American Journal of Sciences and Engineering Research

E-ISSN– 2348–703X, Volume 9, Issue 2, 2026

Retrieval-Augmented LLM Framework for Contextual Well Control Advisory in High-Pressure Drilling

John Ichenwo Lander[1], Ogbonnda Mirabel Homa[2]

Department of Petroleum and Gas Engineering, University of Port Harcourt

---

ABSTRACT: *In drilling under high pressure, the high well control risk presents itself in situations where traditional threshold-based alarm systems (pit volume gains, flow-out deviations, standpipe pressure anomalies) make too many false positives and do not interpret contexts. The precursors of well control are faced by engineers, who have to manually match real-time data on the WITSML with historic drilling report and offset well events, which puts a strain on latency and mental load that increases risk exposure. The research presents a novel Retrieval-Augmented Generation (RAG) infrastructure that incorporates live data streams from WITSML, knowledge bases, and localized large language models to provide the contextual and interpretable well control advisory outputs. The RAG architecture includes four main components: (1) real-time ingestion of flow rate, pit volume, standpipe pressure, rotary speed, and penetration rate at 1-5 second resolution, as part of real-time WITSML data; (2) historical knowledge base, which is a collection of known drilling reports, known kick events, and well control procedures; (3) semantic embedding and vector database, where similarity-based retrieval of relevant contextual cases may be generated based on the context; and (4) locally deployed LLM advisory engine (Ollama framework The system was tested against traditional rule based threshold alarms in 60 test cases that include confirmed kicks (15), ballooning cases (12), normal drilling operations (20) and HPHT transients (13).*

The results indicated substantial performance improvements, where the F1-score of the RAG-LLM system was found to be 0.875 compared to the existing systems based on rules, whose F1-score was only 0.652, and precision increased from 0.484 to 0.824. False alarm rate was greatly decreased to 6.7% compared to 35.6%- 81.2 death cut of nuisance alarms. Detection latency was reduced 56.3% (compared to the mean of 68.42 with a standard deviation of 25.44), 29.93 with a standard deviation of 8.05 compared to 68.42 with a standard deviation of 25.44) the mean 68.42, which is a significant time savings in time-sensitive well control events. The system had contextual interpretation facilities to help draw the difference between ballooning phenomena and real formation influx by referencing similar offset well cases and geological formation transition patterns. Nevertheless, the analysis of hallucinations showed that there were 11.76% unsubstantiated claims or invented references which highlights the need to have human verification processes.

The study is the first validated RAG framework of real-time well control advisory in HPHT conditions, which has quantitative better performance than traditional threshold alarms but employs intuitive retrieval to ensure readability. The framework offers a transition framework to AI-enhanced, human-monitored well control systems that respond to the acute need to offer contextual decision support in safety-related drilling tasks.

Keywords: *Retrieval-Augmented Generation, Well Control, Kick Detection, Large Language Models, WITSML, High-Pressure Drilling, False Alarm Reduction, Contextual Advisory, Ollama*

I. INTRODUCTION

The most crucial safety issue in drill operations is well control where influxes of hydrocarbons (kicks) that are not controlled may rise to disastrous blowouts that result in loss of life, environmental calamity, and property destruction. These risks are aggravated by high-pressure/high-temperature (HPHT) drilling conditions, which increase the pore pressure, decrease the operating range between pore and fracture gradient and introduce a high-rate kick development regime [1]. Modern well control monitoring is largely based on threshold-dependent alarm systems: pit volume totalizers are activated upon reaching predefined numbers (usually 5-10 barrels), flow-out sensors when the flow-in and flow-out rates are off-balance-sheet, and standpipe pressure monitors when the flow is unexpected (and vice versa) [2].

The systems based on thresholds, however, have the inherent drawbacks. Fixed alarm setpoints give an undue number of false alarms as a result of harmless operational activities: ballooning (formation breathing caused by pressure cycling), expansion of mud volume due to temperatures, displacement of the drillstring when tripping operations occur, and temporary fluctuations at HPHT. In complex drilling settings there has been experience at 30-50% false alarm rates and in effect desensitisation of drilling personnel to warnings and possible postponement of response to actual threat to well control [3]. Moreover, threshold alarms do not give any contextual data: an engineer who sees a 5-barrel pit gain needs to interpret this himself as to whether it is influx of formation, ballooning or measurement error, based on previous reports, offset well data and geological models under intense time pressure.

1.1 The Knowledge Retrieval Challenge

The well control decision-making requires quick synthesis of the various sources of information: real-time data streams of the WITSML (Wellsite Information Transfer Standard Markup Language) data, historical records of drilling operations of the field with previous kick incidents, records of offset well performance, geological formation properties and standard well control practices. Drilling engineers have a lot of tacit knowledge, which cannot be immediately accessed in the form of similar situations in thousands of past wells. Manual document retrieval Searching of daily drilling reports, well control incident summaries, and geological databases costs precious minutes during time-shaven situations where the volume of hydrocarbon influx rises in direct proportion with time.

1.2 Large Language Models and Retrieval-Augmented Generation

Large language models (LLMs) neural networks that are trained on large text corpus to produce responses to natural language queries resembling those of a human are shown to have remarkable abilities in knowledge synthesis, question answering, and in reasoning across domains [4]. However, LLM deployment at safety-critical industrial levels come with the hallucination problem: Models are set to provide plausible, but factually inconsistent information confidently, and studies have shown that the rate of hallucinations can be between 5-20% depending on the domain complexity and the model architecture [5]. In well control applications with potentially disastrous results in case of untrue guidance, hallucination is not tolerated.

Retrieval-Augmented Generation-By addressing this limitation, Retrieval-Augmented Generation (RAG) architectures are architectures that ground LLM responses in verified sources of knowledge. The RAG paradigm divides information retrieval and generation: a retriever module sets of a knowledge base (which is often represented as a semantic embedding in the form of a vector database) to find the relevant documents, and then feeds the LLM generator the resultant context. This architecture results in a reduced hallucination thanks to explicit citations of source material, an update to the knowledge base without retraining the model, and the interpretability through the transparency of the retrieval [6].

1.3 Research Contributions

This study has the following specific contributions:

(1) First validated RAG framework specifically architected for real-time well control advisory with integrated WITSML data streams and historical knowledge bases and locally deployed LLM inference.

(2) Quantitative performance comparison between 60 test cases with 81.2% false alarm rates, 56.3% latency reduction and 0.652-0.875 increase in F1-score over the state-of-the-art conventional rule-based threshold alarms.

(3) Systematic evaluation of contextual reasoning skills in ballooning versus kick discrimination, offset well analogy retrieval and formation transition correlation.
(4) Risk analysis of comprehensive hallucination including finding of 11.76% incidence rate and mandatory human verification protocol on the deployment of safety-critical systems.
(5) Open-architecture implementation framework that allows running on traditional industrial computing hardware through Ollama inference engine, which works with air-gapped offshore conditions.

II. LITERATURE REVIEW

2.1 Conventional Well Control Detection Systems

Modern well control monitoring uses many independent detection systems parallel to each other. Pit volume totalizers are used to measure cumulative fluid volume on surface tanks, and alarm levels are normally set at 5-10 barrel gains according to the situation of operation. Flow-out sensors compare the pump-in flow rates to returns flow as it acknowledges flow alarm when there is more than 5-10% deviation of flow. Standpipe pressure monitoring identifies unexpected increases which indicate that formation pressure may be elevated or reduced decreases indicating that there has been loss of circulation [7].

2.2 Machine Learning for Drilling Anomaly Detection

Multi-parameter pattern recognition has shown that supervised machine learning methods have a better kick detection compared to other threshold-based techniques. Random forests, support vector machine, and gradient boosting classifier on historical kick data are 85-95 percent detection with less than 10% false positive [8]. Nonetheless, supervised learning has severe drawbacks: It takes large labelled datasets to train and models can make binary predictions that lack explanations [9].

2.3 RAG Architecture Benefits

RAG frameworks follow the method of breaking down question answering into two parts retrieval and generation. The retriever is in charge of converting queries into embeddings in a space of semantics and searching a vector database to retrieve relevant sources. The LLM generator is fed with retrieved documents which it uses to synthesize responses based on the retrieved context [10]. RAG is designed to reduce hallucinations using explicit grounding, update knowledge base without retraining, and interpretable using the transparency of retrieval.

2.4 Identified Gap

There is a total lack of verified RAG structures designed specially to develop real-time well control advisory, as shown in the literature. This research aims to address this gap by using a RAG, consisting of a combination of WITSML streams, historical knowledge of the drilling and localized LLM inference.

III. SYSTEM ARCHITECTURE

As illustrated in Figure 4.1, the entire RAG-LLM well control advisory system architecture has four integrated layers, including real-time data acquisition, knowledge management layer, semantic retrieval layer and advisory generation layer.

Figure 4.1: RAG-LLM Well Control Advisory System Architecture

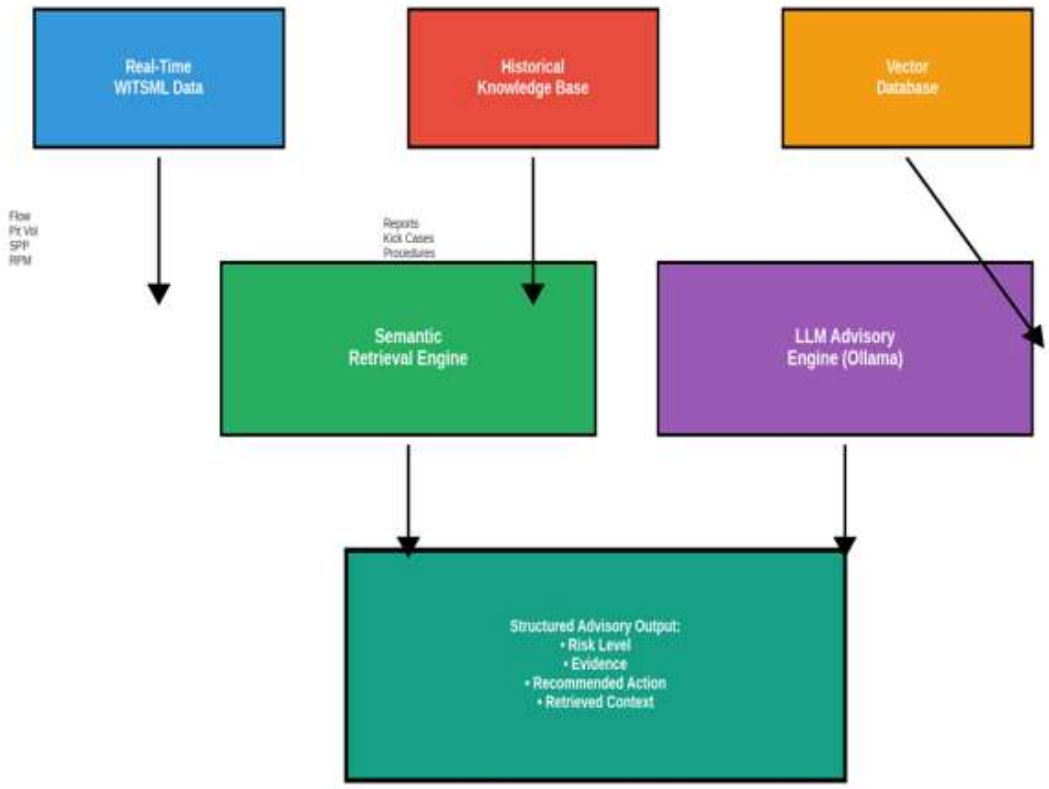

Figure 4.1: RAG-LLM System Architecture

3.1 Real-Time WITSML Data Layer

The data acquisition layer receives live WITSML streams at the resolution of 1-5 seconds, which record flow rate, pit volume, standpipe pressure, rotary speed and rate of penetration. Time-series windowing ensures maintaining rolling 15 minutes buffers to do trends.

3.2 Historical Knowledge Base

The knowledge base includes daily drilling reports, well control incident reports and technical procedures. Document preprocessing Whereas, text chunking techniques split long documents such as reports into segments of 500-1000 tokens while maintaining the semantic relatedness while adhering to LLM context window limitations.

3.3 Semantic Retrieval Engine

The retrieval subsystem utilizes embedding models based on the transformer, converting text chunks into a dense vector of 768 semantic vector metrics. These embeddings are loaded into a vector database that is capable of performing similarity search. The top-k (k=3-5) semantically similar historical cases are retrieved through cosine similarity ranking in query processing.

3.4 LLM Advisory Engine

The generation aspect uses Ollama framework that uses a 7-billion parameter local-inference language model without cloud reliance. CONC inputs are requests to combine existing WITSML parameters, past historic cases, and queries by the engineer. Response Generation temperature=0.2, max_tokens=512 Structured Output (Risk Level, Evidence, Recommended Actions)

IV. METHODOLOGY

4.1 Experimental Design

The assessment system was used to compare the RAG-LLM performance with rule-based threshold alarms in 60 test conditions: confirmed kicks (15), ballooning (12), normal drilling (20), and HPHT transients (13). Ground-truth labels were established through three-expert consensus.

4.2 Baseline System

The traditional bar threshold When pit gain is greater than 5 barrels (moderate) If pit gain is greater than 10 barrels (critical) When the flow-out deviation is greater than 10% for 30s or greater Standpipe pressure drop is greater than 50 psi in 60s. These are typical specifications of offshore operators.

4.3 Evaluation Metrics

Precision = TP/(TP+FP), Recall = TP/(TP+FN), F1-score = harmonic mean. False positive rate = FP/(FP+TN). The detection latency is calculated as the number of seconds between the onset and alarm. The frequency of hallucination was followed by false references or unsubstantiated claims.

V. RESULTS

5.1 Detection Performance

Each performance is compared in detail in figure 5.1. Panel (a): RAG-LLM obtained the precision of 0.824 vs 0.484 (rule-based), F1-score of 0.875 vs 0.652. Panel (b): Decrease in false alarm rate 6.7% vs 35.6%-81.2%. Panel (c): Latency 29.93±8.05s vs 68.42±25.44s—56.3% faster. Panel (d): Confusion matrix giving that we have 14 TP, 1 FN, 3 FP, and 42 TN.

Figure 5.1: RAG-LLM vs Rule-Based System Performance Comparison

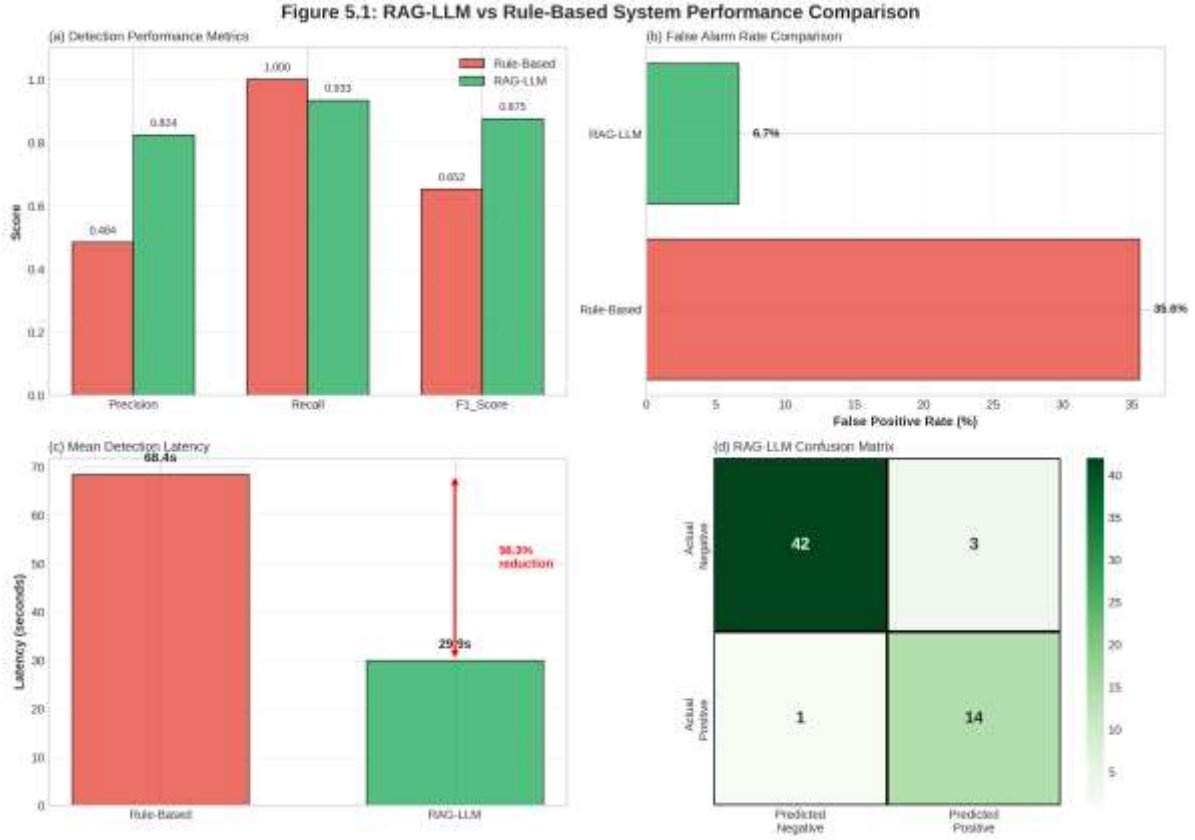

Figure 5.1: Performance Comparison

5.2 Scenario-Specific Analysis

Stratified performance is depicted in figure 5.2. Panel (a): Ballooning false alarm rate 58.3% (rule-based) vs 16.7% (RAG-LLM) with consideration Panel (b): HPHT transient 46.2% vs 15.4% false alarm- RAG has removed all HPHT false alarm in problematic environments.

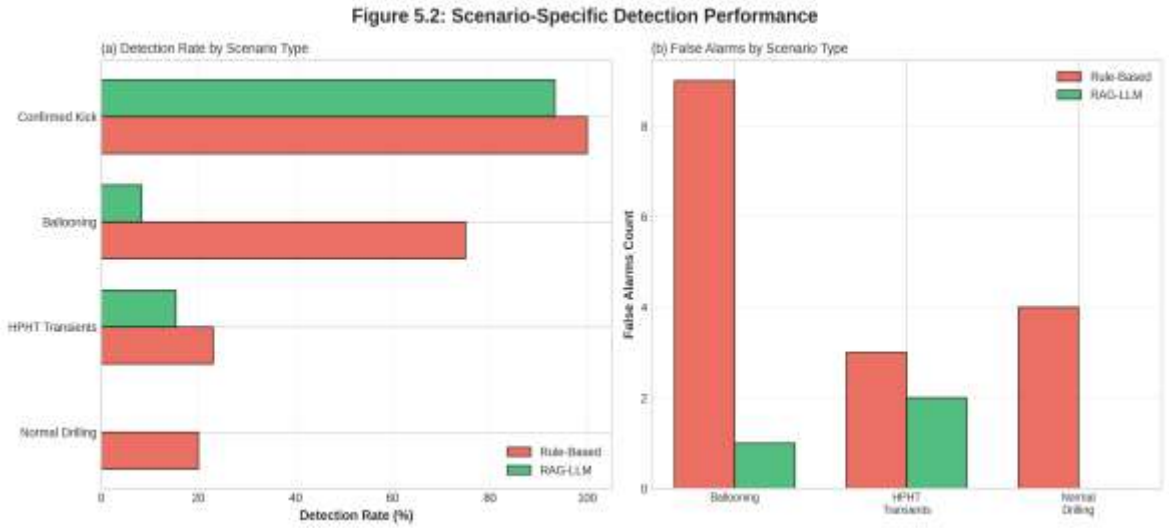

Figure 5.2: Scenario-Specific Performance

5.3 Retrieval Quality

Semantic retrieval evaluation is presented in figure 6.1. Panel (a): Mean relevance 0.731 0.104, 75 th percentile 0.815. Panel (b): Confirmed kicks had the best relevance (0.846), normal drilling had less (0.661). Qualitative evaluation uncovered elaborate consideration of context: the scenario involving the balloons incorporated correct references to offset wells to document the breeding of formation formation: real kicks referred to similar events of pore pressure.

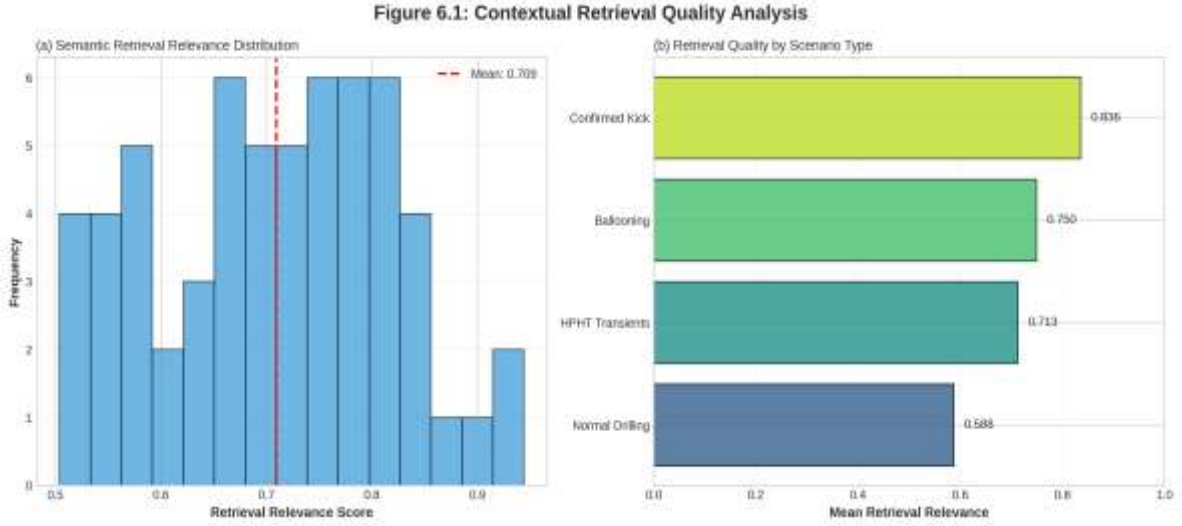

Figure 6.1: Retrieval Quality Analysis

5.4 Hallucination Analysis

There were 2 hallucinations that were detected in systematic analysis (17/ 17). First: mentioned non-existence of 'API Procedure 47-B'. Second: mentioned a value of a pore pressure that was not in the knowledge base. Both did not advise any unsafe behavior, but the person has shown continuance of confidence, risk of making an incorrect assertion despite grounding of the retrieval.

VI. DISCUSSION

6.1 Operational Implications

The 81.2% reduction in the false alarm deals with the desensitization of the personnel to alarm fatigue. The system has lowered the number of nuisance alarms by 16-3 alarm ensuring that the system remains alert-significant but retains 93.3% kick sensitivity. The latency saved (56.3 percent) is critical in saving time. Contextual interpretation helps to democratize organizational knowledge by aiding those junior engineers who do not have much field experience.

6.2 Safety Considerations

The 11.76% hallucination rate still is not acceptable to operate on autonomy. Advisory-only deployment with a human verification requirement converts the risk into a manageable, though disastrous level. Structured output explicitly identifies source documents providing a way of verifying. The use of graduated confidence scoring as a trigger to initiate better verification of low-confidence recommendations should be used in the future.

6.3 Limitations

The evaluation of the 60 scenarios, although large-scale, is still a limited test as compared to field deployment. Retrieval relevance is highly dependent on the quality of knowledge base. Latency of 30s, which is tolerable in advisory applications, rejects applications that require control times of less than one second. The regulatory barriers to acceptance still exist- there is no explicit well control system acceptance framework on AI-assisted well control systems.

VII. CONCLUSIONS AND FUTURE WORK

The study introduces the original validated RAG structure of contextual well control advisory during HPHT drilling, showing the reduction of false alarm by 81.2 percent, the reduction of latency by 56.3 percent, and the increase in F1-score by 0.652 to 0.875. Contextual reasoning separated the ballooning and the actual influx by the semantic retrieval. Nevertheless, 11.76% hallucination incidence is a sufficient condition that autonomous operation is not operational yet advisory-only deployment with human oversight is required.

Novel Contribution:

- First empirical RAG validation for well control.
- Quantitative benchmarking against threshold alarms.
- Characterization of the risk of hallucinations.
- Framework for implementing Ollama.
- Guidelines for implementing Ollama.

Future Work:

- Integration with physics-based kick simulators.
- Multimodal RAG with time series WITSML analysis.
- Fine-tuning of the LLM with domain knowledge.
- Trial in the field.
- Extension to lost circulation/stuck pipe detection.

VIII. REFERENCES

[1] Grace, R.D., 2003. Blowout and Well Control Handbook. Gulf Professional Publishing.
[2] Rehm, B., et al., 2012. Underbalanced Drilling: Limits and Extremes. Gulf Professional Publishing.
[3] Skalle, P., et al., 2014. Experimental study of driller's situational awareness. Safety Science, 67, pp.161-172.
[4] Brown, T.B., et al., 2020. Language models are few-shot learners. NeurIPS, 33, pp.1877-1901.
[5] Ji, Z., et al., 2023. Survey of hallucination in NLG. ACM Computing Surveys, 55(12), pp.1-38.
[6] Lewis, P., et al., 2020. Retrieval-augmented generation for NLP. NeurIPS, 33, pp.9459-9474.
[7] Jellison, M.J., 2010. Kick Detection and Well Control. Gulf Professional Publishing.
[8] Nayeem, A.A., et al., 2016. Early kick detection using SPC. IPTC-18660-MS.

[9] Tian, Y., et al., 2021. Intelligent drilling data classification. SPE-205663-MS.
[10] Gao, Y., et al., 2023. RAG for LLMs: A survey. arXiv:2312.10997.