Instituto Alberto Luiz Coimbra de Pós-Graduação e Pesquisa de Engenharia
COPPE UFRJ

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

Brandão Sabbagh, Vitor

Comparative Analysis of Single and Multi-Agent Large Language Model Architectures for Domain-Specific Tasks in Well Construction/Vitor Brandão Sabbagh. – Rio de Janeiro: UFRJ/COPPE, 2025.

XII, 84 p.: il.; 29, 7cm.

Orientador: Geraldo Bonorino Xexéo

Dissertação (mestrado) – UFRJ/COPPE/Programa de Engenharia de Sistemas e Computação, 2025.

Referências Bibliográficas: p. 56 – 62.

1. Large Language Models. 2. Agents. 3. Oil Well Construction. I. Bonorino Xexéo, Geraldo. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III. Título.

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. X:1 V:2

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] X:3 V: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

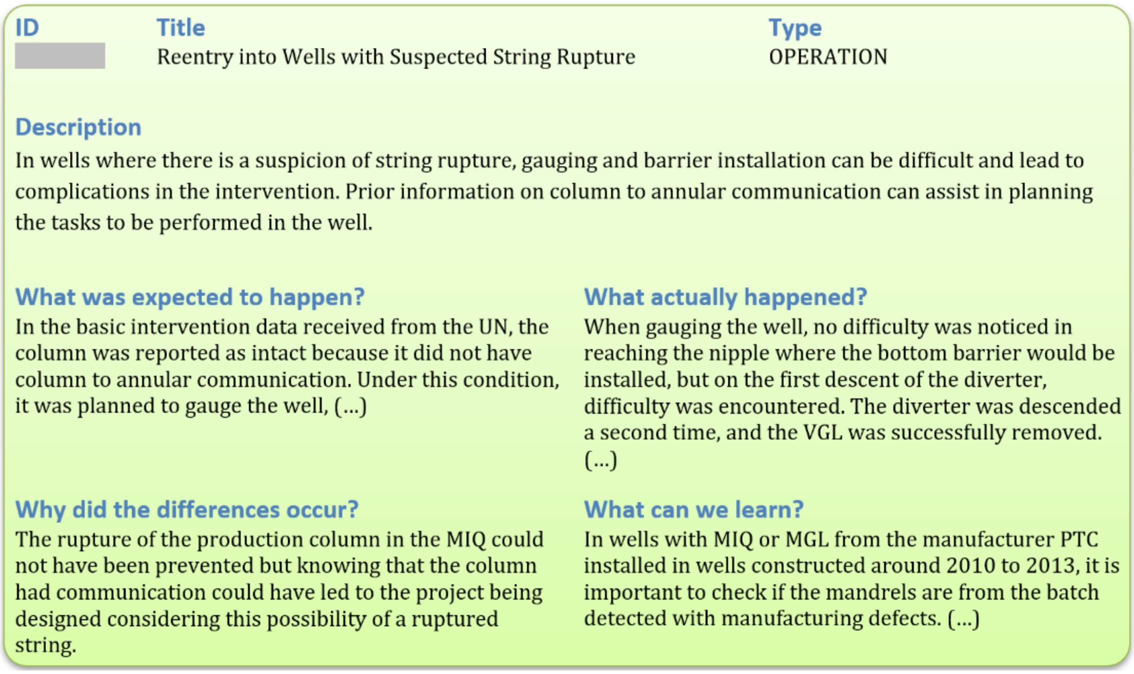
et al., 2023) highlights that Gen-AI is poised to increase global GDP by nearly 7%, increasing productivity growth by 1.5 percentage points over the next decade. This economic uplift is expected due to AI's ability to automate complex workflows and create new business opportunities, significantly impacting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


ID

Title
Reentry into Wells with Suspected String Rupture

Type
OPERATION

Description
In wells where there is a suspicion of string rupture, gauging and barrier installation can be difficult and lead to complications in the intervention. Prior information on column to annular communication can assist in planning the tasks to be performed in the well.

What was expected to happen?
In the basic intervention data received from the UN, the column was reported as intact because it did not have column to annular communication. Under this condition, it was planned to gauge the well, (...)

What actually happened?
When gauging the well, no difficulty was noticed in reaching the nipple where the bottom barrier would be installed, but on the first descent of the diverter, difficulty was encountered. The diverter was descended a second time, and the VGL was successfully removed. (...)

Why did the differences occur?
The rupture of the production column in the MIQ could not have been prevented but knowing that the column had communication could have led to the project being designed considering this possibility of a ruptured string.

What can we learn?
In wells with MIQ or MGL from the manufacturer PTC installed in wells constructed around 2010 to 2013, it is important to check if the mandrels are from the batch detected with manufacturing defects. (...)

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 Knowledge Items 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

««PENDENTE»» V:6

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 RESPONDER A ESTAS PERGUNTAS (QUESTÃO 1, QUESTÃO 2...), EU CONDUZI ESTES EXPERIMENTOS (EXPERIMENTO 1, EXPERIMENTO 2), QUE ME LEVARAM A ESTAS DESCOBERTAS PRINCIPAIS (RESULTADO A, RESULTADO 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).

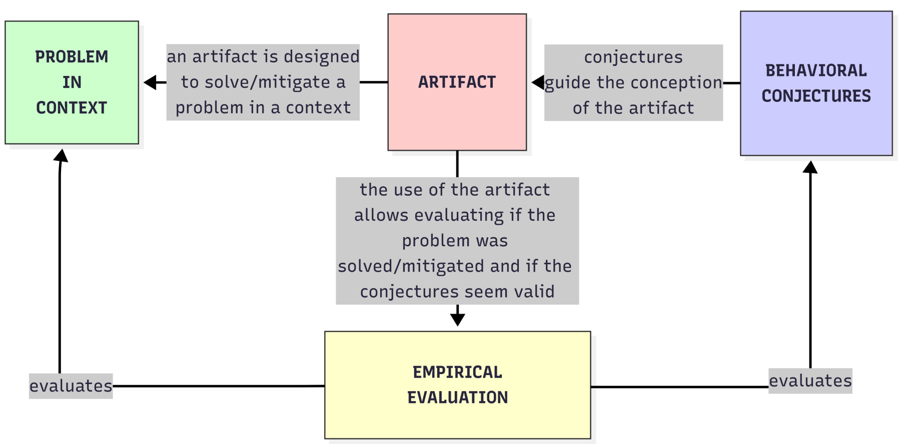


Figure 1.2: Main elements of DSR-Model, translated from OSWALD *et al.* (2023).

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 linguistic 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

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 V:10

However, the trajectory of LLM development in 2025 has signaled a shift in focus. 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

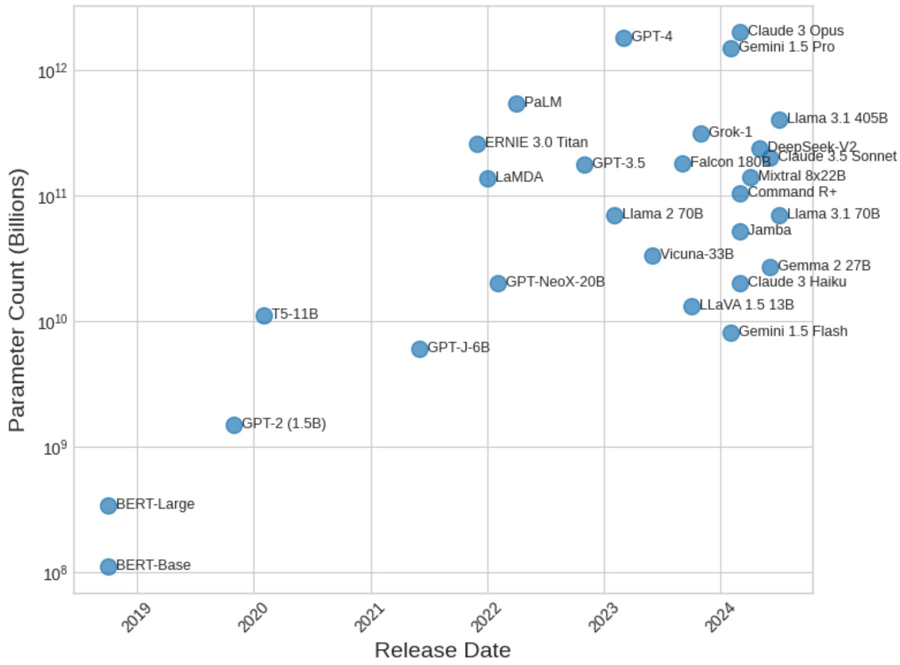


Figure 2.1: The evolution of LLMs.

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

The proliferation of LLMs has led to a diverse array of applications that leverage their ability to understand, generate, and process human language. V:12

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

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

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 information, overcoming the inherent limitations of LLMs that rely solely on the prior knowledge embedded in the language model during the training phase. With

X:13
V:14

[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.

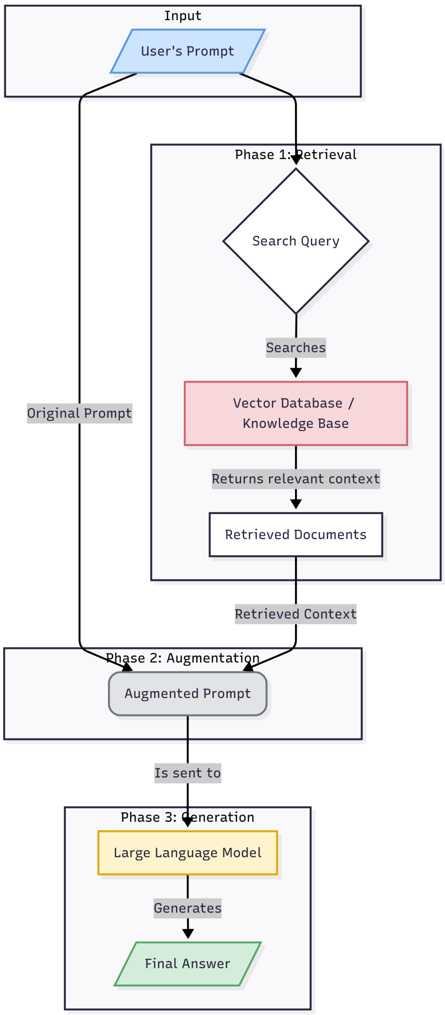


Figure 2.2: A diagram illustrating the RAG process.

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] X:15 V: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] X:17 V: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] X:19 V:20

[19] X:Você usou no prmeiro experimento outras métricas, tem que descreve-las aqui: Truthfulness, Performance, LLMCost

[20] V:Feito

Chapter 3

First Experimental Evaluation Cycle

↯[21] ↯[22] X:21

↯[23] ↯[24] V:22

↯[25] ↯[26] X:23

↯[27] ↯[28] V:24

X:25

This chapter describes the first experimental cycle of this research, as introduced V:26
in Section 1.3.2, conducted to investigate the effectiveness of different LLM based X:27
agent architectures. The primary objective is to address complex, domain-specific V:28
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.

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

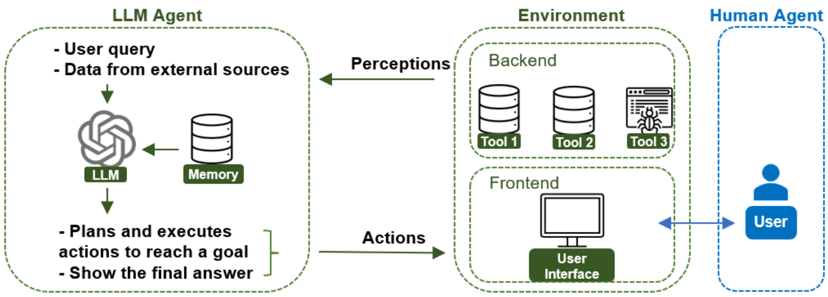


Figure 3.1: Schematic of the LLM-based agent interacting with an environment containing tools for task-specific operations, and the Human Agent interface for user interaction and feedback.

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:

- **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.

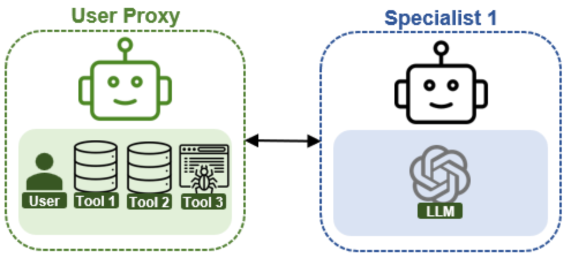


Figure 3.2: Chat setup with one User Proxy (WU *et al.*, 2023) and one Assistant.

- **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.

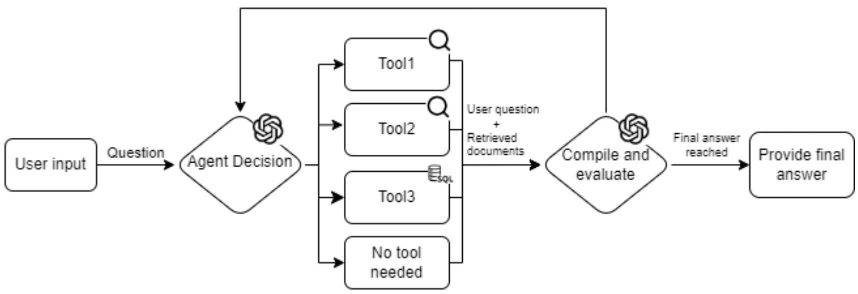


Figure 3.3: Decision process of the agent.

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. 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.

3.3.3 Agent's Tools

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

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

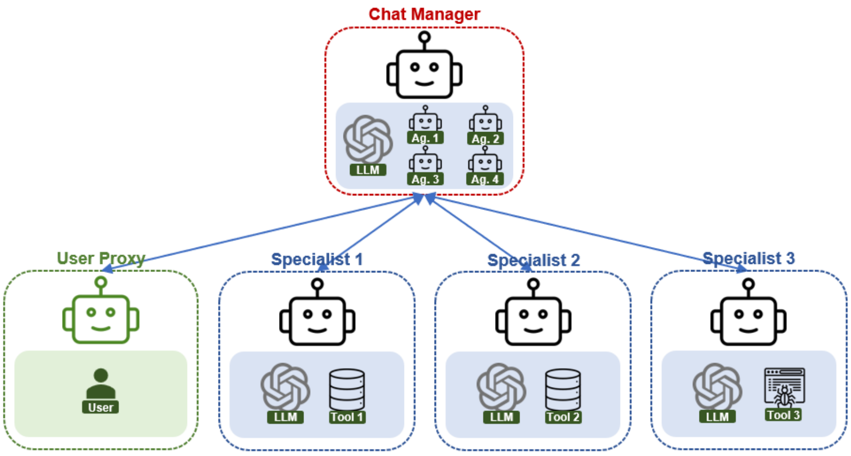


Figure 3.4: Chat setup with one Chat Manager and a group of LLM agents.

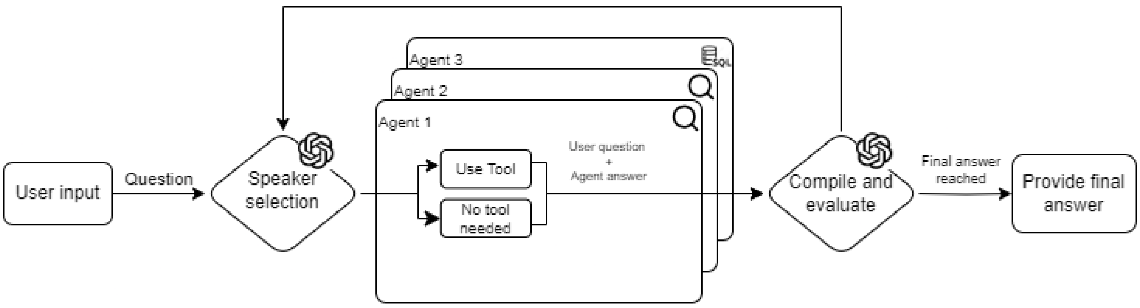


Figure 3.5: Multi-agent decision process.

- **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.

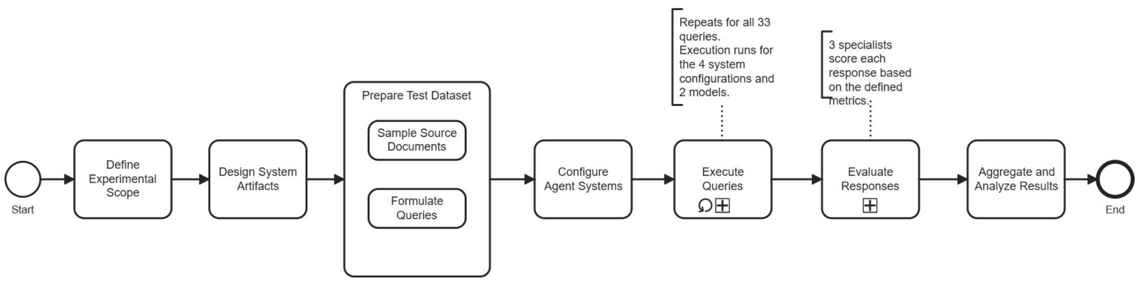


Figure 3.6: Experimental workflow.

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 Knowledge Items (lessons learned, alerts), a structured database of Non-Productive Time (NPT) incidents, and a Collaborator Finder tool, as described in Section 1.1.1.

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

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

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.

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

[32] V:Feito

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.
- **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	Perfor- mance	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

[38] [39] [40] [41] [42] [43] [44] X:38

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. V:39 X:40 V:41 X:42 V:43 X:44

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

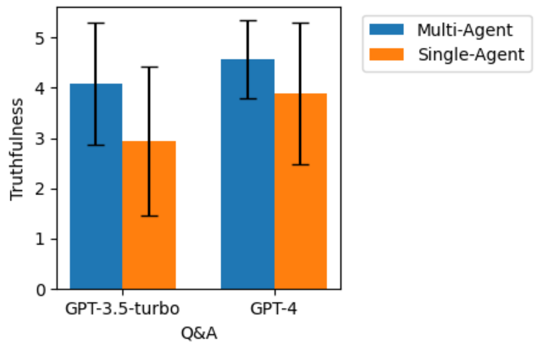


Figure 3.7: Truthfulness and standard deviation in Q&A tasks by LLM model and agent configuration.

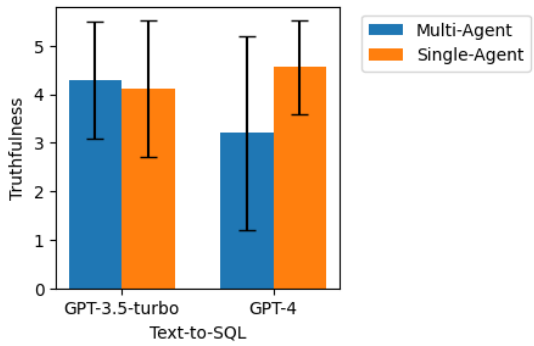


Figure 3.8: Truthfulness and standard deviation in Text-to-SQL tasks by LLM model and agent configuration.

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

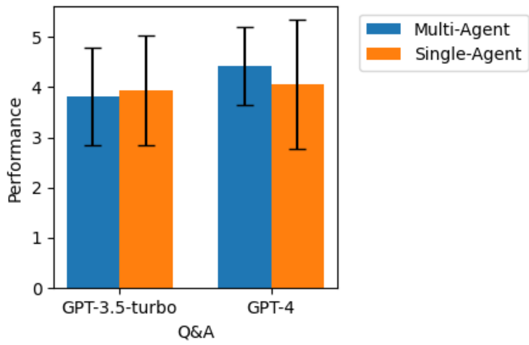


Figure 3.10: Performance and standard deviation in Text-to-SQL tasks by LLM model and agent configuration.

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.

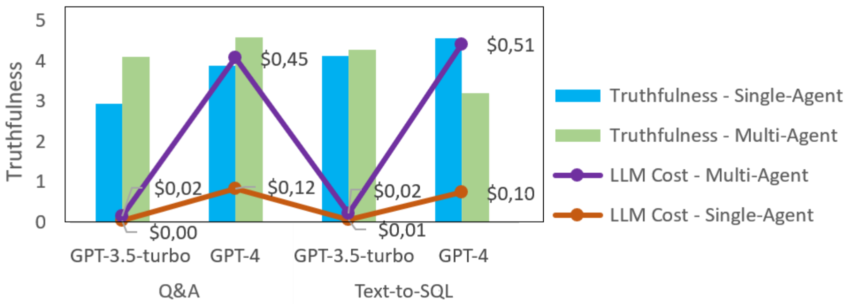


Figure 3.11: Average LLM costs and Truthfulness per completed task according to setup and model.

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 V:46

This trade-off highlights an important implication for any organization considering 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, 2025). ⚡[47] ⚡[48] This compartmentalization can lead to better handling of complex, multi-faceted queries, as each agent can focus on its specific 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. X:47 V:48

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 V:50

The average cost per query for each configuration is presented in Table 3.4. These 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. ⚡[51] ⚡[52] X:51 V: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 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

[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:Vou comentar isso na conclusão

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 knowledge items to the wrong author[53][54]. This highlights the difficulty in accurately retrieving information based on proper names, which can be complicated by variations in accentuation, abbreviation, and formatting. [55] [56] A potential solution to be explored is the use of Self-Query Retriever (LANGCHAIN, 2023), implementing 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. X:53 V:54 X:55 V:56

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 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

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

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 Experiment

↯[57] X:57

This chapter describes the second experiment used to evaluate LLM agents and workflows to answer questions in the well engineering domain. The experiment integrates multiple LLM configurations, agent architectures, and RAG tools, leveraging Petrobras↯[58] datasets. X:58

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, revealing key insights into their performance, cost, and limitations such as hallucination and context interpretation. ↯[59] The rapid evolution of generative AI frameworks and models prompted X:59 a second, more up-to-date experiment in 2025. This second phase built upon the initial findings, also employing non-agentic workflows as a 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 referred to as "LLM-as-a-judge" (GU *et al.*, 2025).

The use of "LLM-as-a-judge" was driven by the sheer volume of responses to be evaluated. With four configurations, two models, and three executions for each, a total of 24 responses were generated for every question in the dataset. Manually assessing this volume of data would have been impractical. Furthermore, previously used metrics like 'truthfulness' had become obsolete. This metric was highly relevant when models frequently hallucinated, a problem that is far less prevalent in the current generation of LLMs.

[57] X:Tem muito LLMnismos aqui, como adjetivos desnecessários propagandísticos. Tá muito genérico, falta dados e deixa muita pergunta. Parece mais um plano do que algo feito. Você já fez e falta escrever?

[58] X:OPA, escapou um Petrobras aqui

[59] X:não é o capítulo anterior? Acho que pode tirar

4.1 Methodology

4.1.1 Experimental Workflow

Dataset Preparation

The experimental workflow was designed to provide a thorough and reproducible evaluation of language model agents within oil well operations. The process begins with the careful preparation of the dataset, which is composed of questions and corresponding ground truth answers derived from a diverse range of operational records, incident reports, and lessons learned. To ensure the quality and relevance of the data, questions undergo a filtering and preprocessing phase where clarity, diversity, and alignment with real-world scenarios are prioritized. This includes removing duplicates, standardizing terminology, and confirming that each question is properly paired with an accurate answer. The dataset is further validated for completeness and consistency, ensuring it represents the full spectrum of operational challenges, such as safety, cementing, and intervention scenarios.

Model and Setup Selection

Following dataset preparation, the experimental design incorporates a variety of agent architectures. These include approaches where questions are routed to specialized agents, single-agent systems that centralize all reasoning and retrieval, and multi-agent frameworks that leverage collaboration among specialized agents under a supervisory structure. Each of these configurations is evaluated using different language models, allowing for a comprehensive assessment of how model choice and agent setup influence performance. The agents are also provided with access to advanced retrieval tools and domain-specific knowledge bases, enabling them to draw on a broad foundation of operational expertise.

Execution Loop

The core of the experimental workflow is an execution loop depicted in Algorithm 1. For each combination of question, agent setup, and language model, the system systematically loads the relevant data, configures the agent, and executes the workflow. Throughout this process, all responses and intermediate reasoning steps are meticulously logged. This approach not only ensures systematic coverage of all experimental conditions but also provides full traceability for subsequent analysis. The automation of these procedures guarantees consistency and reproducibility, while the comprehensive logging facilitates in-depth evaluation and comparison of agent performance across a range of operational scenarios.

Algorithm 1 Experiment Execution Loop

```
Require: questions, setups, models
Ensure: results
function RunExperiment
    results ← {}
    for all question ∈ questions do
        ground_truth ← question.ground_truth
        for all setup ∈ setups do
            for all model ∈ models do
                agent ← InitializeAgent(setup, model)
                response ← agent.ProcessQuestion(question)
                metrics ← EvaluateResponse(response, ground_truth)
                results[question, setup, model] ← {
                        "response" : response,
                        "metrics" : metrics,
                        "execution_trace" : agent.trace
                }
            end for
        end for
    end for
    return AggregateResults(results)
end function
```

Evaluation and Metrics

Following the execution of all experimental combinations, a comprehensive evaluation framework is applied to assess agent performance. The system calculates a suite of quantitative metrics for each question, setup, and model combination by comparing the generated answers against the established ground truth. These metrics include standard performance indicators such as accuracy, precision, recall, and F1 score, which provide a multifaceted view of response quality. For open-ended questions where binary correctness measures are insufficient, a confusion matrix approach is implemented to capture nuances in answer quality and content coverage. Additionally, the system measures answer size ratio relative to ground truth, offering insights into model verbosity and conciseness. These metrics are then aggregated across different dimensions to enable meaningful comparisons between agent architectures and language models, revealing patterns in performance across various operational scenarios and question types.

Reproducibility and Quality Control

To ensure scientific rigor and reproducibility, the experimental methodology incorporates robust tracking of all environmental variables and configuration parameters. The system maintains detailed logs of the computational environment, including software versions, dependency specifications, and hardware characteristics that might influence results. All experimental parameters, from model identifiers to dataset specifications, are systematically recorded alongside the results they generate. Throughout the experimental process, periodic validation checks are performed to maintain data integrity and result consistency, with anomalies flagged for investigation. This comprehensive approach to reproducibility not only facilitates verification of findings but also enables future extensions of the research with comparable baselines. The quality control measures embedded in the workflow ensure that conclusions drawn from the experiments rest on a foundation of methodological soundness and data reliability.

<insert workflow diagram or pseudocode here to illustrate the above stages>[60] X:60

4.1.2 Data Sources

The experimental evaluation relies on a curated collection of data sources that represent the diverse knowledge domains relevant to oil well operations. At the core of the experiment is a comprehensive questions dataset containing structured entries that simulate real-world queries an operator might encounter. This dataset was developed through extensive collaboration with domain experts and analysis of historical operational records. Each entry in the dataset contains a question formulated in natural language, a unique identifier, categorical metadata to facilitate analysis, and a corresponding ground truth answer validated by subject matter experts. The questions span various complexity levels, from factual inquiries to complex reasoning scenarios that require integration of multiple knowledge sources.

To provide the language models with the necessary domain knowledge, the experiment incorporates several specialized knowledge bases that reflect different aspects of oil well operations:

- **Knowledge Bases and Tools**:
  - **Lessons**: A repository of knowledge items capturing insights, best practices, and technical know-how from past oil well operations. These lessons represent institutional memory and expertise accumulated over years of operational experience.
  - **Alertas SMS**: A collection of safety alerts and incident reports documenting past events, near-misses, and accidents, providing critical safety information and preventative measures.

[60] X:OPA!

– **SITOP**: Detailed daily operational logs from drilling rigs, containing technical parameters, operational decisions, and situational reports from active drilling operations.

These knowledge sources were preprocessed to ensure consistency, remove sensitive information, and optimize retrieval performance. The integration of these diverse data sources enables a holistic evaluation of how language model agents navigate the complex informational landscape of oil well operations, from technical specifications to safety protocols and historical precedents.

<insert table or image summarizing datasets and tools here>

4.1.3 System Architecture

The experimental system was implemented using modern Python frameworks specialized for language model orchestration and agent workflows. The architecture leverages the LangChain and LangGraph ecosystems, which provide robust foundations for building complex language model applications with multiple components and state management. This subsection details the modular design of the system, highlighting how different components interact to enable systematic evaluation of language model agents in oil well operations.

Experiment Orchestration

At the core of the system architecture is an experiment orchestration layer responsible for coordinating the entire evaluation process. This component manages the loading of questions from the dataset, systematically iterates through different model and setup combinations, and ensures proper logging of results. The orchestrator maintains experiment state across multiple runs, handles error recovery, and implements checkpointing to allow for resumption of long-running experiments. By centralizing control flow, this component ensures that all experimental conditions are tested consistently and that results are captured in a standardized format for subsequent analysis.

Agent Workflow Frameworks

The system implements multiple agent workflow frameworks to evaluate different approaches to question answering in the oil well domain. These frameworks define the flow of information and decision-making processes within and between language model agents. The implemented workflows include a Linear-Flow with Router (CORTEX) that directs questions to specialized processing paths, a Single-Agent approach that centralizes all reasoning and tool use, and a Multi-Agent Supervisor framework that coordinates multiple specialized agents. Each workflow is defined declaratively, specifying the sequence of operations, decision points, and information exchange patterns that govern agent behavior during question processing.

Nodes and Tool Integration

The system architecture includes specialized nodes that implement specific reasoning steps and tool-calling logic. These nodes serve as the building blocks of agent workflows, encapsulating discrete functionality such as question analysis, knowledge retrieval, and answer synthesis. The tool integration layer provides agents with access to external knowledge sources through a standardized interface, enabling semantic search over domain-specific corpora, structured data queries, and other specialized operations. This modular approach to tool integration allows for consistent evaluation of how different agent architectures leverage available tools and knowledge sources.

Prompt Engineering and System Messages

A critical component of the architecture is the prompt engineering layer, which defines the instructions and context provided to language models. This includes carefully crafted system messages that establish the role and capabilities of each agent, prompt templates that structure inputs consistently across experimental conditions, and few-shot examples that guide model behavior. The system maintains a library of prompt variants optimized for different tasks within the question-answering workflow, ensuring that each agent receives appropriate guidance while maintaining experimental control.

State Management and Metrics

The architecture incorporates a comprehensive state management system that tracks the progress of experiments, maintains contextual information across agent interactions, and captures intermediate reasoning steps. This component is tightly integrated with the metrics calculation subsystem, which computes performance indicators in real-time as experiments progress. The metrics framework implements various evaluation approaches, from simple accuracy measures to sophisticated semantic similarity calculations, providing multi-dimensional assessment of agent performance. All experimental data, including intermediate states and final results, is persisted in structured formats to enable both immediate feedback and in-depth post-experiment analysis.

<insert system architecture diagram here>

4.1.4 Experimental Setups

To comprehensively evaluate language model performance in well construction operations, the experiment employed multiple agent architectures and model configurations. This subsection details the different experimental setups, highlighting their design principles, operational characteristics, and the rationale behind their selection. The experimental design deliberately incorporates contrasting approaches to agent architecture, enabling comparative analysis of different strategies for complex question answering in specialized domains.

Linear-Flow

The Linear-Flow architecture represents the simplest RAG design, where user input is processed in a strictly sequential manner. In this setup, the user's query is handled by a single LLM step, which carries all the instructions (PT1, PT2, PT3 and PT4, as depicted in 4.1) required for the generation of various types of search queries. These instruction prompts are often quite long, as they are carefully crafted to produce high-quality queries for the vector store. Due to the aggregation of all instruction prompts within a single LLM invocation, the resulting context becomes notably extensive. This can lead to performance degradation as the context length increases [O QUE PODE SER VISTO NO GRAFICO TAL EM CONTRASTE COM O SETUP TAL QUE DIVIDE OS PROMPTS EM PARTES].

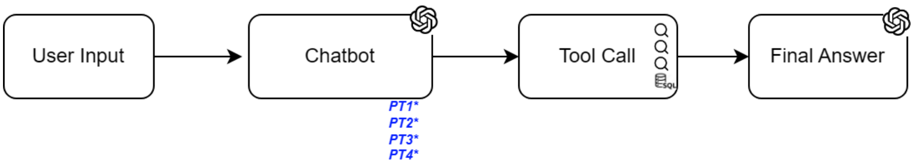


Figure 4.1: Linear-Flow architecture. PT1 indicates Prompt for Tool 1 and so on.

Linear-Flow with Router

[VERSÃO 1] The Linear-Flow with Router architecture implements a sequential processing pipeline with routing, in order to divide and reduce tool instruction prompts. In this setup, questions first pass through a router that analyzes the query content and determines the most appropriate specialized node for the user query at hand.

[VERSÃO 2] The Linear-Flow with Router paradigm extends the basic linear flow by introducing a routing mechanism that enables the distribution of tool instruction prompts into specialized nodes with smaller prompts. User questions first pass through a router that determines the most appropriate specialized node. As illustrated in Figure 4.2, instead of a single node generating different types of retrieval queries, we have several nodes (or sub-queries), each with its own specialized tool. Each sub-query is then processed independently by separate tool invocations.

This approach offers several advantages:

- **Increased Throughput:** By distributing sub-tasks across multiple tools, the system can handle more complex or multi-faceted user requests efficiently.
- **Specialization:** Each tool can be tailored to address a specific aspect of the user's query, allowing for more accurate and relevant results.
- **Scalability:** The architecture naturally supports scaling, as additional tools can be added to handle more sub-queries or specialized tasks.

In practice, the router acts as an orchestrator, analyzing the user input and generating multiple targeted queries (PT1*, PT2*, PT3*, PT4* in the figure). These queries are dispatched to their respective tools, and the results are aggregated to form the final answer. This method is particularly effective for tasks that can be decomposed into independent components, such as multi-part questions or workflows requiring different types of expertise.

Compared to the standard linear flow, the use of a router introduces additional complexity in query generation and result aggregation but enables a significant boost in system flexibility and performance.

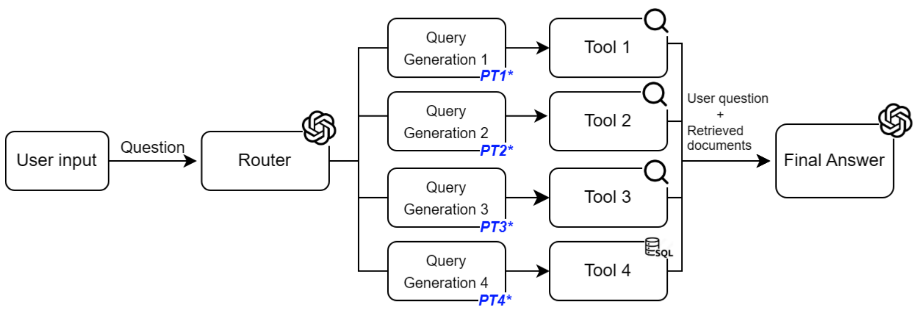


Figure 4.2: Linear-Flow with Router architecture.

Single-Agent

The Single-Agent approach represents a centralized architecture where a single language model agent handles the entire question-answering process. This agent has access to the full suite of retrieval tools and knowledge sources, making independent decisions about which tools to invoke and how to synthesize information into coherent answers. The design emphasizes end-to-end reasoning within a unified context, allowing the model to maintain a consistent understanding throughout the process.

This approach tests the capability of language models to manage complex workflows autonomously, balancing between exploration of different knowledge sources and focused answer generation without the overhead of inter-agent communication.

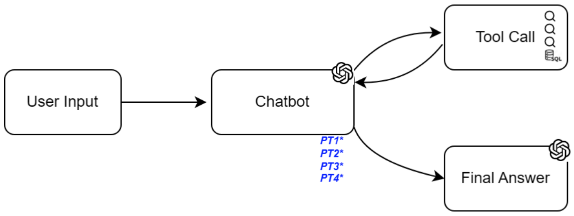


Figure 4.3: Single-Agent architecture

Multi-Agent Supervisor

The Multi-Agent Supervisor setup implements a collaborative approach where multiple specialized agents work together under the coordination of a supervisor agent. Each specialized agent focuses on a specific domain of knowledge or reasoning skill, such as retrieval, analysis, or explanation generation. The supervisor agent orchestrates the collaboration, delegating subtasks to appropriate specialized agents, integrating their contributions, and ensuring coherence in the final answer. This architecture explores the potential benefits of distributed cognition, where complex reasoning is decomposed into manageable components handled by purpose-built agents. The framework includes mechanisms for resolving conflicts between agents and synthesizing potentially divergent perspectives into unified responses.

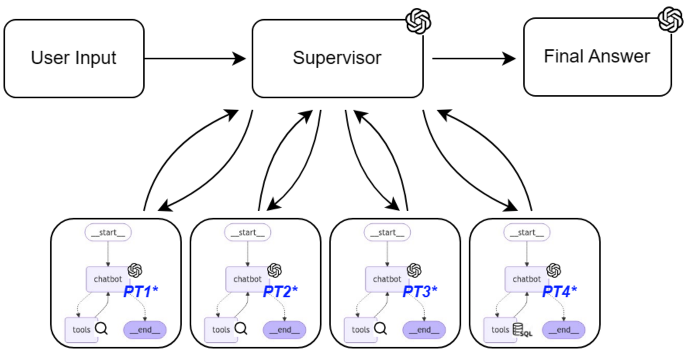


Figure 4.4: Multi-Agent setup with one supervisor and 4 specialist agents.

<insert table summarizing experimental setups and models here>

4.1.5 Execution Details

The experiment was driven by a script without manual intervention during the evaluation process. A main execution loop systematically iterated through all combinations of questions, agent setups, and language models defined in the experimental design.

Tool Integration and Knowledge Access

During execution, the agent systems accessed domain-specific knowledge through a standardized tool interface layer. This layer provided consistent access patterns across all experimental configurations, ensuring that differences in performance could be attributed to agent architecture rather than variations in knowledge availability. The tool integration framework supported a diverse range of knowledge access methods, including semantic search over unstructured text corpora, structured queries against relational databases, and specialized information extraction routines tailored to the oil well operations domain. Each tool invocation was executed within a controlled environment that captured performance metrics such as latency and resource utilization, providing additional dimensions for analysis beyond answer correctness. The standardization of tool interfaces across agent architectures was a critical design decision that enabled fair comparison while still allowing each architecture to implement its own strategy for tool selection and result interpretation.

Comprehensive Logging and Observability

A cornerstone of the experimental methodology was the implementation of comprehensive logging throughout the execution process. The system captured detailed records of each step in the question-answering workflow, from initial question parsing to final answer generation. These logs included intermediate reasoning steps, tool invocations with their inputs and outputs, and internal state transitions within the agent systems. All experimental artifacts were persisted in structured formats that facilitated both automated analysis and manual inspection. The logging system implemented a hierarchical organization that linked high-level metrics to the detailed execution traces that produced them, enabling root cause analysis of performance patterns. This observability infrastructure was essential for understanding not just what results were produced, but how and why different agent architectures arrived at their answers, providing insights into their reasoning processes and failure modes.

<insert code snippet or pseudocode of main execution loop here>

4.1.6 Evaluation Metrics

The evaluation of each experimental run is grounded in a comprehensive set of metrics designed to capture both the correctness and the quality of the system's responses. Standard quantitative measures such as accuracy, precision, recall, and

F1 score are calculated by comparing the answers generated by the agent systems to the established ground truth for each question. These metrics provide a multifaceted view of performance, indicating not only how often the system produces correct answers but also how well it balances false positives and false negatives.

For questions that are open-ended or less amenable to binary correctness, the evaluation framework employs a confusion matrix approach. This allows for a more nuanced assessment, capturing partial correctness and the degree to which the system's response overlaps with the expected content. Additionally, the methodology includes the calculation of the answer size ratio, which measures the verbosity of the generated answer relative to the ground truth. This metric helps to identify tendencies toward overly concise or excessively verbose responses, offering further insight into the models' behavior and suitability for practical deployment.

4.1.7 Limitations

While the experimental methodology strives for rigor and comprehensiveness, several limitations must be acknowledged. One key limitation concerns the coverage of the dataset: although the question set is carefully curated to represent a broad range of operational scenarios, it may not capture the full diversity of real-world challenges encountered in oil well operations. Similarly, the models and agent architectures evaluated are constrained by the available computational resources and the current state of language modeling technology, which may limit their ability to generalize beyond the scenarios tested.

Another limitation arises from the reliance on ground truth answers, which, despite expert validation, may still reflect subjective judgments or incomplete information in certain cases. Furthermore, the evaluation metrics, while robust, may not fully capture qualitative aspects of answer usefulness or clarity, especially in highly technical or ambiguous situations. Recognizing these limitations is essential for interpreting the results and for guiding future research aimed at addressing these gaps.

4.1.8 Summary

This methodology provides a systematic framework for comparing different agent architectures and large language models in the context of complex question answering for oil well operations. By integrating rigorous evaluation metrics, robust reproducibility practices, and a clear acknowledgment of limitations, the approach enables meaningful insights into the strengths and weaknesses of various system designs. The findings derived from this methodology can inform both the deployment of language model agents in operational settings and the ongoing development of more capable and reliable AI systems for specialized industrial domains.

4.1.9 Quality Assessment with LLM-as-a-Judge

To evaluate the quality of the responses generated by the RAG system, an automated evaluation approach known as *LLM-as-a-Judge* (ZHENG *et al.*, 2023) was adopted. This method uses a large language model (LLM) as an impartial judge to compare the system-generated response with a reference response, known as the *Ground Truth* (GT). Using an LLM for evaluation allows for a more granular and scalable analysis than manual human evaluation, while also capturing nuances of semantics and content.

The evaluation process was implemented through a structured *prompt* that instructs the LLM-judge to perform an analysis based on the concepts of a confusion matrix. The evaluation flow for each question-answer pair is as follows:

1. **Decomposition into Statements:** The LLM-judge is instructed to decompose both the system's response and the *Ground Truth* response into a set of atomic and objective statements. This allows for a detailed comparison, rather than a holistic and subjective evaluation.

2. **Classification of Statements:** Based on the comparison between the statements, the LLM-judge classifies them into three categories, following the logic of a confusion matrix:

   - **True Positive (TP):** Correct statements made by the system that are also present in the *Ground Truth* response.
   - **False Positive (FP):** Incorrect, irrelevant, or hallucinated statements made by the system that are not in the *Ground Truth* response.
   - **False Negative (FN):** Statements present in the *Ground Truth* response that were omitted by the system.

   The True Negative (TN) category is not applicable in this context, as the goal is to measure the precision and completeness of the information provided.

3. **Calculation of Metrics:** Based on the count of statements in each category (TP, FP, FN), the following information retrieval metrics are calculated:

   - **Precision:** Measures the proportion of correct statements among all statements made by the system. It is calculated as:

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

- **Recall:** Measures the proportion of correct statements that the system managed to retrieve compared to the total number of statements in the *Ground Truth* response. It is calculated as:

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

- **F1-Score:** Is the harmonic mean of Precision and Recall, providing a single metric that balances both values:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

This methodology allows for a quantitative and objective evaluation of the accuracy and completeness of the system's responses, providing valuable *insights* into its performance in different scenarios.

4.2 Results and Discussion

4.2.1 Performance

In this section, the F1-score charts are presented, which is the main metric for performance evaluation in this work. The corresponding charts for the precision and recall metrics are available in Appendix A.3.

É importante ressaltar que todos os resultados apresentados correspondem à melhor performance obtida em três execuções independentes para cada configuração de agente e modelo. A adoção desta metodologia visa mitigar a variabilidade inerente aos processos estocásticos presentes em muitos algoritmos de aprendizado de máquina. Fatores como a inicialização aleatória de pesos ou a aleatorização de dados podem levar a resultados distintos a cada execução. Ao selecionar o melhor resultado, busca-se apresentar o potencial máximo de cada configuração, reduzindo a chance de que uma performance inferior, causada por um ótimo local ou uma inicialização desfavorável, seja erroneamente interpretada como a capacidade real do modelo.

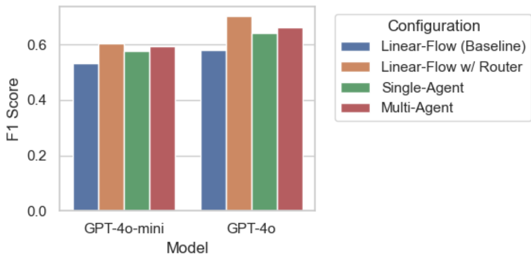


Figure 4.5: F1 score by model and configuration.

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 architectures, 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.

[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/SPEATCE/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.investidorpetrobras.com.br/en/results-and-announcements/results-center/>.

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://onepetro.org/SPEADIP/proceedings/23ADIP/2-23ADIP/D021S065R002/534485`>.

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/article-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/29/01/1/3358626/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/10.1145/331648.331659>.

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.grandviewresearch.com/industry-analysis/large-language-model-llm-market-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”. 6 2025. 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>.

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. 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. 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. 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. 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) 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) 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) 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)

Continued on next page

Table A.1 – Continued from previous page

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: 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"). 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"). 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"). 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"). 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. 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. 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. Referências: Alerta Técnico AT5533: Utilização de cimento com sílica na cimentação de revestimentos de superfície. Autor: Jan Fernandes Aslan. Data: 16/11/2020. 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.

Continued on next page

Table A.1 – Continued from previous page

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

Table A.1 – Continued from previous 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: 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. 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. 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. 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. 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. 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. 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.

Continued on next page

Table A.1 – Continued from previous page

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: - 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. 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. 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. 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. 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. 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: Documento da sonda NS-48 - 2024-07-19 11:50 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: Documento da sonda NS-48 - 2024-05-09 07:30 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: Documento da sonda NS-48 - 2024-05-21 05:24 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: Documento da sonda NS-48 - 2024-04-13 13:29 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: NS-48 - 2024-07-19 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: NS-48 - 2024-04-12

Continued on next page

Table A.1 – Continued from previous page

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.

Continued on next page

Table A.1 – Continued from previous page

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. 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: 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. 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

```
class Confusion_Matrix(TypedDict):  # type: ignore
    true_positive: list[str]
    false_positive: list[str]
    true_negative: list[str]
    false_negative: list[str]

def calculate_metrics(llm, question, history):
    if type(question['Ground Truth']) == str:
        prompt_confusion_matrix = f"""
            Voce recebera os seguintes parametros:
            Pergunta: a pergunta do usuario
            Resposta Ideal: a resposta considerada correta por um humano
            Resposta do sistema: a resposta fornecida pelo sistema baseado em IA

            Pegue a resposta do sistema, separe em afirmacoes e classifique cada afirmacao entre as opcoes abaixo:
            True Positive (TP): as afirmacoes corretas feitas pelo sistema, ou seja, que estao presentes na resposta ideal.
            False Positive (FP): as afirmacoes incorretas ou irrelevantes feitas pelo sistema, ou seja, que nao estao presentes na resposta ideal.

            Pegue a resposta ideal, separe em afirmacoes e classifique cada afirmacao entre as opcoes abaixo:
            True Negative (TN): Nao se aplica, deixar vazio.
            False Negative (FN): as afirmacoes que constam na resposta ideal, mas nao foram feitas pelo sistema.

            Importante:
            - Voce deve gerar listas de afirmacoes para cada categoria.
            - Voce deve quebrar as respostas do sistema e a ideal em afirmacoes objetivas.
            - Se as respostas do sistema ou a ideal contiverem frases grandes com mtas afirmacoes, analisar cada afirmacao separadamente.
```

```
        - Uma afirmacao nao pode estar em mais de uma categoria.
        - Ignore coisas na resposta do sistema que nao sao afirmacoes objetivas, como por exemplo citacoes de fontes e links.

        Vamos la!

        ############

        Pergunta: {{{{
        {question['Question']}
        }}}}

        Resposta ideal: {{{{
        {question['Ground Truth']}
        }}}}

        Resposta do sistema: {{{{
        {history[-1].content}
        }}}}"
    response = llm.with_structured_output(Confusion_Matrix).invoke(prompt_confusion_matrix)

    true_positive_count = len(response.get('true_positive', []))
    false_positive_count = len(response.get('false_positive', []))
    true_negative_count = len(response.get('true_negative', []))
    false_negative_count = len(response.get('false_negative', []))

    try:
        precision = true_positive_count / (true_positive_count + false_positive_count)
    except ZeroDivisionError:
        precision = 0

    try:
        recall = true_positive_count / (
```

```
    true_positive_count + false_negative_count)
        except ZeroDivisionError:
            recall = 0

        try:
            f1_score = 2 * (precision * recall) / (precision
    + recall)
        except ZeroDivisionError:
            f1_score = 0

        print("\\n\\nPrecision: "+str(precision))
        print("Recall: "+str(recall))
        print("F1 Score: "+str(f1_score))

        print("\\nAnswer Size Ratio: "+str(question['Answer
    Size (% of GT)']))

        state.precision = precision
        state.recall = recall
        state.f1_score = f1_score
        state.ground_truth = question['Ground Truth']
        state.statements = response
```
Listing A.1: Código para LLM-as-a-Judge

A.3 Results

Best Precision by Model and Configuration

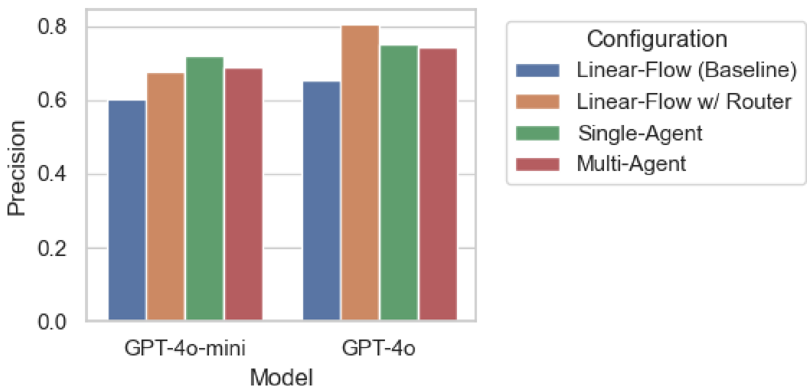


Figure A.1: Best precision by model and configuration.

Best Precision by Question Index and Configuration

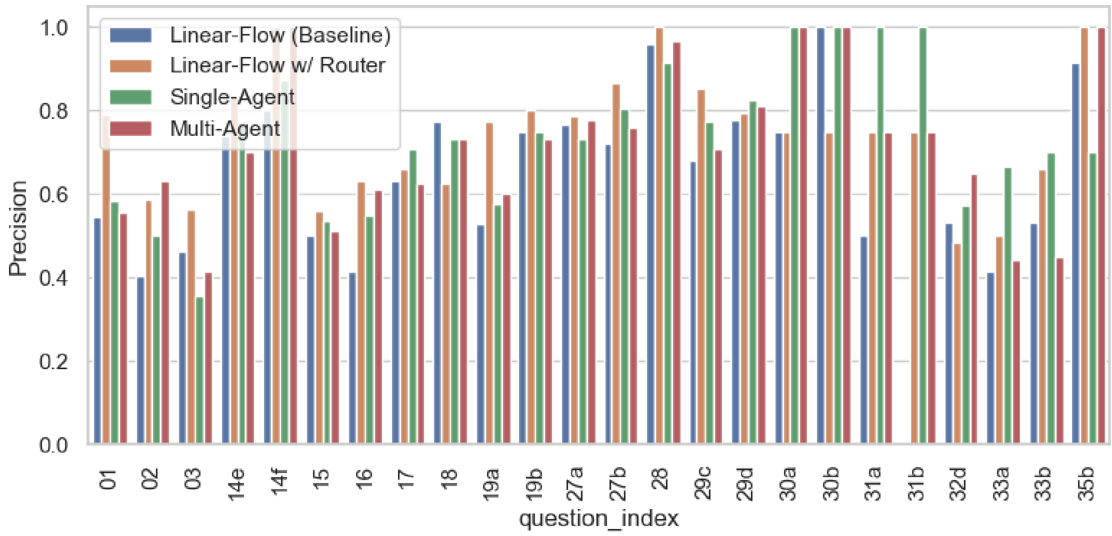


Figure A.2: Best precision by question index and configuration.

Best Precision by Question Index and Model

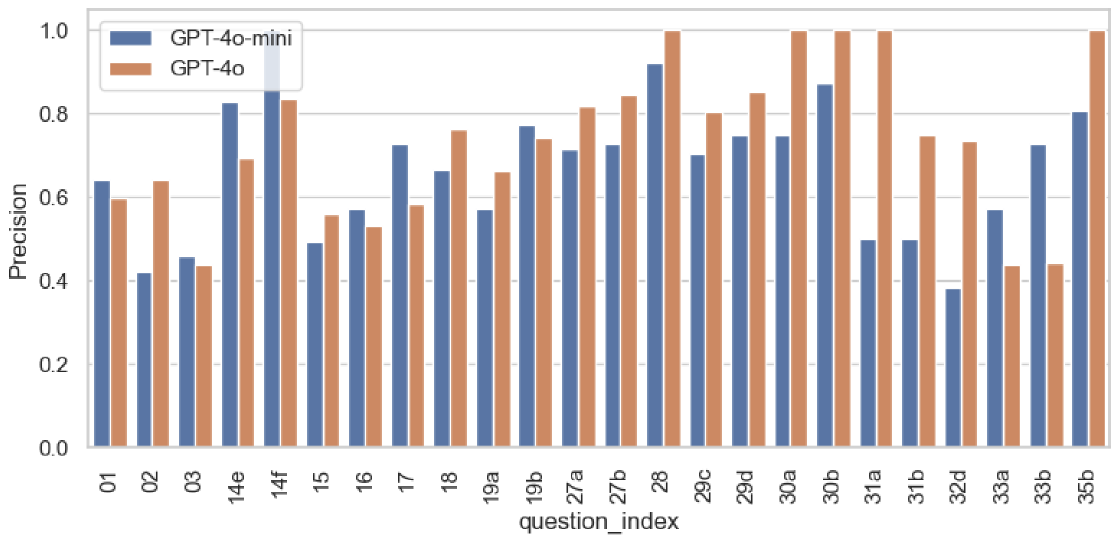


Figure A.3: Best precision by question index and model.

Facet Histogram of Precision by Model

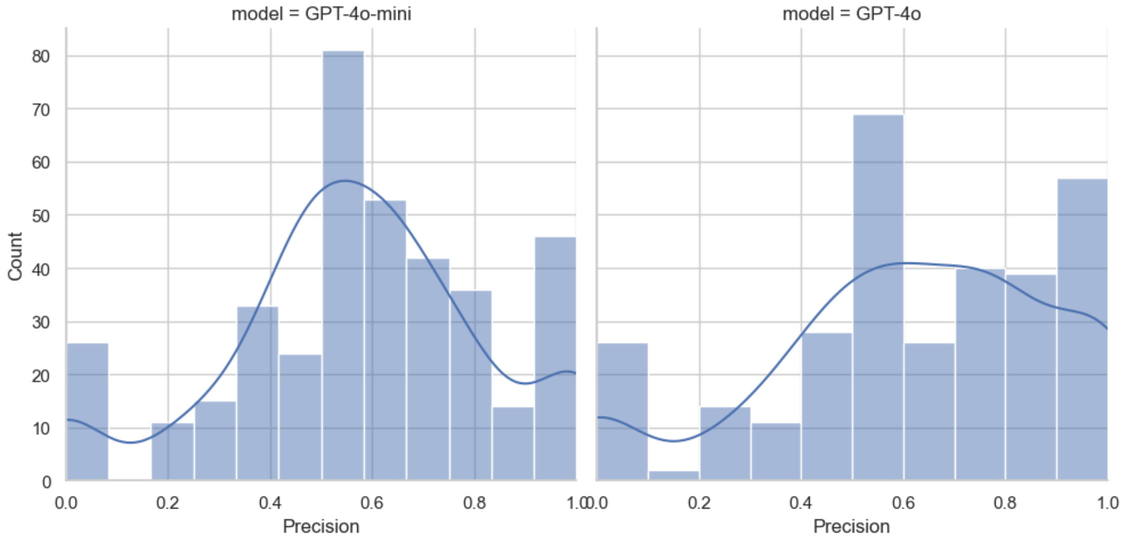


Figure A.4: Facet histogram of precision by model.

Facet Histogram of Precision by Model (best of 3)

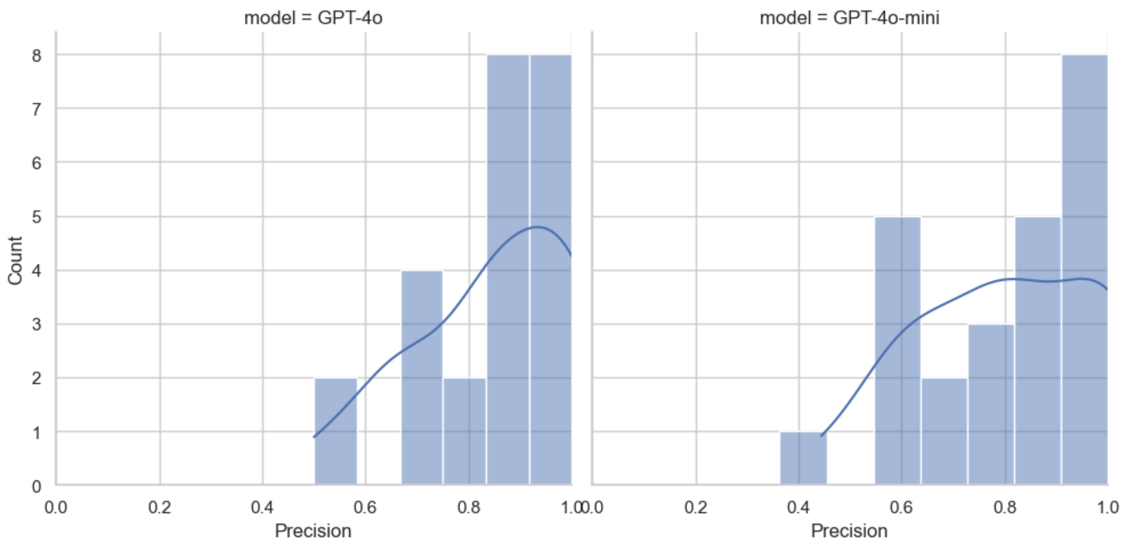


Figure A.5: Facet histogram of precision by model (best of 3).

Histogram of All Precisions

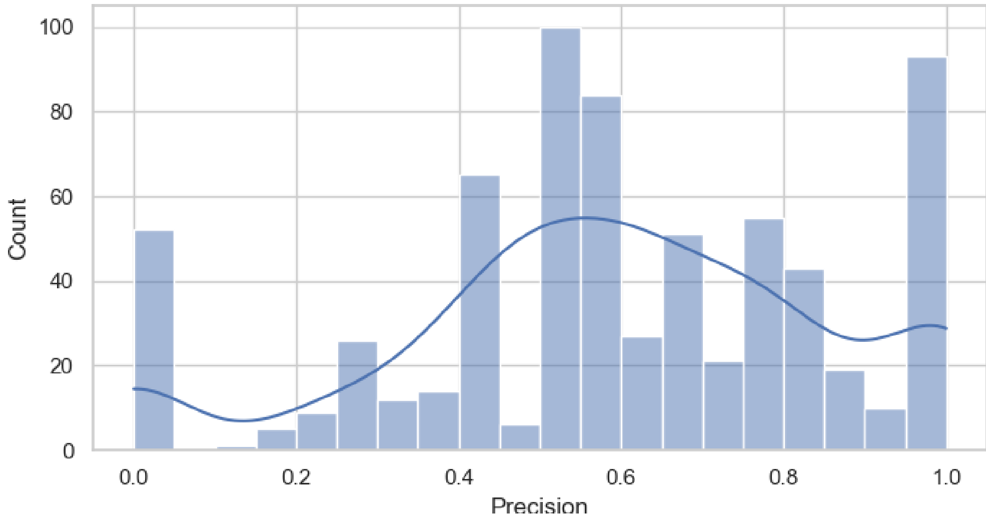


Figure A.6: Histogram of all precisions.

Line Plot of Precision by Question Index and Model

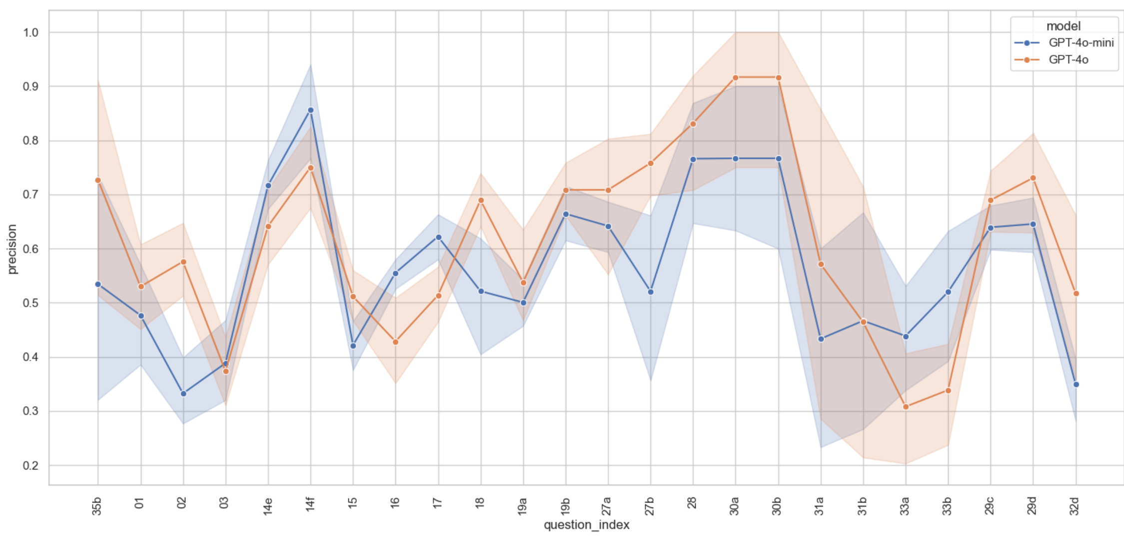


Figure A.7: Line plot of precision by question index and model.

Boxplot of Precision by Model and Configuration

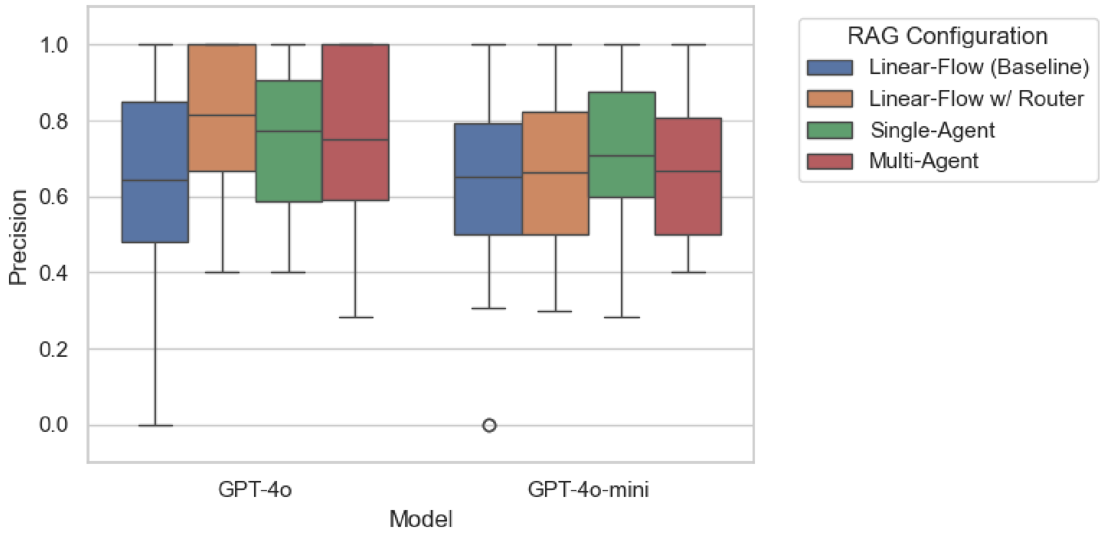


Figure A.8: Boxplot of precision by model and configuration.

Precision by Model

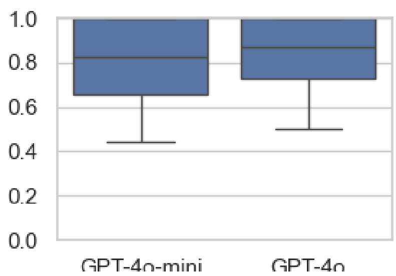


Figure A.9: Precision by model.

Precision by Model and Configuration

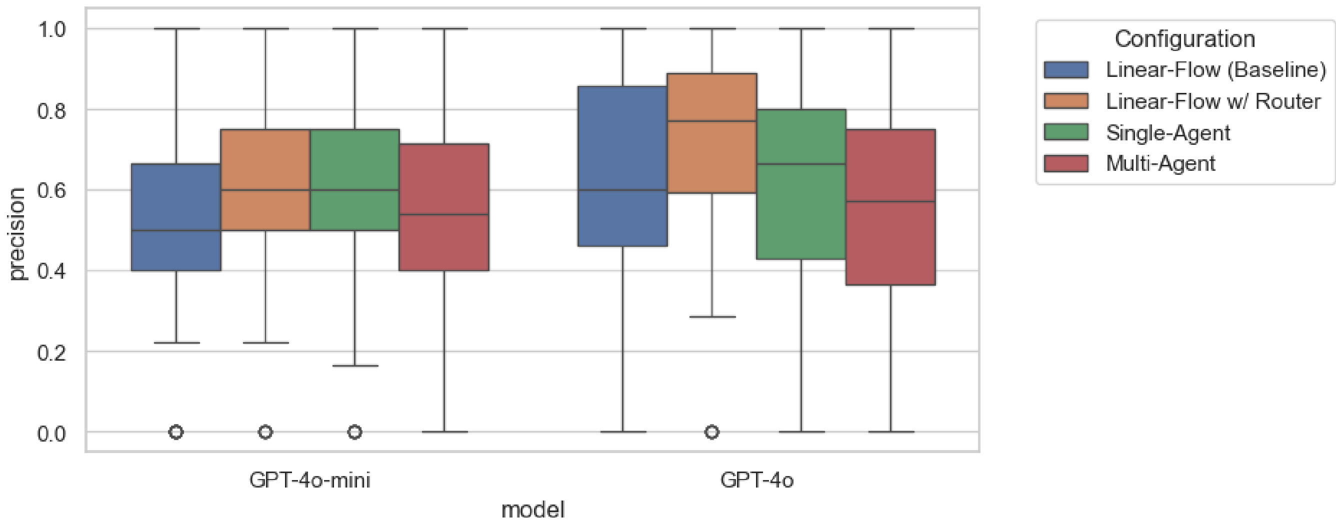


Figure A.10: Precision by model and configuration.

Line Plot of Precision by Question Index and Configuration

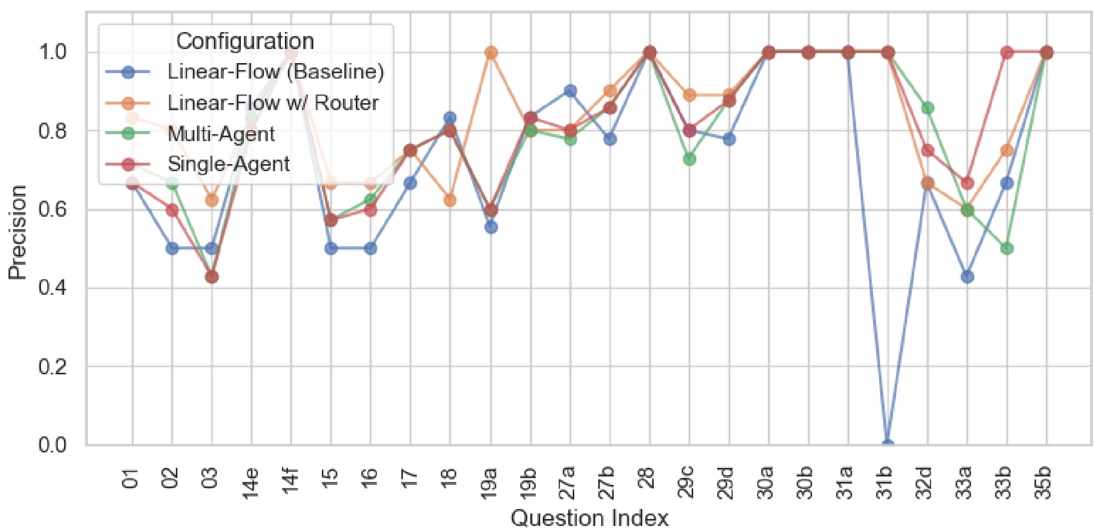


Figure A.11: Line plot of precision by question index and configuration.

Scatter Plot of Precision vs. Total Time

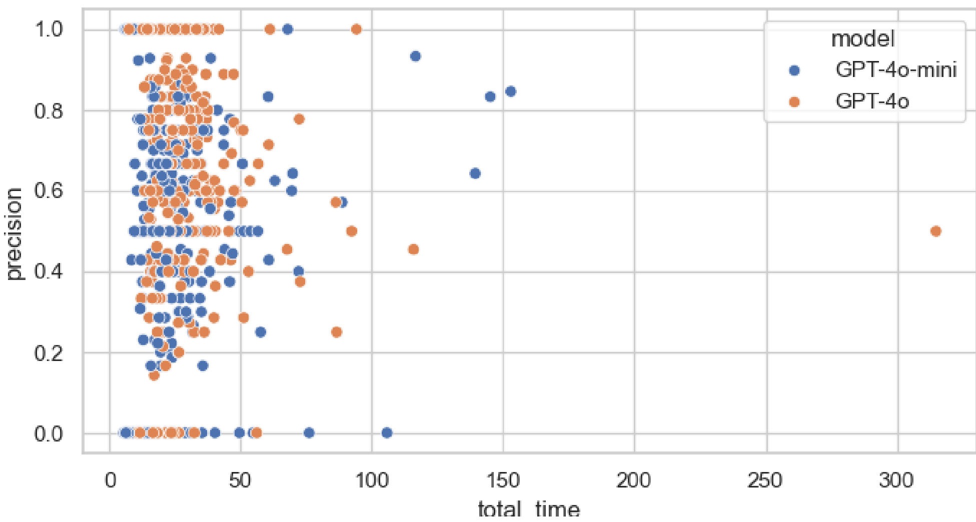


Figure A.12: Scatter plot of precision vs. total time.

Scatter Plot of Precision vs. Total Token Count Input

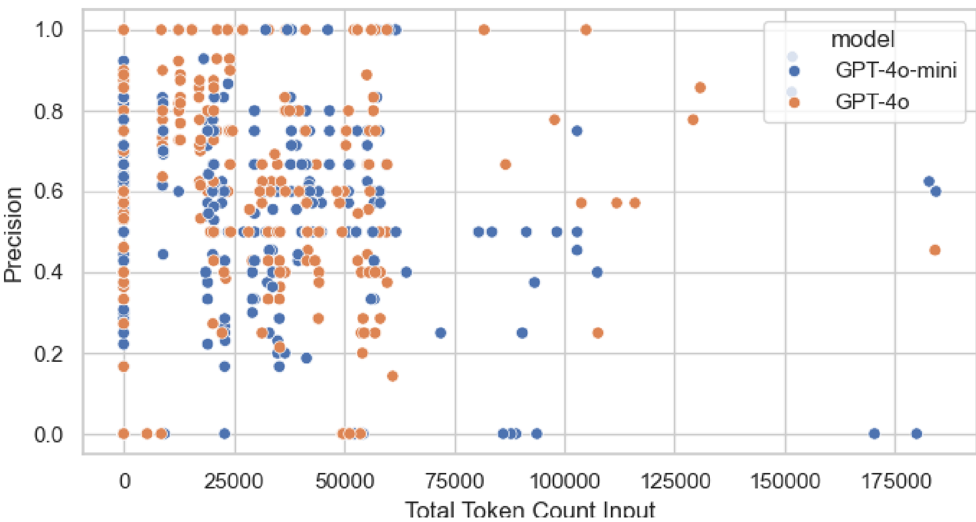


Figure A.13: Scatter plot of precision vs. total token count input.

Swarm Plot of Precision by Model and Configuration

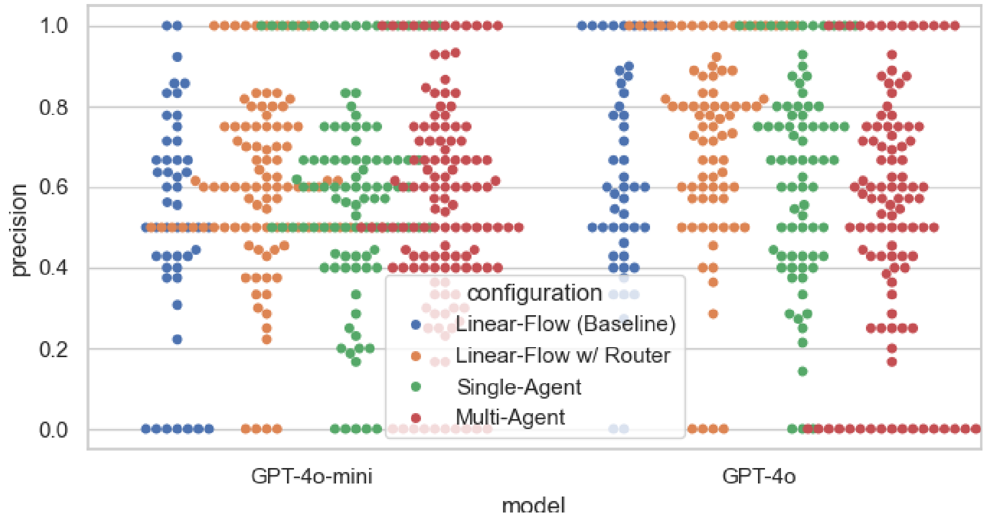


Figure A.14: Swarm plot of precision by model and configuration.

Violin Plot of Precision by Model and Configuration

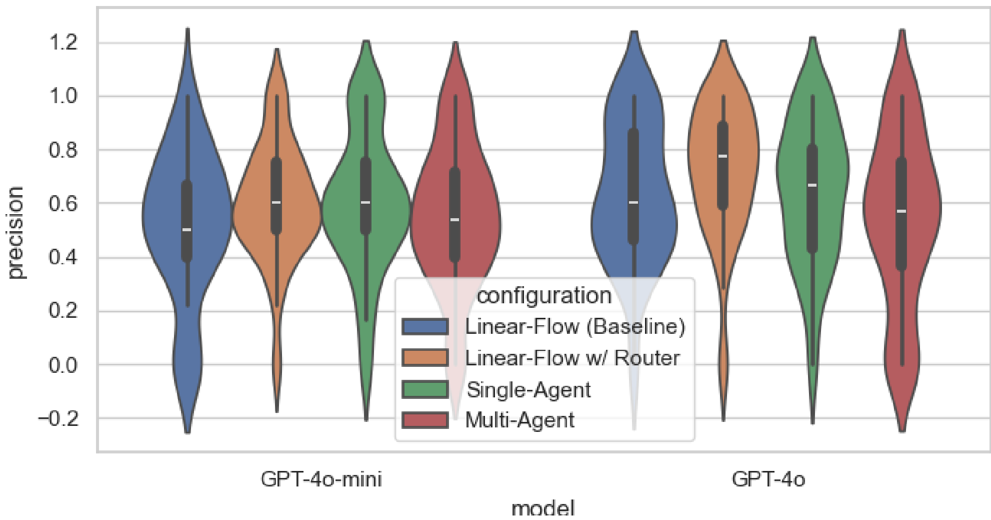


Figure A.15: Violin plot of precision by model and configuration.

List of Comments