

Projeto de Pesquisa para Doutorado

Vitor Brandão Sabbagh

Novembro de 2025

1 Título

Mapeando a Fronteira Irregular: Uma Análise Experimental da Capacidade de Agentes Baseados em LLM em Tarefas de Engenharia de Poços

(*Mapping the Jagged Frontier: An Experimental Analysis of LLM-Based Agent Capabilities in Complex Offshore Well Engineering Tasks*)

WMJ: O título (e a proposta como um todo) poderia ser um pouco mais geral focando na fronteira irregular e no impacto em agentes LLM, deixando a tarefa de engenharia de poços como estudo de caso. Não é claro para mim o quanto essa tarefa difere de outras, justificando que a tese seja tão específica assim. O contexto de agentes baseados em LLM é bem amplo, e entender essa nova classe de problemas sob a perspectiva da fronteira irregular me parece mais interessante. O segundo ponto que me chama a atenção no título é “análise experimental” que novamente é restritivo. Acho que tem que ter alguma formulação teórica, um arcabouço, uma lógica mais

2 Introdução e Contextualização

A evolução recente dos modelos de linguagem marca uma transição da IA concebida predominantemente como ferramenta de consulta pontual para sua configuração como agente autônomo [5] [1] **WMJ: ref?** **VBS: feito.** Nessa nova configuração, sistemas de agentes são capazes de decompor problemas, planejar e executar tarefas de múltiplos passos, articulando raciocínio, memória e uso de ferramentas externas [13] [9] **WMJ: ref?** **VBS: feito.** Na prática, essa transição se materializa em soluções comerciais de agentes autônomos, como Manus AI ¹, OpenAI Operator ², OpenAI Deep Research ³ e Genspark AI ⁴.

¹<https://www.manus.ai/>

²<https://openai.com/pt-BR/index/introducing-operator/>

³<https://openai.com/pt-BR/index/introducing-deep-research/>

⁴<https://www.genspark.ai/>

WMJ: Aqui podem ser referências ou footnotes, o que for mais prático. Por exemplo, se forem só as URLs, melhor que seja footnote. **VBS:** feito, entre outras.

WMJ: As referências **tem** que estar no texto usando cite ou citep. Mais ainda, use o bibtex para evitar ter referências desformatadas ou incompletas.

VBS: feito

Contudo, a rápida adoção dessa tecnologia na indústria, especialmente em setores de alto risco como Óleo e Gás (O&G), supera nossa compreensão de seus reais limites **WMJ:** acho que esse tom de O&G como cenário típico fica melhor, mas precisamos de referências que falem desse assunto. O entusiasmo com as capacidades em tarefas “dentro da fronteira” (*inside the frontier*) muitas vezes ofusca a existência de tarefas “fora da fronteira” (*outside the frontier*), onde a IA falha [4]. **WMJ:** Algum artigo nisso? Senão podemos converter em sub-produto do trabalho a caracterização de picos e vales **VBS:** Nenhum artigo com essa caracterização. Acho melhor usarmos a mesma nomeclatura do artigo da HBS, “inside the frontier” e “outside the frontier”. **Fiz o ajuste no projeto.**

[4] introduziu o conceito de “Fronteira Tecnológica Irregular” (Jagged Frontier) para descrever como o desempenho da IA é irregular, alternando entre tarefas “dentro da fronteira” (*inside the frontier*) e tarefas “fora da fronteira” (*outside the frontier*). No entanto, este estudo focou na produtividade de *humanos usando IA* em tarefas de consultoria.

A lacuna que este projeto busca abordar é a falta de conhecimento a respeito da dita fronteira irregular para agentes autônomos em domínios de engenharia de alta complexidade. **WMJ:** tem mais algum além de engenharia de poços? quais as refs? **VBS:** inserido abaixo Embora estudos recentes tenham documentado limitações similares em outras áreas críticas como medicina [8, 7], direito [10, 6], engenharia civil [2, 14], finanças [3] e sistemas *safety-critical* em aeroespacial [12, 11], não se sabe que tipo de tarefas compreendidas na construção de poços de petróleo (um domínio em que falhas podem frequentemente custar vidas e/ou prejuízos materiais relevantes) estariam “dentro da fronteira” (*inside the frontier*) e quais estariam “fora da fronteira” (*outside the frontier*) de performance e assertividade das ferramentas. **WMJ:** aqui acho que vale explicar um pouco mais o contexto de uso e como tarefas dentro e fora da fronteira se manifestariam no problema de engenharia de poços.

WMJ: um exemplo de como agentes LLM são usados em engenharia de poços seria bem ilustrativo

3 Problema de Pesquisa e Objetivos

3.1 Problema Central

A implantação de agentes de LLM em atividades diversas de O&G é dificultada pela falta de um mapa de risco-capacidade. As métricas de *benchmarks* genéricos (ex: MMLU, AgentBench) não capturam as nuances de tarefas de engenharia do mundo real, que envolvem dados ruidosos, raciocínio físico e adesão estrita a normas de segurança. **WMJ:** Acho que está muito específico, vou tentar reescrever, mas apenas para ilustrar o que seria mais abstrato. Entendo que conceitos como mapa risco capacidade são genéricos e aplicados a vários problemas.

WMJ: O uso de agentes de LLM em atividades diversas, incluindo missões críticas em engenharia, é dificultada pela falta de um mapa risco-capacidade. *Benchmarks* genéricos não capturam nuances do mundo real, que envolvem dados ruidosos, raciocínio físico e adesão estrita a normas de segurança.

3.2 Pergunta Principal de Pesquisa (P1)

Onde se localiza, e qual é a topografia, da “fronteira irregular” de capacidade para agentes de LLM no domínio de planejamento e execução de tarefas da construção de poços offshore?

WMJ: Achei a pergunta central de pesquisa muito específica. Vou tentar ampliar novamente

WMJ: A hipótese principal deste trabalho é que é possível identificar a fronteira irregular de capacidade para agentes LLM e caracterizar a sua localização e topografia em domínios relevantes e significativos, em particular planejamento e execução de tarefas de engenharia.

3.3 Perguntas Secundárias (P2-P4)

WMJ: Achei as perguntas secundárias bem niveladas

- **(P2)** Quais características de uma tarefa (ex: necessidade de raciocínio causal, dependência de dados físicos, conformidade regulatória, planejamento temporal) definem se ela está “dentro da fronteira” (*inside the frontier* – sucesso do agente) ou “fora da fronteira” (*outside the frontier* – falha do agente)?
- **(P3)** Como diferentes arquiteturas de agentes (ex: LLM “puro” vs. RAG vs. Agentes de Planejamento) navegam por essa fronteira?
- **(P4)** É possível desenvolver um *framework* para identificar tarefas “fora da fronteira” *a priori*, permitindo a implantação segura de agentes em tarefas “dentro da fronteira”?

WMJ: Acrescentaria uma pergunta da aplicação desse framework em cenários relevantes e significativos

3.4 Objetivo Geral

Mapear e caracterizar a fronteira irregular de capacidade de agentes de LLM no domínio de engenharia de poços, identificando os fatores que determinam o sucesso e a falha em tarefas complexas.

WMJ: Ficou meio parecido com a minha proposta de problema de pesquisa (me desculpe). Eu iria além de engenharia de poços.

3.5 Objetivos Específicos

1. **OE1:** Desenvolver uma taxonomia de tarefas representativas da construção de poços offshore, classificadas por tipo de cognição e complexidade.
2. **OE2:** Projetar e implementar um *benchmark* experimental baseado nesta taxonomia, com métricas de avaliação e *ground truth* definidos por especialistas.
3. **OE3:** Avaliar sistematicamente diferentes arquiteturas de agentes de LLM neste *benchmark*.
4. **OE4:** Analisar os resultados para construir o “mapa” da fronteira irregular, correlacionando tipos de tarefa com o desempenho dos agentes.
5. **OE5:** Propor um *framework* de decisão para a implantação segura de agentes na indústria de O&G, baseado nas descobertas.

4 Justificativa e Relevância

Este projeto possui relevância em três eixos:

1. **Contribuição para a Ciência da Computação (Teórica):** Estende a teoria da “Fronteira Irregular” do campo de Interação Humano-Computador (HCI) para o campo de Agentes Autônomos. Além disso, critica e avança o estado da arte em *benchmarking* de agentes, saindo de tarefas genéricas para domínios industriais complexos.
2. **Contribuição para a Indústria de O&G (Prática):** Fornece o primeiro estudo rigoroso sobre o que agentes de IA podem (e, crucialmente, *não podem*) fazer com segurança na engenharia de poços. Isso desbloqueia ganhos de eficiência (em tarefas “dentro da fronteira”) e previne falhas catastróficas (em tarefas “fora da fronteira”).

3. **Originalidade:** A intersecção de Agentes LLM, a teoria da “Jagged Frontier” e o domínio de O&G *onshore/offshore* é inteiramente nova na literatura.

5 Fundamentação Teórica

A tese será fundamentada em quatro pilares:

1. **Agentes Baseados em LLM:** Arquiteturas e paradigmas (RAG, ReAct, CoT, Multi-Agentes). Como eles funcionam, planejam e usam ferramentas.
2. **Avaliação de Agentes (Benchmarking):** Estado da arte (ex: Agent-Bench, GAIA, MT-Bench). Análise de suas limitações para tarefas industriais/engenharia.
3. **Produtividade e Limites da IA:** O *paper* seminal de Dell'Acqua et al. (2023) sobre a “Fronteira Irregular”.
4. **Engenharia de Poços e IA:** Aplicações atuais de machine learning em O&G e a lacuna existente na aplicação de *agentes generativos* para atividades diversas do setor.

6 Metodologia Proposta

Este projeto empregará uma **metodologia de pesquisa experimental quantitativa e qualitativa**, dividida em quatro fases:

Fase 1: Definição do Domínio e Taxonomia de Tarefas (OE1)

- **Fonte de Dados:** Análise documental de Normas Técnicas, Padrões Operacionais, Relatórios de Situação Operacional, Lições Aprendidas, Alertas Técnicos e Relatórios Diários de Perfuração (DDRs/Boletins Diários de Operação - BDOs).
- **Amostragem:** Criação de um *dataset* de 20-40 tarefas representativas.
- **Classificação (Taxonomia):** As tarefas serão classificadas por eixos:
 - *Tipo de Ação:* Extração de Informação, Síntese, Diagnóstico, Planejamento, Verificação de Conformidade.
 - *Domínio de Conhecimento:* Geologia, Fluidos, Mecânica, Regulação.
 - *Complexidade:* Nível de raciocínio causal, temporal e espacial exigido.

Fase 2: Design do Benchmark Experimental (OE2)

- **Plataforma:** Desenvolvimento de um ambiente de teste (sandbox) onde os agentes podem atuar.
- **Ferramentas (Tools):** Disponibilização de “ferramentas” simuladas para os agentes (ex: `buscar_norma_api(id)`, `calcular_volume_anular(diametros)`, `ler_ultimo_ddr()`).
- **Ground Truth:** Definição de critérios de sucesso (o “gabarito”) para cada tarefa, validado por Especialistas no Domínio (SMEs - *Subject Matter Experts*).

Fase 3: Execução Experimental (OE3)

- **Variáveis Independentes:** Arquitetura do Agente.
- **Variáveis Dependentes (Métricas):**
 1. *Taxa de Sucesso Binário:* Completou a tarefa com sucesso?
 2. *Qualidade da Resposta:* Avaliação cega (1-5) por SMEs.
 3. *Eficiência:* Custo (tokens), passos de raciocínio.
 4. *Robustez:* O agente “alucina” ou falha?

Fase 4: Análise e Mapeamento da Fronteira (OE4, OE5)

- **Análise Quantitativa:** Correlação estatística entre as *características da tarefa* (da Fase 1) e as *métricas de desempenho* (da Fase 3).
- **Análise Qualitativa:** Análise de causa-raiz das falhas (tarefas “fora da fronteira”). O agente falhou por falta de conhecimento (RAG falho), raciocínio (LLM falho) ou planejamento (arquitetura falha)?
- **Resultado:** O “mapa” da fronteira e o *framework* de decisão.

7 Cronograma Preliminar (48 meses)

- **Ano 1 (Meses 1-12):**
 - Revisão de Literatura aprofundada.
 - Disciplinas obrigatórias.
 - Execução da Fase 1 (Taxonomia de Tarefas).
 - Definição do projeto de tese (Exame de Qualificação).
- **Ano 2 (Meses 13-24):**

- Execução da Fase 2 (Desenvolvimento do Benchmark).
- Testes-piloto.
- Artigo de revisão ou *position paper* sobre o *benchmark*.
- **Ano 3 (Meses 25-36):**
 - Execução da Fase 3 (Bateria principal de testes e avaliação).
 - Coleta de dados (avaliação pelos SMEs).
 - Início da Fase 4 (Análise).
 - Submissão de artigo para conferência principal (ex: NeurIPS, ICML, ou conferência de O&G como a OTC).
- **Ano 4 (Meses 37-48):**
 - Conclusão da Fase 4 (Desenvolvimento do Framework).
 - Redação da Tese.
 - Submissão de artigo em periódico.
 - Defesa.

8 Referências Bibliográficas Preliminares

WMJ: Coloquei a primeira ref em vitor.bib e citei no texto. Acho que deve ser feito com todas e esse itemize e mesmo essa seção perdem sentido

- Dell'Acqua, Fabrizio, et al. (2023). *Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality*. Harvard Business School.
- Yao, S., et al. (2023). *ReAct: Synergizing Reasoning and Acting in Language Models*.
- Lewis, P., et al. (2020). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*.
- Zeng, Y., et al. (2024). *AgentBench: Evaluating LLMs as Agents*.
- Manus AI. (2024). *Manus AI: Autonomous AI agents for complex workflows*. Recuperado de <https://www.manus.ai>.
- OpenAI. (2024). *OpenAI Operator: Building and orchestrating AI-native applications*. Recuperado de <https://platform.openai.com>.
- OpenAI. (2024). *Deep Research: Autonomous research agent by OpenAI*. Recuperado de <https://platform.openai.com>.
- Genspark AI. (2024). *Genspark AI: Autonomous AI agents for research and knowledge work*. Recuperado de <https://www.genspark.ai>.

Referências

- [1] LLMs revolutionized AI: LLM-based AI agents are what's next, February 2021.
- [2] Various Authors. Artificial intelligence in civil engineering: Emerging applications and opportunities. *Frontiers in Built Environment*, 11:1622873, 2025. Discute alucinações de LLMs em engenharia civil e desafios de adoção.
- [3] Fangyi Chen et al. Ai-driven financial analysis: Exploring chatgpt's capabilities and challenges. *Journal of Risk and Financial Management*, 17(3):60, 2024. GPT-4o falha em tarefas financeiras complexas e especializadas.
- [4] Fabrizio Dell'Acqua, Edward McFowland III, Ethan R Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelier, and Karim R Lakhani. Navigating the jagged technological frontier: Field experimental evidence of the effects of ai on knowledge worker productivity and quality. *Harvard Business School Technology & Operations Mgt. Unit Working Paper*, (24-013), 2023.
- [5] Mohamed Amine Ferrag, Norbert Tihanyi, and Merouane Debbah. From llm reasoning to autonomous ai agents: A comprehensive review. *arXiv preprint arXiv:2504.19678*, 2025.
- [6] Saeid Gogani-Khiabani et al. An llm agentic approach for legal-critical software: A case study for tax prep software. In *Proceedings of the 48th IEEE/ACM International Conference on Software Engineering (ICSE)*, 2026. Agentes LLM para software legal-crítico; documenta falhas em software tributário.
- [7] Rami Hatem, Brianna Simmons, and Joseph E Thornton. A call to address ai “hallucinations” and how healthcare professionals can mitigate their risks. *Cureus*, 15(9):e44720, 2023.
- [8] Yubin Kim, Hyewon Jeong, Shan Chen, Shuyue Stella Li, Chanwoo Park, Mingyu Lu, Kumail Alhamoud, Jimin Mun, Cristina Grau, Minseok Jung, Rodrigo Gameiro, Lizhou Fan, Eugene Park, Tristan Lin, Joonsik Yoon, Wonjin Yoon, Maarten Sap, Yulia Tsvetkov, Paul Liang, Xuhai Xu, Xin Liu, Chunjong Park, Hyeonhoon Lee, Hae Won Park, Daniel McDuff, Samir Tulebaev, and Cynthia Breazeal. Medical hallucinations in foundation models and their impact on healthcare. 2025.
- [9] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Pinter, Tim Fan, Patrick Lewis, Douwe Kiela, and Tim Rocktäschel.

Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33, 2020.

- [10] Qiaosi Li, Yuxin Duan, Zheng Gu, Hao Zhu, et al. (a)i am not a lawyer, but...: Engaging legal experts towards responsible llm policies for legal advice. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. ACM, 2024. Estudo com 20 especialistas jurídicos sobre limitações de LLMs em aconselhamento legal.
- [11] Yixiang Liu et al. An empirical study of the code generation of safety-critical software. *Applied Sciences*, 14(3):1046, 2024. Geração de código por IA em domínios nuclear, aviação, automotivo e ferroviário.
- [12] NASA. Examining proposed uses of llms. Technical Report NASA/TM-20250001849, National Aeronautics and Space Administration, 2025. Preocupações da NASA sobre uso de LLMs em desenvolvimento safety-critical.
- [13] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Sascha Shafran, Karthik Griffiths, Ruixin Cao, and Karthik Narasimhan. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2023.
- [14] Zhe Zheng et al. A review of llms and their applications in the architecture, engineering and construction (aec) industry. *Artificial Intelligence Review*, 58:241, 2025.