

O USO DO MODELOS AMPLOS DE LINGUAGEM (LLM) PARA EXTRAÇÃO DE ESQUEMAS DE CONJUNTO DE DADOS SEMIESTRUTURADOS

ANA PAULA HARTMANN^{1,2}, LUAN ALECXANDER KRZYZANIAK³, GEOMAR
ANDRÉ SCHREINER⁴, DENIO DUARTE⁵

1 Introdução

A evolução da tecnologia levou a geração de um grande número de dados em comparação com décadas passadas. Esses, em sua maioria, não seguem um esquema definido, sendo fracamente estruturados, irregulares e incompletos (Kellou-Menouer et al., 2022). Nesse sentido, o uso de formatos flexíveis, como o *JavaScript Object Notation* (JSON), tornam-se essenciais. Entretanto, a falta de padronização, validação e compreensão estrutural definem desafios para trabalhar com esse tipo de documentos (Pezoa et al., 2016).

Os esquemas são essenciais para aplicativos centrados em dados (Kellou-Menouer et al., 2022). Porém, a identificação dos esquemas em dados semi-estruturados e não estruturados, onde os metadados se encontram nos próprios valores, possui dificuldades para realizar uma inferência automática (Banhara et al. 2024). Para realizar essa tarefa, podem ser utilizados Modelos de Linguagem de Grande Escala (LLMs), modelos de aprendizado profundo, capazes de compreender, processar e gerar linguagem natural em diferentes tarefas (Zhao et al. 2023).

Para entender essa lacuna, o presente trabalho propõe e avalia um *pipeline* que recebe documentos não estruturados, realiza seu processamento e utiliza um LLM para identificar as informações relevantes, resultando um esquema JSON representando a estrutura do documento. Para tal, são utilizados documentos legais, visto que o sistema jurídico brasileiro gera um grande volume de documentos que são majoritariamente armazenados em formatos não estruturados (Aquino et al. 2024).

1 Graduação em Ciência da Computação, Universidade Federal da Fronteira Sul, *campus* Chapecó, contato: ana.hartmann@estudante.uffs.edu.br

2 Grupo de Pesquisa: Inovação e Desenvolvimento Tecnológico.

3 Graduando, Universidade Federal da Fronteira Sul.

4 Doutor em Ciência da Computação, Universidade Federal da Fronteira Sul.

5 Doutor, Universidade Federal da Fronteira Sul, **Orientador**.

2 Objetivos

Este trabalho tem como objetivo propor uma abordagem e avaliar o desempenho de LLMs na extração de esquemas JSON de fonte de dados não estruturados, focando em documentos legais.

3 Metodologia

Este trabalho possui três etapas em seu fluxo preparatório: coleta, pré-processamento e preparação dos dados. A etapa de coleta ocorreu com a aquisição de documentos legais retirados de portais oficiais do governo, como o Portal Transparência⁶, tendo em foco as Atas de Registro de Preço (ARP). As ARPs foram escolhidas por possuírem esquemas parecidos mas com diferenças na estrutura e forma de representar os dados. Essa etapa ocorreu, inicialmente, em formato manual, para que se entendesse a estrutura dos documentos. Em seguida, utilizou-se o *framework Selenium*⁷ para automatizar o processo, por conseguir imitar interações e manipular páginas com conteúdo *JavaScript* e carregamento dinâmico.

Com essa etapa concluída, iniciou-se o pré-processamento, em que os documentos duplicados ou compostos em sua maioria por imagens foram retirados do conjunto, resultando em 471 documentos. Após, na etapa de preparação, os arquivos passaram por uma limpeza, normalização e preparação textual. Foi utilizada uma *pipeline* em *Python*, que padroniza o texto sem perder dados. Os documentos resultantes foram reunidos em um único bloco, para que a LLM pudesse acessar o conteúdo de forma sequencial.

Para a realização dos experimentos, foi selecionado o *Mistral 7B Instruct v0.2*, um modelo de linguagem de código aberto que oferece eficiência em tarefas *fine-tuning* (Jiang et al. 2023). Esse modelo opera no nível de subpalavras, permitindo a representação eficiente de termos jurídicos, estruturas numéricas e padrões textuais. Foi proposto um esquema (*golden schema*) que permite verificar as estruturas extraídas, para servir de referência na avaliação. Ele apresenta, em formato JSON, os campos principais e os tipos de dados. Esse modelo foi criado manualmente a partir de 15 documentos aleatórios presentes no conjunto, gerando 25 campos. A Figura 1A apresenta uma versão reduzida do *golden schema*, com os tipos estruturais existentes. O próximo passo foi a criação de um *prompt* com o objetivo de garantir a

⁶ <https://contratos.sistema.gov.br/transparencia>

⁷ <https://www.selenium.dev/>

consistência e guiar a LLM na extração. Esse, consiste em uma exemplificação das estruturas esperadas no documento.

Por fim, foram realizados os experimentos para avaliar o desempenho do *Mistral 7B*, utilizando uma GPU NVIDIA RTX 4070 Ti com 8 GB de memória e quantização em 4 bits. Como o modelo possui um limite de aproximadamente 8.000 *tokens* de contexto, cada documento foi dividido em blocos de 2.000 tokens por meio do *tokenizador* nativo do *Mistral*, evitando truncamentos e mantendo a qualidade da extração. Cada *chunk* foi enviada individualmente ao modelo utilizando o mesmo *prompt* de extração previamente definido. Após a execução, cada documento obteve um JSON, criado a partir dos *chunks* processados, que manteve as chaves e estruturas. Chaves iguais são unificadas, e para chaves diferentes, cria-se uma lista de tipos. A Figura 1B apresenta um pedaço de um esquema gerado, onde existem três objetos completos: (i) *orgao_gerenciar* (linhas 1-9), (ii) *fornecedores* (linhas 11-26) e (iii) *cancelamento* (linhas 29-31).

Figura 1. Exemplo simplificado do golden schema (A) e do esquema unificado gerado pelo modelo (B)

A

```

1 {"orgao_gerenciar": {
2   "type": "object",
3   "properties": {
4     "email": { "type": "string" },
5     "telefone": { "type": "string" },
6     "endereco": { "type": "object",
7       "properties": {
8         "bairro": { "type": "string" },
9         "cep": { "type": "string" }}}},
10  },
11  "fornecedores": {
12    "type": "array",
13    "items": {
14      "type": "object",
15      "properties": {
16        "registro_de_precos": {
17          "type": "object",
18          "properties": { "cancelado": { "type": "boolean" },
19            "precos": {
20              "type": "array",
21              "items": { "type": "object",
22                "properties": { "item": { "type": "string"},
23                  "preco": { "type": "float" }}}}}}
24      },
25    },
26    "cancelamento": { "type": "object", "properties": { "motivo": { "type": "string" } }},
27  ...}

```

B

```

1 {
2   "numero_ata": { "type": "string"},
3   ...
4   "vigencia": {
5     "type": "object",
6     "properties": { "data_inicio": { "type": "date" },
7       "data_fim": { "type": "date" } }},
8   ...
9   "fornecedores_registrados": {
10    "type": "array",
11    "items": { "type": "object",
12      "properties": { "razao_social": { "type": "string"},
13        "cnpj": { "type": "string" } } }},
14  ...}

```

4 Resultados e Discussão

A avaliação qualitativa foi realizada comparando os JSONs extraídos com o *golden schema* de referência por meio de métricas de similaridade semântica e acurácia dos tipos de dados, utilizando o modelo de *embeddings all-MiniLM-L6-v2* e similaridade do cosseno. Os resultados médios obtidos a partir da comparação individual dos documentos são apresentados na Tabela 1.

Tabela 1. Métricas gerais de desempenho do Mistral 7B Instruct v0.2 na extração de esquemas de ARPs por documento.

Métrica	Valor médio
Semantic Accuracy	0,412
Semantic Coverage	0,714
Semantic F1	0,519
Type Accuracy	0,946
Type Precision	0,972
Campos previstos (média)	30,9

A *Semantic Accuracy* indica que aproximadamente 41% dos campos previstos possuíam correspondência semântica igual ou superior a 0,75 com o *golden schema*. A *Semantic Coverage* mostra que 71% dos campos do esquema de referência foram cobertos, incluindo correspondências parciais com similaridade de 0,60 a 0,75. Esses resultados indicam que a compreensão semântica ainda necessita supervisão humana. Um equilíbrio entre precisão semântica e cobertura pode ser visto no resultado da *Semantic F1*. A *Type Accuracy* e *Type Precision* mostram que mais de 90% dos campos extraídos correspondem aos tipos do *golden schema*. Isso demonstra a robustez do modelo mesmo em documentos complexos. Por fim, o modelo gerou em média 30,9 campos por documento, o que se aproxima dos 25 campos do *golden schema*, mostrando um valor aceitável de ruído mas com boa extração de informações importantes. Os resultados indicam potencial para utilização do *Mistral 7B* como ferramenta de apoio, pois apresentou um desempenho satisfatório na extração de estruturas ARPs.

5 Conclusão

Conclui-se que o *Mistral 7B* pode atuar como ferramenta auxiliar na extração de esquemas de documentos não estruturados, principalmente em documentos legais, auxiliando

na preparação e na redução de esforço manual. Porém, ainda necessita de técnicas de pré-processamento, limpeza e correção semântica. Para próximos trabalhos, visa-se aprimorar o modelo com RAG e *fine-tuning*, assim como melhorar o pós-processamentos dos esquemas, para que seja diminuída a necessidade de supervisão humana e aprimorada a correspondência de chaves.

Referências Bibliográficas

AQUINO, I. et al. Extracting information from Brazilian legal documents with retrieval augmented generation. In: SBBD, 39., 2024. *Anais [...]*. Porto Alegre: SBC, 2024.

BANHARA, N. et al. Enumeration, tagged unions, tuples, and collections: a novel approach to extracting JSON Schema. In: SBBD, 2024. *Anais [...]*. Porto Alegre: SBC, 2024.

JIANG, A. Q. et al. **Mistral 7B**. 2023. Disponível em: <https://arxiv.org/abs/2310.06825>. Acesso em: 11 jul. 2026.

KELLOU-MENOUER, K. et al. A survey on semantic schema discovery. *The VLDB Journal*, v. 31, n. 4, p. 675–710, 2022. <https://doi.org/10.1007/s00778-021-00717-x>.

PEZOA, F. et al. Foundations of JSON Schema. In: INTERNATIONAL WORLD WIDE WEB CONFERENCE, 25., 2016. *Proceedings [...]*. Republic and Canton of Geneva: International World Wide Web Conferences Steering Committee, 2016. p. 263–273.

ZHAO, W. X. et al. **A survey of large language models**. *arXiv*, 2023. Disponível em: <https://arxiv.org/abs/2303.18223>. Acesso em: 11 jul. 2026.

Palavras-chave: Large Language Models; Aprendizado de Máquina; Extração de Dados.

Nº de Registro no sistema Prisma: PES-2025-0272

Financiamento:

