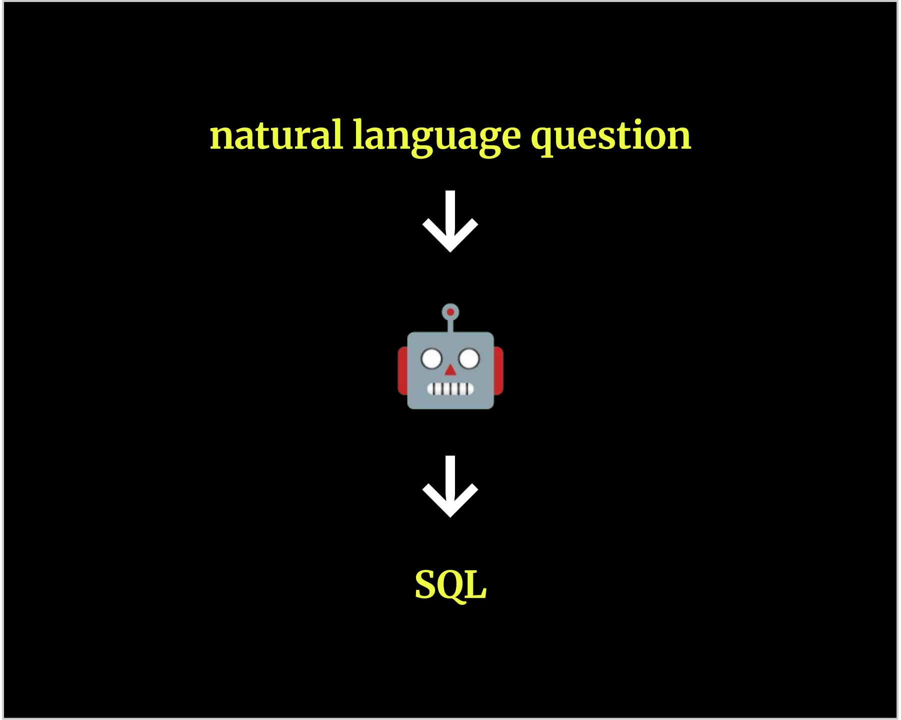
Text-to-SQL experiments with Engineering Data Extracted from CAD Files

Júlio G. Campos
Advisor: Marco A. Casanova
August 2025

Gerência de Engenharia Digital & Bim (GEDi)
Instituto Tecgraf - PUC-Rio

Introduction

- Recent advances in Large Language Models (LLMs) have driven the development of systems capable of translating natural language questions into SQL queries — a task known in the literature as **text-to-SQL** — enabling the construction of natural language interfaces for databases (Qin et al., 2022; Deng et al., 2022; Hong et al. 2025).

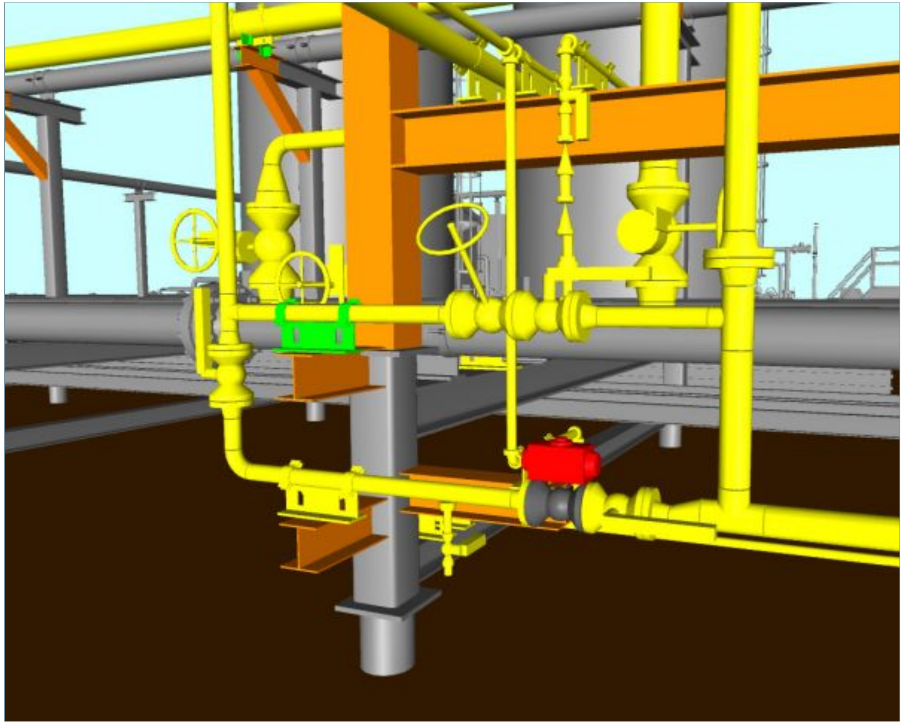


Introduction

- In the oil and gas industry (O&G), LLMs have been used in the development of intelligent assistants, question-answering systems, and in supporting tasks related to data analysis and visualization (Liu et al., 2024).

Introduction

- The O&G sector is responsible for the continuous generation of large volumes of data from various sources such:
    - Seismic surveys;
    - Well logs;
    - Drilling reports;
    - Real-time sensors;
    - Technical documents;
    - Corporate systems;
        - 3D CAD systems



Objective

- This study proposes an process to perform text-to-SQL on real-world data from a company operating in the O&G sector. The data is semi-structured and was extracted from a 3D CAD software, which stores models of two floating production storage and offloading units (FPSOs).

Related Work

Paper	Prompt engineering techniques used	Models used
Al Amin et al. (2025)	Few-shot learning	Not cited
Singh et al. (2024)	Chain-of-Thought (CoT), Retrieval-Augmented Generation (RAG)	Not cited
Ponomareva et al. (2024)	Retrieval-Augmented Generation (RAG)	GPT-4
Sabbagh et al. (2024)	Retrieval-Augmented Generation (RAG)	GPT-3.5-Turbo, GPT-4
Amour et al. (2024)	Retrieval-Augmented Generation (RAG), Dynamic few-shot examples (DFE)	GPT-4, GPT-4o mini, GPT-3.5, Mistral NeMo, Mistral Large
Singh et al. (2023)	Few-shot learning	GPT-3
Lokhande et al. (2023)	Not cited	T5

The Proposed Database Construction Process

```
{
    "type": "...",
    "root": {
        "type": "...",
        "properties": {},
        "geometries": {},
        "children": {
            "type": "...",
            "$values": [
                {
                    "type": "...",
                    "properties": {
                        "type": "...",
                        "$values": [
                            {
                                "type": "metadata",
                                "Name": "id",
                                "Value": "EQ1"
                            },
                            {
                                "type": "metadata",
                                "Name": "Object Class",
                                "Value": "Equipment"
                            },
                            {
                                "type": "metadata",
                                "Name": "Location",
                                "Value": "E 57, N 53, EL+ 51"
                            },
                            {
                                "type": "metadata",
                                "Name": "Weight",
                                "Value": "1235 kg"

                            }
                        ]
                    }
                },
                {},
                {}
            ]
        }
    }
}
```

```
[
    {
        "id": "EQ1",
        "Object Class": "Equipment",
        "Location": "E 57, N 53, EL+ 51",
        "Weight": "1235 kg"
    },
    {
        "id": "ST2",
        "Object Class": "Structure",
        "Material": "Steel",
        "Thickness": "20 mm",
        "Elevation": "4603 mm"
    },
    {
        "id": "PI3",
        "Discipline": "Piping",
        "Pressure": "50 x 50",
        "Length": "235 mm"
    }
]
```

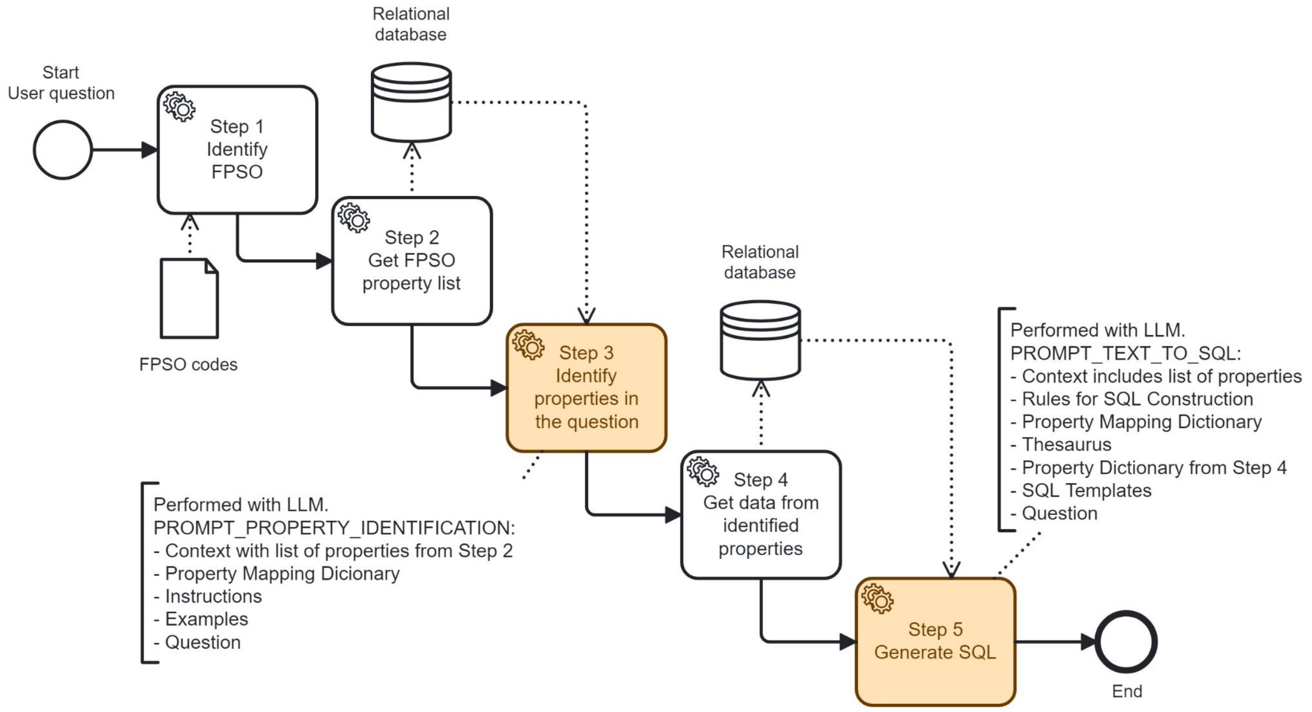
The Proposed Database Construction Process

Component

id	property	value	nval[a]	unit	x	y	z	i[b]	o[c]	fc[d]
EQ1	Object Class	Equipment								1
EQ1	Location	E 57, N 53, EL+ 51			57	53	51			1
EQ1	Weight	1235 kg	1235	kg						1
ST2	Object Class	Structure								1
ST2	Material	Steel								1
ST2	Thickness	20 mm	20	mm						1
ST2	Elevation	4603 mm	4603	mm			4603			1
PI3	Discipline	Piping								1
PI3	Pressure	50 x 50						50	50	1
PI3	Length	235 mm	235	mm						1

[a] Abbreviation for "numeric_value".
[b] Abbreviation for "input".
[c] Abbreviation for "output".
[d] Abbreviation for "fpso_code".

Proposed Prompt Engineering Text-to-SQL Process

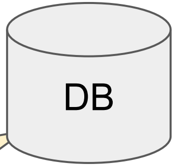


Step 1 - Identify FPSO

1

Quantos equipamentos da FPSO 1 possuem cor de revestimento de alumínio?

Step 2 - Get FPSO property list




```
Properties: [ "Discipline", "Object Class", "Angle", "Category",
"Coating Color", "Length", (...),  "Weight", "Width" ]
```

Quantos equipamentos da FPSO 1 possuem cor de revestimento de alumínio?

Step 3 – Identify properties in question

Property Mapping Dictionary
`{ "Discipline": ["equipamento"] … }`

Discipline

Quantos equipamentos da FPSO 1 possuem cor de revestimento de alumínio?

Coating Color

FPSO property list:
`[ "Discipline", "Object Class", "Angle", "Category", "Coating Color", "Length", (...), "Weight", "Width" ]`

Step 4 - Get data from identified properties

DB

{ "FPSO 1": [{ "name": "Discipline", "categorical_data": ["Equipment", "Structure", "equipment", "Instrument", "Piping"] }] }

List of identified properties:
["Discipline", "Coating Color"]

Quantos equipamentos da FPSO 1 possuem cor de revestimento de alumínio?

{ "FPSO 1": [{ "name": "Coating Color", "categorical_data": ["Aluminum", "Brown", "Gray", "Light Gray", "Blue", "Red", "Undefined", "White"] }] }

DB

Step 5 - Generate SQL

Quantos equipamentos da FPSO 1 possuem cor de revestimento de alumínio?

Generating SQL
from instructions,
using SQL Template 6 to
carry out SELF JOIN...

```
SELECT COUNT(DISTINCT t1.tag) AS count
FROM FPSO_1 t1
     JOIN FPSO_1 t2 ON t1.tag = t2.tag
WHERE lower(t1.property) = lower('Discipline')
     AND lower(t1.value) = lower('Equipment')
     AND lower(t2.property) = lower('Coating Color')
     AND lower(t2.value) = lower('Aluminum')
```

Seven SQL Templates

Question: quais válvulas da FPSO 1 estão localizadas no módulo AB com comprimento maior que 100 mm?

- Consider A, B, C ... W, X, Y e Z as case-sensitive variables.

4) SELF JOIN - Select ids from the same FPSO that have two properties with specific values:

```
SELECT DISTINCT t1.id
FROM table_name t1
JOIN table_name t2 ON t1.id = t2.id
WHERE lower(t1.property) = lower('W')
AND lower(t1.value) = lower('X')
AND lower(t2.property) = lower('Y')
AND lower(t2.value) = lower('Z') ORDER BY t1.id
```

Benchmark

50 natural language questions with
corresponding SQLs and execution results in database

Clauses and Aggregate Functions	Questions	Quantity
SELECT[a]	01-10	10
COUNT	11-20	10
SELF JOIN	21-30	10
UNION	31-40	10
AVG	41-42	2
MAX	43-44	2
MIN	45-46	2
GROUP BY	47-48	2
HAVING	49-50	2

[a] Only SELECT projections.

Experimental Setup

LLM	Params	Context Window	Released	License Type
GPT-4o[a]	-	128K	05/13/2024	Proprietary
OpenAI o3-mini[b]	-	200K	01/31/2025	Proprietary
DeepSeek-V3[c]	671B	128K	12/26/2024	Open Source
Mistral Large 2411[c]	123B	128K	11/18/2024	Proprietary

[a] API version: 2025-01-01-preview.
[b] API version: 2024-12-01-preview.
[c] Version 1.

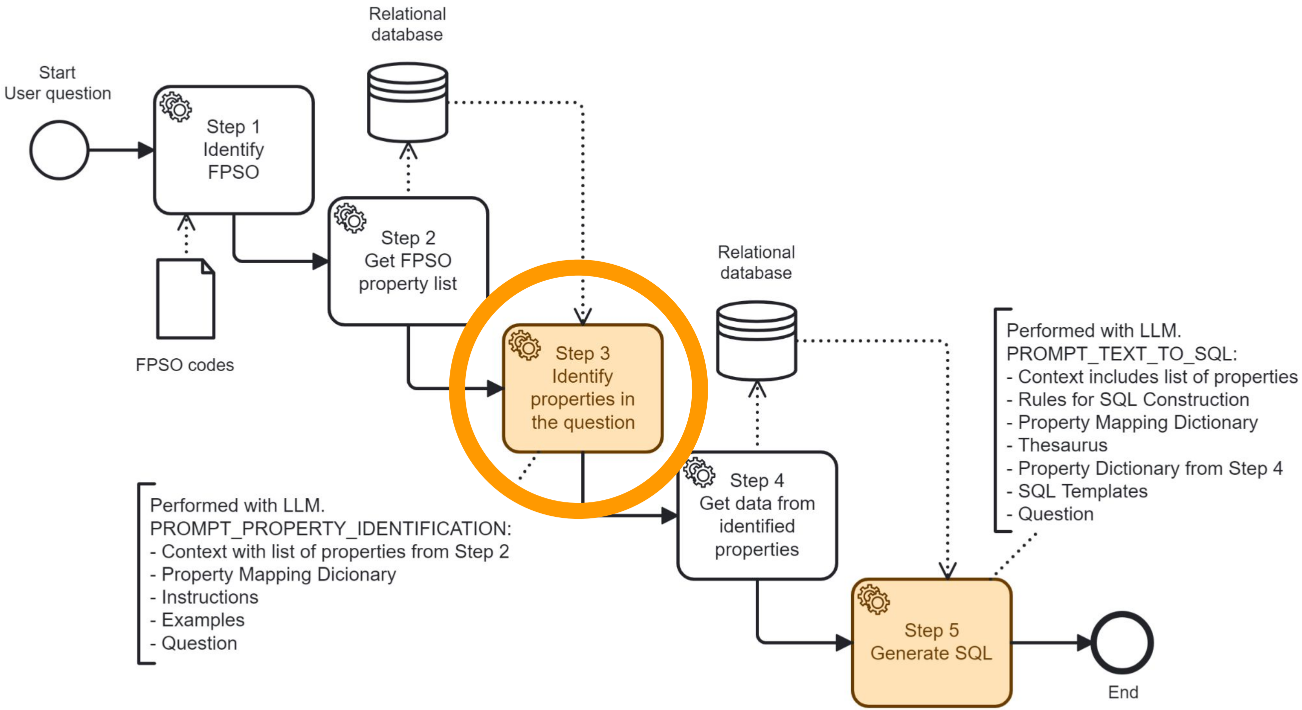
```
temperature = 0
OpenAI 03-mini: “reasoning_effort” = “high”
```

First Set of Experiments

LLM Model	Duration[a]	Hits	Errors	Execution accuracy[b]	Round
GPT-4o	00:03:56	50	0	100%	1
OpenAI o3-mini	00:20:41	45	5	90%	1
DeepSeek-V3	00:06:02	43	7	86%	3
Mistral Large 2411	00:13:13	40	10	80%	3

[a] Approximate time to run all 50 questions in hours, minutes and seconds.
[b] Round in which the model obtained the best result.

Improvements in the Proposed Prompt Engineering Text-to-SQL Process



Second Set of Experiments

LLM Model	Duration[a]	Hits	Errors	Execution accuracy[b]	Round
GPT-4o	00:02:40	50	0	100%	2
DeepSeek-V3	00:03:57	48	2	96%	2
OpenAI o3-mini	00:07:36	47	3	94%	2
Mistral Large 2411	00:08:55	47	3	94%	2

[a] Approximate time to run all 50 questions in hours, minutes and seconds.

[b] Round in which the model obtained the best result.

```
temperature = 0
seed = 26
OpenAI 03-mini: “reasoning_effort” = “high”
```

Comparative Analysis of the Experiments

LLM Model	No. Parameters	EA 1[a]	EA 2[b]
GPT-4o	-	100%	100%
DeepSeek-V3	671B	86%	96%
OpenAI o3-mini	-	90%	94%
Mistral Large 2411	123B	80%	94%

[a] EA 1: Execution accuracy in the first set of experiments.
[b] EA 2: Execution accuracy in the second set of experiments.

Conclusions

- Is possible to perform text-to-SQL on semi-structured engineering data, using open-source LLMs;
- The results presented are not intended to establish a definitive ranking of the best LLMs for performing text-to-SQL;
- The experiments reported in this study also suggested that it is not always possible to rely on the linguistic capabilities of LLMs to execute all tasks, even when contextual information is provided.
- The prompt technique based on using seven SQL templates to help the LLM generate valid SQL queries led to excellent results without resorting to more advanced techniques based on RAG;
- The proposed prompt engineering text-to-SQL process can be adapted to other contexts and research area;

Future Work

- This study is a starting point for replication to other FPSOs, since the O&G company operates dozens of FPSOs;

- New studies must be carried out to perform text-to-SQL using engineering data properties that are common across all FPSOs;

- New prompt engineering techniques can be adopted and new technologies can be integrated into the proposed prompt engineering text-to-SQL process, such as using knowledge graphs.

Publications

- Campos, J., García, G., de Sousa, J.A., Corseuil, E., Izquierdo, Y., Lemos, M., Casanova, M.: **Text-to-SQL Experiments with Engineering Data Extracted from CAD Files**. In: 27th International Conference on Enterprise Information Systems. pp. 343–350 (Apr 2025). https://doi.org/10.5220/0013436800003929

- Submitted for *Lecture Notes in Business Information Processing* (Springer)

Text-to-SQL Experiments with Engineering Data Extracted from CAD Files

Júlio G. Campos[1,a], Grettel M. García[1,b], Jefferson A. de Sousa[1,c], Eduardo T. L. Corseuil[1,d], Yenier T. Izquierdo[1,e], Melissa Lemos[1,2,f] and Marco A. Casanova[1,2,g]
[1]Instituto Tecgraf, PUC-Rio, Rio de Janeiro, CEP 22451-900, RJ, Brazil
[2]Departamento de Informática, PUC-Rio, Rio de Janeiro, CEP 22451-900, RJ, Brazil

Keywords: Text-to-SQL, Large Language Models, Relational Databases, Llama, Gemma, GPT.

Abstract: The development of Natural Language (NL) interfaces to access relational databases attracted renewed interest with the use of Large Language Models (LLMs) to translate NL questions to SQL queries. This translation task is often referred to as text-to-SQL, a problem far from being solved for real-world databases. This paper addresses the text-to-SQL task for a specific type of real-world relational database storing data extracted from engineering CAD files. The paper introduces a prompt strategy tuned to the text-to-SQL task over such databases and presents a performance analysis of LLMs of different sizes. The experiments indicated that GPT-4o achieved the highest accuracy (96%), followed by Llama 3.1 70B Instruct (86%). Quantized versions of Gemma 2 27B and Llama 3.1 8B had a very limited performance. The main challenges faced in the text-to-SQL task involved SQL complexity and balancing speed and accuracy when using quantized open-source models.

1 INTRODUCTION

The development of Natural Language (NL) interfaces to access relational databases attracted renewed interest with the use of Large Language Models (LLMs) to translate NL questions to SQL queries. This translation task is often referred to as *text-to-SQL*.

Companies are increasingly interested in using text-to-SQL to support business processes that require easy access to databases at both the operational and strategic levels. However, they need to address several challenges to make this technology viable, such as semantic data integration, which allows LLMs not only to understand how different databases relate to each other but also to understand the semantics of the data (Campos et al., 2023).

Concerns about cost and data privacy are also driving companies to seek alternatives to proprietary LLMs. Despite their constant evolution, the potential of open-source LLMs is under-exploited, even as notable advances have been made in programming tasks, mathematical reasoning, and text generation (Chiang et al., 2023). Much of the research on text-to-SQL focuses primarily on proprietary LLMs, leaving open-source LLMs behind (Gao et al., 2023).

This paper addresses the text-to-SQL task for one very specific type of real-world relational database storing data extracted from engineering CAD files. The paper introduces a prompt strategy tuned to the text-to-SQL task over such databases and presents a performance analysis of LLMs of different sizes, both proprietary and open source.

The paper assumes that the objects and their property values are first extracted from engineering Computer-Aided Design (CAD) files, which represents 3D models of industrial plants. Then, the data extracted are stored in a relational table with three columns, called an *object-property-value* table. Each row in the table has the format (o, p, v), where o is an object ID, p is the name of one of the object properties, and v is the property value. The use of triples is a familiar strategy to address the enormous variability of the object properties.

[a] https://orcid.org/0000-0002-5023-3836
[b] https://orcid.org/0000-0001-9713-300X
[c] https://orcid.org/0000-0002-5928-9959
[d] https://orcid.org/0000-0002-7543-4140
[e] https://orcid.org/0000-0003-0971-8572
[f] https://orcid.org/0000-0003-1723-9897
[g] https://orcid.org/0000-0003-0765-9636

References

- Androutsopoulos, I., Ritchie, G.D., Thanisch, P.: Natural language interfaces to databases – an introduction. Natural Language Engineering 1(1), 29–81 (Mar 1995). https://doi.org/10.1017/S135132490000005X
- Al Amin, S.T., Wu, K., Mathur, A., Koritala, C.: LLM-In-The-Loop: A Framework for Enhancing Drilling Data Analytics. In: SPE/IADC International Drilling Conference and Exhibition. OnePetro (Feb 2025). https://doi.org/10.2118/223805-MS
- Amour, M., Rachmat, B.K., Rementeria, A.S., Guillot, V., Millan, E.: Empowering Drilling and Optimization with Generative AI. In: ADIPEC. OnePetro (Nov 2024). https://doi.org/10.2118/221862-MS
- Campos, J., García, G., de Sousa, J.A., Corseuil, E., Izquierdo, Y., Lemos, M., Casanova, M.: Text-to-SQL Experiments with Engineering Data Extracted from CAD Files. In: 27th International Conference on Enterprise Information Systems. pp. 343–350 (Apr 2025). https://doi.org/10.5220/0013436800003929
- Deng et al.: Recent advances in text-to-SQL: A survey of what we have and what we expect. In: Proceedings of the 29th International Conference on Computational Linguistics. pp. 2166–2187. International Committee on Computational Linguistics, Gyeongju, Republic of Korea (Oct 2022)
- Hong, Z., Yuan, Z., Zhang, Q., Chen, H., Dong, J., Huang, F., Huang, X.: Next- Generation Database Interfaces: A Survey of LLM-based Text-to-SQL (Mar 2025). https://doi.org/10.48550/arXiv.2406.08426
- Lewis et al.: Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. pp. 9459–9474. NIPS ’20, Curran Associates Inc., Red Hook, NY, USA (Dec 2020). https://doi.org/10.48550/arXiv.2005.11401

References

- Liu, H., Ren, Y., Li, X., Deng, Y., Wang, Y., Cao, Q., Du, J., Lin, Z., Wang, W.: Research status and application of artificial intelligence large models in the oil and gas industry. Petroleum Exploration and Development 51(4), 1049–1065 (2024). https://doi.org/10.1016/S1876-3804(24)60524-0
- Lokhande, A., Pillai, P., Mangsuli, P.: Enabling Contextual Natural Language Search on Oil and Gas Databases. In: ADIPEC. OnePetro (Oct 2023). https://doi.org/10.2118/216349-MS
- Ponomareva, D., Droubi, N.E., Jundi, O.E., Assaf, G., Mustapha, H., Kindi, Z.A.: Domain Driven Methodology Adopting Generative AI Application in Oil and Gas Drilling Sector. In: ADIPEC. OnePetro (Nov 2024). https://doi.org/10.2118/221883-MS
- Sabbagh, V.B., Lima, C.B.C., Xexéo, G.: Comparative Analysis of Single and Multiagent Large Language Model Architectures for Domain-Specific Tasks in Well Construction. SPE Journal 29(12), 6869–6882 (Dec 2024). https://doi.org/10.2118/223612-PA
- Qin et al.: A Survey on Text-to-SQL Parsing: Concepts, Methods, and Future Directions (Aug 2022). https://doi.org/10.48550/arXiv.2208.13629
- Shi, L., Tang, Z., Zhang, N., Zhang, X., Yang, Z.: A Survey on Employing Large Language Models for Text-to-SQL Tasks. ACM Comput. Surv. (Jun 2025). https://doi.org/10.1145/3737873
- Singh, A., Jia, T., Nalagatla, V.: Generative AI Enabled Conversational Chatbot for Drilling and Production Analytics. In: ADIPEC. OnePetro (Oct 2023). https://doi.org/10.2118/216267-MS
- Singh, A., Jia, T., Nalagatla, V., Cunningham, B., Siwani, T.: Application of Generative AI to Derive Insight from Supply Chain & Logistics Contracts. In: ADIPEC. OnePetro (Nov 2024). https://doi.org/10.2118/222932-MS
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., Zhou, D.: Chain-of-Thought Prompting Elicits Reasoning in Large Language Models (Jan 2023). https://doi.org/10.48550/arXiv.2201.11903

Tecgraf
PUC-RIO

Obrigado!

Júlio Campos: juliogcampos@tecgraf.puc-rio.br
Eduardo Thadeu: thadeu@tecgraf.puc-rio.br
Melissa Lemos: melissa@tecgraf.puc-rio.br
Marco A. Casanova: casanova@inf.puc-rio.br