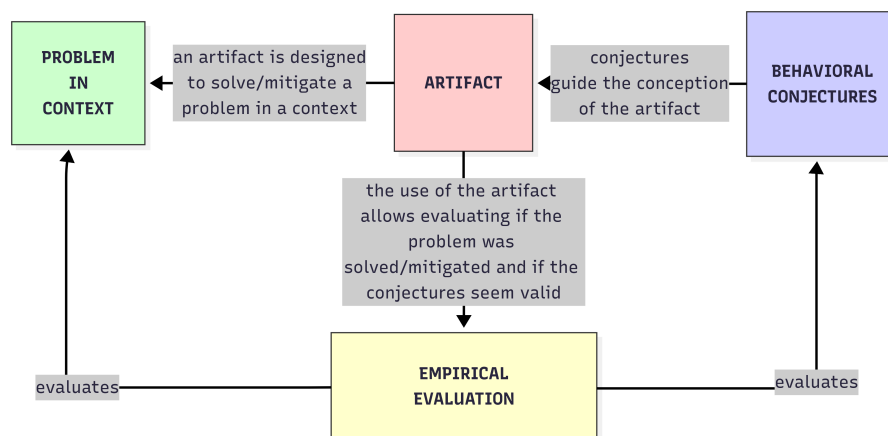


Figure 1.2: Main elements of DSR-Model, translated from OSWALD *et al.* (2023).

## 1.3.1 Design Science Research Framework

The DSR methodology employed in this study consists of four interconnected elements, as illustrated in Figure 1.2:

1. **Problem in Context**: Identifying and defining a relevant organizational challenge within its specific environment

2. **Artifact**: Designing and developing a technological solution to address the identified problem

3. **Behavioral Conjectures**: Formulating hypotheses about how the artifact will function and impact the problem space

4. **Empirical Evaluation**: Systematically testing the artifact to validate its effectiveness and the underlying conjectures

This cyclical framework guides both the research design and execution, ensuring that the developed artifacts are not only technically sound but also practically relevant.

## 1.3.2 Application of DSR in This Research

**Problem in Context**

This study addresses the challenge of efficiently extracting relevant information from extensive technical databases in the oil and gas industry, specifically in well construction and maintenance operations.

Table 1.1: Characteristics of the Problem Context

| Challenge Aspect | Description |
|---|---|
| Data Structure | Large volumes of unstructured data (operational reports, lessons learned documents, NPT reports) |
| Technical Complexity | Domain-specific terminology and complex relationships |
| Business Impact | Significant potential economic impact from improved knowledge access |

**Artifacts**

Two primary artifacts were designed and implemented, illustrated in Figure 1.3, using state-of-the-art language models (GPT-3.5-turbo and GPT-4) and integrated with domain-specific knowledge bases through various retrieval mechanisms.

| Single-Agent LLM System | Multi-Agent LLM System |
|---|---|
| A centralized architecture where one language model agent handles the entire question-answering process with access to multiple tools | A collaborative architecture where multiple specialized agents work together under coordination to process queries |
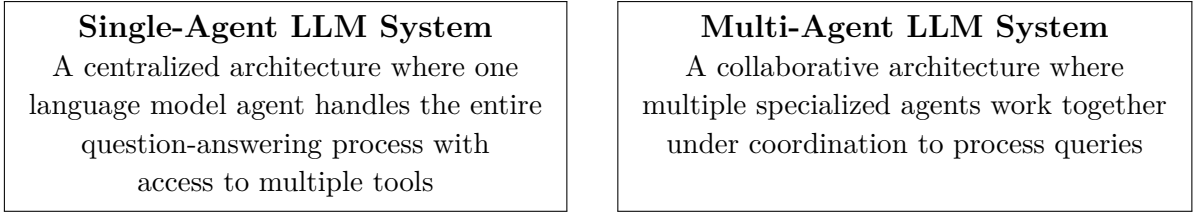
Figure 1.3: Primary Artifacts Developed in This Research

**Behavioral Conjectures**

The research was guided by several key conjectures:

> **Key Research Conjectures**
>
> - Multi-agent systems will demonstrate higher accuracy in complex technical queries due to their ability to distribute cognitive load and specialize in different aspects of the problem
>
> - The performance advantages of multi-agent systems will vary by task type (Q&A vs. Text-to-SQL)
>
> - More advanced language models will yield better performance but at significantly higher LLM financial costs
>
> - The economic efficiency (performance-to-cost ratio) will be a critical factor in determining practical implementation viability

**Empirical Evaluation**

The evaluation was conducted through two distinct experimental phases (summarized in Table 1.2), allowing for iterative refinement of both the artifacts and the evaluation methodology, addressing limitations identified in the first experiment while adapting to the rapid evolution of language model capabilities.

Table 1.2: Comparison of Experimental Phases

| Aspect | First Experiment (2024) | Second Experiment (2025) |
|---|---|---|
| Focus | Comparative analysis of single and multi-agent architectures | Extended evaluation incorporating non-agentic workflows as baseline |
| Evaluation Methods | Expert validation by domain specialists | Automated assessment using "LLM-as-judge" approach |
| Metrics | Truthfulness, performance, and LLM cost | Precision, recall, and F1-score |
| Outcomes | Identification of key challenges and limitations | More rigorous quantitative evaluation methodology |

## 1.4 Thesis Structure

****SERÁ FEITO POR ÚLTIMO****

# Chapter 2

# Literature Review

This chapter provides a comprehensive literature review of the key technologies and concepts that form the foundation of this dissertation. It begins with an overview of the applications of Artificial Intelligence (AI) in the Exploration and Production (E&P) industry. The focus then narrows to Large Language Models (LLMs), discussing their architecture and impact. Subsequently, the chapter delves into the Retrieval-Augmented Generation (RAG) technique, which enhances LLMs with external knowledge. It also explores the use of single and multi-agent setups. Finally, the chapter concludes by examining the 'LLM-as-judge' paradigm for evaluating the performance of generative models.

## 2.1 AI in the Exploration and Production (E&P) industry

The use of AI in the Exploration and Production (E&P) industry has been extensive. In the last decades the majority of AI applications in the industry involved data mining and neural networks (BRAVO *et al.*, 2014). An example is the work by (GUDALA *et al.*, 2021) on optimization of the properties of the heavy oil flow, through the use of neural networks to optimize flow-influencing parameters. Another development was a deep learning workflow proposed by (GOHARI *et al.*, 2024), with the generation of synthetic graphic well logs through the application of transfer learning. These developments illustrate the potential of AI to improve processes and the accuracy and efficiency of data analysis (RAHMANI *et al.*, 2021).

Natural Language Processing (NLP) stands at the intersection of computer science and linguistics, representing a domain within artificial intelligence aimed at enabling computers to understand and process human language in a way that is both meaningful and effective (LIDDY, 2001). This field integrates a diverse range of computational techniques to analyze and represent text at various levels of lin-

guistic detail, striving to emulate human-like language understanding. As an active area of research, traditionally NLP employs multiple layers of language analysis, each contributing uniquely to the interpretation and generation of language, which finds practical applications in various sectors (LIDDY, 2001). In the O&G industry, the management of unstructured data, such as texts, images, and documents, is crucial, with Natural Language Processing (NLP) and Machine Learning playing key roles. Research by ANTONIAK *et al.* (2016) and CASTIÑEIRA *et al.* (2018) has explored the use of NLP to analyze risks and drilling reports.

## 2.2 Natural Language Processing

NLP (Natural Language Processing) is a broad field that covers various tasks to enable computers to process and understand human language. These tasks, which represent specific problems or applications, have been the focus of research for decades, predating the recent surge in Large Language Models. They range from fundamental challenges like part-of-speech tagging to complex applications like machine translation. This section explores two tasks particularly relevant to this dissertation: Question Answering (Q&A) and Text-to-SQL, both of which have been significantly advanced by recent developments in the field.

### 2.2.1 Q&A tasks

Question and Answer (Q&A) represent a method to facilitate knowledge transfer between individuals within organizations (ISKE and BOERSMA, 2005). [7] [8] Conceptually, Q&A systems are designed to connect individuals who possess specific knowledge with those seeking that knowledge through a structured question-and-answer format. The role of Q&A in the documentation landscape, as exemplified by platforms such as Stack Overflow, highlights their significance in technical disciplines (TREUDE *et al.*, 2011). This understanding can guide organizations in making more informed decisions about implementing such systems to enhance knowledge transfer and organizational learning (ISKE and BOERSMA, 2005).

X:7
V:8

### 2.2.2 Text-to-SQL tasks

Text-to-SQL tasks in the context of artificial intelligence involve the automatic translation of natural language questions or commands into structured SQL (Structured

---

[7]X:As tasks vem antes das LLMs, elas sempre existiram como problemas da área de NLP. Inclusive acho que na seção de NLP você pode fazer um parágrafo sobre a existência de várias tasks e usar essas como subseções

[8]V:Feito

11

Query Language) queries (QIN *et al.*, 2022). This is an important area in natural language processing (NLP), allowing users to interact with databases using plain language rather than needing to know how to write complex SQL queries.

The arrival of advanced language models like GPT-3 and GPT-4 (OPENAI, 2023) has marked a significant leap in Text-to-SQL applications (SINGH *et al.*, 2023), demonstrating remarkable capabilities in handling these tasks. This can be attributed to their extensive training on diverse datasets (DENG *et al.*, 2021), which include not only large amounts of text but also structured data like tables and code, enabling the model to understand the intricate relationships between language and data structures. The study by (DENG *et al.*, 2023) introduces a pre-training framework for text to SQL translation, emphasizing the alignment between text and tables in Text-to-SQL tasks.

## 2.3 Intelligent Agents

According to RUSSELL (2020), an agent is something that performs actions. When it comes to computerized agents (in our case, AI-based), these agents are expected to do more: operate autonomously, perceive the environment, persist over time, adapt to changes, create, and strive to achieve goals. The agent program implements the agent function. There is a variety of basic agent program designs that vary in efficiency, compactness, and flexibility. The appropriate design of the agent program depends on the nature of the environment. In this work, a goal-based agent design was implemented, which acts to achieve defined goals (RUSSELL, 2020). Other possible types include simple reflex agents, which directly respond to perceptions, while model-based reflex agents maintain an internal state to track aspects of the world that are not evident in the current perception. Finally, there are utility-based agents, which try to maximize their expected "happiness" (RUSSELL, 2020).

### 2.3.1 Multi-Agent Systems

A Multi-Agent System (MAS) extends the concept of a single agent to a collection of agents that interact within a shared environment (GOKULAN and SRINIVASAN, 2010). A MAS is defined as a loosely coupled network of autonomous problem-solving entities that collaborate to find solutions to problems that are beyond the individual capabilities or knowledge of any single entity (FLORES-MENDEZ, 1999).

The structure of a MAS can vary, with different organizational paradigms such as hierarchical structures or coalitions being employed depending on the application (GOKULAN and SRINIVASAN, 2010). A practical example of a MAS architecture is demonstrated in power system restoration, where a system can be composed

of multiple "bus agents" and a single "facilitator agent" (NAGATA and SASAKI, 2002). In this setup, each bus agent works to restore its local area by negotiating with neighboring agents based on locally available information, while the facilitator agent manages the overall decision-making process, showcasing how a collection of agents with simple, local strategies can cooperate to achieve a complex, global goal (NAGATA and SASAKI, 2002).

## 2.4   Large Language Models

Large Language Models (LLMs) are advanced neural network-based models designed to understand and generate human-like text. They leverage the Transformer architecture introduced in the seminal paper "Attention is All You Need" by VASWANI *et al.* (2017). This architecture relies on self-attention mechanisms, allowing the model to weigh the importance of different words in a sentence effectively.

The emergence of LLMs has made it possible to understand and produce textual information. These systems are expected to revolutionize various industries by supporting complex decision-making processes. GPT models (OPENAI *et al.*, 2023), in particular, take advantage of its vast training data to provide human-like responses (MOSSER *et al.*, 2024), which can be highly beneficial in contexts requiring natural language understanding and generation. The exponential growth in the size and capability of LLMs in recent years has been remarkable. Models like OpenAI's GPT series have shown significant advancements, moving from millions to hundreds of billions of parameters, which gives them increasingly sophisticated natural language understanding and generation. This advancement is illustrated in Figure 2.1. For new models (released after jan/2025), including OpenAI's o3 series and GPT-4.5, Anthropic's Claude 3.7 and 4, and Google's Gemini 2.5 Pro, the exact parameter counts have not been publicly disclosed.

✒[9] ✒[10]                                                                X:9

However, the trajectory of LLM development in 2025 has signaled a shift in fo-   V:10
cus. While previous advancements were often marked by an exponential increase in parameter counts, the latest generation of models emphasizes sophisticated reasoning capabilities over sheer size. This move away from parameter size as the primary metric of progress underscores a new trend: enhancing the models' ability to perform complex, multi-step reasoning. This is evident in features like the private chain-of-thought mechanisms in OpenAI's models and the "extended thinking" mode in Anthropic's Claude series, indicating that language models are advancing through

---

[9]X:Acho que aqui merecia um gráfico do crescimento do tamanho das LLMs e um parágrafo sobre esse crescimento

[10]V:Feito
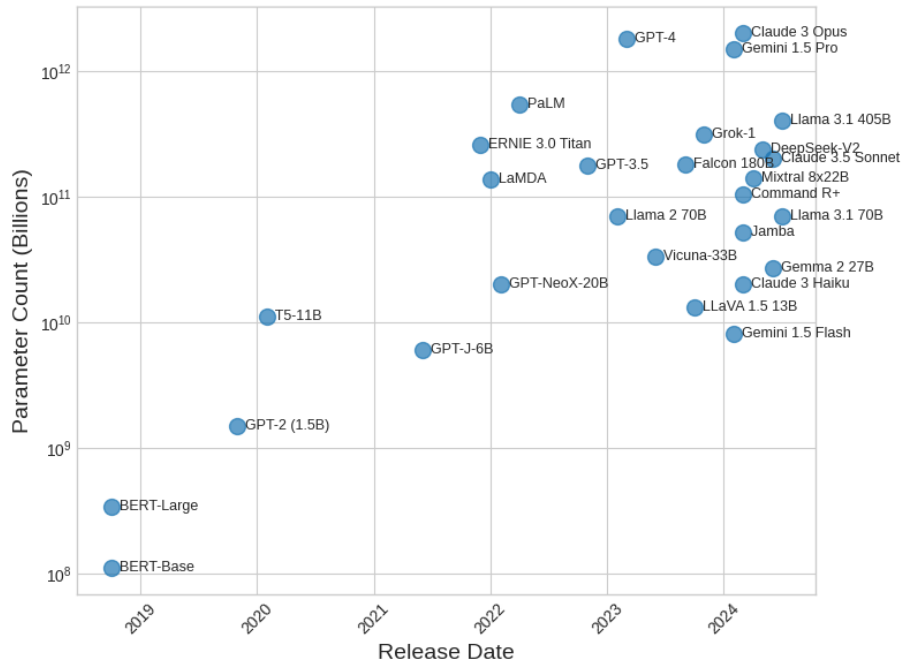
Figure 2.1: The evolution of LLMs.

more intricate cognitive architectures rather than just scaled-up data processing.

As highlighted by SINGH *et al.* (2023), the integration of LLM-based solutions, such as conversational chatbots, offers an approach to optimizing operations across various business segments, including drilling, completion, and production. SINGH *et al.* (2023) uses LLMs models to extract, analyze, and interpret datasets, enabling generation of insights and recommendations.

Despite its widespread impact, language models are not without its limitations. In many industry-specific applications, the critical information required is often proprietary, not shared with third parties, and thus absent from the training data of these LLMs (MOSSER *et al.*, 2024). This gap means that GPT models might not have access to the most up-to-date or sensitive information needed for certain tasks. Moreover, due to their probabilistic nature, LLMs can experience hallucinations, producing confident yet incorrect or nonsensical responses based on user input (OPENAI *et al.*, 2023).

### 2.4.1 LLM applications

[11] [12]                                                                                    X:11

V:12

The proliferation of LLMs has led to a diverse array of applications that leverage their ability to understand, generate, and process human language.

The expansion of the LLM application ecosystem is evident in the significant

---

[11]X:Precisa de um texto aqui
[12]V:Feito

14

market growth projections. For instance, one report projects the global LLM market to grow from \$5.62 billion in 2024 to \$35.43 billion by 2030, with a compound annual growth rate (CAGR) of 36.9% (GRANDVIEWRESEARCH, 2025). This rapid expansion is indicative of the immense value and potential that organizations across industries see in these technologies. The applications themselves are becoming increasingly sophisticated, evolving from simple text generation to complex, multimodal systems capable of processing and integrating text, images, and other data formats (KADDOUR *et al.*, 2023).

The spectrum of LLM-based applications is broad and continually expanding. Early applications focused on tasks such as text summarization, translation, and sentiment analysis. However, the current generation of LLMs powers a much wider range of tools. These can be broadly categorized into several key areas. Conversational AI, in the form of advanced chatbots and virtual assistants, represents a significant segment of the market, enhancing customer service and user engagement (GRANDVIEWRESEARCH, 2025). Content creation is another major application area, where LLMs are employed to generate a variety of materials, from marketing copy and social media posts to technical documentation and even creative writing (LABS, 2025).

Furthermore, LLMs are being integrated into more specialized and high-stakes domains. In the legal field, they assist with tasks like contract analysis and legal research. The financial sector utilizes them for fraud detection and market analysis (LABS, 2025). In software development, LLM-powered tools for code generation and debugging are becoming increasingly prevalent, accelerating development cycles and improving programmer productivity. A key innovation driving the utility of these applications is the advent of techniques like Retrieval-Augmented Generation (RAG), which allows LLMs to retrieve and incorporate information from external knowledge bases, thereby improving the accuracy and relevance of their outputs (AI, 2025). The ongoing development of multimodal LLMs is further pushing the boundaries of what is possible, enabling applications that can understand and reason about the world in a more holistic manner (KADDOUR *et al.*, 2023).

## 2.4.2  Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG) technique combines LLMs with information retrieval to generate accurate and up-to-date responses, as introduced by LEWIS *et al.* (2020). [13] [14] It employs a search in a database to find relevant X:13 information, overcoming the inherent limitations of LLMs that rely solely on the V:14 prior knowledge embedded in the language model during the training phase. With

---

[13]X:Aqui merece um desenho ilustrativo, até para quebrar tanto texto
[14]V:Feito

the ongoing evolution of information retrieval, which has moved from term-based methods to more semantic approaches leveraging deep learning and large datasets to tackle more complex challenges.

A RAG consists of two main components: a retriever and a generator, as illustrated in Figure 2.2. The retriever is responsible for finding relevant information from a knowledge base, and the generator uses that information to create a human-like response.
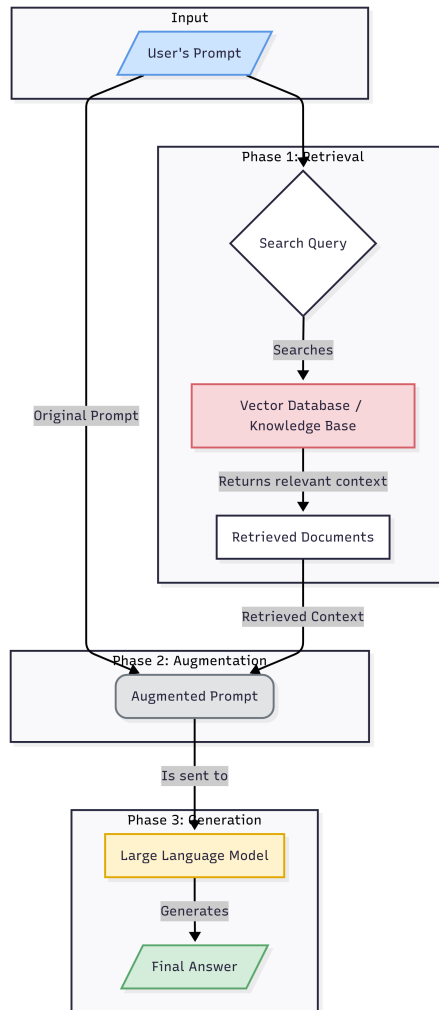


Figure 2.2: A diagram illustrating the RAG process.

As elucidated by LEWIS *et al.* (2020), RAG unites the strengths of pre-trained parametric and non-parametric memory, using a dense vector index and a semantic retriever. As demonstrated by LI *et al.* (2022) in their analysis, RAG is surpassing traditional generative models in terms of performance across a variety of tasks. The study provides a detailed survey on this topic, emphasizing the fundamental concepts and its applicability in specific contexts.

New tools have been developed to facilitate the implementation of RAG solutions. LIU *et al.* (2023) present a toolkit that integrates augmented retrieval

techniques into LLMs, including modules for query rewriting, document retrieval, passage extraction, response generation, and fact-checking, enabling the creation of more factual and specific responses. The recent study by ZHAO *et al.* (2023) extends this horizon by examining the incorporation of multimodal knowledge into generative models, exploring the integration of diverse external sources such as images, code, tables, graphs, and audio, to enhance the grounding context and improve usability. It also explores potential future trajectories in this emerging field, marking a relevant contribution to the evolving narrative of RAG and its applications.

### 2.4.3 Multi-Agent Setup

🔥[15] 🔥[16]

As demonstrated by XI *et al.* (2023), the pursuit of Artificial General Intelligence (AGI) has significantly benefited from the development of LLM-based agents, capable of sensing, decision-making, and acting across diverse scenarios. His study outline a foundational framework for such agents, consisting of brain, perception, and action components, which can be customized for various applications including single-agent scenarios, multi-agent systems, and human-agent collaboration . The comprehensive survey underscores the crucial role of LLMs in moving towards AGI, suggesting a promising horizon for operational efficiency and decision-making processes in complex organizational settings (XI *et al.*, 2023).

LI *et al.* (2024a) demonstrated that, through a sampling and voting method, the performance of LLMs scales with the number of instantiated agents. Another open-source framework is AutoGen (WU *et al.*, 2023), that enables the creation of LLM multi-agent applications, allowing for customization across various modes including. It supports diverse applications in fields such as mathematics, coding, and operations research, demonstrating its effectiveness through empirical studies (WU *et al.*, 2023).

## 2.5 Evaluation

### 2.5.1 Truthfulness

In the evaluation of RAG systems, ensuring the truthfulness of the generated output is a primary concern. LIN *et al.* (2022) introduces a framework for this purpose. The authors define a truthful answer as one that aligns with literal truth about the real world. This is particularly relevant for RAG systems, which can retrieve

---

[15]X:Isso aqui seria uma seção de multi agentes dentro da LLM, mas você precisa escrever pelo menos um parágrafo de multi agentes gerais na seção agentes
[16]V:Feito. Inclui uma subseção sobre MAS.

and incorporate information from vast and varied sources. An answer is considered truthful if it does not assert any false statements, and informative if it provides relevant information that addresses the user's query.

In LI *et al.* (2023), the authors conducted an evaluation to determine the effectiveness of their proposed prompts on the performance of various LLMs. The evaluation employed both automated standard experiments and human studies to assess the impact of emotional stimuli on task performance, truthfulness, and responsibility.

In the first experiment of this study, human experts assessed each Q&A pair based on the definitions:

- ***Truthfulness****: a metric to gauge the extent of divergence from factual accuracy, otherwise referred to as hallucination (LIN* et al.*, 2021).*

  *1="The response promulgates incorrect information, detrimentally influencing the ultimate interpretation"*

  *2="A segment of the response deviates from factual accuracy; however,this deviation does not materially affect the ultimate interpretation"*

  *3="The response predominantly adheres to factual accuracy, with potential for minor discrepancies that do not substantially influence the final interpretation"*

  *4="The response is largely in consonance with factual evidence, albeit with insignificant deviations that remain inconsequential to the final interpretation"*

  *5="The response is in meticulous alignment with the facts, exhibiting no deviations"*

- ***Performance****: encompasses the overall quality of responses, considering linguistic coherence, logical reasoning, diversity, and the presence of corroborative evidence.*

  *1 = "The response fails to address the question adequately"*

  *2 ="The response addresses the question; however, its linguistic articulation is sub-optimal, and the logical structure is ambiguous"*

  *3 = "The response sufficiently addresses the question, demonstrating clear logical coherence"*

  *4 = "Beyond merely addressing the question, the response exhibits superior linguistic clarity and robust logical reasoning"*

  *5 = "The response adeptly addresses the question, characterized by proficient linguistic expression, lucid logic, and bolstered by illustrative examples"(LIN* et al.*, 2021).*

### 2.5.2 Precision, Recall, and F1-Score

Precision, recall, and F1-score are fundamental metrics for evaluating classification tasks, particularly in scenarios with imbalanced datasets. These metrics provide a more nuanced understanding of a model's performance than accuracy alone.

**Precision** measures the accuracy of positive predictions. It is the ratio of correctly predicted positive observations to the total predicted positive observations. A high precision relates to a low false positive rate.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2.1}$$

**Recall** (or Sensitivity) measures the ability of the model to find all the relevant cases within a dataset. It is the ratio of correctly predicted positive observations to all observations in the actual class.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2.2}$$

The **F1-score** is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. It is the harmonic mean of the two and is a good way to show that a model has a good performance on both metrics.

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.3}$$

### 2.5.3 LLM-as-judge

[17] [18]

The LLM-as-Judge paradigm represents a significant shift in the evaluation of NLP systems in general, using a language model as a scalable proxy for human evaluators ((LI *et al.*, 2024b)). This approach was developed to overcome the semantic shallowness of traditional metrics like BLEU or ROUGE and the logistical challenges of extensive human annotation ((ZHENG *et al.*, 2023)). By providing a "judge" LLM with a clear rubric and context, it can perform assessments of qualities like coherence, relevance, and factual accuracy ((LI *et al.*, 2024b)). This method has proven effective for complex, open-ended tasks where simple string matching is insufficient, with models like GPT-4 demonstrating over 80% agreement with human preferences in benchmarking studies (ZHENG *et al.*, 2023).

---

[17]X:Òk, essa seção eu não gostei muito, apesar de não ter nenhum erro. Primeiro tem que fazer uma seção de avaliação com as medidas, onde devem estar todas as medidas que você usar na dissertação (não li ainda, estou indo na ordem). Aí então você pode falar disso, mas apenas se usou

[18]V:Enxuguei esta seção e fiz a inclusão das demais métricas utilizadas antes desta seção. Veja se está melhor.

For evaluating Retrieval-Augmented Generation (RAG) systems, the LLM-as-Judge framework can be adapted to produce structured, quantitative assessments. In this application, the judge LLM is tasked with comparing the RAG-generated answer against a ground-truth dataset. By using a crafted prompt that defines the classification criteria, the judge can systematically categorize each output into classes such as True Positive (TP) (factually consistent with the ground truth), False Positive (FP) (introduces unsupported information), True Negative (TN) (a correct refusal to answer), or False Negative (FN) (missing relevant information). This approach moves beyond subjective scoring towards a more objective evaluation. The prompt used in this work is presented in the code in Appendix A.1.

The advantage of this methodology is its ability to translate qualitative judgments directly into a confusion matrix, allowing the calculation of standard metrics such as precision (Equation 2.1), recall (Equation 2.2), and F1-score (Equation 2.3). This process establishes a replicable pipeline for benchmarking the factual accuracy of a RAG system at scale. While it is important to acknowledge the potential for inherent biases in LLM judges ((GU *et al.*, 2025)), studies show high correlation with human-expert evaluations ((LI *et al.*, 2024b)), making it a useful tool for iterative development and system comparison.

> Como você criou um dataset de teste para uma tarefa, seria bom falar disso. Mas essa é uma coisa adicional a outras mudanças que pedi, seria bom, mas se não der tempo, não deu.
>
> VITOR: O DATASET ENTRARIA NO CAP. 3 E 4 OU AQUI MESMO?

[19] [20]

---

# Chapter 3

# First Experimental Evaluation Cycle

[21] [22]
  [23] [24]
  [25] [26]
  [27] [28]

This chapter describes the first experimental cycle of this research, as introduced in Section 1.3.2, conducted to investigate the effectiveness of different LLM based agent architectures. The primary objective is to address complex, domain-specific queries within the field of well construction and maintenance. This initial cycle serves as a foundational study, comparing single-agent and multi-agent systems to generate empirical insights into their performance, cost, and inherent limitations. The findings from this cycle will inform the more advanced, quantitative evaluation performed in the second experiment.

Following the principles of DSR, this chapter is structured to clearly present the research components. We will begin by defining the business context and the specific problem this experiment aims to solve. Subsequently, we will describe the design of

---

[21]X:Todo capítulo deve ter uma introdução explanatória. "This chapter describes"

[22]V:Feito.

[23]X:Acho que pode até ser Primeiro Ciclo, ou pode ter um nome como "Efetividada das LLMs na Solução.." vai ter que quebrar esse capítulo, que tem muita informação nos conceitos da DSR: no mínimo nos quatro principais: Contexto, Problema, Artefato, Avaliação. Grande parte do contexto e problema já devem ser descritos no capitulo que eu pedi para criar e aqui só faz referência.

[24]V:O Capítulo foi refatorado para ficar alinhado com a DSR.

[25]X:Ok, você foi direto para o experimento mas não disse o que ia fazer. Aqui exatamente cabe o quadro da DSR que eu te mandei: qual o contexto, qual o problema, qual a suposição (de utilidade ou de mundança de contexto), qual os quadro teóricos, qual o(s) artefato(s) proposto(s) e como serão avaliados, vai fechar muito bem.

[26]V:O Capítulo foi refatorado para ficar alinhado com a DSR.

[27]X:isso aqui é a validação, mas qual é o problema, qual a proposta, são essas informações que faltam para ficar bem organizado

[28]V:O Capítulo foi refatorado para ficar alinhado com a DSR.

the proposed technological solutions, referred to as artifacts. Finally, we will detail the evaluation methodology, including the process for data set creation, the metrics used for assessment, and a thorough analysis of the results.

## 3.1 Design Science Research Framework

To provide a clear and organized structure for this experiment, we adopt the DSR framework. The key components of this research cycle are outlined as follows:

**Context** The operational environment of the well construction department within a major oil company, where efficient access to technical knowledge is critical.

**Problem** The challenge faced by engineers and specialists in effectively querying and retrieving accurate information from vast, unstructured, and domain-specific knowledge bases (e.g., operational reports, lessons learned).

**Supposition** Our core supposition is that LLM-based agent systems can improve the efficiency and accuracy of information retrieval for specialized tasks, but that the choice of architecture (single-agent vs. multi-agent) will have a measurable impact on performance and cost.

**Theoretical Frameworks** This work is grounded in the theories of Intelligent Agents, Retrieval-Augmented Generation (RAG), and multi-agent systems, as detailed in the Literature Review.

**Proposed Artifacts** Two distinct LLM-based agent systems are proposed and built:

- A Single-Agent Architecture.
- A Multi-Agent Architecture.

**Evaluation** The artifacts are evaluated by a panel of domain experts who assess the quality of their responses to a curated set of real-world queries. The evaluation is based on predefined metrics for truthfulness, performance, and cost.

## 3.2 Context and Problem Statement

### 3.2.1 Context

As established in the Introduction, this research is situated within the oil and gas industry, a sector characterized by complex, expensive operations. This experiment was carried out specifically within the well construction department of a major

oil company. In this environment, engineers and technical staff frequently need to access specialized information from a variety of internal data sources, including operational reports, learned lessons, and safety alerts. The efficiency and accuracy of this information retrieval process directly impact operational decision-making, safety, and cost-effectiveness.

The set of queries used to test the systems provides a concrete exemplification of the problem space.

### 3.2.2 Problem

The central problem addressed in this experiment is the inefficiency of technical knowledge management and data analysis in the well construction domain. Specialists often struggle to find precise answers to their queries, which are typically buried in large volumes of unstructured or semi-structured documents. This leads to time-consuming manual searches and the risk of overlooking critical information.

This experiment investigates two primary task categories that exemplify this problem, as described in Section 2.2 and summarized here:

- **Q&A Tasks:** Require the system to answer complex technical questions by synthesizing information from documents. For example: "How does the presence of silica in the composition of cement paste affect its thermal stability at high temperatures?"

- **Text-to-SQL Tasks:** Require the system to query structured databases using natural language. For example: "What was the longest-lasting NPT on rig number 05?"

The set of queries used to test the systems, listed in Appendix A, provides a concrete exemplification of the problem space.

## 3.3 Proposed Artifacts

To address the problem, we designed, built, and tested two distinct artifacts: a single-agent solution and a multi-agent solution. Both are goal-based agents designed to accurately respond to user queries by leveraging a suite of tools.

### 3.3.1 Single-Agent Architecture.

In this work, a goal-based agent (RUSSELL, 2020) was implemented with the goal of accurately responding to various queries. The agent operates within an environment
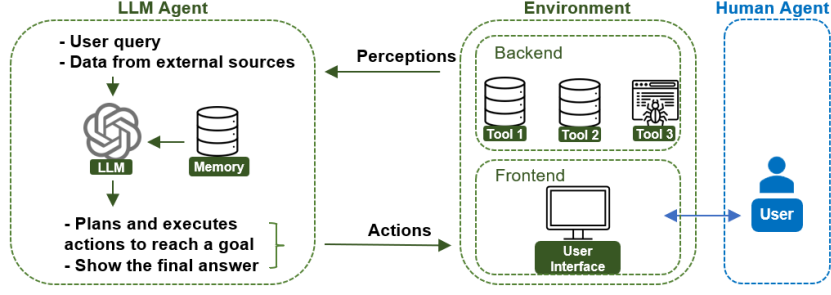
Figure 3.1: Schematic of the LLM-based agent interacting with an environment containing tools for task-specific operations, and the Human Agent interface for user interaction and feedback.

equipped with multiple tools for task-specific operations, as shown in Figure 3.1, and interfaces with users to receive queries.

Initially, a configuration of agents was implemented as described in Figure 3.2 using AutoGen Framework (WU *et al.*, 2023) with an architecture that allows information retrieval and user interaction. This system consists of two agentic setups:
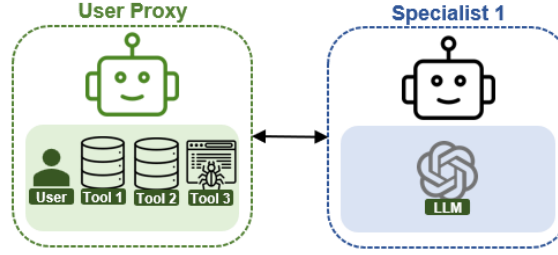


Figure 3.2: Chat setup with one User Proxy (WU *et al.*, 2023) and one Assistant.

- **User Proxy:** represents the interface with the user and with tools to access external databases. The modular nature of the tools allows the User Proxy to be customized and expanded based on the variety of data sources and the specific requirements of the application domain.

- **Agent:** powered by LLMs such as GPT-4 and GPT-3 (the specific model is configurable), is the analytical engine of the system. This agent interprets the queries received from the User Proxy and formulates responses.

For each question in the data set, the agent's decision-making process is executed as described in Figure 3.3, initially selecting the appropriate tool to respond to a query and, finally, compiling the retrieved information to provide a final answer.

### 3.3.2 Multi-Agent Architecture

The second artifact is a multi-agent system where responsibility is distributed among several specialized agents, coordinated by a Chat Manager, as shown in Figure 3.4.
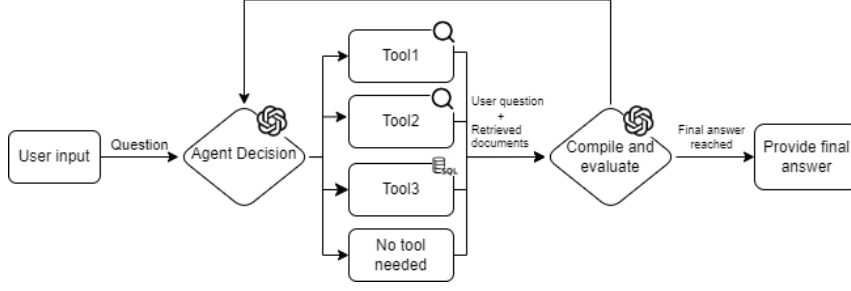
Figure 3.3: Decision process of the agent.

This architecture is designed to handle queries by routing them to the agent best equipped for the task. As depicted in the decision process in Figure 3.5, a "speaker selection" step determines the most suitable agent to act at each turn, promoting a more focused and contextualized approach to problem-solving.
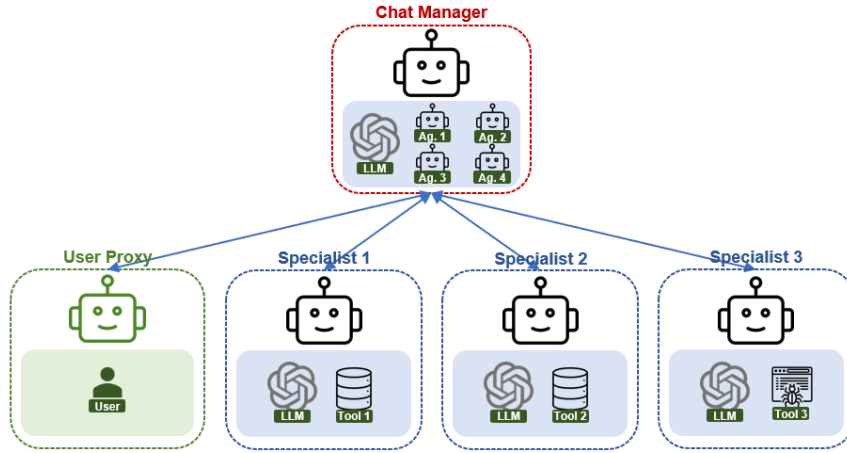


Figure 3.4: Chat setup with one Chat Manager and a group of LLM agents.
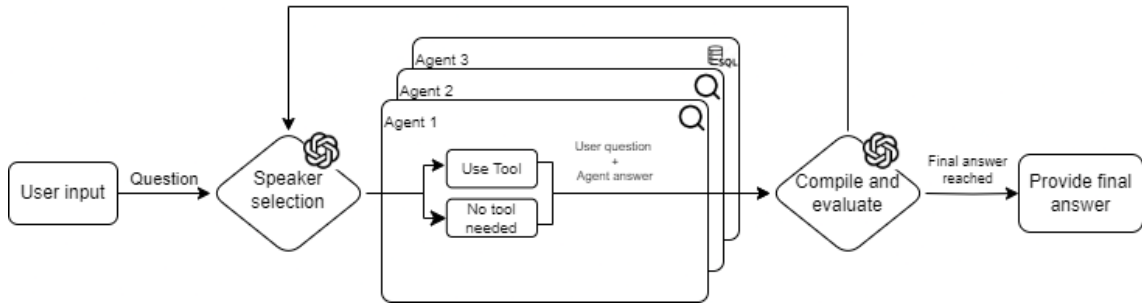


Figure 3.5: Multi-agent decision process.

### 3.3.3 Agent's Tools

In this experiment, three tools were considered in the decision-making process:

- **Tool 1 - Learned Lessons Search:** a tool to search for learned lessons that may be relevant to the query.

- **Tool 2 - Employee Search:** functionality that allows the search for information related to collaborators of an organization.

- **Tool 3 - NPT SQL Query:** Interface for executing SQL queries on a database of operational NPTs.

There is also a pathway that allows the agent to provide a direct response, without the need to resort to other tools, presumably used when the LLM already possesses the necessary information.

## 3.4 Evaluation

The evaluation phase was designed to assess and compare the performance of the two proposed artifacts. This section details the methodology, the data set creation process, the metrics used, and the final results.

### 3.4.1 Evaluation Methodology

The evaluation was conducted by presenting a standardized set of questions to both the single-agent and multi-agent systems, using both GPT-3.5-turbo and GPT-4 models. The responses generated by each configuration were then collected and anonymized.

A panel of three specialist engineers from the well construction department was tasked with analyzing the generated answers. Each specialist independently scored the responses based on the metrics described in Section 3.4.3. The final score for each response was calculated by averaging the scores from the three experts, ensuring a robust and comprehensive assessment.

[29] [30]

X:29
V:30

To provide a clear visual representation of the experimental workflow, a Business Process Model and Notation (BPMN) diagram is presented in Figure 3.6. This diagram illustrates the step-by-step process, from query submission to expert evaluation.

---

[29]X:Aqui seria bom fazer um BPMN do passo a passo do seu experimento, veja a figura 4.1 de`https://www.cos.ufrj.br/uploadfile/publicacao/3172.pdf`
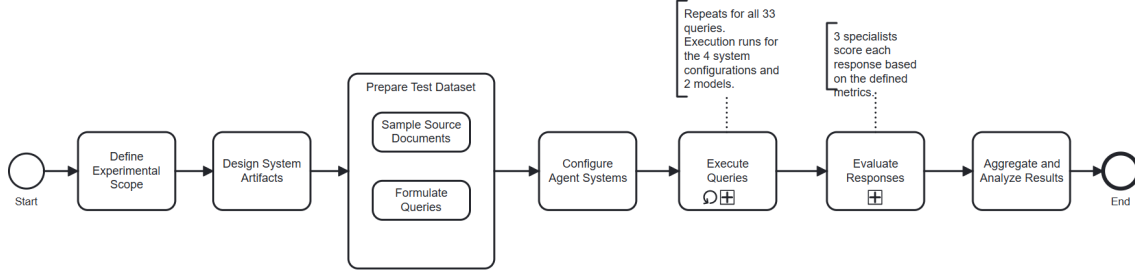
[30]V:Feito

Figure 3.6: Experimental workflow.

## 3.4.2 Data Set Creation

A critical component of this evaluation is the test dataset. The dataset was meticulously created to reflect authentic information needs within the well construction domain. The process was as follows:

**Source Selection** We identified three primary internal data sources: a database of Operational Learned Lessons, a structured database of Non-Productive Time (NPT) incidents, and a Collaborator Finder tool, as described in Section 1.1.1.

**Document Sampling** A random sample of documents and records was selected from each data source to ensure broad coverage of topics and scenarios.

**Query Formulation** This process was performed by the author, leveraging domain expertise and collaboration with colleagues to ensure the questions were realistic, relevant, and challenging.

**Dataset Composition** In total, a dataset of 33 unique queries was created.

This approach to dataset creation, grounded in author experience and real-world documents, provides a valid basis for evaluating the artifacts. Table 3.1 presents a sample of the queries formulated for the experiment.

⚡[31] ⚡[32]

X:31
V:32

## 3.4.3 Evaluation Metrics

To ensure a comprehensive assessment, the expert panel evaluated the artifacts' responses using the following metrics, which are based on the definitions presented in Section 2.5:

- **Truthfulness**: A 1-5 Likert scale score measuring the factual accuracy of the response and the extent of any divergence from the ground truth. A higher score indicates a more factually correct answer with no hallucinations.

---

[31]X:Coloca todas na tabela! E faz uma seção de criação de perguntas, ou subseção
[32]V:Feito

| Task category | Question |
|---|---|
| Q&A | How does the presence of silica in the composition of cement paste affect its thermal stability at high temperatures? |
| | What are the main challenges and risks associated with through tubing plug and abandonment in highly deviated wells? |
| | Give me all the information about employee BFD1. |
| | Who are the employees of the POCOS/EP/SASD team? |
| | How many advisors do we have in the POCOS/SPO department? |
| | Who are the advisors in the departments belonging to the POCOS/EP department? |
| | What data sources do you have? |
| | What functions do you have? |
| | How does well inclination affect the effectiveness of cementing during through-tubing plugging? |
| | What can cause difficulty in locking the handling cap of the coiled tubing BOP? |
| | What can cause anomalous behavior of the AutoTrak with GunDrill during drilling? |
| | What can be done to optimize the assembly of COP/COI for parallel movement of the JRC/THRT? |
| | What strategies can be adopted to improve the quality of cementing in highly inclined wells during through-tubing plugging? |
| | What are the alternatives to accelerate the curing time of cement slurry without compromising its integrity in high-temperature conditions? |
| | What are the risks associated with the improper substitution of cement with silica for pure cement in surface casing cementations in high-temperature wells? |
| | What was the strategy adopted to allow the passage of eccentric and/or large-diameter elements through the BOP quickly and without wedging the string with these elements inside the BOP? |
| Text-to-SQL | What was the longest-lasting NPT on rig number 05? |
| | How many NPTs occurred on rig number 06 during August 2023? |
| | What were the 5 most common abnormalities across all rigs? |
| | What were the abnormalities that occurred on all rigs during the week of September 14th to 20th, 2023? |
| | Which rigs had the most lost time in 2023? Give me a table with the rigs and the sum of hours. |
| | Which rigs had the most lost time in the first half of 2023? |
| | What were the latest abnormalities that occurred on the SS-70 rig? |
| | What was the peak of abnormality occurrences on the NS-52 rig? |
| | What was the total lost time in hours for abnormalities whose description mentions the term "Coiled Tubing"? |
| | What was the total lost time in hours on the NS-38 rig in 2023? |
| | What was the total time lost due to equipment failure on the NS-38 rig in 2023? |
| | How many abnormalities occurred on the NS-31 rig during July 2023? |
| | How many hours of lost time were caused by human error on the NS-47 rig in 2023? |
| | How many hours of lost time occurred on the MS-20 rig during June 2024? |
| | How many hours of lost time occurred on the NS-35 rig in 2024? |

Table 3.1: Queries used in first cycle.

- **Performance**: A 1-5 Likert scale score assessing the overall quality of the response, including its linguistic coherence, logical structure, relevance, and conciseness.

- **LLM Cost**: A quantitative metric representing the financial cost in US dollars (USD) to generate a response for a given query using the OpenAI API. This reflects the computational expense and efficiency of each configuration. While other costs exist (development, infrastructure, maintenance), the API cost is a primary operational expenditure that scales directly with usage and is therefore a key metric for evaluating the economic viability of the artifacts, as established in our DSR framework.

To illustrate the application of the first two metrics, an example of an expert evaluation is provided in Table 3.2. The table shows the responses of different models to the query: "How does the presence of silica in the composition of cement paste affect its thermal stability at high temperatures?". It details the scores for

Truthfulness and Performance, along with the specialists' comments that justify the ratings.

[33] [34] ~~OLHA UMA PALAVRA INÚTIL TÍPICA DE LLM~~[35] ~~Outra~~[36] [37]

X:33
V:34
X:35
X:36
V:37

---

[33]X:Por que essa pergunta? Isso é um exemplo? Como o multi agente entrou aqui se ainda não falou de multi agente? Melhor deixar para mais tarde ou mostrar só para single-agent?

[34]V:Refatorado pra deixar claro q esse eh um exemplo e inserido após a explicação das arquiteturas.

[35]X:comprehensive

[36]X:in-depth

[37]V:Essa parte com palavras típicas de llm foi escrita no artigo, início do ano passado, quando essas expressões ainda não eram indesejáveis kkkk estou retirando tudo. Se ainda tiver passado algo, me avise pfv.

Table 3.2: Query example with inputs, outputs, and human expert evaluations.

| Input query | Model | Setup | Final output | Truth.fulness | Performance | Specialist Comments |
|---|---|---|---|---|---|---|
| How does the presence of silica in the composition of cement paste affect its thermal stability at high temperatures? | GPT-3.5-turbo | Single-Agent | The presence of silica in the composition of cement paste can affect its thermal stability at high temperatures due to some characteristics of silica. Silica is an inorganic material that has a stable crystalline structure at high temperatures. When exposed to high temperatures, the silica present in the cement paste can undergo structural transformations, resulting in changes in the properties of the paste. One of the main changes that can occur is the thermal expansion of the cement paste due to the formation of unstable silica compounds. This thermal expansion can lead to cracks and fissures in the structure, compromising the material's integrity and stability. Moreover, the presence of silica in the composition of cement paste can affect the material's thermal resistance. Silica has low thermal conductivity, which can lead to a reduced ability of the cement to withstand high temperatures. This can result in a decrease in the material's strength and, consequently, a reduction in its thermal stability. In summary, the presence of silica (...) | 4 | 5 | Created a section that is irrelevant to the question: "Moreover, the presence of silica in the composition of cement paste can affect the material's thermal resistance. (...)" |
| | | Multi-Agent | The presence of silica in the composition of cement paste has a positive effect on thermal stability at high temperatures. Silica is added to cement to mitigate the retrogression of cement strength when exposed to temperatures above 230°F (110°C). Deep wells or those with a high geothermal gradient may require the use of silica cement to ensure the thermal stability of the paste. The use of silica cement is particularly important in surface casing cementing, where the temperature can exceed the critical limit. (...) | 2 | 2 | Retrieved an unrelated Lesson Learned on the subject (about plugs). |
| | GPT-4 | Single-Agent | The presence of silica in the composition of cement paste increases thermal stability, especially at high temperatures. Silica is used to transform the cement into a refractory material, capable of withstanding extreme conditions. The cement (...). However, (...). | 3 | 4 | Although it is true, the item on the side is not relevant to the question: "However, it is important to note that (...)." |
| | | Multi-Agent | Silica is added to cement at a ratio of 35% by weight of cement to mitigate the retrogression (loss of compressive strength) of the cement in the well when subjected to temperatures above 230ºF (110ºC). (...) | 5 | 4 | Cited an unrelated (but true) Lesson Learned about plugs, Lesson Learned LA7005. |

### 3.4.4 Results

This section provides an analysis of the data collected during the first experimental cycle. The aggregated results are presented in Table 3.3, followed by a discussion of each evaluation metric established in our DSR framework: Truthfulness, Performance, and LLM Cost.

Table 3.3: Results on Q&A and Text-to-SQL tasks, including standard deviation (Std). The best metrics are highlighted with **bold and underline**. The second best are highlighted with **bold**.

| Task | Single-Agent | | | | | Multi-Agent | | | | |
|------|------|--------|------|------|------|------|--------|------|------|------|
| Model | LLM Cost | Truth. | Std | Perf. | Std | LLM Cost | Truth. | Std | Perf. | Std |
| Q&A | | | | | | | | | | |
| GPT-3.5-turbo | 0.005 | 2.94 | 1.48 | 3.94 | 1.09 | 0.02 | 4.09 | 1.22 | 3.82 | 0.98 |
| GPT-4 | 0.12 | **3.88** | 1.41 | **4.06** | 1.30 | 0.45 | **<u>4.57</u>** | 0.79 | **<u>4.43</u>** | 0.79 |
| Text-to-SQL | | | | | | | | | | |
| GPT-3.5-turbo | 0.009 | 4.13 | 1.41 | 4.44 | 1.03 | 0.02 | **4.29** | 1.20 | **4.29** | 1.33 |
| GPT-4 | 0.10 | **<u>4.56</u>** | 0.96 | **<u>4.63</u>** | 0.81 | 0.51 | 3.20 | 1.99 | 3.70 | 1.89 |

The comparative analysis between single and multi-agent setups for RAG, using GPT-3.5-turbo and GPT-4 models, revealed insights regarding the metrics of truthfulness, performance, and costs of the language model.

**Truthfulness**

In assessing the truthfulness metric, significant differences are noted between the single and multi-agent settings in both Q&A and Text-to-SQL tasks. The results are illustrated in Figures 3.7 and 3.8. For Q&A tasks, GPT-4 in a multi-agent configuration significantly exceeded the performance of the single-agent with a truthfulness score of 4.57 compared to 3.88. The GPT-3.5-turbo model showed distinct results between the two configurations, with the multi-agent surpassing the single-agent with scores of 4.09 and 2.94, respectively. In terms of Text-to-SQL queries, a different outcome was observed. GPT-4 single-agent achieved a score of 4.56, while the

---

[38]X:Isso aqui é uma pergunta de pesquisa tem que entrar de alguma maneira na definição do DSR, lembrando que as avaliações do DSR podem ser mais de uma

[39]V:Feito. Movido p/ definição do DSR

[40]X:Não é represents, já que é o custo mesmo, acho que corresponds to, ou mesmo só is

[41]V:Feito

[42]X:Tem que falar alguma coisa que não é o único custo, e quais são os outros e porque esse é importante, isso pode estar descrito no modelo DSR, antes

[43]V:Feito

[44]X:Esse parágrafo tipicamente aparece na revisão

same model in the multi-agent configuration obtained 3.20, highlighting a limitation for the multi-agent in this task. Conversely, the GPT-3.5-turbo maintained a more balanced performance between configurations, scoring 4.29 for multi-agent and 4.13 for single-agent.
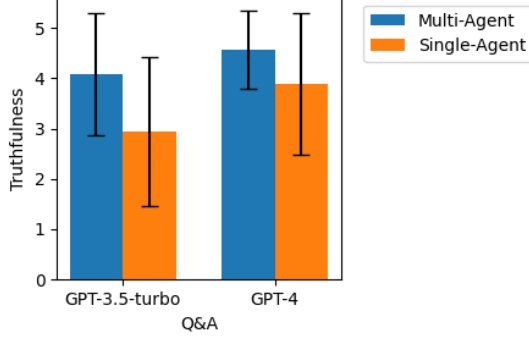


Figure 3.7: Truthfulness and standard deviation in Q&A tasks by LLM model and agent configuration.
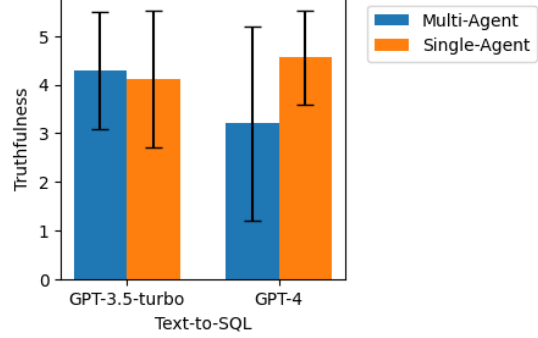
Figure 3.8: Truthfulness and standard deviation in Text-to-SQL tasks by LLM model and agent configuration.

**Performance**

The evaluation of LLM performance (LI *et al.*, 2023) in the tasks of Q&A and Text-to-SQL reveals trends which are similar to the truthfulness results. For Q&A tasks, the multi-agent setup shows a performance boost compared to the single-agent setup. In particular, the multi-agent GPT-4 achieves a performance score of 4.43, which is higher than the single-agent GPT-4 score of 4.06. This pattern is consistent with the GPT-3.5-turbo, where the multi-agent system also surpasses the single-agent system, scoring 3.82 and 3.94, respectively. These findings emphasize the effectiveness of the multi-agent approach in handling technical user queries.

**LLM Cost**

Language model services are typically composed by a values per token. For instance, GPT-4 model costs US$30.00 (input) and US$60.00 (output) per 1 million tokens received and sent, respectively. The single-agent architecture demonstrated substantially lower costs for both Q&A and Text-to-SQL tasks compared to the multi-agent setup as shown in Figure 3.11. For instance, the average cost of the GPT-4 model (OPENAI *et al.*, 2023) for a Q&A task was $0.12 per processed question for the single-agent, while the multi-agent recorded an average cost of $0.45. This trend of higher costs for the multi-agent architecture was also maintained for Text-to-SQL tasks, with an average cost of $0.51 for the multi-agent architecture in contrast to $0.10 for the single agent. The higher token count and cost for multi-agent setting
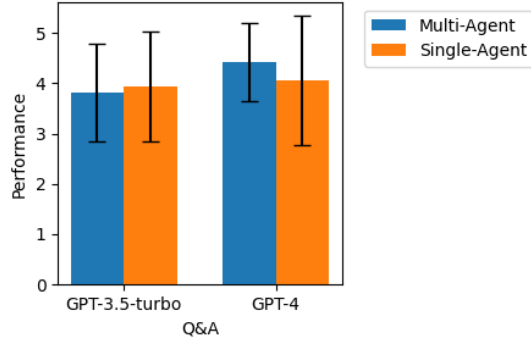
Figure 3.10: Performance and standard deviation in Text-to-SQL tasks by LLM model and agent configuration.

Figure 3.9: Performance and standard deviation in Q&A tasks by LLM model and agent configuration.

is due to the inclusion of intermediate calls, for example, when the "Agent Selector" needs to decide which agent to pass the turn to. All the message history is passed to the LLM at this stage, substantially increasing the number of tokens submitted and response time.
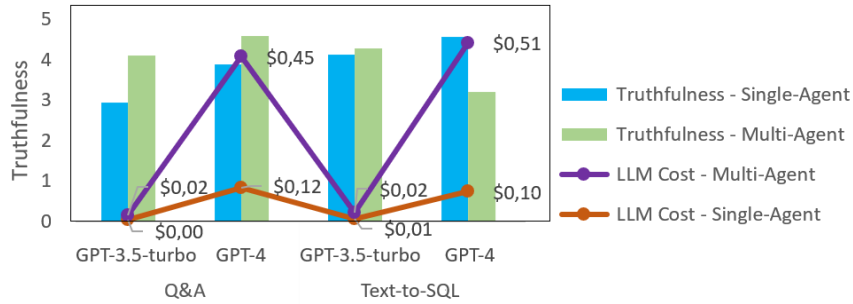


Figure 3.11: Average LLM costs and Truthfulness per completed task according to setup and model.

### 3.4.5 Discussion

The comparison between single and multi-agent systems revealed significant differences in terms of performance and cost:

**General Performance.**

The results indicate that for Q&A tasks in the context of O&G, truthfulness measure was 28% higher with the multi-agent architecture compared to single. However, for Text-to-SQL tasks, this trend was inverted, where the single-agent scored 15% higher.

These findings suggest that for Q&A tasks, the multi-agent setup may be more advantageous in terms of providing truthful information, particularly when utilizing

the more advanced GPT-4 model. Conversely, in Text-to-SQL tasks, the GPT-4 model in a single-agent configuration proved more effective. This might imply that the added complexity of managing multiple agents in some tasks does not necessarily lead to improved performance in responses, underscoring the importance of carefully selecting the agent configuration based on the task type and specific features of the language model used.

**Cost-Performance Analysis.**

While the multi-agent system shows higher truthfulness in Q&A tasks, it is crucial to consider the associated costs. To provide a clearer comparison, let us consider the score/cost ratios. For Q&A tasks using GPT-4, the single-agent configuration yields a ratio of 32.33 truthfulness points per dollar, compared to 10.16 for the multi-agent setup. This indicates that while the multi-agent system shows a 17.8% improvement in truthfulness, it comes at a 275% increase in cost.

⚡[45] ⚡[46]                                                                        X:45

This trade-off highlights an important implication for any organization consid-   V:46
ering the adoption of these technologies. The optimal architecture is not universal; it is highly dependent on specific task requirements and budget constraints. This reality underscores the necessity of conducting a preliminary, cost-performance evaluation. Rather than simply selecting a model, decision-makers must first perform a targeted analysis to establish a cost-benefit threshold. Our work not only provides initial data for the O&G domain but also demonstrates a foundational methodology for this evaluation process, which ultimately motivated the more rigorous and quantitative approach of our second experimental cycle.

**Model Performance Variations.**

Interestingly, our results show that GPT-3.5-turbo outperforms GPT-4 in certain tasks, particularly in the Text-to-SQL multi-agent configuration, despite GPT-4's larger size and more extensive training. This unexpected performance could be attributed to several factors. First, GPT-3.5-turbo may have undergone more specific fine-tuning for structured query tasks, allowing it to excel in Text-to-SQL scenarios. Additionally, GPT-3.5-turbo's training data might be more recent or more relevant to the specific domain of our study. Another possibility is that the smaller model size of GPT-3.5-turbo allows for faster processing and more efficient handling of the multi-agent setup, resulting in better performance in some contexts.

---

[45]X:TEm que deduzir a necessidade de fazer um experimento antes levando essas coisas em consideração

[46]V:Feito abaixo.

However, it is important to note that GPT-4, when used in a multi-agent setup, demonstrated more consistent truthfulness and performance, as evidenced by its reduced standard deviation in results. This consistency can be particularly advantageous in applications where reliability and accuracy are critical. Multi-agent systems have the advantage of maintaining separate contexts for different aspects of a task (LANGCHAIN, 2025a). [47] [48] This compartmentalization can lead to better X:47 handling of complex, multi-faceted queries, as each agent can focus on its specific V:48 context without being overwhelmed by irrelevant information. However, this advantage may be offset in tasks like Text-to-SQL, where maintaining a unified context of the database schema and query structure is crucial, possibly explaining the better performance of single-agent setups in this task. Furthermore, the multi-agent architecture inherently involves multiple stages of information processing, which can serve as natural filtering mechanisms. As information passes from one agent to another, irrelevant or low-quality data may be naturally filtered out, leading to more refined and accurate final outputs. This could explain the superior performance in filtering irrelevant information observed in multi-agent setups.

**Economic Efficiency.**

The multi-agent architecture incurs significantly higher costs compared to the single-agent system, primarily due to additional intermediate calls to the language model and multiple iterations between agents for action planning. Also, the cost differences between using GPT-4 and GPT-3.5-turbo are substantial, with GPT-4 being 20 times more expensive (in early 2024). [49]. [50]                 X:49

The average cost per query for each configuration is presented in Table 3.4. These V:50 figures highlight the direct cost implications of the chosen architecture and model.

Table 3.4: Average LLM Cost Per Query (USD). Values from early 2024.

| Configuration | Cost per Query |
|---|---|
| Single-Agent (GPT-3.5-Turbo) | $0.0068 |
| Single-Agent (GPT-4) | $0.1095 |
| Multi-Agent (GPT-3.5-Turbo) | $0.0197 |
| Multi-Agent (GPT-4) | $0.4896 |

To illustrate the financial implications of adopting different models and architectures, we estimate the annual costs for a large company with 40,000 knowledge

---

[47]X:Você pode suportar essa afirmação com uma citação?
[48]V:Feito.
[49]X:Dizer x vezes mais caro em julho de 2025
[50]V:Feito

workers. Our calculations are based on an average of 5 queries per worker per day, over 250 working days per year.

Under these assumptions, the total annual query volume is 50 million (40,000 workers × 5 queries/day × 250 days). For a single-agent configuration, this results in an annual cost of approximately \$337,843 for GPT-3.5 and \$5.47 million for GPT-4.

In a multi-agent architecture, the costs increase substantially, escalating to approximately \$986,631 for GPT-3.5 and \$24.48 million for GPT-4. These estimates underscore the significant financial trade-offs when adopting a multi-agent system, which, while potentially offering performance benefits, comes with a considerable increase in LLM operational costs.

While multi-agent systems and more advanced models like GPT-4 offer improvements in performance, the economic efficiency, as measured by truthfulness per dollar, may favor single-agent systems and less costly models like GPT-3.5-turbo, depending on the specific application and budget constraints.

It is important to note that, as of July 2025, the landscape of LLMs has evolved substantially. The emergence of more efficient models, has led to a significant decrease in API's costs. This suggests that the financial trade-offs discussed previously may no longer be as pronounced, and that high-performance multi-agent systems could become economically viable much sooner than anticipated.

⚡[51] ⚡[52]

### Challenges and Limitations

During the evaluation of the agents, several challenges and limitations were identified.

***Contextualization and Interpretation.*** In many cases, the single-agent solution had difficulty understanding the context of the question. For example, a question about cementing was interpreted in the context of the construction industry, a theme to which the language models were more exposed during the training phase. However, the multi-agent structure, with its well-defined roles, better understood the questions and showed superior performance in Q&A tasks, corroborating the findings of (LI *et al.*, 2024a).

***Filtering Irrelevant Information.*** The agent often receives irrelevant documents along with important ones in the prompt context, and it is up to the LLM to ignore these. For example, when asked about alternatives to accelerate the curing

---

[51]X:In summary é o parágrafo típico das LLMs... Mas é isso mesmo. Porém tem que colocar um ponto: o custo dos modelos está caindo barbaramente com o aparecimento de novos modelos no topo de desempenho e novas tecnologias tem permitido alcançar resultados de ótima qualidade com máquinas muito menores, o que também derruba o custo. Pode até citar o exemplo do DeepSeek (buscando na literatura o desempenho x custo dele)

[52]V:Feito

time of cement paste without compromising its integrity at high temperatures, the RAG system retrieved a document that included information about batch cementing to ensure homogeneity during manufacturing and pumping. While this information is true, it was not relevant to the specific question asked. In this aspect, the multi-agent solution performed better at discarding such irrelevant information, focusing more accurately on the task at hand. Other possible solutions include improving the accuracy of semantic search by adjusting a minimum threshold for similarity measures or through re-ranking techniques such as those proposed by (CARRARO, 2024) and (SUN *et al.*, 2023).

***Hallucination.*** During the evaluation of our system, we encountered instances where the agent produced hallucinated information instead of utilizing the appropriate tool to retrieve accurate data, as in (BILBAO *et al.*, 2023). For example, when asked, "How many anomalies occurred on rig number 05 during August 2023?" the agent was expected to use the Text-to-SQL tool to query the database. However, it bypassed this tool and generated a fabricated response, stating that 5 anomalies occurred, along with detailed descriptions of fictional events. The correct answer, as retrieved from the database, was that 7 anomalies occurred. This hallucination likely resulted from the agent's reliance on its internal knowledge rather than external data retrieval.

In terms of hallucination statistics, our analysis revealed that for Q&A tasks, hallucinations occurred in 9.6% of cases and 3.8% for partially hallucinated. In contrast, Text-to-SQL tasks exhibited a lower hallucination rate, with only 3.6% of responses containing hallucinated information and 96.4% being accurate. These findings highlight the variation of susceptibility to hallucination in different types of tasks, highlighting the need for targeted strategies to mitigate this problem.

***Industry Jargon:*** Specifically analyzing the activity of drilling and completion of offshore wells, the main challenge is the inherently complex and technical nature of the data involved. There were instances of incorrect interpretation of information, likely due to the use of terms, expressions, and themes specific to well construction, to which the language model had little or no exposure during training phase. A possible solution is the implementation of specialized models, which has been pointed out in gray literature as a trend for the coming years (SHAH, 2024; MEENA, 2023; GHOSH, 2023).

***Tools vs. Performance:*** It was identified during the experiments that agents with a high amount of tools showed a decline in overall performance. This can be attributed to the added context to the prompts. As the context length increases, the model's ability to accurately interpret and respond diminishes. This is a limitation of current language models, where longer contexts can lead to a dilution of relevant information and increased difficulty in maintaining coherence and accuracy.

This conclusion is currently qualitative, as these metrics were not addressed in this experiment.

**Queries Involving Proper Names:** In queries involving people's names, it was not possible to retrieve relevant documents using semantic search. For example, when asked to identify the employee associated with a specific key and list knowledge items they registered in the system, the LLM incorrectly attributed such documents to the wrong author [53] [54] . This highlights the difficulty in accurately <span style="color:red">X:53</span> retrieving information based on proper names, which can be complicated by vari- <span style="color:blue">V:54</span> ations in accentuation, abbreviation, and formatting. [55] [56] A potential solution <span style="color:red">X:55</span> to be explored is the use of Self-Query Retriever (LANGCHAIN, 2023), implement- <span style="color:blue">V:56</span> ing a hybrid search with metadata filters (including proper names) and semantic retrieval of the rest of the query. It is also suggested, in these cases, to use the (LEVENSHTEIN, 1966) distance to handle possible variations in the spelling of names. This approach could improve the accuracy of retrieving documents related to specific individuals, ensuring that the correct information is associated with the right person.

**Practical Implications.**

The findings from our study have significant practical implications for the O&G sector, and potentially for other industries characterized by complex and technical data environments:

- **Enhanced Decision-Making Support:** Our results indicate that multi-agent systems provide a 28% higher truthfulness measure in Q&A tasks. This can be particularly beneficial for decision-making in well engineering, where accurate and truthful information is critical. Implementing multi-agent systems in decision-making processes can lead to more reliable and informed decisions, thereby reducing the risk of errors and enhancing operational safety and efficiency.

- **Balancing Performance and Economic Efficiency:** While multi-agent systems offer superior performance in terms of truthfulness, they come with a cost that is 3.7 times higher on average compared to single-agent systems. This highlights the importance of a strategic approach in selecting agent configurations based on specific tasks and budget constraints. A detailed cost-benefit analysis reveals that for Q&A tasks using GPT-4, the single-agent configuration yields a ratio of 32.33 truthfulness points per dollar, compared to 10.16

---

[53]X:O RAG ou a LLM usando o RAG, não ficou claro
[54]V:Corrigido
[55]X:tem evidências disso em outros artigos?
[56]V:não encontrei

for the multi-agent setup. While the multi-agent system shows a 17.8% improvement in truthfulness, this comes at a 275% increase in cost. The efficiency varies significantly by task type; in Text-to-SQL tasks, the GPT-4 single-agent outperforms the multi-agent by 42.5% in truthfulness while costing 80.4% less.

- **Reflection and Critic Agents:** A promising approach to enhance the performance of these agents is the use of reflection (SHINN *et al.*, 2023), a method where agents verbally reflect on task feedback signals and maintain this reflective text in an episodic memory buffer to improve decision-making in subsequent trials. Critic agents are a way to implement reflection in a multi-agent setup. This type of agent is challenging to apply in Q&A tasks over private technical data, as commercial LLMs (OpenAI, Google Bard, and others) have not been deeply trained in the domain and struggle to provide relevant and precise critiques, reinforcing the trend toward increased use of domain-specific models (SHAH, 2024; MEENA, 2023; GHOSH, 2023).

- **Task-Specific Agent Configuration:** The study highlights that the complexity of managing multiple agents does not always lead to better performance. In some cases, a single-agent setup might be more effective. This insight can guide the development and deployment of AI systems, ensuring that the configuration of agents is tailored to the specific requirements of the task, thereby optimizing both performance and cost.

- **Potential for Broader Application:** The insights gained from this study are not limited to the O&G sector but can be applied to other industries with similar technical complexities, such as aerospace, pharmaceuticals, and renewable energy. By adopting multi-agent systems in these industries, organizations can improve decision-making, knowledge management, and operational efficiency, driving innovation and competitiveness.

**Future Directions.**

This work indicates possible pathways for enhancing RAG architectures in O&G sector.

- **Enhancement of IR Semantic Techniques:** There is a critical need to develop more sophisticated semantic search technologies. Future efforts should focus on enhancing the precision of information retrieval by filtering out irrelevant content more effectively. This will ensure that agents can provide more accurate and contextually appropriate responses, crucial for technical domains such as O&G.

- **Development of Domain-Specific Models:** Specialized models tailored specifically to the O&G and other domains, such as biomedical engineering (PAL *et al.*, 2024), could significantly improve the handling of specific jargon and complex technical data, while reducing LLM costs (AREFEEN *et al.*, 2024). Future research should aim to develop and train these models to better understand and interpret the unique language and data types found in O&G, enhancing the overall accuracy of agent responses.

- **Optimization of Tool Use in Agent Performance:** The relationship between the quantity of tools available to an agent and its performance needs further exploration. Future studies should quantify the impact of tool availability on agent efficacy and efficiency, aiming to optimize tool use without overwhelming the agent or diluting performance quality.

- **Integration of Advanced Name Recognition Techniques:** Queries involving proper names pose a significant challenge in semantic search. Integrating advanced retrieval techniques, such as Self-Query Retrievers (LANGCHAIN, 2023) and (LEVENSHTEIN, 1966) distance algorithms, could improve the handling of these queries. Future research should focus on enhancing name recognition capabilities to ensure that agents can accurately retrieve and utilize correct information, especially in scenarios where precision is paramount.

- **Extension to Other Complex Domains:** The potential applications of multi-agent systems are not limited to the O&G sector. Future research should explore the adaptation and implementation of these systems in other complex and technical domains, such as aerospace, pharmaceuticals, and renewable energy. Investigating how these systems can support decision-making in these areas will provide valuable insights into their versatility and adaptability.

- **Hybrid Model Experimentation:** Combining the strengths of single and multi-agent systems could yield significant benefits. Future directions should include experimenting with hybrid models that integrate the robustness and depth of multi-agent interactions with the simplicity and efficiency of single-agent systems. This hybrid approach could potentially offer a balanced solution, maximizing performance while managing costs and complexity.

By pursuing these directions, future research can significantly advance the development of multi-agent systems, not only enhancing their application in the O&G sector but also expanding their utility across various technologically intensive activities.

# Chapter 4

# Second Experimental Evaluation Cycle

This chapter describes the second experimental cycle of this research, building upon the findings of the first cycle detailed in Chapter 3. The rapid evolution of generative AI frameworks and models, along with the insights gained previously, prompted a more advanced and rigorous evaluation. This second phase employs non-agentic workflows as a baseline, introduces a more quantitative evaluation methodology, and leverages an automated assessment process based on the "LLM-as-judge" concept (GU *et al.*, 2025).

The use of "LLM-as-a-judge" was driven by the sheer volume of responses requiring evaluation. With four configurations, two models, and three executions for each of the 33 [AJUSTAR] questions, a total of 792 [AJUSTAR] responses were generated. Manually assessing this volume of data would have been impractical. Furthermore, previously used metrics like 'truthfulness' had become less critical. This metric was highly relevant when models frequently hallucinated, a problem that is far less prevalent in the current generation of LLMs, shifting the focus to precision and recall of factual information.

[TODO: GERAR BPMN!!]

## 4.1  Design Science Research Framework

This second experimental cycle adheres to the Design Science Research (DSR) methodology, focusing on refining the artifacts and evaluation based on the outcomes of the first cycle.

**Context** The operational environment of well construction engineering, where practitioners require efficient and reliable access to vast amounts of technical and ESG-related information.

**Problem** The first experimental cycle revealed several limitations, including the subjective nature and scalability issues of expert-based evaluation, the need to compare agentic systems against simpler non-agentic baselines, and the challenge of ensuring consistent performance. This second cycle addresses the problem of developing a more robust, scalable, and objective method for evaluating and comparing different LLM-based architectures for domain-specific Q&A.

**Proposed Artifacts** Four distinct architectures were designed and implemented to compare different strategies for information retrieval and reasoning:

- A non-agentic **Linear-Flow** RAG pipeline.

- A non-agentic **Linear-Flow with a Router** to direct queries.

- A **Single-Agent** architecture, refined from the first experiment.

- A **Multi-Agent Supervisor** architecture for distributed reasoning.

**Evaluation** The artifacts are evaluated using an automated pipeline. An "LLM-as-a-judge" assesses the generated answers against a ground-truth dataset. The evaluation is based on quantitative information retrieval metrics: **Precision**, **Recall**, and **F1-Score**.

## 4.2 Context and Problem Statement

### 4.2.1 Context

As established in the previous chapters, this research is situated within the oil and gas industry, specifically in the domain of well construction and maintenance. Engineers and specialists in this field must navigate a complex information landscape, drawing from operational reports, ESG alerts, and documented best practices (Learned Lessons) to make critical decisions. The effectiveness of these decisions hinges on the speed and accuracy with which relevant information can be retrieved and synthesized.

### 4.2.2 Problem

The first experimental cycle confirmed the potential of LLM-based agents but also highlighted key challenges. The manual, expert-led evaluation process was time-consuming and difficult to scale. Furthermore, the performance differences between single and multi-agent systems suggested that a more granular analysis was needed, including a comparison with non-agentic RAG workflows to establish a performance

baseline. Therefore, the central problem for this second cycle is to design and execute a more rigorous, automated, and scalable evaluation to definitively compare the efficacy of various agentic and non-agentic architectures in this specialized domain.

## 4.3   Proposed Artifacts

To address the research problem, four distinct artifacts were developed, representing a spectrum of complexity from simple sequential pipelines to collaborative multi-agent systems.

### 4.3.1   System Architecture Overview

The experimental system was implemented using LANGCHAIN (2025b) and LANG-GRAPH (2025) frameworks specialized in language model orchestration. This modular design allows for the systematic and reproducible evaluation of different components and workflows. Key layers of the architecture include:

- **Experiment Orchestration:** Manages the execution loop, iterating through all combinations of questions, models, and setups.

- **Agent Workflow Frameworks:** Defines the logic for each of the four proposed artifacts using LangGraph to create cyclical graphs for agentic behavior.

- **Tool Integration:** A standardized interface providing agents with access to external knowledge sources. This layer enables consistent semantic search over domain-specific vector stores, ensuring that performance differences are attributable to architectural choices rather than variations in data access.

- **Prompt Engineering:** A library of system messages and prompt templates designed to guide the LLM's reasoning process for each specific task within the workflows.

- **State Management and Logging:** Captures the complete execution trace of each run, including intermediate steps, tool calls, and final outputs. This observability is essential for understanding not just the final output, but the process by which each architecture arrived at its answer.

### 4.3.2   Artifact 1: Linear-Flow

The **Linear-Flow** architecture represents the simplest non-agentic RAG design, serving as a performance baseline. As shown in Figure 4.1, user input is processed in a strictly sequential manner. The user's query is handled by a single LLM step, which

contains all the instructions needed to generate search queries for every available tool.
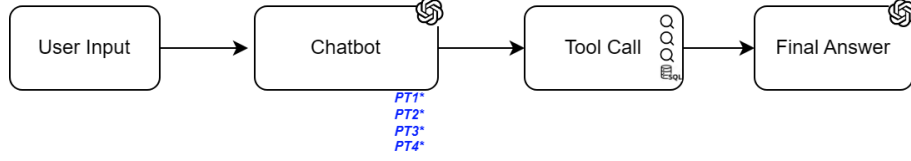


Figure 4.1: Linear-Flow architecture. PTn indicates the prompt for Tool n.

Because the instruction prompts for all tools are aggregated into a single call, the resulting context for the LLM becomes notably extensive and complex. While this approach is straightforward to implement, its primary drawback is the potential for performance degradation as the context length increases, which can dilute the model's focus and lead to less precise retrieval queries.

### 4.3.3 Artifact 2: Linear-Flow with Router

The **Linear-Flow with Router** paradigm (Figure 4.2) extends the basic pipeline by introducing a routing mechanism to create a descentralized, non-agentic workflow. This architecture first directs a user's question to a "router" node, which is a preliminary LLM call tasked with analyzing the query and determining the most appropriate tool or sequence of tools to use.
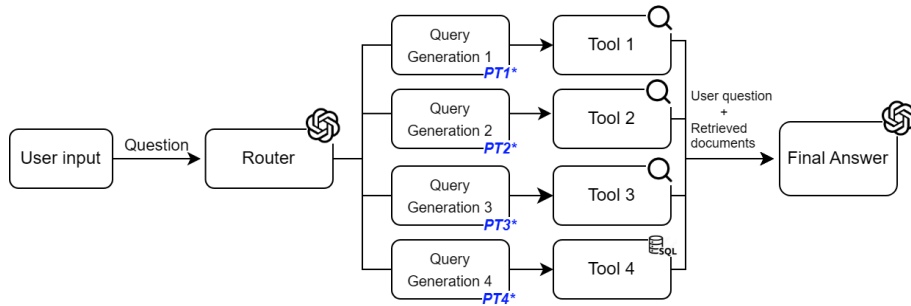


Figure 4.2: Linear-Flow with Router architecture.

This design enables the distribution of complex instruction prompts into smaller, more specialized nodes. Instead of one large prompt, several targeted sub-queries are generated, each dispatched to its respective tool. This approach offers two main advantages:

- **Specialization:** Each tool receives a query tailored to its specific function, leading to more accurate and relevant retrieval results.

- **Reduced Context:** By breaking down the master prompt, each LLM call operates on a smaller, more focused context, mitigating performance issues associated with long context windows.

### 4.3.4 Artifact 3: Single-Agent

The **Single-Agent** architecture (Figure 4.3) embodies a centralized agentic approach, building on the lessons from the first experimental cycle. In this setup, a single LLM agent manages the entire question-answering process. It has access to the full suite of tools and autonomously makes decisions about which to invoke, in what order, and how to synthesize the retrieved information into a final answer.
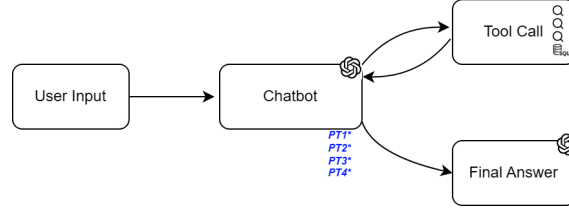
Figure 4.3: Single-Agent architecture.

The design emphasizes **end-to-end reasoning within a unified context**, allowing the model to maintain the same "thought process" from start to finish. This artifact tests the capability of a standalone LLM agent to manage a RAG workflow, balancing the tool calling for different knowledge sources, all without the communication overhead required by multi-agent systems.

### 4.3.5 Artifact 4: Multi-Agent Supervisor

The **Multi-Agent Supervisor** setup (Figure 4.4) implements a collaborative, hierarchical system to explore the benefits of distributed cognition. This architecture consists of two main components:

1. **A Supervisor Agent:** This master agent receives the user's query, analyzes it, and orchestrates the workflow by delegating these tasks to the appropriate specialist agents.

2. **Specialist Agents:** A team of agents, each focusing on a specific domain of knowledge or reasoning skill. For this experiment, each specialist was tied to a single tool (e.g., a "Learned Lessons Agent," an "HSE Alert Agent").

The supervisor orchestrates the collaboration, integrates the findings from each specialist, and synthesizes the potentially divergent information into a single, coherent final answer. This framework is designed to mimic real-world expert collaboration and tests whether decomposing a problem and assigning its parts to dedicated specialists yields a more accurate result.
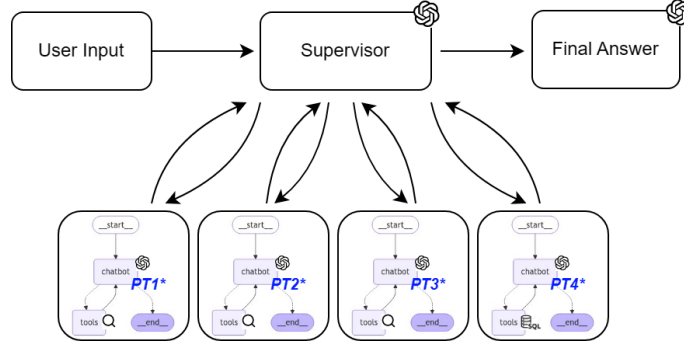
Figure 4.4: Multi-Agent Supervisor architecture with four specialist agents.

## 4.4 Evaluation

The evaluation phase was designed to be automated, scalable, and objective, addressing the limitations of the first experimental cycle.

### 4.4.1 Evaluation Methodology

The core of the evaluation is an automated execution loop (detailed in Algorithm 1) that runs each of the 33 questions through every combination of artifact (4 setups) and model (2 models), repeating each run three times to account for stochasticity.

---

**Algorithm 1** Experiment Execution Loop

**Require:** questions, setups, models
**Ensure:** results
 1: **function** RUNEXPERIMENT
 2:     $results \leftarrow \{\}$
 3:     **for all** $question \in questions$ **do**
 4:         $ground\_truth \leftarrow question.ground\_truth$
 5:         **for all** $setup \in setups$ **do**
 6:             **for all** $model \in models$ **do**
 7:                 **for** $i \in 1 \ldots 3$ **do**         ▷ Execute 3 times for consistency
 8:                     $agent \leftarrow$ InitializeAgent($setup, model$)
 9:                     $response \leftarrow agent.$ProcessQuestion($question$)
10:                     $metrics \leftarrow$ EvaluateResponse($response, ground\_truth$)
11:                     Store $metrics$ and $response$ in $results$
12:                 **end for**
13:             **end for**
14:         **end for**
15:     **end for**
16:     **return** AggregateResults($results$)
17: **end function**

---

The quality of each generated response is assessed using the **LLM-as-a-Judge** approach. A powerful LLM (GPT-4) is prompted to act as an impartial evaluator, comparing the generated answer against the ground-truth answer. The judge

decomposes both texts into atomic statements and classifies them to build a confusion matrix, from which the final metrics are calculated. The full prompt for the LLM-as-a-Judge can be found in Appendix A.1.

## 4.4.2 Data Set Creation

The experiment utilizes a curated dataset developed in collaboration with domain experts.

- **Questions Dataset:** A set of 17 questions reflecting real-world information needs of well engineers. Each question is paired with a manually created, expert-validated "ground-truth" answer.

- **Knowledge Bases:** The artifacts were given access to three distinct, pre-processed knowledge sources from within the organization, vectorized for semantic search:

  - **Learned Lessons:** A repository of learned lessons, best practices, and operational alerts.
  - **HSE Alerts:** A collection of ESG alerts and incident reports.
  - **Operational Reports:** A database of detailed daily operational reports from drilling rigs.

## 4.4.3 Evaluation Metrics

To provide a quantitative and objective assessment, the following information retrieval metrics, detailed in Section 2.5.2, were calculated for each response based on the LLM-as-a-Judge's analysis:

- **Precision:** Measures the accuracy of the information presented in the generated answer. It is the ratio of correct statements (True Positives) to the total number of statements made.

- **Recall:** Measures the completeness of the answer. It is the ratio of correct statements retrieved to the total number of statements available in the ground truth.

- **F1-Score:** The harmonic mean of Precision and Recall, providing a single, balanced measure of overall performance.

### 4.4.4 Results

To ensure a robust evaluation and account for the inherent non-determinism of language models, each of the 17 questions in the dataset was processed three times for every model and configuration combination. This experimental design resulted in a total of 408 executions (17 questions $\times$ 2 models $\times$ 4 configurations $\times$ 3 runs). Each of the 408 generated answers was then compared to a ground truth answer to calculate performance metrics.

The results presented in this section are derived from this set of runs. For each of the 136 unique combinations of question, model, and configuration, the best-performing run (out of three) was selected based on the F1-Score. The final metrics reported in Table 4.1 represent the average of these best-run scores across all 17 questions for each of the eight model-configuration pairs. This approach presents a clear view of the potential of each setup, with the F1-Score serving as the primary metric for performance evaluation.

Table 4.1: Detailed performance metrics by model and agent configuration. The best result for each metric is highlighted in bold and underlined. For the inferior model, the best result is only underlined.

| Model | Configuration | F1-Score | | | | Precision | | | | Recall | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Mean | Std. Dev. | Min | Max | Mean | Std. Dev. | Min | Max | Mean | Std. Dev. | Min | Max |
| GPT-4o | Linear-Flow (Baseline) | 0.581 | 0.204 | 0.000 | 1.000 | 0.656 | 0.262 | 0.000 | 1.000 | 0.548 | 0.201 | 0.000 | 1.000 |
| | Linear-Flow w/ Router | **0.702** | 0.202 | 0.333 | 1.000 | **0.805** | 0.185 | 0.400 | 1.000 | **0.674** | 0.242 | 0.286 | 1.000 |
| | Single-Agent | 0.643 | 0.213 | 0.364 | 1.000 | 0.751 | 0.198 | 0.400 | 1.000 | 0.618 | 0.240 | 0.294 | 1.000 |
| | Multi-Agent | 0.664 | 0.214 | 0.286 | 1.000 | 0.746 | 0.221 | 0.286 | 1.000 | 0.630 | 0.231 | 0.286 | 1.000 |
| GPT-4o-mini | Linear-Flow (Baseline) | 0.534 | 0.208 | 0.000 | 0.923 | 0.604 | 0.262 | 0.000 | 1.000 | 0.516 | 0.216 | 0.000 | 0.923 |
| | Linear-Flow w/ Router | 0.604 | 0.155 | 0.333 | 1.000 | 0.676 | 0.196 | 0.300 | 1.000 | 0.602 | 0.206 | 0.267 | 1.000 |
| | Single-Agent | 0.576 | 0.184 | 0.308 | 1.000 | 0.719 | 0.214 | 0.286 | 1.000 | 0.544 | 0.227 | 0.231 | 1.000 |
| | Multi-Agent | 0.596 | 0.182 | 0.348 | 1.000 | 0.687 | 0.198 | 0.400 | 1.000 | 0.578 | 0.201 | 0.235 | 1.000 |
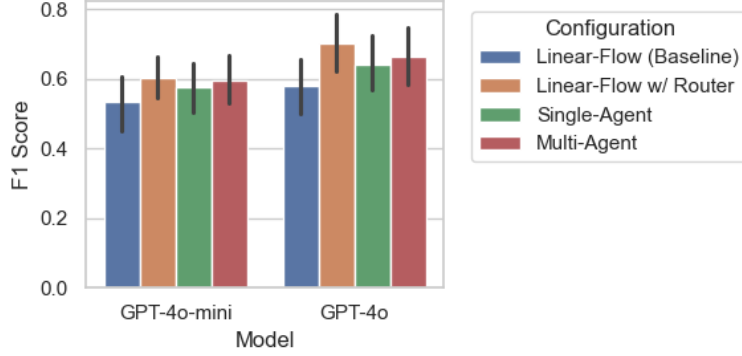
Figure 4.5: Best F1-Score by model and configuration.

The analysis of the results presented in Table 4.1 and visualized in Figure 4.5 reveals several insights into the performance of the different models and configurations. A primary observation is the consistent performance superiority of the GPT-4o model over its counterpart, GPT-4o-mini, across all tested configurations. The most capable configuration for GPT-4o, *Linear-Flow w/ Router*, achieved a mean F1-Score of 0.702. This represents a significant performance uplift of approximately 16.2% compared to the best score achieved by GPT-4o-mini (0.604), which was also with the *Linear-Flow w/ Router* configuration. This gap underscores the impact that the model's reasoning and instruction-following capabilities have on the overall performance of the system.

Furthermore, the results show that more complex configurations brought a notable improvement over the *Linear-Flow (Baseline)* for both models. The *Linear-Flow w/ Router* configuration emerged as the most effective architecture overall. For the superior GPT-4o model, this configuration boosted the mean F1-Score by a relative 20.8% over the baseline (from 0.581 to 0.702). It also increased the mean precision by 22.7% (from 0.656 to 0.805), indicating that the router is highly effective at selecting the correct reasoning path or tool, thereby reducing incorrect or irrelevant responses.

While the *Single-Agent* and *Multi-Agent* configurations also outperformed the baseline, they did not reach the performance level of the router-enhanced linear flow. For GPT-4o, the *Multi-Agent* setup (F1-Score 0.664) slightly outperformed the *Single-Agent* (F1-Score 0.643), but both fell short of the *Linear-Flow w/ Router*. This suggests that for the tasks in this experiment, the added complexity of reflective agent loops or multi-agent collaboration did not yield a proportional benefit over a more direct, intelligent tool-routing approach.

A crucial aspect of the results is the high standard deviation observed across all configurations, typically around 0.20 for the F1-Score. The wide range between minimum and maximum scores indicates that performance is highly variable and

question-dependent. This suggests that even the best-performing systems can fail completely on certain queries.

## 4.4.5 Discussion

The results from this second experimental cycle present a series of compelling, and in some aspects, counter-intuitive insights into the application of LLM-based architectures in specialized technical domains. The most significant finding, which stands in contrast to the prevailing trends in agentic AI and some initial findings from our first experiment, is the superior performance of a non-agentic configuration (*Linear-Flow w/ Router*) over its more complex, cyclical agentic counterparts. This outcome challenges the assumption that increased agent complexity, with its capacity for reflection and iterative refinement, universally leads to better performance. This discussion will explore the primary hypothesis for this phenomenon, explore other contributing factors, and consider the broader implications for designing AI systems in niche domains.

### The Domain Knowledge Deficit: Why Agentic Reflection Fails

The core benefit of an agentic architecture, whether single or multi-agent, lies in its ability to perform cyclical reasoning. An agent can call a tool, assess the output, reflect on its progress, and decide on a new course of action, potentially correcting earlier mistakes or refining its strategy. This iterative process is a form of simulated cognition. However, we hypothesize that the effectiveness of this reflective capability is fundamentally contingent on the Large Language Model's pre-existing, foundational knowledge of the subject matter.

For an LLM to effectively judge the output of a tool or the partial answer from a sub-agent, it must have a robust internal model of what constitutes a "good" or "correct" answer in that domain. Consider, for example, the domain of software engineering. LLMs like GPT-4o are extensively trained on vast repositories of code, documentation, and programming discussions. When an agent generates a piece of code, the LLM can "read" it, understand its logic, identify bugs, and suggest improvements because it has been trained on countless similar examples. In this context, a cyclical, reflective flow is highly effective because the LLM is a competent judge of its own (or its peers') output.

The domain of this study—well construction engineering—presents a starkly different scenario. The knowledge is highly specialized, laden with niche terminology, and often contained within proprietary corporate documents that do not form a significant part of the public web crawl used to train general-purpose LLMs. Consequently, when an agent in our experiment retrieves a technical snippet from a

lessons-learned document, the LLM lacks the deep, specialized knowledge required to effectively critique it. It cannot reliably discern subtle inaccuracies, determine if the context is fully appropriate, or judge whether a sub-agent's reasoning is sound from an engineering perspective.

In this context of a "domain knowledge deficit," the cyclical flow of an agentic system becomes a liability rather than an asset. The reflective loop introduces computational and cost overhead (more LLM calls, more complex state management) without a corresponding improvement in the quality of reasoning. The agent may cycle, but it does so without true insight, making the additional complexity ineffective. This leads to the observed result: the more straightforward, non-agentic approach outperforms it.

### The Unsurprising Efficacy of Intelligent Routing

While the agentic systems underperformed, the success of the *Linear-Flow w/ Router* configuration is, in itself, a significant finding. Its superior performance can be attributed to its focused efficiency. Rather than engaging in a complex, multi-step reasoning process, this architecture excels at a single, critical task: upfront intent classification and tool selection.

The initial "router" call is a highly targeted use of the LLM's reasoning power. Its sole purpose is to analyze the user's query and map it to the most appropriate knowledge base (tool). This is a task that even a general-purpose LLM can perform well, as it relies on semantic understanding rather than deep domain expertise. By correctly identifying the right tool from the outset, the router ensures that the subsequent retrieval step is already on the right path.

This approach of "decide once, execute well" proved more effective for this dataset than the agents' "execute, reflect, re-execute" loop. It avoids the risk of error propagation inherent in cyclical systems. In an agentic loop, a minor misinterpretation in an early step can be amplified in subsequent cycles as the agent doubles down on a flawed path. The linear flow of the router configuration is immune to this, as there are no subsequent cycles to compound an error. The simplicity of its prompt and execution logic also reduces the chance of "meta-errors," where the LLM becomes confused by the complex state and instructions of a multi-turn agentic conversation.

### Other Plausible Contributing Factors

Beyond the primary hypothesis of the domain knowledge deficit, several other factors likely contributed to the observed results:

Nature of the Task: The questions in the experimental dataset, while technically complex, are primarily information retrieval tasks. They demand finding the correct

facts from the knowledge base and synthesizing them into an answer. They do not necessarily require the kind of complex, multi-step deliberation or creative problem-solving where a reflective agent might theoretically excel. For such "retrieve-and-synthesize" queries, optimizing the retrieval step—as the router does—yields the greatest performance gain.

Model Capabilities (GPT-4o vs. GPT-4o-mini): The consistent and significant performance gap between GPT-4o and GPT-4o-mini across all configurations underscores the critical importance of the underlying model's reasoning and instruction-following capabilities. The superior performance of the *Linear-Flow w/ Router* on GPT-4o (F1-Score 0.702) compared to the same configuration on GPT-4o-mini (F1-Score 0.604) highlights that even the most effective architecture is gated by the power of the engine driving it.

High Performance Variance: A crucial observation is the high standard deviation across all configurations. This indicates that performance is highly dependent on the specific question being asked. Even the best-performing system failed completely on some queries (minimum F1-Score of 0.333), while succeeding perfectly on others (maximum F1-Score of 1.000). This suggests that no single architecture is a panacea. Certain query structures or topics may inherently favor one approach over another, reinforcing the conclusion that the optimal architecture is task- and domain-dependent.

In conclusion, this experiment provides strong evidence that for specialized domains where LLMs lack deep pre-existing knowledge, the architectural focus should be on optimizing the interface between the query and the external knowledge tools. A simple, intelligent routing mechanism that accurately directs a query to the correct source can be more robust and effective than a complex, cyclical agentic system that attempts to "reason" in a domain it does not truly understand. The assumption that more complexity equals better performance is a fallacy; true effectiveness comes from aligning the architectural design with both the nature of the task and the inherent capabilities and limitations of the language model itself.

# Chapter 5

# Conclusions

The results of this study highlight the potential of multi-agent architectures based on LLMs in the O&G sector, particularly in the domain of well engineering. The ability to process and respond to complex queries paves the way for a significant digital transformation in the area.

Our comparative analysis of single-agent and multi-agent architectures, using GPT-3.5-turbo and GPT-4, reveals a detailed landscape of trade-offs between performance and economic efficiency. Multi-agent systems demonstrate 28% greater factuality in question-and-answer (Q&A) tasks, especially with GPT-4, compared to single-agent systems. However, they incur LLM costs that are, on average, 3.7 times higher due to the complexities of inter-agent communication. In contrast, single-agent systems excel in Text-to-SQL tasks, showing 15% better performance than multi-agent setups. This cost-benefit dynamic requires careful consideration when implementing RAG in real-world scenarios, where accuracy and financial constraints must be balanced.

We highlight several challenges encountered during our experiments, including issues with contextualization, the need for more refined information filtering, and the persistence of hallucinations. These challenges underscore the need for continuous research in areas such as domain-specific specialized models, advanced semantic search techniques, and hybrid architectures that combine the strengths of single-agent and multi-agent systems.

The practical implications of this study extend beyond the O&G sector. The insights gained here are applicable to any knowledge-intensive domain that deals with large volumes of technical data. By focusing on enhancing retrieval mechanisms, developing domain-specific LLMs, and optimizing interactions between agents and tools, we pave the way for more effective, reliable, and cost-efficient RAG solutions across various sectors.

The main points of the study are as follows: multi-agent systems offer superior factuality in Q&A tasks, albeit at a significantly higher cost. Single-agent archi-

tectures, on the other hand, excel at Text-to-SQL tasks. Despite the advantages, several challenges persist, including issues with contextualization, filtering, hallucination, and domain-specific vocabulary.

Future research should focus on developing specialized models, advancing retrieval techniques, and exploring hybrid architectures. The lessons learned from this study have broader implications and can extend to other complex technical domains. By addressing the limitations identified in this study and embracing emerging trends in multi-agent systems and RAG technology, we can unlock their full potential, revolutionizing decision-making, knowledge management, and operational efficiency in complex industries worldwide.

[CONCLUSIONS FROM EXPERIMENT 2 ARE YET TO BE WRITTEN HERE]

[INCLUIR: COMENTAR QUE OS PRECOS DE LLM ESTAO CAINDO MTO, MODELOS PEQUENOS COM EXCELENTE DESEMPENHO, O QUE TORNA A ANALISE FINANCEIRA DO 1o CICLO OBSOLETA]

# References

OSWALD, M. L. M. B., HOFFMANN, A., DE MELLO E SILVA, D., et al. *Metodologias De Pesquisa Online: Investigando Em Rede Com o Outro*. 1 ed. Rio de Janeiro, Ayvu Editora, 2023.

WU, Q., BANSAL, G., ZHANG, J., et al. "AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation", 2023. doi: 10.48550/arXiv .2308.08155.

KAR, A. K., VARSHA, P. S. "Unravelling the Impact of Generative Artificial Intelligence (GAI) in Industrial Applications: A Review of Scientific and Grey Literature", *Global Journal of Flexible Systems Management*, v. 24, pp. 659–689, 12 2023. ISSN: 09740198. doi: 10.1007/s40171-023-00356-x.

ECKROTH, J., GIPSON, M. "Answering Natural Language Questions with OpenAI's GPT in the Petroleum Industry", pp. 16–18, 2023. doi: 10.2118/214888-MS. Available at: <http://onepetro.org/SPEAT CE/proceedings-pdf/23ATCE/3-23ATCE/D031S032R005/3301837/spe -214888-ms.pdf/1>.

DELLACQUA, F., SARAN, A., MCFOWLAND, R. E., et al. "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality", *Harvard Business School: Technology and Operations Management Unit Working Paper Series*, 2023. doi: 10.2139/ssrn.4573321. Available at: <https://ssrn.com/abstract=4573321>.

PETROBRAS. "Petrobras 2023 Results - Unaudited Condensed Consolidated Interim Financial Statements". 2024. Available at: <https://www.invest idorpetrobras.com.br/en/results-and-announcements/results-c enter/>.

HATZIUS, J., BRIGGS, J., KODNANI, D., et al. "The Potentially Large Effects of Artificial Intelligence on Economic Growth". 2023.

SINGH, A., JIA, T., NALAGATLA, V. "Generative AI Enabled Conversational Chatbot for Drilling and Production Analytics", *Day 2 Tue, October 03, 2023*, 10 2023. doi: 10.2118/216267-MS. Available at: <https://onep etro.org/SPEADIP/proceedings/23ADIP/2-23ADIP/D021S065R002/5 34485>.

HADI, M. U., QASEM AL TASHI, QURESHI, R., et al. "A Survey on Large Language Models: Applications, Challenges, Limitations, and Practical Usage", 2023. doi: 10.36227/techrxiv.23589741.v1. Available at: <https: //doi.org/10.36227/techrxiv.23589741.v1>.

BADIRU, A. B., OSISANYA, S. O. *Project Management for the Oil and Gas Industry: A World System Approach*. Boca Raton, FL 33487-2742, CRC Press, 1 2016. ISBN: 9781420094268. doi: 10.1201/b13755.

THOMAS, J. E. *Fundamentos de Engenharia de Petróleo*. 2nd ed. Av. Presidente Vargas 435, Rio de Janeiro, RJ - 20.077-900, Editora Interciência, 2004.

GU, J., JIANG, X., SHI, Z., et al. "A Survey on LLM-as-a-Judge", 3 2025. Available at: <http://arxiv.org/abs/2411.15594>.

HEVNER, A. R. "A three cycle view of design science research", *Scandinavian journal of information systems*, v. 19, n. 2, pp. 4, 2007.

BRAVO, C., SAPUTELLI, L., RIVAS, F., et al. "State of the art of artificial intelligence and predictive analytics in the E&P industry: A technology survey", *SPE Journal*, v. 19, n. 4, pp. 547–563, 2014. ISSN: 1086055X. doi: 10.2118/150314-pa. Available at: <http://onepetro.org/SJ/arti cle-pdf/19/04/547/2099035/spe-150314-pa.pdf/1>.

GUDALA, M., NAIYA, T. K., GOVINDARAJAN, S. K. "Remediation of heavy oil transportation problems via pipelines using biodegradable additives: An experimental and artificial intelligence approach", *SPE Journal*, v. 26, n. 2, pp. 1050–1071, apr 2021. ISSN: 1086055X. doi: 10.2118/203824-PA.

GOHARI, M. S. J., NIRI, M. E., SADEGHNEJAD, S., et al. "Synthetic Graphic Well Log Generation Using an Enhanced Deep Learning Workflow: Imbalanced Multiclass Data, Sample Size, and Scalability Challenges", *SPE Journal*, v. 29, pp. 1–20, 2024. ISSN: 1086055X. doi: 10.2118/217466-PA. Available at: <http://onepetro.org/SJ/article-pdf/29/01/1/335 8626/spe-217466-pa.pdf/1>.

RAHMANI, A. M., AZHIR, E., ALI, S., et al. "Artificial intelligence approaches and mechanisms for big data analytics: a systematic study", *PeerJ Computer Science*, v. 7, pp. 1–28, 4 2021. ISSN: 23765992. doi: 10.7717/peerj-cs.488.

LIDDY, E. *Natural Language Processing*. Encyclopedia of Library and Information Science, 2001. Available at: <https://surface.syr.edu/istpub>.

ANTONIAK, M., DALGLIESH, J., VERKRUYSE, M., et al. "Natural language processing techniques on oil and gas drilling data", *Society of Petroleum Engineers - SPE Intelligent Energy International Conference and Exhibition*, 2016. doi: 10.2118/181015-MS.

CASTIÑEIRA, D., TORONYI, R., SALERI, N. "Machine Learning and Natural Language Processing for Automated Analysis of Drilling and Completion Data", pp. 23–26, 2018. doi: 10.2118/192280-MS. Available at: <http://onepetro.org/SPESATS/proceedings-pdf/18SATS/All-18SATS/SPE-192280-MS/1246545/spe-192280-ms.pdf/1>.

ISKE, P., BOERSMA, W. "Connected brains. Question and answer systems for knowledge sharing: Concepts, implementation and return on investment". 2005. ISSN: 13673270.

TREUDE, C., BARZILAY, O., STOREY, M.-A. "How do programmers ask and answer questions on the web? (NIER track)". In: *Proceedings of the 33rd International Conference on Software Engineering*, ICSE '11, p. 804–807, New York, NY, USA, 2011. Association for Computing Machinery. ISBN: 9781450304450. doi: 10.1145/1985793.1985907. Available at: <https://doi.org/10.1145/1985793.1985907>.

QIN, B., HUI, B., WANG, L., et al. "A Survey on Text-to-SQL Parsing: Concepts, Methods, and Future Directions", 8 2022. doi: 10.48550/arXiv.2208.13629. Available at: <http://arxiv.org/abs/2208.13629>.

OPENAI. "Embeddings - OpenAI API". 2023. Available at: <https://platform.openai.com/docs/guides/embeddings>.

DENG, X., AWADALLAH, A. H., MEEK, C., et al. "Structure-Grounded Pretraining for Text-to-SQL", pp. 1337–1350, 2021. doi: 10.18653/v1/2021.naacl-main.105. Available at: <http://dx.doi.org/10.18653/v1/2021.naacl-main.105>.

DENG, X., GU, Y., ZHENG, B., et al. "MIND2WEB: Towards a Generalist Agent for the Web", 2023. Available at: <https://osu-nlp-group.github.io /Mind2Web>.

RUSSELL, S. *Artificial intelligence: a modern approach.* Pearson, 2020. ISBN: 0134610997.

GOKULAN, B. P., SRINIVASAN, D. "An Introduction to Multi-Agent Systems". v. 310, pp. 1–27, 07 2010. ISBN: 978-3-642-14434-9. doi: 10.1007/978-3 -642-14435-6_1.

FLORES-MENDEZ, R. A. "Towards a standardization of multi-agent system framework", *XRDS*, v. 5, n. 4, pp. 18–24, Jun. 1999. ISSN: 1528-4972. doi: 10.1145/331648.331659. Available at: <https://doi.org/10.114 5/331648.331659>.

NAGATA, T., SASAKI, H. "A multi-agent approach to power system restoration", *IEEE Transactions on Power Systems*, v. 17, n. 2, pp. 457–462, 2002. doi: 10.1109/TPWRS.2002.1007918.

VASWANI, A., BRAIN, G., SHAZEER, N., et al. "Attention Is All You Need". 2017. Available at: <https://arxiv.org/abs/1706.03762>.

OPENAI, ACHIAM, J., ADLER, S., et al. "GPT-4 Technical Report", v. 4, pp. 1–100, 2023. doi: 10.48550/arXiv.2303.08774. Available at: <http: //arxiv.org/abs/2303.08774>.

MOSSER, L., AURSAND, P., BRAKSTAD, K. S., et al. "Exploration Robot Chat: Uncovering Decades of Exploration Knowledge and Data with Conversational Large Language Models". In: *Day 1 Wed, April 17, 2024*. SPE, apr 2024. doi: 10.2118/218439-MS. Available at: <https://onepetro.org /SPEBERG/proceedings/24BERG/1-24BERG/D011S002R006/544177>.

GRANDVIEWRESEARCH. "Large Language Models Market Size, Share and Trends Analysis Report". 6 2025. Available at: <https://www.grandvie wresearch.com/industry-analysis/large-language-model-llm-mar ket-report>. Accessed: 2025-06-30.

KADDOUR, J., HARRIS, J., MOZES, M., et al. "A Comprehensive Overview of Large Language Models", *arXiv preprint arXiv:2307.06435*, 2023. Available at: <https://arxiv.org/abs/2307.06435>.

LABS, V. "11 Best Applications of Large Language Models (LLMs) [2025]". 5 2025. Available at: <`https://www.v7labs.com/blog/best-llm-applications`>. Accessed: 2025-06-30.

AI, K. "My top 10 LLM research papers in 2024". 1 2025. Available at: <`https://www.keywordsai.co/blog/top-10-llm-research-papers`>. Accessed: 2025-06-30.

LEWIS, P., PEREZ, E., PIKTUS, A., et al. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks". 2020. Available at: <`https://github.com/huggingface/transformers/blob/master/`>.

LI, H., SU, Y., CAI, D., et al. "A Survey on Retrieval-Augmented Text Generation", 2 2022. doi: 10.48550/arXiv.2202.01110.

LIU, J., JIN, J., WANG, Z., et al. "RETA-LLM: A Retrieval-Augmented Large Language Model Toolkit", 6 2023. Available at: <`https://arxiv.org/abs/2306.05212v1`>.

ZHAO, R., CHEN, H., WANG, W., et al. "Retrieving Multimodal Information for Augmented Generation: A Survey", 3 2023. doi: 10.48550/arXiv.2303.10868. Available at: <`http://arxiv.org/abs/2303.10868`>.

XI, Z., CHEN, W., GUO, X., et al. "The Rise and Potential of Large Language Model Based Agents: A Survey". 2023.

LI, J., ZHANG, Q., YU, Y., et al. "More Agents Is All You Need", 2024a. doi: 10.48550/arXiv.2402.05120.

LIN, S., HILTON, J., EVANS, O. "TruthfulQA: Measuring How Models Mimic Human Falsehoods". 2022. Available at: <`https://arxiv.org/abs/2109.07958`>.

LI, C., WANG, J., ZHANG, Y., et al. "Large Language Models Understand and Can Be Enhanced by Emotional Stimuli", 2023. doi: 10.48550/arXiv.2307.11760.

LIN, S., OPENAI, J. H., EVANS, O. "TruthfulQA: Measuring How Models Mimic Human Falsehoods", 2021. Available at: <`https://github.com/`>.

LI, H., DONG, Q., CHEN, J., et al. "LLMs-as-Judges: A Comprehensive Survey on LLM-based Evaluation Methods". 2024b. Available at: <`https://arxiv.org/abs/2412.05579`>.

ZHENG, L., CHIANG, W.-L., SHENG, Y., et al. "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena", 12 2023. Available at: <http://arxiv.org/abs/2306.05685>.

LANGCHAIN. "How and when to build multi-agent systems (accessed 6 July 2025)". 6 2025a. Available at: <https://blog.langchain.com/how-and-when-to-build-multi-agent-systems/>.

CARRARO, D. "Enhancing Recommendation Diversity by Re-ranking with Large Language Models", 2024. doi: 10.48550/arXiv.2401.11506.

SUN, W., YAN, L., MA, X., et al. "Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agents", 2023. doi: 10.48550/arXiv.2304.09542. Available at: <https://arxiv.org/abs/2304.09542>.

BILBAO, D., GELBUKH, A., RODRIGO, A., et al. "A Mathematical Investigation of Hallucination and Creativity in GPT Models", *Mathematics 2023*, v. 11, pp. 2320, 5 2023. ISSN: 2227-7390. doi: 10.3390/MATH11102320. Available at: <https://www.mdpi.com/2227-7390/11/10/2320/htmhttps://www.mdpi.com/2227-7390/11/10/2320>.

SHAH, B. "Large Learning Models: The Rising Demand of Specialized LLM's". 2024. Available at: <https://blogs.infosys.com/emerging-technology-solutions/artificial-intelligence/large-learning-models-the-rising-demand-of-specialized-llms.html>.

MEENA, S. "The Future of Large Language Models: Evolution, Specialization, and Market Dynamics". 2023. Available at: <https://www.linkedin.com/pulse/future-large-language-models-evolution-specialization-shekhar-meena/>.

GHOSH, B. "Emerging Trends in LLM Architecture,". 2023. Available at: <https://medium.com/@bijit211987/emerging-trends-in-llm-architecture-a8897d9d987b>.

LANGCHAIN. "Self-query Retriever". 2023. Available at: <https://python.langchain.com/docs/modules/data_connection/retrievers/self_query/>.

LEVENSHTEIN, V. "Binary codes capable of correcting deletions, insertions, and reversals", *Cybernetics and Control Theory*, v. 10, 1966.

SHINN, N., CASSANO, F., BERMAN, E., et al. "Reflexion: Language Agents with Verbal Reinforcement Learning", 3 2023. Available at: <`http://arxiv.org/abs/2303.11366`>.

PAL, S., BHATTACHARYA, M., LEE, S. S., et al. "A Domain-Specific Next-Generation Large Language Model (LLM) or ChatGPT is Required for Biomedical Engineering and Research". 3 2024. ISSN: 15739686.

AREFEEN, A., DEBNATH, B., CHAKRADHAR, S. "LeanContext: Cost-efficient domain-specific question answering using LLMs", *Natural Language Processing Journal*, v. 7, pp. 100065, 2024. doi: 10.1016/j.nlp.2024.100065. Available at: <`https://doi.org/10.1016/j.nlp.2024.100065`>.

LANGCHAIN. "Introduction to LangChain (accessed 6 July 2025)". 2025b. Available at: <`https://python.langchain.com/docs/introduction/?_gl=1*n3yqvs*_gcl_au*MTI5OTg4NDQ5Mi4xNzQ4MjE2OTYz*_ga*MTM4MDIxMjAyOS4xNzQ4MjE2OTYz*_ga_47WX3HKKY2*czE3NTE4MTExNDEkbzEyJGcwJHQxNzUxODExMTQxJGo2MCRsMCRoMA..`>.

LANGGRAPH. "Overview of Langgraph Framework (accessed 6 July 2025)". 2025. Available at: <`https://langchain-ai.github.io/langgraph/concepts/why-langgraph/`>.

# Appendix A

# Experiment 2

## A.1  Dataset

Table A.1: Dataset used in the thesis experiments (Landscape).

| Question | Ground Truth |
|---|---|
| Cite acidentes envolvendo colisão entre top drive e outros equipamentos. | 1. Colisão do Top Drive com Coluna de Produção na Mesa Auxiliar Durante a montagem de coluna de produção 6 5/8" na mesa auxiliar, o top drive colidiu com o último tubo acunhado na mesa rotativa, causando empenamento do tubo. A causa foi falha de planejamento e avaliação de risco, pois o top drive não foi retraído totalmente antes da descida.<br>2. Colisão entre o Top Drive e a Hydraracker Durante manobra de descida de coluna de Drill Pipe 5", o top drive colidiu com o braço principal da Hydra Racker que estava sendo estendido. A colisão provocou o desprendimento entre o braço principal e o braço inferior da Hydra Racker, com queda de uma peça de 2,7kg de altura de 27m. Causas incluíram operação em modo manual e configuração inadequada do sistema anticolisão.<br>3. Colisão do Top Drive com Seção de Drill Pipe Durante manobra de retirada do BHA de 8 7/8", o top drive colidiu com a seção de Drill Pipe sustentada pelo PRS (Pipe Racking System). Com o impacto, a carenagem frontal do top drive foi cisalhada e caiu no piso do rigfloor de uma altura de 20m, pesando aproximadamente 85kg.<br>4. Interação do Top Drive Auxiliar com Equipamentos de Segurança Durante conexão de mangueira MPD, o cabo da trava quedas de um plataformista ficou preso no arranjo de içamento do top drive auxiliar, provocando a suspensão do trabalhador. O incidente ocorreu por falha na análise de risco e comunicação inadequada. |
| Quais as causas e falhas típicas de acidentes ou incidentes envolvendo guarda corpo? Cite alertas que retratem isso. | Os acidentes ou incidentes envolvendo guarda-corpo geralmente ocorrem devido a falhas de projeto, gestão de processos, especificação inadequada do serviço, e falta de procedimentos específicos. Aqui estão algumas causas e falhas típicas.<br>1. Falha de Equipamentos e Projeto de Engenharia Deficiente: O projeto do guarda-corpo pode não ser adequado para a operação, levando a deslocamentos inesperados e acidentes. (Fonte: Definitivo - Lesão no dorso do pé direito devido ao choque com guarda corpo_ POCOS SM _Abrange+2023-000154)<br>2. Falha na Gestão de Processos e Pessoas: Inclui falhas no processo de qualificação da empresa contratada, violação de procedimentos por supervisores, e falta de documentos específicos para avaliação de operações. (Fonte: Definitivo - Lesão no dorso do pé direito devido ao choque com guarda corpo_ POCOS SM _Abrange+2023-000154)<br>3. Falha de Identificação de Risco Adicional de Queda: Não identificar riscos adicionais, como vãos entre a gaiola da escada e o guarda-corpo, pode resultar em quedas fatais. (Fonte: Acidente fatal-Queda de altura - REVAP-MA-PM_ Abrange 2023-000279)<br>4. Movimento Irregular da Carga e Falha no Freio do Guindaste: Durante movimentações de carga, falhas no equipamento podem causar impactos que danificam o guarda-corpo e resultam em quedas de objetos. (Fonte: Alerta Definitivo - Vermelho - Descida descontrolada de carga - Divulgação equipe movimentação carga) |

| Question | Ground Truth |
|---|---|
| Quais foram as principais falhas identificadas em acidentes com pés, de acordo com os Alertas de SMS? | As principais falhas identificadas em acidentes envolvendo pés incluem:<br>1. Planejamento inadequado e falta de análise de risco: Muitos acidentes ocorreram devido à falta de planejamento adequado e análise de risco antes de realizar atividades, como movimentação de cargas e içamento de peças. Isso foi observado em acidentes como o de fratura em três metatarsos do pé esquerdo durante a fabricação de um corrimão na oficina de solda ("Preliminar AMARELO - Acidente com Lesão - Fratura - Geral - POCOS SM 055 2022").<br>2. Uso inadequado de Equipamentos de Proteção Individual (EPI): Em alguns casos, o uso inadequado ou a falta de EPI apropriado, como botas de segurança com proteção para metatarso, contribuiu para a ocorrência de acidentes. Isso foi destacado em acidentes como o corte contuso no pé esquerdo durante a manutenção de um gerador ("Definitivo AMARELO - Corte contuso no pé_ Geral_ POCOS SM 019_rev01").<br>3. Falta de percepção de riscos e consciência situacional: A falta de percepção de riscos e a escolha de caminhos mais curtos sem avaliar o ambiente contribuíram para acidentes, como a torção no tornozelo ao pisar em uma madeira de apoio de carga ("Definitivo -Torção no tornozelo ao pisar sobre uma peça de madeira - POCOS SM 025 2023").<br>4. Falhas na supervisão e no cumprimento de diretrizes: A supervisão inadequada e o não cumprimento de diretrizes e procedimentos de segurança foram fatores em acidentes, como o de fratura no tornozelo ao tropeçar em um umbilical no convés ("Definitivo VERMELHO - Fratura no tornozelo_Geral_ POCOS SM 016 2021").<br>5. EPI Inadequado: Uso de botas sem características anti-torção e solado adequado para ambientes offshore, como mencionado no acidente de torção de tornozelo ao descer escada (POCOS SM 009 2021). |
| Como a presença de sílica na composição da pasta de cimento afeta sua estabilidade térmica em altas temperaturas, de acordo com as lições aprendidas? | A sílica é adicionada ao cimento na proporção de 35% por peso de cimento com o objetivo de mitigar a retrogressão (perda da resistência compressiva) do cimento no poço quando submetido a temperaturas superiores a 230°F (110°C). Esta prática é comum em poços profundos ou com gradiente geotérmico elevado que, em suas fases finais, ultrapassam a temperatura limite. A presença de sílica na pasta de cimento também pode ser necessária em situações onde a temperatura nos anulares dos revestimentos durante a produção do poço seja bastante elevada, mesmo em menores profundidades.<br>A ausência de sílica no cimento sujeito a temperaturas superiores a 230°F pode resultar numa perda de resistência compressiva no cimento, levando a problemas de isolamento.<br>Entretanto, é importante observar que a má preparação da pasta de cimento pode resultar em propriedades da pasta inhomogêneas durante a fabricação e bombeamento. Isso pode levar a gelificação prematura e possível prisão da coluna. Em caso de anormalidades na preparação da pasta, recomenda-se a realização da cimentação em bateladas.<br>Referências:<br>Alerta Técnico AT5533: Utilização de cimento com sílica na cimentação de revestimentos de superfície. Autor: Jan Fernandes Aslan. Data: 16/11/2020.<br>Alerta Técnico AT10479: Substituição indevida de cimento com sílica por cimento puro em cimentação de revestimento de superfície em poço com temperatura elevada. Autor: Rafael Peralta Muniz Moreira. Data: 14/02/2021. |
| | |

| Question | Ground Truth |
|---|---|
| De acordo com as lições aprendidas, como a inclinação do poço afeta a eficácia da cimentação durante o tamponamento through tubing? | Durante o tamponamento through tubing em poços de alta inclinação (maior que 45 graus), a eficácia da cimentação pode ser afetada por fatores advindos da inclinação do poço. Segundo um Alerta Técnico (AT9613), essa situação aumenta a complexidade das operações de tamponamento devido ao fato de a coluna de produção/injeção estar descentralizada dentro do revestimento de produção. Além disso, outras dificuldades surgem, como: - Dificuldade no assentamento de cement retainer inflável - Dificuldade na descida de ferramentas de perfilagem e plugs para isolamento preliminar abaixo do trecho do poço a ser tamponado Estes fatores podem resultar em poços com qualidade média ou falha na cimentação. Recomenda-se por agora, até que estudos mais aprofundados sejam realizados, uma análise criteriosa de cada cenário e a escolha de alternativas mais robustas para o abandono de poços de alta inclinação e a seleção de escopo de projeto executivo de abandono, detalhamento operacional e estimativa dos riscos e tempos da intervenção. Por outro lado, existem algumas alternativas como o Bombeio Direto de Cimento (BDC) para abandono permanente through tubing (TT). Isso permite o bombeio de pasta de cimento diretamente pelo riser dual bore, passando pela COP/COI até ser posicionada em anular "A" (COP x revestimento). No entanto, essa técnica também tem suas restrições e premissas, como necessidade de emprego de Riser Dual Bore e menor tolerância ao erro no deslocamento de pasta. Em resumo, a inclinação do poço pode afetar a eficácia da cimentação durante o tamponamento através do tubo, tornando a operação mais complexa e aumentando o risco de falhas na cimentação. Diferentes abordagens e tecnologias estão sendo estudadas para otimizar essas operações e minimizar os riscos. |
| De acordo com as lições aprendidas, quais são os riscos associados à substituição de cimento com sílica por cimento puro em cimentações de revestimento de superfície em poços com temperatura elevada? | Estes são os principais riscos associados à substituição indevida de cimento com sílica por cimento puro em poços de alta temperatura: 1. O cimento puro, quando submetido a altas temperaturas (acima de 110°C), sofre uma regressão de sua resistência compressiva ao longo do tempo, podendo comprometer sua função em poços de petróleo como elemento de barreira. 2. Falhas na cimentação do revestimento de superfície podem levar a custos e impactos adicionais, como o vazamento de fluidos de perfuração para o fundo do mar e o aumento das cargas hidrostáticas sobre o revestimento de superfície e os revestimentos subsequentes instalados. 3. Em alguns casos, a bainha de cimento pode ser exposta a temperaturas acima de 110°C em algum momento do ciclo de vida do poço. Nesses casos, o uso de cimento com sílica deve ser mantido para garantir a estabilidade térmica da pasta de cimento. 4. A substituição indevida pode levar a outros riscos como escape de fluidos de perfuração do anular C para o fundo do mar, alterações no dimensionamento do revestimento, alterações no desenho de projeto de início de poço, entre outros. Deve-se ressaltar que a intenção de reduzir o tempo de aguardo de cura do cimento não justifica a substituição do cimento com sílica por cimento puro, pois existem outros métodos para alcançar uma cura mais rápida, como o aumento da densidade da pasta ou o uso de aditivos. Esse fenômeno é conhecido na indústria do petróleo desde 1954 e foi mitigado pela introdução de sílica cristalina na composição da pasta de cimento para garantir a estabilidade térmica. |
| | Continued on next page |

| Question | Ground Truth |
|---|---|
| Quais estratégias podem ser adotadas para melhorar a qualidade da cimentação em poços de alta inclinação durante o tamponamento through tubing, de acordo com as lições aprendidas? | Com base nos itens de conhecimento obtidos, algumas estratégias para melhorar a qualidade da cimentação em poços de alta inclinação durante o tamponamento through tubing incluem: <br> 1. Aumento das vazões no deslocamento quando a cimentação é executada com flexitubo. Este procedimento auxilia na melhor distribuição do cimento no espaço anular e contribui para uma melhor qualidade de cimentação. <br> 2. Utilização de dispositivos para vibração da coluna durante a cimentação. A vibração pode ajudar a melhorar a aderência do cimento no espaço anular tornando a cimentação mais eficaz. <br> 3. Utilização de volumes de excesso de pasta de cimento compatíveis com uma operação que busque robustez máxima. Este procedimento garante que haja cimento suficiente para preencher o espaço anular ao longo de todo o comprimento da coluna. <br> 4. Análise criteriosa de cada cenário e seleção de uma alternativa mais robusta para o abandono (preferencialmente com a retirada da coluna). Isso é particularmente importante para poços de alta inclinação e outros cenários que adicionem complexidade ao tamponamento. <br> 5. Previsão no cronograma da intervenção da perfilagem de avaliação da cimentação como método de verificação. Isso permite a identificação de possíveis falhas ou áreas de cimentação insatisfatória para um possível retoque da cimentação. <br> 6. Uso do Bombeio Direto de Cimento (BDC) para abandono permanente through tubing, uma técnica que pode otimizar o tempo e reduzir os riscos operacionais, embora com restrições específicas. <br> Por último, é importante lembrar que toda situação é única e deve ser avaliada individualmente de acordo com as suas características específicas. |
| Quais são as alternativas para acelerar o tempo de cura da pasta de cimento sem comprometer sua integridade em condições de temperatura elevada, de acordo com as lições aprendidas? | As melhores práticas incluem o aumento da densidade da pasta, o uso de aditivos e a consideração da rampa de cura, mas é importante manter a utilização de cimento com sílica para garantir a integridade da cimentação em condições de temperatura elevada. |

67

| Question | Ground Truth |
|---|---|
| Quais são os principais desafios e riscos associados ao tamponamento through tubing em poços de alta inclinação, de acordo com as lições aprendidas? | Existem vários desafios e riscos associados ao tamponamento through tubing em poços de alta inclinação, conforme revelado no Alerta Técnico "Risco de falha de cimentação em tamponamento through tubing em poços de alta inclinação (45 graus)". Os principais problemas incluem:<br>- Maior complexidade nas operações de tamponamento devido à descentralização da coluna de produção/injeção dentro do revestimento de produção em poços com alta inclinação. - Dificuldade no assentamento de cement retainer inflável em poços inclinados. - Dificuldade na descida de ferramentas de perfilagem e plugs para isolamento preliminar abaixo do trecho do poço a ser tamponado.<br>Além disso, foi observado que em poços de alta inclinação (> 45 graus), a qualidade da cimentação é, muitas vezes, apenas média, cumprindo apenas os requisitos mínimos e, em alguns casos, levando a falhas na cimentação.<br>No atual cenário, para melhorar os resultados, é sugerido o uso de vazões no deslocamento superiores quando a cimentação é realizada com flexitubo, a realização do uso de dispositivos para vibração da coluna durante a cimentação e o uso de volumes excedentes de pasta compatíveis com uma operação da maior robustez possível.<br>Destaca-se também a importância do controle de circulação e perda de fluido durante a perfuração da fase do reservatório. Em situações de perda de circulação severa, a perfuração com água do mar e tampões viscosos pode ser uma alternativa segura para evitar custos e riscos adicionais.<br>É relevante mencionar também o uso de métodos sustentáveis e eficientes de combate à perda de fluido pós-teste de formação, em que um tampão especial de filtrado relaxado pode ser uma solução viável, trazendo economia e eficiência operacional.<br>No entanto, até que um estudo mais aprofundado sobre as causas desses problemas seja realizado e medidas de melhoria sejam implementadas, recomenda-se a seleção cuidadosa do plano de abandono e a análise prévia de cada cenário para poços de alta inclinação. |
| Cite problemas operacionais ocorridos na sonda NS-48 em 2024, de acordo com o SITOP? Liste todas as ocorrências que encontrar. | 1. Falha no Sistema de Rotação do Top Drive: Houve uma investigação contínua sobre uma falha no sistema de rotação do top drive, que impossibilitou a conexão à coluna. Durante o período de espera, o ponto de acunhamento da cunha foi alterado duas vezes. Fonte: <a href="https://wellsom.petrobras.com.br/sitopi/sonda/NS-48/2024-07-19/M" target="blank">Documento da sonda NS-48 - 2024-07-19 11:50</a><br>2. Condições Ambientais Adversas: Em várias ocasiões, as operações foram interrompidas devido a condições ambientais adversas, como ventos fortes, que impediram a montagem de equipamentos de arame e a retirada de plugues. Fonte: <a href="https://wellsom.petrobras.com.br/sitopi/sonda/NS-48/2024-05-09/M" target="blank">Documento da sonda NS-48 - 2024-05-09 07:30</a><br>3. Atrasos na Instalação de Equipamentos: Houve atrasos na instalação de equipamentos, como a ANM CCB-462, devido à espera por embarcações de apoio, o que impactou o cronograma das operações. Fonte: <a href="https://wellsom.petrobras.com.br/sitopi/sonda/NS-48/2024-05-21/M" target="blank">Documento da sonda NS-48 - 2024-05-21 05:24</a><br>4. Problemas com o WSSV Blue Marlim: A sonda enfrentou problemas operacionais devido à indisponibilidade do WSSV Blue Marlim para operar conectado, causada por condições de mar adversas e problemas no gerador do barco. Fonte: <a href="https://wellsom.petrobras.com.br/sitopi/sonda/NS-48/2024-04-13/T" target="blank">Documento da sonda NS-48 - 2024-04-13 13:29</a><br>5. Falha no sistema de rotação do top drive: Houve uma investigação sobre a falha no sistema de rotação do top drive, que impossibilitou a conexão à coluna. Durante o período de aguardo, o ponto de acunhamento da cunha foi alterado duas vezes. Fonte: <a href="https://wellsom.petrobras.com.br/sitopi/sonda/NS-48/2024-07-19/M" target="blank">NS-48 - 2024-07-19</a><br>6. Aguardando troca de turma: Houve um período de espera para troca de turma do WSSV Blue Marlim, durante o qual foram realizadas operações de mini-TI e recalcamento de fluidos. Fonte: <a href="https://wellsom.petrobras.com.br/sitopi/sonda/NS-48/2024-04-12/M" target="blank">NS-48 - 2024-04-12</a> |

| Question | Ground Truth |
|---|---|
| Em quais poços foi utilizado Well Stroker para abertura e fechamento de VIF, de acordo com o SITOP? Liste todas as ocorrências que encontrar. | Foram encontrados 6 poços com ocorrência de uso de Well Stroker para abertura e fechamento de VIF 1. 8-BUZ-55-RJS 2. 7-BAZ-3-ESS 3. 8-RO-146HP-RJS 4. 7-CRT-52-RJS 5. 8-BUZ-58D-RJS 6. 7-JUB-61D-ESS 7. 8-ATP-7D-RJS 8. 3-RJS-688A |
| Em quais poços haverá cimentação de revestimento entre as próximas operações, de acordo com o SITOP do dia 13/05/2025? Informe somente a lista de nomes dos poços. Liste todos os poços que encontrar, não tente resumir. | Nos próximos dias, haverá operações de cimentação de revestimento nos seguintes poços: 1. Poço 7-BUZ-95-RJS 2. Poço 7-JUB-81H-ESS 3. Poço 7-BUZ-94D-RJS 4. Poço 8-MRO-36-RJS 5. Poço 8-BUZ-101-RJS 6. Poço 9-AB-138D-RJS 7. Poço Buena Suerte-1 8. Poço 7-BUZ-94D-RJS |
| Em qual poço está a sonda NS-39, de acordo com o SITOP? Considere que a data de hoje é 05/01/2025. | A sonda NS-39 está operando no poço 7-MRO-37-RJS. |
| Quais as últimas operações efetuadas no poço 7-MRL-244H-RJS, de acordo com o SITOP do dia 05/01/2025? | Operação Principal: MP: Circulado FPBNA 9,4 ppg via coluna a 8 bpm / 270 psi. MP: Bombeado 50 bbl de colchão espaçador 11,0 ppg via bomba da sonda 6 bpm / 230 psi e deslocado com 10 bbl de FPBNA 9,4 ppg a 6 bpm / 230 psi pela bomba da sonda. MP: Bombeado 96 bbl de pasta de cimento 15,8 ppg a 4 bpm e deslocado com 6,9 bbl de colchão espaçador 11 ppg e com 59 bbl de FPBNA 9,4 ppg. MP: Retirada coluna até 939 m (mais de 200 m acima do TOC), circulado para limpeza da coluna com FPBNA 9,4 ppg e fechado BOP anular. MP: Bombeado via UC 68 bbl de FPBNA 9,4 ppg pela coluna e efetuando squeeze do cimento no overlap liner 10,75" x revestimento 13,626" a 1 bpm/ 260 psi, 2 bpm/ 330 psi, 3 bpm/ 340 psi e a 4 bpm/ 410 psi. MP: Iniciado aguardo de pega de 11 h do cimento às 21:30 h. Operação Paralela: MA: Continuada montagem e estaleiramento do BHA 8 1/2". MA: Iniciada montagem do Djar e de ferramenta de condicionamento do topo do liner. |

69

| Question | Ground Truth |
|---|---|
| Quais foram as operações executadas na NS-42 em 22/06/2024, de acordo com o SITOP? | Descido conjunto FDR/TRT com coluna de DPR 6 5/8" até 936 m. Efetuado dummy run com coluna de DPR e ajustado balanceio. Conectada coflex na SFT. Efetuada plumbagem da SFT. Verticalizado arranjo de superfície com SFF, EBE, links articulados e conectado DPR 6 5/8" da SFT na coluna. Em paralelo, limpado TMF da ANM pelo ROV. Montado BOP de arame sobre a SFT. <br> Finalizada montagem das linhas de superfície e manifold auxiliar. Efetuado teste de estanqueidade das linhas de superfície com 300 psi / 4500 psi por 5 min / 10 min. Efetuado teste de estanqueidade do DPR até as valvulas VG5 e XO da FDR com 3500 psi por 10 min. Hold Point aprovado pelo CSD-EQSB Efetuado flush do DPR e HCR para limpeza. Posicionado FCBA inibido com MEG 10 |
| Qual o SITOP de hoje da NS-45? Considere que a data de hoje é 05/01/2025. | A situação operacional da sonda NS-45 em 05/01/2025 é a seguinte: <br> Operação Principal - **Manutenção do BOP**: As manutenções do BOP estão 63- **Situação Atual**: A operação está aguardando a substituição do conector para 16 3/4" e a troca da gaveta de tubos fixa do BOP, além de manutenções e testes. A previsão de conclusão dessas atividades é para 07/01. - **Próximas Operações**: Prosseguir com a substituição do conector e gaveta de tubos do BOP, manutenções e testes (em andamento). Está previsto descer o BOP em 07/01. <br> Operação Paralela - **Atividades**: Atualmente, não há operações em andamento. Estão sendo programadas ferramentas MLWD para montagem do BHA e fabricado fluido FPBA para perfuração BHD. - **Próximas Operações**: Está prevista a montagem do BHA de perfuração com Broca PDC 8 1/2", RSS, Ferramentas MLWDs e Alargador 9 1/2" para 06/01. |
| Qual sonda completou o poço 7-JUB-62DA-ESS, de acordo com o SITOP? | A sonda que completou o poço 7-JUB-62DA-ESS foi a NS-40 |

## A.2   Evaluation Prompt

```
1  class Confusion_Matrix(TypedDict):  # type: ignore
2      true_positive: list[str]
3      false_positive: list[str]
4      true_negative: list[str]
5      false_negative: list[str]
6
7  def calculate_metrics(llm, question, history):
8      if type(question['Ground Truth']) == str:
9          prompt_confusion_matrix = f""
10             Voce recebera os seguintes parametros:
11             Pergunta: a pergunta do usuario
12             Resposta Ideal: a resposta considerada correta
    por um humano
13             Resposta do sistema: a resposta fornecida pelo
    sistema baseado em IA
14
15             Pegue a resposta do sistema, separe em afirmacoes
     e classifique cada afirmacao entre as opcoes abaixo:
16             True Positive (TP): as afirmacoes corretas feitas
     pelo sistema, ou seja, que estao presentes na resposta
    ideal.
17             False Positive (FP): as afirmacoes incorretas ou
    irrelevantes feitas pelo sistema, ou seja, que nao estao
    presentes na resposta ideal.
18
19             Pegue a resposta ideal, separe em afirmacoes e
    classifique cada afirmacao entre as opcoes abaixo:
20             True Negative (TN): Nao se aplica, deixar vazio.
21             False Negative (FN): as afirmacoes que constam na
    resposta ideal, mas nao foram feitas pelo sistema.
22
23             Importante:
24             - Voce deve gerar listas de afirmacoes para cada
    categoria.
25             - Voce deve quebrar as respostas do sistema e a
    ideal em afirmacoes objetivas.
26             - Se as respostas do sistema ou a ideal
    contiverem frases grandes com mtas afirmacoes, analisar
    cada afirmacao separadamente.
```

```
27              - Uma afirmacao nao pode estar em mais de uma
       categoria.
28              - Ignore coisas na resposta do sistema que nao
       sao afirmacoes objetivas, como por exemplo citacoes de
       fontes e links.
29
30              Vamos la!
31
32              ############
33
34              Pergunta: {{{{
35              {question['Question']}
36              }}}}
37
38              Resposta ideal: {{{{
39              {question['Ground Truth']}
40              }}}}
41
42              Resposta do sistema: {{{{
43              {history[-1].content}
44              }}}}"
45     response = llm.with_structured_output(
       Confusion_Matrix).invoke(prompt_confusion_matrix)
46
47     true_positive_count = len(response.get('true_positive
       ', []))
48     false_positive_count = len(response.get('
       false_positive', []))
49     true_negative_count = len(response.get('true_negative
       ', []))
50     false_negative_count = len(response.get('
       false_negative', []))
51
52     try:
53         precision = true_positive_count / (
       true_positive_count + false_positive_count)
54     except ZeroDivisionError:
55         precision = 0
56
57     try:
58         recall = true_positive_count / (
```

```
                true_positive_count + false_negative_count)
59          except ZeroDivisionError:
60              recall = 0
61
62          try:
63              f1_score = 2 * (precision * recall) / (precision
    + recall)
64          except ZeroDivisionError:
65              f1_score = 0
66
67          print("\\n\\nPrecision: "+str(precision))
68          print("Recall: "+str(recall))
69          print("F1 Score: "+str(f1_score))
70
71          print("\\nAnswer Size Ratio: "+str(question['Answer
    Size (% of GT)']))
72
73          state.precision = precision
74          state.recall = recall
75          state.f1_score = f1_score
76          state.ground_truth = question['Ground Truth']
77          state.statements = response
```

Listing A.1: Código para LLM-as-a-Judge

## A.3  Results

**Best Precision by Model and Configuration**
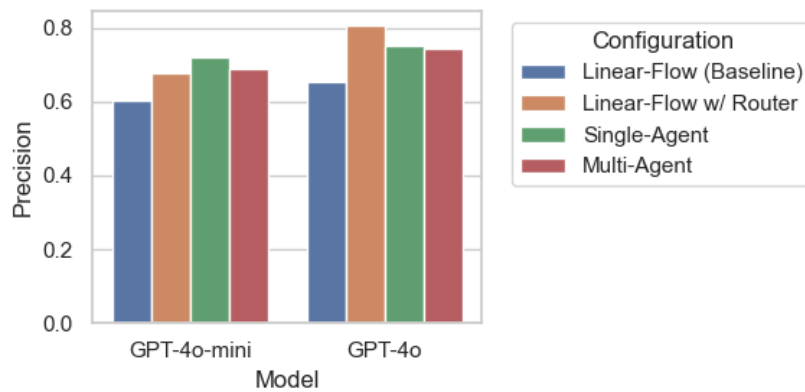


Figure A.1: Best precision by model and configuration.

**Best Precision by Question Index and Configuration**
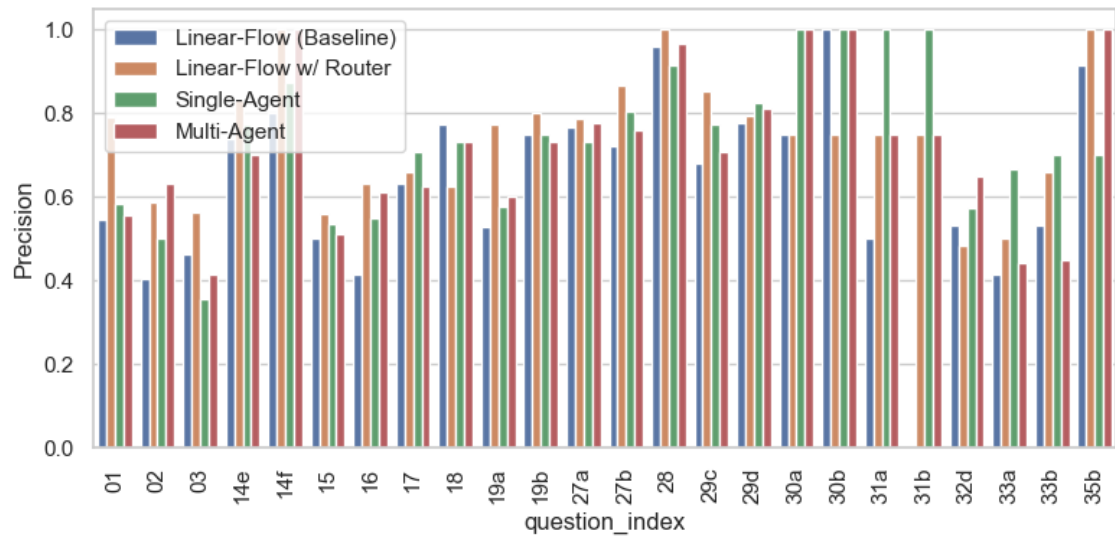


Figure A.2: Best precision by question index and configuration.

**Best Precision by Question Index and Model**
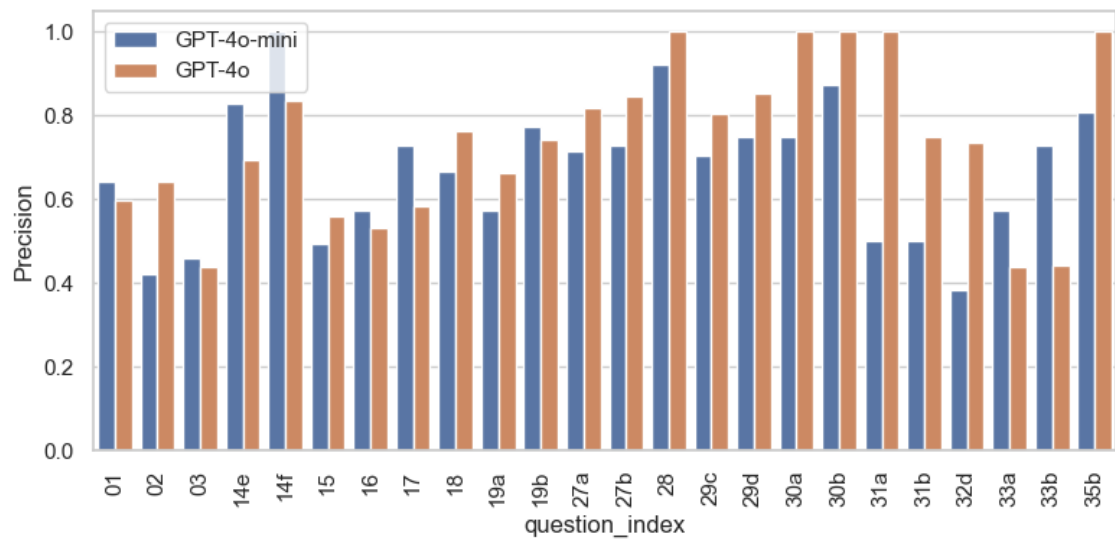


Figure A.3: Best precision by question index and model.

**Facet Histogram of Precision by Model**
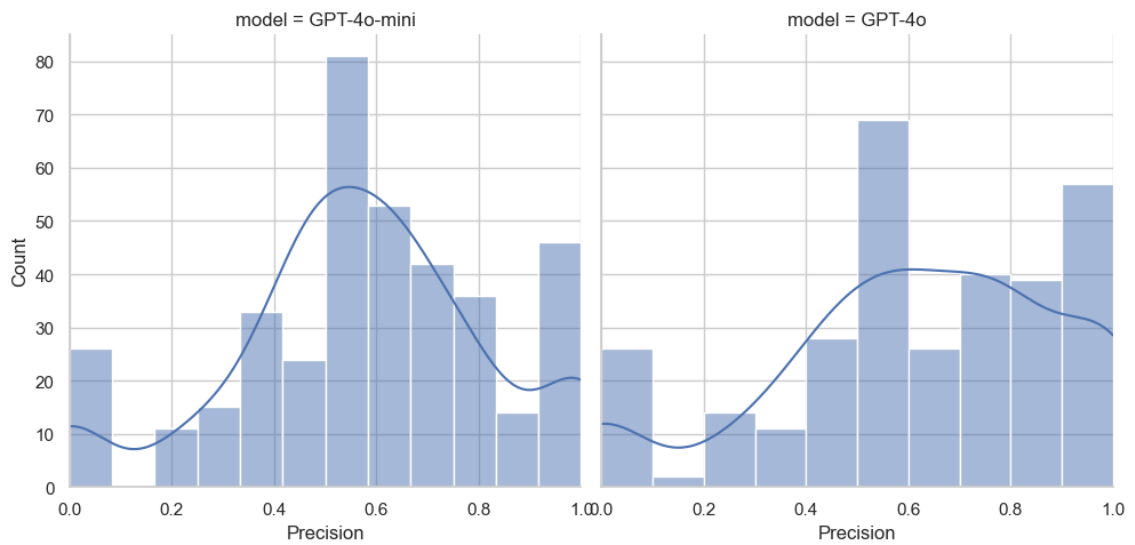


Figure A.4: Facet histogram of precision by model.

**Facet Histogram of Precision by Model (best of 3)**
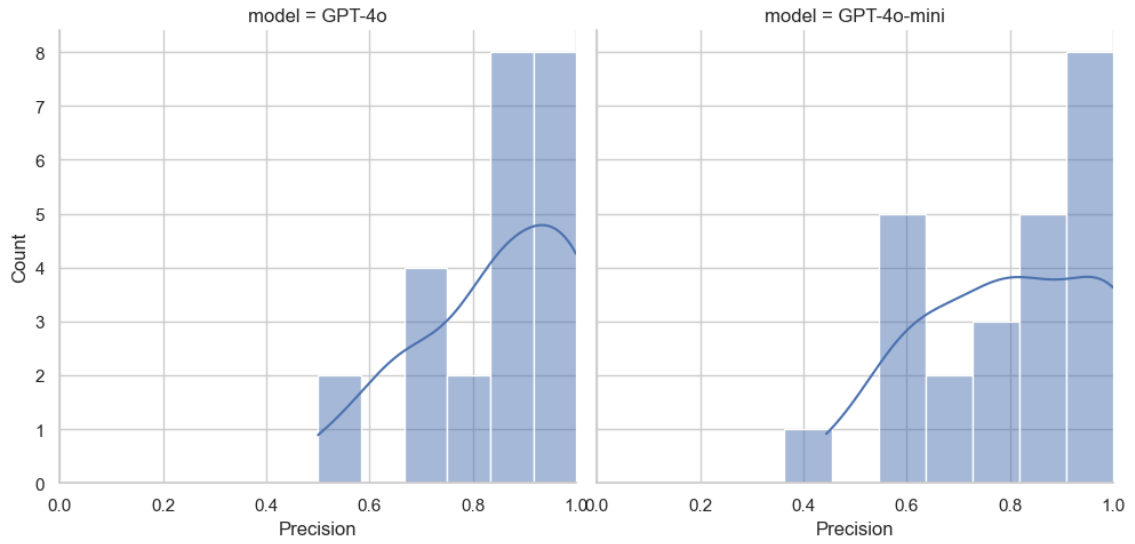


Figure A.5: Facet histogram of precision by model (best of 3).

**Histogram of All Precisions**
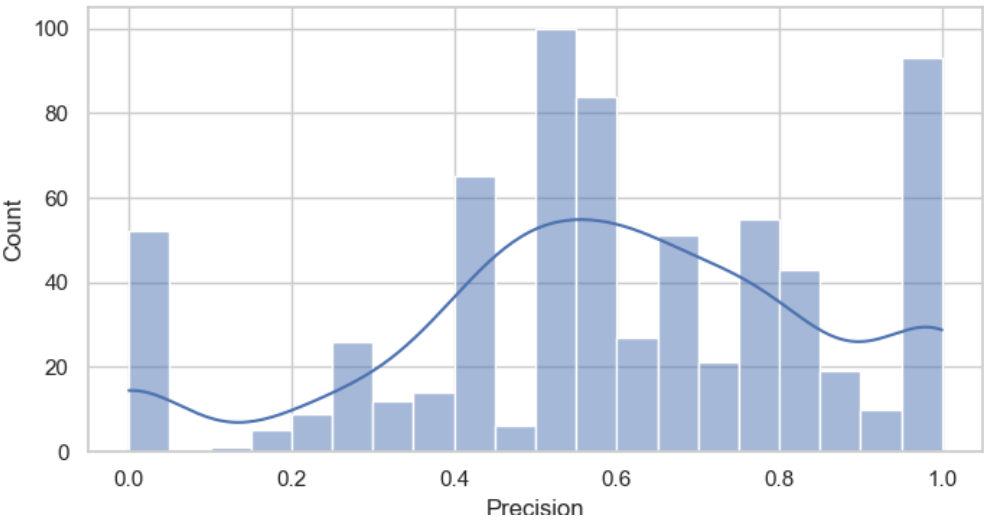


Figure A.6: Histogram of all precisions.

**Line Plot of Precision by Question Index and Model**
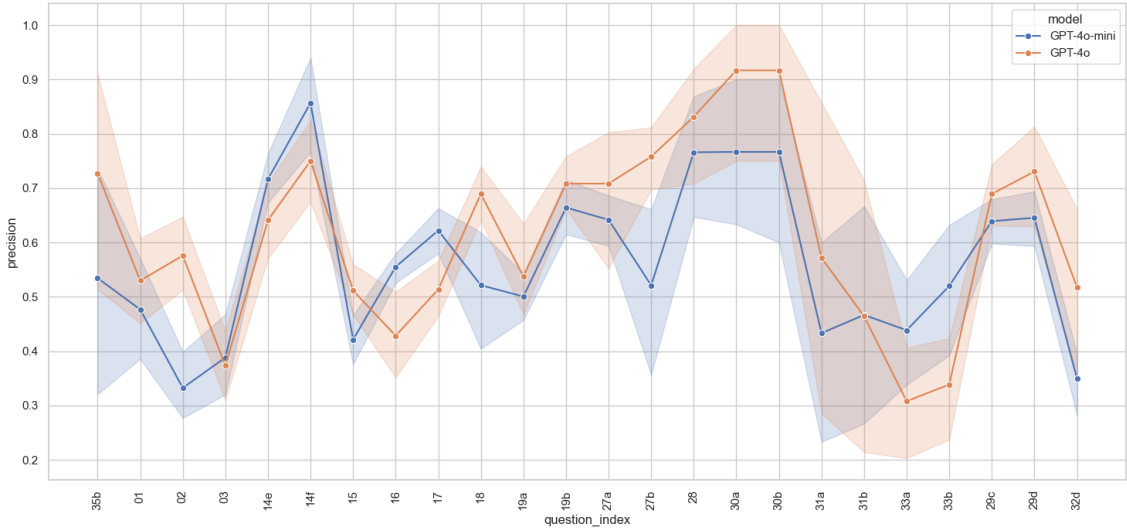


Figure A.7: Line plot of precision by question index and model.

**Boxplot of Precision by Model and Configuration**
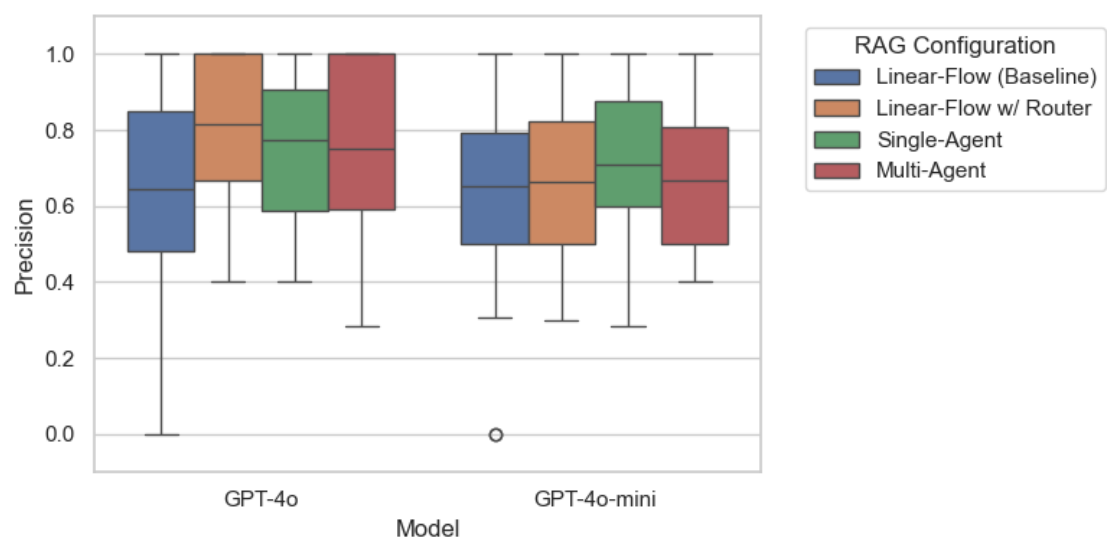


Figure A.8: Boxplot of precision by model and configuration.

**Precision by Model**
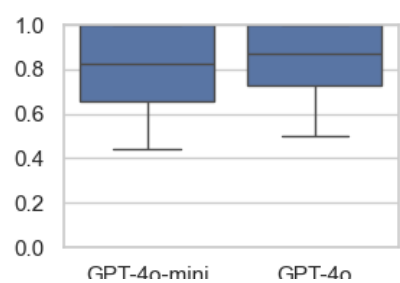


Figure A.9: Precision by model.

**Precision by Model and Configuration**
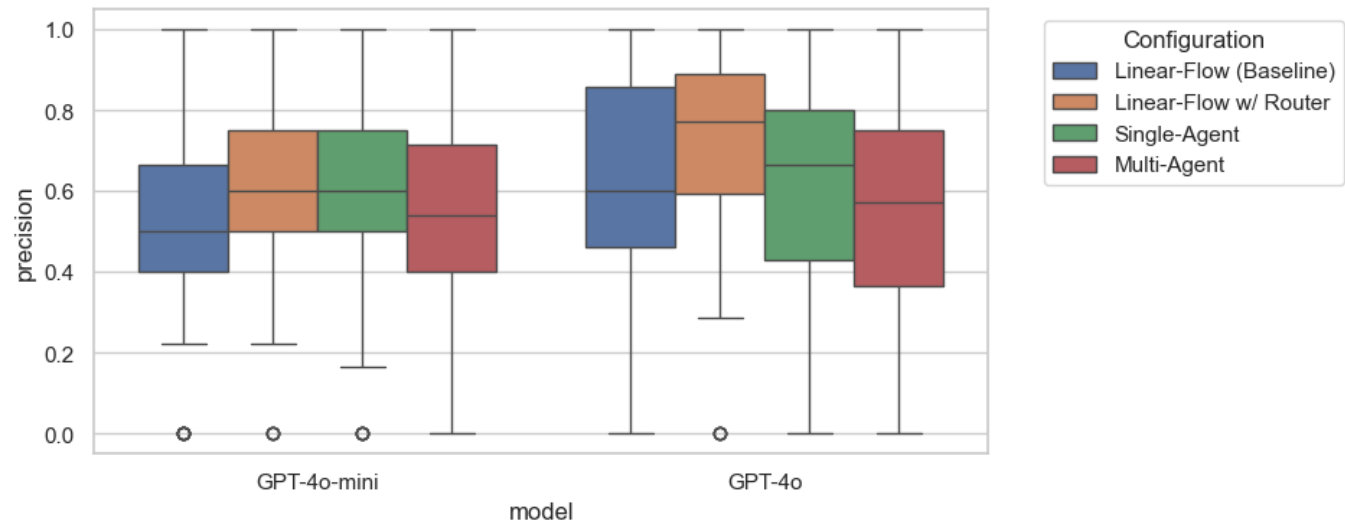


Figure A.10: Precision by model and configuration.

**Line Plot of Precision by Question Index and Configuration**
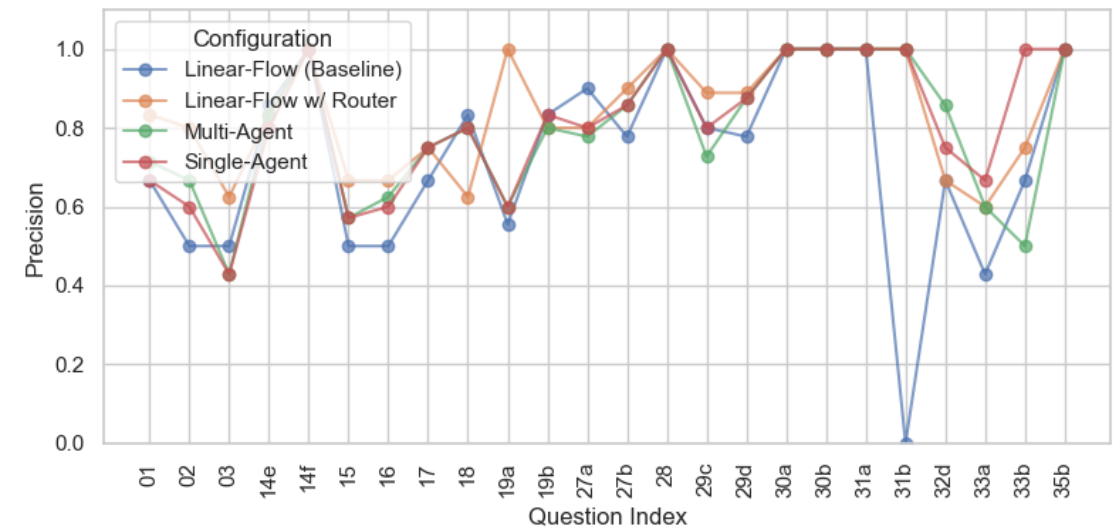


Figure A.11: Line plot of precision by question index and configuration.

**Scatter Plot of Precision vs. Total Time**
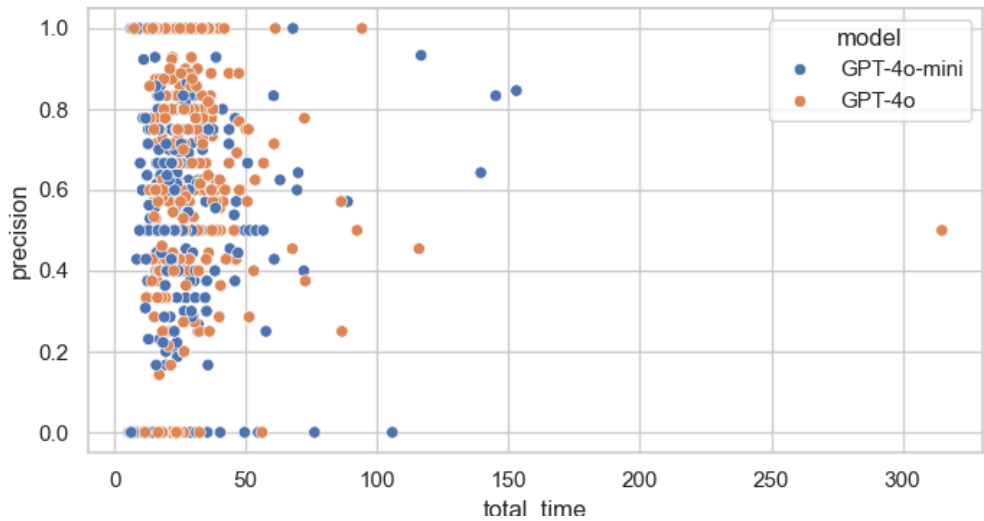


Figure A.12: Scatter plot of precision vs. total time.

**Scatter Plot of Precision vs. Total Token Count Input**
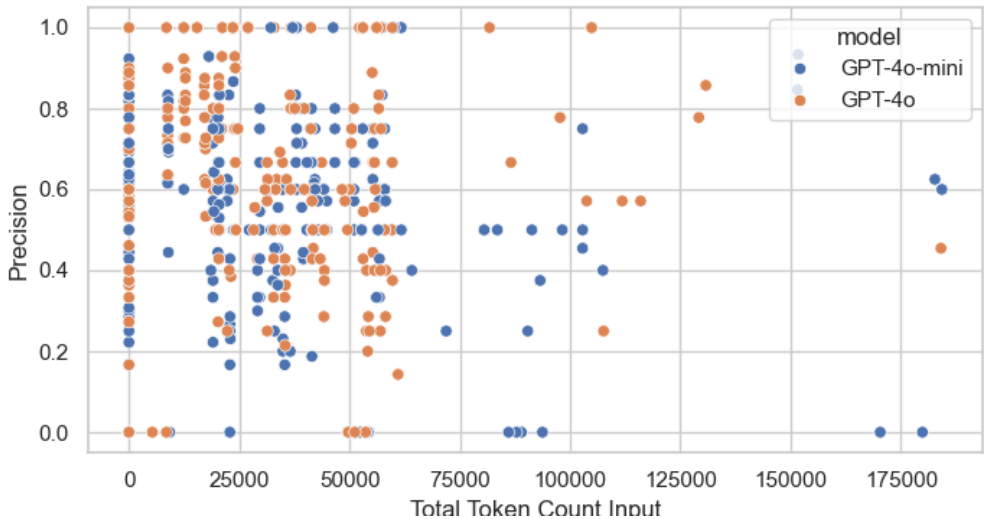


Figure A.13: Scatter plot of precision vs. total token count input.

**Swarm Plot of Precision by Model and Configuration**
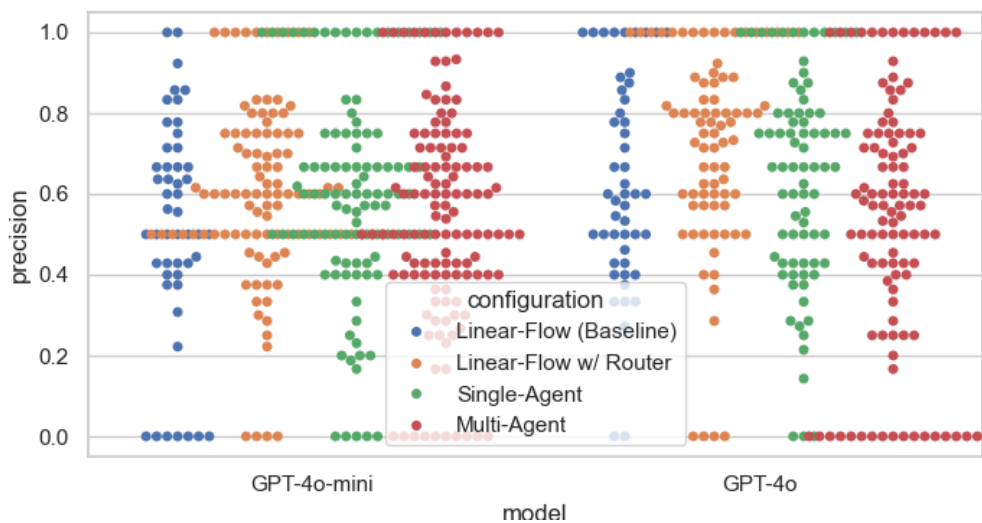


Figure A.14: Swarm plot of precision by model and configuration.

**Violin Plot of Precision by Model and Configuration**
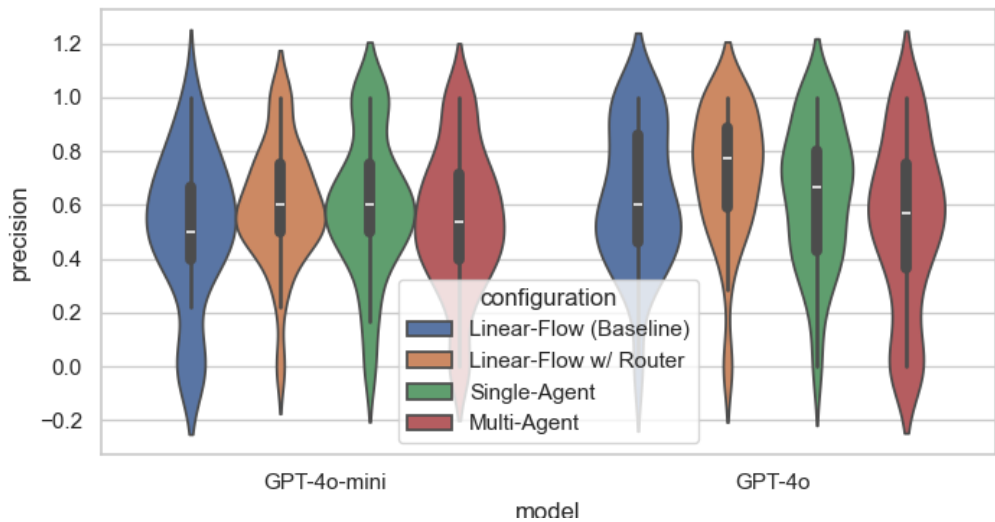


Figure A.15: Violin plot of precision by model and configuration.

# List of Comments