UNIVERSIDADE FEDERAL DO TOCANTINS

PROGRAMA DE PÓS-GRADUAÇÃO EM GOVERNANÇA E TRANSFORMAÇÃO DIGITAL (PPGGTD)

Bruno Mortari

UMA JORNADA DE TRANSFORMAÇÃO DIGITAL NO SERVIÇO PÚBLICO: RUMO AO AGÊNTICO

Uma pesquisa focada na otimização de processos de trabalho

Palmas – Tocantins
2026

Bruno Mortari

UMA JORNADA DE TRANSFORMAÇÃO DIGITAL NO SERVIÇO PÚBLICO: RUMO AO AGÊNTICO

Uma pesquisa focada na otimização de processos de trabalho

Dissertação apresentada ao Programa de Pós-Graduação em Governança e Transformação Digital (PPGGTD) da Universidade Federal do Tocantins, como requisito parcial para obtenção do título de Mestre.

Orientador: Prof. Gentil Veloso Barbosa, Dr.
Coorientador: Prof. Rafael Lima de Carvalho, Dr.

Palmas – Tocantins
2026

Dados Internacionais de Catalogação na Publicação (CIP)
Sistema de Bibliotecas da Universidade Federal do Tocantins

M887j Mortari, Bruno.

Uma jornada de transformação digital no serviço público: rumo ao agêntico: Uma pesquisa focada na otimização de processos de trabalho. / Bruno Mortari. – Palmas, TO, 2026.

55 f.

Dissertação (Mestrado Profissional) - Universidade Federal do Tocantins – Câmpus Universitário de Palmas - Curso de Pós-Graduação (Mestrado Profissional) em Governança e Transformação Digital - PPGGTD, 2026.

Orientador: Gentil Veloso Barbosa

Coorientador: Rafael Lima de Carvalho

1. Inteligência artificial. 2. Agentes de IA. 3. Transformação digital. 4. Automação de processos. I. Título

CDD 004

Elaborado pelo sistema de geração automática de ficha catalográfica da UFT com os dados fornecidos pelo(a) autor(a).

Bruno Mortari

UMA JORNADA DE TRANSFORMAÇÃO DIGITAL NO SERVIÇO PÚBLICO: RUMO AO AGÊNTICO

Uma pesquisa focada na otimização de processos de trabalho

Dissertação apresentada ao Programa de Pós-Graduação em Governança e Transformação Digital (PPGGTD) da Universidade Federal do Tocantins, como requisito parcial para obtenção do título de Mestre.

Data da Aprovação: 04/03/2026
BANCA EXAMINADORA:

Prof. Gentil Veloso Barbosa, Dr.
Orientador – Universidade Federal do Tocantins

Prof. Sergio Manoel Serra da Cruz, Dr.
UFRJ

Prof. Marluce Evangelista Carvalho Zacariotti, Dra.
UFT

Este mestrado representa muito mais do que aquisição de conhecimento técnico. Foi uma jornada de transformação pessoal e profissional, construída com o apoio inestimável de pessoas e instituições que acreditaram neste projeto.

Cada linha desta dissertação carrega a contribuição de mestres, colegas, familiares e instituições que tornaram este sonho possível. Por e para Deus. A todos, minha profunda gratidão.

RESUMO

MORTARI, Bruno. **Uma Jornada de transformação digital no serviço público: rumo ao agêntico**: Uma pesquisa focada na otimização de processos de trabalho. Dissertação (Mestrado) – Programa de Pós-Graduação em Governança e Transformação Digital (PPGGTD), Universidade Federal do Tocantins, Palmas, 2026.

Este trabalho apresenta uma pesquisa sobre a implementação de soluções com uso de Inteligência Artificial no contexto do serviço público brasileiro, como força motriz da transformação digital. A partir da convergência entre o alto nível de padronização dos processos de trabalho, a disponibilidade de grandes volumes de dados, poder computacional acessível e o avanço dos modelos de linguagem, acreditamos que o momento atual é o cenário ideal para a proliferação de agentes inteligentes. Com o formato de pesquisa aplicada, foram desenvolvidos protótipos funcionais, incluindo sistemas para apoio à licitação, auditoria, geração de atas, checagem de fatos, além de chatbots especializados, com o objetivo de provar conceitos. Em complemento, foi desenvolvido um classificador para textos sintéticos a partir da técnica de ajuste fino; criação e compartilhamento de bases de dados e estudos teóricos voltados ao tema. A pesquisa demonstra a viabilidade técnica e o valor potencial desses sistemas, corroborando a ideia de que é possível desenvolver sistemas agênticos eficazes em conformidade com o marco regulatório proposto pelo CNJ. Os resultados indicam ganhos significativos em eficiência operacional e qualidade dos serviços, mantendo o equilíbrio entre a autonomia dos agentes e a supervisão humana, constituindo ferramenta poderosa para otimizar atos administrativos e judiciais.

Palavras-chave: Inteligência Artificial; Agentes de IA; Transformação Digital; Serviço Público; Justiça Eleitoral; Automação de Processos.

ABSTRACT

MORTARI, Bruno. **AI Agents at the Center of Digital Transformation**: An applied research focused on work process optimization. Dissertation (Master's Degree) – Programa de Pós-Graduação em Governança e Transformação Digital (PPGGTD), Universidade Federal do Tocantins, Palmas, 2026.

This work presents research on the implementation of solutions using Artificial Intelligence in the context of the Brazilian public service, as a driving force for digital transformation. Based on the convergence of the high level of standardization of work processes, the availability of large volumes of data, accessible computing power, and the advancement of language models, we believe that the current moment is the ideal scenario for the proliferation of intelligent agents. Using an applied research format, functional prototypes were developed, including systems to support bidding, auditing, minutes generation, fact-checking, and specialized chatbots, with the aim of proving concepts. In addition, a classifier for synthetic texts was developed using fine-tuning techniques; databases were created and shared; and theoretical studies focused on the topic were conducted. The research demonstrates the technical feasibility and potential value of these systems, corroborating the idea that it is possible to develop effective agentic systems in accordance with the regulatory framework proposed by the CNJ (National Council of Justice). The results indicate significant gains in operational efficiency and service quality, maintaining a balance between agent autonomy and human supervision, constituting a powerful tool for optimizing administrative and judicial actions.

Keywords: Artificial Intelligence; AI Agents; Digital Transformation; Public Service; Electoral Justice; Process Automation.

LISTA DE ABREVIATURAS E SIGLAS

IA Inteligência Artificial
LLM Large Language Model (Modelo de Linguagem de Grande Escala)
CNJ Conselho Nacional de Justiça
TRE-GO Tribunal Regional Eleitoral de Goiás
TSE Tribunal Superior Eleitoral
RPA Robotic Process Automation
ML Machine Learning
RL Reinforcement Learning
DFD Documento de Formalização da Demanda
PP Pesquisa de Preço
MR Matriz de Riscos
ETP Estudo Técnico Preliminar
TR Termo de Referência
ED Edital
PNCC Portal Nacional de Contratação Pública
BI Business Intelligence
RAG Retrieval-Augmented Generation
API Application Programming Interface
PDF Portable Document Format
HTML HyperText Markup Language
JSON JavaScript Object Notation
PEP-8 Python Enhancement Proposal 8
ISO International Organization for Standardization
LGPD Lei Geral de Proteção de Dados
ODS Objetivos de Desenvolvimento Sustentável
PCA Plano de Contratações Anual
SPAR Sentir, Planejar, Agir, Refletir
CoT Chain of Thought
ToT Tree of Thoughts
CoVe Chain of Verification
MoE Mixture of Experts
ACT Análise Cognitiva de Tarefas
A2A Agent-to-Agent
PBIA Plano Brasileiro de Inteligência Artificial
EFGD Estratégia Federal de Governo Digital
FAIR Facebook AI Research

PoC Proof of Concept/Prova de Conceito

NIST National Institute of Standards and Technology

ONU Organização das Nações Unidas

SUMÁRIO

1 Introdução **10**

2 Delimitação do Problema **13**
2.1 Justificativa 14
2.2 Objetivos 15
2.2.1 Objetivo Principal 15
2.2.2 Objetivos Secundários 15
2.3 Metodologia 16
2.3.1 O Laboratório de Inovação como Campo de Pesquisa 16
2.3.2 Primeiro Pilar: Design Thinking 16
2.3.3 Segundo Pilar: Análise Cognitiva de Tarefas (ACT) 18
2.3.4 Processo Consolidado de Desenvolvimento 20
2.3.5 Procedimentos de Avaliação 21
2.3.6 Ferramentas computacionais e uso de IA 22

3 Fundamentação Teórica **23**
3.1 Definição e Evolução dos Agentes de IA 23
3.2 Ciclo de Vida SPAR - Sentir, Pensar, Agir e Refletir 24
3.3 Estrutura e Componentes Essenciais 25
3.4 Conceitos Fundamentais 26
3.4.1 Autonomia: O Grau de Independência 26
3.4.2 Agência: Intencionalidade e Poder de Escolha 26
3.4.3 Auto-Governança: Controle e Adaptação 26
3.5 Classificação dos Agentes de IA 27
3.5.1 por arquitetura 27
3.5.2 por capacidade 27
3.5.3 por raciocínio 27
3.6 Mudança do Paradigma 28
3.7 Engenharia de Prompt para agentes de IA 29
3.7.1 Componentes do prompt 29
3.7.2 Técnicas de prompt 30
3.8 Engenharia de Contexto: Geração aumentada por recuperação 31
3.8.1 Técnicas RAG 31
3.9 Resolução 615/CNJ e o Marco Regulatório 33

4 Transformação Digital: rumo ao agêntico **35**
4.1 Ecossistema Nativa IA 35
4.2 Sistemas agênticos 36
4.2.1 LIA: Licitação com IA 36
4.2.2 Agente de Auditoria 36
4.2.3 IATA: Geração de Atas 38
4.2.4 Checagem de Fatos 39
4.3 Chatbots 41
4.3.1 Normativos do CNJ 41
4.3.2 PalanqueIA 42
4.4 Oficina Prática 43
4.5 Ajuste fino 43
4.5.1 Classificador entre Textos Sintéticos e Textos Humanos 43
4.5.2 Classificador Federado para Classes Judiciais 44
4.5.3 Despesas de Campanha e Resultado das Eleições 45
4.5.4 Democracia Inclusiva: Votos dos Eleitores Analfabetos e com Mobilidade Reduzida 45
4.5.5 Resolução 615/2025 do CNJ e Agenda 2030 46
4.5.6 Estudo de normas internacionais para uma governança de IA segura e responsável nas organizações 46
4.6 Discussão 46
4.7 Limitações e Trabalhos Futuros 47

5 Conclusão **49**

Referências **50**

Capítulo 1

Introdução

A administração pública brasileira atravessa a fase preparatória de uma profunda transformação onde seus processos de trabalho, hoje altamente dependentes do esforço humano, serão migrados para uma gestão apoiada por agentes inteligentes. Com o avanço da inteligência artificial (IA), o processo produtivo será migrado para o sistema agêntico, em que os mesmos percebem seu ambiente, planejam ações, executam tarefas e avaliam seu resultado, em um ciclo orientado ao alcance de seu objetivo.

O termo modelo agêntico é considerado uma evolução na engenharia de IA Sapkota et al., 2025, que distingue três paradigmas progressivos: a IA Generativa, os Agentes de IA e o sistema agêntico.

Enquanto a IA Generativa se limita a produzir conteúdo em resposta a prompts, sem autonomia ou capacidade de ação sobre o ambiente, e os Agentes de IA avançam ao integrar ferramentas externas para executar tarefas específicas, o sistema agêntico apresenta uma estrutura em que múltiplos agentes especializados colaboram para alcançar objetivos complexos, por meio de decomposição dinâmica de tarefas, comunicação interagentes, memória persistente e raciocínio reflexivo.

Essa transformação inicia o surgimento de uma nova era na produtividade institucional. A tecnologia deixa de ser uma ferramenta simplesmente auxiliar e secundária, e passa para uma posição de proatividade dentro da rotina de trabalho. Estabelece-se a junção entre pessoas e máquinas, em que os programas de computador estão integrados aos processos de trabalho, com determinado nível de autonomia, articulando planos e executando ações, e os humanos passam a ter um papel de supervisão e de tomada de decisões estratégicas.

Para compreender essa mudança, podemos pensar na transformação de materiais, em que a reação química depende da presença de alguns fatores ajustados: reagentes, temperatura e um catalisador. O resultado não é um acidente, mas uma consequência das leis da natureza.

Essa analogia reflete uma grande mudança nos dias atuais. A transformação digital ocorrerá na presença de quatro fatores: a alta padronização dos processos de trabalho, a disponibilidade massiva de dados institucionais, o poder computacional acessível e o avanço dos modelos de linguagem em grande escala (LLMs).

Nestes anos de emprego público, atuando como analista de dados e estatístico do quadro da Justiça Eleitoral, pude observar inúmeros processos rotineiros que consomem demasiado tempo e esforço humano. Tarefas repetitivas, elaboração de relatórios burocráticos, além de

preenchimento de formulários desnecessários. Esse ambiente consumido pela burocracia cria um cenário ideal para a automação inteligente.

A maneira com que os modelos de linguagem agregam valor em uma instituição é por meio do agenciamento. Enquanto tais modelos são capazes de solucionar problemas, os sistemas agênticos os integram em nossas atividades: percebem o ambiente, escolhem caminhos, acionam ferramentas e acompanham o processo até que a tarefa esteja concluída.

Esta dissertação é o meu esforço para documentar, analisar e desenvolver parte desse novo paradigma, investigando como os funcionários públicos podem manusear essa nova tecnologia com sabedoria, responsabilidade e propósito.

A ideia central do estudo é a de que sistemas inteligentes bem projetados, considerando sua estrutura, sua lógica de trabalho, ferramentas disponíveis, sua memória e demais componentes, devem contribuir para a eficiência pública e a celeridade processual, sem comprometer a conformidade e ética legal.

Em relação à implantação técnica, essa transformação já está acontecendo dentro de algumas organizações. Relatório realizado pela *Langchain* (LangChain, 2024), pouco mais de 50% das organizações entrevistadas já utilizam agentes de IA em ambiente de produção. Essa constatação marca uma evolução dos sistemas determinísticos em agênticos capazes de executar complexas tarefas. A tendência atual é de supervisão humana e sistemas multi-agentes, onde o usuário permanece no controle das principais decisões e múltiplos agentes atuam em conjunto e colaborativamente para aperfeiçoar o sistema.

O mercado de agentes de IA autônomos deixou de ser um sonho distante. Segundo o estudo (Prophecy Market Insights, 2025), estamos no início de uma curva de crescimento agressiva: avaliado em cerca de US$ 6,9 bilhões em 2024, o setor deve atingir US$ 101,8 bilhões até 2035. As empresas buscam autonomia capaz de operar 24/7, tomando decisões em tempo real e liberando o potencial humano para tarefas mais estratégicas.

Essa aceleração fica visualmente evidente ao analisarmos a trajetória de crescimento do mercado. O gráfico da *Grand View Research* ilustra um comportamento exponencial: após um período de crescimento lento entre 2018 e 2023, vemos uma inflexão acentuada a partir de 2024. As projeções indicam que o mercado atingirá a marca de US$ 50 bilhões já em 2030, sugerindo que estamos saindo da fase de experimentação para a adoção em massa (Grand View Research, 2024).

Este estudo inicia-se com a delimitação do problema, relatando que o serviço público é marcado por tarefas que demandam demasiado tempo e esforço humano. Segue a justificativa, descrevendo o uso dos sistemas agênticos como parte da solução para otimizar recursos e aperfeiçoar a eficiência operacional; quanto aos objetivos da pesquisa, os mesmos se relacionam com o desenvolvimento de uma metodologia para produção de sistemas agênticos e posterior avaliação.

A seção de metodologia descreve os passos práticos que seguimos, incorporando a técnica do *Design Thinking* e análise cognitiva de tarefas (ACT) em uma prática para desenvolvimento

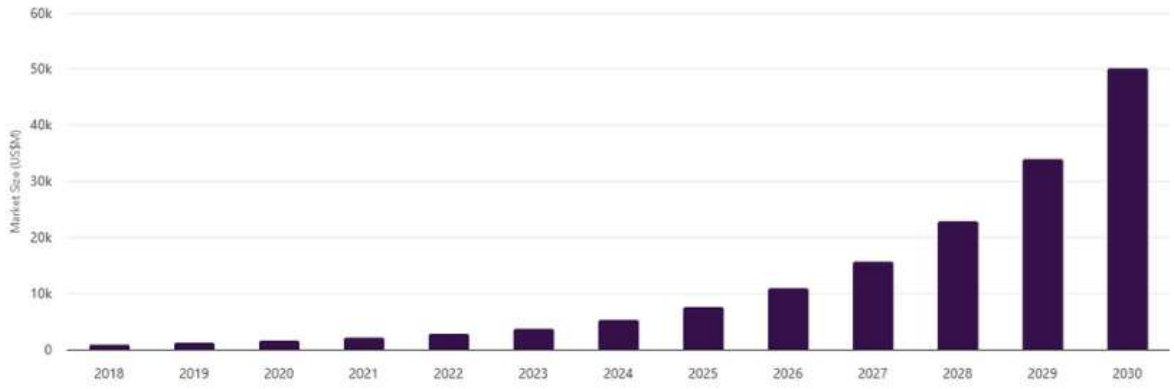

Figura 1.1: Evolução do mercado de agentes de IA

de tais sistemas.

Em outro capítulo, segue a fundamentação teórica da pesquisa, na qual se detalham o que são agentes de IA, sua estrutura, engenharia de *prompt* e de contexto. Nesse escopo, é feita a diferenciação entre sistemas com uso de IA e sistemas agênticos, os quais são proativos, autônomos e autodidatas. Por fim, é considerado um estudo sobre governança de IA, analisando o marco regulatório de IA no Poder Judiciário Brasileiro.

A discussão dos resultados analisa criticamente o desempenho desses protótipos, avaliando a precisão alcançada e sua real vantagem à luz do interesse da administração. Por fim, a conclusão sintetiza os achados, oferece uma reflexão sobre algumas implicações deste trabalho e um cenário sobre o futuro da colaboração entre humanos e IA, sob a perspectiva de um serviço público mais eficiente.

Capítulo 2

Delimitação do Problema

O cotidiano do serviço público brasileiro é marcado por atividades repetitivas operacionais que demandam tempo e esforço humano. Tais atividades variam desde tarefas simples até com alto grau de complexidade, incluindo análise de documentos, formulários, cliques em sistemas, relatórios padronizados, e outros. Apesar dos avanços tecnológicos, os processos de trabalho ainda permanecem majoritariamente manuais, exigindo sobremodo esforço humano.

Apesar da administração pública apresentar avanços na área de gestão de processos e desburocratização do serviço público, mapeando e melhorando os processos, e por mais que sistemas de informação auxiliem a prestação de tais processos, ainda assim observa-se uma alta dependência do fator humano, uma vez que grande parte das atividades permanece centrada em análises manuais, conferências repetitivas, produção textual e tomada de decisões operacionais de baixa complexidade, mas de alto volume.

Nós não eliminamos a burocracia; nós apenas a digitalizamos. O fato é que os sistemas de informação tradicionais, baseados em regras rígidas e fluxos determinísticos, atingiram o seu limite.

Por outro lado, o campo da IA apresentou evolução quando se trata de execução de diversas tarefas, desde as mais simples até raciocínio matemático a nível olímpico. O treinamento de modelos específicos de pensamento profundo e planejamento de tarefas, os conhecidos modos *thinking*, tem permitido realizar exercícios de planejamento no contexto dos agentes. Além disso, a abundância de dados voltados à programação refinou a capacidade dessas ferramentas de gerar e executar códigos, tornando-as aptas a manipular sistemas e bases de dados com precisão.

Esse cenário mostra um diferencial de duas forças, sugerindo uma ruptura no modelo antigo para um novo paradigma no serviço público: a implantação dos sistemas agênticos para compor os recursos tecnológicos organizacionais, além de apoiar a força de trabalho na execução de suas rotinas.

Logo, surge a seguinte pergunta: Como romper com o modelo antigo e migrar o serviço público para um modelo agêntico — onde sistemas possam atuar em tarefas de forma autônoma, integrando-se proativamente à rotina de trabalho?

2.1 Justificativa

Por muitos anos, a transformação digital se baseou em desenvolvimento de sistemas tradicionais buscando a digitalização de nossas atividades ocupacionais. No entanto, a parte burocrática e repetitiva do trabalho continuava presente. Se os sistemas determinísticos chegaram ao seu limite, a saída não é construir softwares com mais regras — é mudar o paradigma.

A entrada na era dos agentes representa um salto evolutivo em comparação aos softwares antigos. Não estamos mais lidando apenas com modelos que preveem a próxima palavra, mas com estruturas projetadas para alcançar missões complexas em ambientes que se modificam constantemente.

Vivemos o momento ideal dessa nova transformação digital: 4 fatores convergem pela primeira vez: alta padronização dos processos, disponibilidade massiva de dados, poder computacional acessível e a maturidade dos LLMs.

Nesta perspectiva, conceitos como "planejamento"e "raciocínio"estabelecem um novo paradigma de computação. Os agentes conseguem quebrar problemas em etapas executáveis, avaliando caminhos alternativos e refletindo sobre erros antes de prosseguir. Tais sistemas ultrapassam os limites de dados estáticos ao se conectarem a APIs, bancos de dados e navegadores web, desbloqueando uma grande classe de tarefas, anteriormente exclusivas ao humano, como a execução de um código ou o agendamento de uma reunião. (Grootendorst & Alammar, 2025; Huyen, 2024)

Além disso, o desenvolvimento de sistemas agênticos no serviço público brasileiro ganha tração com as atuais diretrizes nacionais, fundamentado pelo Decreto nº 12.198/2024 (Brasil, 2024), que institui a Estratégia Federal de Governo Digital (EFGD 2024-2027) e define como princípio um "Governo inteligente e inovador", baseado no uso de dados e tecnologia visando a desburocratização.

A convergência tecnológica citada preliminarmente neste trabalho alinha-se à vontade governamental de implantar núcleos e laboratórios para soluções com IA aplicada a políticas públicas, além da disponibilização de infraestrutura adequada.

O movimento ganha ainda mais força com o Plano Brasileiro de Inteligência Artificial (PBIA), que prevê investimentos na ordem de R$ 1,76 bilhão especificamente para a melhoria dos serviços públicos, validando a ideia de que o momento é ideal para a proliferação de serviços assistidos por IA. (Ministério da Ciência, Tecnologia e Inovação, 2024)

Na seara da Justiça, o uso dessa tecnologia recebeu o aval da recente Resolução CNJ nº 615/2025 (Conselho Nacional de Justiça, 2025), que institucionaliza seu uso para promover a inovação tecnológica e a eficiência dos serviços, desde que observados os princípios de transparência, ética e supervisão humana. A regulamentação valida expressamente o uso de modelos de linguagem como ferramentas auxiliares de apoio à gestão e à decisão, auxiliando magistrados e servidores na utilização de tais tecnologias.

A confiança nesses agentes são reforçados pela exigência de normas técnicas, como auditabi-

lidade e explicabilidade, essenciais tanto na norma ISO/IEC 42001 (International Organization for Standardization & International Electrotechnical Commission, 2023) quanto na resolução nacional do CNJ para mitigar vieses e garantir a precisão dos resultados. Dessa forma, a transformação digital impulsionada por agentes de IA não apenas acelera a burocracia da "papelada", mas também reforça a segurança jurídica ao operar dentro de um ciclo de vida controlado.

2.2 Objetivos

2.2.1 Objetivo Principal

Analisar e propor uma metodologia para o desenvolvimento e implementação de sistemas e agentes de IA como ferramenta de transformação digital no âmbito da administração pública, com foco na otimização de processos e na eficiência do serviço público, em conformidade com os preceitos éticos e normativos estabelecidos pela Resolução nº 615/2025 do Conselho Nacional de Justiça.

2.2.2 Objetivos Secundários

1. Construir o alicerce do conhecimento sobre agentes: Nesta fase, o foco é mergulhar na teoria sobre agentes inteligentes, definir conceitos de agência e autonomia e descrever sua estrutura. Ao final, desenhar um mapa mental sobre todos os assuntos que dizem respeito a sistemas agênticos.
2. Compilar normas e diretrizes: A ideia deste objetivo é buscar as principais normas, diretrizes e outras legislações e realizar uma leitura minuciosa com o propósito de entender onde estão as zonas verdes e as limitações.
3. Identificar gargalos na rotina administrativa e judicial: O trabalho se resume a voltar os olhos para o dia a dia no serviço público, suas rotinas, e identificar onde o trabalho emperra. Para isso, é preciso rastrear tarefas repetitivas e cansativas que consomem tempo humano desnecessariamente.
4. Planejar soluções e idealizar produtos: Projetar a estrutura dos agentes. Os desafios identificados no objetivo anterior são utilizados para criar possíveis soluções. A meta é criar uma construção que funcione de forma modular e a solução precisa ser escalável.
5. Desenvolver soluções e avaliar performance: Nesta etapa, se resume a executar o planejamento e construir versões funcionais dos agentes. O intuito é lapidar a ferramenta através de tentativa e erro controlado. Avaliar esses produtos, ajustar a rota, corrigir falhas, e deixá-la pronta para entrar em produção.

2.3 Metodologia

Este estudo é classificado como **pesquisa aplicada**, de natureza empírica e abordagem mista qualitativa-quantitativa. Esse estudo é focado no desenvolvimento e avaliação de provas de conceito (PoCs) e aplicações funcionais de sistemas agênticos voltados à administração pública. A abordagem é orientada por artefatos: o desenvolvimento faz parte da investigação, e os produtos tecnológicos constituem simultaneamente o meio e o resultado da pesquisa.

A pesquisa qualitativa foi desenvolvida em fases de imersão em campo, onde dados não estruturados, como relatos de servidores, observações diretas de rotinas de trabalho e modelos mentais de especialistas, foram coletados e interpretados. A parte quantitativa está presente na fase de avaliação dos artefatos, onde métricas foram empregadas para mensurar a eficácia dos protótipos desenvolvidos.

2.3.1 O Laboratório de Inovação como Campo de Pesquisa

O campo de pesquisa foram os próprios **laboratórios de inovação** dos órgãos públicos envolvidos. Ao longo da pesquisa, foi estabelecido uma cooperação entre seis órgãos além do TRE-GO — o TRE-AC, o TRE-AP, o TJAP, TJAC e a Universidade Federal do Tocantins (UFT) —, formalizadas pelo Acordo de Cooperação Técnica nº 09/2025, em cumprimento à Meta 9 do CNJ, que fomenta a atuação em rede.

Esses laboratórios funcionaram como um **canteiros de obras**: espaços de experimentação controlada onde era possível testar rapidamente, errar em ambiente seguro e evoluir a abordagem a cada ciclo. A cada nova imersão em um laboratório parceiro, um novo agente era desenvolvido em colaboração com os servidores locais. Esse processo iterativo de tentativa e erro controlado, repetido ao longo de múltiplas imersões, foi o verdadeiro motor de aprendizagem da pesquisa.

Importante ressaltar que a metodologia aqui apresentada não foi definida *a priori*. Ela **evoluiu ao longo do processo**: após inúmeras tentativas, erros e acertos ao longo dos ciclos de desenvolvimento, convergimos em uma metodologia que trouxe atenção à necessidade real e à experiência do usuário. Os dois pilares que se consolidaram nesse processo — o *Design Thinking* e a Análise Cognitiva de Tarefas — foram incorporados gradualmente, à medida que percebemos seu impacto na qualidade dos resultados.

A relevância dessa metodologia foi validada e reconhecida externamente: o ecossistema colaborativo resultante, denominado Nativa IA, foi laureado com o **1º lugar no Prêmio Justiça Exponencial 2025 (J. Ex.)** na categoria Laboratórios de Inovação (J.Ex, 2025).

2.3.2 Primeiro Pilar: Design Thinking

O *Design Thinking* representa o primeiro pilar da presente metodologia, utilizado com o propósito de garantir que as soluções fizessem sentido do ponto de vista de quem as utilizaria (Brown,

2008). Incluir os próprios usuários no processo desde o começo foi essencial para a construção da ideação.

A aplicação do *Design Thinking* seguiu suas cinco fases clássicas: Empatia, Definição, Ideação, Prototipação e Teste, adaptadas ao contexto de desenvolvimento de agentes de IA:

Empatia

A fase de empatia foi baseada em **entrevistas** e **sessões de observação direta** no âmbito dos laboratórios de inovação. Essa abordagem focou na **experiência real do servidor**: suas dificuldades ao executar a tarefa, os principais gargalos de tempo, "segredos"na execução da tarefa e os pontos inalcançados pelo processo atual. O objetivo era mapear as dores reais dos processos de trabalho.

Definição

Com base nos dados coletados na fase de empatia, iniciou o processo de releitura do processo de trabalho e à identificação e ao **reenquadramento dos problemas sob a perspectiva da IA agêntica**. Foram identificados alguns processos candidatos à automação — consulta a normativos, auditoria, licitação, geração de atas, entre outros. Para cada processo candidato, avaliou-se a viabilidade de agenciamento com base em três critérios:

1. **Divisibilidade:** Deve ser possível quebrar o processo principal em sub-tarefas;
2. **Disponibilidade de contexto:** Os dados consultados pelos usuários devem estar disponíveis através de algum meio, como API, bancos de dados, arquivos compartilhados;
3. **Personalidade:** Deve ser possível definir uma personalidade clara sobre o agente, além do fluxo de trabalho;

Caso esses três requisitos estejam presentes, o sistema é considerado válido para o trabalho de agenciamento. Um diferencial desta fase foi a orientação pela ACT (segundo pilar, detalhado a seguir): o estudo partiu do mapa cognitivo real das tarefas, o que gerou arquiteturas fundamentadas em LLMs, RAG e orquestração agêntica.

Ideação

A geração de soluções utilizou **encontros colaborativos** com as partes interessadas. Os laboratoristas, nessa fase, desenhavam a solução baseado em seus anseios e propósitos, gerando arquivos de base para a prototipação, como mapa de estilos, experiência de usuário e planilha de requisitos e regras de negócio.

Prototipação

O ciclo de prototipação seguiu o protocolo do **Produto Mínimo Viável (MVP)**. O processo se deu em 3 etapas:

1. **Testes em IAs comerciais:** Inicialmente, práticas exploratórias foram realizadas diretamente em LLMs comerciais (Gemini, GPT) com os usuários. O objetivo era apenas ajustar as expectativas sobre as capacidades e limitações da tecnologia, além de refinar os prompts em um ambiente controlado.
2. **Protótipos iniciais:** Versões funcionais simplificadas foram desenvolvidas, permitindo que os usuários interagissem com o agente. O foco foi a validação do fluxo do sistema e da qualidade das respostas.
3. **Protótipos funcionais:** As versões validadas foram reconstruídas como aplicações web para testes de usabilidade e integração para produção.

Teste

A validação foi realizada com usuários no ambiente de desenvolvimento. O *feedback* coletado sistematicamente durante a interação dos servidores com os protótipos foi tratado como dado de pesquisa, gerando melhorias para as novas versões dos agentes.

2.3.3 Segundo Pilar: Análise Cognitiva de Tarefas (ACT)

O segundo pilar da metodologia surgiu de uma constatação prática: enquanto sistemas determinísticos tradicionais dependem de **requisitos funcionais** para serem construídos, os agentes de IA dependem do **mapa cognitivo da tarefa**. A diferença é fundamental. Os requisitos funcionais descrevem *o que* o sistema deve fazer. O mapa cognitivo descreve *como* o sistema deve fazer (Crandall et al., 2006).

Essa percepção conduziu à adoção formal da ACT como ferramenta para acessar e representar o conhecimento tácito dos servidores — aquele conhecimento que não está documentado nos manuais de procedimento, mas que reside na experiência acumulada do profissional: sua intuição, suas verificações, seus atalhos mentais e seus critérios de qualidade. A ACT mostrou-se indispensável ao revelar que **grande parte do conhecimento especialista simplesmente não estava documentado, mas estava na cabeça dos servidores** (CognitiveGroup, 2023).

Duas técnicas da ACT foram empregadas de forma sistemática ao longo das imersões:

Análise Cognitiva Aplicada de Tarefas (ACTA)

A ACTA consistiu no mapeamento estruturado das tarefas mais complexas dentro de cada processo de trabalho. O procedimento envolveu sessões com os servidores responsáveis, nas quais se realizou:

1. **Decomposição:** Cada tarefa candidata foi decomposta em subtarefas, pontos de decisão e ações elementares, revelando a estrutura real — e frequentemente não documentada — do processo.
2. **Identificação de pontos críticos de decisão:** Para cada nó decisório, foram documentados os insumos informacionais utilizados, as alternativas consideradas e os critérios de escolha

do especialista.

3. **Mapeamento de exceções:** Foram registradas as situações atípicas ou de exceção que exigem julgamento humano diferenciado — informação essencial para definir os limites de atuação e os *guardrails* do agente.

Os pontos críticos de decisão identificados pela ACTA serviram para fundamentar os componentes dos agentes. A *persona*, em especial, é essencialmente baseada nas habilidades necessárias do trabalho. As ferramentas necessárias, da mesma forma, são extraídas a partir das consultas realizadas identificadas no mapeamento da tarefa.

Pensamento em Voz Alta (*Protocolo em pensar alto*)

Complementarmente à ACTA, foi aplicado o protocolo de **verbalização do pensamento**. Nessas sessões, solicitou-se aos servidores que executassem tarefas reais verbalizando seu raciocínio em tempo real — narrando o que estavam pensando, por que tomavam cada decisão, e o que observavam no material analisado. O pesquisador atuou como mediador, indagando sistematicamente sobre a motivação de cada decisão.

Essa técnica capturou **a profundidade cognitiva que não aparecem em entrevistas convencionais**. Enquanto em uma entrevista o servidor descreve o processo de forma racionalizada e simplificada, a verbalização da tarefa revela o fluxo real de pensamento, com suas hesitações, reflexões e pequenas decisões.

Da ACT à Engenharia de Prompt: a Ponte Metodológica

O principal resultado metodológico da integração entre ACT e desenvolvimento de agentes foi a descoberta de que **converter o conhecimento não documentado dos servidores em prompts foi o diferencial** que permitiu construir agentes alinhados à prática real. Essa conversão se materializou em três elementos:

- **Persona do agente:** A intuição, a forma de agir e pensar do servidor — e até seu comportamento — foram traduzidos na Persona do agente. As instruções comportamentais foram construídas a partir do perfil cognitivo real mapeado pela ACT.
- **Prompts de sistema:** Os fluxos de raciocínio, critérios de decisão e estratégias de verificação documentados nos mapas cognitivos foram codificados como instruções estruturadas, reproduzindo a lógica deliberativa do especialista dentro do motor de linguagem.
- **Ferramentas:** A estratégia de ações dos agentes é baseada em suas ferramentas disponíveis. Sendo assim, a partir do fluxo do trabalho, busca-se o desenvolvimento de utilitários capazes de suprir tais demandas.
- **Guardrails:** As zonas de risco e os limites de competência identificados nas sessões — situações em que o próprio servidor reconhecia incerteza ou buscava supervisão — foram traduzidos em restrições do agente, definindo quando ele deve escalar a decisão para um humano.

O objetivo, em síntese, era **espelhar o mapa cognitivo dos melhores especialistas dentro do motor do sistema**, para obter resultados próximos da realidade. Esse avanço na engenharia de prompt e na definição da persona do agente foi o salto qualitativo mais significativo observado ao longo das imersões.

2.3.4 Processo Consolidado de Desenvolvimento

A integração dos dois pilares — Design Thinking e ACT — em um processo iterativo de criação convergiu, ao longo das imersões, em quatro macroetapas, ilustradas na Figura 2.1:

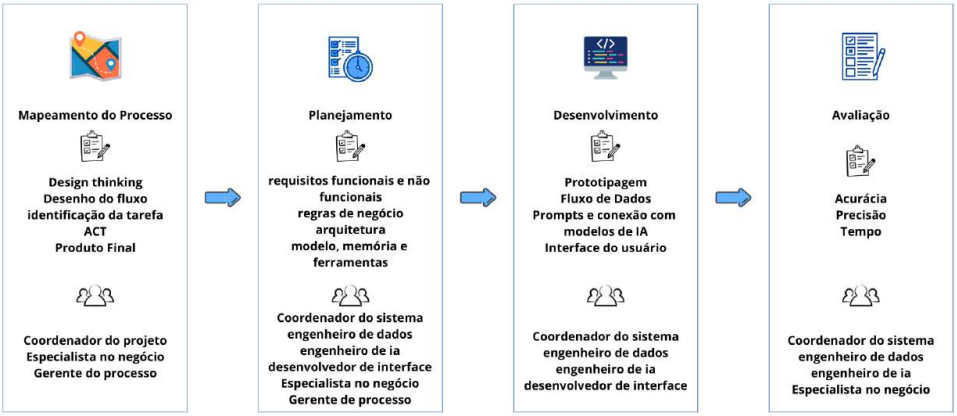

Figura 2.1: Metodologia de prototipação de agentes de IA, integrando Design Thinking e ACT

1. **Mapeamento do processo e identificação da tarefa:** Nesta etapa, a equipe reuniu-se com a unidade interessada na criação do agente para realizar a leitura do processo de trabalho. Dentro do processo, identificou-se a tarefa a ser agenciada, levantando informações básicas: insumos, fluxo, produto, prazos e demais informações relevantes. O diferencial desta etapa, fundamentado pela ACT, é que a imersão não se limitou a documentar *o que* o servidor faz, mas *como ele pensa e executa* a atividade — passo a passo, com suas decisões intermediárias e critérios implícitos.
2. **Planejamento:** A partir do conhecimento profundo da tarefa, foram identificados os requisitos funcionais, não funcionais e as regras de negócio. Uma informação fundamental nesta fase é a definição prévia de como será medido o sucesso da tarefa, para posterior avaliação. Com esses elementos, estruturou-se a arquitetura do agente: modelo de linguagem, estratégia de memória (curto e longo prazo), ferramentas internas e externas, nível de autonomia e, sobretudo, a persona e os prompts de sistema derivados do mapa cognitivo.
3. **Desenvolvimento:** Nesta etapa, realizou-se a construção do agente seguindo a lógica de MVP — versões mínimas, validação rápida com usuários e refinamento contínuo dos prompts. A equipe de desenvolvimento, quando completa, contou com quatro papéis: um engenheiro de dados, um engenheiro de IA, um especialista no domínio (o servidor) e

um desenvolvedor de interfaces. A presença do especialista no domínio como membro ativo da equipe — e não como mero demandante — foi determinante para a qualidade dos resultados.

4. **Avaliação:** Por fim, realizou-se a avaliação do agente, mensurando seu desempenho por meio de métricas objetivas. Um agente foi considerado satisfatório quando apresentou desempenho equivalente ao humano na mesma tarefa, e estado da arte quando apresentou desempenho superior ao de um profissional comum.

Cabe destacar que a metodologia foi destaque no prêmio Justiça Exponencial 2025 (J. Ex.), sendo laureado em primeiro lugar na categoria Laboratórios de Inovação (J.Ex, 2025)

2.3.5 Procedimentos de Avaliação

A avaliação dos artefatos desenvolvidos seguiu uma abordagem mista, combinando métricas quantitativas de desempenho com validação qualitativa junto aos usuários.

Avaliação Quantitativa

Para cada agente e *chatbot*, construiu-se um **conjunto de dados de referência** composto por pares de entrada e resposta esperada, elaborados em colaboração com os especialistas do domínio. As respostas geradas pelos agentes foram confrontadas com as respostas esperadas, e o desempenho foi mensurado por meio das seguintes métricas padrão de aprendizado de máquina:

- **Acurácia:** Proporção de respostas corretas sobre o total de casos avaliados.
- **Precisão (*Precision*):** Proporção de respostas positivas corretas dentre todas as respostas positivas emitidas, medindo a confiabilidade das saídas do agente.
- **Revocação (*Recall*):** Proporção de casos positivos corretamente identificados dentre todos os positivos existentes, medindo a cobertura do agente.
- **F1-Score:** Média harmônica entre precisão e revocação, oferecendo uma medida balanceada.

Adicionalmente, foram construídas **Matrizes de Confusão** para cada agente, permitindo a análise dos tipos de erro predominantes — falsos positivos e falsos negativos — e orientando os ajustes nos prompts e na arquitetura. O processo completo de avaliação de cada produto está documentado em seus respectivos artigos de produção tecnológica.

Avaliação Qualitativa

Complementarmente, os protótipos foram submetidos a sessões de validação com os servidores usuários, que avaliaram a adequação das respostas ao contexto institucional, a qualidade dos documentos gerados, a percepção de economia de tempo em relação ao processo manual e a disposição para incorporar a ferramenta em sua rotina de trabalho.

2.3.6 Ferramentas computacionais e uso de IA

Para a implementação dos protótipos e sistemas inteligentes desenvolvidos nesta pesquisa, adotou-se uma arquitetura de software baseada na linguagem Python. O desenvolvimento do *backend* foi realizado utilizando o framework *FastAPI*, enquanto a interface de usuário (*frontend*) foi estruturada através de *Jinja2 e JavaScript*. A orquestração dos fluxos de dados foram implementadas utilizando a plataforma de automação de fluxos de trabalho n8n. A camada de inteligência artificial foi desenvolvida utilizando a bilioteca OpenAI Agents SDK.

Durante o processo de desenvolvimento, modelos de Inteligência Artificial Generativa — especificamente Claude (Anthropic) e Gemini (Google) — foram utilizados como ferramentas de apoio à programação. O auxílio destas ferramentas concentrou-se na geração de scripts auxiliares, otimização de sintaxe, documentação de código e identificação de erros dentro da arquitetura supracitado.

Adicionalmente, recursos de IA foram empregados para a revisão gramatical e o aprimoramento da fluidez textual desta dissertação. Ressalta-se, contudo, que o uso dessas ferramentas foi estritamente supervisionado e revisado pelo autor. Todas as concepções teóricas, a estrutura da solução, as interpretações dos dados, discussões e conclusões apresentadas neste trabalho são de autoria intelectual exclusiva do autor. A IA não foi utilizada para a geração de ideias originais ou para a redação de conteúdo científico sem a devida verificação humana.

Capítulo 3

Fundamentação Teórica

Durante o processo de ideação e desenvolvimento dos agentes de IA, ocorreram inúmeros debates interessantes acerca de sua natureza. A exemplo do que ocorre em outras tecnologias emergentes e disruptivas, há certa confusão em torno de seus conceitos. A IA tem se consolidado como uma tecnologia transformadora, solucionando inúmeros problemas complexos. No entanto, é essencial estabelecer uma compreensão realista sobre o que são esses sistemas, como funcionam e quais são seus fundamentos conceituais e suas limitações.

Nesta seção, apresentamos os principais conceitos, técnicas e métodos a partir de uma pesquisa bibliográfica em bases de dados acadêmicas e obras de referência sobre o tema. Um fato interessante relacionado a este trabalho é que quase a totalidade das referências datam do início do ano de 2024, o que demonstra o caráter inovador deste campo de estudo. A aceleração das publicações e o interesse global em "Agentes de IA"e "IA Agêntica"tornaram-se particularmente significativos após o lançamento do ChatGPT em novembro de 2022.(Xing, 2025)

3.1 Definição e Evolução dos Agentes de IA

Um agente de IA é fundamentalmente definido como uma entidade baseada em software capaz de perceber seu ambiente, raciocinar sobre seus objetivos, tomar decisões e executar ações, frequentemente de forma autônoma, por meio da interação com sistemas externos. Trata-se de entidades complexas e autogovernadas que percebem seu ambiente e agem para atingir determinados objetivos. (Alto, 2025)

A evolução dos agentes teve origem nos sistemas de Automação de Processos Robóticos (RPA), que se limitavam a fluxos de trabalho predefinidos. Com o avanço da IA, surgiram automações baseadas em Aprendizado de Máquina (ML) e Aprendizado por reforço (RL), que podiam aprender com dados e otimizar ações por tentativa e erro, embora sofressem de generalização.

O surgimento dos grandes modelos de linguagem, escalas de mais de 2 bilhões de parâmetros, foi um marco para a linha de pesquisa. Os agentes modernos podem ser vistos como uma extensão desses modelos, elevando-os de simples modelos estatísticos a sistemas que possuem capacidades de autonomia, reatividade, proatividade e habilidade social. As aplicações são agora capazes de raciocinar e planejar, integrando-se em ambientes externos. (Raieli & Iuculano, 2025)

3.2 Ciclo de Vida SPAR - Sentir, Pensar, Agir e Refletir

Um agente de IA opera através de um ciclo iterativo, buscando maximizar o sucesso na conclusão de uma tarefa. Este processo é frequentemente ilustrado pelo *framework* SPAR, cuja lógica é semelhante ao raciocínio humano. A figura 2.1 descreve o ciclo. (Bornet et al., 2025)

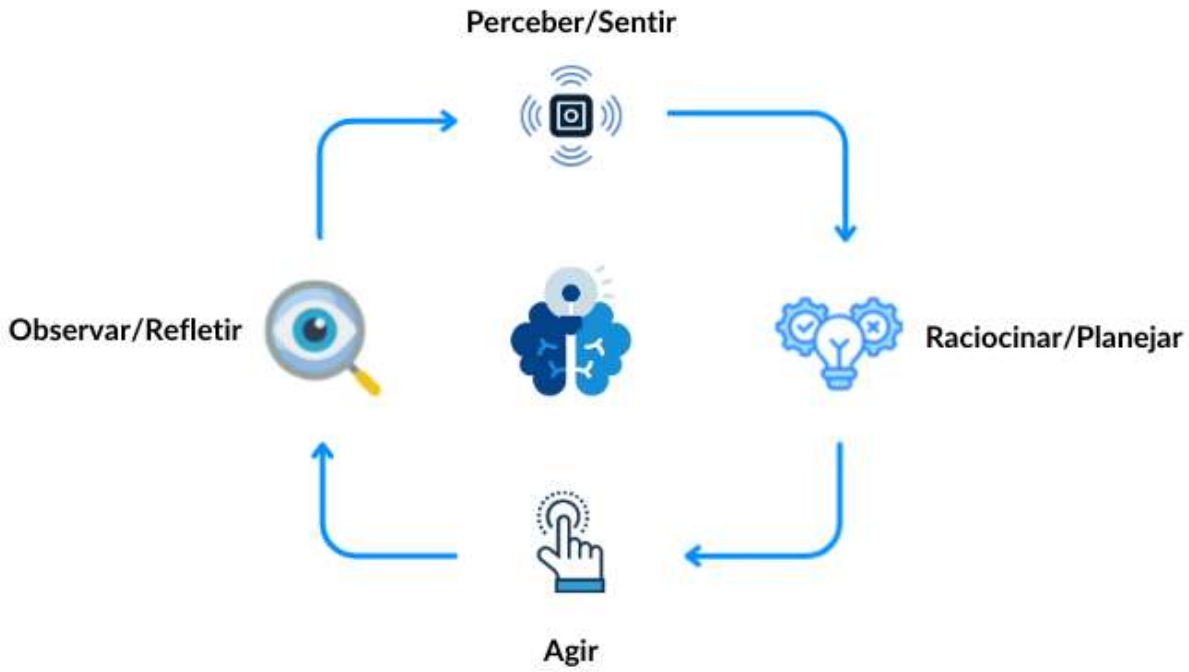

Figura 3.1: Ciclo de vida SPAR

O estágio inicial é o de **Sentir**; atua como a percepção do sistema. Nesta fase, o sistema absorve tudo ao seu redor, buscando informações relevantes para a tarefa; coleta dados de diversas fontes utilizando sensores (APIs, busca na internet, câmeras) e mantém uma memória de curto prazo.

Após a leitura de seu ambiente, inicia-se a etapa de **Planejar e raciocinar**. O agente, então, correlaciona seu objetivo final definido previamente e as informações adquiridas, e a partir disso, projeta inúmeros caminhos para alcance de suas metas. Por fim, decompõe a tarefa complexa em subtarefas gerenciáveis, e decide a melhor estratégia.

O **Agir** é o estágio em que o agente, considerando suas possibilidades e limitações, executa a melhor estratégia decidida. Nesta etapa, ações de monitoramento acompanham a tarefa principal para garantir o efetivo cumprimento da meta. A depender da autonomia do agente, é crucial a possibilidade da intervenção humana, especialmente em agentes de alto risco.

E por fim, o **Refletir**. A partir dos dados coletados nas etapas anteriores, o agente avalia sua performance e o atingimento de seus objetivos e metas. Essas informações são armazenadas em sua memória de longo prazo. O agente avalia esse resultado e evolui sua abordagem para tarefas futuras.

3.3 Estrutura e Componentes Essenciais

Um agente funcional é composto por estruturas que trabalham colaborativamente, orquestradas por um mecanismo de planejamento que transforma intenções em ações concretas (Raieli & Iuculano, 2025). A Figura 2.2 apresenta os elementos principais que tipicamente compõem esta estrutura.

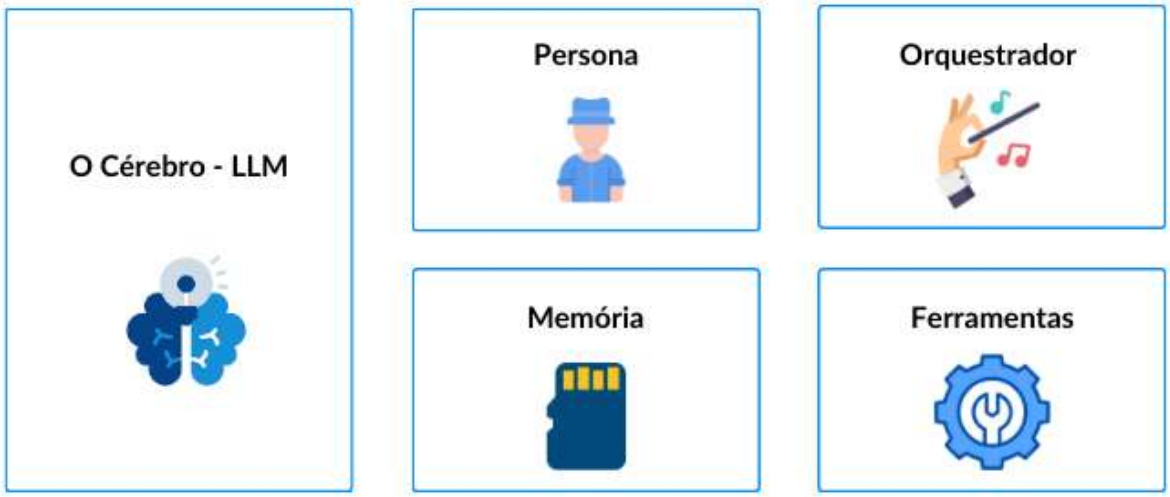

Figura 3.2: Estrutura dos Agentes

- Cérebro - LLM: Atua como o núcleo cognitivo principal, fornecendo compreensão da linguagem natural e geração de respostas. É o componente responsável por simular o raciocínio lógico e o planejamento das tarefas. (Alto, 2025)
- Persona: Instrução primária que define a identidade e o propósito do agente. Ela orienta o perfil do agente, estabelecendo o papel, o tom de voz, a personalidade e quaisquer restrições éticas ou funcionais que o agente deve seguir.
- Orquestrador: funciona como uma camada que governa o fluxo de execução, garantindo a coordenação entre o LLM, as ferramentas e a memória.
- Memória: Base de dados responsável pelo armazenamento do contexto ao longo do tempo. Isso inclui Memória de Curto Prazo, adquirida no processo de recuperação da informação e Memória de Longo Prazo, armazenada ao final do ciclo SPAR, utilizada para adaptar o agente nas próximas tarefas.
- Ferramentas: Representam as extensões externas da camada de inteligência do agente, permitindo-lhe interagir com o mundo exterior. Isso inclui APIs, bases de dados, mecanismos de busca ou scripts de automação.

3.4 Conceitos Fundamentais

3.4.1 Autonomia: O Grau de Independência

O nível de autonomia concedido a um agente é uma escolha a ser definida com base nos requisitos funcionais, a avaliação dos riscos e a necessidade de supervisão humana. A literatura classifica em três níveis principais: (Lipenkova, 2025)

- **Humano no Circuito:** Exige a aprovação humana para autorizar as decisões;
- **Humano no Monitoramento:** O agente atua de forma autônoma, mas um supervisor humano monitora e pode intervir, se necessário;
- **Humano Fora do Circuito:** Autonomia total do agente de IA sem supervisão humana em tempo real.

3.4.2 Agência: Intencionalidade e Poder de Escolha

A agência, por sua vez, diz respeito a capacidade de um sistema agir de forma proativa, dentro de suas capacidades e limitações, visando atingir objetivos específicos. O agenciamento adiciona uma camada de liberdade ao agente, caso seja necessário. Busca responder a pergunta: Qual é a melhor ação para atingir o objetivo? (Bornet et al., 2025)

A agência em sistemas de IA é composta por três elementos:

1. **Autoridade Decisória:** O poder de atuar e executar ações de acordo com uma alternativa escolhida. Os sistemas com agência avaliam diferentes opções e selecionam a ação mais apropriada com base em seus processos internos de tomada de decisão;
2. **Intencionalidade:** Implica a existência de intenções, objetivos ou metas que guiam o comportamento e as ações do sistema. Agentes devem se ajustar para alcançá-las;
3. **Responsabilidade:** Refere-se à responsabilização pelos resultados e consequências das ações tomadas.

Em síntese, enquanto a agência é a capacidade de agir, a autonomia mede o nível de independência com que essa capacidade é exercida.

3.4.3 Auto-Governança: Controle e Adaptação

A auto-governança representa o nível mais avançado, medindo o poder que um agente possui em evoluir. Este conceito desdobra-se nas seguintes competências:

- **Auto-Organização:** É a capacidade do sistema de estruturar e organizar os seus próprios processos internos;
- **Auto-Regulamentação:** Envolve a monitorização e o ajuste das suas próprias saídas com base em feedback;
- **Auto-Adaptação:** É a habilidade de modificar suas próprias estratégias em resposta a mudanças nas suas condições internas ou no ambiente, de modo a atingir seus objetivos

de forma mais eficiente;

- **Auto-Otimização:** É a melhoria contínua do seu desempenho, eficiência ou capacidades de tomada de decisão através de processos evolutivos, aprendizado e experiência;
- **Auto-Determinação:** É a capacidade de estabelecer os seus próprios cursos de ação, prioridades e objetivos com base nos seus processos internos de tomada de decisão, em vez de ser totalmente controlado por forças externas.

3.5 Classificação dos Agentes de IA

Os agentes de IA podem ser classificados em diversos aspectos. Neste trabalho, vamos classificar em três perspectivas: por arquitetura, por capacidade e por raciocínio. (Davies, 2025)

3.5.1 por arquitetura

- simples: são sistemas de agente único;
- multi-agente: são sistemas em que vários agentes de IA colaboram para resolver problemas, podendo haver comunicação A2A (agente para agente) entre os mesmos ou não.

3.5.2 por capacidade

- recuperação de informação: se baseiam no método RAG (Geração Aumentada por Recuperação) e servem para buscar informações relevantes em bases de dados externas;
- tarefas de automação: desenvolvidos para automatizar rotinas de trabalho e executar ações em resposta a comandos ou gatilhos externos, como substituir tarefas repetitivas;
- autônomos: combinam a busca de informações com a automação. Eles conseguem orquestrar várias tarefas, planejar, autoajustar e tomar decisões em tempo real com pouca ajuda humana.

3.5.3 por raciocínio

- reativo: percebem um estímulo do ambiente e fornecem uma resposta imediata, seguindo regras simples de ação-reação;
- deliberativo: seguem um ciclo de sentir-planejar-agir e são muito bons para tarefas que exigem planejamento e solução de problemas;
- híbrido: Eles usam uma camada reativa para as respostas muito rápidas e uma camada deliberativa para o raciocínio mais profundo e o planejamento de alto nível.

3.6 Mudança do Paradigma

A IA agêntica não é definida apenas pelo uso de um modelo de linguagem incorporado em um sistema tradicional. Esse recurso pode ser classificado como sistemas tradicionais com uso de IA, e não agenciamento propriamente dito. O agente é definido pela junção de fatores que torna o sistema capaz de antecipar, se adaptar, traçar estratégias e alterar rotas.

Dentro destes fatores podemos incluir sensores, agindo como a percepção do sistema; o motor principal de raciocínio - os modelos de linguagem, principalmente aqueles treinados em raciocício, traçando o planejamento; atuadores, ferramentas capazes de alterar e buscar informações do mundo externo, responsáveis pela ação.

Especialistas enfatizam que sistemas com uso de IA "não se aprimoram além dos dados de treinamento e não conseguem ajustar seu comportamento além da lógica programada"(Vittie, 2025). Mesmo que utilizem modelos probabilísticos para buscar respostas, são determinísticos em seu fluxo operacional. Por outro lado, a IA agêntica tem o diferencial da proatividade, autonomia e do aprendizado constante.

A proatividade pode ser entendida como a capacidade do agente em monitorar o processo de trabalho e em interferir por meio de uma ação específica, sem precisar de uma ordem humana direta.

Em relação à independência, podemos classificar o sistema em vários níveis. Nos níveis mais baixos, o agente é contido e reativo (Pant & Viswanathan, 2025). No meio da pirâmide, estão níveis que coopilotam e condicionais. E no alto, agentes altamente autônomos.

1. Nível 0 (Sem Independência): O sistema é puramente reativo. O humano tem controle total sobre cada pequena decisão. Exemplo: Um corretor gramatical que sugere mudanças enquanto você digita.
2. Nível 1 (Baseada em regras): O sistema pode executar tarefas, porém segue comandos rígidos e pré-determinados. Exemplo: Um corretor que sempre inclui a pontuação final ao fim dos parágrafos, sendo que o usuário cadastrou tal regra no sistema.
3. Nível 2 (Co-piloto): O agente participa do processo de tomada de decisão, porém a execução depende do humano. Exemplo: O corretor realiza as mudanças gramaticais, porém o humano atua como o aprovador antes de surtir efeito. O sistema passa a ter memória de curto prazo e consciência do contexto.
4. Nível 3 (Independência Condicional): Este nível de independência permite que o sistema gerencie processos completos dentro de um ambiente específico e controlado. Ele opera com independência, a não ser que encontre um caso de exceção. Exemplo: O corretor realiza as mudanças gramaticas e aprova-as, dentro de um domínio específico ou uma certeza alta. Em situações de dúvidas ou assuntos fora do domínio, o humano é notificado para aprovação.
5. Nível 4 (Alta independência): O agente opera com independência na maioria das situação e o humano se torna um observador do processo, recebendo relatórios de progresso e in-

tervindo em crises sistêmicas. Exemplo: O corretor realiza todas as correções gramaticais e as aprova. Em caso de falha grave do agente, o humano intervém e inicia a manutenção do sistema.

6. Nível 5 (Independência Total): De certa forma teórico, representa sistemas capazes de operar em qualquer domínio, sem qualquer supervisão humana.

Por fim, em relação ao aprendizado, tal capacidade está estritamente relacionada à memória do agente. Enquanto a memória de curto prazo, onde é utilizada ao longo daquele ciclo de vida da tarefa e descartada ao final, produz contexto para uma alta precisão na resposta, a memória de longo prazo torna o agente mais inteligente ao longo da execução das tarefas. O diferencial está em armazenar informações de qualidade em cada fim de ciclo, evitando que a memória mantenha ruídos, desinformação, vícios, traumas e outros itens prejudiciais ao agente.

3.7 Engenharia de Prompt para agentes de IA

O prompt pode ser entendido como o comando incluído pelo usuário na LLM, na expectativa de um retorno útil. O que chamamos de prompt não é apenas uma pergunta; é, na verdade, uma ponte entre a intenção humana e o resultado da máquina.

A engenharia de prompt, portanto, é a ciência — e também um pouco de arte — de estruturar seus pensamentos e instruções para reduzir a ambiguidade e garantir que leve exatamente ao lugar desejado. Os LLMs funcionam como motores estatísticos; eles preveem a próxima sequência de palavras baseados no contexto que você fornece. Se o contexto for ruim, a previsão será ruim.

Para entender por que escrever bem é importante, precisamos olhar para como os modelos modernos funcionam. Eles não são apenas um bloco maciço de conhecimento. A estrutura atual, conhecida como *Mixture of Experts (MoE)*, sugere que o modelo é composto por várias sub-redes especialistas — imagine que há um especialista em matemática, outro em poesia, outro em código Python.

O comando funciona, então, como um roteador. Se você faz uma pergunta vaga como "arrume meu código", o roteador pode acionar um especialista genérico. Mas, se você diz "Atue como um Engenheiro Sênior de Python e refatore este código seguindo as normas PEP-8", você obriga o roteador a chamar o especialista certo para o trabalho. Usar o vocabulário correto é como dar coordenadas de GPS precisas dentro do vasto oceano de dados do modelo.

3.7.1 Componentes do prompt

Um prompt robusto geralmente segue uma estrutura que podemos chamar de CPRF: Contexto, Papel, Restrição e Formato.

- Contexto: É o cenário. Em vez de pedir "escreva um e-mail", você diz "estamos analisando os resultados de vendas do terceiro trimestre e precisamos reportar uma queda".

- Papel: É a persona. "Você é um gerente de marketing B2B experiente". Isso ajusta o tom e o vocabulário.
- Restrição: O que não fazer ou limites. "Não use jargões técnicos"ou "máximo de 500 palavras". Isso é vital porque modelos tendem a ser verborrágicos.
- Formato: Como você quer a entrega? Uma tabela, um JSON, um poema ou uma lista em markdown?

3.7.2 Técnicas de prompt

1. *Zero-Shot e Few-Shot*: Às vezes, instrução direta funciona (Zero-Shot). Mas modelos aprendem muito melhor por analogia. A técnica Few-Shot consiste em dar exemplos de "entrada e saída"ideais. (Pisarevskaya & Zubiaga, 2025)

2. Cadeia de Pensamento (*Chain of Thought - CoT*): A técnica CoT força o modelo a "pensar em voz alta"antes de responder. Ao instruir "pense passo a passo"ou pedir para gerar o raciocínio antes da resposta final, você pode aumentar drasticamente a precisão em matemática e lógica, pois o modelo gera passos intermediários que auxiliam no resultado final. (Wei et al., 2022)

3. Árvore de Pensamento (*Tree of Thoughts - ToT*): Para problemas que não são lineares, como estratégias jurídicas ou de negócios, o CoT pode não bastar. O ToT estimula a IA a explorar múltiplos caminhos possíveis, avaliar cada um e descartar os ruins. (Yao et al., 2023)

4. Encadeamento de Prompt: A saída de um prompt é a entrada do próximo prompt, orquestrado por algum controlador.

5. Cadeia de verificação (*Chain of Verification - CoVe*): O modelo gera uma resposta, depois cria perguntas para verificar se o que ele mesmo disse é verdade, responde a essas perguntas e corrige a resposta final. É excelente para precisão factual. (Dhuliawala et al., 2023)

6. Reflexão: É a técnica de "pensar antes de falar". O modelo gera um rascunho, depois assume o papel de um crítico para apontar erros ou tons inadequados nesse rascunho e, por fim, reescreve a versão melhorada. (Shinn et al., 2023)

6. Prompt de Recuo: Diante de uma pergunta muito específica, pedimos para a IA primeiro listar os princípios gerais e só depois responder ao caso concreto. (H. S. Zheng et al., 2024)

7. *System 2 Attention* (S2A): Se o texto de entrada está "sujo"com opiniões inúteis, pedimos para a IA reescrever o contexto deixando apenas os fatos relevantes, limpando o ruído antes de processar a resposta. (Weston & Sukhbaatar, 2023)

8. Esqueleto de Pensamento: Para textos longos, peça para a IA gerar um índice (esqueleto) primeiro. Depois, peça para ela preencher cada tópico. Isso garante que o texto final tenha começo, meio e fim organizados. (Ning et al., 2024)

9. Debate MultiAgente: Consenso através da colaboração e conflito de múltiplas personas. Simula uma discussão entre diferentes agentes (ex: um Otimista e um Pessimista, ou um Cientista e um Ético). Eles propõem respostas e criticam as respostas uns dos outros até chegarem a um consenso comum. (Liang et al., 2023)

3.8 Engenharia de Contexto: Geração aumentada por recuperação

A concepção do RAG não surgiu em um vácuo, mas como resposta direta às limitações observadas nos modelos, tais como:

1. Informações desatualizadas: os LLMs tradicionais foram treinados em um conjunto de dados disponíveis até a data de corte, sendo incapaz de responder problemas atuais;
2. Alucinações: sem acesso a fontes externas, os modelos preenchiam as lacunas com informações erradas;
3. Conhecimento privado: treinar uma LLM desde o início com dados particulares é extremamante caro;
4. Janela de contexto limitada: LLMs possuem limite de entrada de tokens.

O marco definitivo que formalizou e batizou a técnica é o artigo *"Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks"* (Lewis et al., 2020). Este trabalho, desenvolvido por pesquisadores do *Facebook AI Research, University College London e New York University*, propôs uma mudança arquitetônica em que o modelo de linguagem deixa de ser apenas um "enciclopedista"para se tornar um "pesquisador".

Em vez de gerar uma resposta baseada apenas no que aprendeu durante o treinamento, o modelo primeiro converte a pergunta do usuário em um vetor, busca documentos relevantes em outra base externa, e utiliza tais documentos como contexto para gerar a resposta final. É como permitir que um estudante faça uma prova com consulta ao livro, em vez de depender apenas da memória.

3.8.1 Técnicas RAG

RAG Clássico (*Naive RAG*)

O RAG Clássico descreve a implementação padrão que se tornou onipresente em frameworks de desenvolvimento como *LangChain e LlamaIndex*. O processo é linear:

1. Indexação: Seus documentos são quebrados em pedaços menores (chunks), convertidos em vetores que representam seu significado e armazenados em um banco de dados vetorial;
2. Recuperação: Quando você faz uma pergunta, o sistema busca os pedaços mais similares em relação a pergunta, utilizando distância vetorial;
3. Geração: Esses pedaços são enviados junto com o prompt inicial e enviados ao LLM, que escreve a resposta final baseada neles.

GraphRAG

Desenvolvido pela Microsoft Research (Larson & Truitt, 2024), o GraphRAG representa um salto qualitativo ao integrar a natureza não estruturada dos LLMs com a estrutura rígida e

relacional dos grafos de conhecimento.

Esta técnica, antes de responder a qualquer pergunta, lê todos os documentos e cria um um grafo, identificando quem são as entidades (nós) e como eles se conectam (arestas). Então agrupa esses conceitos e gera textos para cada grupo. Quando você faz uma pergunta complexa, ele usa esse mapa organizado e os resumos para construir uma resposta completa.

RAG Híbrido

A busca vetorial entende o significado das palavras, mas às vezes ignora a palavra exata. A busca antiga por palavras-chave encontra a palavra exata, mas não entende o significado. O RAG Híbrido usa as duas ao mesmo tempo: ele faz uma busca pelo significado e outra pela palavra-chave, e depois combina os resultados. É como ter um bibliotecário que entende o tema do livro e outro que sabe exatamente onde está o título específico que você pediu. (Mandikal & Mooney, 2024)

RAG Corretivo (CRAG)

O CRAG adiciona uma camada de conferência que verifica a qualidade dos documentos encontrados antes de enviá-los para a geração da resposta. Se o corretor achar que a informação é boa, ele a refina. Se achar que é ruim ou não tem nada a ver, ele joga fora e faz uma nova busca para encontrar a resposta certa. (Yan et al., 2024)

RAG Adaptativo

Nem toda pergunta precisa de uma pesquisa complexa. Perguntar "Qual a capital do Brasil?"é fácil para a IA; perguntar "Compare a economia de Tocantins e Goiás nos últimos 10 anos"é difícil. O RAG Adaptativo usa um roteador inteligente que analisa a dificuldade da pergunta. Se for fácil, a IA responde prontamente. Se for média, faz uma pesquisa simples. Se for difícil, faz uma pesquisa complexa de vários passos.

RAG Fusion

O RAG Fusion identifica a pergunta inicial e encaminha para a IA reescrevê-la de várias formas diferentes, cobrindo vários ângulos. Depois, ele faz pesquisas para todas essas variações. No final, ele organiza todos os documentos encontrados e os ordena, dando preferência a documentos que apareceram em várias das pesquisas. É como pedir a opinião de cinco especialistas diferentes e confiar mais na resposta a qual todos eles concordam. (Rackauckas, 2024)

Hypothetical Document Embedding (HyDE)

Em vez de usar o prompt inicial para recuperar documentos, o HyDE pede IA "alucinar"uma resposta ideal, mesmo que sem precisão. Ele então transforma essa resposta inicial em um vetor

e usa isso para buscar documentos reais. A ideia é que a resposta imaginada, mesmo com fatos errados, tem o vocabulário e a estrutura muito mais parecidos com o documento real que você quer encontrar do que a sua pergunta original. (L. Gao et al., 2023)

3.9 Resolução 615/CNJ e o Marco Regulatório

A Resolução nº 615/2025 do CNJ é considerada o marco regulatório do âmbito do Poder Judiciário a respeito do uso de IA. O ponto principal é que a IA deve ser usada apenas como ferramenta auxiliar e complementar para dar suporte à decisão, sendo proibido que ela tome decisões judiciais de forma autônoma. O magistrado deve sempre revisar, sendo o responsável por suas decisões.

Considerando o caráter de controle administrativo, a norma pode ser classificada como progressista em sua abordagem de governança e ética, entretanto conservadora no grau de autonomia concedida aos sistemas de IA em funções judiciais. A Tabela 3.1 resume a postura da resolução em diferentes aspectos.

Tabela 3.1: Postura da Resolução 615/CNJ em diferentes aspectos

Aspecto	**Postura**
Governança e risco	Progressista: adota estrutura de risco e cria Comitê supervisor.
Autonomia	Conservadora: Proíbe autonomia decisória judicial e exige supervisão humana.
Ética e vieses	Progressista: Exige mitigação contínua.
Transparência	Progressista: Exige explicabilidade, auditabilidade e publicidade de impacto algorítmico.

Apesar de citar apenas uma vez o termo agente de IA em um contexto secundário, o agente pode ser entendido como o sistema de IA no âmbito da Resolução. A maioria das aplicações permitidas na categoria de Baixo Risco se enquadra nessa definição. Exemplos de baixo risco incluem: execução de atos processuais ordinatórios, extração, classificação e agrupamento de dados e processos, sumarização de documentos e transcrição de áudio.

O CNJ proíbe explicitamente o uso autônomo para a tomada de decisões judiciais. O sistema de IA deve ser de caráter auxiliar e complementar e não pode restringir ou substituir a autoridade final dos usuários internos.

Quanto às formas de controle, podemos citar:

- **Transparência e Explicabilidade:** São princípios centrais. A Resolução exige que os resultados sejam compreensíveis e que os sistemas possuam explicabilidade sempre que tecnicamente possível;
- **Auditabilidade e Contestabilidade:** O Judiciário deve garantir que os sistemas sejam auditáveis, com registro automático das operações, e que os resultados possam ser ques-

tionados e revisados;

- **Mitigação de Vieses:** A Resolução é rigorosa ao exigir a prevenção e mitigação de vieses discriminatórios ilegais ou abusivos. Mais notavelmente, exige medidas corretivas e, se o viés for impossível de eliminar, a solução de IA deverá ser descontinuada;
- **Privacidade:** A Resolução adota os conceitos de *Privacy by Design e Privacy by Default*.

Segue abaixo um mapa mental contendo os principais tópicos de um agente de IA.

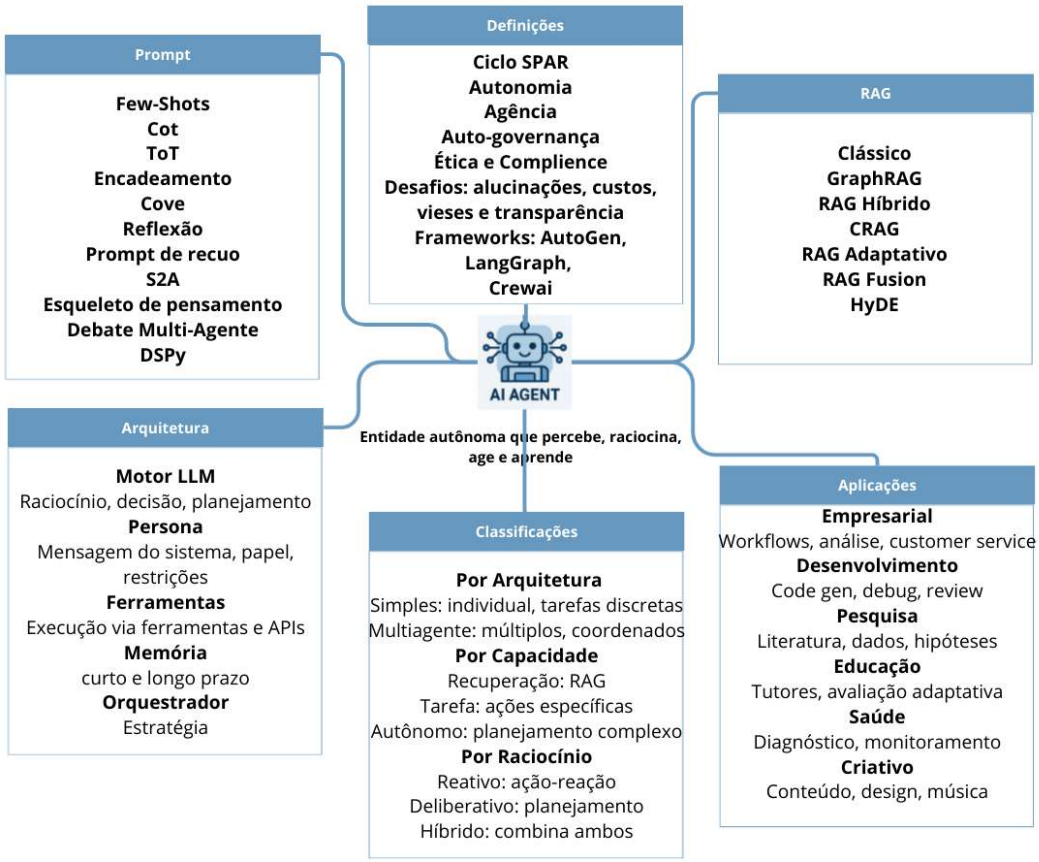

Figura 3.3: Mapa Mental - Agente de IA

Capítulo 4

Transformação Digital: rumo ao agêntico

Os resultados desta pesquisa aplicada consistem em uma coleção de artefatos tecnológicos desenvolvidos ao longo da pesquisa. O escopo dos entregáveis abrange sistemas agênticos de IA, chatbots, painéis gerenciais, conjuntos de dados e estudos complementares.

Adicionalmente, elaborou-se uma oficina prática contendo um resumo analítico para o desenvolvimento de sistemas agênticos, fundamentada na experiência prática adquirida. Por fim, a pesquisa culminou na idealização do Ecossistema Nativa IA, uma proposta conceitual para o serviço público brasileiro que visa fomentar o desenvolvimento colaborativo, pautado em normas padronizadas e focado na escalabilidade das soluções.

Segue um resumo extendido de cada produto abaixo. Para mais detalhes, cada produto tem um artigo referente à produção tecnológica.

4.1 Ecossistema Nativa IA

O ecossistema Nativa IA pode ser representado pela consolidação prática da proposta desta pesquisa. Indo além da criação de ferramentas, a proposta é formar uma rede de colaboração para o desenvolvimento de soluções de IA no âmbito do Poder Judiciário. Na prática, funciona como a junção de diversos laboratórios de inovação, onde é possível testarmos normas, técnicas e métodos para prototipação e futura produção.

A institucionalização desta união foi formalizada por meio do Acordo de Cooperação Técnica nº 09/2025, celebrado entre os Tribunais Regionais Eleitorais do Acre, Amapá e Goiás, o Tribunal de Justiça do Amapá, Tribunal de Justiça do Acre e a Universidade Federal do Tocantins, com o TRE-GO no centro da negociação. Este termo, objetivando o cumprimento da Meta 9 do CNJ que fomenta a atuação em rede, estabelece um fluxo de trabalho onde há o intercâmbio de experiência na área de inovação, envolvendo a academia com suporte infraestrutural e validação científica.

Cabe destacar que o Nativa IA foi destaque no Selo Judiciário Inovador 2025 do CNJ na categoria Gestão Judicial Inovadora, subcategoria Ideias Inovadoras (Conselho Nacional de Justiça (CNJ), 2025).

4.2 Sistemas agênticos

4.2.1 LIA: Licitação com IA

O processo de contratações públicas no Brasil, instituído pela Lei nº 14.133/2021 - Nova Lei de Licitações e Contratos Administrativos, exige rigoroso planejamento e documentação detalhada. Alguns documentos principais desse processo incluem: DFD - Documento de Formalização da demanda, PP - Pesquisa de preço, MR - Matriz de riscos, ETP - Estudo técnico preliminar, TR - Termo de referência; e o ED - Edital.

A elaboração manual destes documentos demanda tempo e recursos consideráveis dos funcionários envolvidos, abrangendo análises de dados históricos, consulta ao Plano de Contratações anual (PCA), alinhamentos ao plano estratégico, dentre outros estudos.

O objetivo do trabalho é desenvolver e implementar um sistema agêntico capaz de gerar automaticamente documentos licitatórios estruturados, em conformidade com a legislação de contratações públicas, reduzindo significativamente o tempo de elaboração e padronizando a qualidade dos documentos produzidos.

Considerando esse desafio, desenvolveu-se um agente especializado na automação em geração de documentos, integrando-se a uma aplicação web construída com *FastAPI* e utilizando a API do Google Gemini para processamento de linguagem natural e geração estruturada de documentos.

A aplicação implementa uma arquitetura assíncrona, com banco de dados PostgreSQL, sistema de autenticação e interface web responsiva. O DFD é o primeiro artefato gerado, estabelecendo as bases para os documentos subsequentes.

Como destaque, foi produzido um algoritmo de pesquisa de preços automatizada no PNCC - Portal Nacional de Contratação Pública. O algoritmo automatiza o processo de busca de itens de interesse recentemente contratados ou levantados e recupera informações úteis para apoiar o estudo técnico comparativo, possibilitando a construção de um valor médio, gerando uma economia de horas de pesquisa manual.

Segue abaixo o fluxograma do sistema agêntico para geração de documentos licitatórios.

4.2.2 Agente de Auditoria

A administração pública enfrenta hoje um grande desafio: realizar auditorias cada vez mais complexas e detalhadas com equipes reduzidas e prazos apertados. Pensando nisso, foi criado o "Agente de Auditoria", uma solução tecnológica que funciona como uma plataforma web. O objetivo principal dessa ferramenta é usar a IA generativa para assistir os auditores em todas as etapas do seu trabalho, desde o planejamento inicial até a entrega dos relatórios finais e planos de ação.

A ideia central não é substituir o auditor, mas retirar dele o esforço das tarefas repetitivas

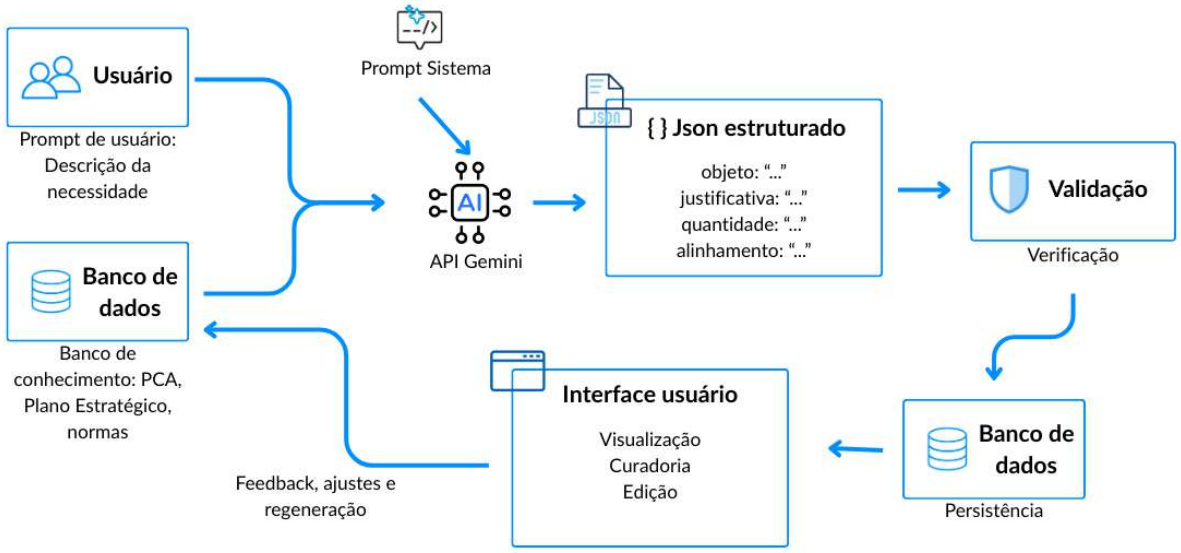

Figura 4.1: fluxograma do sistema agêntico LIA

e manuais. O sistema aceita arquivos em vários formatos, como PDF, Word e Excel. Ele processa essas informações e gera rascunhos de documentos em formato HTML, que podem ser convertidos para PDF com um clique.

Os objetivos específicos do projeto incluem:

- Padronizar documentos: Garantir que todos os relatórios sigam o mesmo modelo e qualidade.
- Cruzar informações: O agente consegue ler o plano anual, a matriz de riscos e outros documentos ao mesmo tempo para encontrar conexões importantes.
- Melhorar a comunicação: O sistema cria um canal formal onde a equipe de auditoria envia perguntas e recebe respostas e documentos das unidades auditadas.
- Revisão humana: Antes de qualquer documento ficar pronto, o auditor pode editar o texto em uma tela intermediária de curadoria. Isso garante que a responsabilidade final seja sempre de um humano.

O desenvolvimento do Agente de Auditoria não foi apenas escrever código de computador; envolveu entender a fundo como os auditores trabalham. O processo foi dividido em etapas:

Primeiro, foi feito um estudo dos requisitos. Foi feita uma análise documental e histórica com o objetivo de compor uma base de conhecimento. Também houve conversas com auditores experientes com o objetivo de encontrar gargalos no processo.

Em seguida, iniciou-se o trabalho de engenharia de prompts. Para cada tipo de documento e em cada campo a ser gerado, foi criada uma instrução específica. O sistema deve instruir o agente a agir como um especialista, ler os riscos listados em um documento, transformá-los em procedimentos de auditoria e sugerir o uso de ferramentas, caso haja necessidade.

Na fase de implementação, foram criadas funcionalidades extras. Além de gerar textos, o sistema tem um assistente virtual com quem o auditor pode conversar. Esse assistente tem

acesso a todo o processo de auditoria, além de fontes externas, permitindo que o auditor seja auxiliado por um agente completamente inteirado do andamento do processo.

O *backend* foi construído utilizando o framework FastAPI. A camada de inteligência do sistema vem da integração com o Google Gemini. A escolha por essa tecnologia específica se deu porque ela consegue ler e analisar grandes quantidades de texto.

Para o usuário, a experiência é simples, com uma interface visual responsiva. A aplicação cobre nove tarefas essenciais do dia a dia de um auditor: criar o Plano de Trabalho, montar a Matriz de Planejamento, fazer requisições de Documentos, gerenciar as respostas de quem está sendo auditado, acompanhar o progresso, organizar a Matriz de Achados e redigir o Relatório de Auditoria, o Parecer e o Plano de Ação.

Isso significa que, em vez de passar dias copiando e colando informações de um documento para outro, o auditor pode focar no que realmente importa: analisar os dados, verificar se as leis estão sendo cumpridas e sugerir melhorias para a gestão pública.

O sistema também promove transparência. As áreas de "Resposta do Auditado"e "Acompanhamento do Auditor"permitem que ambas as partes vejam o andamento das solicitações, prazos e entregas de documentos de forma clara.

4.2.3 IATA: Geração de Atas

Registrar o que acontece conforme a rotina é uma tarefa obrigatória, mas muitas vezes cansativa. Tradicionalmente, o trabalho de gravar a reunião ou sessão, para posteriomente ouvir o áudio pausadamente, e preparar documentos a partir destes textos exige um tempo e esforço incomum. O resultado, muitas vezes, são documentos despadronizados e incompletos.

Este agente traz uma solução para essa dificuldade. Em formato de aplicação web, o agente transcreve áudios, capturados no próprio navegador ou enviados pelo usuário. Posteriormente, a partir da escolha de um modelo de ata pelo usuário, o agente gera uma ata pré-preenchida. E, caso o usuário tenha interesse em personalizar sua própria ata, o sistema disponibiliza um recurso para geração de atas.

O ato de fazer atas manualmente no editor de texto tem dois grandes problemas. Primeiro, a transcrição do áudio. A tarefa de gravar para posteriormente ouvir e transcrever o texto duplica o tempo gasto com a reunião. Segundo, a falta de padrão: cada secretário faz a ata de uma forma diferente.

O objetivo principal é criar um sistema inteligente que realize a tarefa de transcrição de áudio e formatação do documento, deixando o usuário livre para focar apenas na revisão do conteúdo do texto. Para o usuário, o sistema oferece três áreas principais:

- Transcritor: O usuário insere o áudio e o agente transcreve-o.
- Gerador de Atas: O usuário escolhe o modelo de ata entre os disponíveis, e o sistema inicia a geração da ata a partir dda transcrição.
- Criação de templates: O usuário preenche um prompt com o objetivo de fornecer subsídios

para o agente criar um modelo de ata.

Em suma, esta aplicação mostra como os sistemas agênticos podem resolver problemas reais e cotidianos. Ao transformar um processo manual e repetitivo em um fluxo automatizado, a ferramenta não apenas economiza tempo, mas também eleva a qualidade da documentação oficial nas organizações.

4.2.4 Checagem de Fatos

Vivemos a era da "pós-verdade", onde desinformação, boatos, sátiras mal compreendidas, *clickbaits*, teorias da conspiração e afins se misturam ao jornalismo sério, criando um ambiente de insegurança. O problema não é apenas o texto falso, mas a sofisticação da fraude: vídeos manipulados, áudios tirados de contexto e imagens geradas por computador. Diante desse cenário, o trabalho propõe uma solução técnica: um framework que utiliza IA para automatizar a tarefa de classificar uma notícia.

O volume de dados gerado diariamente é humanamente impossível de ser verificado apenas por checadores manuais. Enquanto um jornalista investiga uma única alegação, milhares de novas mentiras são publicadas. Esse desequilíbrio favorece a desordem informacional, que corroi democracias, interfere em eleições e pode até custar vidas em crises de saúde pública.

Além disso, a desinformação evoluiu. Elas exploram vieses cognitivos e apelam para a emoção. Portanto, justifica-se a criação de um sistema que entenda o contexto, com analise multimodal e decisão declarativa.

A abordagem metodológica escolhida é a construção de um sistema agêntico, subdividido em quatro módulos.

1. Módulo I - Extração de recursos: Entrada inicial do framework. O checador foi preparado para receber quatro formatos de mídia: texto, áudio, vídeo e aúdio. Nesta etapa, o agente descreve a mídia em detalhes, a partir da extração de informações sobre o fato a ser verificado.
2. Módulo II - Mineração de argumento: Nesta fase, o sistema irá detectar alegações factíveis que podem ser verificadas. O agente realiza a leitura do fato e da descrição detalhada, e busca identificar hipóteses.
3. Módulo III - Recuperação de Evidências: Utilizando a técnica RAG, o sistema busca na internet evidências que possam corroborar ou contradizer as alegações levantadas. Para esta busca, foi utilizada a biblioteca *BraveSearch*.
4. Módulo IV - Classificação: No estágio final, o framework age como um juiz. Ele cruza a alegação original com o dossiê de evidências coletadas. Usando técnicas de aprendizado chamadas Few-Shot Learning, o modelo classifica a notícia em dez categorias possíveis. O produto final é um relatório legível que justifica a decisão, fechando o ciclo da checagem.

Segue o fluxo de checagem de fatos detalhando cada módulo:

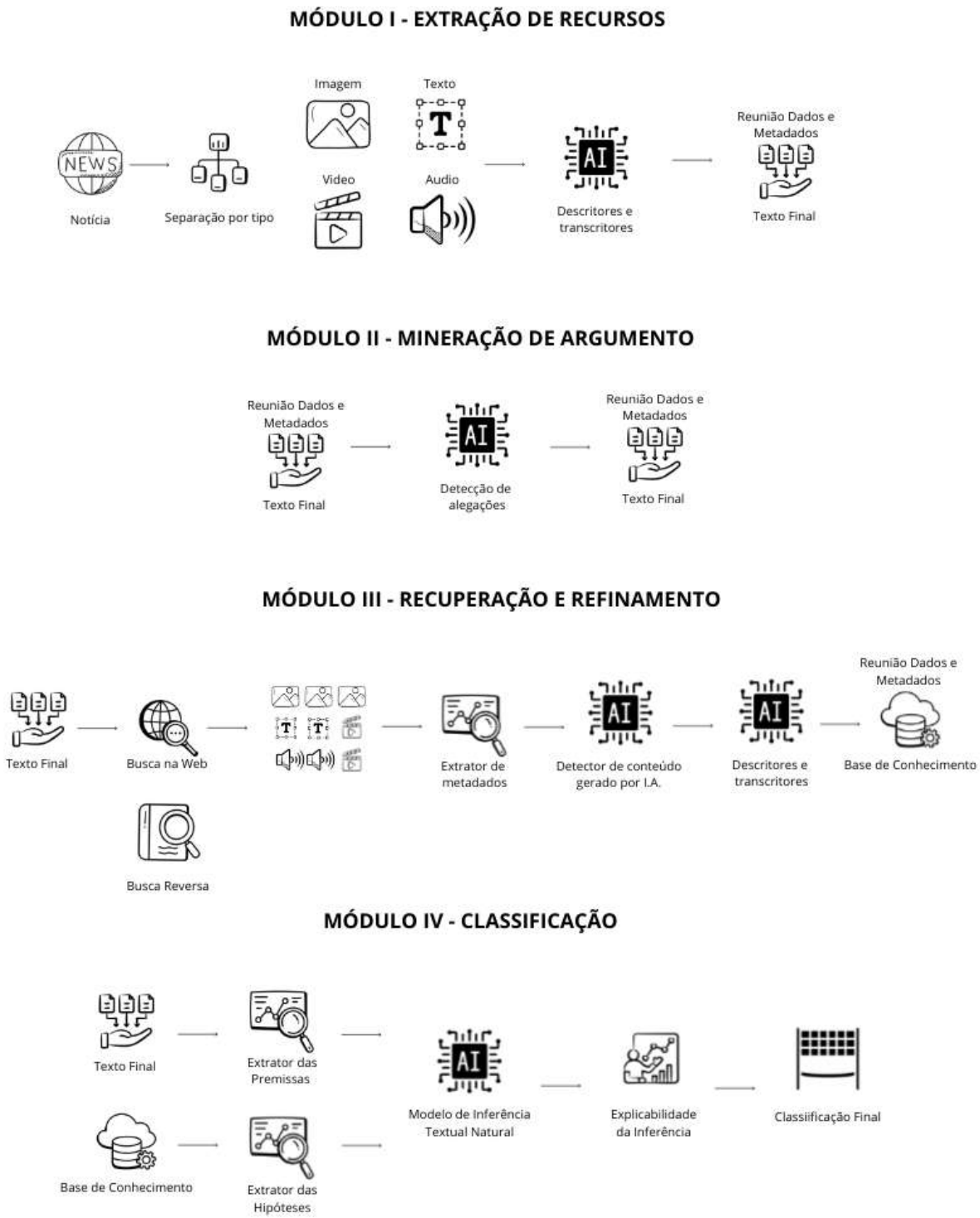

Figura 4.2: Fluxograma Checagem de Fatos

4.3 Chatbots

4.3.1 Normativos do CNJ

Em um dos encontros com pareceristas e analistas judiciários, na busca de oportunidades para automação com IA, identificou-se que buscar atos normativos demanda muito tempo e provoca desgaste. Tendo em vista que essa tarefa se repete ao longo do tempo, foi criado um assistente conversacional com o propósito de buscar atos normativos do CNJ e responder perguntas baseadas nesse contexto. Esse assistente foi desenvolvido de forma escalável, possibilitando que outro compêndio de normas possa ser consultado a depender do contexto da conversa do usuário.

A base de normativos do CNJ compreende a documentação dos atos normativos publicados até um período de referência. A construção dessa base seguiu um fluxo de engenharia de dados, estruturado em etapas sequenciais de coleta, pré-processamento, análise exploratória, fragmentação, vetorização e armazenamento.

A etapa de coleta foi conduzida por meio de técnicas de *web scraping*, implementadas em Python com o auxílio das bibliotecas *requests* e *BeautifulSoup*. O *scraper* foi programado para navegar pela estrutura hierárquica e paginada dos portais oficiais, identificando e extraindo os documentos relevantes de forma automatizada. Em seguida, o pré-processamento textual envolveu a remoção de marcações HTML e metadados irrelevantes, a normalização de caracteres e a padronização de seções jurídicas — tais como artigos, incisos e parágrafos —, além da correção de fragmentações e unificação de blocos textuais correlatos.

Previamente à vetorização, foi realizada uma análise exploratória do corpus com o objetivo de caracterizar sua estrutura estatística e linguística. Dentre os procedimentos adotados, destaca-se a verificação da conformidade da distribuição de frequência lexical com a Lei de Zipf, a qual confirmou a predominância de um reduzido conjunto de termos de alta recorrência — frequentemente associados a expressões procedimentais e deliberativas — em contraste com uma extensa cauda longa de vocábulos raros e específicos do domínio jurídico. Essa análise reforçou a necessidade de uma abordagem baseada em similaridade semântica, em detrimento de métodos tradicionais de correspondência lexical.

Para a fragmentação dos textos (*chunking*), foi empregada uma estratégia de segmentação recursiva baseada no módulo *Recursive Character Text Splitter* do *framework* LangChain, priorizando divisões por estruturas jurídicas — artigos, incisos e preâmbulos — de modo a preservar a coerência e o contexto normativo dos trechos gerados. Foram avaliadas três estratégias de segmentação distintas: Estruturada, Tamanho Fixo e Recursiva.

A etapa de vetorização semântica consistiu na codificação dos fragmentos em representações numéricas de alta dimensão (*embeddings*), utilizando três modelos de incorporação: BAAI/bge-m3, *sentence-transformers*/LaBSE e all-MiniLM-L6-v2. O processamento foi acelerado por GPU em regime de *batch processing*, e os *embeddings* resultantes foram armazenados em

um banco de dados PostgreSQL com a extensão PGVector, que permite buscas eficientes por similaridade utilizando métricas como distância cosseno, além de suportar estratégias híbridas que combinam consulta semântica com filtros estruturados em SQL. Cada registro no banco contém o fragmento textual, o respectivo *embedding*, metadados de origem (data, tipo de normativo) e referências estruturais do documento.

A avaliação do pipeline foi conduzida a partir de um conjunto de teste denominado *Silver Set*, composto por 100 *chunks* amostrados aleatoriamente para cada configuração. Para cada trecho, um modelo de linguagem de última geração (Google Gemini), integrado via N8N, gerou uma pergunta cuja resposta estivesse explicitamente contida no texto e uma paráfrase do mesmo conteúdo, simulando variações naturais de consulta. A combinação das nove configurações — três modelos de incorporação e três estratégias de segmentação — foi avaliada por meio das métricas Hit@K, *Mean Reciprocal Rank* (MRR), *Precision* e *Recall*.

Os resultados demonstraram desempenho expressivamente superior para a combinação BAAI/bge-m3 com a estratégia Estruturada, que alcançou Hit@10 de 78%, Hit@1 de 49% e MRR de 0,585. Uma etapa adicional de refinamento por *Cross-Encoder Reranking* confirmou a robustez dessa configuração, consolidando-a como a escolha técnica ideal para o sistema de recuperação semântica no domínio jurídico.

4.3.2 PalanqueIA

Esse *chatbot* foi desenvolvido para auxiliar pesquisadores políticos, candidatos e a sociedade civil como um todo na busca de planos de governo de candidatos políticos e na criação de propostas políticas. O agente possui dois modos de funcionamento: o modo consulta, onde o chatbot retorna planos de governo cadastrados no TSE na ocasião do registro de candidatura a partir de uma consulta inicial do usuário; e o modo criação, onde o agente, a partir de um tema definido pelo usuário, busca alguns resultados mais similares da base, e retorna um plano de governo em determinado tema.

A base de planos de governo foi construída a partir dos documentos oficiais apresentados pelos candidatos a cargos executivos municipais nas eleições de 2024, disponibilizados em formato PDF no sítio oficial do Tribunal Superior Eleitoral (TSE). A coleta dos arquivos e a extração textual foram realizadas com o auxílio da biblioteca PyPDF2, que permitiu processar e armazenar os documentos como objetos estruturados. Para a construção dos metadados — incluindo nome do candidato, partido e unidade federativa —, foram utilizados dados das candidaturas disponíveis na mesma plataforma.

O pré-processamento revelou uma considerável heterogeneidade nos formatos, tipos e estruturas dos planos de governo, o que representou um desafio para a conformidade da base. Arquivos em formato de imagem foram tratados por meio de técnicas de reconhecimento óptico de caracteres (OCR), embora aproximadamente 8% dos documentos tenham sido descartados em razão de resultados insatisfatórios no tratamento. Técnicas adicionais de limpeza incluíram

a remoção de caracteres indesejados e a normalização textual.

A fragmentação dos textos foi definida com tamanho de 256 *tokens* por fragmento e sobreposição (*overlay*) de 20 *tokens* entre segmentos consecutivos, de modo a preservar o contexto compartilhado entre divisões adjacentes. Para a geração dos *embeddings*, foi utilizado o modelo *sentence-transformers*/all-MiniLM-L6-v2, e os vetores resultantes foram armazenados no banco de dados vetorial ChromaDB, junto aos metadados correspondentes. Ao término do processo, a base registrou 14.197 propostas políticas, fragmentadas em 1.768.177 *chunks*, distribuídos pelas 27 unidades da federação, com predominância temática em áreas como educação, saúde, segurança e assistência social.

4.4 Oficina Prática

Esta oficina tem como foco ensinar e desenvolver habilidades práticas com LLMs e sua aplicação em sistemas agênticos. O programa cobre desde os conceitos iniciais, engenharia de Prompt e diferentes arquiteturas de recuperação de informação até a parte prática utilizando ferramentas como n8n e Python.

Apesar do acesso facilitado aos modelos de linguagem, criar sistemas que integrem memória, ferramentas externas e raciocínio complexo demanda uma nova forma de pensar a engenharia de software. A oficina procura resolver este problema ao transformar conceitos como embeddings, busca vetorial e frameworks aplicados em práticas de programação concretas.

Nesse contexto, os objetivos da oficina concentram-se na aplicação prática do conhecimento para a prototipagem de novos agentes. O intuito é habilitar os participantes a se aprofundarem na concepção teórica, utilizando abordagens de *Design Thinking*, até a construção efetiva de soluções funcionais. Busca-se que, ao final do processo, os desenvolvedores sejam capazes de arquitetar e implementar protótipos de agentes que resolvam problemas reais.

Por fim, este trabalho pode ser considerado um Produto Técnico-Tecnológico, considerando suas características: é inovador, ao introduzir metodologias de limite; é replicável, pois seu material didático permite a reprodução do treinamento em diferentes contextos; possui alto impacto, ao elevar significativamente a maturidade tecnológica das equipes envolvidas; e demonstra elevada aplicabilidade, entregando ferramentas e conhecimentos que podem ser imediatamente convertidos em soluções de software para modernização institucional.

4.5 Ajuste fino

4.5.1 Classificador entre Textos Sintéticos e Textos Humanos

Este estudo apresenta um novo classificador de texto sintético para português brasileiro, utilizando a arquitetura Transformers e aprendizado por transferência. Dada a limitada disponibilidade de modelos PT-BR, os autores desenvolveram um dataset misto, "Database", composto

por textos humanos de diversas fontes (Wikipédia, poemas, notícias, comentários) e textos sintéticos gerados por vários LLMs como Meta-Llama, Mistral e Gemma. O processo de criação do dataset envolveu uma cuidadosa seleção para garantir a diversidade de estilos e a relevância para a língua portuguesa, evitando textos traduzidos por IA.

Os modelos pré-treinados FacebookAI/xlm-roberta-base e neuralmind/bert-large-portuguese-cased (BERTimbau) foram ajustados e avaliados. Ambos demonstraram alta precisão na detecção de textos sintéticos, com o modelo FacebookAI/xlm-roberta-base sendo o de melhor desempenho. Em contraste, o detector openai-community/roberta-base-openai-detector teve um desempenho próximo ao aleatório para o português brasileiro. Os resultados ressaltam a importância de desenvolver LLMs específicos para a língua portuguesa, capazes de capturar suas nuances linguísticas e idiomáticas, e abrem caminho para futuros trabalhos focados na ampliação do corpus e na avaliação da robustez dos modelos contra ataques de ofuscação.

Com o propósito de treinar o modelo , foi desenvolvido um conjunto de dados misto integralmente em português brasileiro, denominado *Database*. A construção dessa base seguiu o princípio de maximizar a diversidade de fontes, estilos e formatos textuais, de modo a favorecer a capacidade de generalização do classificador.

A porção humana da base — composta por aproximadamente 94.300 registros — foi compilada a partir de cinco fontes distintas: *dumps* da Wikipédia em português (33,1%), conjunto de análise de sentimento do Kaggle com comentários de variadas plataformas (30,5%), notícias de portais brasileiros disponibilizadas no repositório HuggingFace (18,9%), poemas de autores brasileiros coletados do *website* Escritas.org (10,2%) e instruções humanas extraídas do *dataset* Cabrita-and-Guanaco-PTBR (7,3%). Foram excluídos textos traduzidos, uma vez que tradutores automáticos frequentemente empregam IA em sua geração, o que comprometeria a integridade da classe humana.

A porção sintética — igualmente balanceada com aproximadamente 94.300 registros — foi gerada a partir dos próprios textos humanos por meio de engenharia de *prompt*, utilizando cinco modelos generativos distintos, cada um representando 20% do subconjunto: Meta-Llama-3-70B-Instruct, Mistral-7B-Instruct-v0.2, Google Gemma-2b, Microsoft Phi-3-mini-4k-instruct e o modelo cnmoro/GPT4-500k. Os *prompts* empregados foram refinados iterativamente para garantir a produção de textos em português brasileiro.

O conjunto de dados final, com 188.600 registros, foi subdividido em treino (80%), validação (10%) e teste (10%).

4.5.2 Classificador Federado para Classes Judiciais

O artigo propõe uma abordagem colaborativa para o treinamento de modelos de IA no Poder Judiciário brasileiro por meio do aprendizado federado, técnica que permite o desenvolvimento conjunto de modelos sem a necessidade de compartilhamento de dados sensíveis. O estudo destaca os desafios enfrentados atualmente, como altos custos operacionais, fragmentação de

iniciativas, vieses regionais e riscos à privacidade previstos na LGPD. O aprendizado federado surge como alternativa viável, promovendo maior eficiência, proteção de dados e padronização entre os tribunais, além de possibilitar a criação de modelos mais representativos e generalizáveis.

A pesquisa experimental comparou o desempenho de modelos treinados de forma centralizada e federada, utilizando o modelo Longformer aplicado à classificação de documentos judiciais. Os resultados indicaram que o aprendizado federado pode ser utilizado no contexto do Poder Judiciário para treinamento de modelos de IA, sem a perda de qualidade na precisão da inferência.

4.5.3 Despesas de Campanha e Resultado das Eleições

O artigo investiga a relação entre gastos de campanha e desempenho eleitoral nas eleições de 2022 para o cargo de governador no Brasil, utilizando análise de correlação de Spearman e modelagem de regressão ridge. Os resultados apontam uma forte correlação positiva (r = 0,86) entre as despesas de campanha e a quantidade de votos válidos, indicando que candidatos que mais investem em suas campanhas tendem a obter maior sucesso eleitoral. Foi também desenvolvido o Indicador de Gasto Médio por Voto (IGMV), que permitiu comparar o custo médio do voto entre cargos e estados, revelando que o Acre apresentou o maior valor (R$ 37,79 por voto) e o Paraná o menor (R$ 3,74).

A modelagem preditiva demonstrou que o gasto eleitoral é a variável mais significativa para explicar a performance dos candidatos, seguido pela condição de reeleição, que favorece os incumbentes. O estudo destaca ainda o desenvolvimento de um painel interativo de inteligência de negócios para análise de diferentes eleições e cargos. Conclui-se que o financiamento de campanha é um fator determinante no desempenho eleitoral no Brasil, reforçando a necessidade de políticas públicas que promovam maior equidade financeira entre os candidatos e assegurem a transparência e a competitividade democrática.

4.5.4 Democracia Inclusiva: Votos dos Eleitores Analfabetos e com Mobilidade Reduzida

O estudo analisa as taxas de abstenção em eleições brasileiras, focando em eleitores com deficiência e analfabetos. A pesquisa quantitativa utilizou dados do TSE e do Instituto Brasileiro de Geografia e Estatística (IBGE) para descrever o evento e identificar um padrão entre os eleitores faltosos. Os resultados indicam um crescimento contínuo do eleitorado geral no Brasil, com as taxas de abstenção entre eleitores com deficiência e analfabetos sendo significativamente maiores do que as do eleitorado geral. Além disso, as regiões Sudeste e Nordeste registraram as maiores médias de eleitores aptos a votar, bem como as maiores quantidades de abstenções.

A análise temporal revelou um aumento nas taxas de abstenção para eleitores com deficiência, atingindo um pico em 2020 (39,21%), e para eleitores analfabetos, que subiram de 40,09% em 2012 para 52,08% em 2022. A análise espacial identificou padrões regionais significativos,

com o Nordeste apresentando a maior concentração de eleitores analfabetos e o Sudeste e Sul predominando na correlação de alta-alta para este grupo em 2024. Para eleitores com deficiência, a distribuição espacial foi menos uniforme, com correlações de baixa-alta no estado de São Paulo em 2022.

4.5.5 Resolução 615/2025 do CNJ e Agenda 2030

O estudo analisa os impactos da Resolução nº 615/2025 do CNJ sobre a governança e o desenvolvimento de soluções de IA no Poder Judiciário brasileiro. Por meio de uma pesquisa qualitativa documental, o trabalho identifica como a norma contribui para a modernização do Judiciário ao estabelecer diretrizes éticas, de transparência e responsabilidade no uso da IA, promovendo inovação sustentável e alinhada aos Objetivos de Desenvolvimento Sustentável (ODS).

Os principais ODS impactados são o 16 (Paz, Justiça e Instituições Eficazes) e o 9 (Indústria, Inovação e Infraestrutura), com reflexos positivos também nos ODS 4, 8, 10 e 17. Conclui-se que a Resolução 615/2025 representa um marco na consolidação de uma governança tecnológica ética e inclusiva, fortalecendo a eficiência, a transparência e o compromisso social do Poder Judiciário.

4.5.6 Estudo de normas internacionais para uma governança de IA segura e responsável nas organizações

O avanço acelerado da IA, especialmente dos LLMs, tem gerado preocupações significativas quanto à segurança e à responsabilidade em seu uso. Governos e organizações, como o Poder Judiciário brasileiro, têm respondido com regulamentações específicas. Este artigo examina os controles de uso seguro e responsável da IA estabelecidos pela Resolução nº 615/2025 do CNJ e os correlaciona com frameworks internacionais de gestão e gerenciamento de riscos, notadamente a norma ISO/IEC 42001:2023 da Organização Internacional de Normalização e Comissão Eletrotécnica Internacional, e o Artificial Intelligence Risk Management Framework (AI RMF) 1.0 do Instituto Nacional de Padrões e Tecnologia (NIST). O estudo demonstra como as diretrizes normativas do Judiciário podem ser operacionalizadas e auditadas por meio da adoção e adaptação de práticas globais, com foco em aspectos críticos como avaliação de impacto algorítmico, governança de dados, transparência e mitigação de vieses.

4.6 Discussão

O presente trabalho partiu da ideia de que a disponibilidade de grandes conjuntos de dados, o poder computacional acessível e a maturidade dos modelos de linguagem criariam as condições favoráveis para uma transformação digital no serviço público. Partindo desta hipótese, iniciamos uma pesquisa aplicada em diversos agentes e outros estudos complementares com o propósito

de validar esta questão.

O principal achado deste estudo é a demonstração prática de que é possível desenvolver e implementar agentes de IA eficazes no contexto do serviço público brasileiro. Os resultados apresentados sob a forma de um conjunto de produtos funcionais e artefatos tecnológicos, não apenas validam essa hipótese, mas também oferecem um panorama claro das oportunidades e desafios inerentes a essa transição. No geral, os resultados foram muito promissores. Houve um retorno muito positivo dos usuários ao verem suas longas tarefas serem desempenhadas com sucesso por uma IA.

O fato da engenharia de IA ser uma área emergente, tivemos muitas mudanças radicais no processo. Ao longo do projeto, novas ferramentas foram surgindo, frameworks foram sendo desenvolvidos, e praticamente, toda a referência utilizada está concentrada a partir do ano de 2024. Os próprios modelos de linguagem, a espinha dorsal dos agentes, tiveram evoluções consideráveis desde o início dos estudos, porém, sempre positivas.

A metodologia de quatro etapas para criação de agentes de IA, melhorada ao longo do projeto, emergiu como um resultado tão relevante quanto os próprios agentes. A incorporação de técnicas de *Design Thinking* na fase de mapeamento foi crucial para garantir que as soluções fossem desenvolvidas a partir das dores reais dos usuários, e não de abstrações tecnológicas. Isso mitiga o risco de criar ferramentas tecnicamente impressionantes, mas funcionalmente inúteis.

De forma geral, os resultados foram bem promissores. Em um curto espaço de tempo, o processo de criação de um agente de IA saiu de muito complexo para muito simples.

4.7 Limitações e Trabalhos Futuros

Em contrapartida, é importante reconhecer as limitações impostas a este estudo. O projeto possui inúmeras, de naturezas diversas. Neste trabalho, citamos as mais relevantes:

1. **Técnicas:** Os agentes desenvolvidos são altamente dependentes de LLMs comerciais. Nos testes realizados, os melhores resultados foram obtidos a partir dos modelos de última geração de empresas privadas. Comparativamente, os modelos de código aberto não possuem a precisão e a flexibilidade técnica para desempenhar inúmeras tarefas. Outra vantagem importante dos modelos de última geração é a multimodalidade altamente avançada. Outra abordagem de limitação técnica relevante é a integração dos produtos com sistemas utilizados pelos órgãos públicos. Essa variedade de integrações dificulta muito o desenvolvimento colaborativo, uma das imposições pela governança imposta pelo CNJ.
2. **Escopo:** O objetivo principal foi analisar e propor um modelo, validado por meio de protótipos funcionais. Esses protótipos demonstram a viabilidade técnica e o valor potencial, mas não são sistemas prontos para produção em larga escala. A implantação é frequentemente mais difícil do que o desenvolvimento.
3. **Propagação do erro:** Cada tarefa realizada pelo agente possui uma probabilidade de

falha. Ao longo do processo como um todo, essa probabilidade se multiplica, tornando alguns agentes inviáveis em sua implementação, apesar de provas de conceitos isolados satisfatórios.

4. **Riscos de ações de escrita:** Fornecer autonomia de escrita no mundo real ao agente expõe o processo a riscos de segurança e falhas com graves consequências. Esse nível de permissão exige alto monitoramento e controle.

Apesar de muitos avanços, ainda é possível vislumbrar um longo caminho pela frente. A adaptação ao modelo de governança proposto pelo CNJ exige um longo processo de melhorias. Os protótipos, desenvolvidos ao longo do processo, deverão ser incorporados no ambiente de produção dos órgãos públicos para consolidar os benefícios. Esse processo de implementação exigirá um grande esforço, esforço este mínimo comparado aos benefícios oferecidos pelos agentes.

Capítulo 5

Conclusão

Ao chegar ao fim desta jornada, a sensação é que estamos diante de uma nova revolução no nosso modo de viver e trabalhar. Em cada reunião realizada e em cada ideia nova, um recurso se desvendava; em cada frustração, uma nova alegria; pudemos então perceber que estávamos diante de um caminho sem volta.

O que se iniciou com uma visão distante, a visão de criar soluções práticas em sistemas inteligentes, agora tem forma e função. Decidimos sair do estado de inércia do serviço rumo a um novo paradigma na execução dos nossos serviços. A prestação de serviços está prestes a mudar, e nos sentimos parte dessa mudança.

Esta pesquisa aplicada demonstrou a viabilidade e a eficácia de uma metodologia para o desenvolvimento e implementação de agentes de IA, validando todos os objetivos propostos. O resultado do trabalho traz ganhos teóricos e práticos, por meio de protótipos funcionais.

A pesquisa indica que estamos no momento certo para essa mudança, pois os avanços tecnológicos estão se alinhando. Um dos principais achados é que é preciso haver um equilíbrio entre a autonomia dos agentes e a supervisão humana, ou seja, a solução deve ser vista como uma ferramenta que potencializa o que os humanos já fazem, e não como um substituto. A IA deve ser utilizada com supervisão humana, com a mente humana sempre tendo a palavra final.

Embora desafios como a redução de vieses e a explicação continuem a existir, a pesquisa conclui que a transformação digital por meio da IA, quando orientada por valores éticos, é uma força significativa para aprimorar a democracia e construir uma sociedade mais justa e eficiente. O sucesso no futuro vai depender da habilidade de adaptar e utilizar essas tecnologias de maneira responsável e focada no bem de todos.

E por fim, olhando para dois anos atrás, este trabalho modificou meu modo de agir e pensar. Não é sobre tecnologia. É sobre servir à sociedade, sendo mais empático, mais produtivo, e mais justo.

Referências

Alla, L., Hmioui, A., & Bentalha, B. (2024). *AI and Data Engineering Solutions for Effective Marketing*. IGI Global Business Science Reference (an imprint of IGI Global).

Alto, V. (2025). *AI Agents in Practice: Design, Implement, and Scale Autonomous AI Systems for Production* [First published: July 2025]. Packt Publishing.

Biswas, A., & Talukdar, W. (2025). *Building Agentic AI Systems: Create intelligent, autonomous AI agents that can reason, plan, and adapt* [First published: April 2025]. Packt Publishing.

Bornet, P., Wirtz, J., Davenport, T. H., De Cremer, D., Evergreen, B., Fersht, P., Gohel, R., & Khiyara, S. (2025). *Agentic Artificial Intelligence* [Ebook version: March 2025]. Pascal Bornet (Self-Published/Internal publication details mentioned).

Brasil. (2024, setembro). Decreto nº 12.198, de 24 de setembro de 2024: Institui a Estratégia Federal de Governo Digital para o período de 2024 a 2027 e a Infraestrutura Nacional de Dados [Acesso em: 18 de julho de 2026]. *Presidência da República*.

Brown, T. (2008). Design Thinking [Disponível em: https://hbr.org/2008/06/design-thinking]. *Harvard Business Review*, *86*(6), 84–92.

Caldwell, T. R. (2025). *The AI Engineering Bible: The Complete and Up-to-Date Guide to Build, Develop and Scale Production Ready AI Systems*. Self-Published/Copyright 2025.

Chan, A., Wei, K., Huang, S., Rajkumar, N., Perrier, E., Lazar, S., Hadfield, G. K., & Anderljung, M. (2025). Infrastructure for AI Agents. *arXiv preprint arXiv:2501.10114*. https://doi.org/10.48550/arXiv.2501.10114

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., et al. (2024). A Survey on Evaluation of Large Language Models. *ACM Transactions on Intelligent Systems and Technology*, *15*(3), 1–45. https://doi.org/10.1145/3641289

CognitiveGroup. (2023). *Cognitive Task Analysis and AI Agents: Extracting Human Expertise for Enhanced AI Performance* [Acessado em 11 dez. 2025]. CognitiveGroup. https://cognitivegroup.com/blog.html?id=cta-and-ai-agents

Conselho Nacional de Justiça. (2025, março). Resolução nº 615, de 11 de março de 2025: Estabelece diretrizes para o desenvolvimento, utilização e governança de soluções desenvolvidas com recursos de inteligência artificial no Poder Judiciário [Acesso em: 18 de julho de 2026]. *Conselho Nacional de Justiça*.

Conselho Nacional de Justiça (CNJ). (2025). Selo Judiciário Inovador 2025 – 2ª Edição do Prêmio Inovação do Poder Judiciário [Acesso e lista de vencedores do Selo Judiciário Inovador na 2ª edição do Prêmio Inovação do Poder Judiciário.]. https://www.cnj.jus.br/gestao-da-justica/politicas-judiciarias-nacionais-programaticas/politica-nacional-de-

gestao-da-inovacao/premio-inovacao-do-poder-judiciario/2a-edicao/selo-judiciario-inovador-2025/

Crandall, B., Klein, G., & Hoffman, R. R. (2006). *Working Minds: A Practitioner's Guide to Cognitive Task Analysis*. The MIT Press.

Davies, R. (2025). *AI Engineering in Practice* [MEAP Edition; Copyright 2025]. Manning Publications.

De Chezelles, T. L. S., Gasse, M., Drouin, A., Caccia, M., Boisvert, L., Thakkar, M., Marty, T., et al. (2024). The BrowserGym Ecosystem for Web Agent Research. *arXiv preprint arXiv:2412.05467*. https://doi.org/10.48550/arXiv.2412.05467

Deng, Z., Guo, Y., Han, C., Ma, W., Xiong, J., Wen, S., & Xiang, Y. (2025). AI Agents Under Threat: A Survey of Key Security Challenges and Future Pathways. *ACM Computing Surveys*, *57*(7), 1–36. https://doi.org/10.1145/3716628

Dennis, A. R., Lakhiwal, A., & Sachdeva, A. (2023). AI Agents as Team Members: Effects on Satisfaction, Conflict, Trustworthiness, and Willingness to Work With. *Journal of Management Information Systems*, *40*(2), 307–337. https://doi.org/10.1080/07421222.2023.2196773

Dhuliawala, S., Komeili, M., Xu, J., Raileanu, R., Li, X., Celikyilmaz, A., & Weston, J. (2023). Chain-of-Verification Reduces Hallucination in Large Language Models [Available at https://arxiv.org/abs/2309.11495]. https://doi.org/10.48550/arXiv.2309.11495

Drouin, A., Gasse, M., Caccia, M., Laradji, I. H., Del Verme, M., Marty, T., et al. (2024). WorkArena: How Capable Are Web Agents at Solving Common Knowledge Work Tasks? *arXiv preprint arXiv:2403.07718*. https://doi.org/10.48550/arXiv.2403.07718

Fung, P., Bachrach, Y., Celikyilmaz, A., Chaudhuri, K., Chen, D., Chung, W., et al. (2025). Embodied AI Agents: Modeling the World. *arXiv preprint arXiv:2506.22355*. https://doi.org/10.48550/arXiv.2506.22355

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., et al. (2024). Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics*, *50*(3), 1097–1179. https://doi.org/10.1162/coli_a_00524

Gao, L., Ma, X., Lin, J., & Callan, J. (2023). Precise Zero-Shot Dense Retrieval without Relevance Labels. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1762–1777. https://doi.org/10.18653/v1/2023.acl-long.99

Gao, S., Fang, A., Huang, Y., Giunchiglia, V., Noori, A., Schwarz, J. R., et al. (2024). Empowering biomedical discovery with AI agents. *Cell*, *187*(22), 6125–6151. https://doi.org/10.1016/j.cell.2024.09.022

Grand View Research. (2024). AI Agents Market Size, Share e Trends Analysis Report By Component, By Deployment, By Enterprise Size, By End-use, By Region, And Segment Forecasts, 2025 - 2030 [Acessado em: 2025]. https://www.grandviewresearch.com/industry-analysis/ai-agents-market-report

Grootendorst, M., & Alammar, J. (2025). *An Illustrated Guide to AI Agents: Visual Intuition for Reasoning, Multimodal, and Diffusion LLMs* [Early Release]. O'Reilly Media, Inc.

Gupta, M., & Sen, N. (2025). *Model Context Protocol Advanced AI Agents for Beginners* [Copyright 2025]. Mehul Gupta; Niladri Sen (First Edition).

Hadfield, G. K., & Koh, A. (2025). An Economy of AI Agents. *arXiv preprint arXiv:2509.01063*. https://doi.org/10.48550/arXiv.2509.01063

He, H., Yao, W., Ma, K., Yu, W., Dai, Y., Zhang, H., et al. (2024). WebVoyager: Building an End-to-End Web Agent with Large Multimodal Models. *arXiv preprint arXiv:2401.13919*. https://doi.org/10.48550/arXiv.2401.13919

Huang, X., Liu, W., Chen, X., Wang, X., Wang, H., Lian, D., et al. (2024). Understanding the planning of LLM agents: A survey. *arXiv preprint arXiv:2402.02716*. https://doi.org/10.48550/arXiv.2402.02716

Huyen, C. (2024). *AI Engineering: Building Applications with Foundation Models* [First Edition: December 2024]. O'Reilly Media, Inc.

International Organization for Standardization & International Electrotechnical Commission. (2023, dezembro). *ISO/IEC 42001:2023: Information technology – Artificial intelligence – Management system* [Acesso em: 18 de julho de 2026]. International Organization for Standardization. https://www.iso.org/standard/81230.html

Ishida, T., Murakami, Y., Lin, D., & Lhaksmana, K. M. (2025). AI Agents: From Concept to Code to Commerce—A Survey Bridging AAMAS and LLMs. *TechRxiv*. https://doi.org/10.36227/techrxiv.175616139.97637301/v1

J.Ex. (2025). Destaques do Prêmio de Inovação J.Ex 2025 [5ª edição do Prêmio de Inovação J.Ex]. https://docs.google.com/document/d/1k-cez8VsGN4JcfIOGGwy-pQreUhzXqxikrW2EuXIiKI/edit

Kapoor, S., Stroebl, B., Siegel, Z. S., Nadgir, N., & Narayanan, A. (2024). AI Agents That Matter. *arXiv preprint arXiv:2407.01502*. https://doi.org/10.48550/arXiv.2407.01502

Koh, J. Y., Lo, R., Jang, L., Duvvur, V., Lim, M. C., Huang, P.-Y., et al. (2024). VisualWebArena: Evaluating Multimodal Agents on Realistic Visual Web Tasks. *arXiv preprint arXiv:2401.13649*. https://doi.org/10.48550/arXiv.2401.13649

Kolt, N. (2025). Governing AI Agents. *arXiv preprint arXiv:2501.07913*. https://doi.org/10.48550/arXiv.2501.07913

Labaschin, B., Wallace, J. A., Brookins, A., & Singh, M. (2025). *Managing Memory for AI Agents: Memory, Models, and the Architecture of Adaptability* [First Edition: October 2025]. O'Reilly Media, Inc.

LangChain. (2024). State of AI Agents Report 2024 [Acessado em: 2025]. https://www.langchain.com/stateofaiagents

Lanham, M. (2025). *AI Agents in Action* [Copyright 2025]. Manning Publications Co.

Larson, J., & Truitt, S. (2024). GraphRAG: Unlocking LLM Discovery on Narrative Private Data [GraphRAG usa grafos de conhecimento derivados por LLMs para melhorar Retrieval-

Augmented Generation em dados privados. Disponível em https://www.microsoft.com/en-us/research/blog/graphrag-unlocking-llm-discovery-on-narrative-private-data/].

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks [Available at https://arxiv.org/abs/2005.11401].

Li, X., Wang, S., Zeng, S., Wu, Y., & Yang, Y. (2024). A survey on LLM-based multi-agent systems: workflow, infrastructure, and challenges. *Vicinagearth*, *1*(1), 9. https://doi.org/10.1007/s44336-024-00009-2

Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R., Yang, Y., Shi, S., & Tu, Z. (2023). Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate [Código disponível em https://github.com/Skytliang/Multi-Agents-Debate].

Lipenkova, J. (2025). *The Art of AI Product Development: Delivering business value*. Manning Publications Co.

Liu, Z., Zhang, Y., Li, P., Liu, Y., & Yang, D. (2023). A Dynamic LLM-Powered Agent Network for Task-Oriented Agent Collaboration. *arXiv preprint arXiv:2310.02170*. https://doi.org/10.48550/arXiv.2310.02170

Mandikal, P., & Mooney, R. (2024). Sparse Meets Dense: A Hybrid Approach to Enhance Scientific Document Retrieval [Available at https://arxiv.org/abs/2401.04055]. https://doi.org/10.48550/arXiv.2401.04055

Ministério da Ciência, Tecnologia e Inovação. (2024, julho). *IA para o Bem de Todos: Proposta de Plano Brasileiro de Inteligência Artificial 2024-2028* (Documento de Proposta) (Apresentado na Reunião do Pleno do CCT em 29 de julho de 2024). Conselho Nacional de Ciência e Tecnologia. Brasília.

Ning, X., Lin, Z., Zhou, Z., Wang, Z., Yang, H., & Wang, Y. (2024). Skeleton-of-Thought: Prompting LLMs for Efficient Parallel Generation [Available at https://arxiv.org/abs/2307.15337v3]. https://doi.org/10.48550/arXiv.2307.15337

Pant, S., & Viswanathan, M. (2025, maio). *5 Levels of Agentic AI Intelligence for Enterprise Use* [Acesso em: 22 jan. 2026]. Outshift by Cisco. https://outshift.cisco.com/blog/agentic-ai-intelligence-for-enterprise-use

Pisarevskaya, D., & Zubiaga, A. (2025). Zero-shot and Few-shot Learning with Instruction-following LLMs for Claim Matching in Automated Fact-checking. *Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025)*, 9721–9736. https://aclanthology.org/2025.coling-main.650/

Prophecy Market Insights. (2025). Autonomous AI Agents Market Size, Share, Trends, Analysis and Forecast till 2035 [Acessado em: 2025]. https://www.prophecymarketinsights.com/market_insight/autonomous-ai-agents-market-6013

Rackauckas, Z. (2024). RAG-Fusion: A New Take on Retrieval-Augmented Generation [Available at https://arxiv.org/abs/2402.03367]. https://doi.org/10.48550/arXiv.2402.03367

Raieli, S., & Iuculano, G. (2025). *Building AI Agents with LLMs, RAG, and Knowledge Graphs: A practical guide to autonomous and modern AI agents* [First published: July 2025]. Packt Publishing.

Riedl, M. O., & Desai, D. R. (2025). AI Agents and the Law. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, *8*(3), 2189–2198. https://doi.org/10.1609/aies.v8i3.36705

Sapkota, R., Roumeliotis, K. I., & Karkee, M. (2025). AI Agents vs. Agentic AI: A Conceptual Taxonomy, Applications and Challenges. *arXiv preprint arXiv:2505.10468*. https://doi.org/10.48550/arXiv.2505.10468

Sevilla, J., & Roldán, E. (2024, maio). *Training compute of frontier AI models grows by 4–5× per year* (rel. técn.) (Accessed: 2025-11-09). Epoch AI. https://epoch.ai/blog/training-compute-of-frontier-ai-models-grows-by-4-5x-per-year

Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., & Yao, S. (2023). Reflexion: Language Agents with Verbal Reinforcement Learning [Presented at the 37th Conference on Neural Information Processing Systems (NeurIPS 2023)]. *Advances in Neural Information Processing Systems 36*. https://proceedings.neurips.cc/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html

Stratis, K. (2025). *AI Agents with MCP Model Context Protocol for Building Clients, Services, and End-to-End Agents* [Early Release]. O'Reilly Media.

Vittie, L. M. (2025, março). *AI Agents vs. Agentic AI: Understanding the Difference* [Acesso em: 22 jan. 2026]. F5. https://www.f5.com/pt_br/company/blog/ai-agents-vs-agentic-ai-understanding-the-difference

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., et al. (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science*, *18*(6), 186345. https://doi.org/10.1007/s11704-024-40231-1

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models [Available at https://arxiv.org/abs/2201.11903].

Weston, J., & Sukhbaatar, S. (2023). System 2 Attention (is something you might need too) [Available at https://arxiv.org/abs/2311.11829]. https://doi.org/10.48550/arXiv.2311.11829

Xing, F. (2025). Designing Heterogeneous LLM Agents for Financial Sentiment Analysis. *ACM Transactions on Management Information Systems*, *16*(1), 1–24. https://doi.org/10.1145/3688399

Yan, S.-Q., Gu, J.-C., Zhu, Y., & Ling, Z.-H. (2024). Corrective Retrieval Augmented Generation [Available at https://arxiv.org/abs/2401.15884]. https://doi.org/10.48550/arXiv.2401.15884

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). Tree of Thoughts: Deliberate Problem Solving with Large Language Models [Presented

at the 37th Conference on Neural Information Processing Systems (NeurIPS 2023), New Orleans, United States]. *Advances in Neural Information Processing Systems 36*. https://proceedings.neurips.cc/paper/2023/hash/271db9922b8d1f4dd7aaef84ed5ac703-Abstract.html

Yehudai, A., Eden, L., Li, A., Uziel, G., Zhao, Y., Bar-Haim, R., et al. (2025). Survey on Evaluation of LLM-based Agents. *arXiv preprint arXiv:2503.16416*. https://doi.org/10.48550/arXiv.2503.16416

Zheng, B., Gou, B., Kil, J., Sun, H., & Su, Y. (2024). GPT-4V(ision) is a Generalist Web Agent, if Grounded. *arXiv preprint arXiv:2401.01614*. https://doi.org/10.48550/arXiv.2401.01614

Zheng, H. S., Mishra, S., Chen, X., Cheng, H.-T., Chi, E. H., Le, Q. V., & Zhou, D. (2024). Take a Step Back: Evoking Reasoning via Abstraction in Large Language Models [Available at https://arxiv.org/abs/2310.06117v2]. https://doi.org/10.48550/arXiv.2310.06117