Revista Brasileira de Cartografia
ISSN 1808-0936 | https://doi.org/10.14393/revbrascartogr
Sociedade Brasileira de Cartografia, Geodésia, Fotogrametria e Sensoriamento Remoto

Explorando o Uso de Large Language Model (ChatGPT) para Alinhamento Semântico entre Esquemas Conceituais de Dados Geoespaciais

Exploring the Use of Large Language Model (ChatGPT) for Semantic Alignment between Geospatial Data Conceptual Schemas

Fabíola Andrade Souza [1], Estephanie Daiane Batista da Silva [2], Silvana Philippi Camboim[3]

1 Universidade Federal da Bahia - UFBA, Salvador-BA, Brasil. fabiola.andrade@ufba.br
ORCID: https://orcid.org/0000-0003-2475-4520
2 Universidade Federal do Paraná - UFPR, Curitiba-PR, Brasil. estephaniedaiane@ufpr.br
ORCID: https://orcid.org/0009-0001-0022-5870
3 Universidade Federal do Paraná - UFPR, Curitiba-PR, Brasil. silvanacamboim@ufpr.br
ORCID: https://orcid.org/0000-0003-3557-5341

Recebido: 09.2024 | Aceito: 04.2025

Resumo: Diante do cenário atual, em que o exponencial crescimento na produção de dados geoespaciais converge com a necessidade de sua disseminação e compartilhamento, torna-se salutar o desenvolvimento de mecanismos que facilitem a interoperabilidade dos dados, cujas fontes de produção podem ser diversas. Assim, questões voltadas a promover os processos de interoperabilidade semântica, entre distintas modelagens conceituais destes dados, tornam-se relevantes. Deste modo, este artigo investiga o potencial de utilização de uma ferramenta de processamento de linguagem natural, construída sobre um grande modelo de linguagem (*Large Language Model* - LLM), como um elemento facilitador para futura automatização dos mecanismos de alinhamento semântico entre esquemas conceituais diversos. Como resultado, a ferramenta utilizada – ChatGPT – apresentou 123 associações semânticas entre os esquemas utilizados: 34 classes da categoria edificações da base cartográfica de referência do Brasil e quaisquer *tags* aplicadas para criação de dados voluntários no *OpenStreetMap* (OSM). Em alguns casos, as associações foram detalhadas, em outros, mais genéricas, sendo possível sua comparação com trabalhos prévios realizados manualmente por humanos. Importante salientar o papel relevante da construção do diálogo de solicitação do alinhamento, com organização estruturada dos dados conceituais, bem como utilização de diálogo claro e sem ambiguidades. Ainda existem limitações no processo, em especial para entendimento da hierarquia dos conceitos utilizados, o que indica necessidade de novos estudos e avaliação de outros LLM disponíveis. Entretanto, o uso de inteligência artificial para interoperabilidade semântica de dados geoespaciais desponta como um caminho viável a ser aplicado.
Palavras-chave: *OpenStreetMap*, bases de referência, processamento de linguagem natural, ChatGPT.

Abstract: Given the current scenario, where the exponential growth in the production of geospatial data converges with the need for its dissemination and sharing, the development of mechanisms that facilitate data interoperability, whose sources of production may be diverse, becomes crucial. Thus, issues aimed at promoting semantic interoperability processes between different conceptual models of these data become relevant. Accordingly, this paper investigates the potential use of a natural language processing tool, built on a Large Language Model (LLM), as a facilitator for the future automation of semantic alignment mechanisms between different conceptual schemas. As a result, the tool used – ChatGPT – presented 123 semantic associations between the utilized schemas: 34 classes from the building category of Brazil's reference cartographic base and various tags applied for creating voluntary data in OpenStreetMap (OSM). In some cases, the associations were detailed, while in others, they were more general, allowing for comparison with previous work manually conducted by humans. It is important to highlight the significant role of constructing the alignment request dialogue, with structured organization of conceptual data, as well as the use of clear and unambiguous dialogue. There are still limitations in the process, particularly in understanding the hierarchy of the concepts used, indicating the need for further studies and evaluation of other available LLMs. Nevertheless, the use of artificial intelligence for the semantic interoperability of geospatial data emerges as a viable path to be applied.
***Key*words:** *OpenStreetMap*, topographic maps, natural language processing, ChatGPT.

1 INTRODUÇÃO

Interoperabilidade é fundamental para resolver problemas de heterogeneidade no compartilhamento e uso integrado de dados geoespaciais. Desafio que se torna mais relevante com o aumento expressivo na geração de dados e discussões sobre infraestrutura de dados espaciais (Ballatore et al., 2013; Robinson et al., 2017; Yu et al., 2018). Entretanto, embora haja discussão e proposição de soluções sobre interoperabilidade de formatos de dados (ISO, 2015; OGC, 2023), questões relacionadas à semântica, ou seja, à compatibilidade conceitual dos dados, tem sido debatida de maneira teórica, com menores avanços práticos, especialmente com respeito à tecnologia, buscando reduzir a dependência humana (Kuhn, 2003; OGC, 2020; Machado & Camboim, 2024).

A literatura indica que todo mapeamento deve primeiro passar pela conceituação de seus objetos, para depois avançar à etapa de produção (Borges et al., 2005; Sluter et al., 2019). Sob a perspectiva tecnológica, a modelagem conceitual de dados é uma técnica elaborada para formalizá-los em esquemas de bancos de dados e, para os dados geoespaciais, tem estreita relação com conhecer o perfil do usuário e seus contextos de uso, a forma de raciocínio sobre o tema, como sugerem Borges et al. (2005). Sob o aspecto cartográfico, um alinhamento de conceitos depende da percepção e cognição humanas baseadas em conhecimento prévio e acordos sociais, podendo variar em diferentes lugares e culturas (Bravo, 2014; Machado, 2020) e são alicerce para a interoperabilidade entre diferentes esquemas de dados, em especial ao tratar de bases de referência.

Considerando o mapeamento topográfico brasileiro, este tem seus elementos de representação definidos, para diferentes escalas, através do plano de ação para implantação da Infraestrutura Nacional de Dados Espaciais (INDE), sendo ponto de partida para os dados da cartografia nacional, cujas definições poderão ser referência aos demais mapeamentos e estão embasadas na legislação do país e normas anteriores. Neste caso, destaca-se dentre as normas propostas, as Especificações Técnicas para Estruturação e Aquisição de Dados Geoespaciais Vetoriais (ET-EDGV e ET-ADGV) que tratam, respectivamente, da modelagem conceitual de dados vetoriais geoespaciais oficiais de referência e da padronização e orientação ao processo de aquisição de suas geometrias de representação, atuando de maneira integrada e servindo de referência para a padronização e disseminação dos dados oficiais (Brasil, 2010; Concar, 2017; Ministério da Defesa, 2018).

Importante salientar que a cobertura cartográfica no Brasil é deficitária, em especial para escalas maiores, conforme estudos de Silva e Camboim (2020), que apontam necessidade de complementar e atualizar esta cobertura através de fontes alternativas, como, por exemplo, dados colaborativos. Neste caso, a integração das bases colaborativas àquelas oficiais implica, principalmente, em compatibilidade semântica, para que as informações incorporadas sejam de natureza conceitual similar. A integração de dados baseada na semântica, considerando tanto a organização de dados internacionais e gerais quanto locais e específicos, embora ofereça esforços, ainda carece de análises e avanços tecnológicos, principalmente com ferramentas operacionais efetivas (Anand *et al.*, 2010; Ballatore *et al.*, 2013; Yu *et al.*, 2018; Silva, 2022; Machado & Camboim, 2024).

Neste contexto, observou-se algumas ações de integração semântica de dados geoespaciais em nível internacional (Anand *et al.*, 2010; Ballatore *et al.*, 2013) e no cenário brasileiro (Bortolini et al., 2018; Machado, 2020; Silva *et al.*, 2022; Silva, 2022; Machado & Camboim, 2024), cujos estudos, em sua maioria, apresentam metodologias e resultados possíveis apenas através de análise humana e ao final geraram figuras, grafos, planilhas ou arquivos textuais. Corroborando com esta situação, Machado & Camboim (2019, p. 13), ao avaliarem o potencial de integração da base oficial do município de Curitiba-PR no esquema da ET-EDGV com as *tags* (*key+value*) do *OpenStreetMap* (OSM), apresentam em sua metodologia a realização de compatibilidade semântica que "consistiu na análise, observação e interpretação minuciosas de cada uma das descrições das camadas de feições e atributos". As autoras apontam que, dentre os desafios para integração de duas modelagens distintas, estão as diferentes estruturas semânticas dos dados, que necessitam de ferramentas que automatizem a compatibilidade e não dependam de análises individuais de interpretação.

Portanto, compatibilidade conceitual é essencial para interoperabilidade e integração de dados visando facilitar sua disponibilidade e acesso abrangentes, devendo ocorrer, inclusive, a nível computacional, para além do entendimento pelos seres humanos, permitindo a identificação, comparação e associação de conceitos entre distintas modelagens de maneira automatizada. Contudo, há uma ampla gama de questões a serem pesquisadas e aprofundadas, especialmente quanto à automação dos processos de integração e seu potencial de uso.

Neste cenário sobre a automação, o uso de ferramentas baseadas em inteligência artificial – IA e

Processamento de Linguagem Natural – PLN pode ter aplicação. Embora a definição de IA tenha surgido na década de 1950 como a capacidade de simular computacionalmente o raciocínio humano, ou seja, "fazer com que as máquinas usem a linguagem, formem abstrações e conceitos, resolvam tipos de problemas hoje reservados aos humanos e se aprimorem" (McCarthy *et al.*, 1955, p. 12), seu uso de maneira abrangente evoluiu nas décadas seguintes, à medida que a capacidade de processamento computacional avançava e novos modelos matemáticos eram propostos para simular este raciocínio (Segaran, 2007; Prince, 2023).

O PLN surgiu no contexto do aprendizado de máquina e das redes neurais, cujas equações matemáticas têm parâmetros ajustáveis e aprendem com novos dados inseridos na base de conhecimento, mesmo permitindo falhas em situações fora do padrão (McCarthy *et al.*,1955; Segaran,2007; Prince,2023). As equações dependem dos pesos, vieses e funções das entradas de dados e os erros podem estar associados à incerteza da tarefa, à quantidade de dados de treinamento e ao modelo (Segaran, 2007; Santhanam e Shaikh, 2019; Prince, 2023).

As bases de conhecimento são construídas a partir da modelagem linguística, que utiliza diferentes algoritmos de redes neurais e aprende a distribuição de probabilidade sobre sequências de símbolos em um idioma, efetuando interpretação sintática e semântica do texto (Jozefowicz *et al.*, 2016). Recentemente, grandes modelos de linguagem (*Large Language Models* - LLM) têm sido aplicados para tratamento de amplas bases de dados textuais, com bilhões de parâmetros treinados (Zhang *et al.*, 2024). Porém, é necessário adequar as expectativas sobre os LLM, estudos indicam que ferramentas que usam este modelo têm maior dificuldade em responder questões complexas, que exigem conhecimento de contexto geográfico, manipulação dos dados e operações em ferramentas geográficas, além da conexão de processos aos conceitos (Mooney *et al.*, 2023).

Dentre os modelos de aprendizado de máquina para tratamento de linguagem, os generativos, que utilizam arquitetura *Transformer*, têm apresentado resultados favoráveis, pois sua estrutura facilita associações de dependência entre termos, mesmo que distantes no texto. Este modelo atribui pesos diferentes às palavras, ajudando a capturar o contexto e relações mais distantes, e simula paralelização do processamento, o que reduz a computação sequencial e melhora a qualidade na interpretação e no tempo de resposta. Os LLM utilizam esta arquitetura, como o BERT (*Bidirectional Encoder Representations from Transformers*) e o GPT (*Generative Pre-Trained Transformer*), e são aplicados para ferramentas específicas (Vaswani *et al.*, 2017; Prince, 2023).

Portanto, ferramentas de PLN não interpretam o texto diretamente, pois "o modelo só conhece as estatísticas da linguagem e não entende o significado de suas respostas" (Prince, 2023, p. 7), ou seja, o usuário não controla o resultado, o que requer eficácia nas interações através da *interface* de comunicação (Dang et al., 2022). Assim, a engenharia de *prompt* busca estratégias que melhorem a comunicação humano-máquina no uso da linguagem natural com LLM. Algumas questões relevantes são abordadas por Mooney *et al.* (2023), Wang *et al.* (2023) e Zhang *et al.* (2024): (i) necessidade de diálogos detalhados, com frases curtas e objetivas, pois os resultados são imprevisíveis em função dos modelos não explicitarem os dados de treinamento nem detalhes da arquitetura aplicada; (ii) construção do *prompt* tradicionalmente ser manual, implicando na experiência do usuário sobre o tema ser abrangente, para evitar ambiguidades; (iii) dificuldade de comunicação com formatos diferentes do texto, como diagramas ou figuras; e (iv) complexidade ao lidar com conhecimento específico de domínio mais restrito ou linguagem regional, o que exigiria geração de *prompt*s mais holísticos.

Concomitante a esta discussão, estudos na psicologia também foram essenciais para os avanços na IA, ao apontar que a cognição humana pode estar relacionada a diversos aspectos – intelectual, musical, visual, espacial, de movimento, lógica, matemática, dentro outros –, nem todos ainda simuláveis nas equações matemáticas para representação computacional, em especial questões relacionadas à consciência humana (Vasconcellos e Machado, 2006; Silva e Teixeira, 2022).

No tocante ao conhecimento espacial, cuja natureza é multissensorial, pesquisas em cartografia se aproximaram igualmente da psicologia, para entender como funciona o pensamento na leitura de mapas. A apropriação de um novo conhecimento por um humano depende de seu nível cognitivo para interpretá-lo e, na comunicação cartográfica, é importante entender como os usuários veem aspectos particulares do mapa e como isso varia entre os indivíduos com diferentes percepções do mundo, especialmente considerando variações culturais e sensoriais, o que impacta nos resultados de mapeamentos colaborativos e na integração de bases de dados (Rosch, 1975; Petchenik, 1977; Bravo, 2014; Machado, 2020). O uso de IA, neste processo, é algo a ser explorado, sendo este artigo uma contribuição embrionária para seu desenvolvimento.

Deste modo, para evoluir na interoperabilidade semântica, é preciso avançar nos estudos que avaliem

tecnologias que sirvam como viés facilitador na automatização, ainda que parcial, da compatibilidade entre bases de dados geoespaciais, conforme desafios apontados por Machado & Camboim (2019). Há necessidade de pesquisas voltadas para interpretação automatizada dos conceitos dos objetos que permitam uma associação semântica com menor dependência da interpretação humana.

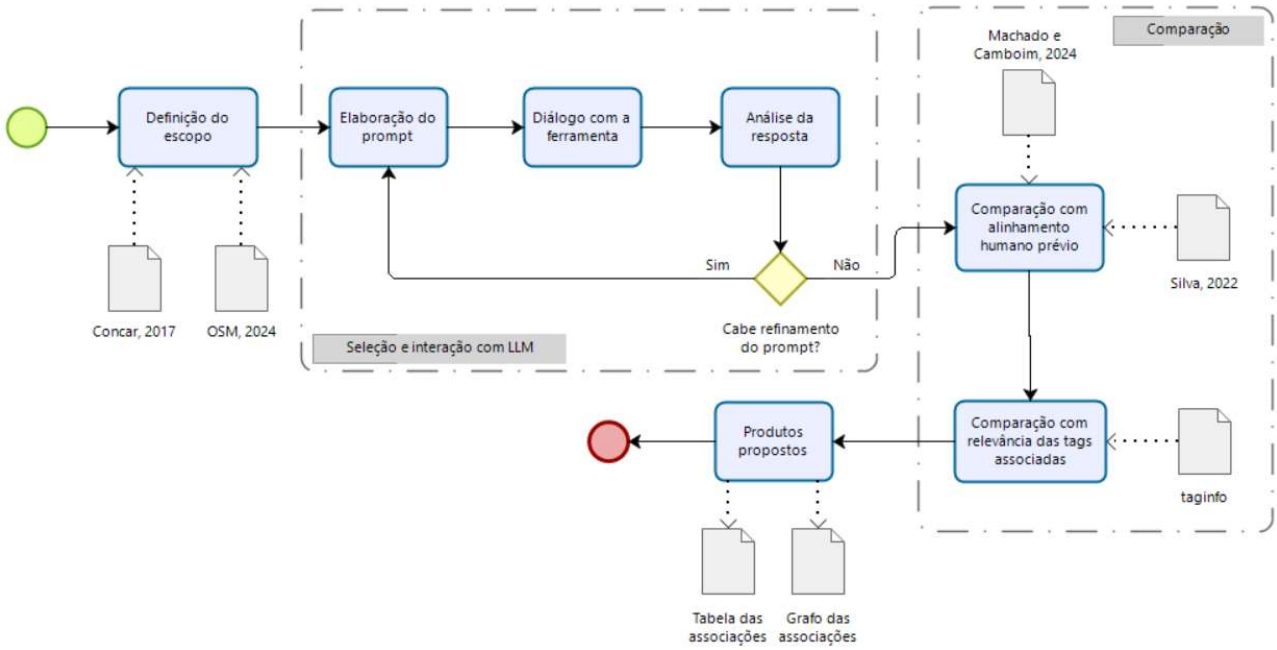
Na literatura pesquisada, não foram identificados exemplos específicos de LLM na compatibilidade semântica de dados geoespaciais. Entretanto, estudos recentes evidenciam o potencial dessas ferramentas para tarefas semelhantes em outras áreas, tais como respostas a perguntas relacionadas a sistemas de informação geográfica por Mooney *et al.* (2023) ou interpretação e correção automática de códigos de programação de Zhang *et al.* (2024). Tais resultados apontam para uma capacidade significativa dos LLMs em compreender linguagem natural e conceitos complexos, sugerindo que essas ferramentas poderiam auxiliar também na interpretação semântica de diferentes modelos de dados geoespaciais, ação atualmente efetuada por humanos.

Portanto, este artigo avança nos estudos anteriores efetuados por (Machado, 2020; Silva, 2022; Machado & Camboim, 2024), investigando especificamente o potencial e os limites do ChatGPT, ferramenta baseada no modelo GPT da OpenAI, para realizar a associação semântica semiautomática entre objetos geoespaciais descritos em diferentes esquemas conceituais. O estudo enfoca especificamente a categoria "Edificações" da ET-EDGV e sua correspondência com os conceitos definidos no projeto colaborativo *OpenStreetMap*. A hipótese central adotada é que o ChatGPT pode simplificar o processo tradicionalmente manual de compatibilidade semântica ao interpretar automaticamente definições textuais dos objetos geoespaciais envolvidos e propor correlações consistentes entre os esquemas estudados. Os resultados obtidos com o ChatGPT serão avaliados comparativamente às associações realizadas por especialistas humanos, permitindo compreender claramente até que ponto a ferramenta contribui para a automatização desse processo.

2 METODOLOGIA

Como mencionado, o estímulo para realização deste artigo foram os trabalhos prévios de Silva (2022) e Machado & Camboim (2024) que realizaram alinhamento semântico da ET-EDGV e do OSM de maneira manual e, em ambos, ficou a provocação de trabalho futuro com automatização a partir do uso de inteligência artificial. A metodologia aplicada segue os passos apresentados na figura 1 e descritos nas sessões 2.1 a 2.5.

Figura 1 – Fluxograma da metodologia.



Elaboração: As autoras (2024).

2.1 Materiais

Desta maneira, para obter associação semântica entre esquemas conceituais de dados geoespaciais com uso de uma ferramenta baseada em LLM, empregou-se a modelagem da ET-EDGV (Concar, 2017) e do *OpenStreetMap* (OSM, 2024). A escolha destes esquemas deu-se por eles terem sido trabalhados na literatura de referência, permitindo seus usos na comparação entre os resultados da IA e os realizados por humanos.

A ET-EDGV foi apresentada anteriormente como a modelagem conceitual dos dados geoespaciais vetoriais das bases de referência no Brasil. Sua estrutura utiliza o modelo orientado a objetos *Object Modeling Technique for Geographic Applications* – OMT-G, que representa entidades geográficas em quatro níveis taxonômicos: escala, categoria, classe (representa os objetos) e atributos (contém domínios de preenchimento e tipos de dados, dentre estes a geometria de representação – ponto, linha ou polígono), além das relações topológicas entre os objetos (Borges *et al.*, 2005; Concar, 2017; Machado & Camboim, 2024).

O OSM atualmente é a ferramenta de contribuição voluntária para mapeamento de maior relevância, sendo um projeto no qual a comunidade cria e mantém um banco de dados livre e editável de representação do mundo e cujos dados estão associados aos contextos pessoais e socioculturais dos colaboradores e comunidade contribuintes. Diversos trabalhos recentes descrevem sua relevância em vários aspectos de representação (Kaur et al., 2017; Grinberger et al., 2022; OSM, 2024). Sua estrutura taxonômica difere da ET-EDGV, sendo o primeiro nível a linguagem utilizada e, posteriormente, o par *key* + *value* (*tag*) que tem como objetivo descrever e identificar um objeto geográfico representado, maiores detalhes em Machado & Camboim (2024).

2.2 Definição do escopo da associação semântica

Iniciou-se este trabalho com a leitura da documentação dos esquemas conceituais (Concar, 2017; OSM, 2024) e a análise dos alinhamentos semânticos já propostos, visando selecionar um subconjunto de dados que fosse representativo para realização dos testes. No caso da ET-EDGV decidiu-se por utilizar todas as 34 classes georreferenciadas da categoria "Edificações", sendo "Edificação" a classe genérica e as demais especializadas a partir desta. Não foram utilizadas as classes Classif_Econ_Administ e Endereco_edif que são convencionais e definem a atividade econômica desenvolvida e o endereço da edificação, respectivamente. Em relação aos conceitos das *tags* do OSM, foram acessados todos disponíveis no seu *site* (OSM, 2024).

A escolha da categoria Edificações deu-se em função da variedade de tipos de classes na ET-EDGV e da diversidade de *tags* e quantidade de objetos representados no OSM que poderiam ser associados. Além disso, é um tema essencial para mapeamento urbano e pode apresentar bastante heterogeneidade em sua representação, conforme visto em Brovelli *et al.* (2016), Elias e Fernandes (2019), Machado & Camboim (2019), Bortolini *et al.* (2020) e Fernandes *et al.* (2020), o que torna relevante a realização de alinhamento semântico adequado. Desta maneira, estando a metodologia definida, esta pode ser replicada para as demais categorias e classes da ET-EDGV sem prejuízo do processo.

Yu et al. (2018) apresentam uma abordagem holística para integrar dados considerando conceitos, propriedades e instâncias simultaneamente, com o uso de quatro métricas distintas para mensurar similaridades semânticas entre esquemas geoespaciais. Os autores definem esses tipos como: (i) similaridade léxica, referente às diferenças terminológicas usadas por diferentes organizações ao publicarem seus dados; (ii) similaridade estrutural, relativa às variações nos padrões e protocolos utilizados, abrangendo desde diferentes formatos de armazenamento até variadas granularidades espaciais e níveis taxonômicos; (iii) similaridade espacial, que ocorre quando as geometrias representativas do mesmo local são significativamente distintas entre as bases de dados; e (iv) similaridade extensional, que avalia a semelhança com base na interseção entre os conjuntos de conceitos, instâncias e propriedades relacionadas aos dados comparados. Uma abordagem semelhante, embora com ênfase adicional em correspondência semântica, de multiplicidade e geométrica, também é discutida por Machado & Camboim (2024).

Entretanto, o presente artigo optou por avaliar exclusivamente a associação semântica utilizando a similaridade léxica. Essa escolha decorre da necessidade de simplificar o processo de análise, permitindo explorar especificamente o potencial do ChatGPT como ferramenta de processamento de linguagem natural (PLN) baseada em *Large Language Model* (LLM). Optou-se por não incluir no escopo deste estudo as questões

associadas às primitivas geométricas, relações topológicas ou variações estruturais dos dados, por serem elementos que introduzem complexidades adicionais à interpretação automática dos conceitos envolvidos.

2.3 Seleção e interação com LLM

No passo seguinte da metodologia, a partir dos preceitos da engenharia de *prompt*, elaborou-se o texto para realização de diálogo com a ferramenta escolhida – ChatGPT[1]. Esta escolha deu-se, principalmente, por usar o GPT, modelo construído na arquitetura *Transformer*, que apresenta ganhos exponenciais de performance e possui uma base de conhecimento ampla, além dos resultados promissores apresentados nos testes preliminares de Souza & Camboim (2023).

Uma questão essencial no *prompt* foi pôr os conceitos da ET-EDGV em um formato de arquivo a ser lido e interpretado pelo ChatGPT, uma vez que ferramentas de PLN não são construídas para ler e interpretar documentos com estruturas diversas. A documentação da ET-EDGV é composta por diagramas OMT-G que representam o modelo conceitual e tabelas de atributos (Concar, 2017). Problema identificado por Mooney *et al.* (2023), uma vez que o PLN interpreta textos em linguagem humana com base no posicionamento e na distância entre as palavras, tendo dificuldade de interpretar outras estruturas de organização do conhecimento.

Cada diálogo realizado com o ChatGPT teve suas respostas analisadas em relação a coerência do resultado e a uma possível necessidade de refinamento na entrada de dados, uma vez que a engenharia de *prompt* sugere o maior detalhamento possível do cenário no momento do questionamento. Diferentes diálogos foram construídos e a pergunta efetuada era ajustada como evolução dos resultados do teste anterior, com variação principalmente no formato de arquivo para a entrada dos conceitos. O resultado foi considerado satisfatório quando no retorno do diálogo havia entendimento dos dados de entrada e indicação de *tags* associadas a todas as classes, sem perguntas complementares. O formato de arquivo mais adequado foi texto sequencial para a ET-EDGV e acesso direto ao *site* OSM, cujas implicações estão na sessão 3.1 deste artigo.

Este último arquivo de texto sequencial foi passado ao ChatGPT, com os conceitos das classes, atributos e domínios da ET-EDGV (figura 2). A seguir, foi solicitada à ferramenta que indicasse as *tags* do OSM que fossem compatíveis semanticamente com cada classe a partir de seus conceitos, atributos e domínios (figura 3). Cada teste de refinamento foi efetuado em *chats* e datas separadas, visando evitar influência na continuidade do diálogo e buscando entender a compreensão do modelo quanto a interpretação dos conceitos.

Figura 2 – Exemplo de texto para os conceitos da ET-EDGV – classe edificação de abastecimento de água.

```
Classe Edificação de abastecimento de água é uma construção componente de um sistema de abastecimento de água.
Atributos de Edificação de abastecimento de água: herda todos os atributos da classe edificação e acrescenta tipoEdifAbast.
tipoEdifAbast - Indica o tipo de edificação de abastecimento e consumo de água. Tendo como domínio: Desconhecido, Administração,
                Captação, Recalque, Tratamento, Misto ou Outros.
```

Fonte: Adaptado de Concar (2017).

Figura 3 – Pergunta efetuada ao ChatGPT (*prompt* de diálogo).

```
Considerando a modelagem de dados geoespaciais ET-EDGV e sua lista de classes, respectivos atributos e domínios de preenchimento, com
os conceitos definidos conforme arquivo anexo:
1- indique as mais adequadas tags (key + value) do OpenStreetMap (conforme conceitos no link
https://wiki.openstreetmap.org/wiki/Map_features) que apresentam compatibilidade semântica com a classe, a partir de seus conceitos,
atributos e domínios;
2- apresente as tags indicadas de maneira ordenada para as associações semânticas efetuadas, de acordo com um ranking de maior
probabilidade de associação de cada tag;
3- faça para todas as 34 classes listadas no anexo.
```

Elaboração: As autoras (2024).

Outro aspecto relevante é que o esquema conceitual da ET-EDGV, sendo a referência para cartografia no Brasil, está em língua portuguesa e, por este motivo, o diálogo com o ChatGPT deu-se nesta língua, enquanto o OSM utiliza o inglês, dada sua aplicação internacional. Apesar das diferenças linguísticas, observou-se que o modelo de linguagem foi capaz de realizar as associações semânticas corretamente entre os esquemas sem necessidade de instruções adicionais relacionadas à tradução, ainda que não tenha sido realizada uma avaliação quantitativa ou qualitativa específica da acurácia destas traduções.

[1] Através de comunicação no site https://chat.openai.com

2.4 Comparação dos resultados

Para avaliar as respostas apresentadas pelo ChatGPT, as associações semânticas entre classes e *tags* indicadas foram comparadas com as associações efetuadas por humanos, de maneira manual – Machado & Camboim (2024) e Silva (2022) –, separando as *tags* que tiveram indicação de associação pela ferramenta e pelas análises humanas, quais foram indicadas apenas pelo ChatGPT ou apenas pelos autores.

Finalmente, para entender se as associações propostas pelo ChatGPT são relevantes em relação às *tags* mais aplicadas no OSM, utilizou-se a ferramenta *taginfo*[2], que apresenta uma análise quantitativa do uso das *tags*. Esta avaliação esteve embasada nos dados cadastrados no OSM no Brasil no período de 16 a 23 de abril de 2024 e buscou entender se o ChatGPT ofertou opções de *tags* com grande volume de aplicação pelos usuários brasileiros. Nos resultados, *tags* indicadas com * no lugar do *value* foram consideradas 'ambíguas', uma vez que este valor pode indicar qualquer tipo de preenchimento além de uma feição que represente uma edificação. Para *tags* que não apresentavam informações no *taginfo* estas foram indicadas como 'sem feição'.

2.5 Produtos propostos

Após a realização dos trabalhos, dois produtos derivados foram gerados: tabela e grafo das associações semânticas sugeridas pelo ChatGPT (sessões 3.2 e 3.3). A tabela indica a classe ET-EDGV e as *tags* sugeridas pelo ChatGPT, por Machado & Camboim (2024) e por Silva (2022) e a quantidade de dados no *taginfo*.

O resultado na forma de grafo pretende apresentar de maneira mais adequada a hierarquia conceitual da ET-EDGV e as associações feitas com as *tags* do OSM, considerando que estes esquemas possuem estruturas taxonômicas bastante diversas e com múltiplas possibilidades de conexão, conforme discutido por Machado & Camboim (2024). Um grafo é uma estrutura matemática que representa relações entre objetos, onde os nós representam estes objetos e as arestas representam suas conexões (Longley *et al.*, 2011).

Neste grafo, considerou-se que os nós representam as classes, atributos ou domínios da ET-EDGV e as *tags* propostas para associação semântica; enquanto as arestas representam os tipos de associação que podem existir dentro do mesmo esquema ou entre diferentes esquemas. Utilizou-se o *software* gratuito Cytoscape e os dados das associações foram organizados em tabela para importação no *software* no sentido origem e destino das relações, tanto na hierarquia interna da ET-EDGV (classe genérica, classe especializada, atributo e domínio) quanto nas relações propostas entre uma classe ou domínio com uma *tag* do OSM.

3 RESULTADOS E DISCUSSÕES

A partir da metodologia, alguns resultados relevantes foram identificados e serão discutidos a seguir.

3.1 Elaboração do *prompt*

Dado o caráter exploratório do estudo, diversos testes foram realizados para refinamento dos comandos (*prompts*) inseridos no ChatGPT. Inicialmente, comandos mais genéricos foram aplicados, porém os resultados apresentavam correspondências semânticas superficiais e pouco específicas. Progressivamente, foi necessário detalhar mais claramente as estruturas conceituais dos esquemas utilizados, e explicitar ao modelo quais aspectos semânticos deveriam ser priorizados. Por exemplo, o quadro 1 apresenta um resumo das estratégias de diálogo e equívocos observados nas respostas, enquanto o Apêndice A[3] detalha alguns diálogos completos, demonstrando como o refinamento progressivo dos comandos e as mudanças de versão da ferramenta (ocorridas no decorrer da realização dos testes), levaram a respostas mais detalhadas e alinhadas semanticamente. A partir dessa abordagem iterativa e com a evolução de suas versões, observou-se significativa melhoria na coerência e detalhamento das respostas do ChatGPT ao longo das interações.

[2] https://wiki.openstreetmap.org/wiki/*Tag*info

[3] https://anonymous.4open.science/r/SemanticachatGPT-072A/).

Quadro 1 – Resumo dos diálogos iterativos realizados com o ChatGPT.

Testes	ChatGPT	Estratégia do diálogo	Exemplo	Análise das respostas
1 a 5	3.5	Os testes iniciais carregavam no ChatGPT uma tabela com a estrutura dos dados da ET-EDGV (edificação geral e de saúde) e outra tabela com a estrutura do OSM, conforme correspondência em Machado & Camboim (2024), e solicitava a associação semântica do conteúdo das tabelas.	Apêndice A Testes 01 e 04	Embora entenda 'edificação' associada a 'building', o ChatGPT cria tipos para associar ao OSM, como 'care level' ou 'primary' ou 'Construction Material Type', que não existem no esquema OSM. Aparentemente efetua apenas uma tradução literal do objeto da ET-EDGV, forçando uma associação.
6 a 14	3.5	Em sequência foram testadas as seguintes alterações: utilizar as tabelas de entrada com o código em formato JSON, realizar as perguntas em inglês, explicitar as colunas da ET-EDGV para associação.	Apêndice A Teste 08	A adoção do formato JSON e do idioma inglês não comprometeu a compreensão do conteúdo, mantendo-se equivalente aos testes anteriores quanto à clareza da leitura. Os resultados seguem indicando dificuldades de interpretação, com associações genéricas - como a vinculação de 'edificação' à tag 'building' - e a sugestão de correlações com tags inexistentes, aparentando possíveis associações forçadas a partir de traduções literais dos termos. O ChatGPT indica a existência de correspondências múltiplas, mas não explicita a relação entre atributos ou domínios como desdobramentos das classes envolvidas. Para se obter respostas relevantes, é necessário realizar diversas perguntas, sendo que, em geral, as respostas apresentadas são semelhantes entre si e de caráter evasivo.
15 a 23	3.5	Em sequência foram testadas as seguintes alterações: retornar as perguntas em português, fragmentar e detalhar em várias perguntas, manter a entrada de dados em formato JSON e solicitar a saída das associações em código Python (a utilização destes formatos pretendia avaliar qual a melhor estrutura de dados para comunicação e automatização futura do diálogo).	Apêndice A Teste 18	Até este momento, observou-se que o ChatGPT consegue ler toda a estrutura de entrada, mesmo com variação de formato. Entretanto, faz-se necessária a realização de mais de uma pergunta para que o raciocínio apresente uma resposta coerente, ainda que parcial. Aparentemente baseada no contexto da escrita, a ferramenta consegue relacionar alguns temas entre os modelos, mas não compreende a hierarquia inerente a um deles individualmente. Por exemplo, para a ET-EDGV, o conceito de 'primário' é parte de 'edificação de saúde' e esta associação hierárquica não é percebida nos comentários do ChatGPT. Embora a ferramenta consiga associar com a edificação de saúde e a seus níveis de atenção, valores do OSM como tipos de locais ligados à saúde (ver pergunta 05 do apêndice A no teste 18). Por fim, o código Python solicitado para a saída dos resultados foi parcial em relação ao raciocínio realizado e apresentou falhas.
24 a 30	4.0	Em sequência foram testadas as seguintes alterações: utilização apenas das classes edificação geral, de saúde ou de ensino; entrada de dados variando nos formatos JSON e OWL; solicitar os resultados na forma de planilha, cujas colunas indiquem a associação, e que estas sejam salvas no drive do Gmail.	Apêndice A Teste 24	Nestes testes, o raciocínio nas associações apresentou evolução, não apontando correlações inexistentes, mas indicando associações básicas, como para edificação geral ('Building') e de saúde ('healthcare'). Contudo, o formato de saída pretendido não foi adequado, pois as planilhas geradas estavam vazias. Outro ponto negativo foi a necessidade de elaborar diversas perguntas para retorno coerente. Dois destes testes foram repetidos por apresentar 'erro de network' do ChatGPT.
31 a 36	4.0	Em sequência foram testadas as seguintes alterações: entrada de dados do esquema ET-EDGV como texto sequencial ou código em XML oriundo de modelagem no OMT-G Designer (apenas o teste 36); entrada de dados do esquema OSM consultado direto no site oficial; expansão para todas as classes da categoria Edificações; solicitação de prioridade nos resultados da	Apêndice A Testes 35 e 36	Ao reestruturar a entrada de dados dos esquemas, sendo o site para o OSM e texto sequencial para a ET-EDGV, as respostas de argumentação e as associações sugeridas ficaram mais adequadas. O teste 34 foi o primeiro com coerência na resposta, mas limitado em apenas cinco classes associadas. O teste 35 apresentou associações entre diferentes classes e tags, contudo, apenas após a realização de várias perguntas. Ao final, responde que outros formatos de entrada dos dados poderiam ser mais

		associação e determinação de limite de tags associadas para a resposta.		adequados (o que não foi comprovado nos testes que aplicaram estes formatos). O teste 36, em XML, apresentou as mesmas limitações anteriores.
37	4o	Manteve a estratégia do teste 35, entretanto, solicitando um máximo de três tags por classe associada.	-	O resultado apresentou todo o conjunto de classes da categoria Edificações com associações a exatamente três tags do OSM. Todavia, a cada grupo de 04 ou 05 classes associadas, perguntava se deveria continuar o trabalho. Esta limitação pode indicar controle na quantidade de processamento.
38	4o	Manteve a estratégia do teste 37, porém, removendo a limitação de três tags por classe e explicitando que a associação deve ocorrer para todas as classes indicadas na listagem de entrada.	-	Resultados foram considerados adequados e a discussão é apresentada neste trabalho.

Elaboração: As autoras (2024).

Os testes preliminares realizados por Souza & Camboim (2023) reorganizaram o formato de texto existente para os conceitos das classes da ET-EDGV e das *tags* OSM em texto sequencial, apenas para algumas classes e *tags* previamente associadas pelos autores de referência. Apesar de resultados promissores, este formato reduzia as possibilidades de novas associações por parte da ferramenta, pois não explorava todas as *tags* do OSM e não incluía atributos e domínios da ET-EDGV, nem suas relações hierárquicas.

Assim, conforme quadro 1, outros testes foram efetuados, inicialmente (1 a 5) com um subconjunto dos dados, agora estruturados em duas planilhas: uma com a indicação do nome da classe, seu conceito, atributos e domínios da ET-EDGV, e outra com o par *key+value* de uma *tag* e seus conceitos, apenas para as *tags* cuja associação com as classes estava baseada nos trabalhos de referência. Com as planilhas os resultados foram superficiais, embora o ChatGPT lesse os conteúdos de entrada, não retornava respostas coerentes na correlação de uma classe, atributo ou domínio com uma *tag*. Efetuou-se novos testes (6 a 30 e 36) com os dados de entrada nos formatos OWL (ontologia), JSON e XML modelado pelo OMT-G Designer, todos com respostas inadequadas: vazias, sem conexão semântica coerente (ex: domínio da ET-EDGV 'Planejado' com *key* do OSM '*college'*) ou com uma associação genérica (ex: edificação na ET-EDGV para a *key* '*building'*, sem considerar outras classes ou *tags*), entre outras incoerências listadas no quadro 1.

Os resultados mais promissores ocorreram quando os testes (31 a 38 do quadro 1) usaram o formato de texto sequencial sem formatação para os conceitos da ET-EDGV e leitura direta no site para as *tags* do OSM. Nesta organização foram feitos alguns testes com variação na pergunta até esta ter uma resposta com associações para todas as classes da categoria Edificações sem necessidade de perguntas complementares (teste 38). Por fim, o ChatGPT retornou um *ranking* com três ou quatro *tags* compatíveis semanticamente, em ordem de maior probabilidade de associação com cada classe, a partir da única pergunta efetuada (figura 3).

Observou-se, neste processo, a importância da pergunta ser clara, abrangente, sem ambiguidades, bem estruturada para os conceitos aplicados, conforme frisa Dang *et al.* (2022) e Wang *et al.* (2023). Além destes fatores, para realizar associação semântica, neste contexto de uso, o formato de texto sequencial mostrou-se mais adequado para interpretação dos conteúdos pelo ChatGPT do que os demais avaliados. Concomitantemente, pode ter havido influência de melhorias na ferramenta em função de sua evolução de versão.

3.2 Produtos derivados

Os resultados apresentados pelo ChatGPT e a comparação com os trabalhos de referência podem ser vistos nos produtos disponíveis em https://anonymous.4open.science/r/SemanticachatGPT-072A/. A tabela das associações semânticas (Apêndice B) está exemplificada no recorte da figura 4, cujas colunas indicam a classe ET-EDGV, as *tags* do OSM indicadas pelo ChatGPT e ordenadas por relevância, as associações apontadas por Machado & Camboim (2024) e Silva (2022) e a quantidade de dados do *taginfo* no Brasil.

De maneira complementar à tabela, o segundo produto gerado foi o grafo das associações semânticas, que pode ser observado de maneira geral na figura 5 e posteriormente com recortes por classe nas figuras 6 a 8. Este grafo representa numa visão genérica a quantidade de associações existentes, tanto na hierarquia dos

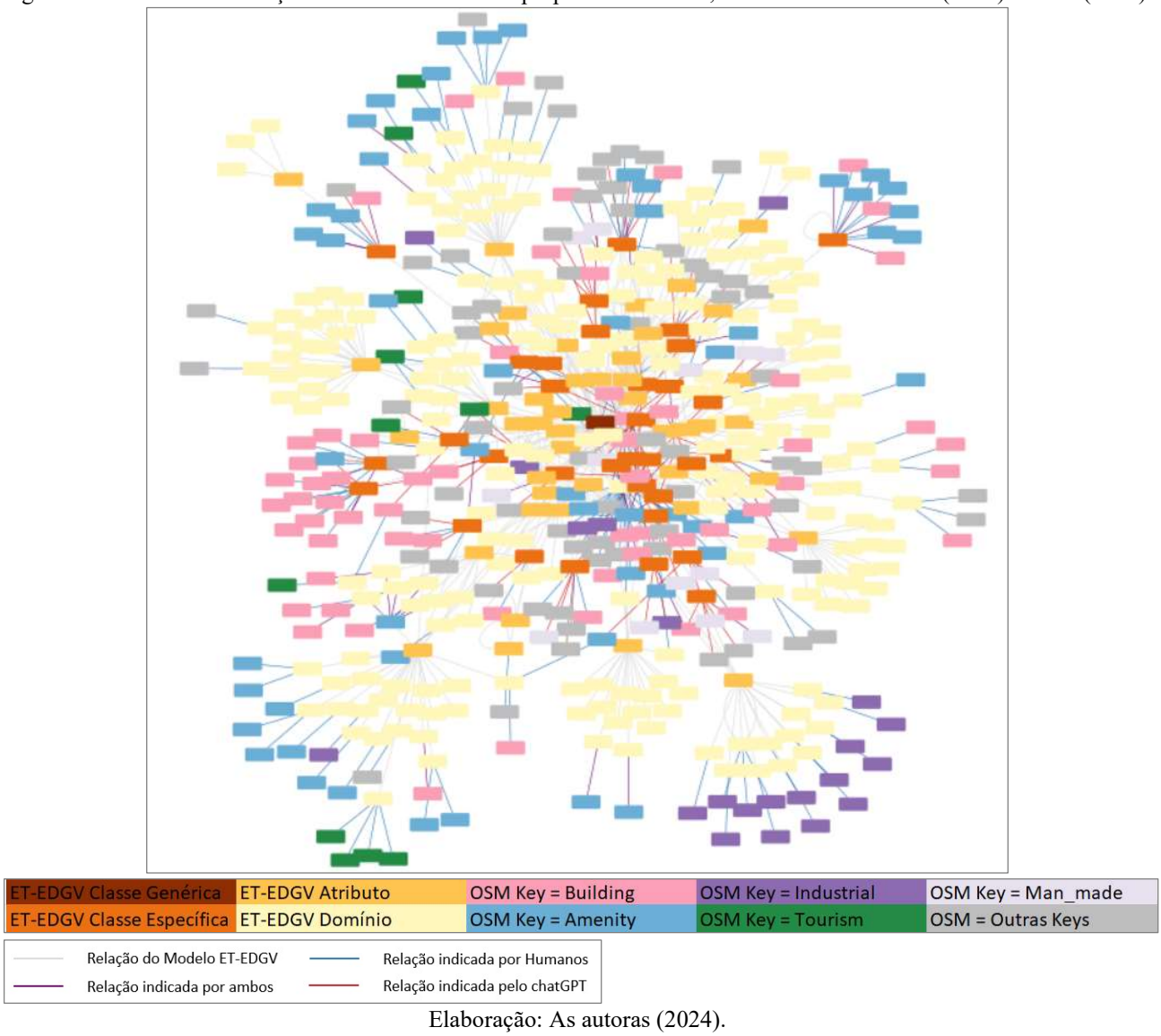
elementos de composição do esquema ET-EDGV, quanto na relação semântica das classes, atributos ou domínios com as *tags* do OSM, dentro da perspectiva proposta somente por Machado & Camboim (2024) e Silva (2022), somente pelo ChatGPT ou tanto pela ferramenta quanto por humanos. O grafo da figura 5 permite visualizar, por exemplo, uma predominância de nós referentes aos domínios dos atributos das classes da ET-EDGV ou a concentração de *tags* de uma mesma *key* associadas a uma determinada classe ou domínio.

Figura 4 – Exemplo da comparação do resultado do ChatGPT com as associações semânticas feitas por humanos.

ET-EDGV / CLASSE	CHATGPT - TAG DO OSM (KEY + VALUE) RANKING DAS TAGS MAIS RELEVANTES	MACHADO E CAMBOIM, 2024 ASSOCIAÇÃO COM A CLASSE OU COM O DOMÍNIO DE ATRIBUTO	SILVA, 2022	TAGINFO - BRASIL
Edificação de Ensino	amenity+school	Pelo atributo classeAtivEcon = ensino fundamental e médio: amenity+school		amenity+school - 55771
	amenity+university	Pelo atributo classeAtivEcon = graduação e pós: amenity + university		amenity+university - 3682
	building+school	building+school	building+school	building+school - 21330
	amenity+kindergarten	Pelo atributo classeAtivEcon = Creche e pré-escola: amenity + kindergarten		amenity + kindergarten - 3527
		Pelo atributo classeAtivEcon = Creche e pré-escola: building + kindergarten Ensino médio: amenity + college Graduação e pós: building + university Outros: amenity + driving_school Outros: amenity + language_school Outros: amenity + music_school	building+kindergarten amenity+college building+university amenity+language_school amenity+music_school	building + kindergarten - 589 amenity+college - 2367 building + university - 9027 amenity + driving_school - 729 amenity + language_school - 628 amenity + music_school - 224
Edificação de Polícia	amenity+police	amenity+police	amenity+police	amenity+police - 6702
	building+public			building+public - 5311
	office+government			office+government - 8951
		amenity+fire_station		amenity+fire_station - 1043
		Pelo atributo tipoEdifPol = Prisional: amenity + prison		amenity+prision - 819

Elaboração: As autoras (2024).

Figura 5 – Grafo das associações ET-EDGV e OSM - propostas ChatGPT, Machado & Camboim (2024) e Silva (2022).



Elaboração: As autoras (2024).

3.3 Análise das associações semânticas resultantes

A resposta apresentada pelo ChatGPT para a pergunta efetuada (figura 3), em geral, trouxe coerência entre a classe e as *tags* apontadas como compatíveis semanticamente, sendo a maioria das *tags* com termos da *key* ou do *value* diretamente relacionados ao tema da classe, podendo em alguns casos ser associada a um atributo ou domínio, mas não indicado de maneira explícita. Por exemplo, Edificação de Comunicação foi associada a *man_made+communications_tower*, Edificação ou Construção de Estação de Medição de Fenômenos com *man_made+monitoring_station*, Edificação Religiosa com *building+church* e *building+ temple*, Edificação de ensino com *amenity+school*, *amenity+university*, *building+school* e *amenity+ kindergarten* (esta última classe vista em detalhe na figura 4).

Ocorreu apenas uma associação cujo significado da *tag* não tem aderência à classe, que foi Edificação Agropecuária, de Extrativismo Vegetal e/ou Pesca com *amenity+marketplace*. O que pode ser indicativo da influência de outros textos próximos ao conceito desta classe nos dados de entrada, uma vez que na estruturação do arquivo de conceitos da ET-EDGV, logo após esta classe vem Edificação de Comércio ou Serviços, na qual o atributo de tipos de comércio tem um domínio de supermercado.

De maneira complementar, observou-se que as traduções (inglês/português) efetuadas para a associação foram coerentes dentro dos resultados apresentados, ainda que algumas associações tenham sido mais genéricas, como é o caso de Posto da Polícia Rodoviária Federal com as *tags*: *amenity+police*, *building+public* e *office+government*. Embora nenhuma das *tags* indicadas aponte relação com fiscalização de tráfego em rodovias, a ferramenta entende que a classe está relacionada à gestão pública e policiamento.

Outro aspecto importante observado foi que as *tags* indicadas pelo ChatGPT neste último teste considerado satisfatório não eram exatamente as mesmas *tags* apontadas nos testes anteriores que já apresentavam alguma coerência na resposta, ainda que com conexão dentro do contexto.

Quando comparando os resultados do ChatGPT com os trabalhos de Machado & Camboim (2024) e Silva (2022), percebe-se que houve nexo nas associações feitas pelo modelo de linguagem em relação às correspondências baseadas em análise humana, mesmo que haja *tags* sugeridas pelo modelo que não façam parte das *tags* apontadas pelos autores. O quadro 2 apresenta um resumo das correlações entre as *tags* sugeridas pelo ChatGPT e a quantidade de *tags* com mesmas associações nos trabalhos de referência, bem como a quantidade de classes onde houve ou não estas associações, para um total de 34 classes.

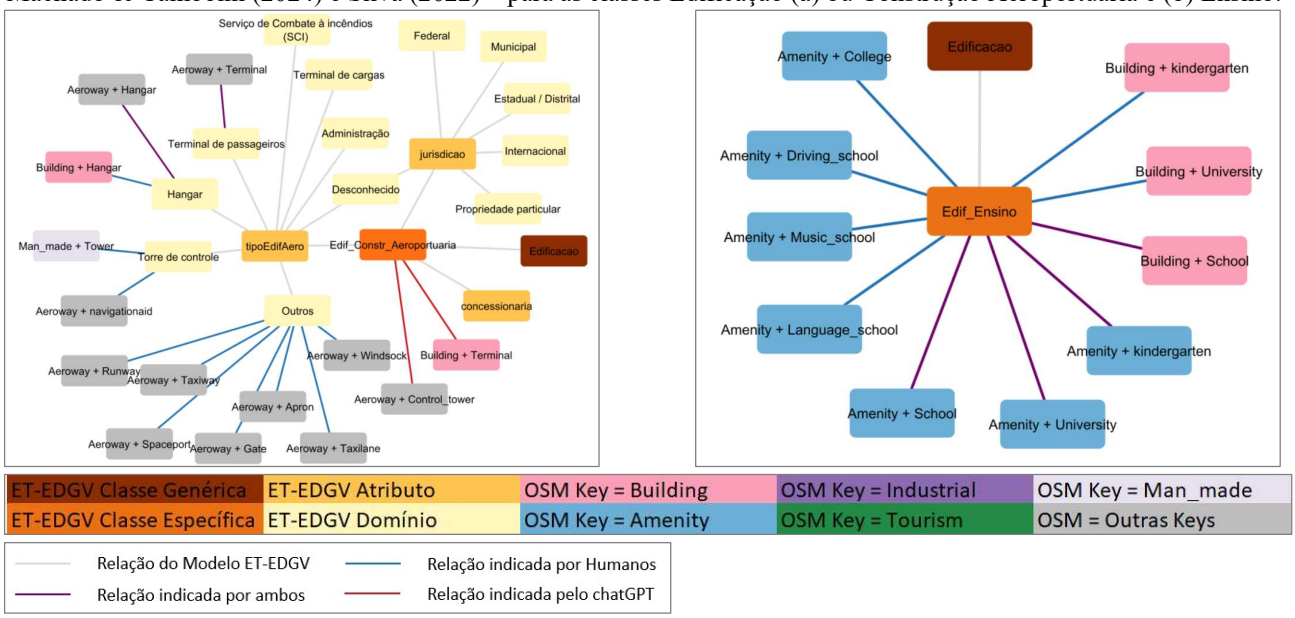
Quadro 2 – Comparação ChatGPT e trabalhos realizados por humanos: Machado & Camboim (2024) e Silva (2022).

ITEM	COERÊNCIA DO CHATGPT COM HUMANOS	QTD DE *TAGS*	QTD DE CLASSES
1	*Tags* indicadas pelo ChatGPT são correspondentes às *tags* indicadas por humanos	44	27 de 34
1.1	Das quais todas as *tags* indicadas pelo ChatGPT são correspondentes às *tags* indicadas por humanos	8	2 de 27
2	*Tags* indicadas pelo ChatGPT são diferentes das *tags* indicadas por humanos	79	32 de 34
2.1	Das quais todas as *tags* indicadas pelo ChatGPT são diferentes das *tags* indicadas por humanos	26	7 de 32
3	*Tags* indicadas por humanos não foram indicadas pelo ChatGPT	152	27 de 34

Elaboração: As autoras (2024).

Observando o item 1 do quadro 2, o resultado do ChatGPT foi coerente com os trabalhos humanos em 27 das 34 classes ET-EDGV utilizadas, para um total de 44 *tags* apontadas, ainda que haja outras associações sugeridas – exemplo na figura 6a para Edificação ou Construção Aeroportuária. Pode-se observar, no item 1.1 do quadro 2, que todas as *tags* indicadas pelo ChatGPT correspondem a indicações humanas nas classes: Edificação Pública Civil e de Ensino (esta última na figura 6b).

Figura 6 – Grafos das associações semânticas entre ET-EDGV e OSM, conforme proposições do ChatGPT e de Machado & Camboim (2024) e Silva (2022) – para as classes Edificação (a) ou Construção Aeroportuária e (b) Ensino.



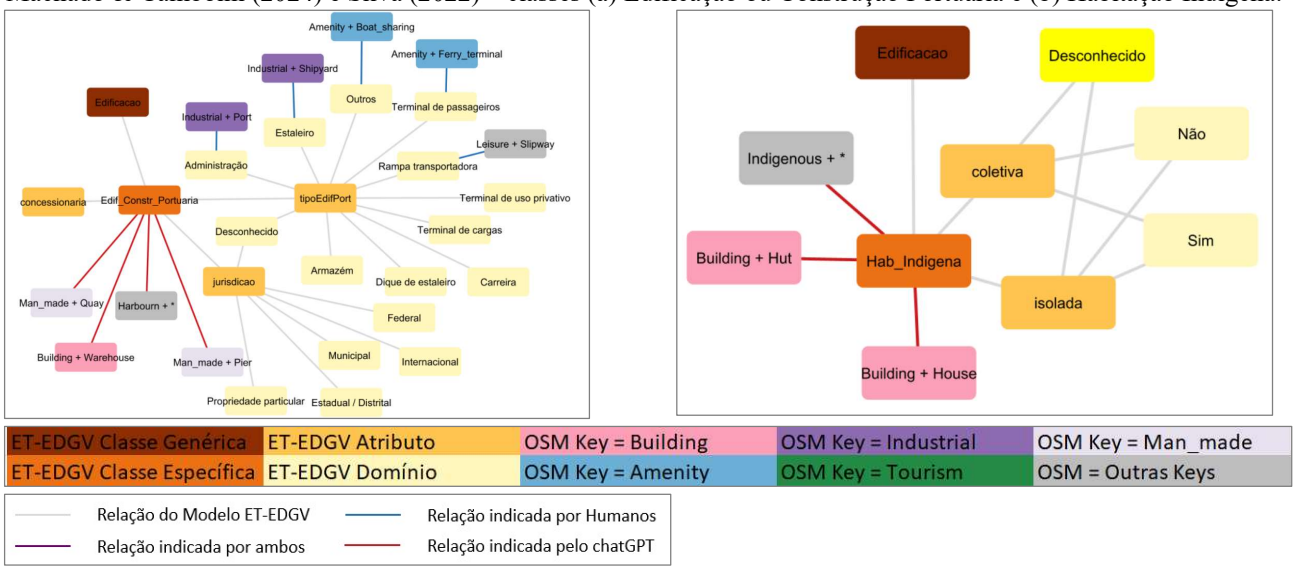
Elaboração: As autoras (2024).

Por outro lado, de acordo com o item 2 do quadro 2, tem-se que 79 *tags* apontadas pelo ChatGPT não foram sugeridas pelos trabalhos humanos, onde 78 tem coerência em relação ao conceito da classe apontada, a exemplo de Edificação ou Construção Aeroportuária associada com *aeroway+control_tower* e *building+terminal* (figura 6a). A exceção neste caso foi a situação já discutida de *amenity+marketplace* com Edificação Agropecuária, de Extrativismo Vegetal e/ou Pesca.

Neste total de 79 novas *tags* apontadas pelo ChatGPT, a análise 2.1 do quadro 2 aponta que 26 *tags* em 7 classes só foram sugeridas pela ferramenta, sendo para as classes Edificação de Abastecimento de Água, Portuária, de Energia, Pública Militar, de Comunicação, Indígena e Posto da Guarda Municipal, onde as três últimas não estavam no escopo dos trabalhos de Machado & Camboim (2024) nem de Silva (2022), mas tiveram resultados coerentes na resposta.

Por exemplo, para Edificação ou Construção Portuária as associações do ChatGPT foram lógicas e detalhistas com a classe, onde o modelo de linguagem associou com *man_made+pier*, *building+warehouse*, *man_made+quay* e *harbourn+** (figura 7a), mas nenhuma destas opções aparecem como *tags* de compatibilidade apontadas nas análises humanas. Enquanto para Habitação Indígena não houve sugestão humana de associação, mas a ferramenta trouxe *building+hut*, *building+house* e *indigenous+** (figura 7b).

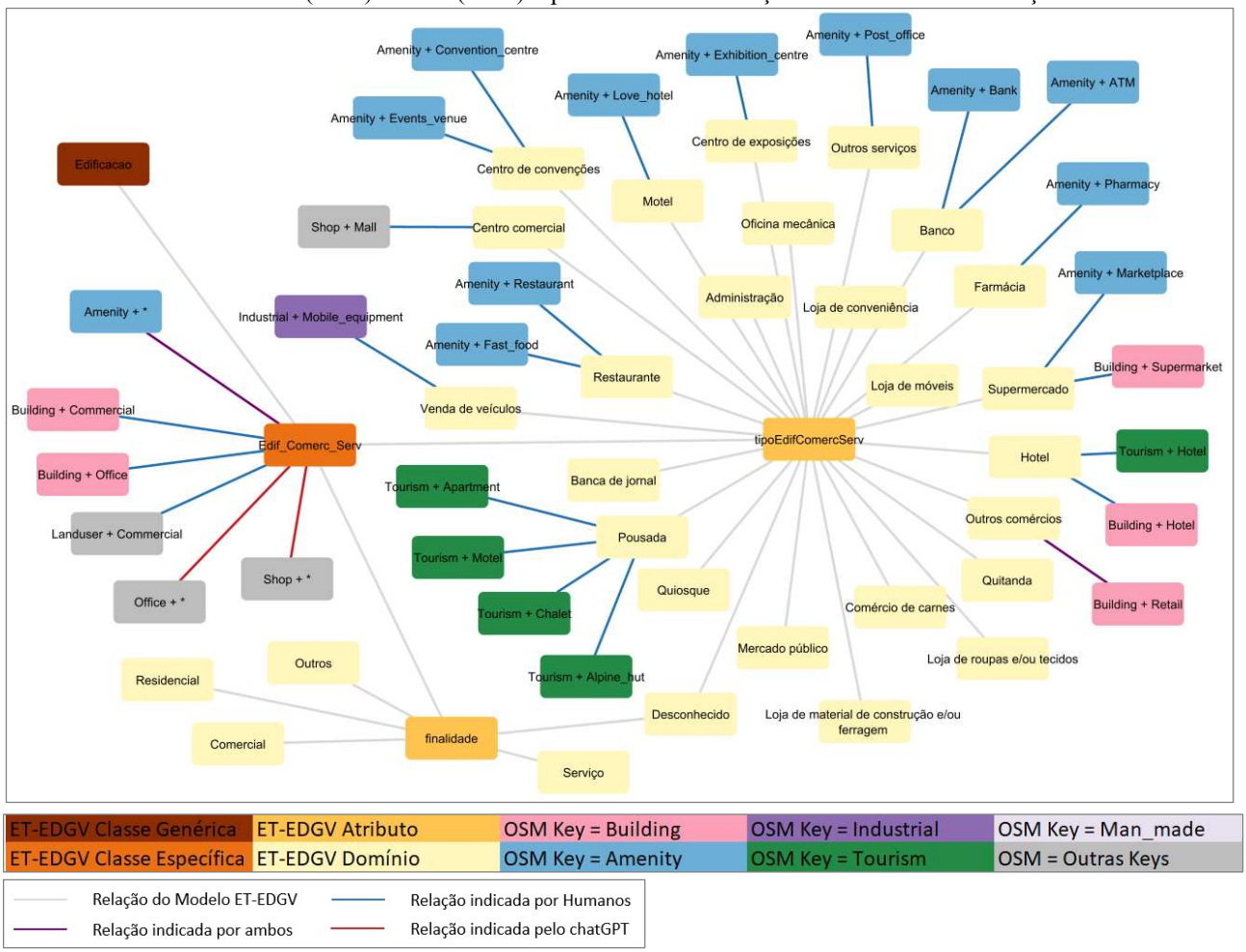
Figura 7 – Grafos das associações semânticas entre ET-EDGV e OSM, conforme proposições do ChatGPT e de Machado & Camboim (2024) e Silva (2022) – classes (a) Edificação ou Construção Portuária e (b) Habitação Indígena.



Elaboração: As autoras (2024).

Ainda no quadro 2, o item 3 indica a quantidade de *tags* (152) e de classes (27) impactadas quando estas foram indicadas por Machado & Camboim (2024) e/ou Silva (2022), mas não aparecem no ChatGPT. Nestes casos, há variações da *key* e não do *value*, como *building+toilets* apontada por ambos para Banheiro Público, mas no ChatGPT aparece como *amenity+toilets*. Mas também há indicação de diferentes *tags*, a exemplo de Edificação ou Construção de Lazer, ou Construção Turística, Industrial e de Comércio ou Serviços (esta última exemplificada na figura 8). Em geral, esta variação ocorre porque os referidos autores trouxeram outras possibilidades de compatibilidade, detalhando a associação no nível de atributo e domínio, enquanto o ChatGPT se limitou a 3 ou 4 *tags* por classe na resposta.

Figura 8. Grafo das associações semânticas entre ET-EDGV e OSM, conforme proposições do ChatGPT e de Machado & Camboim (2024) e Silva (2022) – para a classe Edificação de Comércio ou Serviços.



Elaboração: As autoras (2024).

De maneira geral, observou-se que a ferramenta é capaz de entender adequadamente o significado dos termos em seus contextos e realizar associações semânticas entre esquemas conceituais distintos, como ET-EDGV e OSM, mesmo considerando a barreira linguística. No entanto, sua capacidade parece ser limitada quando precisa interpretar questões implícitas relacionadas à estrutura interna hierárquica dentro do mesmo modelo. Tal limitação pode estar associada ao funcionamento intrínseco dos modelos de linguagem, que operam com base em padrões estatísticos das correlações entre palavras, em detrimento do entendimento profundo das relações conceituais estabelecidas na cognição humana. Nesse sentido, as reflexões trazidas por Rosch (1973, 1975) sobre a categorização cognitiva humana tornam-se fundamentais para compreender tais resultados. A estrutura cognitiva humana, conforme Rosch destaca, organiza o conhecimento em diferentes níveis de abstração e hierarquia, habilidade que aparentemente ainda não é plenamente replicada por ferramentas baseadas em modelos de linguagem, como o ChatGPT.

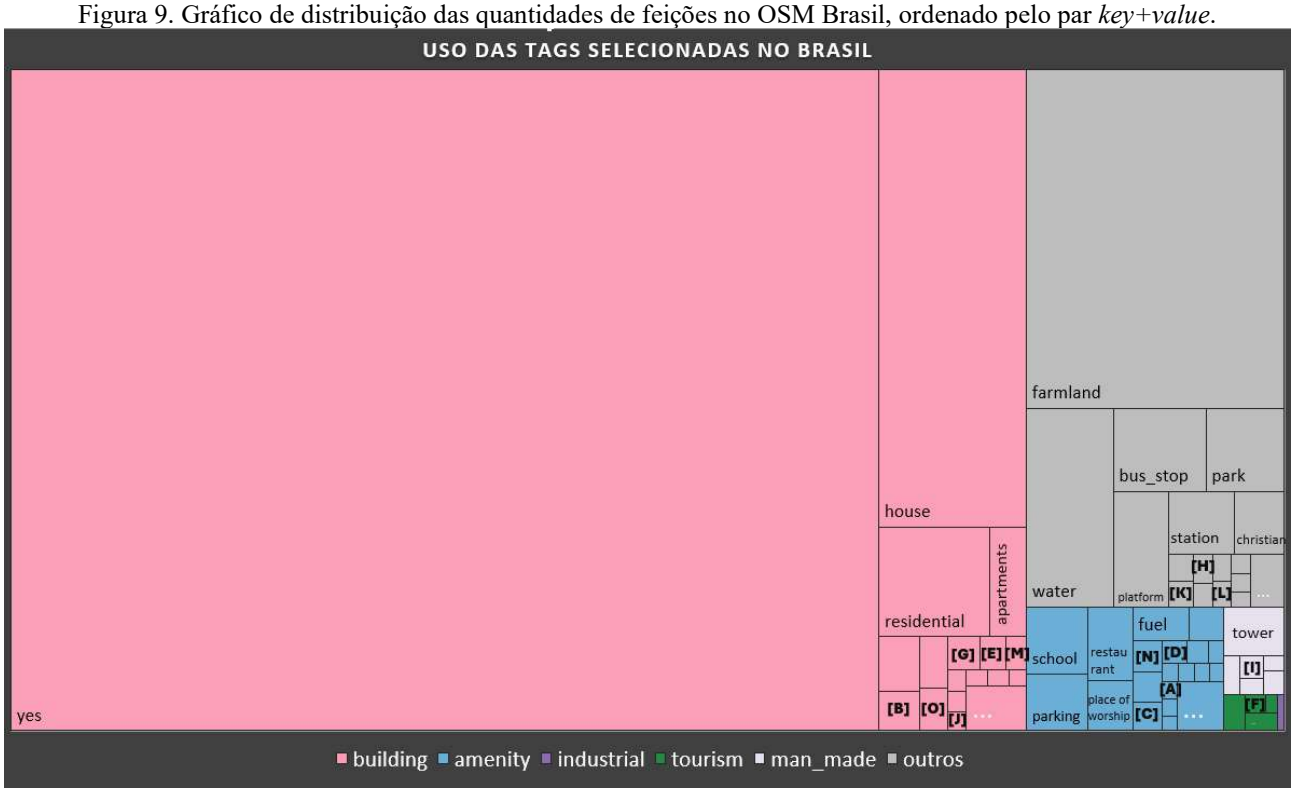
Um exemplo prático dessa limitação é observado no fato de que as classes especializadas da ET-EDGV, ao herdarem atributos gerais da classe Edificação (como elementos culturais ou turísticos), não tiveram essas relações hierárquicas reconhecidas ou explicitamente consideradas nas respostas fornecidas pela

ferramenta. Essa constatação sugere uma dificuldade em interpretar adequadamente o sentido subjacente às relações de herança entre objetos dentro da mesma ontologia, reforçando a necessidade de estudos adicionais que explorem com mais profundidade como os modelos de linguagem possam ou não replicar processos cognitivos mais complexos, como os apontados por Petchenik (1977), Bravo (2014) e Machado (2020).

3.4 Análise da relevância no OSM através da quantidade de tags cadastradas

Finalmente, quando considerando o volume de feições cadastradas no OSM para o Brasil, usando as *tags* apontadas pelo ChatGPT e pelos autores de referência, a partir dos dados do *taginfo* no período analisado, pode-se observar que, de maneira geral, existe uma predominância da *key = building* (figura 9 – rosa), com maior destaque para a *tag* completa *building+yes*, obviamente por ser o objeto mais genérico que representa edificações no OSM. Importante salientar que uma considerável parte das feições está associada às *tags* apontadas pelo ChatGPT, mesmo quando o par *key+value* é mais detalhado, independente do *ranking* de prioridade. Por exemplo, em Banheiro Público, a *tag amenity+toilets* (marcada com a letra **[A]** na figura 9) foi indicada pela IA e pelos humanos e tem 3.639 feições, enquanto a *tag building+toilets* apontada apenas por humanos tem somente 52 feições.

Figura 9. Gráfico de distribuição das quantidades de feições no OSM Brasil, ordenado pelo par *key+value*.



Elaboração: As autoras (2024).

Obviamente *tags* mais genéricas ou que representam elementos predominantes na paisagem, como residências e centros de ensino, saúde e religião, tendem a ter maior representatividade, como *building+yes* (7.786.749), *building+house* (916.982) e *building+residential* (164.164), *amenity+school* (55.771) e *building+school* (21.330) [B], *amenity+clinic* (11.313) [C] e *amenity+hospital* (7.399) [D], *amenity+place_of_worship* (45.307) e *building+church* (11.661) [E], ou ainda *tourism+attraction* (4.948) [F]. Entretanto, *tags* bastante específicas apontadas pelo ChatGPT também apresentam relevância em seu uso no contexto do *taginfo*, como *amenity+fuel* (25.883), *building+retail* (14.869) [G], *leisure+sports_centre* (7.981) [H], *building+farm_auxiliary* (3.103), *amenity+social_facility* (2.340), *aeroway+hangar* (1.388) ou *man_made+monitoring_station* (997).

Há casos em que a *tag* foi apontada apenas pelo ChatGPT e tem resultados expressivos como é o caso de *man_made+pier* (7573) [I] e *building+warehouse* (4956) [J], *power+substation* (8.297) [K], *landuse+quarry* (6299) [L] ou *military+barracks* (324). Embora apenas quatro *tags* apontadas pelo ChatGPT não

tenham feições registradas no Brasil: *communication+transmitter*, *aeroway+control_tower, building+terminal, amenity+theater.*

Também há *tags* indicadas somente por Machado & Camboim (2024) e Silva (2022) que apresentam grande volume de indicações de uso, como *building+university* (9.027) [M], *amenity+pharmacy* (12.279) [N], *amenity+restaurant* (31.059), *building+farm* (16.114) [O], *religion+christian* (42.558) ou *natural+water* (233.840). Esta última, uma *tag* mais genérica sobre água, que os autores indicam o uso conjunto com *water+wastewater*, para associar com edificação de saneamento, tendo a mesma apenas 1.463 registros.

4 CONSIDERAÇÕES FINAIS

Analisando o objetivo principal deste artigo, que buscou avaliar se uma ferramenta de processamento de linguagem natural baseada em grandes modelos de linguagem tem potencial para automatização parcial da associação semântica entre dois esquemas conceituais de dados geoespaciais (ET-EDGV e OSM), observa-se que a ferramenta – ChatGPT – efetuou associações bem-sucedidas, coerentes com os conceitos aplicados e similares aos trabalhos realizados por humanos – Machado & Camboim (2024) e Silva (2022), apontando, inclusive, associações inéditas, com foco na similaridade léxica apresentada por Yu *et al.* (2018).

Cabe destacar que, para realização desta atividade, o detalhamento das informações no *prompt* de diálogo permite resultados melhores e mais específicos, embora nem sempre relacionados diretamente a um atributo ou domínio, corroborando com discussões de Dang *et al.* (2022) e Wang *et al.* (2023) sobre detalhamento do *prompt* e resultados mais refinados. Inclusive, observou-se também, que os resultados mais promissores surgiram com a entrada de dados em texto sequencial, e não para formatos como planilha, OWL, JSON ou XML, validando a observação de Mooney *et al.* (2023) sobre a dificuldade do ChatGPT em compreender outras estruturas de dados que não as textuais.

Por outro aspecto, há indicativo da influência da base de treinamento do LLM sobre os resultados, uma vez que houve associações com *tags* cuja tradução para o português não estava diretamente relacionada a termos apresentados na conceituação das classes, atributos e domínios, a exemplo de *railway+halt*, *military+barracks* ou *building+hut*. Neste contexto, a partir da análise pelo *taginfo* da quantidade de feições, percebeu-se que as *tags* indicadas pelo ChatGPT estão relacionadas àquelas com uma maior frequência de aplicação pelos usuários do OSM no Brasil, podendo ser indício de captura do conhecimento de seus usuários.

Finalmente, embora o modelo tenha demonstrado uma compreensão semântica promissora, seus resultados sugerem uma limitação significativa na interpretação das relações hierárquicas e de herança entre conceitos dentro do mesmo esquema conceitual. Diante disso, seria relevante explorar em estudos futuros formas alternativas ou mais explícitas de apresentação dos dados no *prompt*, buscando avaliar se esta limitação pode ser mitigada.

De maneira complementar, algumas outras questões importantes podem ser aprofundadas em trabalhos posteriores. Por exemplo, seria interessante investigar com mais detalhamento a inadequação da entrada de dados em formatos diversos do texto sequencial e se isso poderia ser melhorado com outros arranjos do diálogo no *prompt* de comunicação.

Saindo da utilização de uma ferramenta pronta, que foi o ChatGPT, pode-se realizar testes mais específicos com treinamento do LLM a partir de uma base de associação semântica conhecida, reduzindo a influência da base de conhecimento do modelo pré-treinado. Nesta mesma linha, seria enriquecedor para a discussão avaliar outros LLM para realização de testes, como o BERT ou o T5, além do GPT, talvez até efetuando análise comparativa dos resultados.

Enfim, outro aspecto a ser considerado em trabalhos futuros é incluir a geometria na análise de similaridade semântica, uma vez que, além da questão léxica, a representação geométrica é fator primordial para o futuro desenvolvimento de aplicações que façam integração automatizada de bases de dados a partir da compatibilidade semântica.

Ponderando, deste modo, a relevância da interoperabilidade semântica para a integração de bases de dados geoespaciais, bem como a necessidade de sua realização em processos mais eficazes, incluindo sua automatização, considera-se que este trabalho avançou um passo nas pesquisas anteriores, ao demonstrar a viabilidade de utilização de inteligência artificial no processo. Em especial o processamento de linguagem

natural, no contexto dos grandes modelos de linguagem, que pode agilizar a identificação de variadas associações possíveis entre modelos conceituais distintos. Apesar de algumas limitações apresentadas, esta é uma área de pesquisa que está em construção e tem avançado em diversas direções em anos recentes, sendo promissora sua aplicação dentro do cenário atual de uso das tecnologias para dados geoespaciais.

Referências

Anand, S., Morley, J., Jiang, W., Du, H., & Hart, G. (2010). *When worlds collide: Combining Ordnance Survey and Open Street Map data*. AGI Geocommunity '10, London, UK.

Ballatore, A., Bertolotto, M., & Wilson, D. C. (2013). *Geographic knowledge extraction and semantic similarity in OpenStreetMap*. Knowledge and Information Systems, *37*(1), 61–81. https://doi.org/10.1007/s10115-012-0571-0.

Borges, K. A. V.; Davis Jr., C. A. & Laender, A. H. F. (2005). Modelagem conceitual de dados geográficos. In: Casanova, M. A.; Câmara, G.; Davis Jr., C. A.; Vinhas, L. & Queiroz, G. R. de (ed). *Bancos de Dados Geográficos*. Curitiba, Editora MundoGEO. http://www.dpi.inpe.br/livros/bdados/

Bortolini, E., Silva, L. S. L., Machado, A. A., Paiva, C. D. A., & Camboim, S. P. (2018). *Potenciais categorias de informações geográficas do mapeamento colaborativo para o mapeamento oficial*. Colóquio Brasileiro de Ciências Geodésicas, X. Curitiba-PR.

Bortolini, E.; Silva, L. S. L.; Elias, E. N. N.; Camboim, S. P. & Schmidt, M. A. R. (2020). *Sinergias entre a produção dos dados geoespaciais de referência oficiais e colaborativos: uma proposição de eixos potenciais*. Simpósio Brasileiro de Infraestrutura de Dados Espaciais, II. Rio de Janeiro-RJ.

Brasil. (2010). *Plano de Ação para Implantação da Infraestrutura Nacional de Dados Espaciais – INDE*. 1° ed. Ministério do Planejamento, Orçamento e Gestão, Comissão Nacional de Cartografia. Brasília-DF.

Bravo, J. V. M. (2014). *A confiabilidade semântica das informações geográficas voluntárias como função da organização mental do conhecimento espacial*. Dissertação de Mestrado. 139 p. Universidade Federal do Paraná, Programa de Pós-Graduação em Ciências Geodésicas, Curitiba-PR.

Brovelli, M. A., Minghini, M., Molinari, M. E., & Zamboni, G. (2016). *Positional accuracy assessment of the openstreetmap buildings layer through automatic homologous pairs detection: the method and a case study*. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, *XLI-B2*, 615–620. https://doi.org/10.5194/isprsarchives-XLI-B2-615-2016.

Concar. (2017). Comissão Nacional de Cartografia. *Especificações Técnicas para Estruturação de Dados Geoespaciais Vetoriais (ET-EDGV 3.0)*. NCB-CC/E 0001B08. Versão 3.0.

Dang, H., Mecke, L., Lehmann, F., Goller, S., & Buschek, D. (2022). *How to Prompt? Opportunities and Challenges of Zero- and Few-Shot Learning for Human-AI Interaction in Creative Applications of Generative Models* (arXiv:2209.01390). arXiv. http://arxiv.org/abs/2209.01390.

Elias, E. N. N., & Fernandes, V. de O. (2019). *Quality Analysis of OpenStreetMap Geospatial Data for Positional Accuracy, Thematic Accuracy and Completeness indicators*. *pp.*, *30*(2).

Fernandes, V. O., Elias, E. N., & Zipf, A. (2020). *Integration of authoritative and volunteered geographic information for updating urban mapping: challenges and potentials.* The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, *XLIII-B4-2020*, 261–268. https://doi.org/10.5194/isprs-archives-XLIII-B4-2020-261-2020.

Grinberger, A. Y., Minghini, M., Juhász, L., Yeboah, G., & Mooney, P. (2022). *OSM Science - The Academic Study of the OpenStreetMap Project, Data, Contributors, Community, and Applications*. ISPRS International Journal of Geo-Information, 11(4), 230. https://doi.org/10.3390/ijgi11040230.

ISO. (2015). *ISO 19103:2015. Geographic information - Conceptual schema language.* International Organization for Standardization (ISO).

Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., & Wu, Y. (2016). *Exploring the Limits of Language Modeling* (arXiv:1602.02410). arXiv. http://arxiv.org/abs/1602.02410.

Kaur, J., Singh, J., Sehra, S. S., & Rai, H. S. (2017). *Systematic Literature Review of Data Quality Within OpenStreetMap*. International Conference on Next Generation Computing and Information Systems (ICNGCIS), 177–182. https://doi.org/10.1109/ICNGCIS.2017.35.

Kuhn, W. (2003). *Semantic reference systems*. International Journal of Geographical Information Science, 17(5), 405–409. https://doi.org/10.1080/1365881031000114116.

Longley, P. A.; Goodchild, M. F.; Maguire, D. J.; Rhind, D. W. (2011). *Geographic Information Systems and*

Science. 3rd ed. Hoboken: Wiley.

Machado, A. A., & Camboim, S. P. (2019). *Mapeamento colaborativo como fonte de dados para o planejamento urbano: Desafios e potencialidades.* urbe. Revista Brasileira de Gestão Urbana, 11, e20180142. https://doi.org/10.1590/2175-3369.011.e20180142.

Machado, A. A. (2020). *Compatibilização Semântica entre o Modelo de Dados do OpenStreetMap e a Especificação Técnica para Estruturação de Dados Geoespaciais Vetoriais (ET-EDGV).* Tese de doutorado em ciências geodésicas. Setor de Ciências da Terra, Universidade Federal do Paraná. Curitiba-PR.

Machado, A. A., & Camboim, S. P. (2024). *Semantic Alignment of Official and Collaborative Geospatial Data: A Case Study in Brazil*. Revista Brasileira de Cartografia, 76. Scopus. https://doi.org/10.14393/rbcv76n0a-72070.

McCarthy, J.; Minsky, M. L.; Rochester, N.; Shannon, C. E. (1955). *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*. AI Magazine. 27(4).

Ministério da Defesa. (2018). *Norma da Especificação Técnica para Aquisição de Dados Geoespaciais Vetoriais (EB80-N-72.005). ET-ADGV*. 1ª Edição. Exército Brasileiro; Departamento de Ciência e Tecnologia; Diretoria de Serviço Geográfico.

Mooney, P., Cui, W., Guan, B., & Juhász, L. (2023). *Towards Understanding the Geospatial Skills of ChatGPT: Taking a Geographic Information Systems (GIS) Exam*. https://doi.org/10.31223/X5P38P.

OGC. Open Geospatial Consortium. (2023). *Standards*. https://www.ogc.org/standards/.

OGC. Open Geospatial Consortium. (2020). *Benefits of Representing Spatial Data Using Semantic and Graph Technologies*. http://www.opengis.net/doc/wp/using-semantic-graph.

OSM. (2024). *Map Features*. https://wiki.OpenStreetMap.org/wiki/Map_features.

Petchenik, B. B. (1977). *Cognition In Cartography*. Cartographica: The International Journal for Geographic Information and Geovisualization, *14*(1), 117–128. https://doi.org/10.3138/97R4-84N4-4226-0P24.

Prince, S. J. D. (2023). *Understanding Deep Learning*. http://udlbook.com.

Robinson, A. C., Demšar, U., Moore, A. B., Buckley, A., Jiang, B., Field, K., Kraak, M.-J., Camboim, S. P., & Sluter, C. R. (2017). *Geospatial big data and cartography: Research challenges and opportunities for making maps that matter.* International Journal of Cartography, 3(sup1), 32–60. https://doi.org/10.1080/23729333.2016.1278151.

Rosch, E. (1975). *Cognitive Representations of Semantic Categories*. Journal of Experimental psychology: General. 104(3).

Santhanam, S., & Shaikh, S. (2019). *A Survey of Natural Language Generation Techniques with a Focus on Dialogue Systems - Past, Present and Future Directions* (arXiv:1906.00500). arXiv. http://arxiv.org/abs/1906.00500.

Segaran, T. (2007). *Programming Collective Intelligence: Building Smart Web 2.0 Applications*. 1ª edição. Ed O'Reilly.

Silva, A. E. de S.; Camboim, S. P.; Delazari, L. S. (2022). *O Problema da Compatibilidade Semântica entre as Representações Cartográficas do OpenStreetMap, ET-EDGV e OHI*. XII Colóquio Brasileiro de Ciências Geodésicas – CBCG e V Simpósio Brasileiro de Geomática – SBG.

Silva, L. S. da; Teixeira, R. R. P. (2022). *Filosofia da mente e inteligência artificial em atividades de divulgação científica mediadas por recursos audiovisuais*. Revista Mundi Sociais e Humanidades. Paranaguá, PR, 07(01), 01-27.

Silva, L. S. L. (2022). *Integração de Dados Provenientes de Mapeamento Colaborativo na Cartografia de Referência do Brasil.* Tese de doutorado em Ciências Geodésicas da Universidade Federal do Paraná.

Silva, L. S. L., & Camboim, S. P. (2020). *Brazilian NSDI ten years later: current overview, new challenges and propositions for national topographic mapping*. Boletim de Ciências Geodésicas, 26(4), e2020018. https://doi.org/10.1590/s1982-21702020000400018.

Sluter, C. R., Camboim, S. P., Iescheck, A. L., & Pereira, L. B. Castro, M. C.; Yamada, M. M.; Araújo, V. S. (2019). *A proposal of topographic map symbols for large-scale maps of urban areas in Brazil*. Abstracts of the ICA, 1, 362-377. https://doi.org/10.1080/00087041.2018.1549307.

Souza, F. A., & Camboim, S. P. (2023). *Semantic Alignment of Geospatial Data Models using ChatGPT: preliminary studies.* Fonseca F. F. da & Vinhas L. (Orgs.), Proc. Brazilian Symp. GeoInformatics (p. 399–404). National Institute for Space Research, INPE; Scopus. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85181118913&partnerID=40&md5=45de9b24f4242bc1e4306f46b84a1ed0.

Vasconcellos, S. J. L., & Machado, S. D. S. (2006). *Construtivismo, psicologia experimental e neurociência.* Psicologia Clínica, 18(1), 83–94. https://doi.org/10.1590/S0103-56652006000100007.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention is All you Need*. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

Wang, J., Liu, Z., Zhao, L., Wu, Z., Ma, C., Yu, S., Dai, H., Yang, Q., Liu, Y., Zhang, S., Shi, E., Pan, Y., Zhang, T., Zhu, D., Li, X., Jiang, X., Ge, B., Yuan, Y., Shen, D., … Zhang, S. (2023). *Review of Large Vision Models and Visual Prompt Engineering* (arXiv:2307.00855). arXiv. http://arxiv.org/abs/2307.00855.

Yu, L., Qiu, P., Liu, X., Lu, F., & Wan, B. (2018). *A holistic approach to aligning geospatial data with multidimensional similarity measuring.* International Journal of Digital Earth, 11(8), 845–862. https://doi.org/10.1080/17538947.2017.1359688.

Zhang, Q., Zhang, T., Zhai, J., Fang, C., Yu, B., Sun, W., & Chen, Z. (2024). *A Critical Review of Large Language Model on Software Engineering: An Example from ChatGPT and Automated Program Repair* (arXiv:2310.08879). arXiv. http://arxiv.org/abs/2310.08879.

Contribuição das Autoras

Conceitualização, F. A. S. e S. P. C.; Análise Formal, F. A. S.; Investigação, F. A. S. e E. D. B. da S; Metodologia, F. A. S. e S. P. C.; Supervisão, S. P. C.; Validação, S. P. C.; Visualização, F. A. S.; Redação – Minuta Inicial, F. A. S.; Redação - Revisão e Edição, F. A. S. e S. P. C. Todas as autoras leram e concordaram com a versão publicada do manuscrito.

Conflitos de Interesse

As autoras informam que não há algum conflito de interesse.

Biografia da autora principal

Fabíola Andrade Souza, nasceu em Jaguaquara-Bahia, Brasil, em 17 de junho de 1978. Bacharel em informática pela Universidade Católica do Salvador (UCSAL) e mestre em Engenharia Ambiental Urbana pela Universidade Federal da Bahia (UFBA), doutoranda no programa de Pós-graduação em Ciências Geodésicas pela Universidade Federal do Paraná (UFPR). Atua como docente na Escola Politécnica da UFBA. Experiência em geotecnologias com ênfase em sistemas de informação geográfica, bancos de dados geográficos e infraestrutura de dados espaciais.