

Navegando em Dados Oceanográficos com Agentes Baseados em LLM: Uma Demonstração das Capacidades de IA na Plataforma SADD*

Almir Monteiro¹, Cláudio Alves¹, Carla Lage³, Fernando Monteiro³, Janice Trotte³
José Roberto Blaschek³, Eduardo Bezerra^{1,2}, Pedro Henrique González¹

¹Programa de Engenharia de Sistemas e Computação
Universidade Federal do Rio de Janeiro (UFRJ) – Rio de Janeiro, RJ - Brasil

{amonteirojr, claudioalves, pegonzalez}@cos.ufrj.br

²Escola de Informática e Computação
Centro Federal de Educação Tecnológica Celso Suckow da Fonseca - Cefet/RJ

ebezerra@cefet-rj.br

³Instituto Nacional de Pesquisas Oceânicas

{janice.trotte, fernando.monteiro, carla.lage}@inpo.org.br, blaschek@gmail.com

Abstract. *The storage, provision and analysis of oceanographic data involve complex workflows that demand domain expertise and technical familiarity with the underlying systems. This paper presents the integration of LLM-based AI Agents into the Sistema de Armazenamento e Disponibilização de Dados Oceanográficos (SADD) modules, covering submission, curation, and data analysis activities. In the submission module, the agent autonomously suggests variable indexing by mapping dataset fields to the platform's search vocabulary, with the user retaining control over the final decision. In the analysis module, two operating modes are supported: a non-specialist mode, in which users without programming knowledge can request oceanographic visualizations through natural language; and a specialist mode, in which oceanographers interact with Python code in an online Jupyter Notebook through natural language commands, accelerating tasks that would otherwise require manual coding. This demonstration showcases how LLM-based agents can lower the barrier to entry for non-expert users while simultaneously enhancing the productivity of experienced ones within a real-world scientific data platform.*

Resumo. *O armazenamento, a disponibilização e a análise de dados oceanográficos envolvem fluxos de trabalho complexos, que demandam conhecimento de domínio e familiaridade técnica com os sistemas subjacentes. Este artigo apresenta a integração de Agentes de IA baseados em LLM nos módulos do Sistema de Armazenamento e Disponibilização de Dados Oceanográficos (SADD), abrangendo atividades de submissão, curadoria e análise de dados. No módulo de submissão, o agente sugere de forma autônoma a indexação de variáveis ao mapear os campos do conjunto de dados para o vocabulário de*

*Link para o vídeo de demonstração: <https://youtu.be/cm9ijC6YRRY>

busca da plataforma, mantendo com o usuário o controle sobre a decisão final. No módulo de análise, são suportados dois modos de operação: um modo para não especialistas, no qual usuários sem conhecimento em programação podem solicitar análises e visualizações oceanográficas por meio de linguagem natural; e um modo especialista, no qual oceanógrafos interagem com código Python em um Jupyter Notebook online por meio de comandos em linguagem natural, acelerando tarefas que, de outra forma, exigiriam codificação manual. Esta demonstração evidencia como agentes baseados em LLM podem reduzir a barreira de entrada para usuários não especialistas, ao mesmo tempo em que aumentam a produtividade de usuários experientes em uma plataforma real de dados científicos.

1. Introdução

O armazenamento, a disponibilização e a análise de dados oceanográficos envolvem fluxos de trabalho complexos, demandando conhecimento de domínio e familiaridade com os processos e sistemas relacionados [Howe et al. 2008]. No que diz respeito à submissão dos dados, o processo deve garantir que todas as informações relevantes sejam corretamente preenchidas. Além de dados básicos, informações precisas sobre os metadados relacionados aos registros são de vital importância. Este material, além de enriquecer a base de dados, impacta positivamente outras atividades, como a pesquisa e a análise dos dados. De forma similar, o sucesso da análise depende de conhecimento de domínio, além do ferramental necessário para a manipulação dos dados brutos, o que pode representar uma barreira para pesquisadores iniciantes. Portanto, ferramentas que auxiliem nessas atividades são indispensáveis, não apenas para reduzir esta barreira de entrada, como também para acelerar a produtividade daqueles que já são mais experientes.

Com o avanço das pesquisas sobre Grandes Modelos de Linguagem (LLM), diversas abordagens e técnicas têm surgido para auxiliar na realização de tarefas. Neste contexto, os chamados “Agentes de baseados em LLMs” [Pati 2025] utilizam a capacidade de raciocínio de LLMs para solução de problemas nos mais variados campos de atuação [Guo et al. 2025, Kessavane 2025, Aylak 2025]. Logo, cria-se a possibilidade do desenvolvimento de recursos que auxiliem nos já citados complexos processos de armazenamento, disponibilização e análise, visando democratizar o acesso a estas atividades.

Este artigo apresenta o desenvolvimento e a incorporação de agentes de IA nos módulos do SADD [Monteiro et al. 2025], abrangendo atividades de submissão, curadoria e análise de dados. Na Seção 2, relatamos os detalhes técnicos destes agentes. Na Seção 3, discutiremos sobre sua integração ao SADD. Por fim, a Seção 4 conclui.

2. Características Técnicas dos Agentes de IA

Conforme [Monteiro et al. 2025], o SADD foi desenvolvido por meio da utilização de *containers* Docker para cada um de seus serviços. Logo, a inclusão de serviços de IA Generativa, sobretudo os agentes de IA, seguiu o mesmo princípio. Em relação ao modelo de linguagem utilizado, os agentes são baseados em modelos da família GPT integrados por meio de *proxies*, *function calling* e *embeddings*, da OpenAI.

No módulo de análise para usuários não especialistas, o agente opera com *function calling* e ferramentas como a *Data Explorer*, para localizar e preparar dados, e a *Ocean*

Viz, para gerar visualizações oceanográficas, combinando geração textual com funcionalidades específicas do SADD. A fim de contextualizar o agente com a tarefa que ele irá executar, orientações devem ser fornecidas, por meio de *prompts*. No caso do agente especializado em vincular metadados da fonte de dados submetida ao vocabulário padrão utilizado pelo SADD, este foi o *prompt* de sistema utilizado:

Você mapeia colunas de um arquivo de usuário para itens do catálogo NERC (tabela Metadados). Retorne SOMENTE JSON válido. Não use markdown. Não use ```.

Da mesma forma, este foi o *prompt* de usuário definido para o agente, onde “*columns*”, “*dictionary_matches*”, “*units_by_column*” e “*candidates*” são variáveis com os conteúdos extraídos dos arquivos submetidos:

```
Retorne exatamente:
{
  "mapping": {
    "coluna": ["valor_real_de_standard_name" | null, "valor_real_de_standard_name_ptBR" | null] | null
  },
  "picks": {
    "coluna": {
      "standard_name": "..." | null,
      "standard_name_ptBR": "..." | null,
      "uri": "..." | null,
      "collection_id": "..." | null,
      "nerc_id": "..." | null,
      "match_type": "exact_nerc_id" | "exact_uri" | "exact_standard_name" | "embedding" | null,
      "distance": number | null
    }
  }
}

Regras:
1) Inclua TODAS as colunas.
2) Se mapping[coluna] não for null, ele DEVE ser um array com os valores reais do candidato escolhido: [candidate.standard_name, candidate.standard_name_ptBR].
3) Nunca retorne as strings literais "standard_name" ou "standard_name_ptBR".
4) Se candidates[coluna] estiver vazio/ausente, retorne mapping[coluna]=null e picks[coluna]={}.
5) Preferência:
a) se existir match_type começando com "exact" -> escolha.
b) senão, escolha menor distance, MAS se top1 e top2 muito próximos (ambíguo), retorne null.
6) Use o dictionary_matches (extraído do PDF) como contexto forte: descrição/unidade/notes ajudam a validar coerência.
7) Preserve o nome original da coluna como chave.
8) Nunca invente uri, collection_id ou nerc_id: copie esses campos somente do candidato escolhido em candidates[coluna].

columns: {columns}
dictionary_matches (PDF): {dictionary_matches}
units_detected: {units_by_column}
candidates: {candidates}
```

Como pode ser observado, os *prompts* preparam o agente para a tarefa que ele deverá executar, pois balizam o seu comportamento e suas respostas.

3. Agentes de IA e o SADD

Nesta seção, apresentamos como os agentes de IA desenvolvidos foram integrados às funcionalidades do SADD. Primeiro, abordamos sua utilização na submissão e curadoria de fontes de dados enviadas ao sistema, onde atuam no auxílio ao vínculo de metadados existentes. Em seguida, mostramos seus usos nas funcionalidades de análise de dados, onde tanto os pesquisadores iniciantes quanto os experientes podem interagir com os dados disponibilizados pelo SADD, auxiliados pelos agentes de IA.

Durante a submissão, é necessário vincular cada variável da fontes de dados com seu nome padrão encontrado no NERC Vocabulary Server (NVS)¹, vocabulário padrão de metadados amplamente utilizado em repositórios de dados oceanográficos e de Ciências

¹<https://vocab.nerc.ac.uk/>

da Terra em geral. Na submissão, o agente sugere o vínculo entre colunas do conjunto submetido e conceitos de coleções do NVS previamente importadas para o catálogo local do SADD. Esse catálogo local é enriquecido por um pipeline que, além de importar as coleções, traduz termos para português e gera *embeddings* em inglês e português, favorecendo tanto a aderência ao vocabulário de domínio quanto sua usabilidade para o público brasileiro.

3.1. Submissão de Dados

A funcionalidade de submissão de fontes de dados constitui o primeiro ponto de contato para pesquisadores que desejam contribuir com o SADD, compartilhando dados produzidos em seus estudos. Nesse processo, o usuário preenche metadados básicos da fonte, como título, descrição e abrangência temporal, e envia tanto os arquivos de dados quanto seu respectivo dicionário de dados.

Para apoiar essa tarefa, o SADD incorpora um agente de IA que sugere correspondências entre as colunas do conjunto submetido e conceitos de coleções do NVS. As sugestões incluem coleção, conceito, URI e identificador NERC, mas a decisão final permanece com o usuário, que pode revisar e confirmar cada vínculo antes da submissão. Com isso, o SADD promove a padronização semântica das fontes de dados, reduzindo o esforço manual e a propensão a erros no mapeamento de metadados. A Figura 1 ilustra esse procedimento.

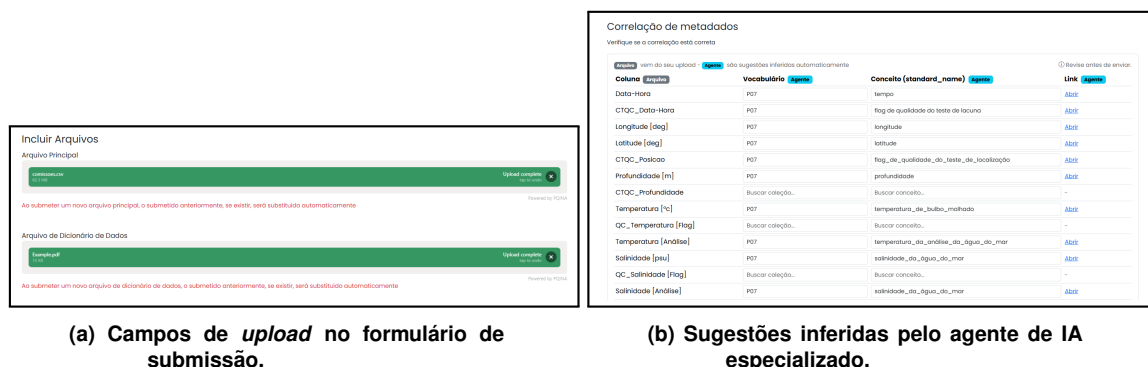


Figura 1. Funcionalidade de submissão de fonte de dados oceanográficos e o resultado da inferência de metadados pelo agente de IA.

3.2. Assistentes de Análise de Dados

Em complemento ao compartilhamento de dados provenientes de seus estudos, um pesquisador geralmente possui interesse em analisar dados compartilhados por outros pesquisadores. Contudo, esta costuma ser uma tarefa complexa, que envolve o *download* dos dados, preparação de um ambiente computacional especializado, pré-processamento dos dados adquiridos e desenvolvimento de *scripts* e outros comandos para realizar as análises desejadas. Logo, torna-se uma barreira natural que exclui usuários menos experientes, sem conhecimentos em programação e/ou atividades mais robustas de pesquisa. Visando contornar esta limitação, dois módulos de análise, suportados por agentes de IA, foram desenvolvidos.

Com foco em usuários que não possuam conhecimento de programação, o assistente não-especialista permite que usuários do SADD interajam com as fontes de dados

disponíveis no sistema, por meio de um assistente, através de linguagem natural. Neste assistente, baseado na interface de código aberto OpenWebUI², o usuário não-especialista pode interagir com um agente de IA, tirar dúvidas e solicitar análises e visualizações diversas, recebendo respostas em uma interface parecida com um *chat*, tornando a experiência bastante natural. A Figura 2 apresenta esta interface.

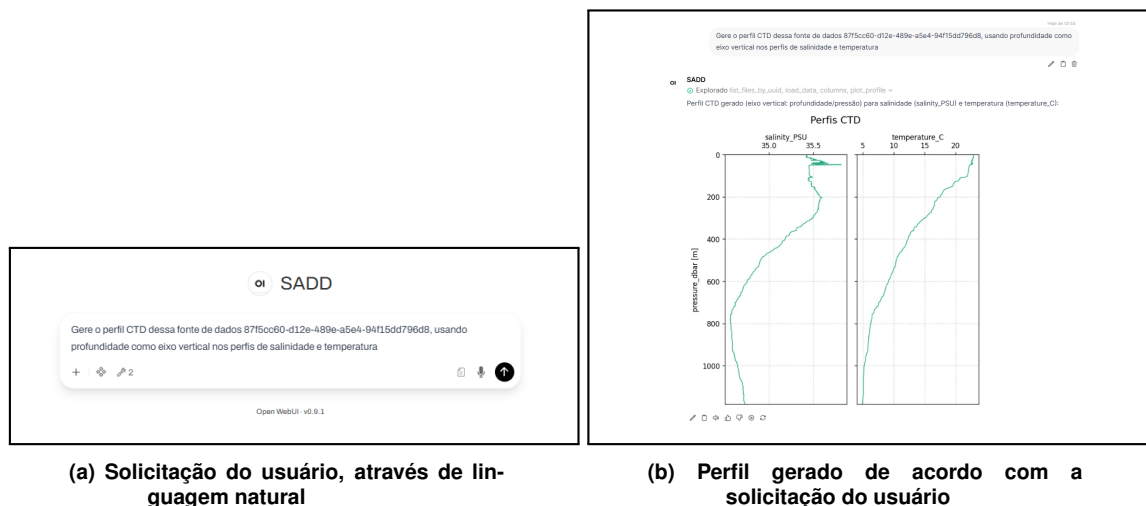


Figura 2. Módulo de análise de dados voltado a usuários não-especialistas

Portanto, esta ferramenta permite que usuários sem qualquer experiência com a geração de códigos possam facilmente gerar visualizações complexas. Logo, o acesso a dados complexos, como dados oceanográficos, se torna simplificado a todo tipo de público.

Embora o ambiente para usuários não-especialistas seja útil para todo tipo de usuário, aqueles mais avançados podem desejar ter um controle mais refinado, assim como acesso a *scripts* já desenvolvidos em estudos anteriores. Contudo, não há a necessidade de se realizar toda a configuração de um ambiente de desenvolvimento. O módulo para especialistas provê um ambiente destinado à análise de dados, através de *Notebooks Jupyter*. Neste ambiente, o usuário pode realizar o *upload* de seus próprios arquivos, além de criar *scripts* e acessar as fontes de dados disponíveis no SADD. De forma complementar, assim como no ambiente para não-especialistas, este módulo para especialistas conta com um agente de IA, acessível através do uso de uma *tag* especial na célula desejada. Por meio deste agente, é possível solicitar, por exemplo, a criação de *scripts* personalizados, utilizando linguagem natural. Portanto, neste módulo, o agente de IA especializado atua como um assistente, acelerando os estudos realizados pelo pesquisador mais experiente. A Figura 3 mostra esta funcionalidade.

4. Considerações Finais

Este artigo apresentou a integração de agentes de IA aos módulos de submissão, curadoria e análise de dados do SADD³. Estas novas ferramentas foram adicionadas com o

²<https://openwebui.com/>

³<https://github.com/AILAB-CEFET-RJ/inpo>

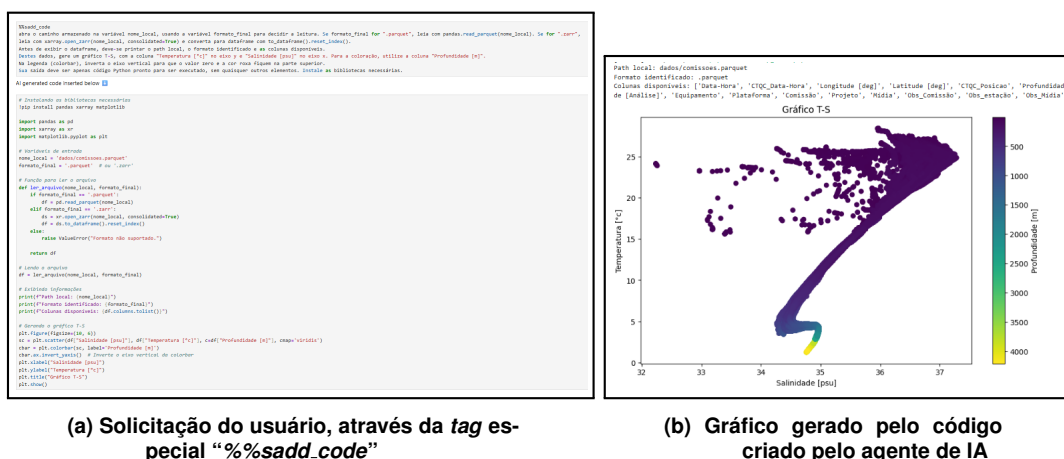


Figura 3. Módulo de análise de dados voltado a usuários especialistas, com familiaridade na criação e execução de códigos

intuito de facilitar o uso das funcionalidades do sistema, acelerando processos, reduzindo a propensão a erros e diminuindo a barreira de acesso a usuários menos experientes. A aceleração de atividades e a diminuição da barreira de acesso se dão pelo apoio ao usuário, sobretudo na análise de dados. As sugestões na funcionalidade de submissão reduzem a chance de falhas de preenchimento dos metadados, que refletiriam em inconsistências e outros problemas de qualidade nos dados posteriormente disponibilizados.

Para trabalhos futuros, investigaremos o refinamento e o desenvolvimento de novos agentes de IA, que possam auxiliar em outras partes do SADD, assim como o aprimoramento do módulo de curadoria com base na participação, utilização e retorno dos atores envolvidos.

Referências

- Aylak, B. L. (2025). Sustai-scm: Intelligent supply chain process automation with agentic ai for sustainability and cost efficiency. *Sustainability*, 17(6):2453.
- Guo, X., Yang, X., Peng, M., Lu, H., Zhu, M., and Yang, H. (2025). Automating traffic model enhancement with ai research agent. *Transportation Research Part C: Emerging Technologies*, 178:105187.
- Howe, B., Lawson, P., Bellinger, R., Anderson, E., Santos, E., Freire, J., Scheidegger, C., Baptista, A., and Silva, C. (2008). End-to-end escience: Integrating workflow, query, visualization, and provenance at an ocean observatory. In *2008 IEEE Fourth International Conference on eScience*, page 127–134. IEEE.
- Kessavane, I. (2025). A study of reducing hr redundancy processes with agentic ai. *Journal of Advances in Developmental Research*.
- Monteiro, A., Alves, C., Trotte, J., Monteiro, F., Bezerra, E., and González, P. H. (2025). SADD - sistema de armazenamento e disponibilização de dados oceanográficos. In *Anais Estendidos do XL Simpósio Brasileiro de Bancos de Dados (SBBD 2025)*, page 118–123. Sociedade Brasileira de Computação - SBC.
- Pati, A. K. (2025). Agentic ai: A comprehensive survey of technologies, applications, and societal implications. *IEEE Access*, 13:151824–151837.