Instituto Politécnico de Gestão e Tecnologia - GAIA

Mestrado em Engenharia de Tecnologias e Sistemas Web

Despacho n.º 9371/2020

Diário da República, 2.ª série, n.º 191, de 30 de setembro de 2020

Relatório de Projeto final de curso

Inteligência Documental Fundamentada em Evidência

Arquitetura Monitorizável para Análise e Validação de Documentos Digitais

Hugo da Silva Nascimento

Relatório de Projeto submetido para satisfação parcial dos requisitos do grau de Mestre em Engenharia de Tecnologias e Sistemas Web sob a orientação do Prof. Doutor Jorge Duque

Vila Nova de Gaia

2026

INSTITUTO POLITÉCNICO DE GESTÃO E TECNOLOGIA

Inteligência Documental Fundamentada em Evidência
Arquitetura Monitorizável para Análise e Validação de Documentos Digitais

Hugo da Silva Nascimento

Aprovado em 25/09/2026

Composição do Júri

_____

Presidente: Prof. Especialista José Moreira

_____

Arguente: Prof. Doutor Pedro Strecht

_____

Orientador: Prof. Doutor Jorge Pereira Duque

Vila Nova de Gaia
2026

Dissertação realizada sob a orientação do Prof. Doutor Jorge Pereira Duque e apresentada ao ISLA – Instituto Politécnico de Gestão e Tecnologia de Vila Nova de Gaia, para obtenção do grau de Mestre em Engenharia de Tecnologias e Sistemas Web, conforme o Despacho nº 9371/2020.

AGRADECIMENTOS

A concretização deste relatório não teria sido possível sem o contributo de várias pessoas, a quem desejo expressar o meu sincero reconhecimento.

Em primeiro lugar, agradeço ao meu orientador, Prof. Doutor Jorge Duque, pela disponibilidade permanente, pelo rigor das suas revisões e pela orientação criteriosa que, em cada fase do trabalho, ajudou a transformar ideias em resultados. As suas observações foram determinantes para a qualidade e a coerência do documento final.

Agradeço igualmente ao ISLA - Instituto Politécnico de Gestão e Tecnologia de Vila Nova de Gaia e ao corpo docente do Mestrado em Engenharia de Tecnologias e Sistemas Web, pelos conhecimentos transmitidos ao longo do curso e pelo ambiente de exigência e apoio que proporcionaram.

Por fim, um agradecimento muito especial à minha família e amigos, pelo apoio incondicional, pela paciência nos momentos de maior dedicação ao trabalho e pelo incentivo constante para levar este projeto até ao fim.

RESUMO

A crescente digitalização dos processos organizacionais intensificou a produção e circulação de documentos digitais não estruturados, cuja extração, interpretação e validação colocam desafios relevantes aos sistemas de informação. Neste contexto, o presente projeto desenvolveu e avaliou uma arquitetura de inteligência documental baseada na integração de Reconhecimento Ótico de Caracteres (OCR), Geração Aumentada por Recuperação (RAG) e Modelos de Linguagem de Grande Escala (LLMs), orientada para a análise e validação de documentos digitais em contexto organizacional. O trabalho adotou uma abordagem de investigação construtiva, materializada num protótipo modular. A solução desenvolvida suporta a ingestão de documentos digitais, a extração e estruturação de conteúdo, a indexação semântica, a recuperação contextual de informação relevante e a geração de respostas fundamentadas em evidência documental. Para reforçar a rastreabilidade e a auditabilidade do processo, a arquitetura integrou ainda uma camada de monitorização documental, inspirada no conceito de gémeo digital, que acompanha o estado do documento ao longo do pipeline e regista eventos e métricas relevantes. A avaliação foi realizada sobre 150 documentos sintéticos e 750 perguntas, incidindo sobre a qualidade da extração textual, a recuperação de evidência e a geração de respostas com e sem contexto documental, recorrendo a métricas quantitativas para a extração e recuperação, complementadas por uma análise qualitativa comparativa das respostas geradas. Os resultados mostram que a extração foi estável na maioria dos documentos pesquisáveis, que a recuperação privilegiou a cobertura em detrimento da seletividade e que, face a um modelo sem acesso ao documento, a ancoragem em evidência tornou as respostas específicas e verificáveis; evidenciam também limitações na recuperação e na avaliação da geração, que se limitou a uma análise qualitativa. O projeto apresenta como contributo principal uma arquitetura de referência para processamento documental inteligente, identificando ainda limitações e oportunidades de melhoria futura.
Palavras-chave: reconhecimento ótico de caracteres; geração aumentada por recuperação*; modelos de linguagem de grande escala; explicabilidade; gémeo digital documental; proteção de dados.*

ABSTRACT

The increasing digitalisation of organisational processes has intensified the production and circulation of unstructured digital documents, whose extraction, interpretation and validation pose relevant challenges for information systems. In this context, this project developed and evaluated a document intelligence architecture based on the integration of Optical Character Recognition (OCR), Retrieval-Augmented Generation (RAG) and Large Language Models (LLMs), aimed at supporting the analysis and validation of digital documents in organisational settings. The study adopted a constructive research approach, materialised in a modular prototype. The developed solution supports digital document ingestion, content extraction and structuring, semantic indexing, contextual retrieval of relevant information and the generation of responses grounded in documentary evidence. To strengthen process traceability and auditability, the architecture also integrated a document monitoring layer, inspired by the concept of a digital twin, which tracks the state of the document throughout the pipeline and records relevant events and metrics. The evaluation was conducted on 150 synthetic documents and 750 questions, covering textual extraction quality, evidence retrieval and response generation with and without documentary context, using quantitative metrics for extraction and retrieval, complemented by a qualitative comparative analysis of the generated responses. The results show that extraction was stable for most searchable documents, that retrieval favoured coverage over selectivity and that, compared with a model without access to the document, evidence grounding made the answers specific and verifiable; they also reveal limitations in retrieval and in the evaluation of generation, which remained qualitative. The main contribution of the project is a reference architecture for intelligent document processing, while also identifying limitations and opportunities for future improvement.
Keywords: optical character recognition; retrieval-augmented generation; large language models; explainable artificial intelligence; document digital twin; data protection.

ÍNDICE

RESUMO ..... **I**

ABSTRACT ..... **I**

ÍNDICE DE FIGURAS ..... **IV**

ÍNDICE DE TABELAS ..... **V**

LISTA DE ABREVIATURAS ..... **VI**

1. INTRODUÇÃO **1**

1.1. CONTEXTUALIZAÇÃO E MOTIVAÇÃO 1

1.2. PROBLEMA E QUESTÕES DE INVESTIGAÇÃO 3

1.3. OBJETIVOS 4

1.4. RELEVÂNCIA E CONTRIBUTOS 5

1.5. ESTRUTURA DO DOCUMENTO 6

2. REVISÃO DA LITERATURA **7**

2.1. ESTRATÉGIA DE REVISÃO DA LITERATURA 7

2.2. PROCESSAMENTO INTELIGENTE DE DOCUMENTOS E OCR 9

2.3. MODELOS DE LINGUAGEM DE GRANDE ESCALA NA ANÁLISE DOCUMENTAL 10

2.4. GERAÇÃO AUMENTADA POR RECUPERAÇÃO 10

2.5. EXPLICABILIDADE, RASTREABILIDADE E MONITORIZAÇÃO 11

2.6. CONFORMIDADE REGULATÓRIA E PROTEÇÃO DE DADOS 13

2.7. SÍNTESE CRÍTICA E IMPLICAÇÕES PARA O ESTUDO 13

3. METODOLOGIA **15**

3.1. ENQUADRAMENTO METODOLÓGICO: DESIGN SCIENCE RESEARCH 15

3.2. APLICAÇÃO DAS GUIDELINES DE DESIGN SCIENCE AO ESTUDO 16

3.3. MODELO DE PROCESSO DSR ADOTADO 17

3.4. DESENHO DO ARTEFACTO 18

3.5. CAMADA DE MONITORIZAÇÃO E RASTREABILIDADE 20

3.6. FASES DO ESTUDO 21

3.7. Corpus Documental e Unidade de Análise 22
3.8. Protocolo de Avaliação 24
3.9. Métricas de Avaliação 27
3.10. Procedimentos Éticos e Conformidade 32
4. DESENVOLVIMENTO E IMPLEMENTAÇÃO DA SOLUÇÃO 34
4.1. Enquadramento do Desenvolvimento do Artefacto 34
4.2. Arquitetura do Sistema 36
4.3. Tecnologias e Ferramentas 42
4.4. Implementação dos Módulos 45
4.5. Configuração e Reprodutibilidade 47
4.6. Logging, Rastreabilidade e Auditabilidade 49
5. RESULTADOS 51
5.1. Procedimento de Obtenção dos Resultados 51
5.2. Avaliação da Extração Documental 53
5.3. Avaliação do Módulo de Recuperação (RAG) 54
5.4. Avaliação Qualitativa da Geração de Respostas 57
5.5. Discussão dos Resultados 59
6. CONCLUSÕES E TRABALHO FUTURO 62
6.1. Conclusões 62
6.2. Limitações 63
6.3. Trabalho Futuro 66
BIBLIOGRAFIA 68

ÍNDICE DE FIGURAS

Figura 1 - Diagrama de fluxo do processo de revisão sistemática (PRISMA 2020) 8

Figura 2 - Arquitetura do pipeline documental: etapas, entradas, saídas e camada transversal de monitorização 19

Figura 3 - Arquitetura e fluxo de processamento documental do protótipo 37

Figura 4 - Evolução média da Precision@K e da Recall@K por profundidade de recuperação (top-k) 56

ÍNDICE DE TABELAS

Tabela 1 - Mapa relacional: dependências entre conceitos e implicações para a arquitetura 14

Tabela 2 - Aplicação das guidelines de Design Science ao presente trabalho 16

Tabela 3 - Operacionalização das fases DSR no projeto 17

Tabela 4 - Fases do estudo e correspondência com a DSR 21

Tabela 5 - Métricas e critérios de avaliação do artefacto 27

Tabela 6 - Dimensões do framework interno de qualidade 28

Tabela 7 - Estratégia de processamento e fallback do OCR 29

Tabela 8 - Métricas de avaliação da recuperação 31

Tabela 9 - Princípios de desenho arquitetural da solução 38

Tabela 10 - Componentes funcionais da arquitetura 39

Tabela 11 - Tecnologias e ferramentas utilizadas no desenvolvimento do protótipo 42

Tabela 12 - Síntese dos componentes tecnológicos do protótipo e respetivas funções 47

Tabela 13 - Organização dos parâmetros de configuração do protótipo 48

Tabela 14 - Configuração final do pipeline utilizada na avaliação experimental 49

Tabela 15 - CER por características estruturais dos documentos 53

Tabela 16 - Desempenho da recuperação por tipo de pergunta e top-k 55

LISTA DE ABREVIATURAS

API: Interface de Programação de Aplicações (Application Programming Interface)

CER: Taxa de Erro de Caracteres (Character Error Rate)

DSR: Investigação em Ciências do Desenho (Design Science Research)

IA: Inteligência Artificial

IDP: Processamento Inteligente de Documentos (Intelligent Document Processing)

LLM: Modelo de Linguagem de Grande Escala (Large Language Model)

MTEB: Referência de Avaliação de Embeddings Textuais (Massive Text Embedding Benchmark)

OCR: Reconhecimento Ótico de Caracteres (Optical Character Recognition)

PDF: Formato de Documento Portátil (Portable Document Format)

PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

RAG: Geração Aumentada por Recuperação (Retrieval-Augmented Generation)

RF: Requisitos Funcionais

RGPD: Regulamento Geral sobre a Proteção de Dados

RNF: Requisitos Não Funcionais

VRAM: Memória de Vídeo de Acesso Aleatório (Video Random Access Memory)

XAI: Inteligência Artificial Explicável (Explainable Artificial Intelligence)

1. INTRODUÇÃO

1.1. Contextualização e Motivação

A crescente digitalização de processos organizacionais intensificou a produção e circulação de documentos em formatos heterogéneos, frequentemente não estruturados, mantendo elevada a complexidade da extração e interpretação automática de informação operacional e juridicamente relevante. No âmbito do Processamento Inteligente de Documentos (IDP), este desafio é reconhecido como um problema de *document understanding*, no qual técnicas de inteligência artificial (IA) e visão por computador procuram transformar documentos digitais em representações utilizáveis por sistemas de análise e apoio à decisão (Loukil et al., 2024). A forte dependência de documentação administrativa, jurídica e regulatória, associada à persistência de fluxos de trabalho intensivos em intervenção humana, tende a traduzir-se em atrasos, estrangulamentos operacionais e ocorrência de erros sistemáticos, reforçando a necessidade de soluções automatizadas robustas (Ramesh Pingili, 2025). A consolidação do PDF como formato predominante nas organizações contribui adicionalmente para esta dificuldade, dado que muitos conteúdos resultam de digitalizações ou incluem elementos não editáveis, exigindo abordagens combinadas de extração baseada na estrutura e de análise do texto resultante (Baviskar et al., 2021).

Neste enquadramento, o Reconhecimento Ótico de Caracteres (OCR) assumiu um papel nuclear ao converter conteúdo textual incorporado em imagens ou em PDFs digitalizados em texto processável, viabilizando a integração destes documentos em *pipelines* de análise e extração de informação (Baviskar et al., 2021). Contudo, quando os documentos apresentam layouts complexos e elevada variabilidade estrutural, a robustez do pipeline depende da capacidade de preservar a estrutura e relações espaciais, sendo o ruído introduzido nas etapas iniciais de reconhecimento, um fator com impacto direto no desempenho subsequente (J. Zhang et al., 2025).

A motivação central do trabalho residiu na necessidade de desenvolver uma arquitetura de análise documental capaz de produzir resultados rastreáveis e tecnicamente justificáveis, reduzindo o risco de respostas não suportadas por evidência recuperada e facilitando a auditoria em cenários organizacionais. Esta exigência de transparência é coerente com as limitações identificadas para a explicabilidade em pipelines multi-etapa, onde se destacam dificuldades de reconstrução da articulação entre componentes e compromissos entre custo,

latência e detalhe explicativo (Habib et al., 2025), bem como a necessidade de combinar diferentes famílias de métodos explicativos em função de objetivos como justificação, depuração e controlo do sistema (Adadi & Berrada, 2018). Neste ponto, o conceito de gémeo digital revelou-se particularmente pertinente, não como tecnologia central autónoma, mas como referencial para o documento no seu percurso de processamento enquanto entidade dinâmica, continuamente atualizada por dados e orientada para monitorização, validação e apoio à decisão, conforme proposta pela National Academies of Sciences et al. (2024). Esta leitura é reforçada por Birk et al. (2022), que destacam a monitorização e a validação contínuas como funções essenciais destas representações, sustentando a observabilidade, a rastreabilidade e o controlo do comportamento do sistema ao longo do ciclo de vida documental.

A adoção organizacional destes sistemas é inseparável das dimensões de conformidade e governação de dados, dado que o tratamento documental envolve frequentemente dados pessoais e informação proprietária. A literatura sublinha riscos específicos quando estes conteúdos são integrados em pipelines com recuperação, incluindo a exposição acidental e a exploração por *prompt injection* para extrair o contexto recuperado (Zhou et al., 2025). Para mitigar estes riscos, Panda & Mukherjee (2025) preconizam medidas de segurança e conformidade distribuídas ao longo de todo o pipeline.

Deste modo, a presente investigação posicionou-se na interseção entre IDP, geração aumentada por recuperação (RAG), explicabilidade e governação documental. A lacuna identificada residiu na ausência de abordagens arquiteturais que articulem, de forma integrada, OCR, recuperação semântica, geração condicionada por evidência e mecanismos de observabilidade orientados ao ciclo de vida do documento em contexto organizacional. Esta lacuna revelou-se relevante porque a análise documental automatizada não depende apenas da qualidade da resposta final, mas também da possibilidade de reconstruir o percurso que liga o documento original, o texto extraído, os excertos recuperados, a resposta gerada e os sinais de validação associados. O Projeto respondeu a esta lacuna através da conceção, implementação e avaliação de uma arquitetura modular que não se limitou à produção de respostas automatizadas, mas tornou observáveis os processos de extração, recuperação, geração e validação que sustentam essas respostas.

1.2. Problema e questões de investigação

O problema central do projeto consiste em determinar como conceber, implementar e avaliar uma arquitetura de inteligência documental que integre OCR, RAG e Modelos de Linguagem de Grande Escala (LLMs), de modo a suportar a extração, análise e validação de informação em documentos digitais não estruturados, conciliando desempenho funcional, explicabilidade, rastreabilidade e conformidade regulatória em contexto organizacional, e incorporando mecanismos de monitorização do ciclo de vida documental inspirados no conceito de gémeo digital documental.

Deste problema decorrem as seguintes questões de investigação secundárias:

1) Como extrair e estruturar informação a partir de documentos com diferentes qualidades de digitalização, layouts complexos e elementos visuais ou tabulares relevantes?
2) De que forma a estratégia de segmentação e a profundidade de recuperação condicionam a recuperação de evidência?
3) De que modo a arquitetura RAG contribui para reduzir respostas não fundamentadas e assegurar a ligação entre as respostas e as fontes documentais recuperadas?
4) De que modo uma camada inspirada no conceito de gémeo digital documental reforça a rastreabilidade, a observabilidade e o controlo do comportamento do sistema ao longo do ciclo de vida do documento?
5) De que forma a arquitetura incorpora princípios de proteção de dados, minimização, limitação da finalidade e auditabilidade, em conformidade com o Regulamento Geral sobre a Proteção de Dados (RGPD)?

1.3. Objetivos

1.3.1. Objetivo Geral

Desenvolver e avaliar uma arquitetura integrada de inteligência documental, baseada em OCR, RAG e LLMs, destinada a apoiar a análise e validação de documentos digitais em contexto organizacional, reforçando a explicabilidade, a rastreabilidade, a conformidade regulatória e a monitorização do ciclo de vida documental.

1.3.2. Objetivos Específicos

1) Avaliar e selecionar abordagens de OCR adequadas à extração e estruturação de informação em documentos com diferentes níveis de qualidade, complexidade estrutural e grau de digitalização
2) Desenvolver estratégias de segmentação documental orientadas para melhorar a recuperação de evidência relevante em tarefas de análise e validação documental.
3) Implementar um pipeline RAG, suportado por indexação vetorial, para recuperação contextual de excertos documentais pertinentes e geração de respostas fundamentadas em evidência.
4) Integrar uma camada de monitorização inspirada no conceito de gémeo digital documental, destinada a acompanhar o estado do documento ao longo do pipeline, registar eventos relevantes e reforçar a observabilidade e o controlo do sistema.
5) Definir e aplicar métricas quantitativas e qualitativas para avaliar o desempenho do sistema, incluindo qualidade da extração, eficácia da recuperação, qualidade das respostas geradas, grau de fundamentação documental e sinais operacionais de confiança.
6) Analisar em que medida a arquitetura desenvolvida satisfaz requisitos de explicabilidade, fiabilidade e conformidade com princípios de proteção de dados aplicáveis ao tratamento documental em contexto organizacional.

1.4. Relevância e Contributos

A área do Processamento Inteligente de Documentos afirmou-se como um espaço de convergência entre extração documental, recuperação de informação, modelos de linguagem e requisitos de explicabilidade, rastreabilidade, monitorização e conformidade regulatória. A relevância deste estudo decorreu não apenas do interesse científico associado à integração de OCR, RAG e LLMs, mas também da necessidade prática de desenvolver soluções capazes de apoiar a análise e validação de documentação não estruturada em contextos organizacionais, onde estas tarefas assumem frequentemente carácter crítico para a operação, o controlo e a tomada de decisão.

Os contributos do estudo organizaram-se em três planos complementares.

- No **plano teórico**, o estudo contribuiu para a clarificação conceptual da articulação entre OCR, RAG e LLMs em cenários de análise documental, com particular atenção às dimensões de fundamentação em evidência, explicabilidade, rastreabilidade e conformidade. Neste âmbito, integrou-se ainda o conceito de gémeo digital documental como referencial de monitorização, observabilidade e controlo do ciclo de vida do documento, não como tecnologia central autónoma, mas como camada de suporte à auditabilidade e à interpretação técnica do comportamento do sistema.
- No **plano metodológico**, o estudo estruturou um desenho de avaliação que articula diferentes níveis do pipeline, incluindo qualidade da extração, eficácia da recuperação, grau de fundamentação da resposta, qualidade final do output gerado e utilidade dos mecanismos de monitorização introduzidos na arquitetura. Esta abordagem permitiu analisar comparativamente estratégias, configurações e componentes da solução desenvolvida, produzindo evidência útil para a avaliação de sistemas de inteligência documental baseados em recuperação de evidência.
- No **plano técnico**, foi desenvolvido um protótipo modular que integra ingestão documental, extração e estruturação de conteúdo, indexação semântica, recuperação contextual, geração de respostas fundamentadas e mecanismos adicionais de monitorização e validação. Esta arquitetura foi concebida como referência para a implementação de sistemas semelhantes, permitindo a adaptação a diferentes

cenários documentais e a evolução dos componentes tecnológicos com impacto limitado na estrutura global da solução.

Em termos aplicados, o projeto forneceu indicações relevantes sobre a incorporação de princípios de *privacy by design*, mecanismos de validação de respostas, estratégias de rastreabilidade documental e formas de observação do comportamento do sistema em arquiteturas de IA aplicadas ao tratamento de documentos. Deste modo, o estudo oferece uma base para o desenvolvimento de sistemas auditáveis e adequados às exigências operacionais e regulatórias de contextos organizacionais.

Importa distinguir o que é contributo do que permanece alegação. A arquitetura é apresentada como referência no sentido de modelo documentado e replicável, componentes, interfaces, parâmetros e protocolo de avaliação, e não como solução comprovadamente superior a alternativas: não foi comparada com outras arquiteturas nem validada em contexto organizacional real.

1.5. Estrutura do Documento

O documento está organizado em seis capítulos. O presente capítulo introduz o tema, o problema de investigação, os objetivos e a relevância do estudo. O Capítulo 2 apresenta a revisão da literatura sobre os conceitos e tecnologias fundamentais para a investigação, incluindo IDP, OCR, LLMs, RAG, explicabilidade, gémeos digitais e conformidade regulatória. O Capítulo 3 descreve a metodologia de investigação adotada, incluindo o desenho do estudo, o corpus documental, o protocolo experimental e as métricas de avaliação. O Capítulo 4 detalha o desenvolvimento e a implementação técnica do projeto, apresentando a arquitetura da solução, os módulos implementados e os mecanismos de rastreabilidade e monitorização. O Capítulo 5 apresenta e discute os resultados obtidos na avaliação do protótipo. Por fim, o Capítulo 6 sintetiza as conclusões do trabalho, discute as suas limitações e apresenta direções para investigação futura.

2. REVISÃO DA LITERATURA

Este capítulo apresenta uma revisão da literatura deliberadamente orientada para os componentes que sustentam o artefacto desenvolvido: Processamento Inteligente de Documentos, Reconhecimento Ótico de Caracteres, Modelos de Linguagem de Grande Escala, Geração Aumentada por Recuperação, explicabilidade, rastreabilidade, gémeo digital documental e conformidade regulatória. A revisão foi condensada para privilegiar a ligação direta entre literatura, desenho da arquitetura e avaliação experimental, evitando uma acumulação descritiva que não acrescentasse valor ao problema de investigação.

2.1. Estratégia de Revisão da Literatura

A revisão foi organizada segundo os princípios do PRISMA 2020 (Page et al., 2021), adotados como referencial de transparência para documentar a pesquisa, a triagem, a elegibilidade e a inclusão dos estudos. A pesquisa incidiu no período de 2020 a 2025, com consulta às bases Web of Science, Google Scholar e Semantic Scholar, complementada por pesquisa dirigida sobre gémeos digitais quando este conceito passou a ser mobilizado como referencial arquitetural para a monitorização documental. A seleção privilegiou estudos com ligação direta à integração entre OCR, RAG, LLMs, explicabilidade, rastreabilidade, validação documental e conformidade em contexto organizacional.

O processo de triagem permitiu excluir trabalhos sem relação direta com o domínio documental, estudos centrados apenas em OCR isolado, aplicações genéricas de LLMs sem recuperação de evidência e publicações sem descrição metodológica suficiente para apreciar os resultados. A revisão resultante não pretendeu ser enciclopédica, mas sim crítica e funcional: cada eixo teórico foi analisado em função da sua utilidade para justificar decisões de desenho, métricas de avaliação e limitações do protótipo. Foram reunidos 404 registos para triagem; após a remoção de 46 duplicados, a triagem de 358 e a avaliação de 65 em texto integral, resultaram 37 estudos considerados relevantes, como sintetiza a Figura 1.

Dada a evolução recente da área, em particular dos sistemas RAG, foram também considerados preprints, relatórios técnicos e uma dissertação, quando reportavam avaliação empírica diretamente relevante para decisões de desenho, por exemplo, estratégias de

segmentação ou configurações de recuperação. Estas fontes são tratadas como literatura cinzenta e usadas para fundamentar opções técnicas, e não como evidência consolidada. A bibliografia inclui apenas as obras efetivamente utilizadas, entre as quais também referências de ferramentas e de método.

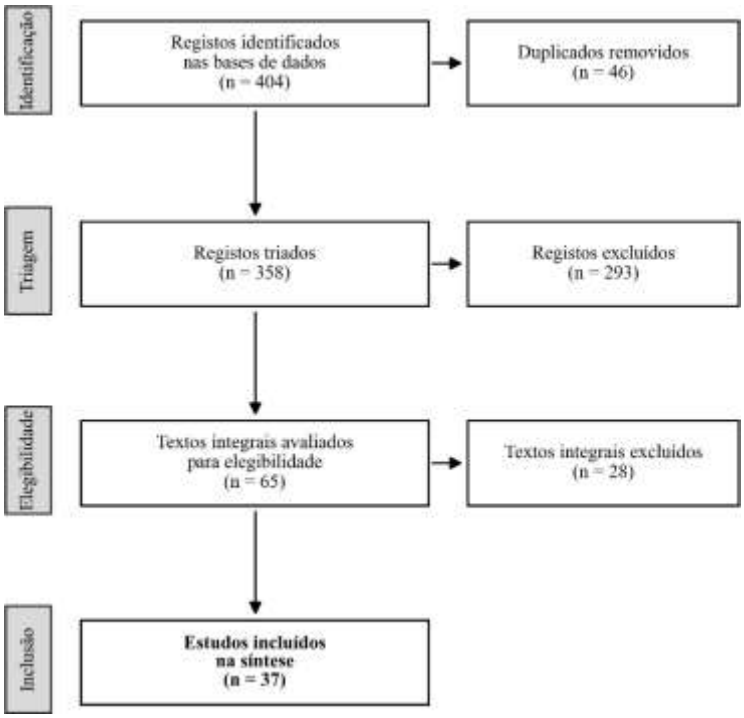

Figura 1 - Diagrama de fluxo do processo de revisão sistemática (PRISMA 2020)

2.2. Processamento Inteligente de Documentos e OCR

O Processamento Inteligente de Documentos procura transformar documentos digitais, frequentemente não estruturados ou semiestruturados, em representações úteis para pesquisa, interpretação, validação e apoio à decisão (Loukil et al., 2024). Como sublinham Baviskar et al. (2021), a ausência de esquemas formais dificulta o tratamento automático desta informação, o que é particularmente crítico em organizações cuja documentação administrativa, jurídica, financeira ou técnica reside sobretudo em PDFs, digitalizações, formulários, tabelas ou relatórios de estrutura visual complexa. Em contextos de administração pública, Ramesh Pingili (2025) associa este predomínio de documentação não estruturada a fluxos intensivos em intervenção humana, com estrangulamentos operacionais e maior suscetibilidade ao erro.

Neste contexto, o OCR funciona como porta de entrada para documentos sem texto diretamente processável, convertendo imagem em texto e viabilizando indexação, recuperação e análise. O reconhecimento é, porém, também uma fonte crítica de erro: mesmo motores maduros como o Tesseract apresentam taxas de exatidão na ordem dos 90-95% (Baviskar et al., 2021), deixando uma fração relevante de carateres incertos. Esta dificuldade é particularmente acentuada na extração de informação a partir de diagramas e componentes gráficos, como salientam Sobhan & Haque (2025). Num plano complementar, J. Zhang et al. (2025) demonstram que o ruído do OCR, semântico e de formatação, se propaga à recuperação e à geração, degradando as respostas. Para o mitigar, Guan et al. (2024) propõem estratégias de correção pós-OCR apoiadas em dados sintéticos, com reduções assinaláveis da taxa de erro.

Daqui resulta uma implicação central para este trabalho: a qualidade do sistema não pode ser atribuída apenas ao modelo generativo, começando antes na extração e na preservação da estrutura. Por esse motivo, a arquitetura proposta assenta o processamento documental no Docling (Auer et al., 2024) e incorpora mecanismos de *fallback*, reparação pós-OCR, separação entre texto bruto e texto indexado e métricas como a Taxa de Erro de Caracteres (CER), que permitem observar empiricamente como a extração condiciona as etapas seguintes.

2.3. Modelos de Linguagem de Grande Escala na Análise Documental

Os Modelos de Linguagem de Grande Escala tornaram-se relevantes para tarefas documentais pela capacidade de interpretar perguntas em linguagem natural, sintetizar conteúdos e produzir respostas contextualizadas. Segundo Mundlamuri et al. (2025), esta trajetória decorre da introdução da arquitetura Transformer e do mecanismo de atenção, que criaram as condições para os modelos pré-treinados de grande escala. No domínio documental, assiste-se a um deslocamento das abordagens discriminativas para formulações generativas (Xu et al., 2024), nas quais o modelo produz diretamente saídas estruturadas a partir de instruções, aproximando a extração das necessidades reais de análise.

Apesar desta utilidade, os LLMs apresentam limitações importantes em contextos críticos, com destaque para a ocorrência de alucinações, respostas fluentes, mas factualmente incorretas (Mokashi et al., 2025). Como observa Paoletti (2025), em tarefas de validação documental esta fragilidade é particularmente sensível, uma vez que a resposta só é útil se estiver ancorada no documento efetivamente processado. Por esse motivo, o trabalho não trata o LLM como fonte autónoma de conhecimento, mas como componente de geração condicionado por evidência recuperada.

Esta leitura justifica a passagem de uma abordagem centrada no modelo para uma abordagem centrada na arquitetura. A função do LLM é transformar evidência recuperada numa resposta clara, mas a confiança da resposta depende da qualidade da extração, da segmentação, da indexação, da recuperação, do *re-ranking* e da validação de *grounding*. Esta opção aparece no Capítulo 4 e é testada no Capítulo 5 através da comparação entre respostas com e sem RAG.

2.4. Geração Aumentada por Recuperação

A Geração Aumentada por Recuperação, proposta por Lewis et al. (2021), liga os modelos generativos a fontes documentais externas: em vez de depender apenas do conhecimento paramétrico do modelo, o sistema recupera excertos relevantes do corpus e fornece-os ao LLM como contexto de geração. Esta estratégia aumenta a atualidade, a verificabilidade e a adequação da resposta ao documento analisado (Gao et al., 2024).

Numa arquitetura RAG típica, o corpus é ingerido, segmentado em *chunks*, convertido em *embeddings*, armazenado numa base vetorial e pesquisado por similaridade semântica (Gao et al., 2024). A definição da granularidade da segmentação é, segundo Smith & Troynikov (2024), determinante para a qualidade da recuperação, exigindo um equilíbrio entre continuidade semântica e redução de ruído. Em configurações mais robustas, a recuperação é complementada por reformulação da pergunta, filtros por metadados, re-ranking e compressão de contexto, uma lógica cuja aplicação a PDFs é reportada por Khan et al. (2024) e que o protótipo adota com a ChromaDB, embeddings locais, *query rewriting* e re-ranking.

A eficácia do RAG não é, contudo, automática. Kuru & Keskin (2025) mostram que o desempenho varia consideravelmente entre diferentes configurações de recuperação. Também a degradação do texto pelo OCR compromete a evidência recuperada, como evidenciam J. Zhang et al. (2025). Se os chunks forem mal definidos, se a evidência ficar fragmentada ou se a recuperação devolver excertos pouco pertinentes, a resposta final permanece vulnerável. Por isso, a avaliação do projeto separa recuperação e geração: primeiro mede-se a capacidade de localizar evidência; depois avalia-se a qualidade da resposta produzida com essa evidência.

2.5. Explicabilidade, Rastreabilidade e Monitorização

Em sistemas de IA aplicados a documentos organizacionais, a explicabilidade não se limita a apresentar a resposta final. A explicação útil deve indicar não apenas que uma fonte foi usada, mas por que razão determinados excertos foram recuperados, em que ordem foram priorizados e de que modo sustentam a resposta (Habib et al., 2025). Assim, Cohen et al. (2025) propõem mecanismos que articulam a avaliação explicável com orientações acionáveis sobre o comportamento do pipeline de recuperação e geração.

Importa delimitar o sentido em que o termo é usado. Neste projeto, a explicabilidade é entendida ao nível do sistema: a camada transversal de monitorização, inspirada no gémeo digital, permite explicar o que acontece a cada documento em todos os módulos do pipeline, que motor extraiu o texto, que reparações foram aplicadas, que excertos foram recuperados

e entregues ao modelo e que verificações foram feitas e, a partir daí, compreender por que razão o sistema produziu uma determinada resposta. Esta aceção difere do sentido estrito da inteligência artificial explicável, em que explicar significa esclarecer por que razão um modelo produziu uma determinada saída, através da atribuição da decisão às entradas, com preocupações de fidelidade da explicação e com a distinção entre explicações intrínsecas e *post-hoc* (Adadi & Berrada, 2018). O trabalho não procura explicar o funcionamento interno do modelo de linguagem, e a atribuição das suas respostas fica fora do âmbito do estudo.

Esta exigência conduz à noção de rastreabilidade operacional. Num sistema documental, rastrear significa manter ligação explícita entre o ficheiro original, o texto extraído, os chunks indexados, os excertos recuperados, o *prompt* construído, a resposta gerada e o resultado da validação. Esta preocupação está diretamente refletida na arquitetura desenvolvida, que integra *logging* e uma camada transversal de monitorização documental.

O conceito de gémeo digital foi mobilizado neste trabalho em sentido arquitetural, e não como réplica física bidirecional completa. O seu valor decorre da articulação entre representação virtual, atualização contínua e utilização operacional dessa representação (National Academies of Sciences et al., 2024). Iliuţă et al. (2024) documentam o alargamento do conceito a domínios para além da indústria. Birk et al. (2022) destacam a geração, calibração, monitorização e validação do modelo como funções essenciais. Transposta para o domínio documental, e na linha explorada por N. Zhang et al. (2024) para decisões explicáveis em gémeos digitais dinâmicos, esta noção permite representar cada documento como entidade observável ao longo do pipeline, registando estados, eventos, métricas de confiança e decisões adaptativas, e reforçando a auditabilidade do sistema.

Importa distinguir esta noção de um registo de proveniência ou de linhagem de dados. A proveniência regista, de forma retrospetiva, a origem de cada artefacto e as atividades que o produziram. A camada aqui proposta inclui essa função, mas acrescenta dois elementos: um estado corrente de cada documento, atualizado ao longo do pipeline, e a utilização desse estado para desencadear decisões sobre o próprio pipeline.

2.6. Conformidade Regulatória e Proteção de Dados

O tratamento automatizado de documentos organizacionais envolve frequentemente dados pessoais, informação contratual ou conteúdos reservados. Segundo Panda & Mukherjee (2025), a conformidade com o RGPD (Regulamento (UE) 2016/679, 2016) exige que o sistema incorpore, por desenho, princípios de minimização, limitação da finalidade, segurança, controlo de acessos, auditabilidade e responsabilização, que não podem ser acrescentados apenas no final da implementação.

Em sistemas RAG, a conformidade coloca desafios específicos, uma vez que informação sensível pode ser indexada, recuperada e introduzida no contexto do modelo, gerando riscos de fuga do contexto recuperado e de prompt injection (Zhou et al., 2025). Em resposta, Panda & Mukherjee (2025) propõem salvaguardas como a encriptação de chunks e embeddings, o controlo de acessos e o isolamento por utilizador. Coerentemente, a arquitetura privilegia a execução local dos modelos, o registo auditável e a separação entre o armazenamento vetorial e a memória operacional.

2.7. Síntese Crítica e Implicações para o Estudo

A revisão da literatura permite concluir que a inteligência documental depende da articulação entre componentes. OCR, LLMs, RAG, inteligência artificial explicável (XAI), monitorização e conformidade são frequentemente estudados de forma separada, mas a sua adoção em organizações exige uma arquitetura integrada. A lacuna que sustenta esta investigação reside precisamente na necessidade de ligar extração, recuperação, geração, validação e rastreabilidade num artefacto avaliável de ponta a ponta. A Tabela 1 sintetiza esta leitura crítica, as relações entre os conceitos, de que forma cada eixo condiciona ou suporta os restantes, e as respetivas implicações para a arquitetura.

Estas implicações justificam o desenho metodológico apresentado no capítulo seguinte. O Capítulo 3 transforma a revisão em protocolo de investigação; o Capítulo 4 materializa a arquitetura; e o Capítulo 5 avalia se as opções tomadas permitem responder ao problema formulado.

Tabela 1 - Mapa relacional: dependências entre conceitos e implicações para a arquitetura

Eixo da literatura	*Conclusão crítica*	**Relação com os restantes eixos**	*Implicação no protótipo*
OCR E Preparação Documental	Erros de reconhecimento e de estrutura propagam-se à recuperação e à geração (Baviskar et al., 2021; J. Zhang et al., 2025).	Condiciona o RAG: a qualidade do texto e da estrutura determina o conteúdo dos chunks e a evidência recuperável.	Docling, fallback, reparação pós-OCR e medição da divergência textual por CER nos PDFs pesquisáveis.
LLMS	A fluidez dos LLMs não garante factualidade nem aderência ao documento (Gao et al., 2024; Mokashi et al., 2025).	Na análise documental, o RAG fornece evidência externa e a rastreabilidade liga a resposta às fontes, sem explicar o funcionamento interno do modelo.	Geração condicionada por excertos, comparação com uma condição sem RAG e validação de grounding.
RAG	O RAG pode aumentar a factualidade e a verificabilidade, mas depende da segmentação e da qualidade da recuperação (Lewis et al., 2021; Gao et al., 2024).	Liga a preparação documental aos LLMs: o conteúdo extraído é segmentado e indexado, e os excertos recuperados condicionam a resposta.	Indexação vetorial, query rewriting, re-ranking e avaliação por Precision, Recall e F1.
Explicabilidade Do Sistema E Rastreabilidade	A rastreabilidade reconstrói etapas e evidências do pipeline, mas não equivale à explicação interna do modelo no sentido estrito de XAI (Adadi & Berrada, 2018; Cohen et al., 2025).	Apoia-se nos registos de monitorização e nos excertos recuperados, contribuindo para a auditabilidade sem alegar uma explicação causal do LLM.	Logging, metadados, excertos de suporte e camada de monitorização.
Camada Inspirada Em Gémeo Digital	Os gémeos digitais articulam representações atualizadas, monitorização, validação e apoio à decisão (National Academies of Sciences et al., 2024; Birk et al., 2022).	No protótipo, esta inspiração materializa-se numa camada transversal que observa OCR, RAG e LLMs e regista decisões, sem constituir uma réplica física bidirecional.	Estados, eventos, métricas de confiança e decisões adaptativas.
RGPD E Segurança	O tratamento de dados pessoais exige limitação da finalidade, minimização, segurança e responsabilização (Regulamento (UE) 2016/679, 2016; Zhou et al., 2025).	Impõe requisitos transversais à ingestão, indexação, recuperação, geração e logging, com controlos proporcionais ao risco.	Execução local via Ollama, filtragem por documento, retenção limitada e registo auditável como opções de privacy by design.

3. METODOLOGIA

Este capítulo apresenta a metodologia adotada para responder ao problema de investigação e aos objetivos definidos no Capítulo 1. Considerando que o projeto se centrou na conceção, implementação e avaliação de uma arquitetura tecnológica para processamento inteligente de documentos, a investigação foi estruturada segundo uma abordagem de Design Science Research (DSR). Esta opção metodológica revelou-se adequada por permitir articular a identificação de um problema relevante em contexto organizacional com a construção, demonstração e avaliação de um artefacto tecnológico orientado para a sua resolução.

A metodologia foi organizada de modo a distinguir três planos complementares: o desenho da investigação, o desenho do artefacto e o protocolo de avaliação. Esta separação permitiu evitar a sobreposição entre fundamentação metodológica, descrição técnica da solução e tarefas operacionais de desenvolvimento, assegurando maior clareza científica e maior rastreabilidade entre problema, objetivos, artefacto e avaliação.

3.1. Enquadramento Metodológico: Design Science Research

A presente investigação adotou o DSR como quadro metodológico principal, por se centrar na conceção, desenvolvimento e avaliação de um artefacto tecnológico orientado para a resolução de um problema prático. Segundo Hevner et al. (2004), esta abordagem é adequada quando o objetivo da investigação não se limita à descrição ou explicação de um fenómeno, mas envolve a criação de uma solução avaliada quanto à sua utilidade, robustez e pertinência científica. Peffers et al. (2007) propõem um processo estruturado que conduz a investigação desde a identificação do problema até à avaliação e comunicação do artefacto, assegurando a sua aplicabilidade em contexto real.

Neste estudo, o problema identificado relacionou-se com a dificuldade de processar, analisar e validar grandes volumes de documentos digitais não estruturados ou semiestruturados em ambientes organizacionais, sobretudo quando estão em causa exigências simultâneas de precisão, rastreabilidade, explicabilidade e conformidade. A escolha da DSR decorreu, assim, da necessidade de responder a este problema através do

desenvolvimento de um protótipo funcional que integrasse diferentes componentes tecnológicos.

Desta forma, a abordagem construtiva não foi entendida apenas como um processo de implementação técnica, mas como um quadro metodológico que articulou problema, fundamentação teórica, construção do artefacto e validação empírica da sua utilidade.

3.2. Aplicação das Guidelines de Design Science ao Estudo

A abordagem construtiva baseia-se nos princípios estabelecidos por Hevner et al. (2004) para Design Science em sistemas de informação, que definem sete guidelines fundamentais aplicados neste trabalho (Tabela 2):

Tabela 2 - Aplicação das guidelines de Design Science ao presente trabalho

Guideline de Design Science	***Aplicação no presente trabalho***
Design como Artefacto	Desenvolvimento de um sistema concreto de processamento inteligente de documentos, baseado numa arquitetura de recuperação e geração, orientado para análise e validação documental.
Relevância do Problema	Resposta a necessidades reais de extração, interpretação e validação de informação em documentos complexos, com requisitos de precisão, rastreabilidade e conformidade em contexto organizacional.
Avaliação do Design	Validação do artefacto através de métricas quantitativas de extração textual nos PDFs pesquisáveis (CER) e de recuperação (Precision, Recall, F1, P@K e R@K sobre 750 perguntas), complementada por uma avaliação qualitativa exploratória de 10 respostas com e sem contexto recuperado. Esta avaliação estabelece viabilidade e coerência interna, mas não mede OCR nos digitalizados nem qualidade da geração em todo o corpus.
Contribuições de Investigação	Proposta de uma arquitetura integrada para processamento documental inteligente, articulando OCR, recuperação semântica, geração condicionada por evidência e mecanismos de controlo de qualidade.
Rigor da Investigação	Fundamentação nas áreas de processamento documental, recuperação de informação, geração aumentada por recuperação e explicabilidade, combinada com avaliação empírica sobre um conjunto documental estruturado.
Design como Processo de Pesquisa	Desenvolvimento iterativo do artefacto, com refinamento progressivo da solução em função da experimentação, dos resultados obtidos e dos requisitos identificados ao longo da investigação.
Comunicação da investigação	Descrição sistemática da metodologia, da arquitetura proposta, das decisões de desenho e dos resultados de avaliação, de modo a assegurar clareza, reprodutibilidade e utilidade científica.

3.3. Modelo de Processo DSR Adotado

O estudo seguiu o modelo de processo de Design Science Research proposto por Peffers et al. (2007), estruturado em seis fases: identificação do problema e motivação; definição dos objetivos da solução; desenho e desenvolvimento do artefacto; demonstração; avaliação; e comunicação. Estas fases foram adaptadas ao contexto do projeto, de modo a enquadrar a conceção e avaliação de uma arquitetura de inteligência documental aplicada à análise e validação de documentos digitais. A Tabela 3 apresenta a operacionalização destas fases no projeto.

Tabela 3 - Operacionalização das fases DSR no projeto

Fase DSR	*Aplicação no projeto*	*Resultado produzido*
Identificação do problema e motivação	Identificação das dificuldades de extração, análise e validação de documentos digitais heterogéneos em contexto organizacional.	Formulação do problema de investigação e das questões secundárias.
Definição dos objetivos da solução	Definição dos requisitos científicos e funcionais da arquitetura, incluindo OCR, recuperação de evidência, geração fundamentada, rastreabilidade, observabilidade e conformidade.	Objetivo geral, objetivos específicos e critérios de desenho do artefacto.
Desenho e desenvolvimento do artefacto	Conceção da arquitetura modular e implementação do protótipo com módulos de ingestão, OCR, segmentação, indexação, recuperação, geração, validação e monitorização documental.	Artefacto tecnológico implementado e descrito no capítulo 4.
Demonstração	Aplicação do protótipo a um corpus documental definido, composto por documentos digitais pesquisáveis e digitalizados com diferentes características estruturais.	Execução do pipeline e produção de evidência empírica sobre o funcionamento da solução.
Avaliação	Avaliação da extração textual, recuperação de evidência, qualidade das respostas, fundamentação documental, rastreabilidade e indicadores operacionais de confiança.	Resultados quantitativos e qualitativos apresentados no capítulo 5.
Comunicação	Sistematização da metodologia, arquitetura, implementação, resultados, limitações e contributos da investigação.	Discussão dos contributos científicos, metodológicos e técnicos.

3.4. Desenho do Artefacto

O artefacto desenvolvido no âmbito desta investigação consistiu numa arquitetura modular de processamento inteligente de documentos orientada para análise, recuperação de informação e geração de respostas fundamentadas em evidência documental. A sua conceção assentou em cinco componentes funcionais principais: ingestão e processamento documental, representação vetorial do conteúdo, recuperação semântica, geração condicionada por contexto e validação das respostas produzidas. Estes componentes foram complementados por uma camada transversal de monitorização documental inspirada no conceito de Gémeo Digital Documental.

O primeiro componente foi responsável pela receção e transformação dos documentos em conteúdo processável, incluindo extração de texto, preservação de estrutura documental e tratamento de documentos digitalizados. O segundo componente converteu esse conteúdo em representações vetoriais suscetíveis de indexação e pesquisa por similaridade. O terceiro componente executou a recuperação de excertos relevantes em função da pergunta colocada. O quarto componente utilizou essa evidência recuperada para condicionar a geração da resposta. O quinto componente introduziu mecanismos de controlo orientados para a validação da resposta face ao conteúdo documental, procurando reduzir inconsistências, omissões e geração não suportada. A Figura 2 representa esta arquitetura, com as entradas e as saídas de cada etapa e a camada transversal de monitorização.

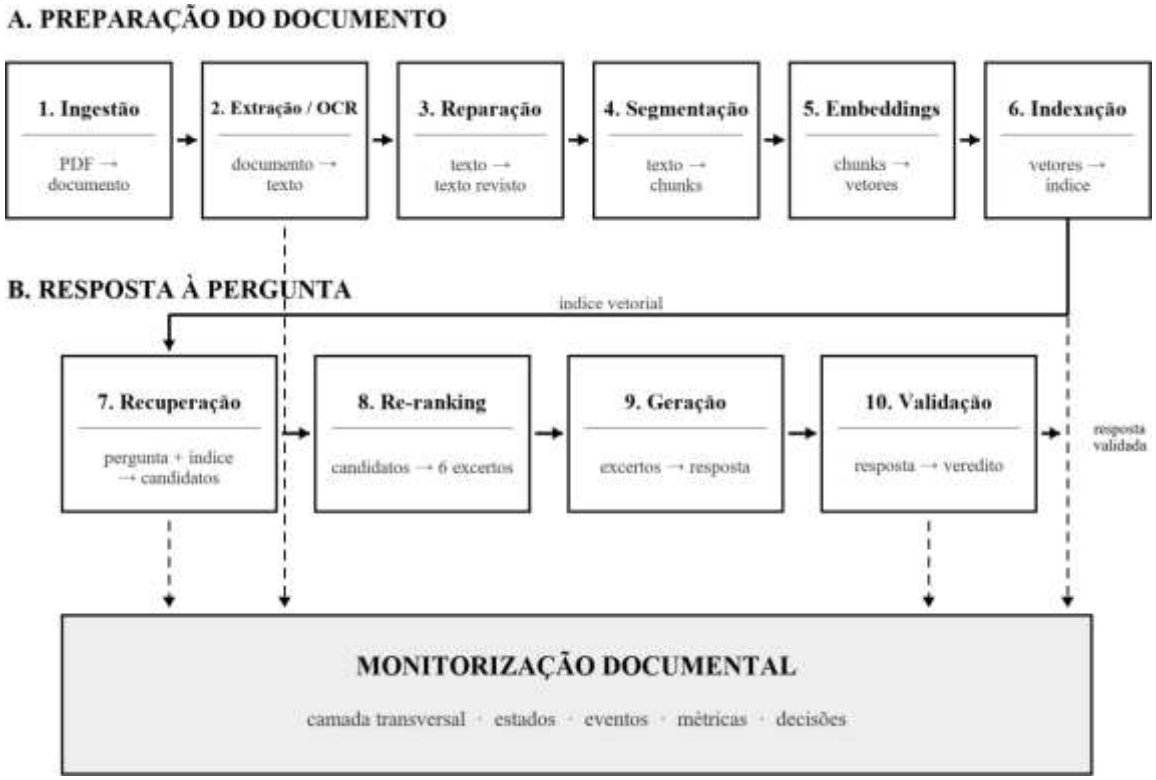

Figura 2 - Arquitetura do pipeline documental: etapas, entradas, saídas e camada transversal de monitorização

A conceção do artefacto foi orientada por seis princípios de desenho: modularidade, rastreabilidade, fundamentação em evidência, observabilidade, reprodutibilidade e conformidade. A modularidade permitiu separar os diferentes componentes do pipeline, facilitando a substituição ou evolução tecnológica de cada módulo. A rastreabilidade assegurou a ligação entre documento original, excertos recuperados e resposta gerada. A fundamentação em evidência procurou reduzir respostas não suportadas por fontes documentais. A observabilidade foi reforçada através da camada de monitorização documental. A reprodutibilidade orientou a parametrização e o isolamento do ambiente de execução, de modo a permitir repetir os ensaios em condições equivalentes. A conformidade foi considerada através da incorporação de princípios de minimização, limitação da finalidade, auditabilidade e proteção de dados.

3.4.1. Componentes Técnicos Considerados no Artefacto

A implementação do artefacto foi realizada em Python 3.10+ e disponibilizada como aplicação baseada em FastAPI. A execução foi suportada por Docker e Docker Compose, garantindo isolamento do ambiente e consistência entre execuções. A descrição detalhada de configuração, serviços, *endpoints*, *timeouts* e variáveis de ambiente é apresentada no Capítulo 4, de modo a manter a metodologia centrada no desenho de investigação e no protocolo de avaliação.

No plano metodológico, as tecnologias principais foram consideradas em função do papel que desempenharam no artefacto: ChromaDB foi utilizada como base vetorial para armazenamento e pesquisa por similaridade; Ollama suportou a execução local de modelos de linguagem, embeddings e análise visual; Docling foi utilizado como motor principal de processamento documental e FastAPI serviu como *framework* para exposição dos serviços da aplicação. Esta combinação permitiu operacionalizar um pipeline modular, reprodutível e avaliável.

3.5. Camada de Monitorização e Rastreabilidade

No desenho da arquitetura, o documento deixou de ser tratado apenas como uma entrada de processamento e passou a ser representado através de uma camada de monitorização documental inspirada no conceito de Gémeo Digital Documental. Neste projeto, este conceito foi adotado em sentido arquitetural e operacional, como mecanismo de observabilidade do ciclo de vida documental, e não como simulação autónoma ou réplica bidirecional completa de um sistema físico. A sua função consistiu em registar estados, eventos, métricas e relações entre documento, evidência recuperada e resposta gerada, reforçando a rastreabilidade e a auditabilidade do pipeline.

Em termos de conceção do sistema, o Gémeo Digital Documental foi definido como uma memória operacional associada a cada documento processado. Essa representação incorporou informação relativa à identidade do documento, metadados de origem, estado de processamento, estratégia de extração utilizada, texto reconhecido, estrutura detetada, unidades segmentadas para indexação, contexto recuperado em interações subsequentes,

respostas produzidas, validações atribuídas, métricas de confiança e histórico de eventos relevantes. Esta organização permitiu observar o comportamento do sistema numa perspetiva longitudinal, evitando que a análise se limitasse às saídas finais ou a momentos isolados do pipeline.

Do ponto de vista arquitetural, esta camada foi concebida como um módulo transversal ao pipeline, articulado com os serviços de ingestão, extração, segmentação, indexação, recuperação, geração e validação. A sua função consistiu em registar transições de estado e atualizar a representação interna do documento à medida que este progredia no sistema. A persistência do Gémeo Digital Documental foi delineada de forma separada da base vetorial utilizada para indexação semântica, preservando a distinção entre o nível semântico da recuperação e o nível observacional do sistema.

Deste modo, a introdução do Gémeo Digital Documental constituiu uma extensão do artefacto principal, orientada não para substituir os mecanismos centrais de processamento documental, mas para os tornar mais observáveis, auditáveis e tecnicamente interpretáveis.

3.6. Fases do Estudo

O desenvolvimento da investigação foi estruturado segundo as fases do processo DSR, mantendo uma organização sequencial e iterativa. Esta organização permitiu ligar a identificação do problema à construção do artefacto, à sua demonstração e à avaliação dos resultados obtidos, conforme sistematizado na Tabela 4.

Tabela 4 - Fases do estudo e correspondência com a DSR

Fase do estudo	*Descrição*	*Correspondência DSR*
Preparação do corpus e das perguntas de avaliação	Seleção do corpus documental, formulação de perguntas de question answering e construção de respostas de referência com indicação dos excertos documentais de suporte.	Definição dos objetivos da solução e preparação da demonstração.
Desenvolvimento do processamento documental	Implementação do módulo de extração de texto, tratamento de digitalizações e preparação de conteúdo pesquisável e indexável.	Desenho e desenvolvimento do artefacto.
Implementação da recuperação semântica	Definição de estratégias de segmentação, geração de representações vetoriais e indexação em base vetorial.	Desenho e desenvolvimento do artefacto.

Integração com LLM e validação da resposta	Geração de respostas condicionadas por evidência recuperada e introdução de mecanismos de controlo orientados para reduzir respostas não fundamentadas.	Demonstração do artefacto.
Avaliação do sistema	Aplicação de métricas quantitativas e apreciação qualitativa estruturada para avaliar extração, recuperação, geração, rastreabilidade e monitorização documental.	Avaliação do artefacto.

3.7. Corpus Documental e Unidade de Análise

A avaliação do protótipo assentou num corpus composto por 150 documentos digitais, construídos de modo a representar diferentes condições técnicas de processamento documental em contexto organizacional. A unidade principal de análise foi o documento digital, entendido como objeto processado pelo pipeline desde a fase de ingestão até à extração, segmentação, indexação, recuperação de evidência, geração de respostas e respetivos registos de rastreabilidade. Os documentos que compõem o corpus foram gerados sinteticamente para a investigação, reproduzindo a diversidade documental do contexto organizacional sem conter dados pessoais reais.

Os documentos foram gerados por um script em Python desenvolvido para o efeito, a partir de modelos (*templates*) de três famílias documentais, requerimentos, formulários e relatórios, 50 de cada, com conteúdo fictício em português, construído a partir de conjuntos de parágrafos predefinidos, e sem intervenção do modelo de linguagem utilizado pelo sistema. O corpus foi organizado em cinco escalões de extensão (1, 3, 5, 7 e 10 páginas), com 30 documentos cada. Cada documento foi produzido numa de cinco variantes de apresentação: PDF pesquisável simples (50 documentos), com tabelas (30), com layout a duas colunas (30), com ruído de fundo a simular papel (15) e digitalizado (25). Os documentos digitalizados foram gerados como imagem rasterizada a 150 dpi, com ruído, desfocagem ligeira e rotação inferior a um grau, inserida num PDF sem camada de texto. Por construção, cada digitalizado corresponde a uma única página, pelo que o conteúdo que excedia essa página não ficou representado na imagem. No total, o corpus integra 125 PDFs pesquisáveis e 25 digitalizados, com tabelas, layouts a duas colunas e campos críticos, como datas, valores, códigos e identificadores.

O corpus foi caracterizado com base em propriedades estruturais e técnicas relevantes para o desempenho do sistema. Entre essas propriedades incluíram-se a existência ou ausência de camada textual pesquisável, a necessidade de OCR, a presença de tabelas, imagens ou elementos visuais, a complexidade do layout, a qualidade visual observável e a existência de campos críticos, como datas, valores, códigos, identificadores ou referências documentais. Esta opção permitiu aproximar a análise das condições que efetivamente influenciam a extração, a segmentação e a recuperação de informação. A distinção entre documentos pesquisáveis e documentos digitalizados foi definida com base na presença real de camada textual no ficheiro PDF, e não a partir do motor de extração utilizado pela pipeline. Esta separação foi metodologicamente necessária para distinguir, na avaliação, as causas dos efeitos: o tipo técnico do documento foi tratado como característica de entrada, enquanto o motor utilizado, o recurso a fallback, a eventual reparação pós-OCR, a reindexação e as métricas de confiança operacional foram analisadas como resultados do comportamento do sistema.

Para cada documento foi definido um conjunto de cinco perguntas transversais de *question answering*, totalizando 750 perguntas de avaliação. Estas perguntas incidiram sobre dimensões comuns aos documentos analisados, nomeadamente: a finalidade principal; as entidades ou serviços mencionados; as datas, prazos, valores, códigos ou referências; o excerto de suporte; e a conclusão, decisão, resultado ou recomendação. Cada pergunta foi associada a uma resposta de referência e a um excerto documental de suporte, extraído do próprio documento. Esta estrutura permitiu separar analiticamente duas dimensões do desempenho do sistema: por um lado, a capacidade de recuperar evidência documental relevante; por outro, a capacidade de gerar uma resposta correta, coerente e fundamentada com base nessa evidência.

As 750 perguntas, as respostas de referência e os excertos de suporte não foram redigidos manualmente: foram produzidos automaticamente pelo mesmo script que gerou os documentos, por regras fixas associadas aos templates. As cinco perguntas têm uma formulação comum a todos os documentos. A resposta de referência é composta a partir dos dados usados para preencher cada documento, como a finalidade, as entidades, a referência e as datas, ou corresponde ao próprio texto do documento, no caso do excerto de suporte e

da conclusão. O excerto de suporte é um parágrafo do próprio documento, selecionado por regra: o primeiro parágrafo do corpo, nas perguntas sobre a finalidade e as entidades; o parágrafo que contém a referência ou a data, na pergunta sobre datas, valores e referências; o parágrafo mais extenso, na pergunta sobre o excerto de suporte; e a conclusão, na última pergunta. Esta geração não recorreu a nenhum modelo de linguagem, pelo que o modelo avaliado não interveio na definição das respostas de referência, garantindo que o critério de correção é independente do sistema avaliado. O conjunto não foi, contudo, validado por uma segunda pessoa, o que constitui uma limitação (secção 6.2).

3.8. Protocolo de Avaliação

O protocolo de avaliação foi estruturado em três níveis complementares: avaliação da extração documental, avaliação da recuperação de evidência e avaliação da geração de respostas. Esta decomposição permitiu analisar o comportamento do sistema tanto ao nível dos seus componentes individuais como ao nível do pipeline integrado, evitando que falhas ocorridas numa etapa fossem atribuídas indevidamente a outra. Deste modo, tornou-se possível observar de que forma a qualidade da extração condicionou a segmentação, como a segmentação influenciou a recuperação e em que medida a evidência recuperada afetou a resposta produzida pelo modelo de linguagem.

A avaliação da extração documental incidiu sobre a qualidade e a utilidade operacional do texto produzido a partir de documentos com diferentes características técnicas e estruturais, incluindo documentos com camada textual pesquisável, documentos dependentes de OCR, presença de tabelas, elementos visuais e diferentes graus de complexidade de layout. Esta avaliação considerou indicadores como volume de texto extraído, número de chunks gerados e indexados, ocorrência de fallback, necessidade de reparação pós-OCR, tempo de processamento, confiança operacional de extração e, nos documentos pesquisáveis, CER. A confiança operacional foi tratada como indicador interno do comportamento da pipeline, não como medida formal de qualidade textual validada.

A avaliação da recuperação de evidência analisou a capacidade do sistema para identificar excertos documentais relevantes em resposta às perguntas previamente definidas. Esta etapa

foi avaliada de forma autónoma, uma vez que a qualidade da resposta gerada depende diretamente da pertinência, cobertura e posicionamento da evidência recuperada. Assim, a recuperação foi analisada antes da geração, permitindo distinguir situações em que uma resposta incorreta resultou de falha na seleção dos excertos de suporte de situações em que o erro ocorreu já na etapa generativa.

A avaliação da geração considerou a qualidade das respostas produzidas pelo modelo de linguagem com base na evidência recuperada, atendendo a critérios como correção factual, completude, clareza, coerência, fundamentação documental e rastreabilidade. Esta dimensão não foi avaliada como desempenho isolado do LLM, mas como resultado final da articulação entre extração, segmentação, recuperação e geração. A existência de mecanismos de grounding e de registo no Gémeo Digital Documental permitiu ainda observar se a resposta gerada se encontrava suficientemente suportada nos excertos fornecidos ao modelo.

Sempre que aplicável, a avaliação quantitativa foi complementada por apreciação qualitativa, através da análise comparativa de respostas geradas com e sem evidência recuperada, orientada por critérios de correção, completude, fundamentação documental, clareza e utilidade da resposta. Esta combinação permitiu captar dimensões técnicas e interpretativas que não são integralmente observáveis através de métricas automáticas, nomeadamente a adequação da resposta ao contexto documental, a suficiência da evidência mobilizada e a utilidade prática do resultado produzido. Ao mesmo tempo, foram mantidas separadas as métricas operacionais calculadas internamente pela pipeline e as métricas de validação formal baseadas em referência externa ou apreciação estruturada.

3.8.1. Procedimento de Teste Automatizado

A avaliação do protótipo integrou uma componente de testes automatizados destinada a observar o comportamento do sistema em condições controladas e repetíveis. Estes testes incidiram sobre diferentes etapas do pipeline, permitindo identificar falhas, comparar configurações de execução e medir a estabilidade do sistema.

A realização de testes automatizados revelou-se particularmente relevante num protótipo modular, na medida em que permitiu avaliar de forma consistente o impacto de diferentes

escolhas de processamento e verificar se o sistema manteve desempenho aceitável perante documentos com características distintas. Para esse efeito, os testes foram conduzidos sobre subconjuntos representativos do corpus documental, abrangendo documentos textuais, documentos digitalizados e documentos com estrutura mista.

Na prática, estes testes assumiram duas formas: os scripts de avaliação automatizada, executados sobre a totalidade do corpus e cujos resultados constituem o Capítulo 5, e testes unitários dos módulos de reparação pós-OCR e da política de decisão do gémeo digital, usados para verificar o comportamento esperado durante o desenvolvimento. Os resultados destes últimos não são reportados, por se tratar de verificação de software e não de avaliação do artefacto.

3.8.2. Estratégia de Benchmarking

Para complementar a avaliação funcional do sistema, foi definida uma estratégia de *benchmarking* orientada para a comparação entre diferentes configurações de processamento documental. Esta comparação incidiu sobre métricas como velocidade de processamento, qualidade da extração e completude do conteúdo obtido, permitindo identificar vantagens e limitações relativas entre motores ou modos de execução alternativos.

O benchmarking foi entendido como um procedimento metodológico de apoio à validação empírica das opções de desenho adotadas no artefacto. Em vez de constituir uma simples comparação técnica entre ferramentas, permitiu analisar em que medida diferentes opções de processamento produziram impacto mensurável na qualidade final do pipeline.

Os ensaios de benchmarking foram realizados durante o desenvolvimento para apoiar a seleção e o ajustamento da configuração final do protótipo. Como tiveram um caráter técnico e exploratório, não integraram a avaliação experimental aplicada ao corpus nem foram utilizados para sustentar as conclusões do estudo; por essa razão, os seus resultados não são apresentados autonomamente no Capítulo 5.

Como exemplo, nos 15 documentos com ruído visual, o Docling registou um tempo médio de processamento de 28,82 segundos e uma qualidade estrutural média de 0,80, enquanto o PyMuPDF e o pdfplumber obtiveram 0,20. Estes resultados mostram que o maior

custo de processamento do Docling foi acompanhado por uma melhor preservação explícita da estrutura documental na saída.

3.9. Métricas de Avaliação

A avaliação do protótipo foi organizada em torno de métricas e critérios adequados às diferentes dimensões do artefacto. A Tabela 5 sintetiza a relação entre dimensão avaliada, métricas ou critérios utilizados e respetiva finalidade. As duas primeiras dimensões, extração e recuperação, foram avaliadas com métricas quantitativas formalizadas; a geração foi objeto de apreciação qualitativa comparativa. A rastreabilidade, a monitorização e a conformidade não foram medidas com métricas formais: foram avaliadas por verificação do desenho, confirmando a presença dos mecanismos correspondentes, e por inspeção dos registos do gémeo digital, cujos resultados se apresentam no Capítulo 5 sob a forma de contagens de decisões adaptativas.

Tabela 5 - Métricas e critérios de avaliação do artefacto

Dimensão avaliada	*Métricas / critérios*	*Finalidade*
Extração textual nos PDFs pesquisáveis	CER, completude do conteúdo, preservação de estrutura e legibilidade computacional.	Avaliar divergências de extração e serialização; o CER aqui calculado não mede OCR nos documentos digitalizados.
Recuperação de evidência	Precision, recall, f1-score, precision@k, recall@k	Avaliar a relevância, cobertura e ordenação dos excertos recuperados.
Geração de respostas	Correção factual, completude, clareza, fundamentação documental e ausência de resposta não suportada.	Ilustrar se a evidência recuperada torna as respostas específicas e verificáveis, sem estimar um efeito para todo o corpus.
Rastreabilidade	Presença de fonte, ligação ao excerto, reconstrução do percurso documento-resposta.	Avaliar a explicabilidade operacional do pipeline.
Monitorização documental	Estados registados, eventos capturados, métricas de confiança e histórico de processamento.	Avaliar a utilidade da camada de monitorização documental.
Conformidade	Minimização, limitação de finalidade, auditabilidade e controlo de dados sensíveis.	Avaliar a incorporação de princípios de proteção de dados.

3.9.1. Framework Interno de Qualidade

Para apoiar a monitorização do processamento documental, foi definido um framework interno de qualidade baseado em oito dimensões complementares. Estas dimensões não substituíram as métricas principais de avaliação, mas funcionaram como indicadores auxiliares de controlo do comportamento do sistema ao longo do pipeline. A Tabela 6 apresenta estas dimensões. Trata-se de indicadores heurísticos calculados internamente pelo módulo de qualidade durante o processamento, sem referência externa, e não de métricas de avaliação: não foram usados nos resultados apresentados no Capítulo 5.

Tabela 6 - Dimensões do framework interno de qualidade

Dimensão	*Descrição*
Correção textual (indicador heurístico)	Sinal interno baseado em padrões de degradação textual, como integridade de caracteres, palavras fundidas e separações anómalas;
Completeness (Completude)	Cobertura do conteúdo extraído face ao documento de origem
Consistency (Consistência)	Uniformidade terminológica, temporal e estrutural
Readability (Legibilidade)	Clareza textual e organização do conteúdo extraído
Structure Quality (Qualidade Estrutural)	Preservação de secções, hierarquia e organização documental
Semantic Coherence (Coerência Semântica)	Continuidade temática e consistência do significado
Visual Alignment (Alinhamento Visual)	Correspondência entre texto extraído e elementos visuais/documentais
Cross-Modal Validation (Validação Cross-Modal)	Consistência entre diferentes métodos de extração e análise

3.9.2. Avaliação da Extração Documental

A divergência da extração textual nos PDFs pesquisáveis foi avaliada através da Taxa de Erro de Caracteres (Character Error Rate - CER), calculada por comparação com a camada textual nativa usada como referência. Neste estudo, esta aplicação do CER mede diferenças de extração, serialização e ordem de leitura; não constitui uma avaliação do reconhecimento ótico nos documentos digitalizados. A métrica é definida da seguinte forma:

$$CER = \frac{S + D + I}{N}$$

Nesta fórmula, S corresponde ao número de substituições, D ao número de deleções ou omissões, I ao número de inserções e N ao número total de caracteres no texto de referência. Valores inferiores indicam menor divergência entre a extração e a referência. A medição de OCR propriamente dita exigiria a comparação dos 25 documentos digitalizados com o respetivo texto-fonte, análise que não foi executada nos resultados disponíveis.

O sistema recorreu a uma estratégia de processamento híbrida, combinando diferentes métodos de extração e mecanismos de fallback hierárquico para maximizar cobertura e robustez. Em termos metodológicos, este desenho permitiu comparar o comportamento de um motor principal de processamento com alternativas de recurso em situações de falha ou degradação do documento. A Tabela 7 resume esta estratégia de processamento e fallback.

Tabela 7 - Estratégia de processamento e fallback do OCR

Nível	*Componente*	*Função*
Nível 1	Docling	Motor principal de processamento documental
Nível 2	OCR tradicional	Recurso para documentos digitalizados ou falhas no motor principal
Nível 3	Extração nativa básica	Último recurso para recuperação mínima de conteúdo

3.9.3. Avaliação da recuperação de evidência

A eficácia da recuperação foi avaliada com recurso às métricas clássicas de sistemas de recuperação de informação: **Precision**, **Recall** e **F1-score** (Manning et al., 2009). Estas métricas permitem medir, respetivamente, a seletividade da recuperação, a sua cobertura e o equilíbrio entre ambas.

- A Precision mede a proporção de excertos recuperados que são efetivamente relevantes. Um valor elevado indica que o sistema devolve poucos resultados irrelevantes.

- A Recall mede a proporção de excertos relevantes existentes que foram efetivamente recuperados pelo sistema, avaliando a completude da recuperação.
- A F1-score corresponde à média harmónica entre Precision e Recall, penalizando situações em que uma das dimensões apresenta valor elevado e a outra permanece reduzida. As métricas foram calculadas individualmente por pergunta e posteriormente agregadas através de média macro.

Neste projeto, um chunk é considerado relevante para uma pergunta quando contém o excerto de evidência de referência dessa pergunta. Como as respostas de referência não identificam os chunks relevantes, esta condição foi operacionalizada de forma automática. Após normalização, um chunk é relevante se contiver integralmente o excerto de evidência, ou se pelo menos 30% das palavras do excerto de evidência, consideradas as que têm mais de dois caracteres, estiverem presentes no chunk. O conjunto de chunks relevantes é determinado sobre todos os chunks do documento.

Sendo R o conjunto de chunks relevantes de uma pergunta e D_k o conjunto dos k primeiros chunks devolvidos pelo sistema, as métricas são definidas da seguinte forma:

$$Precision@k = \frac{|R \cap D_k|}{|D_k|}$$

$$Recall@k = \frac{|R \cap D_k|}{|R|}$$

$$F1 = \frac{2 \times Recall@k \times Precision@k}{Recall@k + Precision@k}$$

As métricas são calculadas por pergunta e agregadas por média macro. As medidas globais de Precision, Recall e F1 correspondem à configuração final do sistema, com k = 6.

A título de exemplo, na pergunta «Qual é a finalidade principal do documento?», aplicada ao documento req_001, o excerto de evidência é o primeiro parágrafo do corpo do

documento: «Vera Santos, titular do número de colaborador […], solicitar a concessão de licença de férias pelo período de 13 dias úteis, com início em 18/10/2023 e término previsto em 31/10/2023». Três chunks do documento satisfazem o critério de relevância. Dos seis chunks devolvidos pelo sistema, dois pertencem a esse conjunto, pelo que Precision@6 = 2/6 = 0,333 e Recall@6 = 2/3 = 0,667.

Exemplos ilustrativos de cálculo: com Precision@6 = 0,333 e Recall@6 = 0,667, o F1 é aproximadamente 0,444. Para o CER, uma referência com 1000 caracteres e uma hipótese com 10 substituições, 5 deleções e 3 inserções teria CER = (10 + 5 + 3) / 1000 = 0,018, ou 1,8%. Este segundo valor é apenas um exemplo da aplicação da fórmula, não um resultado adicional do corpus.

Para complementar esta avaliação, foram ainda consideradas métricas dependentes da profundidade de recuperação. No caso de Precision@K e Recall@K, foram usados K = 3, 5 e 6, por corresponderem a diferentes amplitudes de contexto. Estas métricas indicam seletividade e cobertura até à profundidade K, mas não distinguem suficientemente a posição exata do primeiro resultado relevante; por esse motivo, Hit@K e MRR são identificadas como trabalho futuro. A Tabela 8 sistematiza as métricas calculadas.

Tabela 8 - Métricas de avaliação da recuperação

Métrica	***Descrição***	***Finalidade***
Precision	Proporção de excertos recuperados que são relevantes	Mede a seletividade da recuperação
Recall	Proporção de excertos relevantes que foram efetivamente recuperados	Mede a completude da recuperação
F1-score	Média harmónica entre precision e recall	Mede o equilíbrio entre seletividade e completude
Precision@K	Proporção de resultados relevantes nos primeiros Kresultados	Mede a qualidade dos primeiros resultados
Recall@K	Proporção de resultados relevantes recuperados nos primeiros Kresultados	Mede a cobertura em função do número de resultados considerados (K)

3.9.4. Avaliação da Geração de Respostas e Relação com a Recuperação

A avaliação da geração de respostas não foi tratada de forma isolada, uma vez que a qualidade dos excertos recuperados condiciona diretamente a qualidade da resposta gerada. Assim, a análise do desempenho do sistema considerou a recuperação como base da etapa generativa, permitindo observar de que forma evidência insuficiente, redundante ou mal posicionada afetou a resposta final.

A apreciação das respostas incidiu sobre os critérios já enunciados no protocolo de avaliação. Esta abordagem permitiu avaliar a resposta enquanto resultado final do pipeline integrado, mas também identificar em que etapa se originaram eventuais falhas: extração textual, segmentação, recuperação de evidência ou geração. Importa sublinhar que a extração e a recuperação foram objeto de avaliação quantitativa, enquanto a geração de respostas foi avaliada de forma qualitativa e comparativa, com e sem contexto recuperado.

3.10. Procedimentos Éticos e Conformidade

A avaliação e o desenvolvimento do protótipo exigiram atenção a considerações éticas relacionadas com o tratamento de documentação potencialmente sensível, a proteção de dados e a utilização de modelos generativos em tarefas de análise documental. Neste contexto, a metodologia adotada pressupôs a minimização da exposição de dados pessoais, a utilização controlada do corpus documental e a limitação do tratamento ao estritamente necessário para os objetivos da investigação.

A utilização de modelos de linguagem e mecanismos de recuperação em contexto documental introduz riscos associados à produção de respostas imprecisas, à exposição indevida de informação e à interpretação excessivamente confiante de resultados automatizados. Por essa razão, a avaliação do sistema foi orientada não apenas para medir desempenho, mas também para identificar limites de fiabilidade, necessidades de validação humana e condições em que a automatização deve ser complementada por supervisão.

A conformidade foi tratada como dimensão transversal do pipeline, e não como requisito externo acrescentado no final do desenvolvimento. Esta orientação permitiu incorporar

princípios de privacy by design, limitação da finalidade, segurança, rastreabilidade e auditabilidade na própria lógica de funcionamento do artefacto.

Na preparação deste relatório foi utilizada uma ferramenta de inteligência artificial generativa como apoio à revisão linguística, à reorganização estrutural do texto e à verificação de coerência. A ferramenta não foi utilizada para gerar dados empíricos, executar as análises apresentadas nem substituir o juízo científico do autor. Todas as sugestões foram revistas criticamente e verificadas face aos registos do projeto e às fontes citadas.

4. DESENVOLVIMENTO E IMPLEMENTAÇÃO DA SOLUÇÃO

Este capítulo apresenta o desenvolvimento e a implementação da solução tecnológica concebida no âmbito do projeto. A descrição centra-se nas decisões arquiteturais, na organização modular do protótipo e na forma como os componentes técnicos materializam o artefacto definido na metodologia. Ao contrário de uma documentação de instalação ou de um manual de utilização, o capítulo privilegia a justificação das opções de desenho, a articulação entre módulos e a relação entre implementação, rastreabilidade, explicabilidade e reprodutibilidade experimental.

O protótipo foi desenvolvido como uma aplicação modular para processamento inteligente de documentos, integrando extração e estruturação de conteúdo, indexação vetorial, recuperação semântica, geração de respostas condicionada por evidência documental e mecanismos adicionais de validação e monitorização. Esta organização operacionaliza a arquitetura de inteligência documental definida no capítulo anterior e constitui a base instrumental para a avaliação apresentada no Capítulo 5.

4.1. Enquadramento do Desenvolvimento do Artefacto

O desenvolvimento do artefacto decorreu em duas fases principais. A primeira fase correspondeu à construção do protótipo funcional de ponta a ponta, cobrindo o percurso desde a ingestão do documento até à geração de respostas fundamentadas em evidência recuperada. Nesta fase foram integrados o processamento documental, a segmentação do conteúdo extraído, a indexação vetorial, a recuperação semântica e a interação com modelos de linguagem executados localmente.

A segunda fase incidiu sobre a integração e o refinamento da solução. O trabalho concentrou-se no ajustamento da segmentação documental, na estabilização da recuperação, na melhoria dos mecanismos de validação e no reforço da rastreabilidade e observabilidade do pipeline. Neste contexto, foi integrada uma camada transversal de monitorização documental, inspirada no conceito de Gémeo Digital Documental, destinada a acompanhar o estado de cada documento ao longo do processamento, registar eventos relevantes e apoiar mecanismos de controlo.

Em termos de Design Science Research, este capítulo corresponde à materialização do artefacto. A arquitetura aqui descrita resulta da tradução dos objetivos da solução em componentes implementáveis, mantendo a ligação entre o problema de investigação, os requisitos de desenho, a camada de observabilidade documental e o protocolo de avaliação.

4.1.1. Requisitos funcionais e não funcionais

A tradução dos objetivos da solução em componentes implementáveis partiu de um conjunto de requisitos funcionais (RF) e não funcionais (RNF), identificados, priorizados e acompanhados de um critério de aceitação.

- **RF01 - Submissão de documentos (prioridade alta)**: o sistema aceita um ficheiro PDF, atribui-lhe um identificador único e cria o respetivo registo no gémeo digital.
- **RF02 - Extração e estruturação (alta)**: o texto e a estrutura do documento, incluindo secções e tabelas, são extraídos com o Docling e guardados em Markdown; em caso de falha, é usado um motor de recurso.
- **RF03 - Reparação pós-OCR (média)**: quando é detetada degradação no espaçamento, é gerada uma versão reparada, o texto bruto é preservado e a decisão fica registada no gémeo digital.
- **RF04 - Segmentação e indexação (alta)**: o conteúdo é dividido em chunks de cerca de 1000 caracteres, com sobreposição de cerca de 100, e indexado na ChromaDB com o identificador do documento.
- **RF05 - Recuperação e re-ranking (alta)**: para cada pergunta, são entregues ao modelo até seis excertos do documento selecionado.
- **RF06 - Geração fundamentada (alta)**: a resposta é gerada a partir da pergunta e dos excertos recuperados.
- **RF07 - Validação de grounding (alta)**: cada resposta é verificada face aos excertos; se a verificação falhar, é feita uma nova tentativa com temperatura 0,0.
- **RF08 - Monitorização do ciclo de vida (alta)**: os estados, eventos e métricas de cada documento ficam registados e podem ser consultados.
- **RNF01 - Privacidade (alta)**: nenhum conteúdo documental é enviado para serviços externos; todos os modelos são executados localmente através do Ollama.

- **RNF02 - Reprodutibilidade (alta)**: o sistema é executável com Docker Compose e os parâmetros da avaliação estão documentados na Tabela 14.
- **RNF03 - Modularidade (média)**: os motores de extração, os modelos e os parâmetros de recuperação são configuráveis sem alteração do código.
- **RNF04 - Desempenho interativo (média)**: a resposta a uma pergunta é obtida em menos de 60 segundos no hardware de desenvolvimento.
- **RNF05 - Conformidade por desenho (alta)**: a recuperação é filtrada pelo documento selecionado, a memória conversacional tem retenção configurada de 24 horas e as decisões do pipeline ficam registadas.

4.2. Arquitetura do Sistema

A arquitetura do sistema foi concebida de forma modular e observável, permitindo separar as principais responsabilidades do processamento documental: ingestão, extração de conteúdo, reparação pós-OCR, segmentação, geração de embeddings, indexação vetorial, recuperação de evidência, re-ranking, geração de respostas, validação de grounding e monitorização do ciclo de vida documental. Esta organização permitiu estruturar a solução como um pipeline documental de ponta a ponta, no qual cada componente desempenha uma função específica, mas articulada com as restantes etapas do sistema.

O documento é inicialmente submetido através do módulo de ingestão, sendo-lhe atribuído um identificador único e criada a respetiva representação na camada de monitorização. Em seguida, o módulo de extração converte o documento em texto e estrutura processável, recorrendo ao Docling como motor principal e a mecanismos de OCR, análise visual ou fallback quando necessário. Após a extração, o texto pode ser analisado por uma etapa de reparação pós-OCR, destinada a mitigar problemas como palavras fundidas, sílabas separadas ou falhas de espaçamento. Esta etapa preserva sempre o texto bruto e apenas utiliza a versão reparada para indexação quando existem sinais operacionais de melhoria.

O conteúdo preparado é depois dividido em chunks, convertido em embeddings e armazenado na ChromaDB, permitindo pesquisa semântica por documento. Quando o utilizador coloca uma pergunta, o sistema recupera os excertos potencialmente relevantes,

podendo aplicar reformulação da pergunta, heurísticas de reforço e re-ranking antes de enviar o contexto final ao modelo de linguagem. A resposta é gerada por um LLM local com base nos excertos recuperados e posteriormente submetida a uma verificação de grounding, destinada a avaliar se a resposta se encontra suportada pela evidência fornecida.

Transversalmente, o Gémeo Digital Documental regista estados, eventos, métricas operacionais e decisões relevantes, permitindo reconstruir o percurso entre o documento original, o texto extraído, os chunks indexados, a evidência recuperada e a resposta final. Deste modo, a arquitetura não foi concebida apenas para produzir respostas, mas também para tornar observável e auditável o processo que conduz a essas respostas.

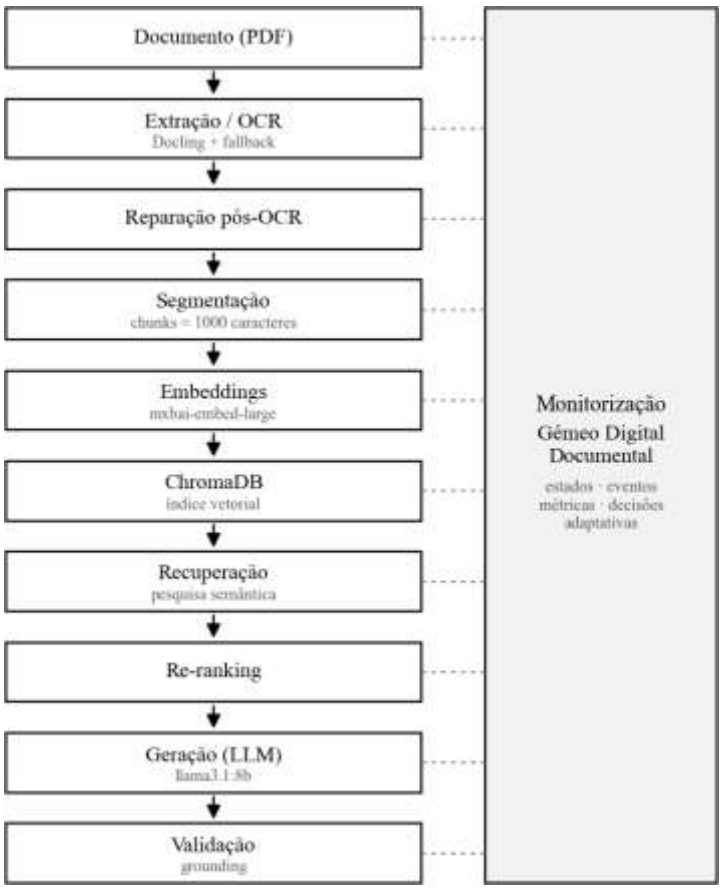

Figura 3 - Arquitetura e fluxo de processamento documental do protótipo

A arquitetura foi orientada por seis princípios de desenho, que funcionaram como critérios transversais de implementação e avaliação: modularidade, rastreabilidade, fundamentação em evidência, observabilidade, reprodutibilidade e conformidade. A Tabela 9 sintetiza estes princípios.

Tabela 9 - Princípios de desenho arquitetural da solução

Princípio de desenho	*Aplicação na arquitetura*	*Finalidade*
Modularidade	Separação entre ingestão, extração, reparação pós-OCR, segmentação, indexação, recuperação, re-ranking, geração, validação e monitorização.	Permitir evolução ou substituição de componentes sem reconfigurar todo o sistema.
Rastreabilidade	Ligação entre documento original, texto extraído, chunks indexados, evidência recuperada, resposta gerada e validação de grounding.	Apoiar auditoria, explicabilidade operacional e análise de falhas.
Fundamentação em evidência	Respostas geradas a partir de contexto recuperado da base documental.	Reduzir respostas não suportadas e reforçar verificabilidade.
Observabilidade	Registo de estados, eventos, métricas e sinais operacionais através da camada de monitorização documental.	Permitir acompanhar o comportamento do sistema ao longo do pipeline.
Reprodutibilidade	Execução containerizada e configuração parametrizada por níveis de precedência.	Facilitar repetição dos ensaios e comparação entre configurações.
Conformidade	Execução local dos modelos, registo auditável e separação entre o armazenamento vetorial e a memória operacional.	Incorporar por desenho a minimização, a limitação da finalidade e a auditabilidade.

A operacionalização destes princípios materializou-se num conjunto de componentes funcionais articulados, apresentados na Tabela 10. Esta decomposição permitiu organizar o sistema em módulos avaliáveis, facilitando a identificação de eventuais falhas, a análise do impacto de cada etapa e a relação entre os resultados experimentais e os componentes concretos do pipeline.

Tabela 10 - Componentes funcionais da arquitetura

Componente	*Função principal*	*Resultado produzido*
Ingestão documental	Receber ficheiros, validar formato e criar identificador único.	Documento registado para processamento.
Extração e estruturação	Extrair texto, tabelas, metadados e estrutura lógica, distinguindo conceptualmente o tipo técnico do documento da estratégia de processamento utilizada.	Conteúdo documental em formato processável, motor utilizado, sinais operacionais de extração e metadados técnicos.
Reparação textual pós-OCR	Detetar e mitigar degradações textuais resultantes do OCR, preservando o texto bruto e registando a decisão de indexação.	Texto bruto preservado, texto reparado quando aplicável, relatório de reparação e indicação da versão indexada.
Segmentação	Dividir o conteúdo em chunks com granularidade adequada.	Unidades documentais indexáveis.
Geração de embeddings	Converter cada chunk numa representação vetorial semântica.	Vetores associados aos fragmentos documentais.
Indexação vetorial	Persistir os embeddings na ChromaDB, com os metadados de origem, e disponibilizá-los para pesquisa por similaridade.	Índice pesquisável por similaridade.
Recuperação semântica	Selecionar excertos relevantes para a pergunta do utilizador.	Contexto documental para o LLM.
Re-ranking	Reordenar os candidatos recuperados de acordo com a sua relevância para a pergunta.	Lista final de excertos a enviar ao modelo gerador.
Geração condicionada	Produzir resposta com base no contexto recuperado.	Resposta textual fundamentada.
Validação	Avaliar coerência entre resposta e evidência recuperada.	Sinal de fundamentação e controlo.
Monitorização documental	Registar estados, eventos, métricas e relações documentais.	Memória operacional auditável do documento.

4.2.1. Camada de Monitorização Documental

A arquitetura desenvolvida integrou uma camada transversal inspirada no conceito de Gémeo Digital Documental. Esta componente foi concebida como uma representação operacional e auditável do estado de cada documento ao longo do pipeline. A sua função consistiu em registar estados, eventos, métricas e relações entre o documento original, o

conteúdo extraído, a evidência recuperada e a resposta gerada. Permitiu tratar cada documento como uma entidade persistente e observável. A partir do momento em que um ficheiro é submetido, é criada uma instância lógica associada ao documento, enriquecida progressivamente com metadados de origem, resultados de extração, informação sobre indexação, evidência recuperada, respostas produzidas, validações atribuídas e eventos relevantes. Esta memória operacional apoia a reconstrução do percurso entre o documento original e a resposta final, reforçando a rastreabilidade e a auditabilidade do sistema.

Do ponto de vista arquitetural, a camada de monitorização foi separada da base vetorial. Enquanto a ChromaDB foi dedicada ao armazenamento e recuperação de embeddings, a camada de Gémeo Digital Documental assumiu a função de memória operacional e de controlo adaptativo. Esta separação preservou a distinção entre o nível semântico da recuperação e o nível observacional do pipeline, favorecendo a modularidade, a extensibilidade e a interpretação técnica do comportamento do sistema. A implementação desta camada foi organizada em três níveis: um modelo de domínio para representar o estado do documento, uma camada de persistência para armazenar o estado e o histórico de eventos, e uma camada de serviço responsável por coordenar as atualizações ao longo do pipeline. Esta organização permitiu monitorizar os motores técnicos de extração, reparação pós-OCR, indexação, recuperação e geração, assegurando que os diferentes módulos interagissem com uma interface estável de registo e atualização do estado documental. Para apoiar a interpretação do comportamento do sistema, foram definidas métricas operacionais de confiança associadas à extração, recuperação e fundamentação das respostas. Estas métricas não foram concebidas como medidas probabilísticas formais, mas como indicadores auxiliares destinados a sinalizar degradações, apoiar auditoria e orientar decisões internas do pipeline. A pontuação global de confiança agregou estes sinais de forma ponderada, permitindo sintetizar a robustez observada em diferentes etapas do processamento documental.

Uma das funções mais relevantes desta camada consistiu no suporte a decisões adaptativas do pipeline. Após a extração inicial realizada pelo motor principal, o resultado é registado na camada de monitorização e utilizado para avaliar a qualidade operacional do conteúdo extraído. Quando são detetados sinais de degradação textual, como palavras

fundidas, quebras anómalas ou falhas de espaçamento, o sistema pode ativar mecanismos de reparação pós-OCR antes da segmentação e indexação. Esta decisão é registada no Gémeo Digital Documental, juntamente com a indicação da versão textual preservada, reparada ou efetivamente utilizada para indexação. Além da reparação textual, a camada de monitorização também suporta o mecanismo de fallback orientado por confiança. Quando a confiança de extração se situa abaixo de um limiar definido, o sistema pode recomendar ou acionar o reprocessamento do documento com um motor alternativo. Assim, o fallback não depende apenas de falhas técnicas explícitas, podendo também ser ativado quando a extração é formalmente concluída, mas apresenta qualidade insuficiente para suportar recuperação e geração fiáveis.

A camada de monitorização apoiou ainda mecanismos de auto-recuperação em situações de falha parcial da indexação. Em determinados cenários, nomeadamente sob pressão de recursos computacionais, a extração textual podia ser concluída sem que a indexação vetorial produzisse segmentos recuperáveis. Nesses casos, o sistema detetava automaticamente a inconsistência entre conteúdo extraído e ausência de segmentos indexados, acionando uma reindexação a partir da representação textual intermédia disponível. Esta estratégia reduziu a dependência de intervenção manual e aumentou a robustez do pipeline perante falhas parciais.

Deste modo, a camada de monitorização documental não substituiu os componentes centrais de OCR, indexação, recuperação e geração, mas acrescentou uma dimensão de observabilidade e controlo operacional sobre o seu funcionamento. O seu contributo principal residiu na capacidade de transformar eventos técnicos dispersos em informação operacional estruturada e acionável, permitindo acompanhar o ciclo de vida documental, reconstruir decisões do sistema e apoiar mecanismos de reparação, fallback, reindexação, validação e auditoria.

4.3. Tecnologias e Ferramentas

A seleção tecnológica procurou conciliar adequação funcional, reprodutibilidade experimental, execução local e integração com o ecossistema de inteligência artificial aplicado ao processamento documental. O protótipo foi implementado em Python 3.10+ e exposto através de uma API desenvolvida em FastAPI. A execução foi containerizada com Docker e Docker Compose, e os componentes de processamento, recuperação e geração foram suportados por Docling, ChromaDB, Ollama, SQLite e Pydantic. A Tabela 11 sintetiza as tecnologias e ferramentas utilizadas.

Tabela 11 - Tecnologias e ferramentas utilizadas no desenvolvimento do protótipo

Tecnologia	*Papel na solução*	*Justificação*	*Relação com o pipeline*
Python 3.10+	Linguagem principal de implementação	Ecossistema maduro para IA, APIs e processamento documental	Base de desenvolvimento dos serviços e módulos
FastAPI	Framework web para exposição de endpoints rest	Suporte a APIs assíncronas, tipagem e documentação openapi	Submissão de documentos, consulta de estado e interação via chat
Docker / Docker Compose	Containerização e orquestração dos serviços	Isolamento, consistência de execução e reprodutibilidade experimental	Execução coordenada da aplicação, chromadb e dependências
Docling	Motor principal de processamento documental	Conversão de documentos em representações estruturadas, incluindo suporte a PDF, tabelas, OCR e estrutura documental	Ingestão, extração e estruturação de conteúdo
Ollama	Execução local de modelos de linguagem, embeddings, visão e validação	Redução de dependência de APIs externas e maior controlo sobre os dados processados localmente	Geração de respostas, embeddings, análise visual, query rewriting e validação
ChromaDB	Base de dados vetorial	Armazenamento de embeddings com metadados e pesquisa por similaridade	Indexação e recuperação semântica de evidência documental
SQLite	Persistência leve local	Simplicidade, ausência de dependências externas e adequação a protótipo experimental	Histórico conversacional, estado operacional e registos auxiliares
Pydantic	Modelação e validação de dados	Definição de modelos e validação de estruturas usadas na API e no gémeo digital documental	Validação de dados, configuração e modelação das entidades do sistema

4.3.1. Python 3.10+ e FastAPI

O *backend* do sistema foi desenvolvido em Python 3.10+, beneficiando da maturidade do ecossistema para inteligência artificial, processamento documental, processamento de linguagem natural e integração de serviços. A adoção de FastAPI decorreu da sua adequação a aplicações baseadas em APIs assíncronas, da integração nativa com tipagem explícita através de Pydantic e da geração automática de documentação segundo a especificação OpenAPI (*FastAPI*, n.d.). Estas características são relevantes num sistema que combina processamento documental assíncrono com interação quase imediata através de endpoints de submissão, consulta e chat.

4.3.2. Docker e Docker Compose

A solução foi containerizada com Docker, de forma a assegurar isolamento do ambiente de execução, consistência entre diferentes máquinas e maior reprodutibilidade experimental. A orquestração por Docker Compose permitiu definir declarativamente os serviços principais, as suas dependências, volumes persistentes e parâmetros de configuração. Esta opção reduziu variações introduzidas pelo ambiente local e facilitou a repetição dos ensaios em condições controladas (Nüst et al., 2020).

4.3.3. Docling como motor de processamento documental

O Docling foi selecionado como motor principal de processamento documental por disponibilizar capacidades de extração estruturada que ultrapassam a lógica de OCR simples, incluindo identificação de hierarquia de conteúdo, deteção de tabelas e preservação de metadados relevantes do documento (Auer et al., 2024). A sua adoção permitiu concentrar o processamento num motor mais unificado, mantendo mecanismos de fallback para documentos cuja qualidade ou estrutura comprometa a extração adequada pelo motor principal.

4.3.4. Ollama e execução local de modelos

A execução dos modelos de linguagem foi suportada por Ollama, solução que permite correr modelos localmente sem dependência de APIs externas. Esta opção responde diretamente a requisitos de privacidade e reprodutibilidade, particularmente relevantes quando o sistema opera sobre documentação potencialmente sensível. No protótipo foram utilizados vários modelos com funções distintas e configuração independente.

O llama3.1:8b foi adotado como modelo principal de conversação, procurando equilibrar qualidade de geração e viabilidade de execução em contexto académico (Grattafiori et al., 2024). O mxbai-embed-large foi utilizado para a geração de embeddings de chunks documentais e de queries, dada a sua adequação a tarefas de recuperação semântica multilingue, incluindo português e inglês, conforme observado em avaliações de referência como o MTEB (Muennighoff et al., 2022). O llava-llama3 foi utilizado como modelo de visão, permitindo apoiar a análise de elementos visuais relevantes em documentos com imagens, gráficos e componentes não textuais (Liu et al., 2023). Adicionalmente, o phi3.5:latest foi utilizado na reformulação de perguntas (query rewriting), conforme descrito na secção 4.4.3. Por se tratar de uma tarefa auxiliar e leve, optou-se por um modelo pequeno e eficiente da família Phi-3, concebido para execução local com elevado desempenho relativo à sua dimensão, o que permitiu melhorar a recuperabilidade das perguntas sem sobrecarregar os recursos nem competir com o modelo principal de geração (Abdin et al., 2024).

4.3.5. ChromaDB como base de dados vetorial

A ChromaDB foi selecionada como base de dados vetorial do sistema devido à sua integração simples com Python, ao suporte de persistência em ambiente containerizado e à disponibilidade de mecanismos de pesquisa por similaridade adequados a cenários de recuperação documental. O seu papel na arquitetura consiste em armazenar os embeddings gerados para os chunks documentais e permitir a pesquisa eficiente de excertos semanticamente relevantes no momento da interação com o utilizador. Deste modo, a ChromaDB constitui o núcleo da componente de recuperação do sistema RAG, apoiando-se nos princípios de pesquisa vetorial por similaridade consolidados por bibliotecas como a

Faiss (Douze et al., 2024). É esta ligação entre a representação vetorial do corpus e a evidência fornecida ao modelo gerador que concretiza a lógica de geração aumentada por recuperação proposta por Lewis et al. (2021).

4.4. Implementação dos Módulos

A implementação seguiu a lógica modular definida na metodologia, evitando que cada componente fosse tratado como uma peça isolada. O objetivo foi construir um pipeline documental reprodutível, no qual a entrada, a transformação do documento, a recuperação de evidência, a geração da resposta e a validação pudessem ser observadas separadamente e analisadas em conjunto.

4.4.1. Ingestão, extração e fallback

O módulo de ingestão recebe documentos PDF, valida o ficheiro, atribui um identificador único e inicia o processamento em segundo plano. A extração principal é realizada com Docling, por permitir obter texto e estrutura documental de forma mais rica do que uma extração meramente linear. Quando o documento apresenta degradação, ausência de camada textual ou instabilidade estrutural, o sistema pode recorrer a OCR tradicional, extração nativa básica ou mecanismos de fallback. Esta estratégia permite lidar com documentos pesquisáveis, digitalizados e estruturalmente heterogéneos.

A decisão de preservar o texto bruto e, quando aplicável, gerar uma versão reparada antes da indexação é importante para a rastreabilidade. O sistema mantém memória do motor usado, das condições de extração e da eventual reparação aplicada, permitindo compreender se uma resposta foi condicionada por problemas de origem documental.

4.4.2. Segmentação e embeddings

Após a extração, o conteúdo é normalizado e segmentado em chunks. A segmentação privilegia fronteiras estruturais, como cabeçalhos e secções, com controlo de dimensão e sobreposição, de modo a preservar coerência semântica sem criar excertos excessivamente longos. Cada chunk é convertido em embedding por um modelo local e armazenado na ChromaDB com metadados de origem, secção, tipo e identificador documental.

Esta etapa é decisiva porque transforma o documento num conjunto de unidades recuperáveis. Uma má segmentação pode separar pergunta e resposta em fragmentos distintos ou misturar informação irrelevante no mesmo chunk. Por isso, a recuperação de evidência é avaliada de forma autónoma, antes da geração.

4.4.3. Recuperação, query rewriting e re-ranking

Quando o utilizador coloca uma pergunta, o sistema pode reformulá-la para melhorar a correspondência com o vocabulário documental. A recuperação inicial consulta a base vetorial e obtém uma janela alargada de candidatos. Em seguida, o re-ranking reordena esses candidatos, procurando priorizar os excertos mais relevantes para a pergunta. A seleção final limita o número de chunks enviados ao LLM, equilibrando cobertura e redução de ruído.

Esta separação entre recuperação inicial e re-ranking permite compreender o comportamento observado nos resultados: aumentar o número de chunks melhora o recall, mas pode reduzir a precision. O sistema foi, por isso, configurado para fornecer contexto suficiente ao modelo sem perder o controlo sobre a pertinência dos excertos.

4.4.4. Geração fundamentada e validação de grounding

A resposta é gerada a partir de um prompt estruturado que contém a pergunta, os excertos recuperados e, quando necessário, contexto conversacional recente. A temperatura reduzida privilegia estabilidade e aderência à evidência. Após a geração, uma etapa de validação verifica se a resposta se encontra suportada pelos excertos fornecidos. Quando a fundamentação não é confirmada, o sistema pode realizar nova tentativa de geração mais conservadora.

Este mecanismo não garante infalibilidade, mas torna explícito o princípio central da arquitetura: a resposta deve ser documentalmente justificável. A validação de grounding funciona, assim, como barreira adicional contra respostas genéricas ou não suportadas.

4.4.5. Gémeo Digital Documental e decisões adaptativas

A camada do Gémeo Digital Documental regista o ciclo de vida de cada documento: ingestão, extração, reparação, segmentação, indexação, recuperação, geração e validação.

Esta memória operacional permite reconstruir o percurso documento-resposta e suporta decisões adaptativas, como recomendação de reprocessamento quando a confiança de extração é baixa, ativação de reparação pós-OCR ou reindexação perante falhas parciais.

A sua função não é substituir OCR, RAG ou LLMs, mas tornar a articulação entre eles observável, auditável e interpretável. O seu objetivo é reforçar a rastreabilidade e a auditabilidade do processo, e não apenas gerar respostas automáticas. A Tabela 12 sintetiza os componentes tecnológicos do protótipo e as respetivas funções no sistema.

Tabela 12 - Síntese dos componentes tecnológicos do protótipo e respetivas funções

Componente	*Tecnologia*	*Função no sistema*
Backend e API	Python 3.10+ / FastAPI	Submissão de documentos, consulta de estado e interação conversacional.
Execução reprodutível	Docker / docker compose	Isolamento do ambiente e repetibilidade dos ensaios.
Processamento documental	Docling + OCR/fallback	Extração de texto, estrutura e tratamento de documentos digitalizados.
Modelos locais	Ollama	Geração de respostas, embeddings, validação e análise auxiliar sem envio externo de dados.
Base vetorial	Chromadb	Indexação de embeddings e recuperação semântica de excertos.
Monitorização	Sqlite + twinmanager	Registo de estados, eventos, métricas e decisões adaptativas.

4.5. Configuração e Reprodutibilidade

A configuração do sistema foi organizada de forma a separar os parâmetros dependentes da infraestrutura dos parâmetros funcionais do pipeline documental. Esta separação permitiu adaptar o protótipo a diferentes ambientes de execução, nomeadamente em contexto de container, sem alterar diretamente o código-fonte e mantendo condições controláveis para efeitos de avaliação.

A solução utilizou dois mecanismos principais de configuração. O primeiro correspondeu a variáveis de ambiente, usadas sobretudo para definir endereços de serviços externos e parâmetros associados ao módulo conversacional. O segundo correspondeu a um ficheiro de

configuração persistente, *app_config.json*, gerido pelo *ConfigurationManager*, onde ficaram concentradas as opções funcionais do pipeline documental, como o modo de processamento, a ativação de funcionalidades adicionais e limites específicos de execução.

Esta distinção foi importante para a reprodutibilidade experimental. As variáveis de ambiente permitiram adaptar a aplicação ao ambiente técnico em que era executada, enquanto o ficheiro de configuração persistente permitiu manter estáveis os parâmetros funcionais entre ensaios. Assim, tornou-se possível identificar com maior clareza se uma alteração observada no comportamento do sistema resultava de uma mudança de infraestrutura, de uma configuração do pipeline documental ou de um parâmetro associado à recuperação e geração.

De forma sintética, a configuração foi organizada em três grupos principais (Tabela 13):

Tabela 13 - Organização dos parâmetros de configuração do protótipo

Grupo de configuração	*Exemplos de parâmetros*	*Finalidade*
Infraestrutura	Endereço do ollama; endereço da chromadb	Permitir a ligação entre serviços em ambiente local ou container
Pipeline documental	Modo de processamento; ativação de análise visual; limites do processamento documental	Controlar o comportamento da ingestão, extração e preparação do documento
Recuperação e geração	Número de candidatos recuperados; número de excertos finais; modelos de geração, validação e reformulação	Controlar a amplitude da recuperação e o comportamento conversacional

Esta organização permitiu comparar configurações do protótipo de forma mais controlada. Por exemplo, os ensaios sobre diferentes modos de processamento documental puderam ser realizados alterando apenas a configuração do pipeline, enquanto os testes sobre recuperação puderam ser conduzidos ajustando parâmetros como o número de candidatos recuperados ou o número de excertos enviados ao modelo gerador.

Em termos de reprodutibilidade, a principal preocupação consistiu em manter explícitas as configurações utilizadas em cada ensaio. A separação entre infraestrutura, pipeline documental e parâmetros conversacionais reduziu a ambiguidade na interpretação dos

resultados, permitindo repetir testes em condições equivalentes e comparar o impacto de alterações específicas no comportamento do sistema.

Deste modo, a estratégia de configuração adotada conciliou flexibilidade operacional e controlo experimental. O sistema pôde ser ajustado a diferentes condições de execução sem comprometer a possibilidade de documentar, repetir e comparar os ensaios realizados sobre o protótipo.

A Tabela 14 sintetiza a configuração final do pipeline utilizada na avaliação experimental, consolidando os parâmetros de segmentação, recuperação e geração adotados.

Tabela 14 - Configuração final do pipeline utilizada na avaliação experimental

Parâmetro	*Valor adotado*
Dimensão máxima do chunk	≈1000 caracteres, com segmentação por fronteiras estruturais
Sobreposição entre chunks	≈100 CARACTERES
Candidatos iniciais recuperados	até 20 por variante da pergunta (1–6 variantes geradas por consulta), agregados antes do re-ranking
Excertos finais enviados ao modelo	6
Modelo de embeddings	mxbai-embed-large
Modelo de geração	LLAMA3.1:8B
Modelo de validação (grounding)	LLAMA3.1:8B (CHAT_JUDGE_MODEL)
Modelo de reformulação de perguntas	PHI3.5
Temperatura de geração	0,2 (0,0 na nova tentativa de geração)

4.6. Logging, Rastreabilidade e Auditabilidade

A arquitetura incorporou mecanismos de logging e rastreabilidade com o objetivo de registar, de forma estruturada, os eventos relevantes ocorridos ao longo do pipeline de processamento documental e geração de respostas. O registo é efetuado com um formato padronizado que inclui timestamp, nível de severidade e identificação do módulo de origem, permitindo uniformidade na observação do comportamento do sistema e facilitando a análise posterior dos eventos produzidos.

O logging cobre os principais momentos do fluxo operacional: motor de extração utilizado, tempo de processamento, número de chunks gerados e indexados, excertos recuperados em resposta a cada query, indicadores de relevância e classificação produzida pelo módulo de validação. Este nível de detalhe permite acompanhar o percurso de cada documento e de cada interação ao longo dos diferentes componentes do sistema.

A presença destes registos assume particular importância num protótipo orientado para análise e validação documental, uma vez que torna possível reconstruir o encadeamento entre entrada, processamento, recuperação e resposta. Esta capacidade de reconstrução tem valor técnico para depuração, mas também valor metodológico e organizacional, na medida em que suporta auditabilidade, explicabilidade operacional e controlo de qualidade.

Durante a fase de avaliação, a rastreabilidade permite correlacionar métricas quantitativas e qualitativas com características específicas dos documentos processados. Esta possibilidade facilita a identificação de padrões de falha, limitações associadas a determinados tipos documentais e degradação do desempenho em etapas específicas do pipeline. Deste modo, o logging deixa de ser entendido como simples registo operacional e passa a integrar a lógica de observabilidade do sistema.

5. RESULTADOS

Este capítulo apresenta e discute os resultados obtidos na avaliação experimental do protótipo. O desempenho de cada componente crítico do sistema foi analisado à luz das métricas definidas na metodologia, articulando-se a interpretação quantitativa com a leitura qualitativa do comportamento observado ao longo do pipeline.

5.1. Procedimento de Obtenção dos Resultados

Antes da apresentação das métricas, importa explicitar como os resultados foram obtidos, uma vez que a avaliação do protótipo não se limitou a demonstrações pontuais nem a uma apreciação meramente impressionista, tendo a extração e a recuperação assentado numa execução sistemática sobre a totalidade do corpus. O procedimento experimental foi estruturado para permitir observar cada etapa crítica do pipeline: primeiro a extração documental, depois a recuperação de evidência e, por fim, a geração da resposta. Esta separação foi essencial para evitar que uma falha observada na resposta final fosse atribuída automaticamente ao modelo de linguagem quando, na realidade, poderia ter origem no OCR, na segmentação, na indexação ou na recuperação.

O ponto de partida foi o corpus de 150 documentos sintéticos, construído para reproduzir cenários organizacionais heterogéneos sem recorrer a dados pessoais reais. Cada documento foi caracterizado quanto ao tipo técnico, presença de camada textual pesquisável, necessidade de OCR, presença de tabelas, elementos visuais, complexidade de layout e existência de campos críticos. Esta caracterização permitiu relacionar o desempenho do sistema com propriedades objetivas dos documentos, em vez de apresentar apenas uma média global descontextualizada.

Na etapa de extração, cada documento foi submetido ao módulo de processamento e registado na camada de monitorização documental. Para os 125 PDFs pesquisáveis, a camada textual nativa foi usada como referência do CER, contabilizando substituições, inserções e deleções. Os 25 documentos digitalizados foram processados, mas não foi preparada e alinhada uma transcrição de referência para medir o reconhecimento ótico; por

isso, não integram o cálculo do CER. Esta ausência limita a avaliação da extração precisamente nos documentos em que o OCR atua.

Na etapa de recuperação, foram usadas as cinco perguntas transversais definidas para cada documento, totalizando 750 perguntas. Para cada pergunta, existia uma resposta de referência e um excerto de suporte. O sistema recuperou excertos da base vetorial e estes foram avaliados de acordo com a sua proximidade à evidência esperada. As métricas Precision, Recall e F1-score foram calculadas para observar, respetivamente, a seletividade da recuperação, a cobertura da evidência relevante e o equilíbrio entre ambas. Foram ainda consideradas métricas por profundidade de recuperação, nomeadamente Precision@3, Recall@3, Precision@5, Recall@5, Precision@6 e Recall@6, porque o comportamento de um sistema RAG depende diretamente do número de chunks disponibilizados ao modelo gerador.

A avaliação da geração foi conduzida de forma qualitativa e comparativa. Para analisar o contributo real do RAG, foram comparadas respostas produzidas com evidência recuperada e respostas produzidas sem contexto documental. Esta comparação permitiu ilustrar se o sistema estava apenas a gerar texto plausível ou se, pelo contrário, estava a usar informação específica do documento. O critério central não foi a elegância textual da resposta, mas a sua aderência à evidência, especificidade, completude, clareza e utilidade operacional.

A camada do Gémeo Digital Documental funcionou como suporte transversal à avaliação. Ao registar estados, eventos e métricas, permitiu reconstruir o percurso de cada documento e perceber se uma resposta final resultou de uma extração estável, de uma recuperação suficiente ou de uma validação de grounding bem-sucedida. Esta camada também permitiu interpretar melhor os casos problemáticos, nomeadamente documentos com tabelas, layouts complexos ou divergência textual elevada. Deste modo, os resultados apresentados nas secções seguintes não são apenas valores agregados; representam evidência sobre o comportamento do pipeline enquanto sistema integrado.

Por fim, a discussão dos resultados foi orientada por uma leitura de ponta a ponta. Em vez de avaliar OCR, RAG e LLMs como módulos independentes e desligados entre si, analisou-se a forma como as decisões de extração e preparação documental influenciaram a

recuperação, e como a qualidade da recuperação condicionou a resposta gerada. Esta lógica responde diretamente ao problema de investigação formulado no Capítulo 1, ao analisar em que medida a arquitetura articula extração, recuperação, geração, validação e monitorização, e com que limitações.

5.2. Avaliação da Extração Documental

A avaliação da extração documental analisou a transformação dos PDFs em representações textuais e estruturais utilizáveis pela segmentação, indexação, recuperação e geração. A análise quantitativa incidiu exclusivamente nos 125 PDFs pesquisáveis. Os 25 documentos digitalizados foram processados pelo pipeline, mas não foram avaliados com uma referência textual alinhada; consequentemente, os resultados seguintes caracterizam extração e serialização nos PDFs pesquisáveis e não medem o desempenho do OCR nos documentos digitalizados.

Tabela 15 - CER por características estruturais dos documentos

Característica	*Categoria*	*Docs. com CER*	*CER médio*	*CER mediano*	*CER mínimo*	*CER máximo*	*Docs. com CER > 30%*
Tipo técnico	PDF pesquisável	125	12,41%	2,80%	1,71%	52,98%	26 / 20,8%
Tabelas	Sim	68	15,37%	3,08%	1,71%	52,12%	19 / 27,9%
Tabelas	Não	57	8,89%	2,78%	2,11%	52,98%	7 / 12,3%

A Tabela 15 apresenta a distribuição do CER em função das principais características estruturais dos documentos analisados. O CER foi calculado para os 125 PDFs pesquisáveis, usando como referência a camada textual nativa de cada documento.

Recorde-se que o CER resulta da comparação entre dois textos: a hipótese, correspondente ao texto produzido pelo sistema de extração, e a referência, correspondente ao texto considerado correto. Nos 125 documentos com CER calculável, o CER médio foi de 12,41%, enquanto o CER mediano foi de 2,80%. Esta diferença indica uma distribuição assimétrica: a maioria dos documentos apresentou valores baixos de divergência textual,

mas um subconjunto de casos com CER elevado aumentou a média global. Esta interpretação é reforçada pela existência de 26 documentos com CER superior a 30%, correspondendo a 20,8% dos documentos pesquisáveis. A presença de tabelas surge como o fator estrutural mais associado ao aumento do CER médio. Os documentos com tabelas apresentaram CER médio de 15,37%, face a 8,89% nos documentos sem tabelas. No entanto, as medianas mantiveram-se próximas, 3,08% e 2,78%, respetivamente. Isto sugere que as tabelas não degradaram uniformemente todos os documentos, mas contribuíram para a existência de casos extremos. Nos documentos com tabelas, 27,9% apresentaram CER superior a 30%, enquanto nos documentos sem tabelas essa proporção foi de 12,3%.

Importa precisar que, nos PDFs pesquisáveis, o Docling utiliza a camada textual existente, pelo que o CER avalia a extração e a serialização do texto, e não o reconhecimento ótico propriamente dito.

A classificação automática simples/médio/complexo foi retirada da análise comparativa porque resultava do número de chunks por página, uma variável produzida pelo próprio processamento, não tinha limiares metodológicos documentados e originava grupos fortemente desequilibrados (0, 6 e 119 documentos). As conclusões sobre extração passam, por isso, a apoiar-se apenas em características diretamente observáveis e discriminativas, como a presença de camada textual e de tabelas.

Esta tabela mostra que o CER médio global foi influenciado sobretudo por documentos com estruturas tabulares e por um subconjunto de casos com elevada divergência textual. A mediana global reduzida indica, contudo, que a extração foi estável na maioria dos PDFs pesquisáveis. Os 25 documentos digitalizados, excluídos desta análise, foram processados pelo sistema e integram a avaliação da recuperação.

5.3. Avaliação do Módulo de Recuperação (RAG)

A avaliação do módulo de recuperação teve como objetivo analisar a capacidade do sistema para identificar excertos documentais relevantes em resposta a diferentes tipos de perguntas. Esta etapa é central numa arquitetura RAG, uma vez que a qualidade da resposta

gerada pelo modelo de linguagem depende diretamente da pertinência da evidência recuperada e disponibilizada no contexto de geração.

A Tabela 16 apresenta os resultados da avaliação do módulo de recuperação para cinco grupos de perguntas, identificados como Q1 a Q5. Para cada grupo são apresentados os valores globais de Precision, Recall e F1-score, bem como as métricas P@3, R@3, P@5, R@5, P@6 e R@6, que permitem observar o comportamento do sistema quando são considerados os três, cinco e seis primeiros chunks recuperados. As colunas globais de Precision, Recall e F1 correspondem à configuração final do sistema, com k = 6, e coincidem por isso com P@6 e R@6.

Tabela 16 - Desempenho da recuperação por tipo de pergunta e top-k

Questão	*Precision*	*Recall*	*F1-score*	*P@3*	*R@3*	*P@5*	*R@5*	*P@6*	*R@6*
Q1	0,2002	0,4386	0,2578	0,2233	0,3177	0,2089	0,4128	0,2002	0,4386
Q2	0,2280	0,5131	0,2961	0,2833	0,3943	0,2409	0,4820	0,2280	0,5131
Q3	0,1980	0,4229	0,2540	0,2300	0,3060	0,2036	0,3918	0,1980	0,4229
Q4	0,3356	0,3938	0,3054	0,3556	0,2537	0,3376	0,3548	0,3356	0,3938
Q5	0,1493	0,5135	0,2119	0,1700	0,3630	0,1551	0,4877	0,1493	0,5135
Média	0,2222	0,4564	0,2650	0,2524	0,3269	0,2292	0,4258	0,2222	0,4564

As métricas apresentadas assumem valores entre 0 e 1, mas não existe neste estudo uma escala externa que permita classificar os valores como baixos, moderados ou elevados. A interpretação é, por isso, estritamente comparativa entre profundidades de recuperação e tipos de pergunta, considerando a granularidade dos chunks, o critério aproximado de relevância e a distribuição da evidência no corpus.

A Figura 4 complementa a Tabela 16, apresentando a evolução média da Precision@K e da Recall@K para k=3, k=5 e k=6. Conforme definido na metodologia, estes valores representam diferentes profundidades de recuperação: k=3 permite avaliar a qualidade dos primeiros excertos devolvidos, k=5 representa uma janela intermédia de contexto e k=6 corresponde à configuração final do pipeline RAG, em que até seis chunks são

disponibilizados ao modelo gerador. Esta representação permite visualizar de forma mais clara o compromisso entre seletividade e cobertura à medida que aumenta o número de chunks recuperados.

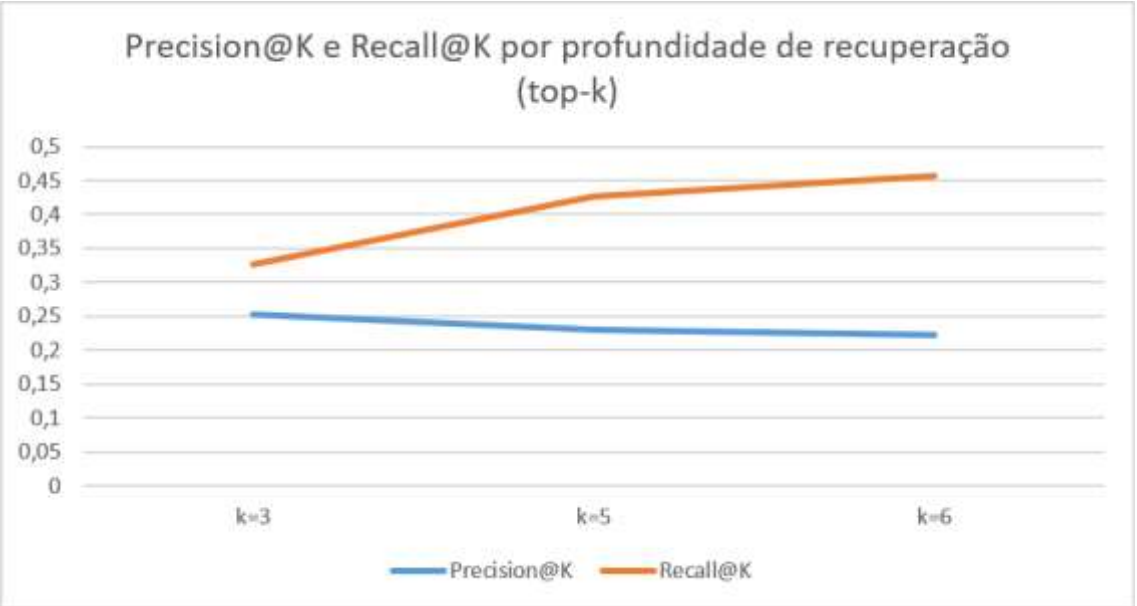

Figura 4 - Evolução média da Precision@K e da Recall@K por profundidade de recuperação (top-k)

Os resultados mostram que o aumento do número de chunks recuperados elevou progressivamente o Recall médio: R@3 = 0,3269, R@5 = 0,4258 e R@6 = 0,4564. Assim, no corpus e segundo o critério de relevância adotado, seis chunks recuperaram uma proporção maior da evidência relevante do que três ou cinco chunks. Estes valores não são classificados por uma escala absoluta externa.

Este ganho de cobertura foi acompanhado por uma redução da Precision média: P@3 = 0,2524, P@5 = 0,2292 e P@6 = 0,2222. Portanto, a configuração com mais contexto incluiu proporcionalmente mais chunks que não satisfaziam o critério automático de relevância. Na comparação entre grupos, Q4 apresentou a maior Precision (0,3356) e o maior F1 (0,3054), enquanto Q5 apresentou o maior Recall (0,5135) e a menor Precision (0,1493). A leitura é relativa ao corpus e não corresponde a categorias universais de desempenho.

Os valores de F1-score variaram entre 0,2119 e 0,3054. Estes resultados mostram que o aumento da cobertura não eliminou o compromisso com a seletividade e justificam investigar mecanismos de recuperação e ordenação mais discriminativos. Não é atribuída uma classificação absoluta aos valores.

Como o F1 é calculado por pergunta e depois agregado por média macro, pode ser inferior tanto à Precision como ao Recall médios: nas perguntas sem qualquer chunk relevante recuperado, o F1 é zero, e a média de valores harmónicos não coincide com a média harmónica das médias. É o que sucede na Q4, com F1 de 0,3054 face a Precision de 0,3356 e Recall de 0,3938.

Os resultados evidenciam um comportamento típico de sistemas RAG baseados em recuperação semântica: o aumento do *top-k* melhora a cobertura da evidência, mas reduz ligeiramente a seletividade. No contexto da arquitetura avaliada, o valor k=6 revelou-se adequado para maximizar a cobertura contextual fornecida ao modelo gerador, embora os valores de Precision indiquem a necessidade de mecanismos complementares de filtragem, re-ranking ou compressão de contexto.

5.4. Avaliação Qualitativa da Geração de Respostas

A avaliação qualitativa da geração de respostas teve como objetivo analisar o impacto do contexto recuperado pelo módulo RAG na precisão, especificidade e fundamentação das respostas produzidas pelo sistema. Para esse efeito, foram comparadas respostas geradas em duas condições: uma condição com RAG, na qual o modelo recebeu excertos documentais previamente recuperados, e uma condição sem RAG, na qual o modelo respondeu sem acesso ao contexto específico do documento.

A comparação incidiu sobre 10 perguntas, de seis documentos das três famílias documentais, cobrindo os cinco tipos de pergunta, e foi analisada pelo autor face à resposta de referência, segundo critérios de aderência à evidência, especificidade, completude e utilidade operacional. Trata-se de uma avaliação exploratória, sem avaliadores independentes nem métrica quantitativa, cujas conclusões não podem ser generalizadas às 750 perguntas. Apresentam-se a seguir dois exemplos representativos.

Um exemplo representativo foi a pergunta: "Qual é a finalidade principal do documento?", aplicada ao documento req_001.

- **Resposta do sistema sem RAG:** "A finalidade principal de um documento pode variar dependendo do tipo e do contexto em que foi criado. No entanto, algumas das principais funções dos documentos incluem: comunicar informações ou ideias, registar eventos ou transações, documentar decisões ou acordos..."
- **Resposta do sistema com RAG:** "A finalidade principal do documento é solicitar a concessão de licença de férias para Vera Santos, titular do número de colaborador COL-1409, pelo período de 13 dias úteis, com início em 18/10/2023 e término previsto em 31/10/2023."

Este exemplo ilustra de forma clara a principal vantagem da abordagem RAG: a capacidade de transformar uma pergunta genérica sobre um documento numa resposta específica, contextualizada e alinhada com a evidência documental recuperada. Enquanto a resposta sem RAG descreve funções gerais que qualquer documento poderia cumprir, a resposta com RAG identifica a finalidade concreta do documento analisado, incluindo o tipo de pedido, a pessoa associada, o período e as datas relevantes.

Um segundo exemplo correspondeu à pergunta: "Que datas, prazos, valores, códigos ou referências relevantes constam no documento?", aplicada ao documento rel_001.

- **Resposta do sistema sem RAG:** "Num documento, os dados relevantes podem incluir: data de emissão ou validade, prazo para pagamento ou entrega, valor monetário, códigos de barras, referências legais..."
- **Resposta do sistema com RAG:** "Data da inspeção: 23/12/2025 | Nível de conformidade: 77% | Prazo para resolução dos itens críticos: 7 dias úteis | Próxima inspeção: 21/06/2026."

Este exemplo reforça a diferença entre geração genérica e geração ancorada em evidência documental. Sem RAG, o sistema limita-se a enumerar categorias possíveis de informação, sem identificar os dados efetivamente presentes no documento. Com RAG, o sistema recupera e apresenta valores concretos, incluindo datas, percentagem de conformidade e

prazo de resolução, tornando a resposta mais útil para tarefas de análise e validação documental.

Na amostra exploratória de 10 perguntas, a comparação qualitativa sugeriu que o contexto recuperado tornou as respostas observadas mais específicas e verificáveis do que a condição sem acesso ao documento. Este resultado ilustra o funcionamento do mecanismo, mas não estima a taxa de respostas fundamentadas, a frequência de alucinações nem um efeito generalizável às 750 perguntas.

Importa ainda delimitar o papel da validação de grounding nesta comparação. As respostas foram geradas diretamente pelo modelo llama3.1:8b, com a recuperação e o re-ranking do sistema, mas sem passar pela etapa de validação. Por esse motivo, não existem contagens de respostas aprovadas, rejeitadas ou submetidas a nova tentativa, e a atuação deste mecanismo não foi quantificada.

5.5. Discussão dos Resultados

A avaliação realizada sustenta a viabilidade funcional do pipeline integrado: os documentos foram transformados em conteúdo processável e a recuperação foi avaliada sobre 750 perguntas. A geração foi observada apenas numa amostra exploratória de 10 perguntas, pelo que as conclusões sobre respostas mais específicas se restringem a essa amostra. A análise da extração mostrou ainda que estruturas tabulares podem aumentar a divergência de serialização e afetar as etapas seguintes.

No módulo de recuperação, o aumento de top-k elevou a cobertura da evidência e reduziu a seletividade, evidenciando o compromisso entre quantidade de contexto e ruído. Na amostra qualitativa de 10 perguntas, as respostas com contexto recuperado incorporaram dados específicos dos documentos, enquanto a condição sem documento produziu respostas genéricas. Esta observação ilustra a função da ancoragem documental, mas não demonstra a qualidade global da geração nem permite estimar o efeito do RAG nas 750 perguntas.

A camada de monitorização inspirada no Gémeo Digital Documental acrescentou valor ao permitir observar o percurso do documento ao longo do pipeline. Esta dimensão é particularmente relevante porque torna possível relacionar a resposta final com etapas

anteriores, como extração, segmentação, recuperação e validação. Deste modo, a arquitetura não se limita a produzir respostas, mas permite compreender em que condições essas respostas foram produzidas.

Quanto às decisões adaptativas, os registos do gémeo digital mostram que, na execução de avaliação, a reparação pós-OCR foi acionada nos 25 documentos digitalizados, e apenas nesses, após deteção de degradação no espaçamento do texto reconhecido. O reprocessamento orientado por confiança e a reindexação automática não foram acionados: nenhuma extração ficou abaixo do limiar de confiança definido, a confiança mínima registada foi de 0,91 e nenhuma indexação falhou. O valor acrescentado pela camada reside, assim, sobretudo na observabilidade do percurso e na capacidade de diagnóstico, e não num efeito medido sobre o desempenho. Um exemplo desta capacidade de diagnóstico: perguntou-se ao sistema quem era o autor de um documento extenso, e a resposta indicou que a informação não constava do documento. O registo do gémeo mostrou que a extração estava correta, o nome constava do texto extraído, e que o excerto com o nome do autor não tinha sido entregue ao modelo, situando a falha na recuperação, e não na extração nem na geração.

Em termos globais, os resultados indicam que a integração entre OCR, RAG e LLMs pode apoiar sistemas de análise documental mais explicáveis e rastreáveis. No entanto, também evidenciam que a robustez do sistema depende da melhoria contínua da extração estrutural, da segmentação e dos mecanismos de recuperação. Assim, a arquitetura desenvolvida deve ser entendida como uma base funcional e monitorizável, com potencial para aplicação em contexto organizacional, mas ainda com margem de evolução ao nível da validação da qualidade das respostas.

Em síntese, os resultados permitem concluir que o artefacto responde ao objetivo geral da investigação, ao demonstrar a viabilidade de uma arquitetura integrada para validação e análise de documentos digitais com evidência recuperada. Os objetivos específicos foram igualmente operacionalizados: a extração documental foi avaliada por características estruturais do corpus; a recuperação foi medida por métricas de precisão, recall e F1; a geração foi comparada em cenários com e sem RAG; e a camada de monitorização

documental reforçou a rastreabilidade do percurso documento-resposta. Persistem, contudo, limitações relevantes na validação quantitativa da geração, aspeto que sustenta as linhas de trabalho futuro apresentadas no capítulo seguinte.

6. CONCLUSÕES E TRABALHO FUTURO

Este capítulo final sintetiza as principais conclusões do trabalho, articulando-as com o problema de investigação e os objetivos definidos no Capítulo 1. Reconhecem-se as limitações do estudo e identificam-se direções de investigação futura coerentes com o protótipo desenvolvido e com a evidência experimental obtida.

6.1. Conclusões

A investigação desenvolvida respondeu ao problema formulado no Capítulo 1, demonstrando que é possível conceber, implementar e avaliar uma arquitetura de inteligência documental que articula OCR, recuperação semântica, RAG, LLMs e mecanismos de monitorização documental para apoiar a validação e a análise de documentos digitais em contexto organizacional.

Os resultados obtidos indicam que a qualidade da extração documental constitui um fator determinante para o desempenho global do sistema. A análise do CER evidenciou que a maioria dos PDFs pesquisáveis apresentou extração estável, mas também demonstrou que documentos com estruturas tabulares e layouts mais exigentes podem introduzir divergências textuais relevantes, com impacto subsequente na segmentação, indexação e recuperação de evidência.

No módulo de recuperação, verificou-se que o aumento do número de chunks recuperados melhorou a cobertura da evidência, mas reduziu a seletividade. Esta constatação confirma a necessidade de equilibrar recall e precision em arquiteturas RAG, uma vez que a disponibilização de mais contexto ao modelo gerador aumenta a probabilidade de incluir a evidência relevante, mas também pode introduzir ruído informacional.

Na amostra exploratória de 10 perguntas, as respostas produzidas com evidência recuperada foram mais específicas e alinhadas com o conteúdo dos documentos do que as respostas sem acesso ao documento. A dimensão reduzida da amostra, a avaliação por um único autor e a ausência de métricas quantitativas impedem generalizar esta observação ou quantificar o efeito do RAG.

A camada inspirada no conceito do Gémeo Digital Documental contribuiu para reforçar a observabilidade, a rastreabilidade e a auditabilidade do pipeline. Ao registar estados, eventos, métricas e decisões adaptativas, esta camada permitiu reconstruir o percurso entre documento original, extração, segmentação, recuperação, geração e validação, tornando o sistema mais transparente e tecnicamente controlável.

Conclui-se, assim, que a fiabilidade de sistemas de inteligência documental não depende exclusivamente da capacidade do modelo generativo, mas da coerência arquitetural entre extração, preparação do corpus, recuperação de evidência, geração fundamentada, validação e monitorização. O protótipo desenvolvido constitui uma base funcional e cientificamente justificada para evolução futura, sendo especialmente relevante para contextos organizacionais que exigem precisão, explicabilidade, proteção de dados e rastreabilidade na análise documental. Neste sentido, o contributo principal do trabalho é uma arquitetura de referência para o processamento documental inteligente em contexto organizacional, replicável e adaptável, que articula extração, recuperação, geração fundamentada e monitorização do ciclo de vida documental.

Este contributo deve ser lido à luz das limitações identificadas: a arquitetura não foi comparada com alternativas nem validada em contexto real.

6.2. Limitações

Apesar dos resultados obtidos, reconhecem-se as seguintes limitações, que delimitam o alcance das conclusões e enquadram o trabalho futuro:

- **Avaliação da geração limitada e qualitativa**: sendo a fundamentação em evidência a questão central do trabalho, a avaliação da geração constitui a sua principal limitação. Assentou numa comparação qualitativa de 10 perguntas, analisada pelo autor. Não permite, por isso, quantificar a taxa de respostas fundamentadas nem a ocorrência de alucinações. Acresce que a validação de grounding não foi exercida nesta comparação, pelo que a sua eficácia não foi medida.
- **Avaliação da extração limitada a documentos pesquisáveis:** o CER foi calculado apenas nos 125 PDFs pesquisáveis e mede sobretudo extração, serialização e ordem

de leitura. O desempenho do OCR nos 25 documentos digitalizados não foi avaliado com uma referência textual alinhada, apesar de o OCR atuar nesses documentos. Esta é uma limitação central da avaliação da extração e impede concluir sobre a qualidade do reconhecimento ótico.

- **Métricas de recuperação aproximadas:** na ausência de anotações de chunks relevantes no ground truth, a relevância foi determinada por correspondência textual aproximada com o campo de evidência, usando um limiar de 30% das palavras do excerto. Esta aproximação pode penalizar chunks de contexto adjacente úteis à geração. Nos documentos com seis chunks ou menos, o sistema devolve o documento inteiro e as métricas deixam de discriminar. Pelo critério adotado, cada pergunta tem, em média, 4,2 chunks relevantes, pelo que Precision@6 não está limitada a 1/6. Não foram calculadas métricas sensíveis à posição, como Hit@K ou MRR. Acresce que o ground truth foi gerado automaticamente, por regras fixas, e não foi validado por uma segunda pessoa.
- **Contributo individual dos componentes não quantificado**: a avaliação não isolou o ganho da reformulação de perguntas, do re-ranking, da reparação pós-OCR e do fallback face a uma configuração de referência, como recuperação vetorial simples ou BM25 sem re-ranking. Por esse motivo, não é possível atribuir a cada componente uma melhoria específica no desempenho. Na avaliação da recuperação, a reformulação de perguntas não foi acionada em nenhuma das 750 perguntas, por estas serem autónomas, e o fallback por confiança também não foi acionado.
- **Escalabilidade e recursos não testados em larga escala:** O comportamento do sistema com corpora na ordem dos milhares de documentos e em condições reais de utilização não foi testado. Acresce que a execução local dos modelos revelou competição por memória entre o modelo de visão e o modelo de embeddings, mitigada através do descarregamento proativo do modelo de visão antes da geração de embeddings, o que evidencia restrições de VRAM relevantes para o dimensionamento da solução.

- **Avaliação fora de contexto de produção**: O sistema foi avaliado em ambiente de laboratório, não tendo sido objeto de estudo de usabilidade com utilizadores finais nem de avaliação de impacto organizacional em cenário real.
- **Âmbito funcional limitado a um único documento:** a interação conversacional opera sobre o documento selecionado de cada vez, recuperando evidência apenas desse documento (filtragem por identificador). Na versão atual, o sistema não suporta consultas transversais a múltiplos documentos ou à totalidade do corpus em simultâneo. Esta delimitação corresponde a uma opção deliberada de âmbito: tendo o protótipo sido desenvolvido e avaliado em hardware pessoal, com capacidade de processamento e de memória limitadas, privilegiou-se a profundidade e a fiabilidade da análise por documento em detrimento da amplitude da consulta simultânea a múltiplos documentos. A interação centrada num único documento permitiu, além disso, avaliar de forma mais controlada a relação entre a evidência recuperada e a resposta gerada.
- **Ausência de memória entre sessões**: o histórico conversacional é mantido apenas ao nível da sessão, através de um ring buffer das últimas interações e de um resumo, com um período de retenção configurado em 24 horas. Não existe memória persistente associada ao utilizador entre sessões distintas, pelo que o contexto de interações anteriores não é retomado numa nova sessão. Esta opção decorre igualmente do foco do trabalho e das condições de execução local: o estudo centrou-se na fundamentação de cada resposta na evidência documental recuperada, e não na personalização do histórico do utilizador. A ausência de memória persistente reduz ainda a acumulação de dados pessoais entre sessões, em coerência com os princípios de minimização e limitação da finalidade adotados, tendo a memória ao nível da sessão sido considerada suficiente para a continuidade conversacional exigida pela análise documental.

6.3. Trabalho Futuro

A partir das conclusões e limitações identificadas, perspetivam-se as seguintes direções de investigação futura, articuladas com o estado atual do protótipo:

- **Melhoria da extração de estruturas tabulares:** Dado que as tabelas foram o principal fator de degradação do CER, justifica-se investigar estratégias de extração e representação que preservem a semântica tabular, reduzindo a divergência face à camada textual nativa e o ruído propagado às etapas de segmentação e recuperação.
- **Validação automática e quantitativa da geração:** Introduzir mecanismos de avaliação da qualidade das respostas, como juízes automáticos baseados em LLM (*LLM-as-judge*) e métricas de fidelidade à evidência, complementados por um estudo com avaliadores humanos, de forma a quantificar o grau de fundamentação e a detetar alucinações de forma sistemática sobre todo o corpus.
- **Avaliação comparativa por componente e anotação de referência**: comparar a configuração completa com configurações de referência, testando de forma controlada a remoção de cada componente, e anotar manualmente uma amostra de perguntas com os chunks relevantes, para calibrar o critério aproximado usado na avaliação e calcular métricas de ordenação como o Hit@k e o MRR.
- **Expansão e validação com documentação real e multilingue**: Substituir ou complementar o corpus sintético por documentação real, devidamente anonimizada, e estender o suporte ao processamento de documentos em múltiplos idiomas, requisito frequente em organizações com operação internacional.
- **Evolução da camada de gémeo digital para apoio à decisão adaptativa**: Aprofundar a exploração dos sinais recolhidos pela camada de monitorização para suportar decisões adaptativas, como o ajuste de parâmetros por documento, e estudar a sua utilidade num ambiente organizacional real, incluindo a usabilidade e o impacto na produtividade.
- **Consulta multidocumento e transversal ao corpus:** estender a recuperação e a geração a vários documentos em simultâneo, permitindo responder a perguntas que

cruzem informação de múltiplos documentos ou de todo o acervo organizacional, em vez de operar apenas sobre o documento selecionado.

- **Memória persistente entre sessões**: introduzir um mecanismo de memória de longo prazo, associado ao utilizador, que preserve o contexto e o histórico relevante entre sessões, mantendo as garantias de privacidade e conformidade adotadas pelo sistema.

BIBLIOGRAFIA

ADDIN ZOTERO_BIBL {"uncited":[],"omitted":[],"custom":[]} CSL_BIBLIOGRAPHY

Abdin, M., Aneja, J., Awadalla, H., Awadallah, A., Awan, A. A., Bach, N., Bahree, A., Bakhtiari, A., Bao, J., Behl, H., Benhaim, A., Bilenko, M., Bjorck, J., Bubeck, S., Cai, M., Cai, Q., Chaudhary, V., Chen, D., Chen, D., … Zhou, X. (2024). *Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone* (arXiv:2404.14219). arXiv. https://doi.org/10.48550/arXiv.2404.14219

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Auer, C., Lysak, M., Nassar, A., Dolfi, M., Livathinos, N., Vagenas, P., Ramis, C. B., Omenetti, M., Lindlbauer, F., Dinkla, K., Mishra, L., Kim, Y., Gupta, S., de Lima, R. T., Weber, V., Morin, L., Meijer, I., Kuropiatnyk, V., & Staar, P. W. J. (2024, August 19). *Docling Technical Report*. arXiv.Org. https://arxiv.org/abs/2408.09869v5

Baviskar, D., Ahirrao, S., Potdar, V., & Kotecha, K. (2021). Efficient Automated Processing of the Unstructured Documents Using Artificial Intelligence: A Systematic Literature Review and Future Directions. *IEEE Access*, *9*, 72894–72936. https://doi.org/10.1109/ACCESS.2021.3072900

Birk, W., Hostettler, R., Razi, M., Atta, K., & Tammia, R. (2022). Automatic generation and updating of process industrial digital twins for estimation and control-A review. *Frontiers in Control Engineering*, *3*. https://doi.org/10.3389/fcteg.2022.954858

Cohen, D., Burg, L., & Barkan, G. (2025). *RAGXplain: From Explainable Evaluation to Actionable Guidance of RAG Pipelines* (arXiv:2505.13538). arXiv. https://doi.org/10.48550/arXiv.2505.13538

Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.-E., Lomeli, M., Hosseini, L., & Jégou, H. (2024, January 16). *The Faiss library*. arXiv.Org. https://arxiv.org/abs/2401.08281v4

FastAPI. (n.d.). Retrieved 29 March 2026, from https://fastapi.tiangolo.com/

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2024). *Retrieval-Augmented Generation for Large Language Models: A Survey* (arXiv:2312.10997). arXiv. https://doi.org/10.48550/arXiv.2312.10997

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., … Ma, Z. (2024, July 31). *The Llama 3 Herd of Models*. arXiv.Org. https://arxiv.org/abs/2407.21783v3

Guan, S., Xu, C., Lin, M., & Greene, D. (2024). Effective Synthetic Data and Test-Time Adaptation for OCR Correction. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 15412–15425. https://doi.org/10.18653/v1/2024.emnlp-main.862

Habib, A., Abdulmahmod, O. F., Raza, M., Gu, Y., Aydogan, M., & Al-antari - Aisslab, M. A. (2025). *Towards Explainable AI in Agentic Retrieval-Augmented Generation: A Systematic Review*. 1–8. https://doi.org/10.1109/IDAP68205.2025.11222281

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, *28*(1), 75–105.

Iliuţă, M.-E., Moisescu, M.-A., Pop, E., Ionita, A.-D., Caramihai, S.-I., & Mitulescu, T.-C. (2024). Digital Twin-A Review of the Evolution from Concept to Technology and Its Analytical Perspectives on Applications in Various Fields. *Applied Sciences*, *14*(13), 5454. https://doi.org/10.3390/app14135454

Khan, A. A., Hasan, M. T., Kemell, K.-K., Rasku, J., & Abrahamsson, P. (2024). Developing retrieval augmented generation (RAG) based LLM systems from PDFs: An experience report (arXiv:2410.15944). arXiv. https://doi.org/10.48550/arXiv.2410.15944

Kuru, E., & Keskin, M. O. (2025). *Evaluating Modern RAG: Textual, Multimodal, Dense, and Late Interaction Pipelines*. monurkeskin.com. https://monurkeskin.com/assets/papers/rag_eval.pdf

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2021). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks* (arXiv:2005.11401). arXiv. https://doi.org/10.48550/arXiv.2005.11401

Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023, April 17). *Visual Instruction Tuning*. arXiv.Org. https://arxiv.org/abs/2304.08485v2

Loukil, F., Cadereau, S., Verjus, H., Galfre, M., Salamatian, K., Telisson, D., Kembellec, Q., & Le Van, O. (2024). LLM-centric pipeline for information extraction from invoices. *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*, 569–575. https://doi.org/10.1109/FLLM63129.2024.10852504

Manning, C., Raghavan, P., & Schuetze, H. (2009). *Introduction to Information Retrieval*.

Mokashi, A., Puthuparambil, B., Daniel, C., & Hanne, T. (2025). Analysis of Large Language Models for Company Annual Reports Based on Retrieval-Augmented Generation. *Information*, *16*(9), 786. https://doi.org/10.3390/info16090786

Muennighoff, N., Tazi, N., Magne, L., & Reimers, N. (2022, October 13). *MTEB: Massive Text Embedding Benchmark*. arXiv.Org. https://arxiv.org/abs/2210.07316v3

Mundlamuri, R., Gunnam, G. R., Mysari, N. K., & Pujuri, J. (2025). The Evolution of AI: From Classical Machine Learning to Modern Large Language Models. *IEEE Access*, *13*, 178302–178341. https://doi.org/10.1109/ACCESS.2025.3621344

National Academies of Sciences, E., Engineering, N. A. of, Studies, D. on E. and L., Sciences, D. on E. and P., Climate, B. on A. S. and, Sciences, B. on L., Board, C. S. and T., Statistics, C. on A. and T., Analytics, B. on M. S. and, & Twins, C. on F. R. G. and F. D. for D. (2024). The Digital Twin Landscape. In *Foundational Research Gaps and Future Directions for Digital Twins*. National Academies Press (US). https://www.ncbi.nlm.nih.gov/books/NBK605499/

Nüst, D., Sochat, V., Marwick, B., Eglen, S. J., Head, T., Hirst, T., & Evans, B. D. (2020). Ten simple rules for writing Dockerfiles for reproducible data science. *PLOS Computational Biology*, *16*(11), e1008316. https://doi.org/10.1371/journal.pcbi.1008316

Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … McKenzie, J. E. (2021). PRISMA 2020 explanation and

elaboration: Updated guidance and exemplars for reporting systematic reviews. *BMJ*, *372*, n160. https://doi.org/10.1136/bmj.n160

Panda, M., & Mukherjee, S. (2025). Enhancing Privacy and Security in Rag-Based Generative AI Applications. *AI, Machine Learning and Applications Advances 2025*, 01–10. https://doi.org/10.5121/csit.2025.150301

Paoletti, V. (2025). *AI-Powered Document Intelligence with Retrieval-Augmented Generation*. webthesis.biblio.polito.it. https://webthesis.biblio.polito.it/37167/

Peffers, K., Tuunanen, T., Rothenberger, M., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, *24*, 45–77.

Ramesh Pingili. (2025). AI-driven intelligent document processing in government and public administration. *World Journal of Advanced Engineering Technology and Sciences*, *15*(3), 1316–1326. https://doi.org/10.30574/wjaets.2025.15.3.1059

Regulamento (UE) 2016/679 do Parlamento Europeu e do Conselho, de 27 de abril de 2016, relativo à proteção das pessoas singulares no que diz respeito ao tratamento de dados pessoais e à livre circulação desses dados (Regulamento Geral sobre a Proteção de Dados). (2016). *Jornal Oficial da União Europeia*, L 119, 1–88. https://eur-lex.europa.eu/eli/reg/2016/679/oj

Smith, B., & Troynikov, A. (2024, June 3). *Evaluating Chunking Strategies for Retrieval*. https://research.trychroma.com/evaluating-chunking

Sobhan, S., & Haque, M. A. (2025). LLM-Assisted Question-Answering on Technical Documents Using Structured Data-Aware Retrieval Augmented Generation. *arXiv Preprint arXiv:2506.23136*. https://arxiv.org/abs/2506.23136

Xu, D., Chen, W., Peng, W., Zhang, C., Xu, T., Zhao, X., Wu, X., Zheng, Y., Wang, Y., & Chen, E. (2024). Large language models for generative information extraction: A survey. *Frontiers of Computer Science*, *18*(6), 186357. https://doi.org/10.1007/s11704-024-40555-y

Zhang, J., Zhang, Q., Wang, B., Ouyang, L., Wen, Z., Li, Y., Chow, K.-H., He, C., & Zhang, W. (2025). *OCR Hinders RAG: Evaluating the Cascading Impact of OCR on Retrieval-Augmented Generation* (arXiv:2412.02592). arXiv. https://doi.org/10.48550/arXiv.2412.02592

Zhang, N., Vergara-Marcillo, C., Diamantopoulos, G., Shen, J., Tziritas, N., Bahsoon, R., & Theodoropoulos, G. (2024). *Large Language Models for Explainable Decisions in Dynamic Digital Twins* (arXiv:2405.14411). arXiv. https://doi.org/10.48550/arXiv.2405.14411

Zhou, P., Feng, Y., & Yang, Z. (2025). *Privacy-Aware RAG: Secure and Isolated Knowledge Retrieval* (arXiv:2503.15548). arXiv. https://doi.org/10.48550/arXiv.2503.15548