

Uma Arquitetura de Software para Integrar Raciocínio Baseado em Casos e Grandes Modelos de Linguagem para Certificação de Produtos Pesqueiros

Gabriel Vieira Casanova¹, Pedro Bilar Montero¹, Gabriel Rodrigues da Silva¹,
Thiago Paulo Both¹, Alencar Machado¹, Vinicius Maran¹

¹ Laboratory of Ubiquitous, Mobile and Applied Computing (LUMAC),
Federal University of Santa Maria, Santa Maria, Brazil

{gabriel.casanova, pedro.montero, gabriel-silva.3, thiago.both}@acad.ufsm.br
alencar.machado@ufsm.br, vinicius.maran@ufsm.br

Abstract. *Processos regulatórios digitais, como o Certificado de Acreditação de Origem Legal (CAOL) na Plataforma Nacional da Indústria Pesqueira (PNIP), exigem múltiplas etapas e conformidade normativa, tornando-se complexos e suscetíveis a erros. Este artigo propõe uma arquitetura de agente inteligente baseada em Model Context Protocol (MCP), Retrieval-Augmented Generation (RAG) e Case-Based Reasoning (CBR) para apoiar a submissão e análise do CAOL. O RAG fornece respostas fundamentadas em normativas oficiais, enquanto o CBR recupera precedentes similares para apoiar estimativas de sucesso. Os resultados preliminares indicam potencial para reduzir carga operacional, melhorar rastreabilidade e apoiar decisões em contextos regulatórios.*

Resumo. *Processos regulatórios digitais, como o Certificado de Acreditação de Origem Legal (CAOL) na Plataforma Nacional da Indústria Pesqueira (PNIP), envolvem múltiplas etapas e elevada complexidade normativa. Este artigo propõe uma arquitetura de agente inteligente baseada em Model Context Protocol (MCP), Retrieval-Augmented Generation (RAG) e Case-Based Reasoning (CBR) para apoiar a submissão e análise do CAOL. O RAG fornece respostas fundamentadas em documentação oficial, enquanto o CBR recupera precedentes administrativos similares para apoiar estimativas de sucesso. Resultados preliminares sugerem potencial para reduzir carga burocrática, melhorar rastreabilidade e apoiar decisões regulatórias.*

1. Introdução

O cumprimento das normas no setor pesqueiro é essencial para garantir origem legal, sustentabilidade e acesso a mercados globais. No Brasil, a Plataforma Nacional da Indústria Pesqueira (PNIP) centraliza processos de certificação, incluindo o Certificado de Acreditação de Origem Legal (CAOL), responsável por validar a origem legal dos produtos da pesca. Embora necessário, o processo envolve múltiplas etapas, instruções normativas e despachos, tornando-se complexo e dependente de conhecimento especializado [Ministério da Pesca e Aquicultura 2024].

Grandes Modelos de Linguagem (LLMs) oferecem potencial para simplificar processos burocráticos por meio de análise documental, preenchimento assistido e suporte

à decisão. Entretanto, sua aplicação em contextos regulatórios enfrenta limitações como conhecimento estático, alucinações factuais e incapacidade de interagir diretamente com sistemas externos.

Para mitigar essas limitações, este trabalho integra Model Context Protocol (MCP), Retrieval-Augmented Generation (RAG) e Case-Based Reasoning (CBR). O MCP conecta o LLM à PNIP de forma controlada, o RAG fundamenta respostas em normativas oficiais e o CBR recupera casos similares para apoiar recomendações e estimativas de sucesso. Essa abordagem integrada visa mudar o papel do usuário de executor manual para supervisor de um fluxo de trabalho assistido por IA. A contribuição deste trabalho está na integração auditável de MCP, RAG e CBR para suporte regulatório no contexto do CAOL/PNIP, combinando recuperação normativa, precedentes administrativos e rastreabilidade em uma arquitetura supervisionada por humanos.

2. Fundamentação Teórica

Esta seção fornece os fundamentos teóricos e tecnológicos que sustentam a solução proposta, incluindo Grandes Modelos de Linguagem, Geração Aumentada por Recuperação, Raciocínio Baseado em Casos, o Model Context Protocol e o contexto institucional da PNIP e do CAOL.

2.1. Grandes Modelos de Linguagem (LLMs)

Grandes Modelos de Linguagem são uma classe de modelos de IA projetados para processar, compreender e gerar linguagem humana em grande escala. Baseados em arquiteturas transformer com mecanismos de autoatenção, os LLMs capturam informações contextuais e suportam diversas tarefas, como classificação de texto, sumarização e resposta a perguntas [Brown 2020, Vaswani 2017]. Normalmente pré-treinados em grandes conjuntos de dados, esses modelos adquirem ampla competência linguística que pode ser adaptada por meio de ajuste fino ou ajuste de instrução [Bommasani 2021].

No entanto, os LLMs apresentam limitações importantes, incluindo sensibilidade às distribuições dos dados de treinamento e uma tendência a produzir saídas plausíveis, mas factualmente incorretas. Essas limitações motivam o uso de mecanismos complementares, como recuperação de conhecimento externo e raciocínio estruturado.

2.2. Geração Aumentada por Recuperação (RAG)

RAG é uma abordagem híbrida que combina geração neural de linguagem com recuperação externa de informações. Ao incorporar conhecimento recuperado de fontes externas no momento da inferência, o RAG melhora a fundamentação factual e a relevância contextual além do que está codificado nos parâmetros do modelo [Lewis 2020].

Em um sistema RAG, as consultas dos usuários são transformadas em solicitações de recuperação contra repositórios de conhecimento externos, com os documentos recuperados sendo injetados no contexto do modelo. Essa separação entre modelagem de linguagem e armazenamento de conhecimento permite a incorporação de informações atualizadas ou específicas do domínio sem retreinamento, reduzindo alucinações e apoiando a rastreabilidade [Chen 2023].

2.3. Raciocínio Baseado em Casos (CBR)

CBR é um paradigma de IA no qual novos problemas são abordados recordando, adaptando e reutilizando soluções de casos previamente resolvidos [Kolodner 1992]. Cada caso captura uma situação problema, a solução aplicada e o resultado, enfatizando o conhecimento experiencial. Um sistema CBR típico segue um ciclo de quatro etapas: recuperação, reutilização, revisão e retenção. O CBR oferece vantagens como aprendizado incremental e explicabilidade, pois as decisões podem ser rastreadas até casos anteriores — características valiosas em domínios regulatórios. Os desafios incluem projetar métricas de similaridade eficazes e manter bases de casos coerentes [Watson 1997].

2.4. Model Context Protocol (MCP)

MCP é um protocolo aberto proposto pela Anthropic para padronizar como os LLMs recebem informações contextuais e interagem com sistemas externos [Anthropic 2024]. Projetado em torno de uma arquitetura cliente-servidor usando JSON-RPC 2.0, o MCP impõe a separação entre componentes de raciocínio (LLMs) e componentes operacionais (serviços externos).

Por meio de três tipos de interface — recursos (dados somente leitura), ferramentas (operações executáveis) e prompts (modelos de interação) — o MCP permite uma integração controlada e auditável do modelo. Seu modelo de segurança baseado em consentimento garante que apenas dados autorizados sejam acessados, apoiando os requisitos de governança em ambientes institucionais.

2.5. Plataforma Nacional da Indústria Pesqueira (PNIP) e CAOL

A PNIP é uma plataforma governamental digital estabelecida pelo Ministério da Pesca e Aquicultura do Brasil como o sistema oficial para certificações relacionadas à pesca. Operacional desde dezembro de 2024, a PNIP centraliza os fluxos de trabalho de certificação para reduzir a complexidade burocrática e aumentar a rastreabilidade [Ministério da Pesca e Aquicultura 2024].

O CAOL é um certificado eletrônico que atesta a origem legal dos produtos da pesca. Sua submissão envolve múltiplas etapas (dados de produção, embarcações, indústria, mapas de bordo, documentos fiscais) e um processo de aprovação sequencial composto por quatro papéis, conforme delineado na Tabela 1. Embora essa validação em vários níveis garanta confiabilidade, ela introduz uma complexidade operacional significativa, tornando o CAOL um caso de uso relevante para assistência inteligente [Arner 2020].

2.6. Trabalhos Relacionados

Sistemas baseados em RAG demonstraram precisão aprimorada em domínios jurídicos e financeiros ao incorporar fontes autoritativas [Lewis 2020]. Pesquisas em governo eletrônico têm explorado agentes orientados por IA apoiando fluxos de trabalho administrativos, enquanto o CBR tem fornecido há muito tempo suporte à decisão por meio de raciocínio analógico [Aamodt and Plaza 1994]. No entanto, trabalhos anteriores não abordaram especificamente os processos CAOL dentro da PNIP. Este artigo contribui ao integrar essas técnicas em um agente baseado em MCP adaptado à certificação pesqueira.

Tabela 1. Papéis de aprovação no processo sequencial do CAOL

Papel	Responsabilidade
Analista	Revisa a documentação submetida quanto à integridade e conformidade com as instruções normativas; verifica a consistência entre dados de produção, mapas de bordo e documentos fiscais
Coordenador	Valida a avaliação do analista; assegura o alinhamento com os procedimentos departamentais; pode solicitar esclarecimentos adicionais
Diretor	Realiza a revisão técnica e administrativa final; aprova ou rejeita a certificação no nível divisional
Superintendente	Emite o certificado final após confirmar a conformidade com todas as aprovações anteriores; conclui o processo

3. Metodologia

Esta seção descreve uma arquitetura baseada em módulos que integra três sistemas complementares — Model Context Protocol (MCP), Geração Aumentada por Recuperação (RAG) e Raciocínio Baseado em Casos (CBR) — para mediar a interação entre usuários (por exemplo, produtores, representantes técnicos, analistas) e a plataforma PNIP. A arquitetura de alto nível é mostrada na Figura 1.

3.1. Módulo MCP (Ferramentas e Recursos)

O servidor MCP atua como a interface entre a camada de inteligência e a PNIP. Ele registra e expõe um catálogo de ferramentas (ações executáveis) e recursos (dados somente leitura) com assinaturas claramente definidas, controle de acesso e semântica de auditoria. Cada invocação de ferramenta deve ser autenticada, autorizada e registrada em um log de auditoria imutável (timestamp, chamador, ferramenta, resumo da entrada, referência da saída). Os recursos (documentos, registros de cadastro) são fornecidos como entradas contextuais somente leitura; as ferramentas realizam operações de alteração de estado sob supervisão explícita. O servidor MCP impõe limites de taxa, validação de entrada e contratos de esquema para limitar solicitações errôneas ou maliciosas.

3.2. Ferramenta RAG

O subsistema RAG ancora as saídas do modelo de linguagem em documentação oficial e fornece citações de fontes rastreáveis.

Fluxo de trabalho RAG (resumo)

1. Pergunta do usuário (em uma etapa específica) → embedding da consulta.
2. Recuperar os top- k blocos do armazenamento de vetores (incluir referências de metadados).
3. Construir um prompt que inclua as evidências recuperadas e um modelo de instrução.
4. Invocar o LLM via MCP com o prompt; o LLM retorna uma resposta resumida mais citações.

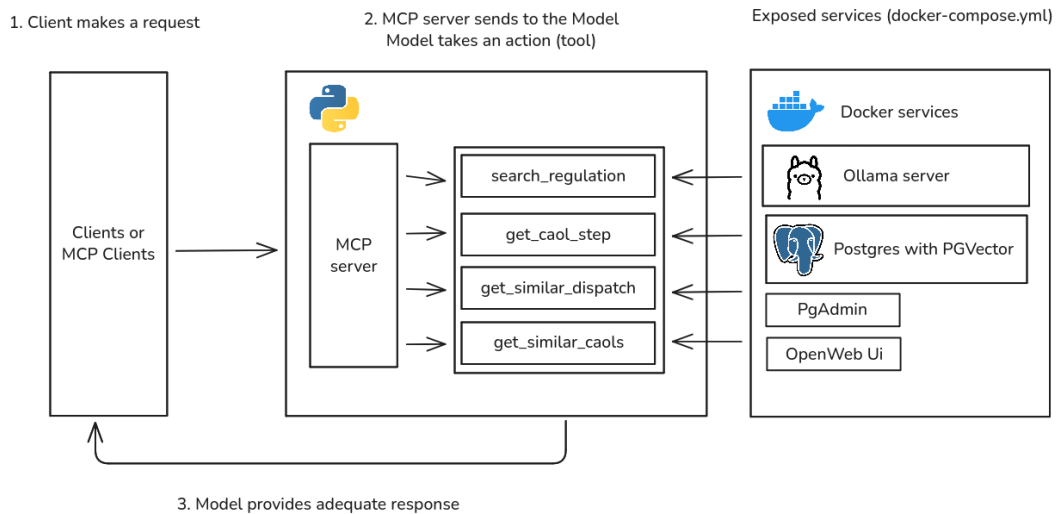


Figura 1. Projeto do sistema proposto para assistência inteligente aos processos CAOL na PNIP.

5. Armazenar a pergunta, os blocos recuperados e a resposta gerada para auditoria e melhorias incrementais.

A Figura 2 apresenta a arquitetura do pipeline RAG utilizada como base conceitual do subsistema de recuperação. Embora derivada de trabalho anterior, sua adaptação neste estudo incorpora integração com MCP e suporte contextual específico para os fluxos regulatórios da PNIP.

3.3. Ferramenta CBR

A ferramenta CBR aproveita casos históricos de CAOL para fornecer assistência inteligente durante o preenchimento do formulário. Cada caso no repositório captura atributos-chave, como espécie, método de aquisição, quantidade produzida, região de origem, despacho final (aprovado, rejeitado ou pendente) e quaisquer pendências identificadas durante a análise. A base de casos contém atualmente 250 CAOLs históricos anonimizados do sistema PNIP.

Os casos históricos foram anonimizados antes do processamento, removendo identificadores pessoais e organizacionais. Cada instância contém atributos categóricos e numéricos, incluindo espécie, método de aquisição, região de origem, quantidade produzida, tipo de despacho e status final da solicitação. Registros incompletos foram tratados por meio de normalização e exclusão de atributos inconsistentes quando necessário. A distribuição entre decisões aprovadas e rejeitadas foi preservada para reduzir distorções durante a recuperação de similaridade.

Quando um usuário começa a preencher um novo formulário CAOL, a ferramenta CBR recupera os casos mais similares com base em uma métrica de similaridade híbrida que combina:

- **Similaridade numérica:** Para atributos contínuos como quantidade produzida, utiliza-se distância euclidiana normalizada.

Tabela 2. Ferramentas MCP e exemplos de assinaturas

Ferramenta	Descrição
<code>search_regulation</code>	Recuperar documentos regulatórios ou Instruções Normativas (INs) por identificador. Entrada: <code>número_da_IN</code> . Saída: trechos de documentos e metadados.
<code>get_caol_step</code>	Retornar o estágio atual do fluxo de trabalho de uma solicitação CAOL dentro da PNIP. Entrada: <code>step_id</code> ou <code>request_id</code> . Saída: descrição da etapa e status.
<code>submit_dispatch</code>	Submeter um despacho técnico ou administrativo como parte do fluxo de avaliação do CAOL. Entrada: dados estruturados do despacho. Saída: confirmação da submissão.
<code>get_similar_cases</code>	Recuperar casos CAOL historicamente similares com base em critérios estruturados. Entrada: critérios de similaridade. Saída: lista ranqueada de casos correspondentes.

- **Similaridade categórica:** Para atributos discretos como espécie e método de aquisição, aplicam-se correspondência exata e similaridade de Jaccard.

A partir dos top- k casos recuperados foi definido como $k = 5$ após testes exploratórios preliminares visando equilíbrio entre diversidade e relevância dos casos recuperados. A ferramenta estima uma taxa de sucesso calculando a proporção de casos aprovados entre o conjunto recuperado, apresentando ao usuário uma estimativa percentual de como submissões similares se saíram historicamente.

O processo CBR segue duas etapas principais:

- **Recuperação:** Identificar casos similares usando a métrica de similaridade híbrida.
- **Retenção:** Após a aprovação ou rejeição final, o novo caso é anexado ao repositório local, permitindo aprendizado contínuo ao longo do tempo.

3.4. Considerações de Implementação

Esta subseção descreve aspectos práticos que influenciam a robustez, escalabilidade e manutenibilidade da arquitetura, com atenção particular à organização dos dados, configuração do modelo e observabilidade do sistema.

Gerenciamento de dados O gerenciamento eficaz de dados é essencial para garantir a separação de preocupações, escalabilidade e governança entre os componentes do sistema. Os dados devem ser isolados logicamente usando esquemas PostgreSQL dedicados e extensões quando apropriado:

- Dados transacionais operacionais (PostgreSQL).

Tabela 3. Parâmetros de configuração do serviço RAG.

Componente	Detalhes de Configuração
Servidor Ollama	Um servidor para executar grandes modelos de linguagem facilmente.
Nome do modelo de embedding	nomic-embed-text.
Nome do modelo LLM	gemma3:12b.
Estratégia de divisão de texto	Divisão de texto baseada em caracteres.
Tamanho do bloco	1000 caracteres.
Sobreposição de blocos	200 caracteres.
Parâmetros de busca por similaridade	Top- k = 3 blocos mais similares recuperados por consulta.

- Embeddings vetoriais e índices de recuperação (PgVector).
- Instantâneos de casos CBR (esquema normalizado no PostgreSQL com descritores vetoriais opcionais).
- Logs de auditoria imutáveis (tabelas somente anexação; opcionalmente protegidas com somas de verificação criptográficas).

Seleção de modelo e embedding As escolhas de modelo e embedding devem ser externamente configuráveis para suportar experimentação, portabilidade e atualizações futuras sem alterações arquiteturais. Slots de configuração reservados devem incluir:

- Identificador do modelo de embedding e dimensionalidade.
- Configuração do endpoint do LLM (implantação local vs. baseada em nuvem), incluindo modelos de prompt e limites de taxa.
- Hiperparâmetros de recuperação, como valores de top- k , limites de pontuação e estratégias opcionais de rerank.

Monitoramento e observabilidade Mecanismos de observabilidade são necessários para avaliar o desempenho do sistema, diagnosticar falhas e melhorar continuamente a qualidade da recuperação e do raciocínio. A instrumentação deve capturar:

- Métricas de invocação de ferramentas, incluindo contagens de chamadas, distribuições de latência e taxas de erro.
- Indicadores de qualidade de recuperação, como precisão dos resultados top- k e sinais de feedback do usuário.

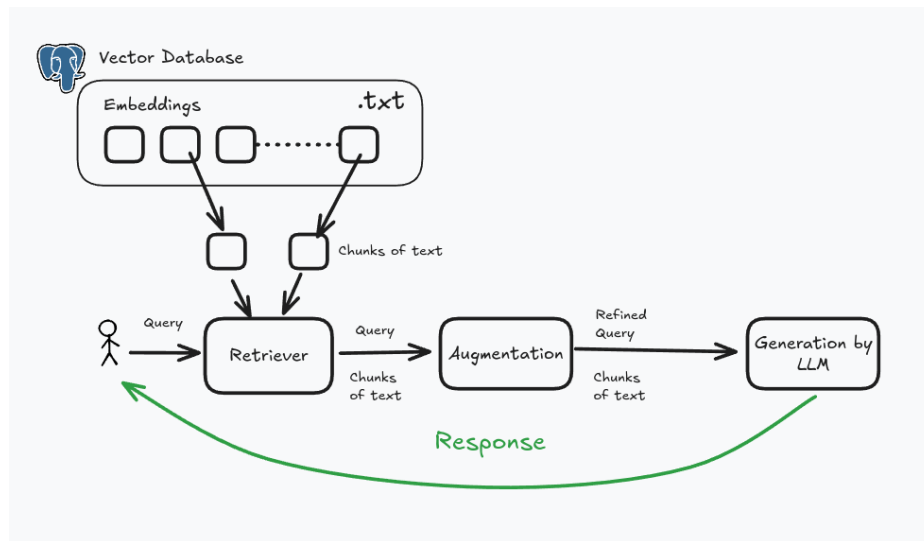


Figura 2. Arquitetura da ferramenta RAG [Casanova et al. 2025].

- Taxas de aceitação das recomendações do CBR e estatísticas de substituição por analistas.

3.5. Avaliação e Validação

Esta subseção define indicadores quantitativos para avaliar a eficácia, usabilidade e confiabilidade do sistema proposto em condições reais. A avaliação deve considerar métricas tanto em nível de sistema quanto centradas no usuário, incluindo:

- Precisão das respostas geradas pelo RAG, medida por meio de precision@k em um conjunto de referência manualmente curado.
- Taxas de aceitação dos usuários e analistas para as sugestões geradas pelo CBR.
- Redução no tempo médio de preenchimento do formulário pelo usuário.
- Correção e pontualidade das submissões de despachos.
- Completude e reprodutibilidade dos registros de auditoria para decisões orientadas por modelos.

3.6. Segurança, Privacidade e Governança

Os requisitos de segurança e governança são críticos para garantir operação confiável, conformidade regulatória e implantação responsável de IA. O projeto do sistema deve aderir aos seguintes princípios:

- Aplicação do princípio do menor privilégio para permissões de ferramentas, com controles explícitos de humano no circuito para operações de escrita ou confirmação.
- Minimização de dados nos prompts, garantindo que apenas informações contextuais estritamente necessárias sejam fornecidas aos modelos.
- Auditabilidade total, em que cada chamada de ferramenta invocada pelo modelo é registrada persistentemente com proveniência clara.
- Controles de acesso baseados em papéis alinhados com as definições da PNIP, combinados com criptografia em repouso e em trânsito.
- Revisão e manutenção periódicas do repositório de casos para remover, anonimizar ou redigir dados pessoais sensíveis de acordo com os requisitos legais.

3.7. Exemplo: O Contrato da Ferramenta CBR no MCP

A ferramenta CBR recupera casos previamente validados similares a uma nova instância do CAOL. Sua entrada consiste em metadados contextuais (tipo de produto, espécie, método de aquisição, quantidade produzida, informações temporais), e sua saída retorna uma lista ranqueada de casos históricos com pontuações de similaridade e proveniência.

O exemplo abaixo mostra um contrato de ferramenta MCP simplificado para consultar casos similares, permitindo inspeção transparente da seleção e reutilização de casos passados.

```
Ferramenta: query_similar_cases
Esquema de entrada: {
  "productId": 13,
  "speciesIds": [32],
  "acquisitionMethod": "Extração",
  "quantity": 120000,
  "hauls": [ { "start": "2025-12-15", "end": "2025-12-16" } ]
}

Esquema de saída: {
  "cases": [
    {
      "case_id": "string",
      "score": 0.92,
      "summary": "...",
      "provenance": "case_db/123"
    },
    ...
  ]
}

Permissões: somente leitura
Auditoria: entrada de log criada com um hash
da entrada e referência da saída.
```

Ao impor esquemas explícitos, permissões de acesso e registro de auditoria, o contrato MCP garante que o uso do CBR permaneça observável e reproduzível, o que é particularmente importante em ambientes do setor público e regulatórios. Em vez de atuar como um módulo de tomada de decisão opaco, o componente CBR opera como um serviço assistivo que fornece precedentes contextualizados aos analistas, complementando explicações regulatórias baseadas em recuperação e o julgamento humano.

4. Simulação, Resultados e Discussão

Esta seção apresenta a avaliação experimental da arquitetura proposta. Um protótipo funcional foi desenvolvido para avaliar o comportamento do sistema integrado em um ambiente controlado, simulando interações reais de usuários com o fluxo de trabalho de certificação CAOL. A avaliação examina três dimensões: eficácia da recuperação do RAG, precisão das sugestões do CBR e impacto geral na eficiência do usuário.

4.1. Protótipo Implementado

Um protótipo funcional foi desenvolvido focando nos módulos RAG e CBR com uma interface PNIP simulada. Os componentes incluem:

- **LLM local:** gemma3:12b via Ollama
- **Backend:** Python com FastMCP
- **Banco de dados vetorial:** PGVector
- **Modelo de embedding:** nomic-embed-text
- **Base de casos:** 250 CAOLs históricos anonimizados

4.2. Métodos de Avaliação

A avaliação aborda três dimensões:

- **Eficácia do RAG:** Precisão@K e Recall@K para recuperação de blocos relevantes; avaliação humana da relevância das respostas
- **Eficácia do CBR:** Precisão das sugestões de campos e precisão da previsão de decisão
- **Impacto no processo:** Tempo médio de conclusão e taxas de erro com e sem assistência

Para avaliação do subsistema RAG, foi elaborado um conjunto de consultas representativas do fluxo CAOL, cobrindo dúvidas procedimentais, normativas e operacionais. O padrão-ouro foi definido com base na documentação oficial da PNIP, sendo a relevância dos blocos recuperados avaliada pela presença de conteúdo normativo suficiente para responder corretamente à consulta.

A avaliação do CBR utilizou os 250 casos históricos anonimizados disponíveis, analisando a capacidade do sistema em recuperar precedentes relevantes e estimar corretamente o desfecho de solicitações similares. Os resultados foram interpretados como apoio à tomada de decisão, e não como substituição do julgamento humano.

A simulação de usuários foi conduzida em ambiente controlado, comparando interações assistidas e não assistidas quanto ao tempo de conclusão e incidência de erros no preenchimento.

4.3. Resultados Preliminares

As principais descobertas incluem:

- **RAG:** A Precisão@5 atingiu 0,87 para recuperação de blocos relacionados a IN; as explicações alcançaram uma classificação média de utilidade de 4,2/5
- **CBR:** A precisão da previsão de decisão atingiu 0,61. Embora moderado, esse resultado é considerado suficiente para apoiar a recuperação de precedentes administrativos similares, não sendo utilizado como mecanismo automatizado de decisão. O objetivo do CBR é oferecer suporte contextual aos solicitantes, preservando supervisão humana no processo regulatório.
- **Simulação de Usuário:** O grupo assistido experimentou uma redução de 34% no tempo de conclusão e uma redução de 60% nas taxas de erro de campo

Os valores apresentados na Figura 3 foram normalizados exclusivamente para fins de comparação visual entre métricas heterogêneas, não implicando equivalência estatística direta entre grandezas distintas.

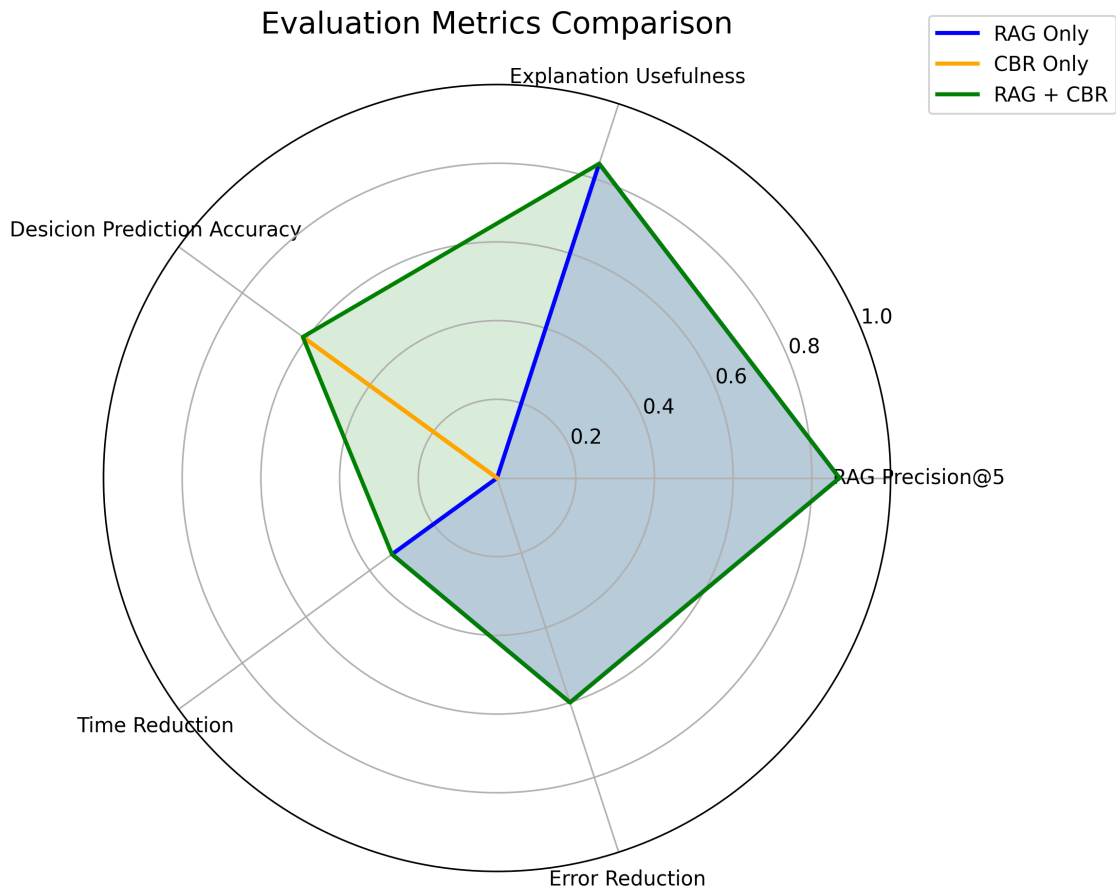


Figura 3. Comparação de desempenho normalizado entre métricas de RAG e CBR.

4.4. Limitações

- Atualmente, o conjunto de dados compreende 250 casos de solicitações CAOL com representações de características homogêneas (similares)
- Os resultados são preliminares; os testes estatísticos ainda estão em andamento
- Vieses históricos podem se propagar através do CBR, a menos que sejam explicitamente mitigados
- A versão atual não inclui comparação com baselines simplificados, como LLM sem RAG, recuperação documental convencional ou modelos sem raciocínio baseado em casos; tais comparações constituem trabalho futuro
- A implementação atual do CBR cobre predominantemente recuperação e retenção, enquanto adaptação e revisão permanecem parcialmente implementadas

5. Conclusão e Trabalhos Futuros

Este artigo apresentou uma arquitetura integrada para assistência inteligente em fluxos de trabalho CAOL dentro da PNIP, combinando MCP, RAG e CBR.

Os resultados preliminares do protótipo demonstram viabilidade técnica e indicam potencial para apoiar a conformidade regulatória e reduzir carga operacional no processo CAOL. A arquitetura busca fornecer suporte auditável e fundamentado para tomada de decisão humana em ambientes regulatórios digitais.

O trabalho futuro inclui validar o sistema em condições operacionais reais, expandir o suporte do MCP para ferramentas adicionais, refinar o CBR como um mecanismo assistivo e melhorar a representação de casos e as medidas de similaridade. O aprendizado contínuo sob governança apropriada será essencial para avaliar a escalabilidade e a confiança do usuário.

Agradecimentos

Esta pesquisa foi financiada pelo Ministerio da Pesca e Aquicultura, projetos UFSM/063089 e UFSM/060642. A pesquisa realizada por Vinícius Maran é suportada pelo projeto CNPq 306356/2020-1 DT-2.

Referências

- Aamodt, A. and Plaza, E. (1994). Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI Communications*, 7(1):39–59.
- Anthropic (2024). Model context protocol: Specification and design. *Technical Report*.
- Arner, D. e. a. (2020). Regtech and the digital transformation of compliance. *Journal of Financial Regulation*.
- Bommasani, R. e. a. (2021). On the opportunities and risks of foundation models. *Stanford Center for Research on Foundation Models*.
- Brown, T. B. e. a. (2020). Language models are few-shot learners. *NeurIPS*.
- Casanova, G. V., Montero, P. B., Machado, A., and Maran, V. (2025). Towards a rag-based whatsapp chatbot for animal certification platform support. In *Proceedings of the 27th International Conference on Enterprise Information Systems - Volume 1: ICEIS*, pages 1077–1084. SciTePress.
- Chen, Y. e. a. (2023). Retrieval-augmented generation: A survey. *IEEE Transactions on Knowledge and Data Engineering*.
- Kolodner, J. (1992). An introduction to case-based reasoning. *Artificial Intelligence Review*, 6:3–34.
- Lewis, P. e. a. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Ministério da Pesca e Aquicultura (2024). Portaria mpa nº 361/2024: Institui a plataforma nacional da indústria do pescado. *Diário Oficial da União*.
- Vaswani, A. e. a. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Watson, I. (1997). Applying case-based reasoning: Techniques for enterprise systems. *Morgan Kaufmann*.