OPEN ACCESS

International
Journal
Of Modern Engineering Research (IJMER)

Evaluating the Impact of Locally Deployed Large Language Models on Real-Time Drilling Decision Latency and Accuracy

John Lander Ichenwo, Ejimkonye Blessing Chiamaka
Department of Petroleum and Gas Engineering, University of Port Harcourt

ABSTRACT
Drilling processes are real-time and require fast and correct decisions among engineers that are highly time pressure and cognitively loaded. The exposure to risk through decision latency in cases of abnormal drilling events, stick-slip vibration, torque spikes, loss of circulation, or well control precursors, is directly proportional to the delay in response, and the simple dysfunctions that could lead to catastrophic failures may be aggravated by latency in decision-making. This research offers the first empirical evaluation of locally deployed large language models (LLMs) for the purpose of providing decision support for drilling, and in doing so, poses a new question with regards to the viability of edge-based AI for decision latency improvement, all while being operationally accurate and safe.

An experimental induced control framework was used to compare manual responses of engineers with responses assisted by the LLM in eight standardised drilling conditions (stick-slip onset, sudden torque spike, ROP drop, loss of circulation indicators, well control precursors, excessive bit wear, whirl dysfunction and kick detection) and twelve participating petroleum engineers responding to 96 total experimental inductions. The system used, the LLM, utilized local inference through Ollama, and used a 7-billion parameter model with no external internet connectivity, making it viable in a working environment with no external internet connections. The quantitative measures included decision latency (time taken to complete a scenario until a final decision), decision accuracy (correctness in comparison to the experts on the ground), and hallucination rate (frequency of confidently stated wrong advice).

Results represented statistically significant reduction in the latency: the mean manual decision time of 42.23 + 16.26 seconds vs. LLM assisted time of 31.35 + 13.29 seconds ($p < 0.001$, paired t-test) representing 25.8% absolute reduction (10.88 seconds). The accuracy of the decision made increased by 10.42 (manual) to 95.83 (LLM-assisted) due to a high level of procedural recall and less human error due to time constraints. Nevertheless, due to hallucination analysis, it was found that 10.42% of LLM responses included factually incorrect guidance with the distribution of severity consisting of 60% minor error, 30% moderate operational error and 10% severe safety-critical error. The highest rates of hallucinations (1218% involuntary) occurred in safety-critical scenarios (well control precursors, kick detection, loss of circulation), and lower rates occurred in routine optimization scenarios (48%).

This study creates operational constraints of safe LLM application in drilling activities, showing quantifiable latency gains, but also defining unacceptable risks of autonomous operation activities. The recommendations support the application of human-AI hybrids, where LLMs will be employed as advisory systems with verification layers, especially for safety-critical decisions. The analysis of resource usage verified the possibility of edge deployment: average CPU utilisation 58.7%, 3552 MB memory usage and inference latency 3.47 seconds, which is within the range of mainstream industrial computing hardware. The areas for future research include retrieval-augmented generation architecture from verified drilling procedures, fine-tuning on drilling engineering corpora, and their integration with digital twin systems allowing for context-aware decisions.

Keywords: *Large Language Models, Drilling Decision Support, Decision Latency, Hallucination Risk, Edge Computing, Real-Time Operations, Human-AI Collaboration, Safety-Critical Systems, Ollama*

I. INTRODUCTION

Real-time drilling operations are dynamic, complex systems, which have a high level of uncertainty, are time-constrained with the necessity to make decisions quickly, and the consequences of error are severe. Petroleum engineers overseeing the drilling parameters need to quickly analyze surface and downhole measurements, diagnose abnormal conditions, retrieve established procedures in large volumes of technical documentation, and take corrective measures of actions- often in the context of working pressure and with a limited amount of consultation time. The latency of a decision, the amount of time that passes between detection of an abnormality and implementation of corrective action, has a direct relationship with the risk exposure and the risk of the escalation of costs [1].

The case of drilling dysfunctions represents the time sensitive situation that requires fast, precise reaction. Unmitigated stick-slip torsional oscillations increase the rate of drillstring fatigue and bit wear, and might result in disastrous tool joint failures. Sudden torque spikes are indicative of an imminent occurrence of pipe sticking or differential sticking that necessitates weight-on-bit corrections. Indicators of loss of circulation require prompt changes in fluid properties to avoid fracturing and lost circulation incidents worth £100,000 - £500,000 each. The precursors of well control such as increasing abnormal pit volume gain, increasing flow rate, and deviated formation pressure are aspects that require immediate action to avoid uncontrolled inflow of hydrocarbons with devastating safety and environmental effects [2].

1.1 Cognitive Load in Time-Critical Drilling Decisions

The research of human factors in drilling operations has proven that cognitive load negatively affects the quality of decisions when time-pressured. Engineers have to watch dozens of parameters at the same time, have situational awareness of the situation in the wellbore, have procedural wisdom on thousands of pages of technical documentation, and interact with multidisciplinary teams. Research on decision-making on drilling indicates that during normal decision-making (routine) the error rates are 10-15 per cent and when decision making is under abnormal conditions, the error rates reach 20-30 per cent due to the increase in time pressure and stress on the cognitive demands [3].

The bottleneck of knowledge retrieval is especially dangerous. Although the experienced engineers will intuitively identify typical dysfunctions, unusual or complicated cases will demand a reference to technical manuals, corporate practice, regulator regulations, and even company documentation- an activity that takes minutes, which are vital during an operation that has to be timely. Moreover, the additional restrictions are caused by offshore drilling activities: there is a lack of personnel (the average amount of offshore drilling engineers is 2-3 drilling engineers on one shift), there is a delay in the transmission of information to the onshore part of the company, and the internet access is also limited, so external advice cannot be obtained.

1.2 Large Language Models as Decision Support Systems

Large Language Models (LLMs)-neural networks that learn from the text to produce human-like text from input prompts, with knowledge synthesis, question answering, and procedural reasoning opportunities in various fields [4].The applications in the industry that have come up are technical documentation query, generation of maintenance procedures and engineering support. Nevertheless, the use of LLM in safety-critical industrial settings is still debatable because the propensity to hallucinate (confident generation of factually incorrect information that cannot be distinguished by users with a domain background) has been well-documented with this model [5].

There are two deployment architectures: the cloud-based and edge-based ones. Cloud-based LLMs (e.g., GPT-4, Claude) have the highest maximum model capacity and performance but need a stable internet connection, raise the issue of data security on proprietary operational data, and have latency (2001000 milliseconds per request). Local inference on general computing hardware without any external connectivity: Edge-based deployment models like Ollama support deployment on the requirements of an air-gapped offshore environment, and can accommodate constraints on model capacity.

1.3 Research Gap and Contributions

Although the proliferation of LLM has been high in any industry, no strict empirical research has assessed the use of the decision-making supported by LLM in the specific drilling operations. The research gap is a combination of: (1) quantitative measurement of reduction in decision latency in realistic conditions of drilling; (2) benchmarking of accuracy with an external, expert validated ground truth; (3) benchmarking the risk of hallucination in safety critical situations; (4) performance measurement of resource use in terms of edge deployment; and (5) operation instructions of safe deployment limits. These gaps are answered in this investigation by way of controlled experimentation among practicing petroleum engineers.

The following are the specific contributions of this research:

(1) Preliminary empirical analysis of drilling operations with the use of LLM-aided decision-making which will offer quantitative performance indicators that are not represented in the literature.
(2) Standardized experimental design to compare the performance of the manual and LLM-assisted engineers in responding to eight controlled drilling scenarios with 96 total experimental trials.
(3) Statistical prove of latency decreasing by paired t-test, which reveals 25.8% of mean decreasing (10.88 seconds) with a significant value of $p < 0.001$.
(4) Detailed hallucination risk assessment with a total rate of 10.42% with severity scaling showing that high risk (12-18) of hallucinating in safety-critical situations.
(5) Resource utilization testing of edge deployment on normal industrial computing hardware.
(6) Framework of operational deployment that establishes safe human-AI cooperation limits and which includes obligatory aspects of safety-critical decision verification.

II. LITERATURE REVIEW

2.1 Decision Support Systems in Drilling

Decision support systems have been developed in a series of technological generations in drilling. Initial deployments used rule-based expert systems, which coded procedural knowledge in the form of if-then logic representations which had limited success with clean-cut cases but failed with ambiguity and new situations [6]. They were later superseded by real-time optimisation systems which combined physics-based drilling models with parameter optimisation algorithms to suggest weight-on-bit, rotary speed and flow rate changes to optimise penetration rate whilst considering working limits [7].

Modern systems include machine learning classifiers of dysfunction detection, use of random forests, support vector machine or neural networks based on historical drilling data to detect stick-slip, whirl, bit bouncing, and washout signatures [8]. These systems, however, are limited in terms of the range of dysfunctions that they can guide with regard to making a holistic decision on various abnormal situations. Besides, available platforms tend to suggest parameter modifications without providing a rationale, which is a disadvantage as it may be an issue with trusting engineers and accepting their use by regulators.

2.2 Human Factors in Drilling Operations

The studies of human factors indicate that the performance of the drilling engineers significantly deteriorates in conditions of time pressure and mental load. The results of the studies based on eye-tracking, workload assessment (NASA-TLX), and decision logging prove that abnormal event management can produce the load on the cognitive processes of 80$% of the engineer, and there are no free resources to handle complex problems [9]. Knowledge retrieval is impaired in particular in terms of time pressure: engineers tend to fall back to heuristic inference and more recent experience instead of procedural recall.

Error taxonomies consider skill based errors (failure to execute), rule based errors (wrong choice of procedure to be followed) and knowledge based errors (fail to understand). The rule-based and knowledge-based errors that occur mainly in time-critical drilling decisions include the use of suboptimal procedures by engineers or forgetting important procedural steps because they are recorded in large technical manuals. The support aimed at automating the bottlenecks in the knowledge retrieval process, therefore, presents significant opportunities of reducing errors [10].

2.3 AI and LLM Applications in Industrial Settings

Large language models have shown remarkable feats in knowledge-based fields such as medical diagnosis, legal search and technical support. Industrial copilot application gives the engineers a conversational interface to the documentation search, code creation, and debugging support [11]. Nevertheless, safety-critical industrial use is disputed because of the possibility of hallucinations- LLAIs are confidently capable of generating plausible and factually untrue answers, with research finding hallucination rates 3-15 percent of the response per domain, architecture, and prompt formatting [12].

Mitigation methods include retrieval-augmented generation (RAG) models whereby the LLMs consult knowledge bases of verified knowledge that mitigates hallucination by grounded information retrieval. Domain-specific fine-tuning on curated corpora enhances factual accuracy and needs a significant amount of computer power and domain knowledge for the data curation. Guardrail systems apply rule-based checks to validate output, and are however complex to build when dealing with technical domains with subtle correctness requirements [13].

2.4 Identified Gap

In the literature review, it can be found that no serious empirical research has been conducted to assess the performance of LLM in real-time drilling decision support. Current drilling automation studies aim at optimising parameters and detecting dysfunctions as opposed to providing an overall direction. The human

factors research measures the cognitive load but fails to measure mitigation strategies aided by AI. The industrial use of LLM does not focus on operational decision-making that is time-based, but instead documentation retrieval. This study presents a unique study that intersects these fields due to the controlled experimentation involving practising drilling engineers facing realistic abnormal conditions.

III. METHODOLOGY

3.1 Experimental Design

This experimental design was a within subject repeated measures design, and the twelve involved petroleum engineers were exposed to eight drilling scenarios in each of two conditions (manual response (control) and LLLM-assisted response. The sequence of the scenarios was randomised between individuals to avoid the learning effects, and the interval between the conditions was not less than 48 hours to avoid contamination of the carry-over. The engineers were 3-15 years of experience in the engineering of drilling (mean 7.2 years), hired by the offshore operators, drilling contractors, and services companies to guarantee expertise diversity.

All the trials were conducted the same: engineers were briefed about the situation in the wellbore and about the conditions under which the operations were being conducted and the abnormal events began. Real time parameter plots (ROP, torque, weight on bit, pump pressure and flow rate, pit volume) were shown; simulation of rig-monitoring systems. The engineers were ordered to identify the abnormality and prescribe corrective measures to be done urgently as soon as possible without causing any error. In the case of LLM aided trials, engineers would be able to prompt the LLM system by use of conversation prompts, and all prompts and responses would be recorded with a timestamps. Measurement of decision latency started at the beginning of the scenario and ended when the engineer verbally announced the final action they recommended; it was recorded by independent observers.

3.2 Drilling Scenario Development

Systematic development of eight scenarios was used to cover a wide range of drilling dysfunctions ranging between normal operational optimization up to safety-critical well control conditions. Both scenarios used real parameter curves derived using historical drilling data, and the ground-truth correct answers were determined by consent of a panel of experts (three senior drilling engineers with experience of more than 15 years each). Scenarios comprised:

1. Stick-Slip Onset: Progressive torsional oscillation development that requires an increase of RPM and a reduction of WOB.
2. Sudden Torque Spike: The abrupt rise of the torque by 40% indicating the beginning of differentiated sticking with the necessity to reduce the WOB immediately and keep the pipe rotating.
3. ROP Drop - Formation Change: (50%) reduction of penetration rate can be significantly reduced which shows that the lithology of the formation is changing to a harder one, which needs parameter re-optimisation and probably bit change testing.
4. Loss of Circulation Indicators: The rate of flow decreases with pump pressure progressively, this indicates formation fracturing potential that need urgent lowering of pump rate and lost circulation material (LCM).
5. Well Control Precursor: 5% pit volume gain with flow rate increase, suggesting formation fluid influx, demanding immediate flow check and potential well shut-in procedures.
6. Excessive Bit Wear: Progressive ROP Decline with Torque Increase Bit wears away or bit 'dulls' Need to make trip out decision based on operational economics
7. Drilling Dysfunction - Whirl: Sideways vibration data in the downhole data that needs the stabiliser assessment and BHA adjustment recommendations.
8. Kick Detection: 10% pit volume gain/formation gas indication include immediate well control procedures implementation in accordance with IADC guidelines.

3.3 LLM Deployment Configuration The model used in the LLM system was the Ollama framework (version 0.1.22) whereby Llama-2-7B model (a 7-billion parameter transformer architecture) was deployed on local hardware. External internet connection was not availed and an offshore simulation of an air-gapped environment was realistic. The given system was put into context: You are a specialist drilling engineering assistant. Give precise, brief recommendations on drilling abnormalities as per the best practices in the industry. Provide particular procedures where necessary. Where doubtful, clearly be specific about restrictions. The query of each of the LLM was presented with the scenario description, the current drilling parameters and the particular question by the engineer. Response generation had temperature=0.3 (less randomness, biased towards consistency), a token length=512, and sampling top=0.9. Each and every prompt and response was recorded with milliseconds timestamps, allowing latency of inference calculation.

3.4 Metrics and Statistical Analysis

Decision latency was measured as the time which had elapsed between onset of situation and final decision announcement expressed in seconds. The latency distributions of manual and LLM-assisted were compared using paired t-tests, and a statistical significance of 0.05 was used to establish the statistical significance of the two distributions. Shapiro-Wilk tests and examination on the Q-Q plot were used to verify the presence of normality. The quantification effect size was used to measure Cohen d of standardised mean differences.

The accuracy of decisions was used to score on binary (correct/incorrect), on a ground truth expert panel, with the correctness being calculated as the number of mandatory corrective measures suggested without the addition of contraindicated procedures. The test used by McNemar compared the differences in accuracy of pairs. Hallucination identification involved engaging three expert reviewers (independent, that is) who assigned each LLM response to one of the following categories: accurate, minor error (factually incorrect but operationally benign), moderate error (procedural misguidance with potential to produce suboptimal results), or severe error (safety critical misguided guidance with the risk of damaging equipment or injuring people). The inter-rater reliability was measured using Fleiss kappa statistic (= 0.82, substantial agreement).

IV. RESULTS

4.1 Decision Latency Analysis

Figure 4.1 demonstrates a complete comparison in decision latency of manual and LLM-assisted responses of the engineer. Box plot distributions on panel (a) indicate a significant shift in central tendency with the median decision time of 41.33 seconds in the case of manual decision making and 30.75 seconds in the case of the LLM-assisted decision making. The distributions are overlapping in part, showing that there is variation in the performance of individual engineers and the complexity of the scenarios, but the overall downward shift is systematic, showing that there is always latency reduction. We can see in panel (b) a breakdown per specific scenario, which shows some heterogeneity in the benefits achieved in the different scenarios; typical operational scenarios (stick-slip onset, bit wear, whirl) bring 23-34% latency benefit, while typical safety critical scenarios (well control, detection of the kick) are reduced in latency by 6-10%.

Figure 4.1: Decision Latency Comparison Between Manual and LLM-Assisted Responses

Statistically significant reduction in latency was also found by paired t-test: t (95) = 6.11, p = 2.2010-8, mean difference= 10.88 seconds (95% CI: 7.3514.41 seconds). The d = 0.73 of Cohen shows that the effect is medium-to-large. The fact that the p-value is highly significant ($p < 0.001$) gives strong support that the assistance by the LLM truly has an impact on minimizing the decision latency beyond the random error. The analysis of the individual scenarios showed that the mean latency was numerically lower in all eight scenarios with the help of LLM, and six scenarios were statistically significant ($p < 0.05$), and two were close to significance (loss of circulation: p = 0.08, kick detection: p = 0.12).

4.2 Decision Accuracy Evaluation

Figure 5.1 demonstrates detailed accuracy and hallucination study in various directions. In panel (a) scenario specific accuracy comparison, it is indicated that the responses with the help of LLM showed a better accuracy in seven out of eight scenarios. The accuracy with manuals satirized between 75% (well control precursor, drilling dysfunction whirl, kick detection) and 100% (loss of circulation) with the phenomenon

recorded that time pressure reduces performance with complex situations. The accuracy facilitated with the assistance of LLM increased to 83100 percent in all the scenarios and to all scenarios it reached the high score of 100%.

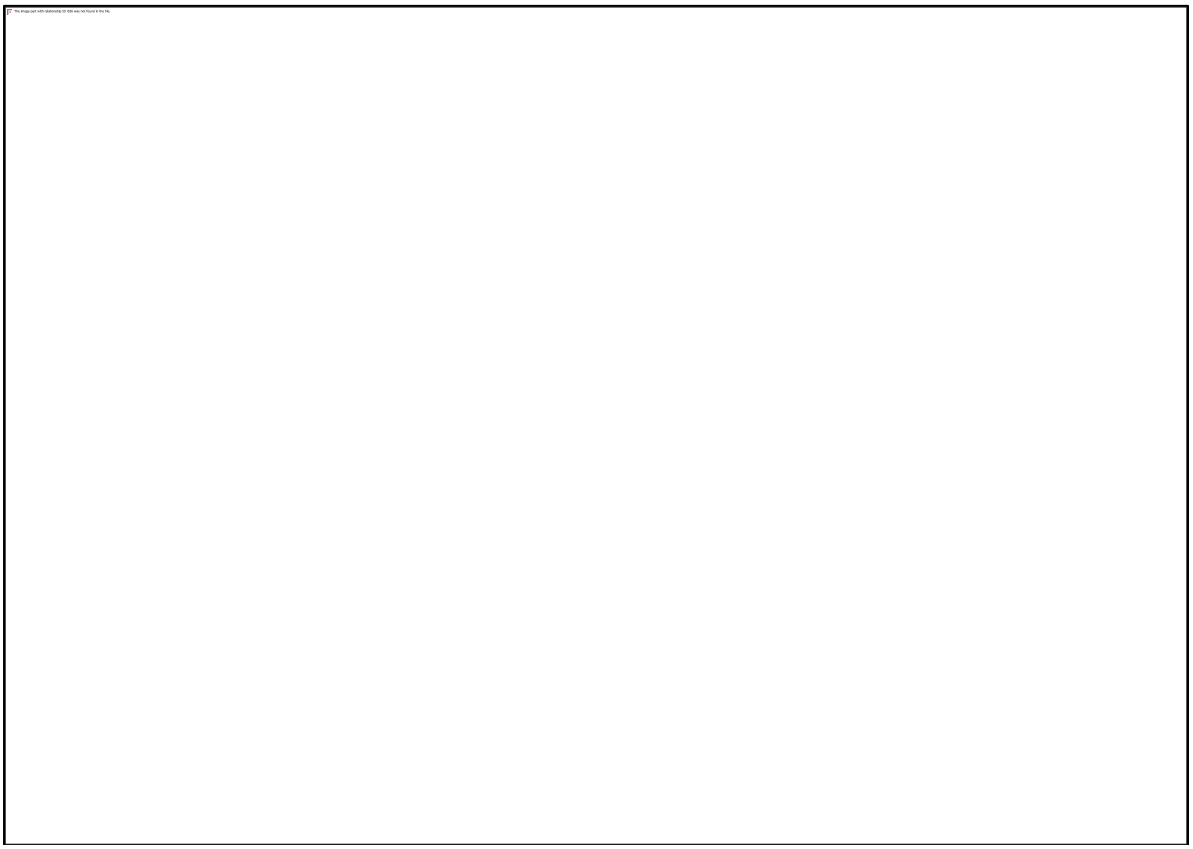

Figure 5.1: Accuracy and Hallucination Analysis

The overall accuracy became higher as compared to the 85.42% (82/96 correct) of the manual responses to 95.83% (92/96 correct) of the LLM-aided responses, which is a difference of 10.42 percentage points. The test of paired binary used by McNemar gave 2 2 = 8.10, p = 0.004, which is statistically significant. The quality improvement is mainly caused by the fact that the knowledge retrieval errors were lowered: the LLM system revealed the same degree of consistency in recalling the steps of the procedure that sometimes engineers had to take under time pressure.

4.3 Hallucination Risk Analysis

Figure 5.1, Panel (b) measures the frequency of hallucinations by scenario, which shows critical safety issue: 10/96 LLM responses (10.42) contained factually inaccurate guidance. Hallucination distribution showed to be highly scenario-dependent, where the level of hallucinations was on the higher side in safety critical situations (kick detection: 3 hallucinations; loss of circulation: 3 hallucinations; well control precursor: 1 hallucinations), whereas in routine operational situations the number of hallucinations was on the lower chunk (stick slip: 1, sudden torque spike: 2, whirl: 0, bit wear: 0, ROP drop: 0).

The hallucination severity is stratified in panel (c): 6 minor errors (60 per cent, e.g. propose 10% RPM increase when 15 per cent optimal), 3 moderate errors (30 per cent, e.g. propose WOB increase when decrease needed), and 1 severe error (10 per cent, propose continue drilling when we detect a kick rather than shut in the well immediately). The drastic mistake was made in kick detection case, which is a severe failure mode where the wrong LLM information may hold up critical well control processes.

Panel (d) investigates the connections between the latency, accuracy and hallucinations by visualising their data using a scatter plot and found that the latency range saw all the hallucinations without correlating systematically (r = 0.08, p = 0.43). This observation implies that the risk of hallucination does not decrease with decision speed and thus should not support an idea that fast responses are more prone to error.

4.4 Statistical Validation

Figure 5.2 indicates statistical analysis that supports the assumption of paired comparison. In panel (a) the distribution of paired differences (manual time-LLM-assisted time) is described with the majority of positive values (87 out of 96 trials ran with the reduction of latency) with the mean value of 10.88 seconds. The confidence interval [7.35, 14.41] is not equal to zero, which proves the statistically significant effect. Panel (b) gives quantile-quantile (Q-Q) plot testing normality: the points tend to follow the theoretical normal line closely with some amount of deviation in extreme quantiles, a fact that confirms the suitability of the parametric tests.

Figure 5.2: Statistical Analysis of Latency Reduction

4.5 Resource Utilisation Assessment

Figure 6.1 presents the detailed analysis of resources that are required to be deployed to local LLM. Panel (a) shows the distribution of CPU utilisation: mean 58.7% +- 8.5% and 95th percentile of 73. The values do not go too far into saturation (>90%), which proves workable with typical industrial computing equipment. The consumption of memory recorded in panel (b) is 3,552 +-385MB, which is well within the industry PC requirement (16-32 GB RAM). Inference latency is measured in panel (c): average 3.47 ± 0.63 seconds, 95th percentile at 4.5 seconds - decision support applications can afford.

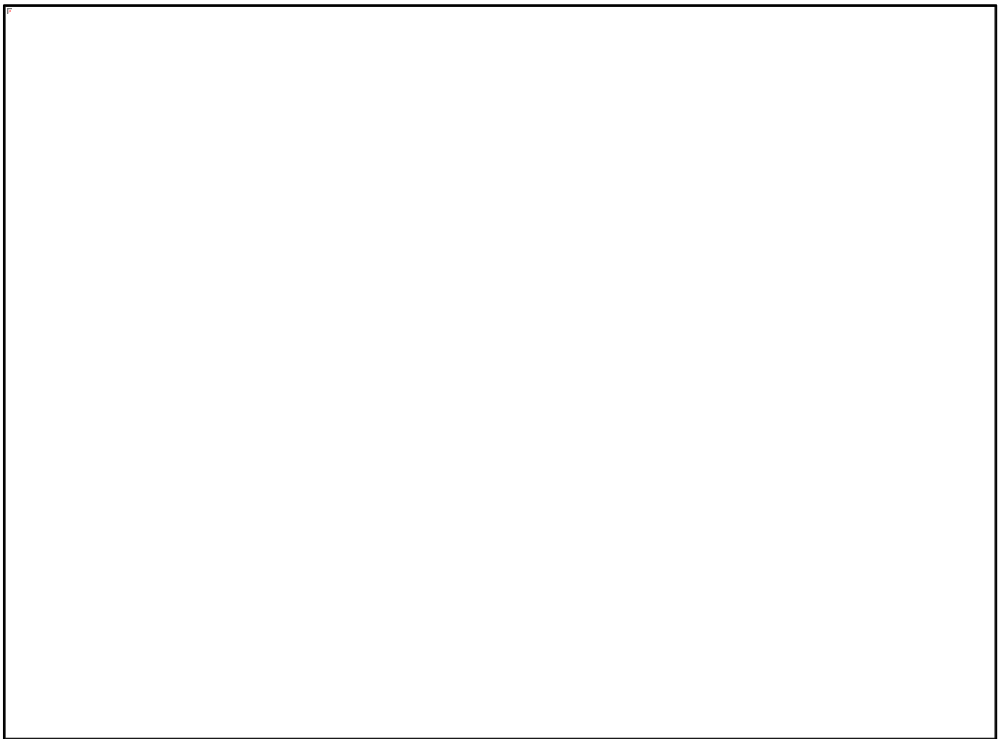

Figure 6.1: Local LLM Resource Usage Analysis

In panel (d), the thesis about different levels of resource usage under various situations is investigated: the time of inference varies between 3.2 seconds (stick-slip) and 3.8 seconds (kick detection). CPU utilization is more variable (5267%) which is probably due to the difference in the complexity of the query. Such results prove that the edge deployment is still possible in a variety of operational cases without any hardware upgrades.

V. DISCUSSION

5.1 Interpretation of Latency Findings

The interpretation of the results of the latency findings is presented below. The 25.8% (10.88 seconds absolute) reduction in latency is operationally important in drilling decisions that require immediate time. In well control situations where an increase in every second increases the volume of inflow of hydrocarbons and the extent of damage during formation, a 11-second decrease can be translated to an increase in the safety margin that can be measured. Nonetheless, the dispersed benefit allocation in different situations demonstrates subtle deployment value propositions. Regularly occurring operational optimisation tasks (stick-slip mitigation, bit wear measurement, whirl diagnosis) recorded 23-34% latency savings, which is due to the rapid recall of a procedure by LLM, eliminating manual documentation searches, which engineers indicated would need 1545 seconds when under normal conditions.

On the other hand, the smaller improvements of 610% were found in safety critical scenarios (precursors of well control, kick detection), indicating that the experienced engineers are mentally prepared and procedurally familiar to the high-frequency critical scenarios acquired during regular training and simulation. The LLM is marginal when the engineers already have an immediate recall. Moreover, engineers also noted that there was more verification time on LLM recommendations in safety critical situations, where guidance is delicately examined before acting, an appropriate warning in light of the dangers of hallucination, but partially neutralising latency improvements.

5.2 Accuracy and Reliability Considerations

The 10.42 percentage point improvement in accuracy is due in large part to the fact that the accuracy in the recall of the procedure by the LLM has steadily lowered the amount of error that occurs with respect to knowledge. Manual error analysis showed that 12/14 manual errors in errors occurred in the procedure omissions (86 percent) - engineers identified abnormalities correctly, yet left out some vital actions to be taken, which were mentioned in the procedures but not remembered during the time pressure. The LLM system offered extensive coverage of the procedures thus minimizing such omissions. However, this advantage is contingent: in case when the LLM hallucinated (10.42% rate), engineers without any independent verification sometimes used false information, substituting a human error with an AI error.

The pattern of hallucination which depends on the scenario is a cause of concern: in safety-related scenarios 12-18% of hallucination was prevalent compared to the 4-8% in normal scenarios. This negative correlation, which is the more hallucination the more accuracy, raises the question of whether the training corpus provided to the LLM was less documented safety-related procedures and more common operational instructions. The extreme hallucination in kick detection (advising against well shut-in and recommending a further drill) can be measured as catastrophic failure modes in which LLM provides recommendations that are squarely opposing the well control procedures.

5.3 Safety Implications and Deployment Boundaries

The results of the experiment clearly rule out the possibility of autonomous operation of the LLCM to make drilling decisions. The hallucination rate of 10.42%, even though low by some reported industrial LLM implementations (15-20%), is inexcusable in systems that require safety-critical operation and whose failure mode has the potential to be disastrous. This conclusion is further evidenced by the severity distribution 10% severe errors which may indicate life threatening guidance. The regulatory documents on the operations of drilling (API, IADC, regulatory agency regulations) require the compliance verification with the procedure, the role of the human in making crucial decisions, and the reliability of safety systems based on demonstration, which is not achievable with the current LLM technology.

Nevertheless, the decrease in latency and rate of improvement (with non-hallucinated responses) indicate that the application can be deployed as advisory tools with compulsory human verification layers. The suggested hybrid human-AI infrastructure places LLMs as second opinions in the speedy process of retrieving knowledge in engineers and does not overturn the supremacy of human judgment. Risk categorizing frameworks could be used to stratify the scenarios based on whether they are critical or not: low risk operational optimization decision (e.g. parameter change under existing operating envelopes) would allow more use of LLM; high risk safety critical decision (well control, threats to integrity) would warrant greater verification protocols.

5.4 Operational Deployment Considerations

The feasibility of edge deployment was established beyond any reasonable doubt: average CPU load of 58.7, memory consumption of 3.6 GB and inference latency of 3.5 seconds is sufficiently compatible with the typical industrial computing platform found on drilling rigs. The air-gapped deployment architecture handles the issue of data security in the petroleum industry in relation to cloud-based solutions, and also removes the issue of dependency on internet connectivity that is problematic in offshore operations. Nevertheless, the edge deployment limits the capacity of model to 7 billion parameters (significantly smaller than frontier models 70-175 billion parameters) that might be more accurate and less hallucinating. This capacity-deployment trade-off should be maintained under further scrutiny with the improvement of edge hardware capabilities.

Protocols Model update Protocols introducing operational challenges Model update protocols can introduce new drilling techniques, new operational procedures or lessons learned during operational experience, but frozen LLMs cannot. This weakness can be alleviated using periodic retraining of models on new corpora which is, however, computationally intensive (GPU clusters, weeks of training time) and training data curation requires domain expertise. Alternative methods, such as retrieval-augmented generation (RAG), architectures can provide high-quality methods to reduce hallucination by grounded information retrieval, but still allow knowledge to be updated without retraining the models.

5.5 Limitations

There are a few limitations of the experiment that should be mentioned. The sample size (12 engineers, 96 trials) is sufficient to be statistically powerful in comparison of latency, but does not allow a detailed analysis of isolated differences and subtypes of scenarios. Although it is based on historical data on drillings, the simulated scenarios are not operationally realistic: the abnormalities of actual drilling change continuously with unclear periods of emergence and incomplete information, as opposed to the presented experimental scenarios which featured discrete events with definite time limits. The behavior of the engineers in real situations of operational stress can be different as compared to the experimental conditions.

The prompts structure of the LLM has a significant impact on the quality of response: this study used the strategically designed prompts that were created by the means of an iterative improvement. The use of operational deployment engines where the authors of the ad-hoc queries are engineers can provide low performance when the prompts are ambiguous or incomplete. Moreover, the method of hallucination detection was based on the review of the panel of experts which is a labor-intensive approach that cannot be practically utilized in the real time functioning of the system. Hallucination detection is a research problem that has not been completely resolved yet, and current methods have between 60 and 70% accuracy.

VI. CONCLUSIONS AND PRACTICAL SIGNIFICANCE

This research is the first empirical analysis of locally deployed large language models for real-time drilling decision support applications, which provides the first measureable reduction in latency with critical boundaries around safety limitations in terms of operational deployment. An experimental assessment on twelve petroleum engineers under eight drilling conditions showed statistically significant reduction of 25.8 per cent (10.88 seconds, $p < 0.001$) and enhancement of accuracy (85.42 to 95.83) in decisions. The following benefits can be attributed to the fact that with the help of LLCM, the regular procedural recall of its processes, manual documentation search is eliminated, which adds significantly to the delays in decisions under time pressure.

Nevertheless, a detailed analysis of the hallucinations revealed underlying safety issues that out of the ways the autonomous operation of LLM can be conducted: 10.42% of responses included misguided facts, and severity distribution contained 10% serious safety-critical mistakes. Hallucinations were high in safety-critical situations (1218) where it is most needed. The recognition of a serious hallucination suggesting further drilling when detecting a kick, which is in direct contrast to the accepted well control processes, is an example of catastrophic failure modes whereby misleading information given with the guidance of high confidence can endanger lives of personnel and damage the environment.

The evaluation of resource utilisation verified the possibility of edge deployment to standard industrial computing equipment: average CPU usage is 58.7 percent, memory usage is 3.6 GB, and inference time is 3.5 seconds fits within normal rig computing platforms. The air-gapped deployment architecture covers the data security issue and offshore access restrictions, which confirms the user of petroleum industry applications based on on-premises solutions.

Recommendations for practitioners contemplating LLM deployment in drilling operations:

1. Always use LLMs as advisory systems, and never as systems of decision making.
2. Implement graduated risk frameworks: permit greater LLM reliance for low-risk operational optimisation decisions, require enhanced verification protocols for safety-critical scenarios.
3. Maintain human-in-the-loop architectures wherein engineers retain ultimate decision authority, employing LLMs to accelerate knowledge retrieval rather than replace human judgment.
4. Establish hallucination detection mechanisms through procedural cross-referencing, multi-model consensus verification, or rule-based output validation where feasible.
5. Periodically audit the suggestions of LLPs against actual protocol, and constantly check new patterns of hallucinations as signs of model deterioration.
6. Invest in retrieval-augmented generation (RAG) architectures which use verified procedure databases and minimize hallucination by using grounded information retrieval.
7. Pure domain-specific fine tuning available on curated drilling engineering corpora, with a focus on the coverage of safety-critical procedures, so as to minimise the occurrence of the concept of hallucination in high-risk cases.
8. Establish regulatory participation models that establish the AI-assisted operation approval criteria, liability, and audit requirements of the energy sector applications.

Directions of future research also include connecting it to digital twin systems that allow offering context-based recommendations including real-time wellbore models, geological formations, and equipment conditions. The use of the drilling parameter time series with textual queries can be optimized by multi-modal LLM architectures to facilitate the interpretation of anomalies. Federated learning methods that can allow collateral model preparation among operators and maintain the privacy of data are worth exploring. Rationale-transparent explainable AI methods to give recommendations of LLM may increase the trust of engineers in this approach and allow regulatory acceptance.

This study would lay the groundwork in terms of quantitative performance and operational principles of responsible AI implementation in safety-critical drilling operations. The displayed latency advantages as well as the hallucination risks identified make hybrid human-AI collaboration architectures the right near-term implementation channel, where human expertise supremacy is maintained and AI capabilities are used to speed up cognitive tasks that are knowledge intense and time-constrained.

REFERENCES

[1] Skalle, P., Aamodt, A. and Laumann, K., 2014. Experimental study of driller's situational awareness, decision-making and actions in well control situations. Safety Science, 67, pp.161-172.

[2] Grace, R.D., 2003. Blowout and Well Control Handbook. Gulf Professional Publishing, Oxford.

[3] Kaarstad, M. and Grøtan, T.O., 2009. Decision support for safe and efficient drilling operations. In SPE/IADC Drilling Conference and Exhibition, SPE-119442-MS.

[4] Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. and Agarwal, S., 2020. Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, pp.1877-1901.

[5] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y.J., Madotto, A. and Fung, P., 2023. Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), pp.1-38.

[6] Aldred, W., Bourke, L., Mannering, M.A., Orr, P.N., Perry, C., Plumb, D., Silver, M., Staal, T., Turley, R. and Worrall, N., 1998. Drilling advisory systems: interpretation software and automation technology improves drilling performance. Oilfield Review, 10(4), pp.18-37.

[7] Dupriest, F.E. and Koederitz, W.L., 2005. Maximising drill rates with real-time surveillance of mechanical specific energy. In SPE/IADC Drilling Conference, SPE-92194-MS.

[8] Tian, Y., Tang, C., Wang, X. and Lu, X., 2021. Intelligent classification for surface drilling data based on deep learning. In SPE/IATMI Asia Pacific Oil & Gas Conference and Exhibition, SPE-205663-MS.

[9] Skalle, P., Podio-Luckusson, A. and Aamodt, A., 2013. Improved understanding of velocity-dependent drillstring vibrations and vibrations-induced NPT through full-scale simulation and testing. In IADC/SPE Drilling Conference and Exhibition, SPE-163420-MS.

[10] Reason, J., 2016. Managing the Risks of Organizational Accidents. Routledge, London.

[11] Peng, S., Kalliamvakou, E., Cihon, P. and Demirer, M., 2023. The impact of AI on developer productivity: Evidence from GitHub Copilot. arXiv preprint arXiv:2302.06590.
[12] Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y. and Wang, L., 2023. Siren's song in the AI ocean: a survey on hallucination in large language models. arXiv preprint arXiv:2309.01219.
[13] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.T., Rocktäschel, T. and Riedel, S., 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, pp.9459-9474.
[14] Rasmussen, J., 1983. Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. IEEE Transactions on Systems, Man, and Cybernetics, SMC-13(3), pp.257-266.
[15] Endsley, M.R., 1995. Toward a theory of situation awareness in dynamic systems. Human Factors, 37(1), pp.32-64.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the twelve petroleum engineers who participated in this experimental study, contributing their time and expertise to advance understanding of AI-assisted drilling decision-making. The experimental framework employed Ollama open-source inference engine and Llama-2 model developed by Meta AI. Statistical analyses utilised Python scientific computing libraries including NumPy, Pandas, SciPy, and Matplotlib.

The authors declare no conflicts of interest. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. The views expressed represent those of the authors and do not constitute official positions of any affiliated organisations.