Abordagem LGRAG-IoT: Contribuição para a Conformidade do Middleware EXEHDA com a Lei Geral de Proteção de Dados

Rogério Albandes[1,2], Rafael Azevedo[1], Renata Abib[1], Anderson Passos[1], Ana Marilza Pernas[2], Adenauer Yamin[2]

[1]Universidade Católica de Pelotas – Pelotas – RS – Brasil

[2]Universidade Federal de Pelotas – Pelotas – RS – Brasil

Abstract. *This work presents the LGRAG-IoT approach, a RAG-based architecture to support automated verification of compliance with the Brazilian General Data Protection Law (LGPD) in Internet of Things environments. The proposal integrates monitoring of data flows from the EXEHDA middleware with semantic retrieval of normative documents for analysis by language models. The architecture identifies privacy violations and generates justifications grounded in legal evidence. The evaluation, using simulated IoT data processing scenarios, compared a standalone LLM with the RAG-based approach. Results indicate improvements in classification performance and decision explainability, showing potential to support audits and data governance.*

Resumo. *Este trabalho apresenta a abordagem LGRAG-IoT, uma arquitetura baseada em RAG para apoiar a verificação automatizada de conformidade com a LGPD em ambientes de Internet das Coisas. A proposta integra o monitoramento de fluxos de dados do middleware EXEHDA à recuperação semântica de documentos normativos para análise por modelos de linguagem. A arquitetura permite identificar violações de privacidade e gerar justificativas fundamentadas em evidências legais. A avaliação, com cenários simulados de tratamento de dados em IoT, comparou um modelo LLM isolado com a abordagem RAG. Os resultados indicam melhora na classificação e na explicabilidade das decisões, mostrando potencial para apoiar auditorias e governança de dados.*

1. Introdução

A Internet das Coisas (IoT) tem se consolidado como uma infraestrutura essencial para coleta e processamento contínuo de dados provenientes de dispositivos distribuídos, sensores e sistemas embarcados. Esses dispositivos frequentemente coletam dados pessoais ou potencialmente identificáveis, muitas vezes de forma contínua e heterogênea, o que levanta desafios importantes relacionados à privacidade e à governança de dados.

No contexto brasileiro, o tratamento desses dados deve observar as disposições da Lei Geral de Proteção de Dados Pessoais (LGPD), que estabelece princípios, direitos dos titulares e obrigações para organizações que realizam tratamento de dados pessoais. A LGPD exige que organizações adotem medidas capazes de demonstrar conformidade com princípios como finalidade, adequação, necessidade, transparência e segurança.

Entretanto, garantir conformidade com a LGPD em ambientes IoT apresenta desafios significativos. Sistemas IoT são caracterizados por fluxos contínuos de dados,

múltiplos dispositivos distribuídos e integrações complexas entre serviços e plataformas. Nessas condições, a verificação manual de conformidade torna-se impraticável em larga escala.

Recentemente, modelos de linguagem baseados em Inteligência Artificial (IA) têm sido utilizados para auxiliar na interpretação de normas jurídicas e na análise documental. Contudo, o uso direto desses modelos pode resultar em respostas imprecisas ou não fundamentadas. A abordagem de *Retrieval-Augmented Generation* (RAG) surge como uma alternativa promissora, pois permite que modelos generativos produzam respostas baseadas em documentos previamente indexados.

A abordagem discutida neste artigo, denominada Abordagem LGRAG-IoT: Contribuição para a Conformidade do Middleware EXEHDA com a Lei Geral de Proteção de Dados, tem como objetivo central abordar o problema de pesquisa relacionado a integração de mecanismos automatizados de interpretação normativa baseados em IA a um middleware IoT baseado em serviços, de forma a auxiliar na verificação contínua de conformidade com a LGPD durante o processamento de dados.

A principal contribuição deste trabalho é a proposição de uma arquitetura que integra ao *middleware* EXEHDA monitoramento de fluxos de dados IoT, recuperação semântica de documentos legais e geração de justificativas auditáveis para decisões de compliance. A LGRAG-IoT tem como objetivos: (i) modelar formalmente fluxos de dados em ambientes IoT sob a perspectiva de compliance; (ii) projetar uma arquitetura capaz de monitorar eventos de tratamento de dados; (iii) integrar um mecanismo RAG para interpretação normativa e geração de justificativas; (iv) produzir relatórios explicáveis de conformidade e (v) gerar alertas de potencial não conformidade.

Este artigo está organizado em sete seções. A segunda Seção apresenta conceitos julgados interessantes quando da revisão de literatura em relação à proposta. Na terceira Seção são discutidos os trabalhos relacionados. Na quarta Seção é apresenta a abordagem LGRAG-IoT tratando suas principais características. A quinta Seção apresenta as tecnologias utilizadas. Na sexta Seção são apresentados os testes realizados da abordagem LGRAG-IoT. Por fim, a sétima e última Seção apresenta as considerações finais e os trabalhos futuros.

2. Escopo do Trabalho

Nesta seção são apresentados conceitos, julgados relevantes, quando da revisão da literatura em relação à proposta desenvolvida.

Lei Geral de Proteção de Dados

A Lei Geral de Proteção de Dados Pessoais (LGPD) [da República 2018], Lei nº 13.709/2018, regula no Brasil o tratamento de dados pessoais e altera dispositivos do Marco Civil da Internet. Atualmente, mais de 126 países possuem legislações semelhantes, voltadas à regulamentação do uso dessas informações, prevenção de abusos e responsabilização das organizações.

Aplicações que utilizam o EXEHDA, especialmente na área da saúde, armazenam dados pessoais e também dados pessoais do tipo sensível (Artigo 11º). A presente abordagem visa proporcionar um monitoramento de eventos de tratamento de dados pessoais, utilizando RAG e gerando alertas quando da ocorrência de alguma inconformidade.

Large Language Models (LLMs)

Os modelos de LLMs (do inglês *Large Language Models*), representam um avanço relevante no processamento de linguagem natural (PLN), alcançando desempenho de estado da arte em tarefas como tradução automática, sumarização e resposta a perguntas. Treinados em grandes corpora textuais e compostos por bilhões de parâmetros, esses modelos apresentam elevada capacidade de generalização [Brown et al. 2020]. Destaca-se também sua capacidade de adaptação a novas tarefas com base em poucos exemplos ou apenas por instruções em linguagem natural, sem necessidade de treinamento adicional específico [OpenAI 2023].

Apesar de seu potencial, os LLMs apresentam limitações como a geração de respostas incorretas, conhecida como alucinação, e dificuldades em lidar com informações atualizadas, pois seu conhecimento permanece estático após o treinamento. Essas limitações têm impulsionado arquiteturas híbridas, como a RAG, que combinam recuperação de informação e geração textual para aumentar a precisão das respostas [Lewis et al. 2020]. Estudos recentes também discutem aspectos de explicabilidade, auditabilidade e impactos éticos desses modelos [Kandpal et al. 2023].

Retrieval-Augmented Generation (RAG)

A RAG pode ser compreendida como uma técnica de otimização da saída de LLMs em que a geração textual é condicionada a informações provenientes de uma base de conhecimento externa e confiável, não restrita aos dados de treinamento [Lewis et al. 2020]. Essa abordagem permite adaptar modelos a domínios específicos ou bases institucionais sem necessidade de retreinamento, constituindo uma alternativa eficiente para aumentar a relevância e a precisão das respostas.

Apesar da elevada capacidade dos LLMs em tarefas de raciocínio e geração textual, ainda persistem limitações como alucinações e a ausência de atualização contínua do conhecimento. A RAG busca mitigar esses problemas ao combinar mecanismos de recuperação de informação, baseados em busca semântica, com a geração contextualizada por LLMs [Lewis et al. 2020].

Sua arquitetura, conforme a Figura 1, é geralmente estruturada em três etapas: (i) indexação de documentos em um banco vetorial por meio de *embeddings*; (ii) recuperação de passagens relevantes com base em similaridade semântica; e (iii) geração de respostas pelo LLM utilizando os trechos recuperados como contexto [Izacard and Grave 2021].

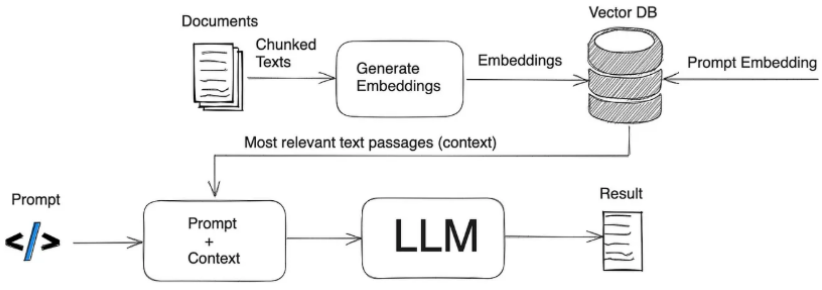

Figura 1. Visão Geral da RAG. Fonte: [Youssef 2023]

3. Trabalhos Relacionados

A preocupação com privacidade, governança de dados e conformidade regulatória em ambientes IoT tem motivado pesquisas recentes em diferentes frentes, incluindo *frameworks* para compartilhamento seguro de dados, ontologias para representação de requisitos legais, mecanismos automáticos de verificação de compliance e aplicações de modelos de linguagem no domínio jurídico.

Chhetri et al. [Chhetri et al. 2024] (Art. 1) apresentam um *framework* semântico para compartilhamento de dados IoT com foco em interoperabilidade, qualidade e conformidade com o GDPR. Embora avancem na governança de dados, não contemplam interpretação normativa automatizada nem geração de justificativas explicáveis.

Kurteva et al. [Kurteva et al. 2023] (Art. 2) propõem a ontologia *smashHitCore* para compartilhamento de dados sensoriais compatível com o GDPR. A abordagem fornece uma base formal para representação de requisitos legais, mas não integra técnicas de IA generativa para análise normativa.

Azam et al. [Azam et al. 2025] (Art. 3) apresentam uma solução para modelagem de conformidade com o GDPR baseada em *Defeasible Logic Programming*, capaz de identificar não conformidades e sugerir mitigação. Contudo, não é voltada a middlewares IoT nem utiliza recuperação documental com LLMs.

Guha et al. [Guha et al. 2023] (Art. 4) apresentam o *LegalBench*, um benchmark para avaliação de raciocínio jurídico em modelos de linguagem. Embora evidencie o potencial dos LLMs no domínio legal, não aborda fundamentação documental nem compliance em IoT.

Gao et al. [Gao et al. 2024] (Art. 5) revisam a arquitetura RAG, destacando sua capacidade de integrar conhecimento externo e reduzir alucinações. Apesar disso, não tratam diretamente de IoT nem de conformidade legal.

A Tabela 1 compara os trabalhos relacionados com a proposta deste artigo. Observa-se que os estudos existentes abordam aspectos relevantes, como compartilhamento seguro de dados IoT, modelagem formal de compliance, raciocínio jurídico com LLMs e recuperação aumentada por documentos. Contudo, nenhum integra simultaneamente monitoramento de fluxos IoT, verificação automatizada de conformidade regulatória, geração explicável de justificativas e uso de RAG. A arquitetura proposta combina esses elementos em um *framework* de compliance explicável para ambientes IoT.

Tabela 1. Comparação entre trabalhos relacionados e a arquitetura proposta

Artigo	IoT	Privacidade	Compliance	IA	Explicabilidade	RAG
1	✓	✓	parcial	✗	✗	✗
2	✓	✓	✓	✗	parcial	✗
3	parcial	✓	✓	✗	parcial	✗
4	✗	✗	parcial	✓	parcial	✗
5	✗	✗	✗	✓	✓	✓
Este trabalho	✓	✓	✓	✓	✓	✓

4. Arquitetura da LGRAG-IoT

A arquitetura de software concebida para abordagem LGRAG-IoT está apresentada na Figura 2. Na continuidade desta seção são tratadas as funcionalidades dos diferentes módulos, sendo discutidos seus perfis operacionais.

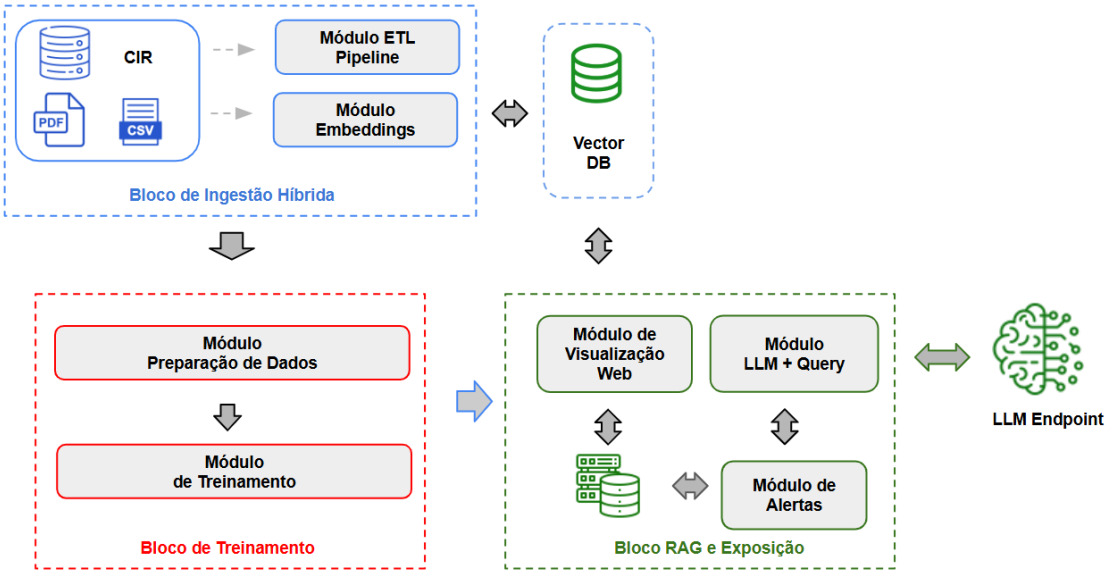

Figura 2. Visão Geral da Arquitetura da LGRAG-IoT

.

4.1. Middleware EXEHDA

EXEHDA é um *middleware* baseado em serviços, projetado para criar e gerenciar um ambiente de computação altamente distribuído, além de viabilizar a execução de aplicações sensíveis ao contexto sobre o mesmo. Este artigo traduz o avanço dos trabalhos grupo registrado nas publicações [Albandes et al. 2024]. O EXEHDA organiza-se em células autônomas interconectadas. Cada célula, para prover suporte a *Situation Awareness*, inclui um *Context Server* (CS) e múltiplos *Edge Servers* (ES) e/ou *Gateways*. Os *Gateways*, usualmente empregam hardware embarcado, coletam dados de sensores e padronizam protocolos para os ES. O processamento contextual é distribuído entre o *Edge Server* e o *Context Server*.

Os *Edge Servers* transmitem dados para o *Context Server*, responsável pelo seu processamento. O *Context Server* integra esses dados com registros históricos no seu *Contextual Information Repository* (CIR), o qual é persistido em banco de dados.

4.2. Bloco de Ingestão Híbrida

O bloco de Ingestão Híbrida é constituído pelo Módulo ETL Pipeline e pelo Módulo de Embeddings, responsáveis pela transformação e indexação vetorial dos dados.

4.2.1. Módulo ETL Pipeline

O módulo *ETL Pipeline*, é responsável pela extração, transformação e carga dos dados provenientes do CIR e de fontes de dados relevantes para a análise de conformidade,

incluindo bases estruturadas provenientes de sistemas legados, arquivos CSV contendo registros de eventos e documentos normativos em formato textual ou PDF, como a Lei Geral de Proteção de Dados, resoluções da Autoridade Nacional de Proteção de Dados (ANPD) e políticas institucionais.

4.2.2. Módulo de Embeddings

O módulo de *embeddings* converte os documentos textuais e registros previamente tratados pelo Módulo ETL Pipeline em representações vetoriais densas, capazes de capturar relações semânticas entre termos presentes em textos legais, políticas internas e dados da LGRAG-IoT. Essas representações possibilitam buscas semânticas eficientes, superando limitações de métodos baseados apenas em correspondência literal. Esse módulo é essencial para o mecanismo RAG, permitindo a recuperação de informações com base em similaridade semântica.

O *Vector Database* armazena os vetores gerados pelo módulo de embeddings e fornece mecanismos eficientes de busca por similaridade. Diferentemente de bancos de dados tradicionais, é otimizado para consultas em espaços vetoriais de alta dimensionalidade, permitindo a rápida recuperação de documentos ou trechos normativos relacionados à consulta. Na arquitetura proposta, armazena representações da LGPD, regulamentos da ANPD e políticas organizacionais, possibilitando a recuperação de evidências jurídicas relevantes para fundamentar respostas e análises de conformidade.

4.3. Bloco de Treinamento

O Bloco de Treinamento reúne os componentes responsáveis pela preparação dos dados e pela execução dos modelos analíticos utilizados na análise de conformidade.

4.3.1. Módulo de Preparação de Dados

O módulo de preparação de dados organiza e estrutura as informações processadas pelo Módulo ETL Pipeline para uso em modelos analíticos e sistemas de inteligência artificial. Nessa etapa são aplicadas técnicas de seleção de atributos, agregação e anonimização ou pseudonimização quando necessário, formando conjuntos de dados adequados para treinamento ou inferência. O módulo também incorpora pré-processamentos voltados à privacidade, assegurando o tratamento de informações sensíveis conforme princípios da LGPD, como minimização e proteção de dados pessoais.

4.3.2. Módulo de Treinamento

O módulo de treinamento é responsável pela criação e ajuste de modelos analíticos ou de aprendizado de máquina para detectar padrões de uso de dados e possíveis violações de políticas de privacidade. Nessa etapa podem ser aplicadas técnicas supervisionadas ou não supervisionadas para identificar comportamentos anômalos, fluxos de dados inadequados ou inconsistências no tratamento de dados pessoais. O resultado são modelos capazes de apoiar a análise automatizada de conformidade do *middleware* EXEHDA.

4.3.3. Módulo LLM + Query

O módulo *LLM + Query* atua como núcleo cognitivo da arquitetura baseada em RAG. Ele recebe consultas sobre conformidade, privacidade ou auditoria e realiza recuperação semântica no banco vetorial para obter trechos relevantes da base normativa. Esses trechos são usados como contexto para um modelo de LLM, que gera respostas fundamentadas nos documentos recuperados, reduzindo o risco de alucinações e aumentando a confiabilidade das análises de compliance.

O *LLM Endpoint* representa o serviço responsável por hospedar o modelo de linguagem utilizado pela LGRAG-IoT. Esse componente recebe consultas enriquecidas com contexto recuperado do banco vetorial e gera respostas interpretativas que auxiliam na análise de conformidade. Dependendo da implementação, pode corresponder a um serviço de IA em nuvem ou a um modelo hospedado localmente.

4.3.4. Módulo de Visualização Web

O módulo de visualização web da LGRAG-IoT fornece uma interface interativa para que administradores, auditores de privacidade e responsáveis pela governança de dados explorem os resultados produzidos pela arquitetura. A interface permite visualizar consultas, evidências normativas recuperadas, análises de conformidade e indicadores de risco associados ao uso de dados em dispositivos IoT, facilitando a interpretação dos resultados e a tomada de decisão.

4.3.5. Módulo de Alertas

O módulo de alertas monitora continuamente os resultados das análises produzidas pela arquitetura e gera notificações quando potenciais violações de privacidade ou inconsistências com a LGPD são detectadas. Esses alertas podem ser enviados a administradores ou equipes de governança de dados, permitindo respostas rápidas a incidentes ou comportamentos suspeitos. O módulo também mantém registros de eventos para fins de auditoria e rastreabilidade.

5. Tecnologias Utilizadas

A implementação da abordagem LGRAG-IoT fundamenta-se em um conjunto de tecnologias complementares, selecionadas de acordo com critérios de desempenho, escalabilidade e conformidade com a LGPD.

5.1. Bancos de Dados Vetoriais e Qdrant

O crescimento de dados não estruturados, como textos, imagens e áudio, tem impulsionado o uso de bancos de dados vetoriais para buscas semânticas eficientes. Diferentemente dos bancos relacionais, esses sistemas armazenam representações em alta dimensão (*embeddings*), permitindo consultas baseadas em similaridade [Johnson et al. 2019]. Entre as soluções disponíveis, destaca-se o Qdrant[1], adotado neste trabalho por sua natureza

[1]Disponível em: https://qdrant.tech/documentation

open source, alta performance e suporte a *payloads* estruturados, o que possibilita consultas híbridas relevantes em cenários de RAG.

5.2. Armazenamento de Objetos

Para complementar a arquitetura proposta, foi adotado o Amazon Simple Storage Service (Amazon S3), serviço de armazenamento de objetos em nuvem disponibilizado pela Amazon Web Services (AWS). O Amazon S3 é amplamente reconhecido por sua escalabilidade, durabilidade e integração nativa com diversos serviços de nuvem, características que o tornam adequado para aplicações distribuídas de grande porte [2].

Na LGRAG-IoT, o Amazon S3 é utilizado para o gerenciamento de documentos institucionais, relatórios exportados em formatos como PDF e CSV, bem como para o armazenamento de *snapshots* de embeddings. Dessa forma, atua como repositório complementar às bases relacionais e vetoriais, assegurando persistência e disponibilidade de dados que não se enquadram diretamente no modelo tabular ou vetorial.

6. LGRAG-IoT: Avaliação

A verificação de compliance na arquitetura proposta é realizada por um processo híbrido que combina recuperação semântica de documentos normativos com inferência contextual por modelo de linguagem. O fluxo completo do algoritmo utilizado nesse processo está disponível no repositório do projeto no GitHub[3] . Nesse fluxo, cada evento ou fluxo de dados IoT é analisado com base em seus atributos, confrontado com a base normativa recuperada via RAG e, por fim, classificado quanto à conformidade com a LGPD.

O algoritmo evidencia que a decisão de conformidade não é produzida exclusivamente pelo modelo de linguagem, mas por sua interação com evidências normativas previamente recuperadas do banco vetorial. Essa estratégia melhora a rastreabilidade das respostas, reduz o risco de alucinação e favorece a explicabilidade do processo decisório, aspecto essencial em cenários de compliance legal e auditoria de ambientes IoT.

A avaliação da abordagem proposta analisa a capacidade da LGRAG-IoT em identificar situações de conformidade e não conformidade em fluxos de dados de ambientes IoT. Para isso, utiliza-se um conjunto de cenários experimentais que simulam diferentes formas de coleta, processamento e compartilhamento de dados pessoais. Cada cenário é previamente anotado com o estado esperado de conformidade segundo a LGPD e políticas institucionais.

O algoritmo é aplicado a esses fluxos utilizando RAG para recuperar evidências normativas e o modelo de linguagem para interpretar o contexto operacional. O desempenho é avaliado por métricas de classificação, como acurácia, precisão, revocação e F1-score, além de critérios qualitativos relacionados à explicabilidade das respostas. Assim, a avaliação permite verificar a capacidade da abordagem em detectar violações de privacidade e justificar decisões de compliance.

Os experimentos utilizam um *dataset*[4] que simula fluxos de dados gerados por dispositivos IoT e seus respectivos metadados de privacidade. Cada registro representa

[2]Disponível em: https://aws.amazon.com/s3
[3]Disponível em: https://github.com/albandes/LGRAG-IoT/blob/main/docs/LGRAG-IoT-Algorithm.pdf
[4]Disponível em: https://github.com/albandes/LGRAG-IoT

um cenário de processamento de dados, que pode ou não estar em conformidade com os princípios da LGPD, descrevendo informações como origem do dispositivo, tipo de dado coletado e finalidade do tratamento.

A arquitetura foi avaliada em dois cenários: (i) LLM puro: o modelo responde às consultas sem acesso a documentos normativos e (ii) RAG + LLM: o modelo recebe contexto adicional recuperado de um banco vetorial contendo documentos da LGPD, resoluções da ANPD e políticas organizacionais.

6.1. Métricas de Avaliação

O desempenho da abordagem foi avaliado utilizando métricas tradicionais de classificação: (i) Acurácia: proporção de classificações corretas, (ii) Precisão: proporção de classificações positivas corretas, (iii) Revocação: capacidade de identificar casos de não conformidade e (iv) F1-score: média harmônica entre precisão e revocação.

Também foi realizada análise qualitativa da explicabilidade das respostas, observando a presença de justificativas baseadas em trechos normativos recuperados. A Tabela 2 apresenta a comparação entre os cenários avaliados.

Tabela 2. Comparação entre LLM puro e arquitetura RAG proposta

Abordagem	Acurácia	Precisão	Revocação	F1-score
LLM puro	0.78	0.74	0.71	0.72
RAG + LLM (proposta)	0.91	0.89	0.88	0.88

Os resultados indicam que o uso de RAG melhora significativamente o desempenho da verificação de compliance. A recuperação de documentos normativos permite que o modelo produza respostas mais consistentes e alinhadas às regras jurídicas.

Além disso, observou-se maior explicabilidade das respostas no cenário com RAG, onde as decisões puderam ser justificadas por trechos específicos da legislação ou de políticas institucionais. Esses resultados reforçam o potencial da abordagem RAG para reduzir problemas de alucinação em modelos generativos e aumentar a confiabilidade de sistemas baseados em inteligência artificial aplicados a contextos legais e regulatórios.

7. Conclusão

Este trabalho apresentou uma arquitetura para verificação automatizada de conformidade com a LGPD em ambientes IoT utilizando RAG. A proposta integra monitoramento de fluxos de dados, recuperação de documentos legais, geração de justificativas explicáveis e emissão de alertas.

As principais contribuições deste trabalho são: (i) A proposição de uma arquitetura voltada à verificação automatizada de compliance do EXEHDA; (ii) Formalização do problema de conformidade em fluxos de dados IoT; (iii) Integração de RAG para análise normativa automatizada; (iv) Geração de relatórios explicáveis de conformidade e (v) Emissão de alertas quadno da quebra de conformidade.

Como trabalhos futuros, pretende-se implementar um protótipo da arquitetura e avaliar sua eficácia em cenários reais de IoT, analisando métricas como precisão na detecção de não conformidades, qualidade das explicações geradas e escalabilidade da LGRAG-IoT.

8. Declaração sobre o Uso de Inteligência Artificial

Em atendimento ao Código de Conduta para autores da SBC, declaramos explicitamente o uso de ferramentas de Inteligência Artificial Generativa exclusivamente para auxiliar na revisão ortográfica e gramatical do texto.

Referências

Albandes, R., Lambrecht, R., Pieper, L., Barcellos, F., Pernas, A. M., and Yamin, A. (2024). Abordagem iot db-audit: uma contribuição a adequação do middleware exehda à lei geral de proteção de dados. In *Simpósio Brasileiro de Computação Ubíqua e Pervasiva (SBCUP)*, pages 51–60. SBC.

Azam, N., Chak, A., Michala, A., Ansari, S., and Truong, N. B. (2025). A practical solution for modelling gdpr-compliance based on defeasible logic reasoning. *Expert Systems with Applications*, 271:127140.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:1877–1901.

Chhetri, T. R., Dehury, C. K., Varghese, B., Fensel, A., Srirama, S. N., and DeLong, R. J. (2024). Enabling privacy-aware interoperable and quality iot data sharing with context. *Future Generation Computer Systems*, 157:164–179.

da República, P. (2018). Lei geral de proteção de dados pessoais. Último acesso 17 março 2023.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., and Wang, H. (2024). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Guha, N., Nyarko, J., Ho, D. E., Ré, C., Chilton, A., et al. (2023). Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *arXiv preprint arXiv:2308.11462*.

Izacard, G. and Grave, E. (2021). Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 874–880.

Johnson, J., Douze, M., and Jégou, H. (2019). Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Kandpal, N., Zhao, V., and Roberts, A. (2023). Towards explainable retrieval-augmented generation: Challenges and opportunities. *arXiv preprint arXiv:2304.12345*.

Kurteva, A., Chhetri, T. R., Tauqeer, A., Hilscher, R., Fensel, A., Nagorny, K., Correia, A., Zilverberg, A., Schestakov, S., Funke, T., and Demidova, E. (2023). The smashhitcore ontology for gdpr-compliant sensor data sharing in smart cities. *Sensors*, 23(13):6188.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

OpenAI (2023). Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

Youssef, H. (2023). The rag spectrum: Exploring 7 distinct approaches. https://youssefh.substack.com/p/the-rag-spectrum-exploring-7-distinct. Accessed: 27 Sep. 2025.