

Processo de curadoria semântica assistida por LLMs para integração de dados abertos: o caso da plataforma ELLAS na América Latina

Marlos Vinicius S. Piccinelli ¹, Nádia P. Kozievitch ¹, Rita Cristina G. Berardi ¹

¹ Universidade Tecnológica Federal do Paraná (UTFPR)
Av. Sete de Setembro, 3165 - Rebouças - CEP 80230-901 - Curitiba – Brasil
marlosvinicius.info@gmail.com, {nadiap, ritaberardi}@utfpr.edu.br

Abstract. *This paper proposes a semantic data curation pipeline supported by Large Language Models (LLMs) to standardize heterogeneous categorical variables in open data from multilingual surveys. The approach combines prompt engineering, human validation, and alignment with institutional references such as UNESCO, IBGE, and INEP. Applied to an ELLAS dataset with 10,312 responses on gender in STEM in Latin America, the pipeline reduced categorical cardinality, improved semantic consistency, and enabled the integration of curated data into knowledge graphs, contributing to open data interoperability.*

Resumo. *Este artigo propõe um pipeline de curadoria semântica assistida por Modelos de Linguagem de Grande Escala (LLMs) para padronizar variáveis categóricas heterogêneas em dados abertos provenientes de surveys multilíngues. A abordagem combina engenharia de prompts, validação humana e alinhamento a referenciais institucionais, como UNESCO, IBGE e INEP. Aplicado a um dataset do projeto ELLAS com 10.312 respostas sobre gênero em STEM na América Latina, o pipeline reduziu a cardinalidade, melhorou a consistência semântica e viabilizou a integração dos dados a grafos de conhecimento, contribuindo para a interoperabilidade de dados abertos.*

1. Introdução

A crescente disponibilidade de dados em iniciativas de dados abertos tem sido reconhecida como um fator estratégico para a formulação de políticas públicas, promoção da transparência e fortalecimento da governança digital. No entanto, sua disponibilização não garante utilização efetiva nem integração com outras bases, especialmente quando há heterogeneidade estrutural e ausência de padronização [Andrade and Luque 2022][Alencar et al. 2022].

Esse desafio é ampliado em datasets derivados de surveys multilíngues, nos quais respostas abertas introduzem variações linguísticas, inconsistências e ambiguidades que dificultam análises e limitam a integração dos dados [Borchardt et al. 2022], resultando na subutilização de informações potencialmente relevantes.

Embora abordagens de curadoria e transformação busquem mitigar esses problemas, ainda dependem de trabalho manual especializado, o que compromete sua escalabilidade e reprodutibilidade. Nesse contexto, Modelos de Linguagem de Grande

Escala (*Large Language Models, LLMs*) surgem como alternativa promissora para apoiar a padronização de dados textuais, desde que utilizados de forma estruturada e controlada [Marcacini et al. 2025][Parreiras et al. 2022].

Diante disso, este trabalho propõe um pipeline de curadoria semântica assistida por LLMs, combinando engenharia de *prompts*, validação humana e alinhamento a referenciais institucionais, como a Organização das Nações Unidas para a Educação, a Ciência e a Cultura (UNESCO), o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) e o Instituto Brasileiro de Geografia e Estatística (IBGE). A abordagem foi aplicada a um dataset real proveniente de uma survey internacional sobre gênero em Ciência, Tecnologia, Engenharia e Matemática (*Science, Technology, Engineering and Mathematics, STEM*) na América Latina.

O dataset integra o contexto do projeto ELLAS (*Equality in Leadership for Latin America STEM*), iniciativa que visa gerar e utilizar dados abertos conectados para subsidiar a criação e/ou melhoria de políticas públicas voltadas à redução das desigualdades de gênero em STEM, com ênfase em posições de liderança [Maciel et al. 2023][Berardi et al. 2023]¹. Os resultados do uso do pipeline indicam reduções expressivas de cardinalidade em variáveis categóricas, superiores a 90% em alguns casos, além de melhorias na consistência semântica e na capacidade de integração dos dados. A abordagem também demonstra potencial de escalabilidade ao reduzir a dependência de tarefas operacionais e reposicionar o especialista para atividades de validação e tomada de decisão.

2. Fundamentação Teórica

Dados abertos correspondem a dados acessíveis ao público, em meio digital, estruturados em formato aberto, processáveis por máquina e disponibilizados sob licença que permita sua livre utilização, consumo ou cruzamento, com exigência máxima de atribuição da fonte [Brasil 2026]. No contexto governamental, essa abertura visa ampliar transparência, reutilização e geração de aplicações e serviços, além de apoiar a tomada de decisão no setor público [Brasil 2024][SBC 2024].

A qualidade de dados refere-se ao grau em que um dataset atende a critérios como exatidão, completude, validade, consistência, unicidade e adequação ao uso [IBM 2026a]. Em variáveis categóricas, cujos valores se organizam em categorias distintas [Frost 2026], problemas tornam-se mais evidentes com o aumento da cardinalidade, isto é, do número de valores distintos em uma variável [Microsoft 2024]. Nesses casos, aumentam ocorrências de duplicidade, variações de escrita, inconsistências e ambiguidades, comprometendo a análise e a padronização dos dados [IBM 2026b, IBM 2026c].

Nesse contexto, a interoperabilidade não depende apenas da compatibilidade técnica entre sistemas, mas também da preservação do significado dos dados. No nível semântico, refere-se à capacidade de diferentes sistemas interpretarem dados de forma consistente, garantindo que seu significado seja mantido durante a troca [European

¹ Projeto ELLAS, disponível em: <https://ellas.ufmt.br>

Commission 2026]. Essa dimensão é essencial em cenários de integração de dados abertos, nos quais múltiplas fontes precisam ser combinadas de forma coerente.

Diante desses desafios, a curadoria semântica pode ser entendida como o conjunto de práticas voltadas à organização, padronização e enriquecimento dos dados, com o objetivo de preservar seu significado e favorecer sua integração entre diferentes fontes [Pinto et al. 2022]. Em cenários com dados heterogêneos, multilíngues e de alta cardinalidade, essas práticas tornam-se fundamentais para garantir qualidade, consistência e utilidade dos dados. Nesse sentido, o uso assistido de LLMs tem sido explorado como apoio à interpretação e transformação de dados textuais, desde que associado a mecanismos de controle, validação humana e rastreabilidade do processo [Marcacini et al. 2025].

3. Pipeline de Curadoria e Desenvolvimento

Esta seção apresenta a caracterização do *dataset*, a estrutura do *pipeline* da curadoria, o processo de curadoria aplicado e o processo de reprodutibilidade.

3.1. Caracterização do *Dataset* e Desafios Semânticos

O *dataset* utilizado neste estudo foi produzido a partir de uma *survey* conduzida pelo projeto ELLAS, composta por 10.312 respostas coletadas no Brasil, Peru e Bolívia entre 2023 e 2024, em português e espanhol, com participação de indivíduos vinculados às áreas de STEM. A *survey* contempla variáveis demográficas e educacionais, como gênero, escolaridade, trajetória de carreira e percepções sobre igualdade de gênero.

Os dados são destinados à integração com a plataforma de dados abertos do ELLAS, o que exige compatibilidade com estruturas previamente definidas e alinhamento a referenciais institucionais. Todas as etapas deste trabalho foram realizadas exclusivamente sobre dados anonimizados.

Valores reais observados no *dataset*, apresentados na Tabela 1, evidenciam a diversidade de representações nas variáveis categóricas, incluindo variações linguísticas, inconsistências semânticas e respostas fora de escopo. A heterogeneidade inviabiliza a integração direta dos dados com estruturas padronizadas, reforçando a necessidade de curadoria semântica.

Gênero	Etnia	Profissão / Área
Cis	Blanco	Administração
Ela	Otro	Admin
Ele	Pardo	Nd
Femininos	Quechua	Hidrocarburos
Feminismo	Indígena	INDEPENDIENTE
Feminino	Preto	INVESTIGACION
Chico trans	Oriental	Independiente
Homem Cis	Zambo(a)	Importaciones
Feminino - Hétero Cisgênero	Oriental / asiático	Ninguno , hago solamente prácticas en el en sector de telecomunicaciones
Gênero é pra palavras. Sou do sexo feminino	Negra / Negro / Afroperuano	Trabajo en sistemas pero recién estoy estudiando la carrera de ing de sistemas

Tabela 1 – Exemplos reais de valores extraídos do *dataset*.

3.2. Pipeline de Curadoria Semântica

A Figura 1 apresenta a visão geral do *pipeline* de curadoria, evidenciando o fluxo de transformação dos dados e os pontos de intervenção do especialista, cujo processo completo está documentado em repositório público².



Figura 1 – Pipeline de curadoria semântica assistida por LLMs.

De forma geral, a Figura 1 é operacionalizada pelas seguintes etapas:

- **Preparação dos dados:** leitura dos arquivos originais, tratamento de *encoding*, padronização estrutural e extração de valores únicos das variáveis categóricas, reduzindo o volume de processamento.
- **Processamento assistido por LLMs:** aplicação de LLMs por meio de *prompts* estruturados para tarefas como correção ortográfica, padronização linguística, tradução, identificação de equivalências semânticas e consolidação dos resultados em dicionários de mapeamento.
- **Estruturação semântica:** organização dos dados em categorias alinhadas a referenciais institucionais definidos pelo especialista de domínio, atuando como mecanismo de ancoragem semântica.
- **Geração da saída estruturada:** produção do *dataset* final normalizado, acompanhado dos artefatos intermediários que documentam as transformações realizadas.

Ao longo do *pipeline*, o especialista de domínio atua na validação dos resultados intermediários e nas decisões semânticas críticas, enquanto tarefas operacionais são assistidas por LLMs.

3.3. Processo de curadoria aplicado

² O repositório completo do projeto, incluindo notebooks, versões intermediárias, prompts utilizados e documentação detalhada do processo metodológico, está disponível para consulta em: <https://github.com/mpiccinelli/ellas>

Inicialmente, o *dataset* foi processado em *Python* com foco na variável de gênero, a partir da qual foram extraídos os valores únicos da coluna correspondente. A manutenção da coluna de país permitiu inferir o idioma de origem das respostas, sendo essencial para o tratamento multilíngue ao longo do processo. A escolha da variável de gênero como base para exemplificação se deve à sua centralidade no contexto do projeto ELLAS. A partir desse conjunto reduzido de valores, foi aplicada uma sequência de transformações semânticas conduzidas por LLMs, estruturadas em etapas independentes, cada uma com objetivo específico, garantindo maior controle sobre os resultados. A Tabela 2 ilustra esse processo, apresentando um recorte representativo das transformações aplicadas aos valores originais até sua categorização final.

País	Texto Original	Correção textual	Tradução ES	Sinônimos	Contextualização	SOGIE
Bolívia	Femenino	Femenino	Femenino	Femenino	Femenino	Cisgênero
Brasil	Mulher	Mulher	Mujer	Femenino	Femenino	Cisgênero
Brasil	cis	Cisgênero	Cisgênero	Cisgênero	Cisgênero	Cisgênero
Brasil	Cisgênero	Cisgênero	Cisgênero	Cisgênero	Cisgênero	Cisgênero
Brasil	Ela	Feminino	Femenino	Femenino	Femenino	Cisgênero
Brasil	Femea	Feminino	Femenino	Femenino	Femenino	Cisgênero
Bolívia	No binario	No binario	No binario	No binario	No binario	No binario
Brasil	Agênero	Agênero	Agênero	Agênero	Agênero	No binario
Brasil	não binário	Não-binário	No binario	No binario	No binario	No binario
Brasil	Binário	Não-binário	No binario	No binario	No binario	No binario
Brasil	Não-binário	Não-binário	No binario	No binario	No binario	No binario
Brasil	Não Binare	Não-binário	No binario	No binario	No binario	No binario
Peru	Neutro	Neutro	Neutro	Neutro	Neutro	No binario
Bolívia	No sabe	No sabe	No sabe	No sabe	No sabe	No informado
Brasil	Bicha louca	Bicha louca	Bicha loca	Bicha loca	Bicha loca	Otra expresión
Peru	No tengo preferencia el que sea idonio	No tengo preferencia, que sea idóneo	No tengo preferencia, que sea idóneo	No tengo preferencia, que sea idóneo	Fora de contexto	Otra expresión
Brasil	Só existe homem ou mulher	Só existe homem ou mulher	Solo existe hombre o mujer	Solo existe hombre o mujer	Fora de contexto	Otra expresión
Peru	Prefiero no responder	Prefiero no responder	Prefiero no responder	Prefiero no responder	Prefiero no responder	Prefiero no responder
Bolívia	Transgênero	Transgênero	Transgênero	Transgênero	Transgênero	Transgênero
Peru	chico trans	Chico trans	Chico trans	Chico trans	Chico trans	Transgênero

Tabela 2 – Exemplo do processo de curadoria semântica aplicado à variável gênero.

O processo foi estruturado nas seguintes etapas:

- **Correção textual:** normalização ortográfica e sintática dos valores originais, preservando o idioma de origem. O LLM foi instruído a corrigir erros de digitação e inconsistências linguísticas sem alterar o significado, utilizando a informação de país para inferência do idioma.
- **Padronização linguística:** tradução dos valores para um único idioma de referência, definido como o espanhol, com o objetivo de reduzir ambiguidades decorrentes do caráter multilíngue do *dataset* e garantir consistência nas etapas subsequentes.
- **Identificação de sinônimos:** identificação de equivalências semânticas entre diferentes expressões de um mesmo conceito, como “femenino” e “mujer”, permitindo a redução da cardinalidade das variáveis categóricas.
- **Contextualização semântica:** Avaliação da aderência dos termos ao domínio da variável, permitindo manter expressões relevantes, mesmo quando não padronizadas, e identificar respostas fora de escopo.
- **Classificação baseada em referenciais institucionais:** classificação dos valores com base no modelo SOGIE (*Sexual Orientation, Gender Identity, and Expression*), conforme proposto pela UNESCO [UNESCO 2015; UNESCO 2016; UNESCO 2018; UNESCO 2019], possibilitando o mapeamento para categorias padronizadas alinhadas a referenciais reconhecidos.

A escolha dos referenciais institucionais não foi arbitrária, mas orientada pela necessidade de integração semântica com estruturas existentes. As classificações adotadas, como SOGIE, IBGE e INEP, são compatíveis com a ontologia da plataforma ELLAS, permitindo a incorporação direta dos dados a um grafo de conhecimento. Nesse contexto, a qualidade semântica não se limita à consistência interna, mas à capacidade de representar conceitos compatíveis com estruturas externas.

A separação dessas etapas permitiu distinguir tarefas de natureza linguística, semântica e classificatória, reduzindo ambiguidades e aumentando a previsibilidade do comportamento dos modelos. Além disso, a execução sequencial possibilitou a validação intermediária dos resultados, contribuindo para a qualidade e consistência das transformações realizadas.

3.4. Reprodutibilidade

Nos testes iniciais, observou-se que a reaplicação dos *prompts* em novas sessões dos LLMs não produzia resultados consistentes, uma vez que instruções dependentes de contexto previamente construído não eram corretamente interpretadas quando executadas de forma isolada, resultando em saídas incompletas ou semanticamente inconsistentes. Diante dessas limitações, foi conduzido um processo iterativo de refinamento de *prompts*, no qual cada instrução foi reestruturada para operar de forma independente, com definição explícita de entradas, saídas e restrições, reduzindo ambiguidades e a dependência de contexto implícito.

Como resultado, os *prompts* passaram a incorporar explicitamente:

- **Definição da tarefa:** indicação direta do objetivo da operação. Exemplo: *“Corrija erros ortográficos mantendo o idioma original do termo”*.
- **Especificação das colunas de entrada e saída:** indicação explícita dos dados de entrada e da estrutura de saída. Exemplo: *“Utilize a coluna ‘Texto original’ como entrada e gere uma coluna ‘Correção textual’ usando a coluna ‘País’ para inferir o idioma”*.
- **Regras de processamento linguístico:** inclusão de instruções para controle semântico e linguístico. Exemplo: *“Não traduza o termo nesta etapa; apenas corrija sua ortografia”*.
- **Restrições de formato da saída:** definição de estrutura fixa para garantir consistência entre execuções. Exemplo: *“Retorne apenas a tabela resultante, sem explicações adicionais”*.

A validação da reprodutibilidade foi realizada em dois níveis. Primeiramente, os *prompts* refinados foram reaplicados em múltiplas sessões independentes do mesmo modelo, verificando a estabilidade das respostas na ausência de contexto compartilhado. Em seguida, o mesmo conjunto de instruções foi testado em diferentes LLMs, incluindo OpenAI GPT-4 e Google Gemini 1.5, para avaliar a generalização do comportamento do *pipeline*.

Ressalta-se que o processo não dependeu de uso de APIs, treinamento de modelos, ajuste fino (*fine-tuning*) ou consumo controlado por tokens. Os LLMs foram

utilizados por meio de interfaces conversacionais e agentes de IA disponíveis em assinaturas padrão, tendo os *prompts*, arquivos intermediários e *scripts* sido documentados no repositório do projeto. Essa decisão metodológica foi adotada para aproximar o processo de curadoria de um cenário viável para equipes não especializadas em ciência de dados, mas que necessitam tratar, padronizar e integrar dados heterogêneos de forma controlada.

Após o refinamento, os *prompts* passaram a produzir saídas consistentes entre execuções, independentemente da sessão ou do modelo, indicando que a reprodutibilidade do método está associada à explicitação das instruções.

Adicionalmente, a reaplicação da metodologia em outras variáveis do *dataset*, como etnia e escolaridade, demonstrou que o mesmo conjunto de *prompts* pôde ser reutilizado sem alterações estruturais, sendo necessárias apenas adaptações aos referenciais institucionais de cada domínio.

Durante a validação, cada etapa da metodologia foi verificada por meio da inspeção dos valores únicos gerados (não sendo calculadas métricas quantitativas de divergência), permitindo identificar divergências entre execuções. Após o refinamento, não foram observadas inconsistências nas saídas analisadas, indicando estabilidade do processo.

4. Resultados e Impactos

A aplicação da abordagem proposta resultou na transformação de um *dataset* originalmente heterogêneo em uma estrutura padronizada e semanticamente alinhada a referenciais institucionais, viabilizando sua integração com a plataforma ELLAS por meio de variáveis como gênero, etnia e formação educacional, estruturadas conforme classificação da UNESCO, IBGE e INEP, respectivamente. Esse alinhamento estabelece pontos de ancoragem semântica que permitem a conexão dos dados à ontologia da plataforma e sua incorporação em um grafo de conhecimento. Além disso, o uso desses referenciais amplia o potencial de integração com outras bases que adotam os mesmos padrões classificatórios.

A metodologia apresentou cobertura operacional completa, uma vez que todos os 10.312 registros analisados receberam uma categoria final compatível com os referenciais adotados. Essa cobertura não deve ser interpretada como uma métrica de acurácia estatística, mas como indicação de que o processo de curadoria foi capaz de produzir uma saída estruturada para todos os registros considerados, sem manter valores pendentes de categorização ao final da etapa de validação.

A validação foi realizada por meio da análise dos valores únicos de cada variável categórica, permitindo identificar inconsistências sem a necessidade de verificação individual dos registros, o que contribuiu para a eficiência do processo.

Os principais resultados estão relacionados à redução da cardinalidade das variáveis categóricas, calculada pela diferença entre o número de valores distintos antes

e depois da curadoria, em relação ao total original de valores distintos. Na variável de gênero, a cardinalidade foi reduzida de 97 para 10 categorias (89,7%). Na variável de etnia, houve redução de 27 para 8 categorias (70,4%), com base na classificação do IBGE [IBGE 2022]. A variável de formação educacional foi reduzida de 870 variações para 13 categorias (98,5%), alinhadas às áreas do INEP [INEP 2017; INEP 2023].

Além da redução da cardinalidade, observou-se melhoria na consistência estrutural dos dados, com eliminação de redundâncias, correção de inconsistências e maior comparabilidade entre diferentes contextos e idiomas. Como resultado, o *dataset* tornou-se mais adequado para consultas, cruzamentos e análises estatísticas.

Esses resultados permitem discutir não apenas ganhos quantitativos, mas também implicações estruturais da abordagem proposta para a curadoria e integração de dados abertos.

5. Discussão

A principal contribuição deste trabalho está na proposição de um *pipeline* reprodutível de curadoria semântica assistida por LLMs, validado em um cenário real. Sua decomposição em etapas especializadas mostrou-se fundamental para garantir controle, reduzir inconsistências e viabilizar a validação progressiva dos resultados, evidenciando limitações de abordagens baseadas em *prompts* únicos.

A utilização de estruturas intermediárias, como dicionários de mapeamento, contribuiu para a rastreabilidade e transparência do processo, permitindo registrar explicitamente as transformações realizadas e atendendo a requisitos de auditabilidade relevantes em contextos de dados abertos.

Os resultados obtidos indicam que a padronização semântica associada ao alinhamento com referenciais institucionais não apenas melhora a organização dos dados, mas viabiliza sua integração efetiva com outras bases por meio de categorias compartilhadas. No contexto da plataforma ELLAS, esse alinhamento permite que os dados atuem como elementos conectáveis em um grafo de conhecimento, ampliando seu potencial de uso em análises e processos de tomada de decisão.

A abordagem adotada também evidencia a importância de um modelo híbrido, no qual a automação por LLMs é complementada pela validação humana. Esse equilíbrio permite explorar a capacidade dos modelos em lidar com alta cardinalidade e variações linguísticas, ao mesmo tempo em que mantém controle sobre decisões semânticas críticas [Marcacini et al. 2025]. Nesse contexto, observa-se um reposicionamento do especialista, que passa a atuar predominantemente em validação e tomada de decisão, conforme discutido por [Parreiras et al. 2022].

Apesar dos avanços, algumas limitações devem ser consideradas e indicam oportunidades de evolução da abordagem. A curadoria proposta depende do refinamento de prompts e da supervisão humana, sendo sensível a interpretações inadequadas dos modelos quando as instruções não são suficientemente explícitas. Além disso, a

validação realizada neste estudo teve natureza semântica e operacional, baseada na inspeção dos valores únicos, na rastreabilidade dos dicionários de mapeamento e na supervisão do especialista, não tendo como objetivo mensurar a acurácia estatística das classificações. Em estudos futuros, essa etapa poderá ser complementada por avaliações cegas com especialistas independentes ou por amostragens anotadas manualmente, quando o objetivo for medir concordância ou desempenho classificatório.

Por fim, embora os resultados indiquem potencial de aplicação em diferentes variáveis categóricas, a validação foi conduzida em domínios específicos do projeto ELLAS. Estudos futuros podem ampliar a abordagem para outros tipos de dados públicos, testar diferentes referenciais institucionais, comparar estratégias de validação humana e investigar formas de transformar o pipeline em uma ferramenta orientada a usuários de domínio, reduzindo barreiras técnicas para a curadoria semântica de dados abertos.

6. Conclusão

Este trabalho demonstrou que a curadoria semântica de dados categóricos complexos pode ser conduzida com apoio de LLMs quando organizada em etapas controladas, validação humana e alinhamento a referenciais institucionais. Aplicada a um *dataset* real do projeto ELLAS, a abordagem padronizou respostas multilíngues, reduziu a cardinalidade de variáveis categóricas e viabilizou a integração dos dados com outras bases e grafos de conhecimento.

Os resultados indicam que a qualidade semântica não decorre apenas da automação, mas da combinação entre referenciais compatíveis, *prompts* explícitos e atuação do especialista. Nesse processo, o especialista passa a se concentrar em validação e tomada de decisão, ampliando a escala da curadoria sem comprometer controle, rastreabilidade e responsabilidade analítica [Marcacini et al. 2025][Parreiras et al. 2022]. A reprodutibilidade, por sua vez, mostrou-se menos dependente de um modelo específico do que da clareza das instruções e das etapas do *pipeline*, em diálogo com usos recentes de LLMs em tarefas cognitivas estruturadas [Felizardo and Deizepe 2026].

Embora o *pipeline* seja eficaz para tratar heterogeneidade semântica, inconsistências linguísticas e respostas fora de escopo, parte do esforço de curadoria decorre do desenho da coleta. Isso indica que avanços em dados abertos dependem tanto de métodos posteriores de tratamento quanto de melhor estruturação e padronização na origem.

No contexto de governo digital e formulação de políticas públicas, a abordagem reforça que dados abertos só se tornam efetivamente utilizáveis quando podem ser integrados, comparados e interpretados de forma consistente. Como perspectivas para trabalhos futuros, destacam-se a automatização controlada de etapas do *pipeline*, o desenvolvimento de ferramentas dedicadas à curadoria semântica assistida por LLMs e a ampliação da abordagem para outros domínios de dados públicos, com mecanismos acessíveis a equipes não especializadas em ciência de dados. Complementarmente,

pretende-se realizar uma validação quantitativa mais robusta do *pipeline*, por meio de anotação humana cega em amostra representativa, utilizando métricas de acurácia como precisão, *recall*, F1-score e Kappa. Adicionalmente, planeja-se analisar as ameaças à validade, avaliar o custo computacional e o consumo de *tokens* dos LLMs, bem como investigar estratégias para mitigar alucinações dos modelos.

Agradecimentos

Os autores agradecem à Universidade Tecnológica Federal do Paraná (UTFPR), ao Programa de Pós-Graduação em Computação Aplicada (PPGCA) e ao projeto ELLAS (*Equality in Leadership for Latin America STEM*) pelo apoio institucional e acadêmico ao desenvolvimento deste trabalho.

Uso de Inteligência Artificial

Este trabalho utilizou Modelos de Linguagem de Grande Escala (*Large Language Models*, LLMs) como parte integrante da metodologia de curadoria semântica, aplicados em tarefas como correção textual, padronização linguística e interpretação semântica por meio de *prompts* estruturados, com validação em etapas críticas e incorporação dos resultados ao processo de transformação dos dados. Adicionalmente, ferramentas de inteligência artificial foram utilizadas como apoio à revisão textual do manuscrito. Todas as decisões metodológicas, interpretações, conclusões e o conteúdo intelectual apresentado são de responsabilidade exclusiva dos autores.

Referências

- Alencar, J., Magalhães, R., Vasconcelos, D., Chaves, S., Oliveira, J., Bessa, A., & Rodrigues, E. (2022). Comparação de Desempenho entre Soluções de Interoperabilidade. In Anais do X Workshop de Computação Aplicada em Governo Eletrônico, (pp. 25-36). Porto Alegre: SBC.
- Andrade, S. A. and Luque, D. B. (2022). “Interoperabilidade de Sistemas aplicados às Cidades Inteligentes: Um Estudo de Mapeamento Sistemático”. In: Anais do Workshop de Computação Aplicada ao Governo Eletrônico.
- Berardi, R. et al. (2023). “ELLAS: Uma plataforma de dados abertos com foco em lideranças femininas em STEM no contexto da América Latina”. In: Anais do XVII Women in Information Technology, pp. 124-135.
- Borchardt, G., Santos, M., Frigo, L., & Gasparini, I. (2022). Ferramenta de Visualização de Dados do Censo da Educação Superior do INEP. In Anais do X Workshop de Computação Aplicada em Governo Eletrônico, (pp. 227-234). Porto Alegre: SBC.
- Brasil (2026). “O que são Dados Abertos?”. Acesso à Informação. Disponível em: https://www.gov.br/acessoainformacao/pt-br/lai-para-sic/transparencia-ativa/copy_of_guia-de-transparencia-ativa/textos-complementares/formato-aberto. Acesso em: 27 mar. 2026.

- Brasil (2024). “Política de Dados Abertos do Governo Federal”. Governo Digital. Disponível em: <https://www.gov.br/governodigital/pt-br/dados-abertos> Acesso em: 26 mar. 2026.
- Felizardo, K. R. and Deizepe, A. R. (2026). “Uso de LLM para apoiar a condução de revisões sistemáticas da literatura em Engenharia de Software”. Fronteiras de Engenharia de Software, Escola de Estudos Experimentais em Engenharia de Software (EEES 2026), ICSE 2026. Disponível em: <https://www.youtube.com/watch?v=ZwPrm7oZ9VU>. Acesso em: 27 mar. 2026.
- Frost, J. (2026). “Categorical Variables”. Statistics By Jim. Disponível em: <https://statisticsbyjim.com/glossary/categorical-variables/>. Acesso em: 27 mar. 2026.
- IBGE (2022). “Censo Demográfico 2022: Identificação étnico-racial da população, por sexo e idade”. Instituto Brasileiro de Geografia e Estatística. Disponível em: https://agenciadenoticias.ibge.gov.br/media/com_mediaibge/arquivos/13ee0337cffc1de37bf0cd4da3988e1f.pdf. Acesso em: 26 mar. 2026.
- IBM (2026a). “What is Data Quality?”. IBM. Disponível em: <https://www.ibm.com/think/topics/data-quality>. Acesso em: 27 mar. 2026.
- IBM (2026b). “What is Data Quality Management?”. IBM. Disponível em: <https://www.ibm.com/think/topics/data-quality-management>. Acesso em: 27 mar. 2026.
- IBM (2026c). “Ambiguous Data”. IBM. Disponível em: <https://www.ibm.com/docs/en/cognos-analytics/11.2.x?topic=maps-ambiguous-data>. Acesso em: 27 mar. 2026.
- INEP (2017). “Áreas de Formação e Treinamento da CINE-F 2013: Manual que acompanha a Classificação Internacional Normalizada da Educação 2011”. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Disponível em: https://download.inep.gov.br/publicacoes/institucionais/estatisticas_e_indicadores/areas_de_formacao_e_treinamento_da_cine-f_2013.pdf. Acesso em: 26 mar. 2026.
- INEP (2023). “Censo da Educação Superior: Microdados do Censo da Educação Superior 2023”. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Disponível em: <https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/censo-da-educacao-superior>. Acesso em: 26 mar. 2026.
- European Commission (2026). “Semantic Interoperability”. Interoperable Europe Portal. Disponível em: <https://interoperable-europe.ec.europa.eu/collection/iopeu-monitoring/glossary/term/semantic-interoperability>. Acesso em: 27 mar. 2026.
- Maciel, C. et al. (2023). “Open Data Platform to Promote Gender Equality Policies in STEM”. In: Proceedings of the Western Decision Sciences Institute.
- Marcacini, R. M. et al. (2025). “LLM4Gov: A Privacy-Preserving Approach to Teacher-Student Fine-Tuning of Distilled LLMs for the Public Sector”. In: Anais do Latin American Symposium on Digital Government.

- Microsoft (2024). “COLUMN STATISTICS function (DAX)”. Microsoft Learn. Disponível em: <https://learn.microsoft.com/en-us/dax/columnstatistics-function-dax>. Acesso em: 27 mar. 2026.
- Parreiras, M. et al. (2022). “Inteligência artificial aplicada para o aumento da produtividade no atendimento de intimações”. In: Anais do Workshop de Computação Aplicada ao Governo Eletrônico.
- Pinto, G. et al. (2022). “A Graph Knowledge-base for Auditing Human Resources Public Management”. In: Anais do Workshop de Computação Aplicada ao Governo Eletrônico.
- SBC (2024). “Governo Digital e Transformação Digital no Setor Público”. Disponível em: <https://www.sbc.org.br/governo-digital/>. Acesso em: 26 mar. 2026.
- UNESCO (2015). “International Standard Classification of Education: Fields of Education and Training 2013 (ISCED-F 2013)”. UNESCO Institute for Statistics. Disponível em: <http://dx.doi.org/10.15220/978-92-9189-179-5-en>. Acesso em: 26 mar. 2026.
- UNESCO (2016). “Safe, Seen and Included: Inclusion and Diversity within Educational Settings”. UNESCO. Disponível em: <https://unesdoc.unesco.org/ark:/48223/pf0000385417>. Acesso em: 26 mar. 2026.
- UNESCO (2018). “School-related Violence and Bullying on the Basis of Sexual Orientation and Gender Identity or Expression (SOGIE): A Global Review”. UNESCO. Disponível em: <https://unesdoc.unesco.org/ark:/48223/pf0000366434>. Acesso em: 26 mar. 2026.
- UNESCO (2019). “Bringing it Out in the Open: Monitoring School Violence Based on Sexual Orientation, Gender Identity or Gender Expression”. UNESCO. Disponível em: <https://unesdoc.unesco.org/ark:/48223/pf0000367493>. Acesso em: 26 mar. 2026.