DISSERTAÇÃO DE MESTRADO PROFISSIONAL

PIPELINE DE INTELIGÊNCIA ARTIFICIAL PARA DECOMPOSIÇÃO E ANÁLISE ESTRUTURADA DE CENÁRIOS COMPLEXOS VIA METODOLOGIA 5W3H

Caio César dos Santos Cruz

Presidente - Orientador: Prof. Dr. Robson de Oliveira Albuquerque
Coorientador: Prof. Dr. João José Costa Gondim

Programa de Pós-Graduação Profissional em Engenharia Elétrica

UNIVERSIDADE DE BRASÍLIA
Faculdade de Tecnologia

DISSERTAÇÃO DE MESTRADO PROFISSIONAL

PIPELINE DE INTELIGÊNCIA ARTIFICIAL PARA DECOMPOSIÇÃO E ANÁLISE ESTRUTURADA DE CENÁRIOS COMPLEXOS VIA METODOLOGIA 5W3H

Caio César dos Santos Cruz

Dissertação de Mestrado Profissional submetida ao Departamento de Engenharia Elétrica como requisito parcial para obtenção do grau de Mestre em Engenharia Elétrica

Banca Examinadora

Prof. Dr. Robson de Oliveira Albuquerque _____
Presidente - Orientador

Prof. Dr. João José Costa Gondim _____
Coorientador

Prof. Dr. Laerte Peotta de Melo, ENE/FT/UnB _____
Examinador Interno

Prof. Dr. Cristoffer Leite, JADS / Technische Universiteit Eindhoven (TU/e), Países Baixos _____
Examinador Externo

Prof. Dr. Georges Daniel A. Nze, ENE/FT/UnB _____
Examinador Suplente

FICHA CATALOGRÁFICA

CRUZ, CAIO
PIPELINE DE INTELIGÊNCIA ARTIFICIAL PARA DECOMPOSIÇÃO E ANÁLISE ESTRUTURADA DE CENÁRIOS COMPLEXOS VIA METODOLOGIA 5W3H [Distrito Federal] 2026.
xvi, 80 p., 210 x 297 mm (ENE/FT/UnB, Mestre, Engenharia Elétrica, 2026).
Dissertação de Mestrado Profissional - Universidade de Brasília, Faculdade de Tecnologia.
Departamento de Engenharia Elétrica
Publicação: PPEE.MP.127

1. Inteligência Artificial Agêntica	2. RAG Híbrido
3. Engenharia de Prompt	4. Sistemas de Apoio à Decisão
I. ENE/FT/UnB	II. Título (série)

REFERÊNCIA BIBLIOGRÁFICA

CRUZ, C. (2026). *PIPELINE DE INTELIGÊNCIA ARTIFICIAL PARA DECOMPOSIÇÃO E ANÁLISE ESTRUTURADA DE CENÁRIOS COMPLEXOS VIA METODOLOGIA 5W3H*. Dissertação de Mestrado Profissional, Departamento de Engenharia Elétrica, Universidade de Brasília, Brasília, DF, 80 p.

CESSÃO DE DIREITOS

AUTOR: Caio César dos Santos Cruz
TÍTULO: PIPELINE DE INTELIGÊNCIA ARTIFICIAL PARA DECOMPOSIÇÃO E ANÁLISE ESTRUTURADA DE CENÁRIOS COMPLEXOS VIA METODOLOGIA 5W3H.
GRAU: Mestre em Engenharia Elétrica ANO: 2026

É concedida à Universidade de Brasília permissão para reproduzir cópias desta Dissertação de Mestrado e para emprestar ou vender tais cópias somente para propósitos acadêmicos e científicos. Do mesmo modo, a Universidade de Brasília tem permissão para divulgar este documento em biblioteca virtual, em formato que permita o acesso via redes de comunicação e a reprodução de cópias, desde que protegida a integridade do conteúdo dessas cópias e proibido o acesso a partes isoladas desse conteúdo. O autor reserva outros direitos de publicação e nenhuma parte deste documento pode ser reproduzida sem a autorização por escrito do autor.

Caio César dos Santos Cruz
Depto. de Engenharia Elétrica (ENE) - FT
Universidade de Brasília (UnB)
Campus Darcy Ribeiro
CEP 70919-970 - Brasília - DF - Brasil

AGRADECIMENTOS

Aos meus orientadores, Prof. Dr. Robson de Oliveira Albuquerque e Prof. Dr. João José Costa Gondim, pela condução segura, rigor metodológico e direcionamento acadêmico indispensável ao longo desta pesquisa.

Ao Dr. Dino Macedo Amaral, pelo tempo dedicado e pelas contribuições técnicas que ajudaram a pavimentar e enriquecer os resultados deste trabalho.

Ao Sr. Geraldo Couto, da Via Appia, por confiar em mim e na minha capacidade de entrega, oferecendo suporte contínuo e parceria estrutural, especialmente durante os momentos mais desafiadores.

Aos meus pais, pelo apoio incondicional e por constituírem a base inabalável da minha formação pessoal e profissional.

À minha esposa, Lara, e aos meus dois filhos, Catarina e Martim, por compreenderem as minhas ausências e pela paciência inesgotável durante as horas dedicadas a este estudo. Esta conquista é, acima de tudo, nossa.

RESUMO

O emprego de Grandes Modelos de Linguagem (LLMs) em tarefas analíticas complexas apresenta desafios inerentes, como a propensão a alucinações factuais e a perda de informações em contextos extensos (fenômeno *lost-in-the-middle*). Para mitigar essas limitações no escopo de análise e coleta para apoio à inteligência estratégica, este trabalho apresenta um *pipeline* de Inteligência Artificial estruturado sob a metodologia *Design Science Research* (DSR). O artefato desenvolvido operacionaliza o Ciclo de Inteligência da OTAN por meio de uma arquitetura agêntica de seis etapas, utilizando a metodologia 5W3H para decompor consultas do usuário e separar o planejamento tático da execução textual. As principais contribuições da arquitetura incluem um motor de RAG Híbrido Agêntico, que integra bases relacionais, vetoriais e grafos de conhecimento, e a estratégia *Hierarchical Chain-of-Density* (H-CoD), que combina planejamento prévio de tópicos com densificação iterativa. O sistema induz um raciocínio deliberativo (*System 2*) focado em ancorar a predição probabilística estritamente nas evidências recuperadas. A avaliação experimental ocorreu em um ambiente computacional isolado, utilizando modelos de pesos abertos limitados a 70 bilhões de parâmetros. O artefato foi confrontado com modelos comerciais de fronteira (GPT-5.1, Claude Sonnet 4.5 e Gemini 3) na resolução de 18 cenários de inteligência variados. Adotou-se um protocolo de *LLM-as-a-Judge* com múltiplos avaliadores, no qual cada avaliador emite um veredito após discussão liderada por um agente supervisor, com rotação cega de papéis entre rodadas. Os resultados sugerem que o *pipeline* local obteve resultados mais favoráveis (ou comparáveis, a depender da tarefa) em precisão factual, densidade informacional e coerência lógica em relação aos sistemas baseados em nuvem. Conclui-se que o rigor no controle do fluxo de dados pode compensar parte da assimetria paramétrica, viabilizando sistemas de apoio à decisão autônomos, seguros e com relação custo-benefício favorável frente a arranjos híbridos e em nuvem.

Palavras-chave: IA Agêntica; RAG Híbrido; Hierarchical Chain-of-Density (H-CoD); Ciclo de Inteligência; Inteligência Estratégica.

ABSTRACT

The deployment of Large Language Models (LLMs) in complex analytical tasks presents inherent challenges, such as the propensity for factual hallucinations and information loss in extensive contexts (the *lost-in-the-middle* phenomenon). To mitigate these limitations within the scope of analysis and collection in support of strategic intelligence, this work presents an Artificial Intelligence pipeline structured under the *Design Science Research* (DSR) methodology. The developed artifact operationalizes the NATO Intelligence Cycle through a six-stage agentic architecture, utilizing the 5W3H methodology to decompose user queries and decouple tactical planning from textual execution. The main architectural contributions include an Agentic Hybrid RAG engine, which integrates relational, vector, and knowledge graph databases, and the *Hierarchical Chain-of-Density* (H-CoD) strategy, which combines prior topic outlining with iterative densification. The system induces deliberative reasoning (*System 2*) focused on grounding probabilistic prediction strictly in retrieved evidence. Experimental evaluation took place in an isolated computational environment, using open-weight models limited to 70 billion parameters. The artifact was compared against frontier commercial models (GPT-5.1, Claude Sonnet 4.5, and Gemini 3) in resolving 18 varied intelligence scenarios. A multi-evaluator *LLM-as-a-Judge* protocol was adopted, in which each evaluator issues a verdict after a discussion led by a supervisor agent, with blind role rotation across rounds. Results suggest that the local pipeline achieved more favorable (or task-dependent comparable) results in factual accuracy, informational density, and logical coherence relative to cloud-based systems. It is concluded that strict data flow control can offset part of the parametric asymmetry, enabling autonomous, secure, and cost-effective decision support systems relative to hybrid and cloud-based arrangements.

Keywords: Agentic AI; Hybrid RAG; Hierarchical Chain-of-Density (H-CoD); Intelligence Cycle; Strategic Intelligence.

SUMÁRIO

1 INTRODUÇÃO 1
1.1 MOTIVAÇÃO 2
1.2 OBJETIVOS 3
1.2.1 OBJETIVO GERAL 3
1.2.2 OBJETIVOS ESPECÍFICOS 3
1.3 HIPÓTESE E JUSTIFICATIVA 4
1.4 TRABALHOS PUBLICADOS 5
1.5 ESTRUTURA DA DISSERTAÇÃO 6

2 METODOLOGIA, FUNDAMENTAÇÃO TEÓRICA E TRABALHOS RELACIONADOS 7
2.1 ASPECTOS METODOLÓGICOS E CARACTERIZAÇÃO DA PESQUISA 7
2.1.1 CARACTERIZAÇÃO DA PESQUISA 7
2.1.2 *Design Science Research* COMO MÉTODO ADOTADO 8
2.1.3 MAPEAMENTO DAS FASES DSR 8
2.1.4 POSICIONAMENTO FRENTE À LINHA DE PESQUISA 8
2.2 FUNDAMENTAÇÃO TEÓRICA 9
2.2.1 SISTEMAS COMPOSTOS E REPRESENTAÇÃO SEMÂNTICA 9
2.2.2 DECOMPOSIÇÃO DA INTENÇÃO E TOPOLOGIAS DE PLANEJAMENTO 10
2.2.3 LAÇO DE RACIOCÍNIO, AÇÃO E RECUPERAÇÃO DE EVIDÊNCIAS 13
2.2.4 SUPERVISÃO METACOGNITIVA E REFINAMENTO ITERATIVO 14
2.2.5 GERAÇÃO DE TEXTO LONGO E AVALIAÇÃO DA SAÍDA 16
2.3 TRABALHOS RELACIONADOS E ESTADO DA ARTE 17
2.3.1 GERAÇÃO AGÊNTICA DE RELATÓRIOS LONGOS 17
2.3.2 LACUNA IDENTIFICADA E CRITÉRIOS DE AVALIAÇÃO 18

3 ARQUITETURA DO SISTEMA 19
3.1 ETAPAS DO PIPELINE 19
3.1.1 ETAPA 1: OBSERVAÇÃO 19
3.1.2 ETAPA 2: DIREÇÃO (PLANEJAMENTO) 21
3.1.3 ETAPA 3: COLETA 22
3.1.4 ETAPA 4: PROCESSAMENTO 24
3.1.5 ETAPA 5: ANÁLISE 24
3.1.6 ETAPA 6: GERAÇÃO DE RELATÓRIO 29
3.2 MOTORES E TÉCNICAS DE APOIO UTILIZADAS 30
3.2.1 MOTOR DE RAG HÍBRIDO AGÊNTICO 30
3.2.2 REPRESENTAÇÃO SEMÂNTICA E VETORIZAÇÃO 31
3.2.3 SUPERVISOR METACOGNITIVO 32
3.2.4 A ENGENHARIA DE CONTEXTO DINÂMICA 34

3.2.5 Evolução Autônoma por Auto Reflexão Crítica 34
3.2.6 Aprendizado em Contexto (*In Context Learning*) 35

4 Avaliação Experimental 36
4.1 Configuração do Ambiente e Amostragem 36
4.1.1 Infraestrutura e Configuração do Ambiente 36
4.1.2 Conjunto de Tarefas Analíticas 38
4.1.3 Métricas e Dimensões de Avaliação 38
4.2 Delimitações Metodológicas 39
4.3 Otimizações Computacionais e Eficiência 39
4.3.1 Utilização de Modelos Esparsos (MoE) 39
4.3.2 Implementação de *Key-Value Cache* 40
4.3.3 Quantização de Modelos 40
4.4 Processamento de Tarefas *Multi-hop* 41
4.5 Desenho Experimental e Protocolo de Julgamento 42
4.5.1 Engenharia do *Prompt* de Avaliação 43
4.5.2 Relação Consolidada das Tarefas Avaliadas 44
4.6 Análise Arquitetural de Confronto Isolado 47
4.6.1 *Pipeline* UnB *versus* Anthropic (Claude) 47
4.6.2 *Pipeline* UnB *versus* OpenAI (GPT-5.1) 48
4.6.3 *Pipeline* UnB *versus* Google (Gemini) 49
4.6.4 Autonomia e Isolamento Operacional 50
4.7 Resultados Agregados da Avaliação 51
4.8 Análise Qualitativa por Dimensão 52
4.9 Discussão: Arquitetura e Escala 52
4.10 O Gargalo da Validação Humana e o *Time to Intelligence* 53
4.11 Trajetória de Otimização e Viabilidade Operacional 54

5 Conclusões 57
5.1 Implicações Operacionais e Soberania Tecnológica 57
5.2 Limitações da Pesquisa 58
5.3 Trabalhos Futuros 59

Referências Bibliográficas 60

Apêndice A – Relatório de Inteligência Competitiva e Financeira: Análise Estratégica do Banco do Brasil (junho de 2025) 66
A.1 Introdução 66
A.2 Sumário Executivo 66
A.3 Visão Geral da Companhia 67
A.4 Análise do Desempenho Financeiro (Foco 1T25) 67
A.5 Desempenho das Ações e Avaliação (BBAS3) 68
A.6 Cenário Competitivo 69

A.7 Avaliação de Riscos e Mitigação 70
A.7.1 Risco de Crédito 70
A.7.2 Risco de Mercado, Operacional e Regulatório 71
A.7.3 Abordagem da Administração para Mitigação de Riscos 71
A.8 Estratégia e Desempenho em Sustentabilidade (ESG) 72
A.8.1 Compromissos e Estrutura ESG 72
A.8.2 Iniciativas Ambientais 72
A.8.3 Iniciativas Sociais 72
A.8.4 Governança 73
A.8.5 Relatórios Chave 73
A.9 Governança Corporativa 73
A.9.1 Conselho de Administração 74
A.9.2 Diretoria Executiva (Conselho Diretor e Diretores) 74
A.10 Perspectivas Estratégicas, Desafios e Oportunidades 75
A.10.1 Perspectivas de Curto e Médio Prazo 75
A.10.2 Desafios Principais 75
A.10.3 Oportunidades de Crescimento 76
A.10.4 Tendências do Setor Bancário (Cenários PwC) 76
A.11 Recomendações Estratégicas 77
A.12 Referências 78

LISTA DE FIGURAS

2.1 Evolução das topologias de raciocínio: do CoT ao *Graph of Thoughts* (DAG). Adaptado de [1]..... 11
2.2 Esquema conceitual do *Self-Refine*. Fonte: [2]. 15

3.1 Pipeline do sistema, organizado em seis etapas adaptativas (Observação, Direção, Coleta, Processamento, Análise e Geração de Relatório) sob uma camada metacognitiva centralizada (Supervisor). Adaptado de [3]. 20
3.2 Arquitetura do RAG Híbrido Agêntico, com processamento multimodal e laços de *feedback*. A estratégia de armazenamento combina dados relacionais, vetoriais contextualizados e grafo de entidades..... 31

4.1 Protocolo de avaliação, da seleção da tarefa à apuração dos vereditos. As setas tracejadas indicam a rotação de papéis dos modelos comerciais entre rodadas. 43
4.2 Evolução do consumo de memória de vídeo (VRAM) ao longo dos trimestres de desenvolvimento..... 55
4.3 Redução do tempo de processamento necessário para finalizar uma tarefa analítica em cada trimestre..... 55

LISTA DE TABELAS

2.1 Posicionamento dos trabalhos relacionados..... 18

3.1 Decomposição 5W3H aplicada ao escopo operacional do Pipeline de IA proposto - montado pelo agente supervisor (IA)..... 23

4.1 Modelos *open-weights* utilizados no *pipeline* UnB e respectivas características..... 37
4.2 Modelos comerciais de fronteira avaliados como linhas de base no protocolo experimental. . 38
4.3 Relação completa das 18 tarefas executadas no protocolo experimental e respectivos resultados. 46
4.4 Detalhamento dos confrontos e resultados frente ao modelo Claude Sonnet 4.5. 47
4.5 Detalhamento dos confrontos e resultados frente ao modelo GPT-5.1. 48
4.6 Detalhamento dos confrontos e resultados frente ao modelo Gemini 3. 49
4.7 Avaliação preliminar via *LLM-as-a-Judge* multi-juiz sobre 18 tarefas analíticas, com rotação cega de papéis e crítica/contracrítica mediada por supervisor. 51
4.8 Comparativo conceitual (projeção do autor, não medida experimentalmente) de *Time to Intelligence* (TTI) entre abordagens de IA em Inteligência Estratégica. 54

A.1 Destaques financeiros do Banco do Brasil — 1T25..... 68
A.2 Panorama da ação BBAS3..... 69
A.3 Métricas financeiras comparativas — Banco do Brasil *versus* principais pares (projeções/resultados 1T25 ou mais recente disponível). 70
A.4 Comparativo de taxas de juros de empréstimo pessoal (referência 2023)..... 70
A.5 Resumo das principais metas ESG e progresso reportado (base 1T25/2024). 73

LISTA DE ABREVIAÇÕES E SIGLAS

5W3H	What, Why, Who, When, Where, How, How Much, How Many (metodologia de decomposição de perguntas)
ACH	Análise de Hipóteses Concorrentes (Analysis of Competing Hypotheses)
ADAS	Automated Design of Agentic Systems (projeto automatizado de sistemas agênticos)
AGU	Advocacia-Geral da União
AI	Artificial Intelligence (inteligência artificial)
AIME	American Invitational Mathematics Examination (benchmark de matemática)
AJP	Allied Joint Publication (publicação de doutrina conjunta aliada da OTAN, p. ex. AJP-2.1)
API	Application Programming Interface (interface de programação de aplicações)
ASG	Ambiental, Social e Governança (ver também ESG)
AWQ	Activation-aware Weight Quantization (quantização de pesos ciente da ativação)
BFCL	Berkeley Function Calling Leaderboard (benchmark de chamada de funções)
BGE-M3	BAAI General Embedding (modelo de embeddings vetoriais)
BM25	Best Matching 25 (função de ranqueamento para recuperação de informação)
BRICS	Brasil, Rússia, Índia, China e África do Sul
BRS	Brainstorming (tempestade de ideias; código interno do catálogo de técnicas analíticas)
BTG	BTG Pactual (banco de investimento)
CBDC	Central Bank Digital Currency (moeda digital de banco central)
CIA	Central Intelligence Agency (Agência Central de Inteligência dos EUA)
CNN	Convolutional Neural Network (rede neural convolucional)
CoALA	Cognitive Architectures for Language Agents (arquiteturas cognitivas para agentes de linguagem)
CoD	Chain-of-Density (cadeia de densidade)
COP	Conferência das Partes (Convenção-Quadro da ONU sobre Mudança do Clima)
CoT	Chain-of-Thought (cadeia de pensamento)
CTI	Cyber Threat Intelligence (inteligência de ameaças cibernéticas)
DAG	Directed Acyclic Graph (grafo acíclico direcionado)
DATG	Dynamic Attribute Graphs (grafos dinâmicos de atributos)

DOME	Dynamic Hierarchical Outlining with Memory-Enhancement (esboço hierárquico dinâmico com memória)
DPO	Direct Preference Optimization (otimização direta de preferências)
DSR	Design Science Research (metodologia de pesquisa em ciência do projeto)
DSS	Decision Support System (sistema de apoio à decisão)
ENE	Departamento de Engenharia Elétrica (UnB)
ESG	Environmental, Social and Governance (ambiental, social e governança; ver também ASG)
FLOPs	Floating-Point Operations (operações de ponto flutuante)
FP16	representação em ponto flutuante de 16 bits (floating-point 16-bit)
GPQA	Graduate-Level Google-Proof Q&A (benchmark de perguntas e respostas em nível de pós-graduação)
GPT	Generative Pre-trained Transformer
GPTQ	método de quantização pós-treinamento para Transformers generativos pré-treinados
GPU	Graphics Processing Unit (unidade de processamento gráfico)
GQA	Grouped-Query Attention (atenção por consultas agrupadas)
GSM8K	Grade School Math 8K (benchmark de problemas de matemática de nível escolar)
H-CoD	Hierarchical Chain-of-Density (cadeia de densidade hierárquica)
HAJailBench	benchmark de avaliação de segurança de LLMs com instâncias anotadas por humanos (Human-Annotated Jailbreak Benchmark)
HiMATE	Hierarchical Multi-Agent framework for machine Translation Evaluation (framework multiagente hierárquico)
IA	Inteligência Artificial
ICF	Inconsistency Finder (detector de inconsistências; código interno do catálogo de técnicas analíticas)
ICL	In-Context Learning (aprendizado em contexto)
IFEval	Instruction-Following Evaluation (benchmark de seguimento de instruções)
IFR	Índice de Força Relativa (equivalente ao RSI)
INR	Bureau of Intelligence and Research (Departamento de Estado dos EUA)
INT	disciplina de inteligência (intelligence discipline)
INT4 / INT8	formatos numéricos de inteiros de 4 e 8 bits (quantização)
JSON	JavaScript Object Notation
KAC	Key Assumptions Check (verificação de premissas-chave; código interno do catálogo de técnicas analíticas)
KV Cache	Key-Value Cache (cache chave-valor da autoatenção)
LAMBADA	LAnguage Modeling Broadened to Account for Discourse Aspects (benchmark de modelagem de linguagem)
LLM	Large Language Model (grande modelo de linguagem)
LLM-as-a-Judge	protocolo de avaliação em que um LLM atua como juiz das respostas

LSEG	London Stock Exchange Group (consenso LSEG)
MATH	benchmark de matemática de nível de competição
MBPP	Mostly Basic Python Problems (benchmark de programação)
MMLU	Massive Multitask Language Understanding (benchmark de compreensão multitarefa)
MMLU-Pro	versão estendida e mais difícil do benchmark MMLU
MMMU	Massive Multi-discipline Multimodal Understanding and Reasoning (benchmark multimodal)
MoE	Mixture-of-Experts (mistura de especialistas)
MPME	micro, pequenas e médias empresas
MSG	Multiple Scenarios Generation (geração de múltiplos cenários; código interno do catálogo de técnicas analíticas)
NIAH	Needle-in-a-Haystack (benchmark de recuperação de informação em contexto longo)
OCR	Optical Character Recognition (reconhecimento ótico de caracteres)
ODNI	Office of the Director of National Intelligence (Gabinete do Diretor de Inteligência Nacional dos EUA)
ODS	Objetivos de Desenvolvimento Sustentável (ONU)
ONU	Organização das Nações Unidas
OOM	Out of Memory (falha por falta de memória)
OSINT	Open Source Intelligence (inteligência de fontes abertas)
OTAN	Organização do Tratado do Atlântico Norte (NATO – North Atlantic Treaty Organization)
PDF	Portable Document Format
PPEE	Programa de Pós-Graduação Profissional em Engenharia Elétrica (UnB)
PPL	perplexidade (perplexity)
PRI	Princípios para o Investimento Responsável (ONU)
PRISMA-ScR	Preferred Reporting Items for Systematic Reviews and Meta-Analyses — extensão para Scoping Reviews (protocolo de revisão sistemática)
RAG	Retrieval-Augmented Generation (geração aumentada por recuperação)
RAPID	framework de geração de texto longo com planejamento de escrita e descoberta de informação (retrieval-augmented long text generation with writing planning and information discovery)
ReAct	Reasoning and Acting (raciocínio e ação sinérgicos)
RL	Reinforcement Learning (aprendizado por reforço)
RLHF	Reinforcement Learning from Human Feedback (aprendizado por reforço com feedback humano)
RNN	Recurrent Neural Network (rede neural recorrente)
RSI	Relative Strength Index (índice de força relativa, IFR)
RWKV	Receptance Weighted Key Value (arquitetura de rede neural linear)
S.A.	sociedade anônima
SATs	Structured Analytic Techniques (técnicas analíticas estruturadas)

Selic	Sistema Especial de Liquidação e de Custódia
SLFF	Sustainability-Linked Financing Framework (framework de dívidas vinculadas à sustentabilidade)
SLM	Small Language Model (pequeno modelo de linguagem)
SOPs	Standard Operating Procedures (procedimentos operacionais padrão)
SSC	Structured Self-Critique (autocrítica estruturada; código interno do catálogo de técnicas analíticas)
SWOT	Strengths, Weaknesses, Opportunities and Threats (forças, fraquezas, oportunidades e ameaças)
T1–T4	trimestres de desenvolvimento do projeto (T1 a T4)
TGI	Text Generation Inference (runtime de inferência de modelos)
TII	Technology Innovation Institute
TLN	Timelines (linhas do tempo; código interno do catálogo de técnicas analíticas)
ToT	Tree-of-Thoughts (árvore de pensamentos)
TrOCR	Transformer-based OCR (OCR baseado em Transformer)
TST	Tribunal Superior do Trabalho
TTFT	Time to First Token (tempo até o primeiro token)
TTI	Time to Intelligence (tempo até a inteligência; ciclo completo de produção analítica)
VRAM	Video Random Access Memory (memória de vídeo)
WIA	What-If Analysis (análise "e se"; código interno do catálogo de técnicas analíticas)

1 INTRODUÇÃO

Este trabalho apresenta um *pipeline* de IA voltado à análise de tópicos complexos, ancorado na metodologia 5W3H e organizado como um sistema multiestágio de apoio à decisão. A proposta integra técnicas de engenharia de *prompt*, meta-agentes e RAG Híbrido (*Retrieval-Augmented Generation*). O objetivo é produzir relatórios analíticos sobre temas complexos, reduzindo o tempo até a produção de inteligência e apoiando processos decisórios.

A ideia central do *pipeline* é decompor, de forma contextual, a pergunta de entrada. Para isso, a metodologia 5W3H (*What, Why, Who, When, Where, How, How much, How many*) orienta o ciclo de inteligência, ampliando contexto e ajudando a ajustar a saída ao problema. Essa decomposição, somada ao mapeamento de intenção e a mecanismos de autogestão, alimenta um motor de RAG Híbrido Agêntico em duas etapas. Na primeira, dados são coletados de fontes abertas, bases estruturadas e documentos internos, conforme os requisitos do usuário, e persistidos em um armazenamento unificado que combina modalidades relacional, vetorial e em grafo. Na segunda etapa, lacunas contextuais são identificadas para orientar a geração, acionando novas iterações de busca e consulta ao RAG até compor o conjunto de dados necessário à resposta. Cada etapa subsequente opera sobre um objeto de estado compartilhado, permitindo retroalimentação iterativa e verificações de contexto e consistência ao longo do *pipeline*, com o intuito de reduzir alucinações.

Para lidar com gargalos de desempenho e limitações dos LLMs de base, o sistema orquestra métodos voltados a aproximar um raciocínio deliberativo do tipo *System 2* [4]. A consistência é reforçada por protocolos de raciocínio recursivo com *In-Context Learning*, *Self-Consistency*, *Self-Refinement* e planejamento por *Chain of Thought* (CoT), sob supervisão meta-agêntica. Em termos arquiteturais, o *pipeline* incorpora *frameworks* recentes da literatura para coerência em contexto longo, como o esboço hierárquico com memória do DOME [5], o fluxo orientado à descoberta do RAPID [6] e as análises sobre degradação em contexto longo [7]. Por fim, a qualidade do texto é refinada por meio do *Hierarchical Chain-of-Density* (H-CoD) e de passos automatizados de reescrita, calibrados com restrições objetivas como *Flesch Reading Ease*, perplexidade e similaridade vetorial.

O *pipeline* é orquestrado por um LLM supervisor que gerencia mecanismos de *In-Context Learning* por meio de *prompts* metaestruturados ("*metaprompts*"), garantindo portabilidade entre modelos *open-source* e permitindo selecionar dinamicamente o modelo mais adequado às necessidades do usuário.

Mesmo quando executado em GPUs de baixo custo e com LLMs de menor porte, o sistema apresenta desempenho comparável ao de soluções como GPT-5.1, Claude Sonnet 4.5 e Gemini 3 em tarefas de geração e síntese. Os textos produzidos a partir de contextos extensos tendem a apresentar maior coesão e melhor sustentação lógica, além de indicadores mais altos de precisão factual, profundidade analítica e fluidez textual, tanto em *benchmarks* (como o LAMBADA) quanto em avaliação automatizada (*LLM-as-a-Judge*).

Os resultados indicam que a abordagem proposta melhora a qualidade de relatórios produzidos por LLMs de menor porte. Isso contribui para ampliar o acesso à análise de temas complexos, com prioridade

para sistemas menores e para requisitos de segurança da informação, em cenários de recursos limitados e demandas para resposta a temas complexos.

A contribuição principal deste trabalho é demonstrar, na prática, a viabilidade de construir sistemas de IA capazes de entregar resultados específicos e competitivos com tecnologias líderes de mercado, utilizando infraestrutura de baixo custo, com menor investimento e manutenção. Também são preservados requisitos operacionais como soberania de dados e previsibilidade de custo. Como consequência, o desenho tende a reduzir desperdício computacional e impactos energéticos associados à operação contínua.

1.1 MOTIVAÇÃO

O problema que motiva este trabalho pode ser formulado de forma direta: como produzir inteligência analítica de qualidade, em tempo hábil, em um cenário no qual custo computacional, soberania sobre os dados, acesso à tecnologia e equivalência com o trabalho humano são, com frequência, tratados como exigências mutuamente excludentes. Essas dimensões são analisadas neste estudo, com o objetivo de propor um modelo arquitetural (ou *framework* metodológico) capaz de equilibrá-las de forma viável, segura e sustentável.

É fato que o custo computacional de grandes modelos cresce continuamente, e isso se traduz em custos recorrentes de operação (infraestrutura, latência e despesas variáveis por uso) que podem tornar inviável o emprego sustentado de modelos de fronteira em certos cenários. Em paralelo, há uma dimensão energética: estimativas para treinamento em larga escala apontam emissões relevantes de CO_2 equivalente [8], e análises mais recentes indicam que a inferência, antes considerada marginal, passou a representar parcela importante da pegada de carbono em produtos de uso massivo [9, 10]. Ferramentas como o SPROUT [11] sugerem que intervenções relativamente pequenas (diretivas de geração, escolha de modelo e roteamento condicional) podem reduzir consumo e emissões sem perda de qualidade. Nesse sentido, arquiteturas que fazem mais com menos endereçam primeiro a viabilidade operacional e, adicionalmente, dialogam com a preocupação ambiental.

Além do custo, há a soberania sobre os dados. O ciclo de inteligência moderno opera cada vez mais sobre fontes abertas. Nesse contexto, cabe destacar a estratégia de OSINT 2024–2026 da comunidade de inteligência dos Estados Unidos [12], que colocou o tema como prioridade institucional, e a estratégia do Departamento de Estado norte-americano publicada em 2024 [13]. Em paralelo, levantamentos sobre privacidade em LLMs documentam que cada chamada a um modelo hospedado por terceiros pode, em alguns cenários, expor informações por inversão, extração ou inferência de presença [14]. Para países e organizações que produzem conhecimento sensível, a dependência de modelos sob jurisdição estrangeira é um risco que se soma ao custo ambiental. Estratégias nacionais recentes em inteligência artificial têm articulado, em paralelo, modelos soberanos, comerciais e de código aberto.

Há ainda a questão do acesso. Nem toda instituição capaz de produzir análise relevante consegue pagar a infraestrutura necessária para usar modelos de fronteira em escala, e o quadro tende a piorar conforme os modelos crescem. Pequenos modelos de linguagem (SLMs), caracterizados por Belcak et al. [15] como modelos abaixo de aproximadamente dez bilhões de parâmetros, mudaram o cenário. O Phi-3.5-mini [16],

com 3,8 bilhões de parâmetros, rivaliza com gerações anteriores de modelos de fronteira em tarefas de raciocínio. A família DeepSeek-R1 [17] levou esse ponto adiante ao destilar capacidade de raciocínio de um modelo grande para variantes menores, executáveis em GPUs de consumo. Técnicas de quantização (AWQ [18], GPTQ [19]) ampliam esse acesso, viabilizando inferência local em *hardware* que, há poucos anos, seria insuficiente para o porte dos modelos utilizados.

No cenário proposto, surge a pergunta: o que esses sistemas produzem se compara, em qualidade analítica, ao que um analista experiente produziria? Avaliar essa proximidade com rigor é uma frente constante de pesquisa. O ExpertLongBench [20] mostra que mesmo modelos comerciais de fronteira, quando avaliados em tarefas longas de domínio especialista com listas estruturadas de verificação, ainda apresentam desempenho modesto em comparação ao padrão humano. Esse dado tem duas leituras: evidencia que a equivalência total ainda não foi atingida e sugere que, se um sistema bem orquestrado de modelos menores se aproximar desses desempenhos, a justificativa para usar um único modelo muito grande enfraquece em parte dos cenários práticos.

Por fim, houve uma mudança de foco na literatura técnica que torna o argumento desta dissertação mais plausível. Em fevereiro de 2024, Zaharia et al. [21] formalizaram o conceito de *Sistema de IA Composta*: parte do ganho de qualidade em produtos recentes de IA decorre da composição entre componentes especializados, e não apenas do aumento de escala de modelos isolados. Belcak et al. [15] complementam esse argumento ao defender que, para invocações repetitivas em sistemas agênticos, SLMs tendem a ser suficientemente capazes e mais econômicos. A mudança de foco (do modelo para o sistema e do gigante para o especializado) define a abordagem central apresentada nesta dissertação. O *pipeline* proposto materializa essa combinação: um Sistema de IA Composta, com predominância de SLMs, executado em *hardware* de consumo, com o objetivo de melhorar a qualidade dos processos de produção de inteligência.

1.2 OBJETIVOS

1.2.1 Objetivo Geral

Desenvolver um Sistema de IA Composta capaz de produzir relatórios analíticos sobre temas complexos a partir de fontes abertas, operável em *hardware* doméstico (de pequeno porte), com precisão factual, densidade informacional e coerência lógica competitivas com as obtidas por modelos comerciais de fronteira, gerando impacto no tempo até a produção de inteligência, na qualidade dos processos de apoio à decisão, na portabilidade e na sustentabilidade computacional da operação.

1.2.2 Objetivos Específicos

Os objetivos específicos que sustentam o objetivo geral são:

1. Estruturar um *pipeline* agêntico capaz de produzir um relatório analítico de alto nível em tempo hábil.
2. Implementar um motor de *Retrieval-Augmented Generation* Híbrido Agêntico capaz de combinar

bases relacionais, vetoriais e em grafo (GraphRAG), com *cache* semântico, segmentação adaptativa e retroalimentação orientada por lacunas informacionais, de modo a sustentar um estado de conhecimento unificado e auditável ao longo de um *pipeline* de múltiplos estágios.

3. Propor o *Hierarchical Chain-of-Density* (H-CoD) como estratégia de síntese hierárquica, capaz de gerar relatórios de extensão variável (entre cinco e trinta páginas, conforme a densidade de evidências) sem incorrer no fenômeno *lost-in-the-middle* ou na diluição informacional típica da geração de texto longo por LLMs.

4. Avaliar o *pipeline* em condições controladas contra modelos comerciais de fronteira por meio de um protocolo de *LLM-as-a-Judge* com múltiplos juízes, com rotação cega de papéis entre rodadas e com crítica/contracrítica mediada por um agente supervisor.

1.3 HIPÓTESE E JUSTIFICATIVA

A combinação de decomposição contextual via 5W3H, RAG Híbrido Agêntico e densificação hierárquica (H-CoD), orquestrada sobre um conjunto de pequenos modelos de linguagem, alinhados a técnicas modernas de construção de contexto, rodando em hardware convencional (de uso doméstico), pode impactar significativamente a qualidade dos relatórios analíticos produzidos, elevando a precisão factual, a densidade informacional e a coerência de raciocínio a patamar competitivo com o obtido por modelos comerciais de fronteira (a exemplo de GPT-5.1, Gemini 3 e Claude Sonnet 4.5) operados em infraestrutura computacional de grande porte.

A plausibilidade dessa hipótese repousa sobre uma confluência de avanços recentes na literatura que vale percorrer antes de testá-la empiricamente. O ponto de partida está nos próprios pequenos modelos. Resultados como os do Phi-3.5-mini [16], com 3,8 bilhões de parâmetros e desempenho comparável ao GPT-3.5 em tarefas de raciocínio, mostraram que a capacidade de raciocínio, antes considerada propriedade emergente exclusiva de modelos com mais de cem bilhões de parâmetros, passou a estar acessível em escalas muito menores. A família DeepSeek-R1 [17] levou o ponto adiante ao destilar capacidade de raciocínio de um modelo grande para variantes pequenas, executáveis em GPUs de uso doméstico. Trabalhos como o rStar-Math [22] mostraram que técnicas algorítmicas como a busca em árvore por Monte Carlo [23] têm a capacidade de elevar SLMs a níveis de raciocínio matemático antes restritos a modelos muito maiores.

A discussão se desloca, com isso, do tamanho do modelo para a arquitetura do sistema e engenharia de contexto. Zaharia et al. [21] formalizaram esse deslocamento ao definir o paradigma dos Sistemas de IA Composta. Pouco depois, Chen et al. [24] forneceram fundamento matemático para a tese, estabelecendo propriedades de escala desses sistemas e mostrando que múltiplas chamadas a modelos menores podem, sob condições amplas, superar uma chamada única a um modelo maior. Frameworks como o DSPy [25] operacionalizam esse insight em pipelines declarativos. Belcak et al. [15] fecham o argumento ao defender, em *position paper*, que para invocações repetitivas em sistemas agênticos os SLMs são suficientemente capazes, mais adequados e economicamente preferíveis.

Há ainda um mecanismo que compensa, no nível operacional, o porte reduzido dos modelos: o *In-*

Context Learning (ICL), discutido em detalhe na Seção 2.2.1.2 do Capítulo 2. Por meio da inserção dinâmica de exemplos e instruções estruturadas no contexto de inferência, o ICL permite que um SLM atue, *ad hoc*, como um especialista do domínio sem qualquer atualização de parâmetros. Hendel et al. [26] sugerem que o ICL opera ativando vetores de tarefa latentes no espaço interno do modelo, com efeito próximo ao de um *fine-tuning* efêmero em tempo de inferência. Em um pipeline como este, em que diferentes etapas exigem diferentes "personalidades" do mesmo modelo subjacente, o ICL é o que viabiliza essa especialização funcional sem multiplicação do custo de inferência. Sem ele, o argumento de SLMs em hardware modesto perderia uma das suas sustentações centrais.

Resta o problema da avaliação. Validar empiricamente a hipótese passa por um protocolo capaz de substituir, ao menos parcialmente, a presença de especialistas humanos, cada vez mais escassos, caros e difíceis de mobilizar em escala. A literatura tem convergido para protocolos multiagentes como alternativa: Du et al. [27] e Liang et al. [28] mostraram que a agregação estruturada de julgamentos de múltiplos modelos pode ampliar a correção dos veredictos em relação à votação majoritária simples. ChatEval [29] foi um dos primeiros *frameworks* a operacionalizar essa ideia, e o JudgeBench [30] catalogou o estado da arte em juízes baseados em LLM. Neste trabalho, essa classe de abordagem é instanciada como um protocolo de *LLM-as-a-Judge* com múltiplos juízes e rotação cega de papéis, no qual cada juiz conduz um ciclo de crítica/contracrítica de forma bilateral e mediada por um agente supervisor antes de emitir o veredito. O trabalho HAJailBench [31] mostra que avaliações confiáveis podem ser obtidas com modelos menores quando o processo de julgamento é cuidadosamente estruturado. Por equivalência conceitual, esse achado sustenta a possibilidade de produção de conteúdo competitivo quando o processo de produção é igualmente bem desenhado.

A confluência desses avanços torna razoável esperar que o sistema proposto, ainda que baseado em modelos muito menores que os hoje considerados estado da arte, produza saídas analiticamente competitivas. Os resultados preliminares apresentados no Capítulo 4 sustentam, em escala reduzida, essa expectativa.

1.4 TRABALHOS PUBLICADOS

A proposta que deu origem a esta dissertação foi apresentada à Universidade de Brasília em 2024, na forma de um "crawler autônomo adaptativo"voltado à coleta e processamento de dados de fontes abertas (OSINT), e desde então evoluiu para o pipeline aqui descrito. Nesse percurso, foram produzidos dois trabalhos, apresentados na ordem em que se inserem no projeto.

O primeiro, ainda na fase de revisão de literatura, contempla uma revisão sistemática sobre o emprego de inteligência artificial em sistemas militares de apoio à decisão (DSS). O estudo seguiu o protocolo PRISMA-ScR e comparou dezesseis programas-referência em quatro blocos estratégicos: Estados Unidos, China, Rússia e OTAN/União Europeia. Partindo das classificações tradicionais de cinco fatores, propôs um quadro de avaliação com múltiplas dimensões. A principal conclusão foi que a agilidade para integrar novos modelos a processos operacionais já validados é o maior vetor de vantagem na adoção de IA, com destaque para a arquitetura de copiloto cognitivo, que entrega recomendações contextualizadas e com incerteza calibrada, preservando a supervisão humana. Esse levantamento mapeou o estado da arte e a lacuna

que o *pipeline* desta dissertação busca preencher. O artigo foi aceito para publicação no MICRADS 2026 (9th Multidisciplinary International Conference of Research Applied to Defense and Security).

O segundo consolida a proposta central da dissertação, apresentando o *pipeline* em sua forma integral e detalhando a arquitetura de seis estágios, baseada no Ciclo de Inteligência da OTAN e na metodologia 5W3H. O artigo também formaliza a introdução do *Hierarchical Chain-of-Density* (H-CoD) e demonstra como a execução do *pipeline* em GPUs de pequeno porte superou *baselines* comerciais em precisão factual, densidade de entidades e coerência de raciocínio. O manuscrito encontra-se no prelo em *Engineering Proceedings* (MDPI).

- CRUZ, C. C. S. *et al*. O Copiloto Cognitivo: Inteligência Artificial Aplicada a Sistemas Militares de Apoio à Decisão. *In*: Multidisciplinary International Conference Of Research Applied To Defense and SEcurity (MICRADS), 9., 2026, Guayaquil, Equador. Publicação em edição especial da RISTI: Revista Ibérica de Sistemas e Tecnologias de Informação.
- CRUZ, C. C. S. *et al*. AI Pipeline for Deep Analysis of Complex Topics Using the 5W3H Methodology Applied to an Adaptive Multistage Decision Support System. *In*: *Engineering Proceedings* (MDPI), no prelo, 2026.

1.5 ESTRUTURA DA DISSERTAÇÃO

A organização desta dissertação acompanha o raciocínio de construção do *pipeline* e mantém o foco no vínculo entre escolhas de projeto e validação empírica.

O Capítulo 2 reúne os aspectos metodológicos da pesquisa, a fundamentação teórica e os trabalhos relacionados, articulando o método adotado e as técnicas que sustentam cada decisão de projeto ao longo do sistema. O Capítulo 3 descreve a arquitetura do artefato e o funcionamento do *pipeline* em seis etapas, incluindo os motores de apoio e a governança exercida pelo Supervisor Metacognitivo.

O Capítulo 4 apresenta a validação empírica do *pipeline*, detalhando o desenho experimental, o *hardware* utilizado, os modelos tomados como linha de base (GPT-5.1, Gemini 3 e Claude Sonnet 4.5) e o protocolo de avaliação, que pontua as saídas segundo precisão factual, densidade informacional e coerência de raciocínio. Por fim, o Capítulo 5 sintetiza as contribuições, reconhece as limitações observadas e aponta caminhos para trabalhos futuros.

2 METODOLOGIA, FUNDAMENTAÇÃO TEÓRICA E TRABALHOS RELACIONADOS

Este capítulo reúne a metodologia adotada na pesquisa, a fundamentação teórica que sustenta as decisões de projeto do *pipeline* e os trabalhos relacionados. A organização do texto segue o percurso do trabalho: primeiro, descreve-se como o artefato foi concebido e desenvolvido; em seguida, apresentam-se as bases técnicas que viabilizam cada etapa da arquitetura (Capítulo 3); por fim, discute-se a literatura recente e a lacuna abordada. Esse encadeamento também fixa o vínculo entre hipótese, desenho do sistema e critérios de avaliação retomados no Capítulo 4.

2.1 ASPECTOS METODOLÓGICOS E CARACTERIZAÇÃO DA PESQUISA

Esta seção descreve o método de pesquisa e a caracterização adotada no PPEE, estabelecendo como o problema foi enquadrado e como o artefato foi desenvolvido e avaliado. O foco aqui é o percurso metodológico, e não o funcionamento interno do sistema, que é detalhado no Capítulo 3. Ao explicitar essas escolhas, busca-se fixar o vínculo entre a hipótese, o processo de projeto e os critérios de verificação empírica empregados ao longo da dissertação.

2.1.1 Caracterização da Pesquisa

Com base na tipologia de Cervo, Bervian e Silva [32], adotada como referência metodológica no PPEE, esta pesquisa pode ser caracterizada como aplicada quanto à natureza, exploratório-descritiva quanto aos objetivos e mista quanto à abordagem. Ela é aplicada porque busca produzir um artefato funcional para um problema concreto: gerar inteligência analítica sobre temas complexos, em tempo hábil, com infraestrutura computacional acessível. É exploratória porque investiga um espaço de soluções ao combinar pequenos modelos de linguagem (SLMs), Sistemas de IA Composta e a metodologia 5W3H aplicada ao ciclo de inteligência. É descritiva porque apresenta, ao longo do trabalho, a caracterização de cada componente do *pipeline*, suas dependências e pontos de falha.

A abordagem é mista. A fase qualitativa orienta o projeto e o desenvolvimento do artefato, com decisões de engenharia derivadas do estado da arte e da inspeção do comportamento do sistema. A fase quantitativa sustenta a avaliação experimental, na qual a hipótese central é testada com parâmetros objetivos, descritos no capítulo de avaliação. Adota-se um estudo de caso instrumental: uma situação concreta de análise é usada como meio para validar o método proposto.

2.1.2 *Design Science Research* como Método Adotado

A pesquisa segue o *framework* de *Design Science Research* (DSR), proposto por Peffers et al. [33] para sistemas de informação e adaptado ao contexto brasileiro por Pimentel, Filippo e Santoro [34]. A DSR é adequada a pesquisas baseadas no desenvolvimento de artefatos computacionais. Conforme a literatura, ela se aplica a trabalhos em que o entregável central é um sistema funcional, em contraste com metodologias que tratam o objeto como pré-existente (como o estudo de caso clássico) ou que dependem de intervenções no ambiente (como a pesquisa-ação).

A DSR organiza a pesquisa em seis fases: identificação do problema, definição dos objetivos da solução, projeto e desenvolvimento do artefato, demonstração, avaliação e comunicação. Cada fase possui critérios de saída. A identificação do problema exige fundamentação em literatura recente; a avaliação requer métricas bem definidas e comparação com *baselines*. Esse encadeamento permite auditar não apenas o resultado, mas também o processo. Essa rastreabilidade é relevante quando o artefato é um sistema de inteligência analítica, em que decisões de projeto afetam a fidelidade do produto final e a qualidade das análises geradas.

A literatura brasileira sobre DSR aplicada a artefatos computacionais [34] destaca dois pontos que orientam esta pesquisa: (i) a centralidade dos elementos teóricos que motivam o projeto e (ii) a necessidade de avaliar o sistema e seus efeitos (o sistema satisfaz os requisitos e o problema foi resolvido?). Ambas as perspectivas são tratadas nesta dissertação, na qual a coerência funcional do *pipeline* é verificada componente a componente e a hipótese é submetida à avaliação empírica.

2.1.3 Mapeamento das Fases DSR

A correspondência entre as fases da DSR e a estrutura desta dissertação é direta e permite que cada capítulo seja lido também como evidência metodológica da fase correspondente. A identificação do problema é desenvolvida na introdução, na qual são caracterizadas as pressões que estabelecem a relevância do problema: custo computacional, soberania dos dados, acesso à tecnologia e equivalência com o trabalho humano. A definição dos objetivos da solução é apresentada juntamente com a formulação da hipótese, que fixa o critério mensurável de sucesso. O projeto e desenvolvimento do artefato são detalhados no capítulo de arquitetura, no qual o fluxo de seis estágios é descrito componente a componente. A fase de demonstração, juntamente com a avaliação rigorosa contra as métricas estipuladas, consolida-se na avaliação experimental. A fase de comunicação compreende a conclusão e as publicações associadas.

2.1.4 Posicionamento Frente à Linha de Pesquisa

A produção recente do PPEE na área de concentração em Segurança da Informação tem consolidado uma linha de pesquisa em torno da inteligência artificial aplicada a contextos cibernéticos, inteligência de ameaças, da exploração estruturada de fontes abertas e do uso de *frameworks* analíticos para a produção de conhecimento. Dentro desse contexto mais amplo, métodos baseados em questionamentos sistemáticos como o 5W3H têm sido explorados como mecanismos para estruturar requisitos de inteligência e enriquecer indicadores operacionais, conforme exemplificado em trabalhos correlatos do programa [35].

O presente trabalho permeia diversas disciplinas do programa de pesquisa do PPEE, utilizando a abordagem 5W3H no direcionamento e preparação do *pipeline* de análise, que se estabelece como principal elemento de expansão de contexto e delimitação de escopo. O artefato resultante é um sistema de IA composto que articula RAG Híbrido, raciocínio multiestágio e síntese hierárquica, atendendo à demanda mais geral por inteligência analítica de propósito amplo. O domínio de aplicação se expande da inteligência de ameaças cibernéticas para problemas analíticos de propósito geral, mantendo o eixo metodológico comum: a organização disciplinada do raciocínio analítico em torno de questões complexas e domínio multidisciplinar.

2.2 FUNDAMENTAÇÃO TEÓRICA

A seção reúne os conceitos e técnicas que sustentam as escolhas de projeto do *pipeline*. A apresentação é organizada por blocos temáticos, agrupando-se as técnicas conforme a etapa do *pipeline* que cada uma endereça, e conectando essa discussão ao desenho arquitetural descrito no Capítulo 3. Com isso, a progressão acompanha a arquitetura de seis estágios e fixa o papel funcional de cada técnica no comportamento do sistema.

2.2.1 Sistemas Compostos e Representação Semântica

2.2.1.1 Sistemas de IA Composta e Pequenos Modelos

Antes de apresentar as técnicas específicas que sustentam o *pipeline*, é necessário explicitar o paradigma teórico que organiza o conjunto. Em fevereiro de 2024, pesquisadores de Berkeley, Stanford, MIT e Databricks formalizaram o conceito de Sistema de IA Composta (*Compound AI System*, CAIS) [21]. A ideia central é que parte relevante do ganho de qualidade observado em produtos recentes de IA decorre menos do aumento de escala de modelos isolados e mais da composição entre componentes (modelos, recuperadores, ferramentas externas e validadores), orquestrados por uma camada de controle. O foco passa do modelo ao sistema.

Esse deslocamento conceitual tem implicações práticas para este trabalho. Ele sustenta a decisão de priorizar arquitetura, e não apenas parâmetros: se a qualidade emerge da composição, o esforço de engenharia se concentra no desenho do *pipeline*, na engenharia de contexto e nas técnicas empregadas. Também alinha a contribuição desta dissertação a uma agenda recente que cataloga variações arquiteturais sob esse enfoque [36]. Por fim, justifica a substituição seletiva de chamadas a modelos de fronteira por chamadas a modelos menores, desde que a composição se mantenha funcional.

Os pequenos modelos de linguagem, ou SLMs, foram caracterizados por Belcak et al. [15] como modelos cujo tamanho permite implantação em dispositivos de consumo, com latência baixa o suficiente para uso interativo. Em *position paper* de junho de 2025, esses autores argumentam que SLMs tendem a ser suficientemente capazes e mais econômicos para muitas invocações em sistemas agênticos. A combinação dos dois paradigmas (Sistemas de IA Composta com predominância de SLMs) define o enquadramento no qual as técnicas discutidas nas seções seguintes são aplicadas neste trabalho.

As seções seguintes aprofundam aspectos técnicos desse paradigma: o *In-Context Learning* como mecanismo de especialização sem treinamento adicional; a recuperação aumentada via grafo e vetor; as topologias de raciocínio em árvore e em grafo; o refinamento iterativo e a autoconsistência; e a densificação hierárquica como estratégia de síntese para contextos extensos.

2.2.1.2 Semântica Contextual e *In-Context Learning*

A eficácia dos Grandes Modelos de Linguagem depende da capacidade de sustentar coerência semântica em contextos extensos. Avanços em contexto longo ajudam a reduzir perdas de informação e ampliam a capacidade de síntese [37]. Nesse quadro, o *In-Context Learning* (ICL) permite ao modelo adaptar a resposta a partir de demonstrações inseridas no *prompt*, sem atualização de parâmetros. Em termos teóricos, o ICL pode ser entendido como a ativação, em tempo de inferência, de representações internas associadas a uma tarefa, produzindo um efeito semelhante ao de *fine-tuning* [26].

2.2.1.3 Semântica em Espaço Vetorial

Para quantificar a relação semântica entre as perguntas do usuário, os documentos recuperados e as saídas geradas, o sistema utiliza similaridade vetorial em representações vetoriais de alta dimensão. Em vez de usar correspondência lexical, *embeddings* vetoriais codificam a intenção semântica de trechos de texto. Métricas como a similaridade do cosseno são usadas não apenas na recuperação densa (RAG), mas também como mecanismo de controle de qualidade nas fases de refinamento do *pipeline*. Ao medir a distância angular entre as evidências e o texto gerado, o sistema pode detectar desvios semânticos ou alucinações, mantendo o relatório alinhado à base de fatos [38].

A proximidade vetorial, porém, mede semelhança de significado, não veracidade: dois trechos podem ocupar regiões vizinhas do espaço vetorial e ainda assim divergir nos fatos. Por isso, no *pipeline*, a similaridade do cosseno é usada como um filtro inicial, cujo resultado é cruzado com o grafo de entidades e com a base de origem antes de um trecho ser admitido como evidência. Assim, aproveita-se a varredura rápida por similaridade, sem ignorar suas limitações.

Esses mecanismos também podem ser combinados com reordenação baseada em avaliações posteriores por LLMs (*reranking*), o que tende a melhorar a qualidade dos resultados.

2.2.2 Decomposição da Intenção e Topologias de Planejamento

2.2.2.1 Estrutura 5W3H

Como *framework* interpretativo para identificar e desdobrar a intenção do usuário, o 5W3H operacionaliza pedidos em linguagem natural, transformando-os em especificações estruturadas que um agente pode usar como base de planejamento (*what, who, where, when, why, how, how much, how many, how long*). Com isso, criam-se entradas explícitas para objetivos intermediários. Essa abordagem por perguntas se alinha aos padrões atuais de desenho agêntico, abrindo espaço para melhorias em planejamento, mapeamento de intenção, decomposição de tarefas e expansão de contexto. Na prática, tratar o 5W3H como

framework de planejamento do *pipeline* regulariza a captura de intenção, restringe o uso de ferramentas e calibra explicações, produzindo respostas específicas e rastreáveis até o raciocínio e as evidências subjacentes [39, 40]. Essa rastreabilidade se alinha à discussão de IA explicável, na medida em que permite justificar a saída com base em evidências e passos intermediários, em vez de apenas expor um veredito final [41].

A escolha do 5W3H também atende a um requisito operacional: ao fixar campos para agente, tempo, lugar, motivo e meios, ela espelha as perguntas que já orientam o ciclo de inteligência, o que facilita encaixar o *pipeline* em fluxos de trabalho conhecidos por analistas. Esse formato fechado contribui para previsibilidade e rastreabilidade. Cada subpergunta torna-se um nó de tarefa nas etapas seguintes, permitindo que afirmações do relatório final sejam rastreadas até o item da matriz que as originou. Trata-se de uma estrutura simples o suficiente para não onerar a operação e rígida o bastante para reduzir a perda de escopo ao longo do processo e reduzir alucinações.

2.2.2.2 *Chain of Thought* e *Tree of Thoughts*

O paradigma do *Chain of Thought* (CoT) alterou a inferência em LLMs ao induzir passos intermediários de raciocínio, separando a geração de lógica da predição imediata de *tokens*. Enquanto o CoT segue uma trajetória linear, o *Tree of Thoughts* (ToT) generaliza essa abordagem ao introduzir um mecanismo de busca heurística (por exemplo, em largura ou em profundidade) sobre unidades coerentes de texto. Isso permite ao modelo executar *lookahead* deliberado e *backtracking*, explorando múltiplos ramos de raciocínio antes de comprometer-se com uma resposta. Análises recentes apontam que, embora o CoT seja suficiente para recuperação semântica, o ToT é especialmente relevante em tarefas que dependem de planejamento e exploração de múltiplos caminhos [42]. A Figura 2.1 sintetiza as diferenças estruturais entre essas abordagens.

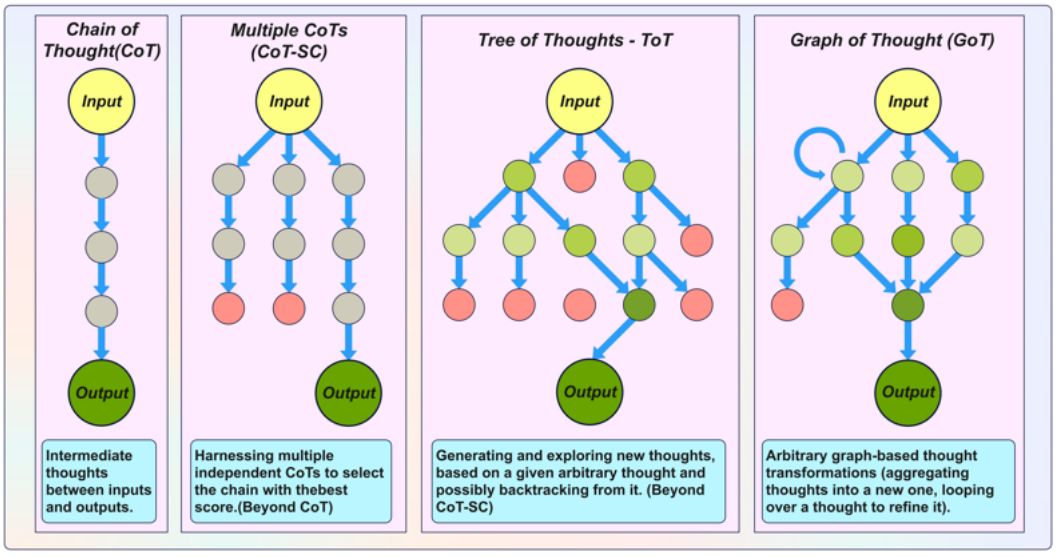

Figura 2.1: Evolução das topologias de raciocínio: do CoT ao *Graph of Thoughts* (DAG). Adaptado de [1].

O ganho do ToT, contudo, tem custo: cada ramo aberto multiplica chamadas de inferência, e abrir ramos em excesso pode transformar deliberação em desperdício computacional. Por isso, o *pipeline* não

adota a busca em árvore como modo padrão. O ToT é reservado à fase de planejamento, na qual decisões inadequadas tendem a contaminar etapas posteriores, enquanto o raciocínio linear, mais barato, é mantido nos trechos em que uma única trajetória é suficiente. Em termos operacionais, trata-se de alocação de esforço computacional: concentra-se busca onde o custo do erro é mais alto e economiza-se onde o custo é menor.

2.2.2.3 Grafos Acíclicos Direcionados (DAG)

Do ponto de vista matemático, a evolução do CoT ao ToT representa a passagem de listas encadeadas lineares para grafos acíclicos direcionados (DAG). Nessa topologia, pensamentos são nós e dependências lógicas são arestas. A estrutura em DAG permite agregar caminhos de raciocínio distintos em uma solução unificada, otimizando a relação entre amplitude de exploração e profundidade de inferência. O *framework Graph of Thoughts* demonstrou que modelar o raciocínio como um DAG supera, com folga, estruturas mais simples em árvore em tarefas de ordenação e análise de redes complexas, justamente por permitir o fluxo de informação entre ramos não adjacentes [1].

Para o tipo de pergunta que este trabalho aborda, essa diferença é importante. Análises de inteligência raramente seguem uma linha clara: uma conclusão sobre intenções depende de outra sobre capacidades, que por sua vez se apoia em evidências coletadas por caminhos independentes. Modelar o raciocínio como DAG permite reunir essas conclusões parciais vindas de ramos distintos e cruzá-las entre si, expondo contradições internas antes que cheguem ao relatório. Uma topologia em árvore, presa a uma hierarquia estrita, perderia justamente as conexões laterais entre evidências que não se tocam de forma direta.

2.2.2.4 Raciocínio Deliberativo (*System 2*)

O conceito de *System 2*, fundamentado na obra de Kahneman [4], designa processos cognitivos associados a esforço analítico, aplicação de lógica, deliberação e planejamento consciente. No contexto de arquiteturas de inteligência artificial voltadas a sistemas de missão crítica, a emulação desse modo de raciocínio funciona como um requisito operacional, no sentido de que é ele que sustenta a resolução de problemas complexos quando há necessidade de validação rigorosa e controle explícito do encadeamento lógico.

A adoção de mecanismos de deliberação marca a transição de respostas predominantemente reativas para procedimentos estruturados. Essa capacidade permite decompor cenários complexos em etapas lógicas, simular rotas alternativas de ação e antecipar consequências antes de oferecer suporte à decisão. No *pipeline* proposto, esse processamento deliberativo se articula diretamente com mecanismos de reflexão e autoconsciência operacional, que atuam como controle de qualidade do raciocínio e como gatilho de realimentação quando as evidências são insuficientes. O resultado esperado é que as saídas geradas sejam precisas, justificáveis e auditáveis, em conformidade com requisitos operacionais e de segurança.

2.2.3 Laço de Raciocínio, Ação e Recuperação de Evidências

2.2.3.1 Reflexão e Auto-Consciência

Por trás dessas metodologias está a emergência do pensamento *System 2* em LLMs, marcado por raciocínio deliberado e de processo lento [4]. A reflexão funciona como um mecanismo de realimentação verbal, no qual o modelo analisa o próprio traço de raciocínio e induz correções pontuais antes de gerar a saída final (abordagem descrita como benéfica em tarefas agênticas complexas [43]). Em paralelo, a autoconsciência operacional, frequentemente discutida como calibração de confiança, refere-se à capacidade do modelo de reconhecer limites do próprio conhecimento. *Frameworks* recentes para desacoplar metacognição de cognição têm permitido prever falhas potenciais e governar a incerteza com maior precisão [44].

Na prática, os dois mecanismos atuam sobre um risco comum: em cadeias agênticas longas, um erro não corrigido cedo tende a se propagar e contaminar passos seguintes. A reflexão opera como um freio, revisando o raciocínio interno antes que ele seja convertido em saída. A autoconsciência fecha o circuito ao permitir que o sistema admita o que não sabe, em vez de preencher lacunas com invenção. É essa capacidade que, no *pipeline*, dispara o sinal de *Knowledge Gaps*. Quando o Supervisor identifica que as evidências não sustentam uma conclusão, abre-se nova rodada de coleta, em vez de forçar uma resposta. O principal ponto de atenção é impor limites: reflexão sem critério de parada tende a virar um laço que consome inferência sem agregar precisão.

2.2.3.2 Raciocínio e Ação Sinérgicos (ReAct)

Para aproximar raciocínio interno e recuperação de conhecimento externo, o paradigma ReAct (*Reason + Act*) é adotado. Diferentemente do CoT puro, que se apoia apenas no conhecimento paramétrico interno, o ReAct induz um ciclo intercalado de traços de raciocínio e ações específicas (por exemplo, consultar uma base de dados ou filtrar resultados). Isso permite que o agente ajuste a trajetória de planejamento em tempo de execução, com base em observações vindas de ferramentas externas, reduzindo a propagação de erros e alucinações em tarefas de recuperação de informação em múltiplos passos [40].

A eficácia do ReAct depende da qualidade das observações retornadas pelas ferramentas. Se a busca devolve ruído, o ciclo de raciocínio pode ajustar a rota na direção errada e consumir chamadas sem convergir. O *pipeline* contorna esse risco ao não deixar o ReAct improvisar escopo: ele opera dentro do roteiro em DAG fixado na fase de planejamento, de modo que cada ação intercalada executa um nó previsto, e não uma decisão criada do zero a cada passo. Assim, o raciocínio em tempo real fica reservado a como cumprir a tarefa, e não a escolher qual tarefa executar.

2.2.3.3 RAG e Integração com Grafos

O RAG tradicional vem sendo gradualmente complementado por abordagens do tipo *Graph-Enhanced RAG* (*GraphRAG*). Ao estruturar o conhecimento recuperado não como blocos vetoriais isolados, mas como um grafo de conhecimento, o modelo pode percorrer relações entre entidades e responder a consultas globais que exigem entendimento mais holístico. Essa abordagem tende a reduzir alucinações ao ancorar

a geração em dependências de dados estruturadas e verificáveis [45].

Na prática, *GraphRAG* tem custo operacional, pois montar e manter o grafo de conhecimento requer processamento, e consultas simples (por exemplo, uma data ou um número) não precisam percorrer relações complexas. Daí a opção do *pipeline* por um motor híbrido, que escolhe a modalidade conforme a pergunta: recuperação vetorial quando basta localizar trechos semanticamente próximos e travessia em grafo quando a resposta exige seguir conexões entre entidades.

2.2.4 Supervisão Metacognitiva e Refinamento Iterativo

2.2.4.1 Meta-Agentes e Metacognição

No nível arquitetural, essas capacidades convergem na metacognição: uma camada de governança que monitora e regula as estratégias adotadas pelo sistema. Sua implementação se dá por meio de meta-agentes, isto é, unidades autônomas hierárquicas capazes de operacionalizar o conceito de design automatizado de sistemas agênticos (*Automated Design of Agentic Systems – ADAS*), gerando e otimizando dinamicamente agentes subordinados para subtarefas específicas [46, 47]. Para isso, o *meta-prompting* funciona como técnica de apoio (*scaffolding*), abstraindo a camada de instruções e viabilizando raciocínio de alto nível. Essa tríade contribui para transformar LLMs estáticos em sistemas multiagente hierárquicos e adaptativos (*HiMATE*), capazes de planejamento autônomo de longo horizonte e de autoevolução [48].

A viabilidade dessa abstração é reforçada pela transição de instruções textuais puras para chamadas com lógica programática. *Frameworks* recentes como o *DSPy* [25] indicam que compilar cadeias de linguagem em *pipelines* declarativos tende a estabilizar o comportamento dos agentes, enquanto arquiteturas como o *MetaGPT* [49] mostram que impor procedimentos operacionais padrão (*SOPs*) a um conjunto de agentes colaborativos reduz alucinações de planejamento. Sob essa ótica, o *meta-prompting* deixa de ser apenas uma lógica de formatação e passa a atuar como um compilador do sistema, traduzindo intenção estratégica em rotinas executáveis por unidades menores e especializadas.

No sistema proposto, essa tríade se materializa em um Supervisor cognitivo: a camada que, diante de cada subpergunta, decide qual técnica instanciar e qual modelo de base acionar. Ao tratar o *meta-prompting* como uma camada de orquestração isolada, a decisão torna-se portável. Isso abstrai a instrução do modelo concreto, permitindo que a mesma lógica de governança opere sobre diferentes arquiteturas *open-source*. O sistema pode trocar de modelo conforme a exigência da tarefa, sem reescrever o código de orquestração.

O contrapeso dessa autonomia é impor limites à auto-organização. Gerar agentes subordinados sem freios claros pode levar a explosão combinatória ou a uma governança pouco auditável [50]. Por isso, o Supervisor seleciona técnicas a partir de um catálogo definido, em vez de criar unidades operacionais indefinidamente. Essa restrição mantém adaptabilidade cognitiva em tempo de execução, mas sem abrir mão da rastreabilidade epistemológica, que é requisito central em um sistema de inteligência.

2.2.4.2 Planejamento Cognitivo Auto-Evolutivo

No topo da escala de autonomia do sistema está o planejamento cognitivo autoevolutivo, em que os agentes refinam, de forma iterativa, seus planos de execução com base em *feedback* do ambiente, criando um laço de aprendizado dinâmico distinto do treinamento estático. Esse comportamento costuma ser implementado via coordenação multiagente, com agentes especializados (planejador, executor e crítico) atuando em conjunto. *Frameworks* recentes utilizam ajustes hierárquicos dinâmicos, permitindo que o conjunto de agentes se auto-organize e otimize protocolos de comunicação em tempo real para resolver tarefas de maior entropia [49, 50].

Essa autonomia tem custo em sistemas de produção de conhecimento e apoio à decisão. Quando o conjunto se auto-organiza sem restrições, o comportamento tende a ficar menos previsível e reprodutível: duas execuções sobre a mesma pergunta podem divergir sem explicação clara. Como rastreabilidade é requisito, o *pipeline* adota uma versão controlada do planejamento autoevolutivo. Mantém-se o trio planejador–executor–crítico, mas o crítico segue heurísticas fixas e cada decisão de descarte é registrada. Assim, o sistema adapta o plano dentro de limites que podem ser auditados posteriormente.

2.2.4.3 *Self-Refinement* e *Self-Consistency*

Self-Refinement e *Self-Consistency* são estratégias de engenharia de *prompt* usadas para aumentar a confiabilidade de respostas geradas por LLMs. No *Self-Refinement*, o modelo revisa e reescreve a própria saída em ciclos de *feedback*, com rotinas de autoverificação aplicadas em tempo de teste [51]. Já a *Self-Consistency* explora múltiplos caminhos de raciocínio e seleciona os mais consistentes, com variações recentes buscando reduzir custo e melhorar a precisão [52]. Em conjunto, essas abordagens podem reduzir erros e alucinações quando integradas a critérios de verificação.

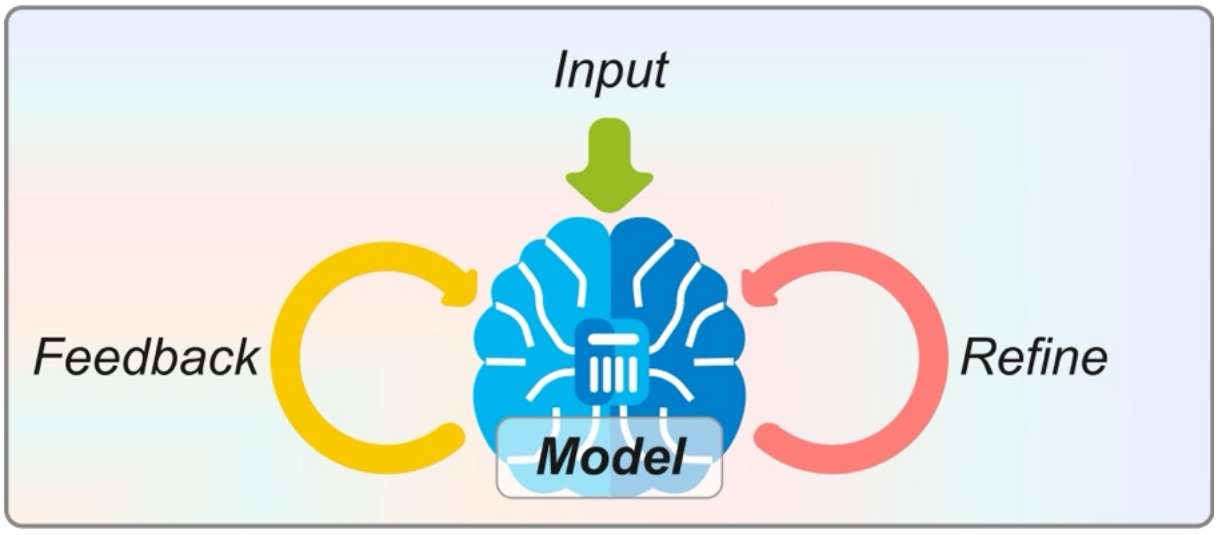

Figura 2.2: Esquema conceitual do *Self-Refine*. Fonte: [2].

Há limites conhecidos nessas técnicas. Quando o modelo julga a si mesmo, seus pontos cegos podem se transferir para a avaliação. Um erro cometido com convicção tende a sobreviver à autoverificação, e a *Self-Consistency* pode convergir, de forma estável, para uma resposta consistentemente errada. Por isso,

o *pipeline* não depende de introspecção isolada. Os laços de refinamento são correlacionados a sinais externos ao modelo, incluindo evidências de origem e métricas objetivas, de modo que a correção seja verificada de forma mensurável, e não apenas pela "opinião" do gerador.

2.2.5 Geração de Texto Longo e Avaliação da Saída

2.2.5.1 Arquiteturas Hierárquicas para Texto Longo

A decodificação autorregressiva padrão tende a sofrer com desvio semântico e com o fenômeno *lost-in-the-middle* na geração de relatórios extensos [7]. Para mitigar esse problema, *frameworks* recentes têm migrado da geração linear para planejamento hierárquico. A arquitetura DOME (*Dynamic Hierarchical Outlining with Memory-Enhancement*) endereça essa limitação ao dissociar a fase de planejamento da fase de geração de conteúdo e ao usar um módulo de memória dinâmica para manter consistência entre seções durante a redação [5]. De forma complementar, o RAPID propõe um fluxo orientado à descoberta, no qual o modelo identifica lacunas de informação e recupera evidências específicas antes de sintetizar conteúdo de longo fôlego, buscando garantir que a narrativa final seja factualmente ancorada e estruturalmente coesa [6].

A hierarquia supera a geração linear por um motivo prático. Ao planejar primeiro um esqueleto e preencher cada seção localmente, o sistema não obriga o modelo a manter o relatório inteiro em uma única janela de contexto, justamente onde a perda de informação tende a ocorrer. Essa abordagem demanda controle adicional para manter o esqueleto coerente e integrar seções escritas separadamente. Para relatórios longos, essa troca costuma compensar. É sobre essa base que o *pipeline* constrói sua etapa de geração, combinando o planejamento do DOME com a densificação iterativa.

Gerenciar a dispersão informacional em janelas de contexto extensas exige ir além do carregamento estático de dados. Como apontam Mao et al. (2025) [53] em um estudo sobre refinamento hierárquico, a estruturação deliberada do fluxo informacional em etapas intermediárias é determinante para a estabilidade semântica do contexto.

2.2.5.2 Densificação Iterativa e *Chain of Density*

Em relatórios de inteligência, a sumarização padrão pode falhar ao priorizar brevidade em detrimento de detalhe, omitindo entidades críticas. A metodologia *Chain of Density* (CoD) introduz um *framework* de *prompting* iterativo que aumenta progressivamente a densidade informacional do texto gerado. Diferentemente da geração *one-shot*, o CoD obriga o modelo a identificar e incorporar entidades ausentes na versão anterior do rascunho, sem aumentar o total de *tokens*. O resultado é uma narrativa mais densa, que eleva a carga informacional por *token* e se ajusta a exigências de sistemas de apoio à decisão, nos quais ambiguidade e verbosidade são prejudiciais [54].

A densidade, contudo, tem limites e precisa ser calibrada. Levado longe demais, o CoD pode produzir o efeito oposto ao pretendido: um parágrafo saturado de entidades vira um bloco de fatos com pouca articulação, elevando a carga cognitiva do leitor. A densificação, portanto, não pode ser cega; deve ser maximizada até um limiar de legibilidade. Por isso, o *pipeline* associa o CoD a uma restrição de legibilidade (Flesch), regulando o resultado obtido.

2.2.5.3 Métricas Quantitativas

Para garantir a confiabilidade de agentes autônomos, métricas objetivas atuam como sinais de *feedback* no refinamento iterativo. A Perplexidade (PPL) funciona como proxy para a incerteza do modelo, em que valores baixos correlacionam com fluência e aderência à distribuição de treinamento, ainda que não impliquem em maior precisão factual. Para avaliar a capacidade do modelo de lidar com dependências de longo alcance, utiliza-se o *benchmark* LAMBADA, que testa a predição de texto exigindo contexto de passagens anteriores, sendo crítico para sustentar arcos narrativos em relatórios [55]. O *Flesch Reading Ease* é aplicado como parâmetro de otimização restrita na fase de reescrita, garantindo que a saída se ajuste à carga cognitiva do público-alvo [56].

É preciso ser claro sobre o que essas métricas não fazem: nenhuma delas mede verdade. Perplexidade, LAMBADA e Flesch capturam fluência, coerência de longo alcance e legibilidade. Um relatório pode pontuar bem nos três e ainda estar factualmente errado. No pipeline, elas operam como balizas automáticas para a qualidade da forma durante o refinamento, enquanto a correção factual é abordada através da ancoragem de cada afirmação relevante às evidências de origem.

2.3 TRABALHOS RELACIONADOS E ESTADO DA ARTE

A geração automática de relatórios analíticos extensos passou a ser tratada como problema de arquitetura, e não de sumarização. As propostas recentes separam a fase de investigação da fase de redação e distribuem o trabalho entre agentes especializados. Esta seção situa o artefato proposto frente a essa literatura e deriva os critérios aplicados na avaliação do Capítulo 4.

2.3.1 Geração Agêntica de Relatórios Longos

O sistema STORM, de Shao *et al.* [57], foi um dos primeiros a tratar a etapa anterior à escrita como objeto de projeto. Em vez de solicitar o texto diretamente ao modelo, levanta perspectivas distintas sobre o tema, simula diálogos entre um redator e um especialista ancorado em fontes recuperadas, e converte o material em um esboço que só então é expandido. Os autores registram duas falhas persistentes: a transferência do viés das fontes para o texto final e a associação indevida entre fatos sem relação entre si.

OmniThink [58] parte da mesma constatação por outro caminho. Alterna expansão e reflexão sobre o tema, mantendo uma árvore de informação e um repositório de conceitos enriquecidos a cada iteração, de modo a ampliar o escopo que o modelo considera antes de escrever. Do trabalho vem a métrica de densidade de conhecimento, que mede a proporção de informação útil no texto produzido.

DOME [5] e RAPID [6] atacam o mesmo gargalo pelo lado da estrutura, o primeiro com esboço hierárquico ajustado durante a geração, o segundo acoplando o planejamento da escrita à descoberta de informação. Do lado da recuperação, as abordagens discutidas na Seção 2.2.3.3 respondem à degradação em contexto longo por caminhos distintos, seja pela navegação sobre grafo de conhecimento, seja pela crítica do próprio texto quanto ao apoio nas evidências.

O denominador comum é a separação entre planejamento e redação, princípio que a arquitetura do Capítulo 3 também adota. A diferença de partida está em outro ponto. Nenhum desses sistemas foi projetado para operar sem acesso à internet aberta, e quase todos pressupõem modelos comerciais de grande porte acessados remotamente. A Tabela 2.1 resume o posicionamento.

Tabela 2.1: Posicionamento dos trabalhos relacionados

Sistema	**Planejamento prévio**	**Recuperação**	**Execução local**
STORM [57]	Perspectivas e esboço	Web aberta	Não
OmniThink [58]	Árvore de informação	Web aberta	Não
DOME [5]	Esboço dinâmico	Não	Não
RAPID [6]	Acoplado à descoberta	Vetorial	Não
GraphRAG [45]	Não	Grafo e vetorial	Parcial
Self-RAG [59]	Não	Crítica de contexto	Sim
Pipeline proposto	DAG 5W3H e H-CoD	Relacional, vetorial e grafo	Sim, isolado

Fonte: Elaborada pelo autor (2026).

2.3.2 Lacuna Identificada e Critérios de Avaliação

Três observações decorrem dessa leitura, e delas derivam as dimensões usadas na avaliação experimental. A primeira é que a ancoragem factual permanece o ponto de falha mais recorrente, presente tanto na associação indevida relatada pelos autores de STORM quanto na degradação em contexto longo que motiva a literatura de recuperação, o que sustenta a dimensão de precisão factual. A segunda é que a quantidade de informação útil por unidade de texto é propriedade distinta da extensão, formalizada por OmniThink na métrica de densidade de conhecimento e tratada pela densificação iterativa discutida na Seção 2.2.5.2, o que sustenta a dimensão de densidade informacional. A terceira é que a estrutura de textos longos se degrada conforme o texto avança, motivo das arquiteturas hierárquicas de DOME e RAPID, o que sustenta a dimensão de coerência lógica.

A lacuna que este trabalho ocupa está na interseção. Nenhum dos sistemas revisados reúne planejamento hierárquico prévio, recuperação híbrida sobre três formas de armazenamento e execução integralmente local em ambiente isolado. Os dois primeiros elementos aparecem separadamente e sempre apoiados em modelos remotos. O terceiro raramente é considerado, por não constituir requisito nos domínios em que essa literatura costuma ser avaliada. Em inteligência estratégica, ele é condição de emprego, e é essa combinação que o Capítulo 3 descreve e o Capítulo 4 submete a teste.

3 ARQUITETURA DO SISTEMA

Este capítulo descreve a implementação arquitetural do sistema proposto, conectando os *frameworks* teóricos discutidos no capítulo anterior a restrições práticas de engenharia. A filosofia de projeto adota princípios alinhados a metodologias consagradas de produção de conhecimento, em particular o Ciclo de Inteligência da OTAN definido em AJP-2.1 [60]. Embora respeite as fases canônicas de Direção, Coleta e Processamento analisadas por Lowenthal [61], a arquitetura proposta eleva o nível de abstração do controle de fluxo: em vez de apoiar-se em *scripts* procedurais rígidos, o pipeline delega autoridade decisória a agentes autônomos baseados em LLM, organizados em torno de um contexto operacional flexível.

Esse deslocamento em direção à abstração agêntica amplia a capacidade de generalização do processo de resolução de problemas. Ao tratar o pipeline não como uma sequência de funções, mas como uma cadeia de estados com transições auditáveis, a arquitetura passa a lidar com perguntas complexas de raciocínio em múltiplos estágios, que exigem adaptabilidade dinâmica. Cada etapa foi concebida para mitigar gargalos cognitivos ou computacionais identificados na literatura, mapeando técnicas recentes como GraphRAG, CoT e ReAct para a fase de processamento em que oferecem maior utilidade e melhor desempenho.

A arquitetura resultante é um laço adaptativo de seis etapas: Observação, Direção, Coleta, Processamento, Análise e Geração de Relatório. Como mostra a Figura 3.1, esses estágios operam sob uma camada de governança metacognitiva, entendida aqui como o controle metacognitivo do fluxo de execução do *pipeline* e de seus laços de retroalimentação, e não como governança institucional. As seções a seguir descrevem a lógica operacional de cada fase, mostrando como metodologias específicas de IA viabilizam o ciclo de inteligência e transformam dados brutos não estruturados em narrativas analíticas densas e rastreáveis.

A coerência operacional do pipeline é garantida por um repositório de estado de conhecimento unificado, que funciona como memória de trabalho persistente entre os estágios. Essa escolha arquitetural permite ao sistema escapar do processamento linear de consultas e atuar, de fato, como um motor eficiente de compreensão textual. Ao refinar esse contexto compartilhado de forma iterativa e transformar observações brutas da Coleta em conhecimento estruturado na Análise, o sistema busca atingir a profundidade cognitiva exigida em decisões de alto risco, tratando a natureza e os detalhes da pergunta inicial com precisão.

3.1 ETAPAS DO PIPELINE

3.1.1 Etapa 1: Observação

A fase de Observação transforma a entrada bruta do usuário em um pedido claro e desambiguado. Antes de qualquer raciocínio, o sistema sanitiza a mensagem, removendo artefatos sem valor informacional (assinaturas de e-mail, avisos legais e similares) e preservando a intenção central. Esse pré-processamento reduz ruído, diminui a superfície de *prompt injection* e melhora a relação sinal-ruído para as etapas seguin-

tes [62].

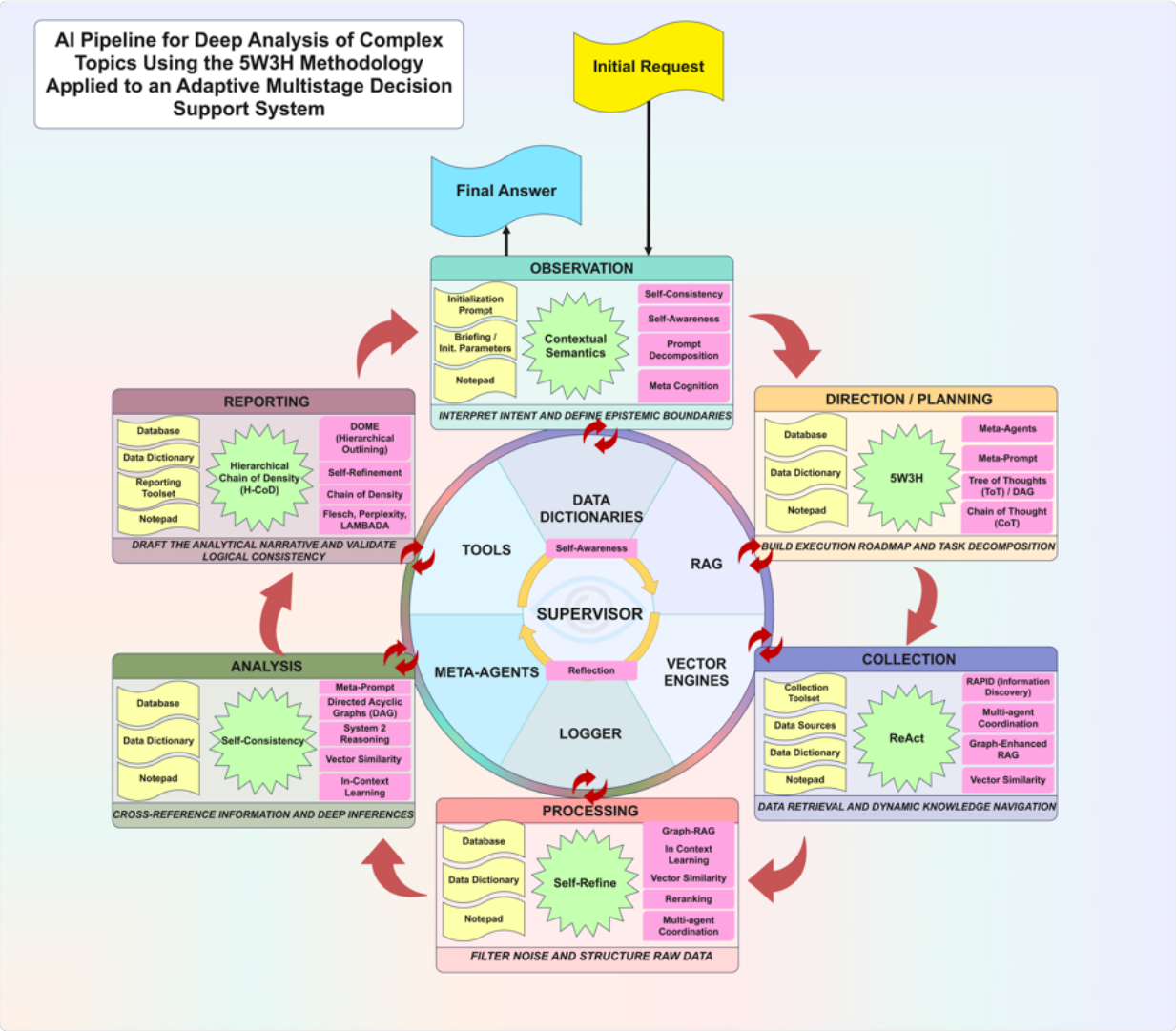

Figura 3.1: Pipeline do sistema, organizado em seis etapas adaptativas (Observação, Direção, Coleta, Processamento, Análise e Geração de Relatório) sob uma camada metacognitiva centralizada (Supervisor). Adaptado de [3].

Após a sanitização, o sistema aplica uma estratégia computacional adaptativa que funciona como roteador *zero-shot*. Esse mecanismo classifica a complexidade do pedido em camadas hierárquicas: Resposta Direta (*Tier* 3), Consulta *Web* (*Tier* 2) ou Análise Profunda (*Tier* 1). A lógica de roteamento se alinha a *frameworks* recentes do tipo "*RouteLLM*", que indicam que alocar dinamicamente o orçamento de inferência conforme a complexidade da consulta reduz latência e custo sem perder precisão [63]. Ao identificar gatilhos como "discussão multiperspectiva" ou "planejamento estratégico", o sistema evita gastar recursos de raciocínio *System 2* em consultas factuais triviais.

Por fim, a intenção interpretada é normalizada em um contrato em formato JSON, com *schema* rígido e *parseável* por máquina. Diferente da decomposição de linguagem natural padrão, esse passo impõe uma saída estruturada que mapeia restrições implícitas do usuário (formato de saída, perfil de público, idioma, localidade, limites temporais e outros) em parâmetros explícitos de execução. A abordagem se

apoia em técnicas de *Constrained Decoding* para garantir que a fase de planejamento subsequente receba um objeto estruturado e determinístico, em vez de texto ambíguo, reduzindo a propagação de erros nas cadeias agênticas *multi-step* [64].

3.1.2 Etapa 2: Direção (Planejamento)

A fase de Direção é responsável pelo planejamento. Ela consome o contrato JSON gerado em Observação e o converte em um plano de execução. Em vez de concentrar planejamento e resposta em uma única passada de geração, o *pipeline* separa essas funções para reduzir a propagação de erros de interpretação.

A intenção é desdobrada segundo a matriz 5W3H, gerando subperguntas que cobrem atores, tempo, recorte geográfico e critérios de sucesso. Cada subpergunta recebe metadados de execução e já é mapeada para uma tarefa de recuperação ou de análise, evitando que a definição de escopo fique para as etapas seguintes.

Com base nesse conjunto estruturado de questões, o sistema instancia um mecanismo *Tree of Thoughts* (ToT), supervisionado por um agente metaplanejador. Em lugar de produzir um único roteiro, o planejador expande múltiplos ramos candidatos de execução e os submete a um agente crítico, que avalia cada plano contra regras objetivas, tais como a cobertura da matriz 5W3H, profundidade estimada de raciocínio, custo de inferência e risco de redundância. Os ramos descartados não desaparecem do estado epistêmico compartilhado; eles permanecem registrados como hipóteses fracas, acompanhadas da justificativa do descarte, o que permite às fases posteriores retomar caminhos alternativos caso surjam evidências contraditórias durante a Coleta. Essa lógica de *backtracking* explícito é apontada pela literatura recente como condição mínima para planejamento autoevolutivo em sistemas multiagente hierárquicos [46, 48]. Essa mesma lógica também compõe os registros de auditoria que sustentam a rastreabilidade das decisões e dos caminhos tomados.

Como saída, a fase produz um roteiro formal em DAG, no qual os nós representam tarefas atômicas (consulta vetorial, busca em grafo, validação de hipótese, comparação cruzada) e as arestas codificam dependências de dados e ordem de execução. Esse grafo é serializado de volta ao mesmo contrato estruturado iniciado em Observação, agora estendido com o plano de execução, garantindo continuidade lógica entre os estágios. É aqui que o planejamento se separa formalmente da geração, de forma que o que segue para a Coleta não é uma instrução em linguagem natural, mas uma especificação determinística e auditável, na qual cada decisão de busca, refinamento e síntese pode ser rastreada até a subpergunta 5W3H que a originou. Esse protocolo estendido reduz a propagação de ambiguidade ao longo da cadeia agêntica e cria condições para que os agentes posteriores operem como executores especializados, e não como improvisadores semânticos [64, 44].

Para tangibilizar a capacidade de desdobramento investigativo do sistema, a Tabela 3.1 ilustra a decomposição estruturada de uma macro-intenção. Neste cenário metalinguístico, a ferramenta aplica o modelo 5W3H para investigar as propriedades de operação, implementação e impacto do próprio objeto desta pesquisa: um *Pipeline de IA para Análise Profunda de Tópicos Complexos*.

A exigência procedimental força a geração de exatamente três questionamentos atômicos para cada um dos oito eixos da matriz. Esse processo impede a formulação de consultas rasas e garante que as etapas

subsequentes de Coleta e Análise recebam um grafo de tarefas exaustivo.

A tabela demonstra como a rigidez imposta pela formulação estruturada inibe a ambiguidade. Ao transformar uma única diretriz em vinte e quatro vetores de investigação direcionados, o sistema assegura que a exploração informacional não dependerá da intuição estocástica do modelo na fase de coleta.

3.1.3 Etapa 3: Coleta

A fase de Coleta opera como o braço executor do pipeline, consumindo o roteiro em DAG produzido pela Direção e materializando, em paralelo ou em sequência (conforme as dependências entre nós), cada uma das tarefas atômicas de recuperação. Em vez de tratar a busca como uma chamada única e indiferenciada a um índice vetorial, esse estágio adota o paradigma ReAct, intercalando ciclos de raciocínio e ação em que o agente decide, a cada passo, se aciona uma fonte *web*, uma base local ou o cache semântico já populado [40]. Esse roteamento adaptativo é mediado por um *Search Agent*, que primeiro consulta o cache de respostas curadas, viabilizando recuperação imediata para consultas semanticamente equivalentes, e só então despacha buscas mais caras a *engines* externas ou a repositórios internos, reduzindo latência e custo computacional, sem comprometer a cobertura informacional.

A coleta propriamente dita é executada por uma camada multimodal de ingestão, capaz de assimilar páginas *web*, arquivos estruturados, documentos PDF e imagens curadas como entradas válidas. *Crawlers* adaptativos, instanciados como subagentes especializados pelo *Search Agent*, navegam links de profundidade variável conforme a relevância estimada, enquanto um *Markdown/OCR Processor* converte conteúdo não textual em representações legíveis pela máquina por meio de redes neurais convolucionais (CNNs, *Convolutional Neural Networks*) e do *Transformer-based OCR* (TrOCR), além de modelos multimodais. Sobre essa massa heterogênea atua um *Extraction Agent*, encarregado de isolar entidades, tabelas, imagens e blocos textuais junto ao respectivo contexto de origem, preservando proveniência, autoria e cadeia de custódia da informação, requisitos relevantes em fluxos de inteligência estratégica em que a rastreabilidade das fontes é tão importante quanto o conteúdo recuperado.

Antes de cruzar a fronteira para o estágio seguinte, todo o material extraído passa por um *chunker* contextual dinâmico, que segmenta os blocos por fronteiras semânticas (e não por contagem arbitrária de *tokens*) antes de serem vetorizados e persistidos na estratégia de armazenamento híbrido descrita em detalhe na Seção 3.2.1. A triangulação de dados entre bases relacionais, vetoriais e de grafo evita a perda de relações estruturais característica do RAG ingênuo e prepara o terreno para consultas *multi-hop* dos estágios posteriores. Quando o *Query Constructor* identifica lacunas informacionais, entidades centrais não recuperadas, contradições não resolvidas ou *gaps* temporais, um sinal de *Knowledge Gaps* reativa a Coleta em modo iterativo, fechando um laço de *feedback* autônomo análogo ao *Self-RAG*, em que o sistema critica a suficiência da própria recuperação antes de declarar o material maduro para o prosseguimento do pipeline [59].

Tabela 3.1: Decomposição 5W3H aplicada ao escopo operacional do Pipeline de IA proposto - montado pelo agente supervisor (IA)

Eixo	Desdobramento interrogativo (subtarefas)
What	1. O que constitui o estado epistêmico unificado persistido entre os seis estágios? 2. O que engatilha autonomamente o laço de re-coleta no vetor de *Knowledge Gaps*? 3. O que caracteriza matematicamente o limiar de densidade aceitável no algoritmo H-CoD?
Why	1. Por que o desacoplamento arquitetural entre planejamento e execução minimiza o desvio semântico? 2. Por que a travessia em grafo resolve consultas analíticas que o RAG vetorial puro falha em mapear? 3. Por que a validação de divergências prioriza métricas de similaridade de cosseno frente ao julgamento estocástico?
Who	1. Quem supervisiona e aprova as atualizações das rotinas de reflexão no *cache* global? 2. Quem são os atores institucionais impactados pelo ganho de latência gerado pela execução local? 3. Quem assume o papel de validador de desempate durante o protocolo de *LLM-as-a-Judge* multi-juiz?
When	1. Quando o motor híbrido prioriza a correspondência exata (BM25) em detrimento dos *embeddings* densos? 2. Quando ocorre o expurgo, por baixa frequência de uso, de entradas ociosas da memória compartilhada? 3. Quando a arquitetura hierárquica DOME determina que uma seção de texto atingiu sua saturação informativa?
Where	1. Onde ocorrem os maiores gargalos de processamento durante a etapa paralela de *Self-Refine*? 2. Onde a topologia armazena o acervo de evidências para garantir o bloqueio de estado de concorrência? 3. Onde o isolamento *air-gapped* se prova mais crítico do que o uso de *datacenters* globais?
How	1. Como o In-Context Learning adapta a competência do agente sem a necessidade de reescrever pesos sinápticos? 2. Como a representação vetorial concilia variantes lexicais (e.g., "EUA"e "Estados Unidos") em uma entidade canônica única? 3. Como a estratégia *coarse-to-fine* comprime o volume textual sem sacrificar a legibilidade (índice Flesch)?
How Much / Many	1. Quanto do orçamento computacional (FLOPs) é economizado pela ativação exclusiva de modelos esparsos? 2. Quantas fontes heterogêneas distintas o extrator consolida, em média, por nó de evidência no grafo? 3. Quantos *tokens* de contexto a memória *Key-Value Cache* retém de maneira eficiente sem degradação na janela de inferência?
How Long	1. Por quanto tempo a latência global da inferência é estendida a cada ciclo de retroalimentação disparado pelo supervisor metacognitivo? 2. Qual é a duração de validade (tempo de vida) dos artefatos retidos no *cache* semântico antes que a poda autônoma os descarte? 3. Por quanto tempo o sistema sustenta a exploração de ramificações no planejamento (ToT) antes de forçar a convergência da tarefa?

3.1.4 Etapa 4: Processamento

A fase de Processamento transforma o material recolhido pela Coleta em representações limpas, normalizadas e prontas para a etapa de Análise. Em termos práticos, esse estágio concentra rotinas como deduplicação entre fontes, normalização de formatos (datas, unidades, moedas, transliterações), tradução automática quando necessária e harmonização de codificações textuais. Sem esse saneamento, ruídos do meio de coleta degradariam o *reranking* e, por consequência, a inferência.

Essas rotinas são executadas por subagentes instanciados por *metaprompts*. A adaptação à tarefa ocorre, portanto, por *In-Context Learning* (ICL; ver Seção 2.2.1.2), isto é, por especialização em tempo de inferência sem atualização de pesos e sem *fine-tuning* persistente, preservando portabilidade entre modelos de base [26].

Sobre o material já saneado atua a camada de vetorização (Seção 3.2.2). Cada *chunk* semanticamente delimitado é projetado em um espaço latente por modelos de *embedding* contextual (BGE-M3 ou *text-embedding-3*), enquanto as entidades extraídas passam por resolução referencial, reconciliando variantes ("EUA", "Estados Unidos", "*United States*") em identificadores canônicos. Esses identificadores alimentam, em paralelo, a construção do grafo de conhecimento, viabilizando consultas *multi-hop* que dependem de estrutura, e não apenas de proximidade semântica.

Para manter fidelidade à fonte original, o *pipeline* aplica laços de refinamento/verificação em tempo de teste (Seção 2.2.4.3), comparando *embeddings* e relações inferidas com o trecho de origem e descartando ou regenerando entradas abaixo de um limiar de similaridade [2, 51].

Como saída, a fase entrega um estado epistêmico indexado e distribuído pelas três modalidades de armazenamento (relacional, vetorial contextualizada e grafo de entidades), com metadados de proveniência preservados. Esse contrato de saída distingue um repositório de fragmentos correlatos de uma base de conhecimento pesquisável: cada *chunk* pode ser localizado por similaridade semântica, por correspondência exata (BM25 sobre o relacional) ou por travessia de relações no grafo, conforme o tipo de subpergunta gerada pela Direção. É essa triangulação que habilita o *reranking* deliberativo da etapa seguinte, em que o *Curator/Reranker* aplica *Self-Consistency* e o *framework* RAPID sobre os candidatos retornados pelo *Query Constructor*, isolando a informação relevante antes de servi-la à Análise [52, 6].

3.1.5 Etapa 5: Análise

A fase de Análise é o estágio em que o sistema deixa de operar sobre dados e passa a operar sobre evidências, executando raciocínio deliberativo do tipo *System 2* sobre o material curado nas etapas anteriores. É importante frisar que essa deliberação é implementada na orquestração do agente (seleção de técnicas, filtros, laços de verificação e replanejamento), e não por qualquer alteração interna dos pesos do modelo.

O processo começa pelo *Curator/Reranker*, que consome os candidatos retornados pelo *Query Constructor*, combinando busca vetorial, BM25 e travessia em grafo, com filtros alinhados às subperguntas 5W3H originadas na Direção. Esse *reranking* aplica três camadas de validação: (i) similaridade vetorial determinística entre o candidato e a subpergunta, (ii) *Self-Consistency* entre múltiplos traços de raciocínio independentes [52] e (iii) o *framework* RAPID, que confronta cada evidência com o plano de cobertura

informacional explicitado na etapa de Direção [6]. Apenas o que sobrevive a esse filtro é admitido como evidência para a inferência.

Sobre essa base evidencial, o sistema aciona técnicas analíticas registradas em seu catálogo interno (análise comparativa, ACH, construção de cenários, identificação de pontos de inflexão [65], entre outras), selecionadas pelo Supervisor Metacognitivo conforme o perfil da subpergunta e o domínio inferido em Observação. A parametrização das técnicas é feita por *metaprompts* (ICL; Seção 2.2.1.2), mantendo portabilidade entre modelos de base [26]. O raciocínio não se desdobra de forma linear: é organizado como um DAG no estilo *Graph of Thoughts*, em que nós representam conclusões parciais e arestas codificam dependências lógicas, permitindo agregar caminhos de inferência e detectar inconsistências antes do relatório final [1].

Como saída, a fase produz um conjunto consolidado de achados: afirmações estruturadas, com nível de confiança calibrado e ancoradas nas evidências de origem, prontas para a fase de síntese e construção narrativa. Se, durante a inferência, o sistema identifica hipóteses subdeterminadas ou contradições materiais entre fontes que não podem ser resolvidas com o material disponível, um sinal de *Knowledge Gaps* reativa a Coleta de forma direcionada, sem reprocessar o que já foi validado. Esse fechamento condicional do laço de execução distingue o *pipeline* de uma cadeia agêntica linear: o sistema pode recusar o estado epistêmico vigente e exigir nova evidência antes de prosseguir.

3.1.5.1 Dicionário de Técnicas Analíticas

Para operacionalizar o raciocínio deliberativo (*System 2*), o módulo de Análise não depende da intuição estocástica do modelo de base. Em vez disso, ele acessa um objeto de configuração estrito que mapeia as Técnicas Analíticas Estruturadas (SATs). O Supervisor Metacognitivo consome este dicionário em tempo de execução para selecionar dinamicamente a abordagem ideal, com base na natureza da subtarefa e no perfil das evidências recuperadas.

Abaixo, apresenta-se a estrutura de dados real (*payload*) que parametriza o comportamento analítico do sistema:

```
TECNICAS_DE_ANALISE: dict[str, TecnicaAnalise] = {
  "1": {
    "code": "KAC",
    "name": "Verificação de Premissas-Chave",
    "description": (
      "Um processo estruturado para identificar e desafiar as premissas "
      "fundamentais sobre as quais um plano ou análise se baseia."
    ),
    "primaryQuestion": (
      "As premissas centrais que sustentam este esforço são realmente "
      "verdadeiras? O que acontece se elas forem falsas?"
    ),
    "whenToUse": [
      "**Contexto:** [Validação], [Redução de Riscos], [Individual/Grupo]",
      "- Análise Post-Mortem: Descobrir se uma premissa falha levou ao fracasso.",
      "- Antes de um Grande Compromisso: Verificação de sanidade da lógica base.",
```

```
    "- Quando a Análise Está Estagnada: Descobrir premissas não declaradas.",
    "- Quando uma Análise Parece 'Perfeita Demais': Desafiar proativamente."
  ],
  "commonPitfalls": (
    "Pode ser ineficaz se o grupo não estiver aberto a críticas "
    "(pensamento de grupo). Não gera novas ideias, apenas testa as existentes."
  )
},
"2": {
  "code": "ACH",
  "name": "Análise de Hipóteses Concorrentes",
  "description": (
    "Método sistemático e baseado em evidências para avaliar explicações "
    "concorrentes, visando refutar hipóteses em vez de confirmar uma."
  ),
  "primaryQuestion": (
    "Qual destas possíveis explicações é a mais provável de ser verdadeira, "
    "com base nas evidências que temos?"
  ),
  "whenToUse": [
    "**Contexto:** [Diagnóstico], [Baseado em Dados], [Objetividade]",
    "- Problemas de Causalidade Complexa: Evidências ambíguas/contraditórias.",
    "- Superação de Vieses Fortes: Estrutura objetiva contra vieses.",
    "- Investigação de Alto Risco: Conclusão rigorosamente defensável.",
    "- Tentativa de Provar uma Negativa: Mostrar inconsistência de evidências."
  ],
  "commonPitfalls": (
    "Consome muito tempo. Impraticável para dezenas de hipóteses. "
    "Depende inteiramente da qualidade da matriz de evidências."
  )
},
"3": {
  "code": "SSC",
  "name": "Autocrítica Estruturada",
  "description": (
    "Uma revisão disciplinada, no estilo checklist, do próprio processo "
    "analítico e conclusões para encontrar falhas antes da finalização."
  ),
  "primaryQuestion": (
    "Eu fui rigoroso? Onde estão os pontos fracos em meu próprio "
    "argumento, fontes e lógica?"
  ),
  "whenToUse": [
    "**Contexto:** [Controle de Qualidade], [Individual], [Estágio Final]",
    "- Revisão Final: Último passo para capturar erros ou perspectivas ausentes.",
    "- Preparação para Audiência Hostil: Antecipar desafios possíveis.",
    "- Quando um Argumento Parece Frágil: Encontrar o elo fraco."
  ],
  "commonPitfalls": (
    "Suscetível aos pontos cegos do próprio analista. Não substitui "
    "uma revisão externa (ex: análise 'Red Team')."
  )
```

```
},
"4": {
  "code": "WIA",
  "name": "Análise 'E Se'",
  "description": (
    "Explora consequências potenciais de um evento único, específico "
    "e muitas vezes de baixa probabilidade/alto impacto."
  ),
  "primaryQuestion": (
    "Presumimos que X não vai acontecer. E se acontecer? Quais seriam "
    "as consequências e como deveríamos nos preparar?"
  ),
  "whenToUse": [
    "**Contexto:** [Planejamento de Contingência], [Estratégia], [Proativo]",
    "- Teste de Estresse: Testar resiliência contra um evento disruptivo.",
    "- Exploração de 'Cisne Negro': Ameaça improvável, mas catastrófica.",
    "- Desafiar a Complacência: Forçar a organização a pensar no desconfortável."
  ],
  "commonPitfalls": (
    "Foca em apenas um evento, potencialmente ignorando outros riscos. "
    "Pode levar à paralisia por análise se não for focado no alto impacto."
  )
},
"5": {
  "code": "MSG",
  "name": "Geração de Múltiplos Cenários",
  "description": (
    "Desenvolve vários futuros plausíveis e divergentes para moldar "
    "uma estratégia robusta de longo prazo."
  ),
  "primaryQuestion": (
    "Quais são os diferentes futuros possíveis e como construir "
    "uma estratégia robusta em todos eles?"
  ),
  "whenToUse": [
    "**Contexto:** [Estratégia de Longo Prazo], [Alta Incerteza], [Proativo]",
    "- Planejamento Estratégico Longo Prazo: Futuro incerto demais.",
    "- Navegando em Mudanças Transformadoras: Disrupção tecnológica/clima.",
    "- Grandes Investimentos Irreversíveis: Testar decisão em vários futuros."
  ],
  "commonPitfalls": (
    "Pode ser muito abstrato. Frequentemente resulta em cenários não "
    "verdadeiramente diferentes (bom, ruim e status quo)."
  )
},
"6": {
  "code": "TLN",
  "name": "Linhas do Tempo",
  "description": (
    "Organiza cronologicamente eventos e pontos de dados para entender "
    "sequência, duração, causalidade e lacunas."
  ),
```

```
  "primaryQuestion": (
    "Qual é a sequência dos eventos e o que essa sequência nos diz "
    "sobre o que aconteceu?"
  ),
  "whenToUse": [
    "**Contexto:** [Diagnóstico], [Investigação], [Organização de Dados]",
    "- Reconstrução de Incidente: Juntar peças de uma falha de sistema.",
    "- Análise de Causa Raiz: Mapear cadeia de eventos até o resultado.",
    "- Compreensão de Processo: Mapear passos para encontrar gargalos."
  ],
  "commonPitfalls": (
    "Correlação não é causalidade. Pode ser enganosa se houver lacunas "
    "significativas não explicitamente marcadas."
  )
},
"7": {
  "code": "BRS",
  "name": "Brainstorming",
  "description": (
    "Técnica de criatividade focada em gerar grande quantidade de "
    "ideias em um ambiente livre de julgamentos."
  ),
  "primaryQuestion": (
    "Quais são todas as possíveis hipóteses que conseguimos pensar, "
    "independentemente da viabilidade?"
  ),
  "whenToUse": [
    "**Contexto:** [Geração de Ideias], [Pensamento Divergente]",
    "- Fase Inicial Incerta: Explorar toda a gama de possibilidades.",
    "- Empacado em Solução Criativa: Sair de padrões de pensamento convencionais.",
    "- Geração Inicial de Hipóteses: Primeiro passo antes de técnicas como ACH."
  ],
  "commonPitfalls": (
    "Pode ser caótico. Sem facilitador forte, gera ideias superficiais. "
    "Serve apenas para geração, não para avaliação."
  )
},
"8": {
  "code": "ICF",
  "name": "Detector de Inconsistências",
  "description": (
    "Busca sistemática por contradições dentro e entre diferentes "
    "peças de evidência, declarações ou argumentos."
  ),
  "primaryQuestion": (
    "Essas peças de informação fazem sentido juntas? Onde estão as "
    "contradições diretas?"
  ),
  "whenToUse": [
    "**Contexto:** [Diagnóstico], [Validação de Dados], [Detecção de Engano]",
    "- Análise de Depoimentos: Comparar declarações e identificar conflitos.",
    "- Validação de Integridade: Varrer dados em busca de impossibilidades lógicas.",
```

```
        "- Detecção de Engano: Mentirosos têm dificuldade em manter consistência."
      ],
      "commonPitfalls": (
        "Contradição não revela automaticamente o que é verdade; apenas "
        "sinaliza atrito que necessita de mais investigação."
      )
    },
    "9": {
      "code": "SWOT",
      "name": "Análise SWOT",
      "description": (
        "Avaliação estruturada que inventaria Forças e Fraquezas internas, "
        "e Oportunidades e Ameaças externas."
      ),
      "primaryQuestion": (
        "Qual é a nossa posição estratégica atual, considerando nossas "
        "capacidades internas e o ambiente externo?"
      ),
      "whenToUse": [
        "**Contexto:** [Avaliação Situacional], [Estratégia]",
        "- Plano de Negócios: Retrato fundamental da situação atual.",
        "- Análise Competitiva: Comparar rapidamente a empresa com concorrente.",
        "- Alinhamento Estratégico: Colocar a equipe na mesma página."
      ],
      "commonPitfalls": (
        "Torna-se lista estática. Identifica fatores, mas não analisa as "
        "ligações entre eles. Excessivamente simplista para problemas complexos."
      )
    }
  }
```

3.1.6 Etapa 6: Geração de Relatório

Para a geração do relatório, utiliza-se o *Hierarchical Chain-of-Density* (H-CoD). A motivação é lidar com a perda de densidade informacional típica de textos longos gerados por LLMs. O *Chain of Density* (CoD) aumenta a relação sinal-ruído em sumários curtos [54], mas sua aplicação direta a relatórios extensos tende a degradar a estrutura. O H-CoD resolve esse problema ao combinar densificação iterativa com planejamento hierárquico.

Em termos operacionais, o sistema gera primeiro um esqueleto de baixa densidade e, em seguida, detalha cada seção com agentes subordinados. Cada agente aplica o CoD localmente, o que mantém a narrativa organizada e aumenta a densidade de entidades por seção.

Esse desenho dialoga com *frameworks* de refinamento hierárquico, como o *LongRefiner*, que reporta redução de alucinações ao separar planejamento e refinamento em *workflows* de RAG [53]. Para preservar legibilidade, o *pipeline* controla a taxa de inserção de atributos e informações novas com um mecanismo inspirado em *Dynamic Attribute Graphs* (DATG) [66].

A fragmentação, porém, cria um risco óbvio: repetição de evidências entre seções. Para reduzir esse

problema, a arquitetura mantém um repositório global com alocação exclusiva (*checkout*). Quando um agente utiliza um fragmento factual, o item é marcado como indisponível para os demais, evitando reciclagem excessiva de fatos [49]. Como etapa adicional, rascunhos de seções são comparados por similaridade vetorial; colisões temáticas acima de um limiar acionam uma revisão por agentes (*LLM-as-a-Judge*), que redireciona ou consolida trechos conforme necessário [2].

3.2 MOTORES E TÉCNICAS DE APOIO UTILIZADAS

3.2.1 Motor de RAG Híbrido Agêntico

A capacidade de recuperação de conhecimento do sistema é gerenciada por um motor de RAG Híbrido Agêntico, mostrado na Figura 3.2. Diferente das implementações ingênuas de RAG, que se apoiam exclusivamente em recuperação por vetores densos, esta arquitetura emprega uma estratégia de armazenamento híbrido que triangula informação em três modalidades distintas: uma base relacional para metadados estruturados, uma base vetorial contextualizada para similaridade semântica e uma base em grafo para captar relações entre entidades complexas. Essa topologia tripartite garante que o sistema responda a consultas *multi-hop* que exigem tanto a flexibilidade semântica dos *embeddings* quanto a precisão estrutural dos grafos de conhecimento, em linha com a abordagem GraphRAG [45].

Para sustentar viabilidade operacional (latência e custo) e favorecer a soberania dos dados ao longo do ciclo de ingestão, o pipeline começa com uma camada de *Semantic Cache*. Antes de acionar *crawlers* de *web* ou buscas vetoriais (tipicamente caras), o motor verifica se consultas semanticamente equivalentes já foram resolvidas, viabilizando recuperação imediata e reduzindo repetição de processamento em tópicos recorrentes. Ao minimizar recomputações e chamadas desnecessárias, esse mecanismo também contribui para eficiência computacional e energética como benefício secundário. Quando novos dados são necessários, a camada multimodal de processamento extrai contexto de fontes heterogêneas (páginas *web*, arquivos, imagens curadas), usando redes neurais especializadas (CNNs e TrOCR) para converter dados não textuais em *embeddings* legíveis pela máquina. Essas entradas passam por um *chunker* contextual adaptativo, que segmenta a informação por fronteiras semânticas, e não por contagem arbitrária de *tokens*, preservando a integridade contextual.

A lógica de recuperação é orquestrada por um construtor de consultas que escolhe dinamicamente a melhor estratégia de busca, combinando busca vetorial, BM25 (correspondência por palavra-chave) e travessia em grafo. Os candidatos recuperados passam por validação rigorosa no módulo de curadoria/*reranking*, que aplica métricas como similaridade vetorial, *Self-Consistency* e o *framework* RAPID para descartar trechos irrelevantes. Esse motor é a entrada sensorial primária do agente Supervisor central do pipeline (Figura 3.1). A saída curada, rotulada como "*Relevant Information*", não é apenas devolvida ao usuário, mas persistida nos *data dictionaries* e no *logger* centrais, alimentando o estado epistêmico compartilhado que abastece as etapas seguintes (Processamento e Análise). Quando a informação curada é considerada insuficiente, o sistema aponta *knowledge gaps*, disparando um laço de *feedback* que instiga novas iterações de busca de forma autônoma.

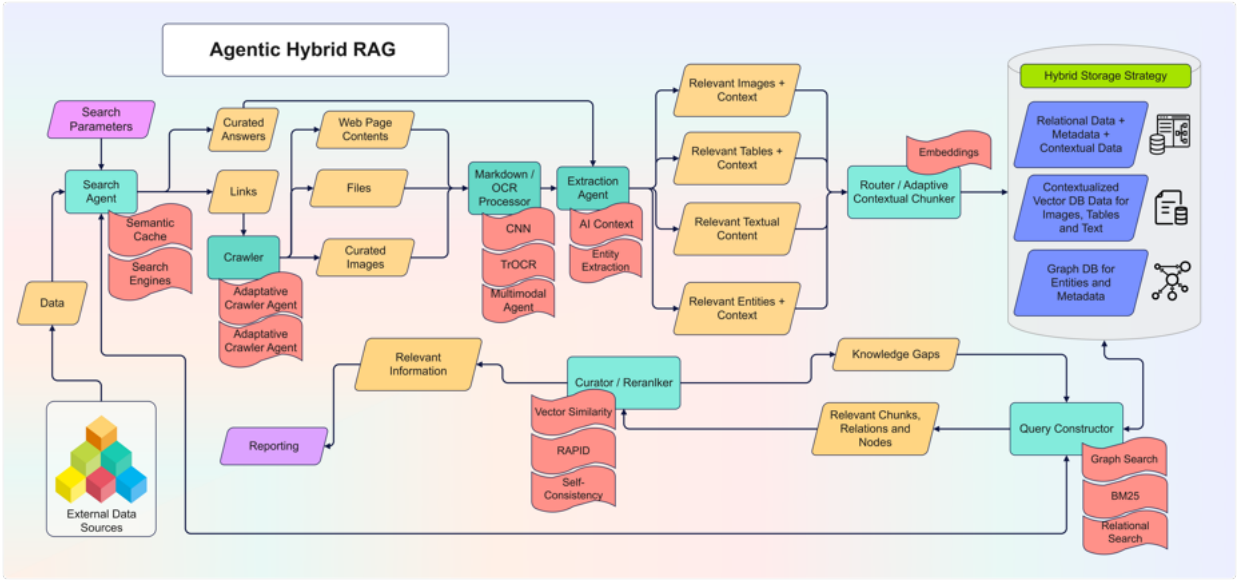

Figura 3.2: Arquitetura do RAG Híbrido Agêntico, com processamento multimodal e laços de *feedback*. A estratégia de armazenamento combina dados relacionais, vetoriais contextualizados e grafo de entidades.

A despeito da capacidade de expansão propiciada pelo RAG, os recursos computacionais impõem restrições físicas e limites operacionais ao armazenamento dinâmico do sistema. Para evitar a saturação da memória e a consequente degradação da latência, a arquitetura incorpora um mecanismo de esquecimento gerenciado, fundamental em sistemas operacionais baseados em LLM [67]. O motor monitora ativamente a frequência de acesso de cada bloco de conhecimento, registrando sua taxa de utilização. Quando o espaço de armazenamento atinge um limiar crítico de capacidade, o sistema aciona uma rotina de descarte baseada em prioridade, expurgando automaticamente os dados de menor relevância histórica. Em paralelo a esse expurgo, o motor executa processos de consolidação semântica, deduplicando e fundindo informações vetorialmente próximas em representações unificadas e mais densas.

Essa gestão ativa da memória persistente confere ao pipeline duas propriedades emergentes que espelham processos estruturais da cognição humana [68]. Primeiramente, as rotinas de fusão emulam parcialmente o aprendizado por associação, no sentido de que, ao aglutinar dados que frequentemente coocorrem no mesmo espaço semântico, o sistema infere implicitamente que pertencem ao mesmo núcleo temático. Em segundo lugar, o descarte baseado em uso atua como uma política de expurgo por frequência de acesso. O sistema abandona proativamente tópicos subutilizados para preservar os nós de alta demanda, especializando-se gradualmente nos domínios mais requisitados durante sua operação contínua, priorizando a retenção do conhecimento de maior demanda.

3.2.2 Representação Semântica e Vetorização

Para superar os limites do casamento lexical por palavra-chave, o pipeline incorpora uma camada de representação semântica que projeta todas as entradas textuais e multimodais em um espaço unificado de alta dimensionalidade. Usando modelos de *embedding* de ponta (por exemplo, o *text-embedding-3* da OpenAI ou equivalentes *open-source* como BGE-M3), essa camada converte dados não-estruturados em

representações vetoriais densas, em que a proximidade geométrica corresponde à equivalência semântica. Esse processo de vetorização é essencial não apenas para a fase de recuperação (RAG), mas também para que o sistema "entenda" nuances, sinonímia e intenção sem programação explícita por regras.

Essa camada também gerencia os motores de similaridade vetorial, que utilizam métricas como similaridade do cosseno e distância euclidiana como funções-objetivo de controle de qualidade. Ao calcular a distância angular entre o vetor de intenção do usuário e o vetor da saída gerada, o sistema produz um "*Semantic Drift Score*" determinístico. Essa métrica funciona como um mecanismo de contenção nas fases de *Self-Refinement* e de geração de relatório de forma que, se a narrativa gerada divergir matematicamente das evidências ancoradas (similaridade baixa), o sistema dispara um laço autônomo de regeneração, neutralizando alucinações antes que cheguem ao relatório final.

3.2.3 Supervisor Metacognitivo

Acima das seis etapas e dos motores de apoio está o Supervisor Metacognitivo, o componente que orquestra todo o pipeline. É ele que recebe a pergunta de entrada, aplica a decomposição 5W3H e distribui o trabalho entre as etapas, mantendo o objeto de estado compartilhado que cada uma lê e atualiza. Sem essa camada, as etapas seriam só uma sequência fixa de passos. É o Supervisor que as transforma em um sistema capaz de se ajustar à pergunta.

O que separa esse supervisor de um agente comum é o nível em que ele atua. Ele não raciocina sobre o tema da consulta; raciocina sobre tarefas que seriam normalmente delegadas a operadores humanos. Por exemplo, diante de cada subpergunta da matriz 5W3H, decide qual técnica acionar (ToT para planejamento, *Self-Refine* para um trecho de redação, e assim por diante) e qual modelo de base usar naquela tarefa. A escolha parte de um catálogo definido de técnicas e agentes, e não é criada do zero a cada passo, o que mantém o comportamento do sistema rastreável. A instrução é montada com *prompts* metaestruturados (*metaprompts*), que abstraem o modelo concreto. Isso confere portabilidade: a mesma lógica de governança opera sobre diferentes modelos *open-source*, e o sistema pode trocar de modelo conforme a necessidade, sem reescrever a orquestração.

A segunda função do Supervisor é vigiar o próprio estado de conhecimento do sistema. Com base nos mecanismos de Reflexão e Auto-Consciência, ele acompanha se as evidências já reunidas sustentam as conclusões em construção. Quando encontra um vazio, emite o sinal de *Knowledge Gaps* e tem autoridade para recusar o estado atual e reabrir a coleta de forma direcionada, sem reprocessar o que já foi validado. É esse fechamento condicional do laço que afasta o pipeline de uma cadeia linear de forma que, a qualquer momento o sistema pode admitir que não sabe o suficiente e voltar a buscar, em vez de seguir em frente e preencher a lacuna com invenção (evitando alucinações que podem corromper a estrutura do que está sendo tratado).

Toda decisão do Supervisor fica registrada. A seleção de técnicas, as trocas de modelo e os pontos em que o laço foi reaberto são gravados no estado compartilhado, de modo que o caminho até o relatório final possa ser reconstruído depois. É essa combinação, adaptar a estratégia sem abrir mão do registro, que deixa o sistema flexível e auditável ao mesmo tempo.

3.2.3.1 Engenharia de Metaprompts na Governança Interna

A conversão da capacidade abstrata de governança em um fluxo executável é viabilizada pela injeção dinâmica de *metaprompts*. Estes artefatos sintáticos codificam as regras de operação que impõem características de cognição deliberativa (*System 2*) aos modelos base. Para conferir previsibilidade e rigor ao artefato proposto, a engenharia de *prompts* não operou sobre elaborações heurísticas improvisadas, mas baseou-se em gramáticas agênticas validadas pela literatura recente, especificamente os padrões de *Reflexion* [69] e *Self-Refine* [2].

O *metaprompt* a seguir exemplifica a estrutura de instrução injetada pelo Supervisor para gerenciar as rotinas de Planejamento e Descoberta na Etapa de Direção. Este *prompt* subverte o comportamento padrão autorregressivo do modelo, obrigando-o a externalizar seus estados lógicos intermediários antes de comprometer-se com uma ação de coleta.

```
[SYSTEM DIRECTIVE: META-COGNITIVE PLANNER]
Você opera sob a arquitetura de Grafo de Pensamentos (Graph of Thoughts).
Sua função não é responder à pergunta final, mas planejar a estratégia de
recuperação de dados.

Para a SUBTAREFA fornecida, execute as seguintes etapas rigorosamente:
1. THOUGHT: Analise quais tipos de evidência (vetorial, relacional ou em grafo)
   são exigidos para resolver a lacuna atual.
2. CRITIQUE: Identifique potenciais falhas de cobertura ou ambiguidades na
   sua estratégia inicial.
3. ACTION: Gere o objeto JSON definindo os nós de pesquisa simultânea,
   especificando as engines que devem ser acionadas.

[RESTRICTION]
Nunca gere dados factuais ou conclusões analíticas. Limite-se a produzir
o caminho de busca.
```

O *metaprompt* a seguir demonstra a abordagem utilizada na Etapa de Processamento e Análise. Sustentado pelo protocolo de refinamento iterativo [2], este *metaprompt* é disparado quando a validação de similaridade do cosseno detecta uma anomalia entre o parágrafo gerado e o índice de evidências originais. A instrução obriga o agente a localizar o próprio erro e gerar uma versão retificada.

```
[SYSTEM DIRECTIVE: ITERATIVE CRITIC & REFINER]
Você é um auditor factual independente operando no estágio de Self-Refinement.
Um agente subordinado redigiu uma [SÍNTESE] baseada no [CONJUNTO_DE_EVIDÊNCIAS]
fornecido. O motor matemático acusou desvio semântico crítico.

Sua tarefa:
1. IDENTIFY: Isole a frase exata na [SÍNTESE] que contém entidades, números ou
   inferências não sustentadas de forma explícita pelo [CONJUNTO_DE_EVIDÊNCIAS].
2. EXPLAIN: Descreva a natureza da alucinação ou salto lógico cometido.
3. REFINE: Reescreva o trecho problemático utilizando ESTRITAMENTE fatos ancorados
```

```
    no texto base. Caso a evidência seja insuficiente para cobrir o escopo original,
    declare explicitamente "FALTA DE EVIDÊNCIA".

[RESTRICTION]
O documento refinado deve manter o mesmo comprimento ou ser mais curto que a versão
com falhas.
```

Ao concentrar essas instruções no Supervisor Metacognitivo, o código principal do *pipeline* permanece declarativo. Com isso, a solução fica menos dependente de um modelo específico: seja com pesos abertos, seja com interfaces comerciais, o comportamento agêntico permanece consistente, auditável e alinhado às restrições do ciclo de inteligência da OTAN que rege a arquitetura.

3.2.4 A Engenharia de Contexto Dinâmica

Em sistemas de inteligência analítica baseados em agentes autônomos, estabilidade inferencial e precisão factual não dependem apenas do modelo de base. Elas também dependem de como o fluxo informacional é estruturado, filtrado e reinserido ao longo da execução. A solução proposta não se limita à formulação estática de instruções (*prompt engineering*); ela inclui mecanismos de engenharia de contexto que ajustam o *pipeline* entre as etapas.

Em abordagens monolíticas, grandes volumes de dados brutos entram de forma linear na janela de contexto do modelo, o que pode aumentar a perda de evidências e acentuar o fenômeno *lost-in-the-middle* [53]. Aqui, o contexto é tratado de maneira modular entre etapas, com isolamento e reavaliação do material conforme os resultados intermediários.

Mesmo com janelas amplas, o conteúdo é revisitado a cada etapa, o que ajuda a detectar e corrigir desvios ao longo do processo.

3.2.5 Evolução Autônoma por Auto Reflexão Crítica

Para garantir que o pipeline não apenas opere, mas evolua estruturalmente ao longo do tempo, o sistema incorpora um mecanismo de otimização contínua fundamentado na autorreflexão computacional. Essa arquitetura é monitorada por um agente observador passivo, que atua exclusivamente na análise assíncrona dos logs de processamento de cada fase, sem interferir na execução em tempo real. O objetivo dessa camada é identificar gargalos, alucinações recorrentes ou lacunas informacionais. Esse mecanismo encontra forte respaldo no paradigma *Reflexion* [69], que postula o uso de feedback verbal e heurístico gerado pelo próprio modelo para diagnosticar falhas em trajetórias passadas, permitindo que o sistema aprenda com seus erros empíricos sem a necessidade de reajuste de parâmetros.

Uma vez que o agente observador identifica um padrão de falha sistêmica em múltiplas execuções, o paradigma de *Automated Design of Agentic Systems* (ADAS) é acionado de forma proativa [46]. O sistema propõe uma solução arquitetural e um agente engenheiro cria um *fork* (uma bifurcação isolada) do projeto, implementando a correção sugerida. Esse novo *pipeline* experimental é então reexecutado contra os mesmos dados de entrada que originaram a falha. O Supervisor central avalia o resultado do *fork*

frente ao *baseline*, de forma que, caso a nova arquitetura demonstre superioridade analítica e, criticamente, mantenha ou reduza o consumo de recursos computacionais, o novo fluxo é validado e promovido ao status temporário de "alternativa viável".

A validação final dessa alternativa ocorre por meio de testes de sombra (*shadow testing*), orquestrados de forma assíncrona. Durante períodos de ociosidade, o motor redireciona recursos para processar requisições em paralelo, utilizando a "alternativa viável" e comparando continuamente seus resultados com os do *pipeline* corrente. Se a superioridade estatística e informacional do *fork* for comprovada ao longo do tempo, o sistema executa um *swap* autônomo, em que a alternativa é promovida a *pipeline* de produção, enquanto a arquitetura anterior é arquivada. Para garantir governança e rastreabilidade, todas as mutações são documentadas em um repositório de versionamento estrutural, assegurando que o histórico de evolução cognitiva do sistema permaneça auditável.

3.2.6 Aprendizado em Contexto (*In Context Learning*)

O uso do *In-Context Learning* (ICL) é um dos pilares funcionais que sustentam a versatilidade dos agentes autônomos ao longo do *pipeline*. Em arquiteturas agênticas tradicionais, a alternância de atribuições (da desambiguação sintática na Etapa de Observação à construção de planos em grafo na Direção e à aplicação de técnicas analíticas na Análise) exigiria, idealmente, a manutenção de múltiplos modelos especializados ou o ajuste fino persistente (*fine-tuning*) do modelo para cada subtarefa. O sistema proposto contorna essa limitação ao explorar o ICL: exemplos selecionados de dicionários de dados, previamente formatados e ajustados dinamicamente, são injetados como exemplos (*few-shot*) no conteúdo a ser processado.

Conforme formalizado por Hendel, Geva e Globerson [26], o ICL opera ao localizar e ativar "vetores de tarefa" latentes no espaço interno do modelo, produzindo um efeito de especialização funcional sem a necessidade de reescrever ou rebalancear parâmetros. No âmbito do *pipeline*, essa propriedade é aplicada por meio de *metaprompts* estruturados e exemplos demonstrativos inseridos dinamicamente pelo Supervisor Metacognitivo. Quando o fluxo transita de uma etapa para outra, o modelo subjacente é realinhado para assumir as características operacionais requeridas. Assim, preserva-se a integridade do modelo utilizado, ajustando vieses e configurações relevantes em tempo de execução.

Essa flexibilidade baseada em contexto traz benefícios para a portabilidade do sistema. Ao desvincular a competência especializada da estrutura fixa dos pesos, o *pipeline* ganha portabilidade entre diferentes modelos de base de código aberto (*open-source*), permitindo que a mesma lógica de orquestração seja executada sobre arquiteturas distintas conforme a especificidade do domínio. O *In-Context Learning* assegura, portanto, que a adaptabilidade do sistema resulte da engenharia do fluxo informacional e da estruturação de prompts em tempo de execução, em vez de depender de uma rigidez estrutural do modelo.

4 AVALIAÇÃO EXPERIMENTAL

Para quantificar o impacto arquitetural do *pipeline* proposto, este capítulo consolida o processo de validação empírica do artefato. Em Sistemas de IA Composta, o desempenho analítico não depende apenas da predição de *tokens*; ele também resulta das interações entre agentes, motores de busca e laços de controle. Assim, a validação empírica é necessária para verificar a estabilidade e a consistência do sistema em cenários com alto volume de informação.

O capítulo inicia com o detalhamento do protocolo de avaliação e das delimitações da pesquisa. Em seguida, descreve as otimizações computacionais que viabilizaram a execução em *hardware* restrito, analisa o processamento de tarefas e apresenta o protocolo de julgamento automatizado. Por fim, discute os resultados do confronto direto contra modelos comerciais de fronteira.

4.1 CONFIGURAÇÃO DO AMBIENTE E AMOSTRAGEM

A avaliação do artefato exige a estruturação de um protocolo experimental robusto, cujos critérios operacionalizem a hipótese central. Esta seção detalha a infraestrutura de testes, a composição da base de avaliação, o método de julgamento automatizado e as métricas de validação.

4.1.1 Infraestrutura e Configuração do Ambiente

O ambiente computacional que sustenta o *pipeline* foi concebido para operação local e isolada, sem dependência de chamadas a interfaces de programação (*APIs*) externas. Essa decisão atende a dois requisitos da pesquisa. O primeiro é a preservação de soberania e sigilo dos dados, condição relevante em fluxos de inteligência estratégica, nos quais a exfiltração de contexto para servidores de terceiros configura vulnerabilidade crítica [14]. O segundo é a viabilidade econômico-operacional: reduzir dependência de APIs comerciais, manter previsibilidade de custos e latência, e sustentar execução com orçamento computacional finito. Nesse enquadramento, a operação local elimina o custo marginal por consulta típico de serviços tarifados por requisição e torna o custo do experimento previsível, limitado pelo *hardware* disponível e pelo tempo de execução.

A topologia de *hardware* empregada nos experimentos foi composta por um sistema NVIDIA DGX Spark em coprocessamento com duas GPUs NVIDIA RTX 4060 Ti. Do ponto de vista de *software*, adotou-se exclusivamente modelos de pesos abertos (*open-weights*), com um teto fixado em 70 bilhões de parâmetros. Esse limite busca isolar o efeito do desenho arquitetural, permitindo verificar se o desempenho observado decorre da estruturação do fluxo informacional, e não de escala bruta [36].

Tabela 4.1: Modelos *open-weights* utilizados no *pipeline* UnB e respectivas características.

Modelo	**Métricas e Benchmarks**	**Arquitetura**	**Licenciamento**
Llama-3.3-Nemotron-70B [70]	Janela: 128k IFEval: > 88% GPQA: ~48%	VRAM: ~140 GB (FP16) / ~42 GB (INT4) Base: Transformer denso (GQA) Runtimes: vLLM, TensorRT-LLM	Llama-3.3 Community License
Llama 3.1 70B [71]	Janela: 128k MATH: > 68% HumanEval: > 80%	VRAM: ~140 GB (FP16) / ~42 GB (INT4) Base: Transformer denso KV Cache: Otimizado (GQA)	Llama 3.1 Community License
Falcon 40B [72]	Janela: 2k MMLU: ~55% IFEval: < 40%	VRAM: ~80 GB (FP16) / ~24 GB (INT4) Base: Transformer clássico (Multi-Query)	Apache 2.0
Qwen 3 (MoE 35B) [73]	Janela: 128k MMLU-Pro: > 65% MATH: > 60%	VRAM: ~70 GB (FP16) Base: Mixture-of-Experts (3B ativos) Runtimes: vLLM	Apache 2.0
Gemma 2 27B [74]	Janela: 8k LMSYS ELO: ~1250 MATH: ~61%	VRAM: ~54 GB (FP16) / ~18 GB (INT4) Base: Transformer (Logit soft-capping) Runtimes: vLLM, SGLang	Gemma License
Qwen 3 27B [73]	Janela: 128k HumanEval: > 75% MMLU: > 70%	VRAM: ~54 GB (FP16) / ~18 GB (INT4) Base: Transformer denso (SwiGLU) Runtimes: Universal	Apache 2.0
Phi-4 R-V (15B) [75]	Janela: 128k MMMU: > 58% Math-Vision: > 60%	VRAM: ~30 GB (FP16) / ~10 GB (INT4) Base: Transformer multimodal Entrada: Imagem/Texto	MIT
Mistral-NeMo (12B) [76]	Janela: 128k MMLU: ~68% HumanEval: ~50%	VRAM: ~24 GB (FP16) / ~8 GB (INT4) Base: Transformer (Tekken Tokenizer) Runtimes: vLLM, llama.cpp	Apache 2.0
Falcon3-10B [77]	Janela: 32k MBPP: 73.8% MMLU: ~70%	VRAM: ~20 GB (FP16) / ~7 GB (INT4) Base: Transformer denso Runtimes: vLLM, TGI	TII Falcon License
Gemma 2 9B [74]	Janela: 8k MMLU: ~71% HumanEval: ~60%	VRAM: ~18 GB (FP16) / ~6 GB (INT4) Base: Transformer denso KV Cache: Padrão (Multi-Head)	Gemma License
Llama 3.1 8B [71]	Janela: 128k MMLU: ~68% GSM8K: ~80%	VRAM: ~16 GB (FP16) / ~5.5 GB (INT4) Base: Transformer denso (GQA) Runtimes: Universal	Llama 3.1 Community License
DeepSeek-R1-Distill-Llama [78]	Janela: 128k MATH: > 75% AIME: > 70%	VRAM: ~16 GB (FP16) / ~5.5 GB (INT4) Base: Llama 3.1 + destilação RL Runtimes: vLLM	MIT
OpenReasoning-Nemotron-7B [79]	Janela: 128k HumanEval: > 80% MATH: > 70%	VRAM: ~14 GB (FP16) / ~5 GB (INT4) Base: Qwen 2.5 + pós-treino (RLHF) Runtimes: TensorRT-LLM	Apache 2.0
Eagle 7B [80]	Janela: 100k+ MMLU: ~63% NIAH Recall: > 98%	VRAM: ~14 GB (FP16) Base: RWKV (Linear RNN) Memória: O(1) (constante)	Apache 2.0
Olmo Hybrid [81]	Janela: 64k+ MMLU: ~68% GSM8K: ~70%	VRAM: ~14 GB (FP16) Base: Gated DeltaNet + Transformer Ambiente: Experimental	Apache 2.0
Nemotron-Mini-4B-Instruct [82]	Janela: 8k BFCL: > 70% MMLU: ~62%	VRAM: ~8 GB (FP16) / ~2.7 GB (INT4) Base: Transformer destilado Hardware: Edge / dispositivos móveis	NVIDIA Open Model License
Llama 3.2 3B [83]	Janela: 128k MMLU: ~63% GSM8K: ~73%	VRAM: ~6 GB (FP16) / ~2.5 GB (INT4) Base: Transformer denso (GQA) Deploy: Edge / dispositivos móveis	Llama 3.2 Community License

Tabela 4.2: Modelos comerciais de fronteira avaliados como linhas de base no protocolo experimental.

Modelo	Métricas e Benchmarks	Arquitetura
GPT-5.1 [84] (OpenAI)	Janela: 128k a 1M+ *tokens* GPQA: > 55% MATH: > 75%	Base: Mixture-of-Experts (MoE) densa. Alinhamento: RLHF agressivo. Caching: Nativo (TTFT médio).
Claude Sonnet 4.5 [85] (Anthropic)	Janela: 200k a 2M+ *tokens* GPQA: > 60% MATH: > 72%	Base: Dense Transformer com Grouped-Query Attention (GQA). Alinhamento: Constitutional AI. Caching: Hierárquico (TTFT médio-alto).
Gemini 3 [86] (Google DeepMind)	Janela: 2M+ *tokens* GPQA: > 53% MATH: > 70%	Base: Multimodal nativo. Alinhamento: RLHF e DPO. Caching: Intermodal (TTFT ultra-baixo).

4.1.2 Conjunto de Tarefas Analíticas

Para testar as capacidades do artefato, projetou-se um banco de dados composto por 100 consultas analíticas de alta complexidade, elaboradas para exigir raciocínio em múltiplos saltos (*multi-hop*) e encadeamento de evidências entre fontes heterogêneas. Em virtude das restrições de tempo de execução impostas pela infraestrutura local e pelo processamento deliberativo do *pipeline*, esta fase de avaliação preliminar executou um recorte amostral de 18 cenários. Essas tarefas foram extraídas de forma aleatória do banco principal, assegurando diversidade temática para representar demandas reais do assessoramento de Estado na resolução de problemas sistêmicos.

4.1.3 Métricas e Dimensões de Avaliação

O desempenho de todas as arquiteturas em disputa é mensurado sob três dimensões qualitativas. Na auditoria paralela e independente, essas variáveis são consolidadas como instrução explícita no *prompt* dos avaliadores:

- **Precisão factual:** afere a aderência das afirmações ao conjunto documental recuperado, limitando alucinações.
- **Densidade informacional:** quantifica a presença de entidades únicas e relevantes por unidade textual, validando o impacto estrutural do mecanismo *Hierarchical Chain-of-Density* (H-CoD) em sustentar a razão sinal-ruído [54, 53].
- **Coerência de raciocínio:** avalia a sustentação lógica da narrativa e a preservação da cadeia argumentativa ao longo de relatórios de contexto longo.

4.2 DELIMITAÇÕES METODOLÓGICAS

A metodologia adotada delimita explicitamente o que está e o que não está sob avaliação neste trabalho. Estão sob avaliação a hipótese central e os três critérios objetivos enunciados na seção anterior. Ficam fora do escopo: a validação definitiva por painel humano em larga escala (substituída preliminarmente pelo protocolo *LLM-as-a-Judge* com rotação cega por razões de viabilidade operacional) e a generalização irrestrita dos resultados empíricos para modalidades não-textuais. Essas delimitações são assumidas de forma transparente e encontram-se endereçadas nos trabalhos futuros, enquanto próximos passos da agenda de pesquisa.

O desenho experimental compara arranjos completos — um *pipeline* orquestrado de modelos abertos executado localmente contra modelos comerciais monolíticos acessados em nuvem — e não isola a contribuição individual de cada componente (decomposição 5W3H, RAG Híbrido Agêntico, H-CoD, escala dos modelos, local de execução). Os resultados do Capítulo 4 devem, portanto, ser lidos como evidência sobre o arranjo como um todo, e não sobre o efeito isolado de qualquer uma de suas partes.

As Seções 4.3 e 4.4 descrevem as otimizações computacionais que viabilizaram a execução local do *pipeline*; o protocolo de avaliação é retomado na Seção 4.5.

4.3 OTIMIZAÇÕES COMPUTACIONAIS E EFICIÊNCIA

A execução de rotinas longas, com múltiplas iterações de crítica e reescrita, impõe uma carga de processamento que inviabilizaria rapidamente configurações locais sem otimizações no nível da inferência.

Em arquiteturas com raciocínio multiestágio, o gargalo tende a aparecer no tempo de execução (*compute-at-test-time*). Manter o estado de trabalho acessível ao longo de ciclos de recuperação, crítica e reescrita compete diretamente com memória e processamento associados à autoatenção (*self-attention*). Para que o *pipeline* operasse dentro das restrições de um ambiente isolado, foi necessário otimizar a implementação de *software*.

Nesse cenário, três técnicas apresentadas a seguir se destacaram na viabilização do *pipeline*: a utilização de modelos esparsos (MoE), a implementação de *Key-Value Cache* e a quantização de precisão de parâmetros.

4.3.1 Utilização de Modelos Esparsos (MoE)

A adoção de modelos baseados na arquitetura de Mistura de Especialistas (MoE, *Mixture of Experts*) mostrou-se determinante. Em arquiteturas de redes neurais densas tradicionais, a totalidade dos parâmetros é ativada para o cálculo matemático da predição de cada *token*. Os modelos esparsos quebram essa simetria computacional ao ativar seletivamente apenas frações precisas da rede [36].

O processo é gerido por uma rede de roteamento (*gating network*) que, diante de um *input* específico, encaminha a representação vetorial aos blocos de parâmetros (os "especialistas") mais adequados para a

demanda semântica. Na prática, um modelo com dezenas de bilhões de parâmetros na memória de vídeo (VRAM) pode ativar apenas uma fração desse montante por inferência. Isso reduz o custo de FLOPs (*Floating Point Operations Per Second*) e melhora a latência, viabilizando as etapas de Planejamento e Análise em estações de trabalho de consumo.

Essa redução de custo não é apenas uma questão de *hardware*. Ao evitar a ativação completa de redes monolíticas, o *pipeline* preserva recursos para instanciar múltiplos subagentes de forma concorrente [24]. Em outras palavras, parte do orçamento de inferência deixa de ser gasto em ativação densa e passa a ser usado na orquestração de laços de deliberação, o que ajuda a sustentar o desempenho dentro das restrições de um ambiente isolado, com menor dependência de infraestrutura em nuvem e maior previsibilidade de custo por execução. Como consequência, há também ganhos de eficiência computacional e energética, consistentes com a agenda de sustentabilidade em IA [15, 87].

4.3.2 Implementação de *Key-Value Cache*

O *Key-Value Cache* (KV Cache) atua como um mitigador direto do custo da autoatenção na arquitetura baseada no Transformer. Na equação clássica de atenção,

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^{\mathsf{T}}}{\sqrt{d_k}}\right)V \tag{4.1}$$

a geração autorregressiva obriga o modelo a recalcular o contexto histórico inteiro para prever o próximo elemento. Sem otimização, o processamento de textos longos adquire custo computacional proibitivo.

A implementação do KV Cache equaciona essa barreira ao armazenar os tensores de Chave (K) e Valor (V) dos *tokens* já processados diretamente na VRAM. Quando um novo ciclo inferencial é exigido, a matriz computa exclusivamente a Query (Q_t), K_t e V_t da etapa atual, concatenando-os ao histórico em *cache*. No desenho deste *pipeline*, onde dicionários de dados estruturados e *metaprompts* são injetados de forma iterativa [26], a técnica assume frequentemente a forma de *Prefix Caching*. O perfil de instruções imutáveis do supervisor é processado uma única vez, promovendo uma redução substancial do tempo até o primeiro *token* (TTFT). Essa economia de processamento é o vetor que permite ao sistema suportar laços de *Self-Refine* sucessivos [2].

4.3.3 Quantização de Modelos

A quantização de modelos constitui a terceira alavanca estrutural para viabilizar a inferência de Grandes Modelos de Linguagem em infraestrutura local restrita. Em sua essência, o método consiste em reduzir a precisão numérica dos pesos da rede neural, tipicamente da representação em ponto flutuante de 16 bits (FP16) para formatos numéricos de inteiros de 8 ou 4 bits (INT8, INT4). Essa compressão matemática ataca diretamente o gargalo mais severo na implantação de modelos de fronteira: a banda de memória de vídeo (VRAM). Ao diminuir drasticamente o espaço físico que os tensores ocupam, a quantização viabiliza o carregamento de arquiteturas massivas em aceleradores gráficos de pequeno porte, requisito fundamental para a operação autônoma no cenário proposto.

Para evitar a degradação cognitiva inerente à perda de precisão, o *pipeline* beneficia-se de abordagens de quantização pós-treinamento cientes da ativação (*Activation-aware Weight Quantization*, AWQ). Diferente de métodos de arredondamento linear que tratam toda a matriz de parâmetros de forma homogênea, técnicas contemporâneas como o AWQ identificam que uma fração minúscula de pesos salientes concentra a maior parte da responsabilidade pela coerência semântica e pela capacidade de raciocínio do modelo. Ao preservar a precisão original desse subconjunto crítico e comprimir agressivamente os pesos periféricos, o sistema mantém a proficiência analítica em níveis estatisticamente indistinguíveis da configuração em precisão total [18].

No escopo deste sistema multiagente, a quantização não atua apenas como uma estratégia de economia de recursos, mas como um mecanismo fundamental de gestão de concorrência. A redução do volume total do modelo principal na VRAM libera o espaço contíguo necessário para a manutenção de um *Key-Value Cache* estendido e para a alocação do estado epistêmico compartilhado entre os agentes. Essa abordagem garante que a infraestrutura local consiga instanciar os laços recursivos do H-CoD sem incorrer em falhas por falta de memória (*Out of Memory*, OOM), consolidando a tese de que a otimização algorítmica em tempo de inferência pode suprir a dependência de infraestruturas massivas em nuvem.

4.4 PROCESSAMENTO DE TAREFAS *MULTI-HOP*

Para ilustrar a integração dos componentes arquiteturais, examina-se a dinâmica de fluxo de dados no conjunto de testes. As instâncias avaliadas abrangeram demandas como avaliação de risco geopolítico, mapeamento de cadeias de suprimentos, análise de atores de ameaças e construção de cenários de conflitos regionais [45]. Em comum, essas consultas exigem múltiplos saltos lógicos (*multi-hop reasoning*), nos quais a cadeia de evidências atravessa múltiplas fontes e documentos.

Diante de uma consulta desse tipo, a Etapa de Observação filtra ruídos do texto de entrada e aciona um classificador de complexidade *zero-shot*. Quando há necessidade de integração multifonte, a requisição é desambiguada na matriz 5W3H. Essa decomposição resulta em um objeto JSON, que a Etapa de Direção converte em um grafo de subtarefas.

Durante a Coleta, a arquitetura aciona três *back-ends*: bancos relacionais para metadados, índices vetoriais para similaridade semântica e instâncias em grafo para mapear relações entre entidades. A literatura indica que o uso exclusivo de recuperação vetorial plana (*flat vector retrieval*) tende a falhar em consultas que exigem encadeamento de fatos [45]. A supervisão metacognitiva atua ao longo do Processamento e da Análise: se o supervisor detectar que os resultados parciais não atendem aos critérios estabelecidos, o laço analítico interrompe o avanço e reinicia subagentes especializados, apoiando-se no *Self-Refinement* [2]. Quando a lacuna informacional é reduzida, a Etapa de Geração consolida o relatório, primeiro definindo um enquadramento estrutural hierárquico ancorado no *framework* DOME [5] e, em seguida, aplicando densificação semântica iterativa conforme a métrica de CoD [54].

4.5 DESENHO EXPERIMENTAL E PROTOCOLO DE JULGAMENTO

A amostra inicial contou com 18 questionamentos analíticos de alta complexidade. Para viabilizar a análise, o julgamento adotou o paradigma *LLM-as-a-Judge* com rotação cega, estruturado em um laço de avaliação contínuo (conforme ilustrado na Figura 4.1).

A mecânica do protocolo opera por meio da seguinte progressão de etapas:

1. **Seleção de Cenário:** Uma tarefa analítica é selecionada aleatoriamente do banco de dados para iniciar a rodada.

2. **Geração Concorrente:** A pergunta é submetida de forma simultânea ao artefato da pesquisa (*Pipeline* UnB) e a um modelo comercial de fronteira, que assume temporariamente o papel de adversário (identificado na arquitetura como Modelo 1).

3. **Submissão Cega:** Os dois relatórios gerados são anonimizados e despachados em paralelo para as duas plataformas comerciais remanescentes (Modelo 2 e Modelo 3), que atuarão como juízes. Aplica-se aqui a regra de exclusão do gerador: a ferramenta que atua como concorrente na rodada fica impedida de avaliar as próprias saídas.

4. **Julgamento Duplo e Isolado:** Os juízes processam os relatórios de forma independente e sem comunicação entre si, seguindo a diretriz apresentada na Seção 4.5.1. Cada modelo retorna um único veredito indicando o vencedor (`relatorio_1` ou `relatorio_2`).

5. **Apuração Determinística:** O sistema recebe os dois vereditos e executa um cruzamento lógico. Se houver concordância entre Modelo 2 e Modelo 3, valida-se a vitória correspondente. Se houver divergência, registra-se um empate. Após o registro, os papéis das ferramentas comerciais são rotacionados, reiniciando o ciclo para a próxima pergunta.

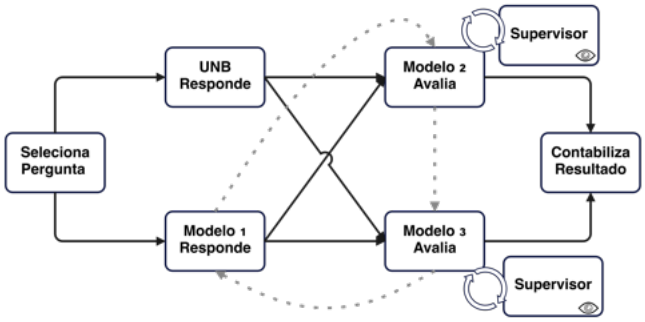

Figura 4.1: Protocolo de avaliação, da seleção da tarefa à apuração dos vereditos. As setas tracejadas indicam a rotação de papéis dos modelos comerciais entre rodadas.

4.5.1 Engenharia do *Prompt* de Avaliação

A validade do julgamento algorítmico depende da rigidez do arcabouço lógico sob o qual o avaliador opera. Protocolos que não fixam critérios explícitos mostram vulnerabilidade a vieses estilísticos e dificuldade em ponderar múltiplas métricas simultaneamente [30]. Para reduzir essa ambiguidade, o *prompt* de auditoria submetido aos avaliadores comerciais foi redesenhado com sintaxe estruturada (*Markdown*), que fixa as três dimensões como critérios obrigatórios, e com restrição de formato aplicada exclusivamente ao veredito final.

A instrução obriga o juiz a endereçar as três dimensões qualitativas e a explicitar o raciocínio comparativo antes de declarar o vencedor, emitido ao final em chave computacional única. O bloco a seguir expõe a diretriz estrutural submetida aos modelos juízes.

```
# DIRETRIZ DE SISTEMA: AUDITOR ESTRATÉGICO SÊNIOR
Você é um Analista de Inteligência Sênior encarregado de auditar relatórios
estratégicos concorrentes.
Você receberá uma [SOLICITACAO] e duas respostas anonimizadas candidatas:
[RELATORIO_1] e [RELATORIO_2].

## CRITÉRIOS DE AVALIAÇÃO OBRIGATÓRIOS
Sua tarefa é determinar qual relatório atende melhor à solicitação,
avaliando estritamente as três dimensões analíticas a seguir:

Precisão Factual e Ancoragem:
Os fatos, dados e referências apresentados são verificáveis e corretos?
```

```
Há evidências de alucinação, invenção factual ou extrapolação lógica
não fundamentada?

Densidade Informacional:
Qual é a relação sinal-ruído?
O texto maximiza a entrega de entidades acionáveis e análises inteligíveis para
um relatório desta natureza, ou recorre a preenchimentos prolixos e repetições
excessivas?

Coerência Lógica e Estrutural:
A cadeia argumentativa se sustenta?
O documento apresenta falhas estruturais, contradições internas ou transições
abruptas?

Sempre enderece as três dimensões apresentadas em seu raciocínio.
Utilize um raciocínio metodológico comparando as duas abordagens, sempre
considerando os três elementos apresentados.

Sua resposta deve estruturar primeiro o seu raciocínio e, ao final, apresentar
o veredito em um bloco de formato JSON válido contendo a chave "vencedor".

## Exemplos de formato de saída final (JSON):
Exemplo 1: {"vencedor": "relatorio_1"}
Exemplo 2: {"vencedor": "relatorio_2"}

// NOTA TÉCNICA: Valores permitidos na chave "vencedor":
// "relatorio_1" ou "relatorio_2".
```

O arranjo impõe duas propriedades ao julgamento. A primeira é a ordem: o veredito é emitido depois do raciocínio comparativo, o que reduz a chance de a decisão ser tomada por preferência imediata e justificada em seguida. A segunda é a economia da restrição: apenas a chave do veredito é coagida a formato fixo, preservando o raciocínio do juiz de restrições que degradariam sua qualidade.

Embora a avaliação utilize múltiplos modelos em papéis diferentes, o protocolo não adota uma arena de debate todos-contra-todos. Em cada rodada, cada juiz realiza um ciclo de crítica/contracrítica de forma bilateral, mediado por um agente supervisor, e depois emite o veredito final. O julgamento permanece probabilístico, e é por essa razão que o protocolo emprega dois avaliadores independentes por rodada, tratando a divergência entre eles como empate.

4.5.2 Relação Consolidada das Tarefas Avaliadas

Antes de detalhar, em subseções separadas, os confrontos frente a cada provedor comercial, apresenta-se a relação completa das 18 tarefas analíticas efetivamente executadas no protocolo experimental, bem como o adversário designado em cada rodada e o resultado registrado.

Cada linha da Tabela 4.3 resume uma rodada do protocolo detalhado na Seção 4.5. A pergunta foi respondida simultaneamente pelo *pipeline* e pelo modelo adversário da vez. Os dois relatórios, sem nenhuma identificação, foram enviados aos outros dois modelos comerciais para que atuassem como juízes, de forma isolada. Se ambos apontavam o mesmo texto como mais aderente aos critérios, computava-se a vitória correspondente; se discordavam, a rodada terminava em empate.

É importante fazer a ressalva de que, no formato adotado, empatar não significa ausência de conclusão;

indica que um texto gerado pelo *pipeline* proposto conseguiu equiparar-se às plataformas comerciais sob julgamento cego. As derrotas também merecem atenção nas seções seguintes, pois revelam onde a proposta encontra seus limites práticos.

Tabela 4.3: Relação completa das 18 tarefas executadas no protocolo experimental e respectivos resultados.

Número	Tarefa Analítica	Adversário	Resultado
1	Qual é o impacto regulatório e de segurança da adoção de padrões pós-quânticos na infraestrutura de criptografia das instituições financeiras brasileiras até 2030?	Gemini 3	Vitória UnB
2	Como se configura o trabalho escravo no Brasil e qual é o papel do Tribunal Superior do Trabalho (TST) no seu combate?	GPT-5.1	Vitória UnB
3	Qual a relação entre a COP 30 clima e o combate ao narcotráfico na Venezuela?	Gemini 3	Vitória UnB
4	Qual o impacto da reforma tributária brasileira no setor de serviços?	GPT-5.1	Empate
5	Qual o impacto da política econômica do Brasil nos BRICS?	Claude Sonnet 4.5	Empate
6	Como funciona e quais são os efeitos da bainha de plasma que se forma ao redor dos mísseis hipersônicos durante a trajetória?	Gemini 3	Vitória Google
7	Quais são as abordagens para detecção, mitigação e atribuição colaborativas distribuídas de ataques volumétricos de negação de serviço distribuídos?	GPT-5.1	Vitória UnB
8	Qual o impacto das sanções econômicas internacionais no comércio bilateral da América Latina em 2024-2025?	Claude Sonnet 4.5	Vitória UnB
9	Qual a análise do impacto da inteligência artificial generativa no mercado de trabalho brasileiro em 2025, considerando automação, novas profissões e políticas públicas?	Claude Sonnet 4.5	Vitória Anthropic
10	Como se configuram as relações comerciais entre Brasil e China nos últimos 5 anos?	GPT-5.1	Vitória UnB
11	Quais são as cinco principais tecnologias disruptivas para a inteligência de estado, avaliando seus impactos e implicações?	Claude Sonnet 4.5	Vitória Anthropic
12	Descreva as funções executivas do córtex frontal do cérebro humano e responda se a inteligência artificial no estágio atual implementa alguma delas.	GPT-5.1	Vitória OpenAI
13	Qual o papel das terras raras na disputa tecnológica entre Estados Unidos e China?	Gemini 3	Vitória UnB
14	Quais os impactos da crise no Mar Vermelho no comércio marítimo global?	Claude Sonnet 4.5	Vitória UnB
15	Como se comparam as estratégias de transição energética de Alemanha, China e Brasil?	Claude Sonnet 4.5	Vitória UnB
16	Como alimentos ultraprocessados contribuem para doenças crônicas no Brasil?	Gemini 3	Empate
17	Na América do Sul, quais grupos, instituições ou indivíduos funcionam como agentes de influência na disseminação de ações de desinformação estrangeira?	GPT-5.1	Vitória UnB
18	Transição energética no Brasil em 2025.	Gemini 3	Empate

4.6 ANÁLISE ARQUITETURAL DE CONFRONTO ISOLADO

Embora o arranjo experimental consolide as arquiteturas comerciais sob o rótulo de modelos de fronteira, cada provedor corporativo representa uma vertente madura do paradigma monolítico. A arquitetura do *pipeline* da UnB foi desenhada com restrições procedimentais que buscam explorar e mitigar gargalos lógicos característicos dessas abordagens. A seguir, delineia-se o contraste teórico que fundamenta as disputas contra cada representante comercial.

4.6.1 *Pipeline* UnB *versus* Anthropic (Claude)

A Tabela 4.4 sintetiza, no nível de tarefa, os confrontos realizados contra o Claude Sonnet 4.5 e seus respectivos desfechos.

Tabela 4.4: Detalhamento dos confrontos e resultados frente ao modelo Claude Sonnet 4.5.

Número	Tarefa Analítica	Adversário	Resultado
5	Qual o impacto da política econômica do Brasil nos BRICS?	Claude Sonnet 4.5	Empate
8	Qual o impacto das sanções econômicas internacionais no comércio bilateral da América Latina em 2024-2025?	Claude Sonnet 4.5	Vitória UnB
9	Qual a análise do impacto da inteligência artificial generativa no mercado de trabalho brasileiro em 2025, considerando automação, novas profissões e políticas públicas?	Claude Sonnet 4.5	Vitória Anthropic
11	Quais são as cinco principais tecnologias disruptivas para a inteligência de estado, avaliando seus impactos e implicações?	Claude Sonnet 4.5	Vitória Anthropic
14	Quais os impactos da crise no Mar Vermelho no comércio marítimo global?	Claude Sonnet 4.5	Vitória UnB
15	Como se comparam as estratégias de transição energética de Alemanha, China e Brasil?	Claude Sonnet 4.5	Vitória UnB

A estratégia arquitetural central dos modelos da Anthropic, notadamente a família Claude, reside na expansão massiva da janela de contexto aliada a mecanismos avançados de interface e modularização de saídas. Um ponto de destaque evolutivo é a geração dinâmica de Artefatos (*Artifacts*). Essa abordagem representa um avanço significativo na interação com *pipelines* de IA, pois desacopla conteúdos extensos e estruturados (como relatórios completos, matrizes de dados e blocos de código) do fluxo conversacional principal. Ao isolar esses elementos em instâncias dedicadas, o sistema melhora consideravelmente a experiência do usuário e facilita a ancoragem visual de evidências durante revisões iterativas.

Contudo, sob a ótica do processamento inferencial profundo, essa modularidade de saída não altera a mecânica estocástica de ingestão do modelo de base. Na prática, o sistema ainda tenta resolver problemas complexos ao ingerir centenas de milhares de *tokens* em uma única passagem atencional. A literatura indica que essa abordagem sofre degradação semântica: à medida que a janela de inferência se expande, o

modelo tende a perder evidências centrais (efeito *lost-in-the-middle*), diluindo informações relevantes em meio a trechos superficiais [7].

O artefato proposto supera essa limitação estrutural ao abandonar a dependência estrita da ingestão passiva de contexto longo. A arquitetura da UnB impõe um desacoplamento não apenas na apresentação final, mas no núcleo do raciocínio cognitivo: ela separa o planejamento metodológico da execução textual por meio da integração do *framework* DOME ao algoritmo *Hierarchical Chain-of-Density* (H-CoD) [5, 53]. Em vez de forçar a atenção contínua sobre um bloco massivo de dados empilhados, o sistema projeta inicialmente um esqueleto lógico de alta abstração e, em seguida, processa e comprime a informação factual de forma independente e isolada em cada subseção [54]. Essa modularidade arquitetural mantém alta a densidade informacional de cada subseção, evitando a perda de densidade observada em gerações longas produzidas em passagem única por modelos de janela expansiva.

O placar do confronto contra o Claude Sonnet 4.5 (3 vitórias, 2 derrotas e 1 empate) indica que a vantagem argumentada se confirmou de forma marginal, sendo a arquitetura de contexto expandido o adversário mais difícil entre os três provedores.

4.6.2 *Pipeline* UnB *versus* OpenAI (GPT-5.1)

A Tabela 4.5 consolida, por tarefa, os confrontos realizados contra o GPT-5.1 e os respectivos resultados.

Tabela 4.5: Detalhamento dos confrontos e resultados frente ao modelo GPT-5.1.

Número	Tarefa Analítica	Adversário	Resultado
2	Como se configura o trabalho escravo no Brasil e qual é o papel do Tribunal Superior do Trabalho (TST) no seu combate?	GPT-5.1	Vitória UnB
4	Qual o impacto da reforma tributária brasileira no setor de serviços?	GPT-5.1	Empate
7	Quais são as abordagens para detecção, mitigação e atribuição colaborativas distribuídas de ataques volumétricos de negação de serviço distribuídos?	GPT-5.1	Vitória UnB
10	Como se configuram as relações comerciais entre Brasil e China nos últimos 5 anos?	GPT-5.1	Vitória UnB
12	Descreva as funções executivas do córtex frontal do cérebro humano e responda se a inteligência artificial no estágio atual implementa alguma delas.	GPT-5.1	Vitória OpenAI
17	Na América do Sul, quais grupos, instituições ou indivíduos funcionam como agentes de influência na disseminação de ações de desinformação estrangeira?	GPT-5.1	Vitória UnB

As arquiteturas da família GPT são otimizadas para atuar como sistemas de propósito geral, generalizando o conhecimento e prevendo sequências de *tokens* com elevada fluência estilística e coerência sintática [88]. Em termos gerais, versões recentes dessa família empregam mecanismos para ajustar o

esforço computacional em tempo de inferência (*compute-at-test-time*) em função da complexidade da consulta, acelerando interações simples e reservando maior orçamento de processamento para tarefas mais exigentes.

Contudo, mesmo quando dispõe de um orçamento estendido de reflexão adaptativa, a decodificação autorregressiva do modelo comercial permanece confinada a uma passagem estocástica fechada sobre seus próprios parâmetros internos. Quando submetida a demandas de inteligência rigorosas em ambiente monolítico, essa arquitetura ainda manifesta a propensão matemática de preencher lacunas informacionais com extrapolações lógicas semanticamente plausíveis, porém factualmente incorretas, incorrendo em episódios de alucinação severa [38].

O *pipeline* da UnB mitiga essa estocasticidade ao deslocar o raciocínio deliberativo (*System 2*) para a própria topologia do agente, fora do modelo [4]. Com isso, reduz-se a margem para que o sistema "adivinhe"respostas ou dependa apenas de reflexão paramétrica interna. Se o cálculo de similaridade vetorial indicar divergência entre a síntese e a base documental subjacente, ou se o Supervisor Metacognitivo detectar um vazio nas evidências (*Knowledge Gap*), o laço de execução é interrompido e subagentes de retro-coleta são acionados para recuperar a variável ausente. Esse mecanismo força o refinamento iterativo (*Self-Refine*) [2] a permanecer ancorado em evidências, em cenários de maior incerteza.

O placar do confronto contra o GPT-5.1 (4 vitórias, 1 derrota e 1 empate) é consistente com o argumento de que a passagem única estocástica tende a ser penalizada em tarefas analíticas longas, nas quais o *pipeline* pode reancorar e refinar iterativamente.

4.6.3 *Pipeline* UnB *versus* Google (Gemini)

A Tabela 4.6 consolida, por tarefa, os confrontos realizados contra o Gemini 3 e os respectivos resultados.

Tabela 4.6: Detalhamento dos confrontos e resultados frente ao modelo Gemini 3.

Número	Tarefa Analítica	Adversário	Resultado
1	Qual é o impacto regulatório e de segurança da adoção de padrões pós-quânticos na infraestrutura de criptografia das instituições financeiras brasileiras até 2030?	Gemini 3	Vitória UnB
3	Qual a relação entre a COP 30 clima e o combate ao narcotráfico na Venezuela?	Gemini 3	Vitória UnB
6	Como funciona e quais são os efeitos da bainha de plasma que se forma ao redor dos mísseis hipersônicos durante a trajetória?	Gemini 3	Vitória Google
13	Qual o papel das terras raras na disputa tecnológica entre Estados Unidos e China?	Gemini 3	Vitória UnB
16	Como alimentos ultraprocessados contribuem para doenças crônicas no Brasil?	Gemini 3	Empate
18	Transição energética no Brasil em 2025.	Gemini 3	Empate

Os sistemas de LLM do Google, durante os testes realizados, foram tratados como referência de vanguarda em processamento multimodal nativo [37]. Diferentemente de *pipelines* clássicos que delegam transcrição de áudio e reconhecimento ótico de caracteres (OCR) a ferramentas externas, o modelo foi projetado para ser multimodal desde o *encoding* inicial. Arquivos de áudio, vídeo, imagens e texto são convertidos simultaneamente e projetados no mesmo espaço vetorial. Esse desenho permite relacionar diferentes mídias com menor perda de nuances semânticas relevantes.

Contudo, na produção de inteligência estratégica, a capacidade multimodal não elimina a dificuldade de concatenar múltiplos saltos dedutivos (*multi-hop reasoning*). Mecanismos de recuperação apoiados estritamente em similaridade semântica de *embeddings* (RAG ingênuo), mesmo em espaços vetoriais multimodais unificados, tendem a falhar ao percorrer hierarquias de causa e efeito, cronologias implícitas ou linhagens corporativas fragmentadas [45].

O *pipeline* da UnB mitiga essa limitação por meio do motor de RAG Híbrido Agêntico. A arquitetura não limita o raciocínio investigativo a distâncias vetoriais no espaço latente; ela combina busca semântica com travessia estruturada de Grafos de Conhecimento (*Knowledge Graphs*) [45]. Além disso, a Etapa de Direção não conduz o planejamento analítico de forma linear: ela constrói um Grafo Acíclico Direcionado (DAG), inspirado no paradigma *Graph of Thoughts* [1]. Ao exigir o traçado explícito de relações e dependências antes da geração textual, o sistema UnB resolve correlações de inteligência que a sobreposição atencional simples de modelos como o Gemini tende a não mapear com o mesmo rigor.

O placar do confronto contra o Gemini 3 (3 vitórias, 1 derrota e 2 empates) é consistente com o argumento de que a combinação de RAG híbrido e planejamento explícito via DAG favorece tarefas que exigem encadeamento de evidências, ainda que o resultado inclua empates e uma derrota.

4.6.4 Autonomia e Isolamento Operacional

Além das dinâmicas cognitivas discutidas anteriormente, o artefato atende a um requisito operacional relevante na esfera da inteligência de Estado: soberania algorítmica e isolamento informacional. Modelos proprietários, independentemente de sua capacidade de síntese, dependem estruturalmente de infraestrutura em nuvem. Essa dependência implica a exfiltração contínua de metadados, intenções analíticas e arquivos sensíveis para jurisdições corporativas estrangeiras, configurando vulnerabilidade a vazamentos por inferência ou inversão de vetores [14].

Em contraste, a engenharia do *pipeline* da UnB foi desenhada como uma camada de isolamento operacional. Por meio de quantização ciente de ativação (AWQ) [18] e particionamento lógico via arquitetura de Especialistas (MoE), o sistema viabiliza capacidades de raciocínio de fronteira em estações de trabalho e servidores isolados da rede pública. Esse isolamento assegura que a matriz investigativa permaneça fisicamente segregada, atendendo a requisitos de classificação e soberania de dados e, ao mesmo tempo, melhorando a viabilidade operacional por reduzir dependências externas e custos variáveis de execução. Como benefício adicional, esse conjunto de otimizações tende a reduzir desperdício computacional e consumo energético na inferência [87, 15].

4.7 RESULTADOS AGREGADOS DA AVALIAÇÃO

As execuções de prova do artefato da UnB foram operadas integralmente em infraestrutura local restrita (um sistema de computação acelerada NVIDIA DGX Spark pareado em coprocessamento com duas GPUs comerciais RTX 4060 Ti). Os resultados consolidados da apuração às cegas emitidos pelos juízes de fronteira encontram-se expostos na Tabela 4.7.

Tabela 4.7: Avaliação preliminar via *LLM-as-a-Judge* multi-juiz sobre 18 tarefas analíticas, com rotação cega de papéis e crítica/contracrítica mediada por supervisor.

Rodada de Confronto	**Vitórias UnB**	**Vitórias Baseline**	**Empates**
GPT-5.1 *vs.* Pipeline UnB	4	1	1
Gemini 3 *vs.* Pipeline UnB	3	1	2
Claude Sonnet 4.5 *vs.* Pipeline UnB	3	2	1
Total Agregado	**10**	**4**	**4**

O desempenho apurado totaliza dez vitórias, quatro empates e quatro derrotas frente a sistemas corporativos. O saldo positivo em cada rodada isolada reforça a validade do desenho procedimental adotado, indicando que a qualidade das saídas decorre principalmente do modelo arquitetural do *pipeline*, e não de efeitos de julgamento associados a combinações específicas de *prompts* e nem de LLMs em particular.

Um ponto central da validação empírica é a assimetria no tempo de resposta. Enquanto modelos de fronteira, apoiados em infraestruturas de nuvem, resolveram as consultas em menos de um minuto, o *pipeline* proposto exigiu até duas horas de processamento local por tarefa.

Esse tempo não é um limite estrutural imutável do desenho, mas um retrato do protótipo avaliado e das restrições adotadas para manter rastreabilidade e isolamento. Como o *pipeline* é multiestágio, parte relevante do custo se concentra em chamadas repetidas nos laços de *ReAct* e de *Self-Refinement*, além da geração hierárquica via H-CoD, que executa ciclos sucessivos de escrita e densificação. Há caminhos concretos para reduzir esse custo sem alterar o princípio do *pipeline*: reutilização mais agressiva de *cache* semântico e de *embeddings* entre subperguntas, paralelização de nós independentes do DAG quando não há dependência de dados, redução do número de iterações de reflexão e refinamento por critérios de parada mais estritos, e roteamento adaptativo de modelos, alocando modelos menores nas subtarefas de baixa entropia. Em conjunto, essas medidas deslocam o gargalo do *compute* bruto para o controle do fluxo, preservando a qualidade do relatório enquanto encurtam a janela total de execução.

Essa diferença não implica, por si só, ineficiência; ela reflete divergência metodológica entre as abordagens. Modelos comerciais entregam respostas quase instantâneas porque operam por predição direta, em uma única passagem de decodificação (*single-pass*). Em contraste, o tempo consumido pelo sistema proposto reflete um processo deliberativo iterativo: o *pipeline* formula hipóteses, executa buscas ancoradas em evidências, mapeia relações e aplica rotinas sucessivas de autocorreção (*Self-Refinement*) antes de consolidar o relatório final.

Na rotina da inteligência de Estado, a *baseline* operacional não é a velocidade conversacional de um *chatbot*, mas o horizonte metodológico da força de trabalho humana. A síntese fidedigna de dezenas de

fontes em um *briefing* estruturado, operando sobre *hardware* isolado de influência de nuvem, exige esforços que variam de horas a semanas. Nesse contexto, assegurar governança plena sobre os dados em ambiente local e isolado, e manter desempenho qualitativo competitivo, sugere que abrir mão de baixa latência em troca de maior controle e ancoragem factual é uma decisão arquitetural coerente com cenários de alto impacto e criticidade.

4.8 ANÁLISE QUALITATIVA POR DIMENSÃO

A análise dos resultados explicita como as escolhas arquiteturais do sistema afetaram as três dimensões qualitativas avaliadas no confronto com modelos comerciais.

Na dimensão de precisão factual e ancoragem, o desempenho do *pipeline* decorre da restrição metodológica imposta aos agentes. Ao exigir que afirmações sejam sustentadas pelos trechos recuperados, a arquitetura reduz a tendência de modelos de linguagem preencherem lacunas com dados inventados. Nas justificativas registradas pelos juízes, essa validação algorítmica aparece de forma recorrente como fator de diferenciação na dimensão de precisão factual, com poucos apontamentos de invenção factual mesmo em contextos que ultrapassaram milhares de *tokens* [38].

Quanto à densidade informacional e à coerência estrutural, o desempenho observado está associado ao módulo de redação (H-CoD). Em relatórios extensos, LLMs tendem a perder evidências centrais, fenômeno descrito como *lost-in-the-middle* [7]. O *pipeline* mitiga esse efeito ao separar planejamento e escrita: primeiro projeta um esboço lógico da estrutura (DOME) [5] e, em seguida, adensa os fatos de cada seção de forma isolada (*Chain-of-Density*) [54]. O resultado é um documento com maior concentração de informações relevantes por parágrafo, preservando clareza e fluidez.

Por outro lado, as quatro derrotas evidenciam limites práticos de uma abordagem mais rígida. Em tarefas que demandavam abstração teórica, criatividade ou maior fluência narrativa, em que a exatidão das evidências era menos determinante, os *baselines* comerciais tiveram vantagem. Nesses casos, a maior escala paramétrica e o volume de treinamento em nuvem favorecem respostas mais sofisticadas e um estilo de escrita que o *pipeline* proposto ainda não iguala integralmente.

4.9 DISCUSSÃO: ARQUITETURA E ESCALA

Os resultados experimentais são consistentes com a premissa central deste trabalho: em tarefas complexas de síntese de cenários, a arquitetura do sistema e o controle do fluxo de dados são mais determinantes para o sucesso do que a simples quantidade de parâmetros do modelo de linguagem [24, 15].

Do ponto de vista da arquitetura, o *pipeline* integra conceitos recentes da literatura especializada para estabelecer seu funcionamento. O sistema adota a separação de funções proposta no *framework* CoALA, utiliza os ciclos de raciocínio e ação do ReAct e aplica a validação iterativa baseada no método Reflexion [40, 69]. O avanço metodológico obtido, no entanto, resulta do controle estrito imposto a esses mecanismos, que foram organizados segundo as fases do Ciclo de Inteligência da OTAN [60].

A vantagem observada decorre do desenho arquitetural — a restrição da liberdade do modelo por decomposição metodológica, recuperação híbrida e laços de refinamento — e não do local de execução: o mesmo *pipeline* poderia operar sobre modelos remotos. Ao processar dados de forma linear em uma única passagem (*single-pass* ou *zero-shot*), grandes modelos ficam sujeitos a erros de inferência e alucinações, especialmente em tarefas longas e complexas. Ao inserir o LLM em um fluxo de trabalho estruturado, a capacidade do modelo é direcionada para resolver uma subtarefa específica por vez.

Essa transferência do esforço de raciocínio, da rede neural para a arquitetura do *pipeline*, é o que viabiliza o uso de modelos computacionalmente mais eficientes. O fato de o *pipeline* composto, executado localmente com modelos abertos, superar arquiteturas monolíticas de maior escala paramétrica nas tarefas testadas sugere a eficácia de abordagens baseadas em raciocínio deliberativo (*System 2*) [4] na produção de conhecimento.

4.10 O GARGALO DA VALIDAÇÃO HUMANA E O *TIME TO INTELLIGENCE*

A avaliação de sistemas de apoio à decisão exige mudar o foco da eficiência puramente computacional (o tempo de inferência) para uma métrica operacional mais ampla, como o *Time to Intelligence* (TTI). O TTI representa o ciclo completo, desde a formulação do problema até a entrega de um produto analítico validado e pronto para uso.

Sob essa ótica, o principal gargalo na produção automatizada de conhecimento não está na capacidade de geração da máquina, mas no trabalho humano. Especificamente, o problema recai sobre o tempo e a carga cognitiva exigidos do analista para revisar e validar o material gerado [20]. Modelos comerciais de fronteira otimizam apenas a primeira metade dessa equação, pois entregam textos longos em poucos segundos. Contudo, por operarem como "caixas-pretas", eles transferem todo o peso da verificação para o analista. Validar um relatório extenso gerado sem rastreabilidade de fontes exige que o especialista faça uma engenharia reversa das afirmações, buscando manualmente os documentos originais para descartar alucinações. Na prática, a agilidade aparente da máquina transforma-se em desperdício de horas de trabalho qualificado, um problema que se relaciona à literatura de confiança e calibração de confiança em automação [89].

O *pipeline* desenvolvido busca inverter essa lógica. Embora a produção de um documento analítico de dez páginas exija até duas horas de processamento computacional ininterrupto, a vantagem da ferramenta está na transparência do processo. Ao entregar um material documentado, com matrizes de relacionamento, referências cruzadas explícitas e validação prévia de informações divergentes, o sistema reduz o esforço de revisão.

Nesse cenário, o analista deixa de ser um mero verificador de fatos isolados para atuar como um auditor estratégico. A Tabela 4.8 ilustra conceitualmente essa mudança na distribuição do tempo dentro do ciclo de inteligência.

A estimativa sugere que o maior tempo de processamento exigido pelo artefato seria compensado pela redução nas horas de auditoria humana — projeção a ser verificada na etapa de validação com painel humano, prevista como trabalho futuro. Em ambientes de missão crítica, o custo do processamento local é

Tabela 4.8: Comparativo conceitual (projeção do autor, não medida experimentalmente) de *Time to Intelligence* (TTI) entre abordagens de IA em Inteligência Estratégica.

Etapa do Ciclo (TTI)	**Modelos de Fronteira (Caixa-Preta / Single-pass)**	**Pipeline UnB (Processo Deliberativo / Execução local)**
1. Tempo de Geração (Máquina)	< 1 minuto	≈ 120 minutos
2. Nível de Rastreabilidade	Baixo (Ausência de mapeamento das fontes originais)	Alto (Ancoragem estrita via DAG e RAG Híbrido)
3. Esforço de Validação (Humano)	Alto (≈ 4 a 6 horas para checagem factual manual)	Baixo (≈ 30 a 45 minutos de auditoria direcionada)
Tempo Total Estimado (TTI)	≈ 4 a 6 horas	≈ 2 a 3 horas

Fonte: Elaboração do autor (2026), baseada em projeções operacionais práticas e em conceitos de carga cognitiva [89, 20].

irrelevante se comparado ao custo do tempo de um analista sênior. Dessa forma, exigir que a inteligência artificial documente e valide seus próprios passos antes da entrega final mostrou-se uma abordagem mais eficiente para reduzir o *Time to Intelligence* e acelerar a tomada de decisão.

4.11 TRAJETÓRIA DE OTIMIZAÇÃO E VIABILIDADE OPERACIONAL

A capacidade de rodar o *pipeline* localmente e entregar relatórios em até duas horas não foi uma característica inata do projeto. O sistema começou como uma prova de conceito funcional, porém computacionalmente inviável para o uso diário. Para transformá-lo em uma ferramenta viável, foi necessário um trabalho focado em substituir a dependência de poder de processamento bruto por eficiência na arquitetura de *software*.

As Figuras 4.2 e 4.3 mostram o impacto dessas otimizações ao longo de quatro trimestres de desenvolvimento (T1 a T4). Os gráficos acompanham a evolução de duas métricas essenciais: a redução do espaço exigido em memória de vídeo (VRAM) e a queda no tempo de processamento necessário para finalizar uma tarefa.

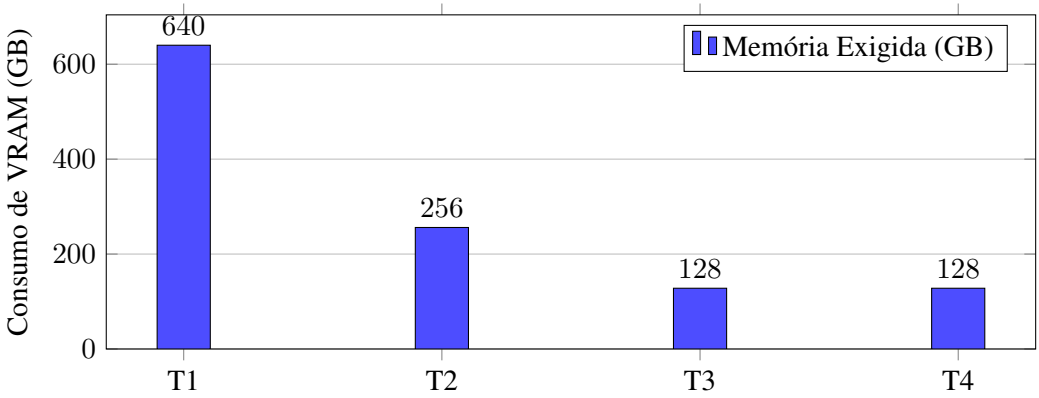

Figura 4.2: Evolução do consumo de memória de vídeo (VRAM) ao longo dos trimestres de desenvolvimento.

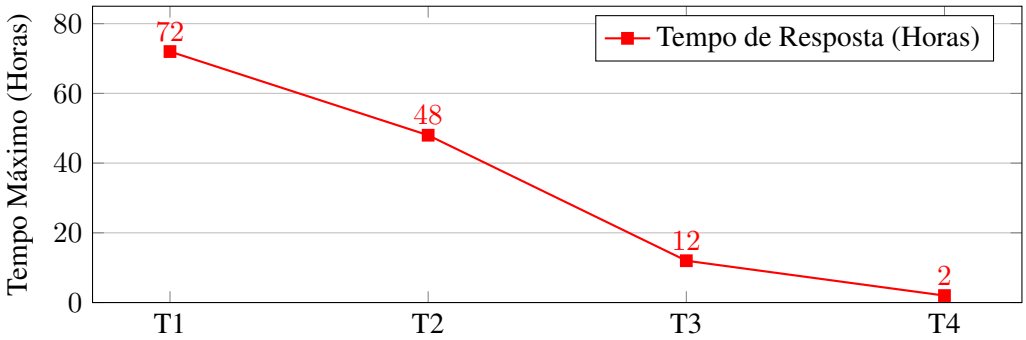

Figura 4.3: Redução do tempo de processamento necessário para finalizar uma tarefa analítica em cada trimestre.

A leitura dos gráficos indica que a eficiência do sistema avançou através de implementações arquiteturais diretas. Cada queda acentuada nos indicadores corresponde a uma etapa específica do desenvolvimento:

- **Primeiro Trimestre (T1), *Pipeline* Inicial com 5W3H:** A versão inicial gerou uma exigência de memória agregada estimada de 640 GB e até 72 horas de processamento por tarefa. O sistema carregava todas as evidências disponíveis e processava as subperguntas de forma sequencial. Ao tentar manter o contexto completo em memória simultaneamente, o protótipo tornou-se incompatível com a infraestrutura efetivamente empregada nos testes (DGX Spark em coprocessamento com duas RTX 4060 Ti).
- **Segundo Trimestre (T2), Metodologias de Refinamento Iterativo:** A primeira redução no uso de memória (Figura 4.2) ocorreu ao fragmentar a carga de trabalho. Com rotinas iterativas, o sistema passou a processar uma etapa por vez, persistir resultados intermediários e liberar memória entre ciclos. Essa separação reduziu a exigência agregada para cerca de 256 GB, embora o tempo de processamento (Figura 4.3) ainda se mantivesse em torno de 48 horas.
- **Terceiro Trimestre (T3), H-CoD e Quantização de Modelos:** No terceiro trimestre, o tempo de processamento recuou para cerca de 12 horas. Esse ganho decorreu da adoção da arquitetura H-CoD

(*Hierarchical Chain-of-Density*) e do uso de quantização, reduzindo a pressão de memória durante a inferência. Nesse estágio, a exigência agregada estabilizou-se em torno de 128 GB.

- **Quarto Trimestre (T4), RAG Híbrido e Modelos Esparsos (MoE):** O trimestre final manteve a estabilidade de memória e reduziu o tempo operacional para a marca atual. Essa melhoria decorre do motor de RAG híbrido, que restringe a injeção de evidências ao subconjunto relevante, e do uso de modelos esparsos (MoE), que reduz ativação efetiva por *token*. Com isso, a exigência agregada permaneceu em torno de 128 GB, e a janela de processamento caiu para aproximadamente 2 horas.

A evolução dos indicadores nas Figuras 4.2 e 4.3 demonstra a viabilidade técnica de rodar capacidades analíticas em sistemas locais e isolados. O ganho decorreu de mudanças no fluxo de informações e no controle do raciocínio do sistema ao longo dos quatro trimestres, e não da simples introdução de *hardware* adicional como variável de comparação. A trajetória indica, portanto, quanto desempenho pode ser recuperado por engenharia de arquitetura sob restrição de recursos.

5 CONCLUSÕES

Este trabalho atingiu seu objetivo geral ao projetar e validar um Sistema de IA Composta capaz de produzir relatórios analíticos de inteligência com precisão factual e boa densidade informacional, mesmo sob restrições severas de *hardware*. Os resultados apresentados decorrem do atendimento dos quatro objetivos específicos delineados no início da pesquisa.

Em primeiro lugar, foi estruturado um *pipeline* agêntico de seis etapas. Observou-se, nos experimentos, que a decomposição da intenção do usuário pela matriz 5W3H é um mecanismo viável para reduzir alucinações em Grandes Modelos de Linguagem (LLMs). Ao separar planejamento tático e execução textual, o sistema manteve a produção de inteligência em um roteiro metodológico auditável e reprodutível. Em segundo lugar, a implementação do motor de RAG Híbrido Agêntico reduziu a fragmentação de contexto. A triangulação entre bases relacionais, índices vetoriais e grafos de conhecimento permitiu explorar relações entre entidades, mitigando limitações da busca por similaridade em consultas de múltiplos saltos.

O terceiro objetivo foi atendido com a proposição e integração da estratégia *Hierarchical Chain-of-Density* (H-CoD). Os testes indicaram que a combinação de planejamento hierárquico prévio com densificação iterativa reduz o fenômeno *lost-in-the-middle*. O método produziu documentos executivos extensos e coesos, com menor verbosidade e melhor razão sinal–ruído. Por fim, o quarto objetivo foi consolidado no protocolo de avaliação empírica. O desenho de *LLM-as-a-Judge* com múltiplos juízes submeteu o sistema à comparação com modelos líderes da indústria, com rotação cega de papéis e com crítica/contracrítica mediada por um agente supervisor.

Tomados em conjunto, esses resultados sustentam a hipótese central da pesquisa: limitações cognitivas e factuais de modelos menores não são, necessariamente, barreiras intransponíveis; elas podem ser tratadas como desafios de engenharia de *software* e arquitetura de dados. O desempenho de dez vitórias, quatro empates e quatro derrotas contra arquiteturas como GPT-5.1, Gemini 3 e Claude Sonnet 4.5 sugere que a orquestração de raciocínio deliberativo (*System 2*) pode compensar, em parte, a simples escala paramétrica.

5.1 IMPLICAÇÕES OPERACIONAIS E SOBERANIA TECNOLÓGICA

Além das métricas computacionais estabelecidas na avaliação experimental, as contribuições do sistema têm implicações para a doutrina de inteligência e para a autonomia tecnológica institucional. A dependência de infraestruturas em nuvem comercial pode ampliar riscos de segurança, sobretudo quando metadados táticos e diretrizes de busca são processados em servidores sob jurisdição estrangeira. Esse arranjo aumenta a exposição de intenções analíticas e restringe o emprego de modelos de fronteira em conjuntos de dados sensíveis.

O *pipeline* desenvolvido reduz esse risco ao viabilizar a execução de capacidades analíticas em regime de isolamento (*air-gapped*). Ao adequar o raciocínio de alto nível aos limites de *hardware* de consumo, em especial instâncias aceleradas por placas RTX 4060 Ti, o projeto indica que o Estado e organizações

estratégicas podem operar sem depender de *data centers* massivos em determinados cenários. Com isso, o ciclo de produção do conhecimento, da ingestão de fontes abertas (OSINT) à emissão de documentos, tende a permanecer sob maior controle institucional.

Além disso, a adoção desse *framework* pode alterar a distribuição do trabalho nas agências de inteligência. Na rotina convencional, o analista dedica grande parte do tempo à coleta, deduplicação e ao cruzamento manual de fontes. A arquitetura proposta automatiza parte desse esforço ao delegar o mapeamento de entidades, a verificação de contradições e a redação estruturada a agentes autônomos. Assim, o analista passa a atuar mais como auditor, concentrando-se na validação do raciocínio gerado e no suporte à tomada de decisão. Essa redistribuição contribui para justificar o custo computacional do processo iterativo, ao priorizar o tempo do especialista em tarefas de maior valor.

5.2 LIMITAÇÕES DA PESQUISA

As restrições identificadas ao longo da pesquisa concentram-se em três dimensões: escopo, protocolo de validação e arquitetura operacional. Quanto ao escopo, o *pipeline* foi concebido e testado exclusivamente em tarefas textuais, nos idiomas português e inglês. Como a inteligência de fontes abertas contemporânea depende, com frequência, de ingestão multimodal, a ausência de processamento de fluxos de áudio, vídeo e telemetria em tempo real restringe o conjunto de cenários que o sistema pode analisar de forma autônoma. Além disso, o foco em demonstrar viabilidade arquitetural deixou em segundo plano otimizações de baixo nível. Ganhos adicionais em latência e consumo energético ainda podem ser obtidos com refatorações voltadas ao paralelismo de tensores, considerando a arquitetura de *hardware* utilizada.

No âmbito da validação metodológica, a delegação do julgamento a juízes algorítmicos preserva incertezas. A rotação cega de papéis e a mediação do supervisor mitigaram anomalias evidentes, mas modelos de linguagem podem apresentar vieses que afetam o veredito, como favorecer saídas prolixas (*verbosity bias*) ou narrativas alinhadas ao próprio treinamento (*self-enhancement bias*). Embora o protocolo ajude a reforçar coerência semântica e ancoragem de entidades, ele não mede diretamente o impacto pragmático da inteligência gerada sob a ótica de um tomador de decisão sob forte pressão de tempo.

Quanto à amostra experimental, o protocolo cobriu 18 tarefas analíticas, seis por confronto, com enunciados heterogêneos em tema e em profundidade. O tamanho reduzido e a variabilidade das tarefas limitam a generalização dos placares agregados, que devem ser lidos como evidência preliminar. Adicionalmente, a taxa de alucinação factual não foi medida diretamente, sendo inferida apenas indiretamente pelos critérios do julgamento.

Do ponto de vista operacional, a execução em infraestrutura física isolada reduz o risco de vazamentos, mas pode introduzir atrasos na ingestão de fatos dinâmicos. Na prática, o estado de conhecimento do sistema fica limitado ao momento da última atualização de seus índices. Em cenários de crise, a ausência de conectividade autônoma em tempo real pode reduzir a utilidade de relatórios que dependem de atualização contínua.

Por fim, a elevada autonomia do *pipeline* limita a interatividade humana. A arquitetura atual processa o ciclo de inteligência de forma assíncrona e em lote, com pouca possibilidade de intervenção do usuário

entre as etapas de coleta e análise. Um copiloto cognitivo mais efetivo exigiria que o especialista pudesse recalibrar premissas ou refutar evidências durante o processamento (*Human-in-the-Loop*). Nesta versão, o formato em lote tende a produzir respostas mais fechadas, dificultando a colaboração investigativa passo a passo.

5.3 TRABALHOS FUTUROS

A agenda de pesquisa derivada deste trabalho organiza-se em três frentes: calibração empírica, aprimoramentos de engenharia e evolução arquitetural. A curto prazo, a prioridade é expandir a base de testes para abranger as cem tarefas analíticas previstas, submetendo as saídas a um painel humano composto por analistas seniores de inteligência. O cruzamento dos vereditos humanos com os resultados do protocolo *LLM-as-a-Judge* multi-juiz permitirá auditar o viés do juiz algorítmico e estabelecer um fator de correção para avaliações futuras. Em paralelo, é necessário aprofundar a avaliação de viabilidade operacional em regimes de uso contínuo e em ambientes *on-premises*, com ênfase em soberania de dados (governança, trilhas de auditoria e superfícies de exposição) e em previsibilidade de custo. Por fim, medir o comportamento térmico e elétrico do hardware em operação permite quantificar ganhos de eficiência e caracterizar impactos energéticos como efeito colateral das otimizações propostas, em diálogo com a agenda de sustentabilidade em IA.

No campo da engenharia de dados, um próximo passo é construir canais de ingestão criptografados que permitam ao motor de RAG Híbrido consumir fluxos dinâmicos de Inteligência de Ameaças Cibernéticas (CTI). Paralelamente, a aplicação de grafos dinâmicos de atributos pode ajudar a modular o limite de densificação textual do H-CoD, evitando relatórios condensados além do que é adequado para o leitor. Explora-se também o potencial da destilação reversa, utilizando as saídas refinadas deste pipeline para treinar modelos menores e mais eficientes na borda (Edge Computing).

Uma linha de pesquisa adicional é reduzir a rigidez do pipeline. A arquitetura de seis estágios cumpriu seu papel de isolamento cognitivo, mas há espaço para uma orquestração mais flexível. Com base no paradigma de Automated Design of Agentic Systems (ADAS), pretende-se permitir que o supervisor metacognitivo sintetize o próprio grafo de execução. Diante da intenção inicial do usuário, o sistema poderia instanciar dinamicamente apenas as ferramentas de busca, subagentes e camadas de validação exigidas pela complexidade do problema. Com isso, a arquitetura evoluiria de um fluxo procedimental fixo para uma solução mais adaptativa e contínua.

REFERÊNCIAS BIBLIOGRÁFICAS

1 BESTA, M. et al. Graph of thoughts: Solving elaborate problems with large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, v. 38, n. 16, 2024.

2 MADAAN, A. et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, v. 36, 2024.

3 CRUZ, C. C. d. S.; ALBUQUERQUE, R. O. d.; GONDIM, J. J. C.; AMARAL, D. M.; MENDONÇA, F. L. L. d. *AI Pipeline Applied to an Adaptive Multistage Decision Support System for Strategic Intelligence*. 2026. Proceeding paper submitted to Engineering Proceedings (MDPI). No prelo.

4 KAHNEMAN, D. *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux, 2011.

5 GU, Y. et al. DOME: Dynamic hierarchical outlining with memory-enhancement for long-form text generation. *arXiv preprint arXiv:2402.04501*, 2024.

6 WANG, Z. et al. RAPID: Efficient retrieval-augmented long text generation with writing planning and information discovery. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. [S.l.: s.n.], 2024.

7 LIU, N. F. et al. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, v. 12, p. 157–173, 2024.

8 PATTERSON, D.; GONZALEZ, J.; LE, Q.; LIANG, C.; MUNGUIA, L.-M.; ROTHCHILD, D.; SO, D.; TEXIER, M.; DEAN, J. Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350*, 2021.

9 LUCCIONI, A. S.; JERNITE, Y.; STRUBELL, E. Power hungry processing: Watts driving the cost of AI deployment? In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. [S.l.: s.n.], 2024. p. 85–99.

10 WU, C.-J.; RAGHAVENDRA, R.; GUPTA, U.; ACUN, B.; ARDALANI, N.; MAENG, K. et al. Sustainable AI: Environmental implications, challenges and opportunities. In: *Proceedings of Machine Learning and Systems (MLSys 2022)*. [S.l.: s.n.], 2022. ArXiv:2111.00364.

11 LI, B.; JIANG, Y.; GADEPALLY, V.; TIWARI, D. Sprout: Green generative AI with carbon-efficient LLM inference. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)*. [S.l.: s.n.], 2024. p. 21799–21813. ArXiv:2403.12900.

12 Office of the Director of National Intelligence (ODNI) and Central Intelligence Agency (CIA). *The IC OSINT Strategy 2024–2026: The INT of First Resort; Unlocking the Value of OSINT*. 2024. <https://www.dni.gov/files/ODNI/documents/IC_OSINT_Strategy.pdf>. Acessado em 2025.

13 U.S. Department of State, Bureau of Intelligence and Research (INR). *Open Source Intelligence Strategy*. 2024. <https://2021-2025.state.gov/open-source-intelligence-strategy/>. Acessado em 2025.

14 DAS, B. C.; AMINI, M. H.; WU, Y. Security and privacy challenges of large language models: A survey. *ACM Computing Surveys*, v. 57, n. 6, p. 1–39, 2025.

15 BELCAK, P.; HEINRICH, G.; DIAO, S.; FU, Y.; DONG, X.; MURALIDHARAN, S.; LIN, Y. C.; MOLCHANOV, P. Small language models are the future of agentic AI. *arXiv preprint arXiv:2506.02153*, 2025.

16 ABDIN, M.; JACOBS, S. A.; AWAN, A. A. et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

17 DeepSeek-AI; GUO, D.; YANG, D. et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

18 LIN, J.; TANG, J.; TANG, H.; YANG, S.; CHEN, W.-M.; WANG, W.-C.; XIAO, G.; DANG, X.; GAN, C.; HAN, S. AWQ: Activation-aware weight quantization for LLM compression and acceleration. In: *Proceedings of Machine Learning and Systems (MLSys 2024)*. [S.l.: s.n.], 2024. ArXiv:2306.00978, Best Paper Award.

19 FRANTAR, E.; ASHKBOOS, S.; HOEFLER, T.; ALISTARH, D. GPTQ: Accurate post-training quantization for generative pre-trained transformers. In: *International Conference on Learning Representations (ICLR 2023)*. [S.l.: s.n.], 2023. ArXiv:2210.17323.

20 RUAN, J.; NAIR, I.; CAO, S.; LIU, A.; MUNIR, S.; POLLENS-DEMPSEY, M.; CHIANG, T.; KATES, L.; DAVID, N.; CHEN, S. et al. ExpertLongBench: Benchmarking language models on expert-level long-form generation tasks with structured checklists. *arXiv preprint arXiv:2506.01241*, 2025.

21 ZAHARIA, M.; KHATTAB, O.; CHEN, L.; DAVIS, J. Q.; MILLER, H.; POTTS, C.; ZOU, J.; CARBIN, M.; FRANKLE, J.; RAO, N.; GHODSI, A. *The Shift from Models to Compound AI Systems*. 2024. Berkeley Artificial Intelligence Research (BAIR) Blog, <https://bair.berkeley.edu/blog/2024/02/18/compound-ai-systems/>. Acessado em 2025.

22 GUAN, X.; ZHANG, L. L.; LIU, Y.; SHANG, N.; SUN, Y.; ZHU, Y.; YANG, F.; YANG, M. rStar-Math: Small LLMs can master math reasoning with self-evolved deep thinking. In: *International Conference on Machine Learning (ICML 2025)*. [S.l.: s.n.], 2025. ArXiv:2501.04519.

23 BROWNE, C. B.; POWLEY, E.; WHITEHOUSE, D.; LUCAS, S. M.; COWLING, P. I.; ROHLFSHAGEN, P.; TAVENER, S.; PEREZ, D.; SAMOTHRAKIS, S.; COLTON, S. A survey of monte carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in Games*, v. 4, n. 1, p. 1–43, 2012.

24 CHEN, L.; DAVIS, J. Q.; HANIN, B.; BAILIS, P.; STOICA, I.; ZAHARIA, M.; ZOU, J. Are more LLM calls all you need? Towards the scaling properties of compound AI systems. In: *Advances in Neural Information Processing Systems (NeurIPS 2024)*. [S.l.: s.n.], 2024. ArXiv:2403.02419.

25 KHATTAB, O.; SINGHVI, A.; MAHESHWARI, P.; ZHANG, Z.; SANTHANAM, K.; VARDHAMANAN, S.; HAQ, S.; SHARMA, A.; JOSHI, T. T.; MOAZAM, H.; MILLER, H.; ZAHARIA, M.; POTTS, C. DSPy: Compiling declarative language model calls into state-of-the-art pipelines. In: *International Conference on Learning Representations (ICLR 2024)*. [S.l.: s.n.], 2024. ArXiv:2310.03714.

26 HENDEL, R.; GEVA, M.; GLOBERSON, A. In-context learning creates task vectors. In: *Findings of the Association for Computational Linguistics (ACL 2024)*. [S.l.: s.n.], 2024.

27 DU, Y.; LI, S.; TORRALBA, A.; TENENBAUM, J. B.; MORDATCH, I. Improving factuality and reasoning in language models through multiagent debate. In: *International Conference on Machine Learning (ICML 2024)*. [S.l.: s.n.], 2024. ArXiv:2305.14325.

28 LIANG, T.; HE, Z.; JIAO, W.; WANG, X.; WANG, Y.; WANG, R.; YANG, Y.; SHI, S.; TU, Z. Encouraging divergent thinking in large language models through multi-agent debate. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)*. [S.l.: s.n.], 2024. p. 17889–17904. ArXiv:2305.19118.

29 CHAN, C.-M.; CHEN, W.; SU, Y.; YU, J.; XUE, W.; ZHANG, S.; FU, J.; LIU, Z. ChatEval: Towards better LLM-based evaluators through multi-agent debate. In: *International Conference on Learning Representations (ICLR 2024)*. [S.l.: s.n.], 2024. ArXiv:2308.07201.

30 TAN, S.; ZHUANG, S.; MONTGOMERY, K.; TANG, W. Y.; CUADRON, A.; WANG, C.; POPA, R. A.; STOICA, I. JudgeBench: A benchmark for evaluating LLM-based judges. In: *International Conference on Learning Representations (ICLR 2025)*. [S.l.: s.n.], 2025. ArXiv:2410.12784.

31 LIN, D.; SHEN, G.; YANG, Z.; LIU, T.; ZHAO, D.; ZENG, Y. Efficient LLM safety evaluation through multi-agent debate. *arXiv preprint arXiv:2511.06396*, 2025.

32 CERVO, A. L.; BERVIAN, P. A.; SILVA, R. *Metodologia científica*. 6. ed. São Paulo: Pearson Prentice Hall, 2006.

33 PEFFERS, K.; TUUNANEN, T.; ROTHENBERGER, M. A.; CHATTERJEE, S. A design science research methodology for information systems research. *Journal of Management Information Systems*, v. 24, n. 3, p. 45–77, 2007.

34 PIMENTEL, M.; FILIPPO, D.; SANTORO, F. M. Design science research: fazendo pesquisas científicas rigorosas atreladas ao desenvolvimento de artefatos computacionais projetados para a educação. In: JAQUES, P. A.; PIMENTEL, M.; SIQUEIRA, S.; BITTENCOURT, I. (Ed.). *Metodologia de Pesquisa Científica em Informática na Educação: Concepção de Pesquisa*. Porto Alegre: Sociedade Brasileira de Computação (SBC), 2020, (Série Metodologia de Pesquisa em Informática na Educação, v. 1). Disponível em: <https://ceie.sbc.org.br/metodologia/livro-1/>.

35 SOUSA, C. E. *Proposta de um Framework para Enriquecimento de Inteligência de Ameaças Cibernéticas Utilizando Informações a partir de Inteligência de Fontes Abertas*. Dissertação (Dissertação de Mestrado Profissional) — Universidade de Brasília, Programa de Pós-Graduação Profissional em Engenharia Elétrica (PPEE), Brasília, DF, 2025. Publicação PPEE.MP.094.

36 LIN, M.; SHENG, J.; ZHAO, A.; WANG, S.; YUE, Y.; HUANG, V. S. J.; LIU, H.; LIU, J.; HUANG, G.; LIU, Y.-J. LLM-based optimization of compound AI systems: A survey. *arXiv preprint arXiv:2410.16392*, 2024.

37 REID, M. et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

38 LI, X. et al. Semantic consistency checks in retrieval-augmented generation: A vector-based approach. *IEEE Transactions on Knowledge and Data Engineering (Early Access)*, 2024.

39 WEI, J. et al. Chain-of-thought prompting elicits reasoning in large language models. In: *Advances in Neural Information Processing Systems (NeurIPS)*. [S.l.: s.n.], 2022.

40 YAO, S. et al. ReAct: Synergizing reasoning and acting in language models. In: *International Conference on Learning Representations (ICLR)*. [S.l.: s.n.], 2023.

41 RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. ẅhy should i trust you?̈: Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2016)*. [S.l.: s.n.], 2016. p. 1135–1144.

42 YANG, L. et al. Buffer of thoughts: Thought-augmented reasoning with large language models. *arXiv preprint arXiv:2406.04271*, 2024.

43 RENZE, M. et al. Self-reflection in large language model agents: Effects on problem-solving performance. In: *IEEE International Conference on Foundation and Large Language Models (FLLM)*. [S.l.: s.n.], 2025.

44 WANG, Y.; ZHAO, H. Decoupling metacognition from cognition: A framework for quantifying metacognitive ability in LLMs. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. [S.l.: s.n.], 2025.

45 EDGE, D. et al. From local to global: A graph RAG approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.

46 HU, Y. et al. Automated design of agentic systems (ADAS). In: *Proceedings of the International Conference on Learning Representations (ICLR)*. [S.l.: s.n.], 2025.

47 YE, Z. et al. Meta-agents: Scalable self-organizing hierarchies. *arXiv preprint arXiv:2502.04501*, 2025.

48 WU, D. et al. HiMATE: A hierarchical multi-agent framework for machine translation evaluation. *arXiv preprint arXiv:2505.16281*, 2025.

49 HONG, S. et al. MetaGPT: Meta programming for a multi-agent collaborative framework. In: *International Conference on Learning Representations (ICLR)*. [S.l.: s.n.], 2024. Publicado originalmente como arXiv preprint em 2023.

50 ZHUGE, M. et al. Mindstorms in natural language-based societies of mind. *Advances in Neural Information Processing Systems (NeurIPS)*, v. 37, 2024.

51 ZHANG, J. et al. ReVISE: Learning to refine at test-time via intrinsic self-verification. *arXiv preprint arXiv:2501.01237*, 2025.

52 TAUBENFELD, A. et al. Confidence improves self-consistency in LLMs. In: *Findings of the Association for Computational Linguistics (ACL 2025)*. [S.l.: s.n.], 2025.

53 MAO, K. et al. Hierarchical document refinement for long-context retrieval-augmented generation. In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025)*. [S.l.: s.n.], 2025.

54 ADAMS, G.; FABBRI, A.; LADHAK, F.; LEHMAN, E.; EL-SAYED, N. From sparse to dense: GPT-4 summarization with chain of density prompting. *arXiv preprint arXiv:2309.04269*, 2023.

55 PAPERNO, D. et al. The LAMBADA dataset: Word prediction requiring a broad discourse context. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 2016.

56 CHANG, Y. et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, v. 15, n. 3, 2024.

57 SHAO, Y.; JIANG, Y.; KANELL, T. A.; XU, P.; KHATTAB, O.; LAM, M. S. Assisting in writing Wikipedia-like articles from scratch with large language models. In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2024)*. [S.l.: s.n.], 2024. p. 6252–6278.

58 XI, Z.; YIN, W.; FANG, J.; WU, J.; FANG, R.; JIANG, Y.; XIE, P.; HUANG, F.; CHEN, H.; ZHANG, N. OmniThink: Expanding knowledge boundaries in machine writing through thinking. In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025)*. [S.l.: s.n.], 2025. p. 956–976.

59 ASAI, A. et al. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In: *International Conference on Learning Representations (ICLR)*. [S.l.: s.n.], 2024.

60 NATO Standardization Office. *Allied Joint Doctrine for Intelligence, Counter-Intelligence and Security (AJP-2.1)*. Brussels, Belgium, 2016.

61 LOWENTHAL, M. M. *Intelligence: From Secrets to Policy*. 8th. ed. Thousand Oaks, CA: CQ Press, 2019.

62 INAN, H. et al. Llama Guard: LLM-based input-output safeguard for human-AI conversations. *arXiv preprint arXiv:2312.06674*, 2024.

63 ONG, D. et al. RouteLLM: Learning to route LLMs with preference data. In: *Proceedings of the 41st International Conference on Machine Learning (ICML)*. [S.l.: s.n.], 2024.

64 WANG, X. et al. Executable code actions elicit better LLM agents. *arXiv preprint arXiv:2402.01030*, 2024.

65 JR., R. J. H.; PHERSON, R. H. *Structured Analytic Techniques for Intelligence Analysis*. 3. ed. Thousand Oaks: CQ Press, 2020.

66 LIU, X. et al. Controlled text generation for large language model with dynamic attribute graphs. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*. [S.l.: s.n.], 2024.

67 PACKER, C.; FANG, V.; PATIL, S. G.; GONZALEZ, J. E. et al. MemGPT: Towards LLMs as operating systems. In: *Proceedings of the International Conference on Learning Representations (ICLR)*. [S.l.: s.n.], 2024. ArXiv:2310.08560.

68 PARK, J. S.; O'BRIEN, J. C.; CAI, C. J.; MORRIS, M. R.; LIANG, P.; BERNSTEIN, M. S. Generative agents: Interactive simulacra of human behavior. In: *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*. [S.l.: s.n.], 2023. ArXiv:2304.03442.

69 SHINN, N.; LABASH, B.; GOPINATH, A. et al. Reflexion: Language agents with verbal reinforcement learning. In: *Advances in Neural Information Processing Systems*. [S.l.: s.n.], 2024. v. 36. Publicado originalmente como arXiv preprint em 2023 (arXiv:2303.11366).

70 NVIDIA. *Llama-3.3-Nemotron-70B-Instruct*. Santa Clara: [s.n.], 2025. Acesso em: 1 set. 2026. Disponível em: <https://huggingface.co/nvidia/Llama-3.3-Nemotron-70B-Instruct>.

71 Meta AI. *The Llama 3.1 Collection*. 2024. Preprint. Acesso em: 1 set. 2026. Disponível em: <https://arxiv.org/abs/2407.21783>.

72 Technology Innovation Institute. *The Falcon Series of Open Language Models*. 2023. Preprint. Acesso em: 1 set. 2026. Disponível em: <https://arxiv.org/abs/2311.16867>.

73 Alibaba Cloud. *Qwen3 Technical Report*. 2026. Preprint. Acesso em: 1 set. 2026. Disponível em: <https://qwenlm.github.io/blog/qwen3/>.

74 Google. *Gemma 2: Improving Open Models*. 2024. Preprint. Acesso em: 1 set. 2026. Disponível em: <https://arxiv.org/abs/2408.00118>.

75 Microsoft. *Phi-4 Reasoning-Vision*. Redmond: [s.n.], 2026. Acesso em: 1 set. 2026. Disponível em: <https://www.microsoft.com/en-us/research/publication/phi-4/>.

76 Mistral AI; NVIDIA. *Mistral-NeMo Technical Report*. Paris; Santa Clara: [s.n.], 2024. Acesso em: 1 set. 2026. Disponível em: <https://mistral.ai/news/mistral-nemo/>.

77 Technology Innovation Institute. *Falcon3-10B-Instruct*. Abu Dhabi: [s.n.], 2024. Acesso em: 1 set. 2026. Disponível em: <https://huggingface.co/tiiuae/falcon-3-10b-instruct>.

78 DeepSeek AI. *DeepSeek-R1 Technical Report*. 2025. Preprint. Acesso em: 1 set. 2026. Disponível em: <https://arxiv.org/abs/2501.12948>.

79 NVIDIA. *OpenReasoning-Nemotron-7B*. Santa Clara: [s.n.], 2025. Acesso em: 1 set. 2026. Disponível em: <https://huggingface.co/nvidia/OpenReasoning-Nemotron-7B>.

80 RWKV Project. *Eagle 7B*. 2025. Acesso em: 1 set. 2026. Disponível em: <https://huggingface.co/RWKV/v5-Eagle-7B>.

81 Allen Institute for AI. *Olmo Hybrid*. Seattle: [s.n.], 2026. Acesso em: 1 set. 2026. Disponível em: <https://allenai.org/olmo>.

82 NVIDIA. *Nemotron-Mini-4B-Instruct*. Santa Clara: [s.n.], 2024. Acesso em: 1 set. 2026. Disponível em: <https://huggingface.co/nvidia/Nemotron-Mini-4B-Instruct>.

83 Meta AI. *Llama 3.2: Edge and Mobile Models*. 2024. Preprint. Acesso em: 1 set. 2026. Disponível em: <https://ai.meta.com/research/publications/llama-3-2-edge-and-mobile-models/>.

84 OpenAI. *GPT-5.1 Technical Report*. San Francisco: [s.n.], 2026. Acesso em: 1 set. 2026. Disponível em: <https://cdn.openai.com/papers/gpt-5-1-technical-report.pdf>.

85 Anthropic. *The Claude 4.5 Model Family*. San Francisco: [s.n.], 2026. Acesso em: 1 set. 2026. Disponível em: <https://www.anthropic.com/research/claude-4-5-technical-report>.

86 Google DeepMind. *Gemini 3: Next Generation Multimodal Models*. Londres: [s.n.], 2026. Acesso em: 1 set. 2026. Disponível em: <https://storage.googleapis.com/deepmind-media/gemini/gemini_3_report.pdf>.

87 SCHWARTZ, R.; DODGE, J.; SMITH, N. A.; ETZIONI, O. Green AI. *Communications of the ACM*, v. 63, n. 12, p. 54–63, 2020.

88 OpenAI. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*, 2023. Disponível em: <https://arxiv.org/abs/2303.08774>.

89 LEE, J. D.; SEE, K. A. Trust in automation: Designing for appropriate reliance. *Human Factors*, v. 46, n. 1, p. 50–80, 2004.

APÊNDICE A

RELATÓRIO DE INTELIGÊNCIA COMPETITIVA E FINANCEIRA: ANÁLISE ESTRATÉGICA DO BANCO DO BRASIL (JUNHO DE 2025)

Nota: Este relatório foi gerado de forma autônoma pelo pipeline de IA desenvolvido nesta pesquisa (Pipeline UnB), em ambiente computacional isolado, em junho de 2025.

A.1 INTRODUÇÃO

Este relatório visa fornecer uma análise estratégica atualizada do Banco do Brasil S.A. (BB), com foco em seu desempenho financeiro recente, posicionamento competitivo, iniciativas estratégicas e perspectivas futuras. A análise baseia-se em informações públicas divulgadas pelo banco, notícias de mercado, relatórios de analistas e dados do setor financeiro, abrangendo o período até o primeiro trimestre de 2025 (1T25) e informações relevantes subsequentes.

A.2 SUMÁRIO EXECUTIVO

O Banco do Brasil apresentou um lucro líquido ajustado de R$ 7,4 bilhões no 1T25, com um Retorno sobre o Patrimônio Líquido (RSPL) de 16,7%. A carteira de crédito expandida demonstrou crescimento robusto, impulsionada por todos os segmentos, embora o custo do crédito tenha sido pressionado, em parte, pela inadimplência no setor do agronegócio. A instituição continua a ser reconhecida por sua liderança em sustentabilidade e mantém uma sólida posição de capital. As ações BBAS3 refletem um sentimento de mercado cauteloso, com a cotação atual próxima das mínimas de 52 semanas, apesar de um preço-alvo médio de analistas que sugere um potencial de valorização. O principal desafio de curto e médio prazo reside na gestão do risco de crédito da carteira de agronegócio, que tem apresentado sinais de deterioração e preocupa analistas. Estrategicamente, o Banco do Brasil avança em sua transformação digital, com foco no desenvolvimento de seu Super App e na exploração de novas linhas de crédito, como o bem-sucedido "Crédito ao Trabalhador". A parceria com a Advocacia-Geral da União (AGU) para aprimorar a gestão de riscos jurídicos e a participação em reestruturações de dívida, como a da Casas Bahia, demonstram uma abordagem proativa na mitigação de riscos e na busca por soluções financeiras. O compromisso com a agenda ESG permanece central, com metas ambiciosas e reconhecimento internacional. O cenário competitivo é intenso, com o Banco do Brasil disputando participação de mercado com grandes bancos privados e públicos. Embora mantenha uma posição de destaque em ativos e lucros, a competitividade em produtos específicos, como crédito pessoal, pode ser um ponto de atenção. O futuro do banco dependerá

de sua capacidade de navegar pela volatilidade macroeconômica, gerenciar eficazmente os riscos de sua carteira de crédito e capitalizar as oportunidades em finanças sustentáveis e inovação digital.

A.3 VISÃO GERAL DA COMPANHIA

Fundado em 1808, o Banco do Brasil S.A. é uma das instituições financeiras mais antigas e proeminentes do Brasil, desempenhando um papel crucial no desenvolvimento econômico e social do país. Como uma sociedade de economia mista, com o governo federal como seu acionista controlador, o BB combina objetivos comerciais com o fomento de políticas públicas, especialmente no setor agrícola. Sua missão transcende a oferta de produtos e serviços bancários tradicionais, englobando o apoio a diversos setores da economia, a promoção da inclusão financeira e o desenvolvimento sustentável. O banco possui uma vasta rede de agências físicas distribuídas por todo o território nacional, complementada por uma crescente presença digital através de seu aplicativo e plataformas online. No mercado brasileiro, o Banco do Brasil consistentemente figura entre os líderes em termos de ativos totais e volume de crédito, detendo uma participação de mercado significativa em diversos segmentos, com destaque para o financiamento do agronegócio. Sua atuação abrange desde o pequeno produtor rural e o microempreendedor até grandes corporações e o setor público.

A.4 ANÁLISE DO DESEMPENHO FINANCEIRO (FOCO 1T25)

No primeiro trimestre de 2025 (1T25), o Banco do Brasil reportou um lucro líquido ajustado de R$ 7,4 bilhões, resultando em um Retorno sobre o Patrimônio Líquido (RSPL) de 16,7%. Este desempenho reflete a contínua expansão de suas operações de crédito e a gestão de suas despesas, embora também evidencie desafios no custo do crédito. A carteira de crédito expandida atingiu um crescimento de 14,4% em doze meses e 1,1% no trimestre. O segmento de Pessoa Física alcançou R$ 335,8 bilhões, um aumento anual de 6,6%, impulsionado pelo crédito consignado e não consignado, incluindo o novo produto "Crédito ao Trabalhador". A carteira de Pessoa Jurídica cresceu 22,4% em um ano, totalizando R$ 459,9 bilhões, com R$ 141,3 bilhões para Grandes Empresas, R$ 123,8 bilhões para MPME e R$ 74,6 bilhões para o setor Governo. O agronegócio, um pilar para o banco, viu sua carteira crescer 9,0% anualmente, alcançando R$ 406,2 bilhões, com R$ 152,5 bilhões desembolsados nos nove meses do plano safra 24/25 (julho/24 a março/25). Adicionalmente, a carteira de crédito sustentável encerrou março de 2025 com R$ 393,5 bilhões, um aumento de 9,6% em 12 meses. A margem financeira bruta totalizou R$ 23,9 bilhões, influenciada pela elevação da taxa Selic e pelo aumento dos volumes de ativos e passivos. No entanto, o custo de crédito atingiu R$ 10,2 bilhões, refletindo principalmente a dinâmica agravada da carteira de agronegócios, cuja inadimplência alcançou 3,04%. Este é um ponto de atenção significativo, como será detalhado na seção de riscos. As receitas de prestação de serviços finalizaram o trimestre em R$ 8,4 bilhões, com destaque para administração de fundos e consórcios. As despesas administrativas somaram R$ 9,5 bilhões, evidenciando controle de gastos e investimentos em tecnologia. A solidez do balanço foi corroborada pelo Capital Principal, que encerrou março de 2025 em 10,97%. É importante notar que a entrada em vigor

da Resolução CMN 4.966/2021 em janeiro de 2025, que alterou a contabilização de ativos financeiros e a provisão para perdas esperadas, compromete a comparabilidade direta de alguns indicadores com períodos anteriores, especialmente na margem financeira bruta, custo do crédito e receitas de prestação de serviços. O Banco do Brasil revisou suas projeções para Margem Financeira Bruta, Custo de Crédito e Lucro Líquido Ajustado, mantendo inalteradas as projeções para os demais indicadores de desempenho de crédito e itens comerciais e operacionais.

Tabela A.1: Destaques financeiros do Banco do Brasil — 1T25.

Indicador	**Valor (1T25)**	**Variação Anual (%)**	**Variação Trimestral (%)**
Lucro Líquido Ajustado	R$ 7,4 bilhões	N/D	N/D
RSPL (Retorno S/ Patrimônio Líquido)	16,7%	N/D	N/D
Carteira de Crédito Expandida	N/D (Cresc. 14,4%)	14,4%	1,1%
Pessoa Física	R$ 335,8 bilhões	6,6%	1,2%
Pessoa Jurídica	R$ 459,9 bilhões	22,4%	1,6%
Agronegócios	R$ 406,2 bilhões	9,0%	N/D
Crédito Sustentável	R$ 393,5 bilhões	9,6%	1,8%
Margem Financeira Bruta	R$ 23,9 bilhões	N/D	N/D
Custo de Crédito	R$ 10,2 bilhões	N/D	N/D
Inadimplência Agronegócios	3,04%	N/D	N/D
Receita de Prestação de Serviços	R$ 8,4 bilhões	N/D	N/D
Despesas Administrativas	R$ 9,5 bilhões	N/D	N/D
Capital Principal	10,97%	N/D	N/D

Nota: N/D: não divulgado diretamente no resumo do resultado para variações específicas, mas os valores absolutos e crescimentos percentuais foram fornecidos.

O impacto da Resolução CMN 4.966/2021 merece atenção. Ao exigir uma abordagem prospectiva para perdas esperadas, a resolução pode levar a uma maior volatilidade no reconhecimento de provisões, especialmente em um cenário de incerteza econômica. A dificuldade de comparação com períodos anteriores exige uma análise mais cautelosa das tendências de rentabilidade e qualidade de ativos.

A.5 DESEMPENHO DAS AÇÕES E AVALIAÇÃO (BBAS3)

As ações ordinárias do Banco do Brasil (BBAS3) são negociadas na B3 e compõem o Ibovespa, principal índice do mercado de capitais brasileiro. Em dados recentes, a cotação da BBAS3 estava em R$ 22,03, com o ativo negociando próximo à sua mínima das últimas 52 semanas, que variou entre R$ 21,97 e R$ 30,04. Em outra consulta, o valor era de R$ 22,04, com uma variação de 10,09% nos últimos 12 meses e 5,46% no mês corrente até a data da consulta. Este desempenho sugere um período de pressão sobre o papel, refletindo, possivelmente, as preocupações do mercado com o cenário macroeconômico e os riscos específicos da carteira do banco, como os do agronegócio. O volume médio de negociação em 3 meses era de 25,72 milhões de ações, indicando boa liquidez para o papel. Um ponto técnico relevante é o Índice de Força Relativa (IFR ou RSI de 14 períodos) em 24,23, o que sugere que a ação pode estar em território de

sobrevenda, podendo indicar uma potencial reversão de curto prazo ou, alternativamente, um pessimismo persistente. Apesar da performance recente da cotação, o consenso de analistas para as ações BBAS3 é "Neutro", com base em 13 avaliações nos últimos 3 meses. O preço-alvo médio projetado por analistas para os próximos 12 meses é de R$ 31,09, o que representaria um potencial de valorização de aproximadamente 39,80% a 41,14% em relação aos preços atuais observados. As estimativas de preço-alvo variam, com uma mínima de R$ 24,00 e uma máxima de R$ 45,00. Estas projeções são derivadas de métricas como lucro por ação, receita, médias móveis e indicadores técnicos. O Banco do Brasil tem um histórico de aumento de dividendos por 4 anos consecutivos. A política de remuneração aos acionistas, incluindo dividendos e Juros sobre Capital Próprio (JCP), é um fator atrativo para investidores focados em renda.

Tabela A.2: Panorama da ação BBAS3.

Indicador	**Valor/Status**
Cotação Atual (aprox.)	R$ 22,03 – R$ 22,04
Variação em 52 Semanas	R$ 21,97 – R$ 30,04
Variação em 12 Meses	10,09% a 11,7%
Volume Médio (3m)	25,72M
RSI (14)	24,23 (Sobrevenda)
Consenso de Analistas (últimos 3 meses)	Neutro (13 avaliações)
Preço-Alvo Médio (12 meses)	R$ 31,09
Potencial de Valorização (vs. alvo)	+39,80% a +41,14%
Histórico de Dividendos	Aumento por 4 anos consecutivos

A divergência entre a cotação atual, próxima das mínimas, e o preço-alvo dos analistas sugere que o mercado pode estar precificando riscos de forma mais severa do que alguns modelos de avaliação, ou que há uma expectativa de melhora nos fundamentos do banco que ainda não se materializou nos preços. A resolução das incertezas, principalmente em relação à carteira de agronegócio e ao impacto da nova resolução contábil, será crucial para a reavaliação do ativo pelo mercado.

A.6 CENÁRIO COMPETITIVO

O Banco do Brasil opera em um setor bancário brasileiro altamente concentrado e competitivo, onde disputa mercado com grandes conglomerados financeiros privados nacionais e internacionais, além de outras instituições financeiras públicas. Em 2023, o Banco do Brasil figurava como o segundo maior banco do país tanto em lucro anual (R$9,5 bilhões, atrás do Itaú com R$ 10,1 bilhões) quanto em ativos totais (R$2,028 trilhões, atrás do Itaú com R$ 2,469 trilhões). Seus principais concorrentes diretos incluem Itaú Unibanco, Bradesco e Santander, que, juntamente com o BB e a Caixa Econômica Federal, formam o grupo dos cinco maiores bancos do país. A Caixa, embora também pública, tem um foco mais direcionado para políticas habitacionais e sociais, mas compete em diversas linhas de produtos de varejo. Bancos de investimento como o BTG Pactual também representam uma concorrência crescente, especialmente em segmentos de maior renda e corporativo. Uma análise comparativa do 1T25, baseada nas expectativas de mercado, indicava um início de ano morno para o Banco do Brasil, com projeção de lucro líquido de R$9,5 bilhões (XP Investimentos) e R$ 9,32 bilhões (consenso LSEG), implicando um RSPL ao redor de

20,9%. Estas projeções colocavam o BB com um desempenho de rentabilidade robusto, potencialmente superior ao Bradesco (RSPL esperado de 13,0Um fator competitivo relevante são as taxas de juros praticadas. Em dados de 2023 para empréstimo pessoal, o Banco do Brasil apresentava uma taxa de 4,14

Tabela A.3: Métricas financeiras comparativas — Banco do Brasil *versus* principais pares (projeções/resultados 1T25 ou mais recente disponível).

Instituição	**Lucro Líquido (1T25 proj./resultado)**	**RSPL (1T25 proj./resultado)**	**Ativos Totais (ref. 2023)**
Banco do Brasil	R$ 9,32bi – R$ 9,5bi (Proj. LSEG/XP)	~20,9% (Proj. XP)	R$ 2,028 trilhões
Itaú Unibanco	R$ 11,0bi – R$ 11,1bi (Proj. LSEG/XP)	~22,3% (Proj. XP)	R$ 2,469 trilhões
Bradesco	R$ 5,42bi – R$ 5,46bi (Proj. Genial/LSEG)	~13,0% – 13,4% (Proj. XP/Genial)	R$ 1,779 trilhões
Santander	R$ 3,76bi – R$ 3,8bi (Proj. LSEG/Genial)	~16,2% – 16,4% (Proj. XP/BBI)	R$ 1,048 trilhões

Nota: projeções para 1T25; ativos totais referentes a dados de 2023.

Tabela A.4: Comparativo de taxas de juros de empréstimo pessoal (referência 2023).

Instituição	**Taxa de juros (empréstimo pessoal % a.m.)**
Banco do Brasil	4,14%
Itaú Unibanco	1,50%
Bradesco	5,89%
Caixa Econômica Federal	1,99%
Santander Brasil	3,27%

A competição no setor bancário brasileiro também é acirrada pela ascensão das fintechs, que desafiam os modelos tradicionais com soluções digitais ágeis e custos operacionais mais baixos. A resposta do Banco do Brasil, através de investimentos em tecnologia, digitalização de serviços e o desenvolvimento de um "Super App", é crucial para manter sua relevância e participação de mercado a longo prazo.

A.7 AVALIAÇÃO DE RISCOS E MITIGAÇÃO

O Banco do Brasil está exposto a diversos riscos inerentes à sua atividade. Uma gestão eficaz desses riscos é fundamental para sua estabilidade financeira e desempenho sustentado.

A.7.1 Risco de Crédito

O risco de crédito é atualmente a principal preocupação para o Banco do Brasil, especialmente em sua carteira de agronegócio. No 1T25, a inadimplência neste segmento atingiu 3,04A estrutura de muitos desses contratos, com pagamentos concentrados no final da safra ("bullet payments"), cria um risco elevado de inadimplência em períodos de colheita, tornando os meses de abril, maio e junho críticos para a avaliação da real extensão do problema. A Mantaro Capital sugere que o mercado pode estar subestimando essa "tempestade perfeita", pois a prorrogação de dívidas, mesmo com autorização governamental, indica dificuldades financeiras dos produtores rurais. A percepção é que os produtores que prorrogam suas dívi-

das não estão financeiramente saudáveis, e o pico de vencimentos desses pagamentos concentrados pode revelar uma inadimplência maior do que a atualmente reportada. O fato de esta carteira prorrogada não ter começado a diminuir e já estar se aproximando de seu pico novamente é um sinal de alerta. A última vez que essa carteira diminuiu, houve um repique na inadimplência. Este cenário configura uma potencial contingência que pode impactar significativamente os resultados para além do 1T25, caso se materialize. A implementação da Resolução CMN 4.966/2021, que altera a forma de provisionamento para perdas esperadas, também adiciona uma camada de complexidade à análise da qualidade da carteira de crédito e à comparabilidade dos dados ao longo do tempo.

A.7.2 Risco de Mercado, Operacional e Regulatório

O banco está exposto a flutuações nas taxas de juros (como a Selic), que afetam sua margem financeira, e a variações cambiais, dado seu envolvimento em operações internacionais. Riscos operacionais, incluindo falhas em sistemas, processos internos e fraudes, são inerentes a uma instituição de grande porte. A crescente digitalização das operações bancárias também eleva a exposição a riscos cibernéticos, um ponto que a parceria com a AGU busca endereçar, entre outros, ao contemplar a cibersegurança. Riscos regulatórios são uma constante no setor financeiro, exigindo adaptação contínua a novas normas e exigências dos órgãos fiscalizadores.

A.7.3 Abordagem da Administração para Mitigação de Riscos

A gestão do Banco do Brasil demonstra uma abordagem multifacetada para a mitigação de riscos. A diversificação da carteira de crédito é uma estratégia chave, exemplificada pelo lançamento e expansão do "Crédito ao Trabalhador", que já atingiu R$ 3,0 bilhões em desembolsos desde março de 2025, visando um público com menor risco percebido (servidores públicos, aposentados, pensionistas e, mais recentemente, trabalhadores da iniciativa privada). No âmbito corporativo, o banco participa ativamente da reestruturação de dívidas de grandes clientes, como no caso da Casas Bahia, onde, juntamente com o Bradesco, planeja converter R$ 1,5 bilhão da dívida da varejista em ações. Esta operação, se aprovada, reduziria significativamente a alavancagem da Casas Bahia e geraria uma economia anual de R$ 230 milhões em despesas financeiras para a empresa, mitigando o risco de crédito para os bancos credores. A parceria estratégica firmada com a Advocacia-Geral da União (AGU) em junho de 2025 visa o intercâmbio de soluções jurídicas e tecnológicas para reduzir a judicialização, valorizar métodos consensuais de resolução de conflitos e combater a litigância abusiva. A cooperação abrange áreas como jurimetria, inteligência jurídica e recuperação de ativos, o que pode otimizar a gestão de riscos legais e reduzir custos processuais. Internamente, a estrutura executiva do banco inclui uma Vice-Presidência de Controles Internos e Gestão de Riscos, liderada por Felipe Guimarães Geissler Prince, responsável por áreas como crédito, riscos, controles internos, compliance, segurança institucional e cibernética, e prevenção a fraudes. Esta estrutura dedicada sinaliza o foco da instituição na supervisão e controle dos riscos. Embora estas iniciativas demonstrem um arcabouço de gestão de riscos sofisticado e adaptativo, a eficácia dessas medidas frente à magnitude do desafio representado pela carteira de agronegócio ainda precisa ser comprovada nos próximos trimestres. A capacidade do banco de navegar por este período crítico, equilibrando o apoio a um setor vital para a economia

brasileira com a prudência na gestão de seus ativos, será determinante para seu desempenho futuro.

A.8 ESTRATÉGIA E DESEMPENHO EM SUSTENTABILIDADE (ESG)

O Banco do Brasil consolidou uma posição de destaque global em sustentabilidade, integrando princípios ambientais, sociais e de governança (ESG) em sua estratégia de negócios e operações.

A.8.1 Compromissos e Estrutura ESG

A estratégia de sustentabilidade do banco é guiada pela "Agenda 30 BB", um plano revisado bienalmente e alinhado aos Objetivos de Desenvolvimento Sustentável (ODS) da ONU. Este plano estrutura-se em cinco pilares: Pessoas, Planeta, Prosperidade, Paz e Parceria. O reconhecimento internacional é notável: pela sexta vez, o Banco do Brasil foi nomeado o banco mais sustentável do mundo pelo ranking das 100 Corporações Mais Sustentáveis do Mundo 2025 da Corporate Knights. Além disso, integra o Índice de Sustentabilidade Empresarial (ISE) da B3 por 18 anos consecutivos e o FTSE4Good Index Series da Bolsa de Londres. Tais representam um ativo intangível significativo, capaz de atrair investidores com foco em ESG, fortalecer a reputação da marca e potencialmente reduzir o custo de captação de recursos através de financiamentos atrelados à sustentabilidade.

A.8.2 Iniciativas Ambientais

A carteira de crédito sustentável do BB atingiu R$ 393,5 bilhões em março de 2025, um crescimento de 9,6% em 12 meses. As metas para 2030 são ambiciosas:

- Alcançar um saldo de R$ 500 bilhões na carteira de crédito sustentável.
- Destinar R$ 30 bilhões para projetos de energia renovável.
- Atingir R$ 200 bilhões em financiamentos para agricultura sustentável.
- Originar R$ 100 bilhões em recursos sustentáveis para o BB e seus clientes.

Em conservação florestal, a meta é alcançar 1 milhão de hectares conservados e/ou reflorestados até 2025, com um compromisso de desmatamento ilegal zero nos financiamentos. Nas ações climáticas, o banco visa compensar 100% das emissões de escopo 1 e 2, utilizar 100% de energia renovável a partir de 2023 (meta já em curso) e reduzir em 42% as emissões diretas (escopo 1) até 2030.

A.8.3 Iniciativas Sociais

O BB promove a inclusão e a diversidade com metas claras, como alcançar 30

A.8.4 Governança

O Banco do Brasil é signatário dos Princípios para o Investimento Responsável (PRI) desde 2010 e desenvolveu uma metodologia própria de avaliação de crédito e de ações de empresas que incorpora critérios ASG. O banco também possui um Framework de Finanças Sustentáveis e um Framework de Dívidas Vinculadas à Sustentabilidade (SLFF BB), que alinha a captação de recursos com metas corporativas de sustentabilidade.

A.8.5 Relatórios Chave

O banco disponibiliza publicamente relatórios sobre sua performance ESG, incluindo o Relatório de Riscos e Oportunidades Sociais, Ambientais e Climáticas, cuja versão mais recente (2024) foi disponibilizada. A consulta a este relatório específico não confirmou informações sobre um investimento de R$ 53 bilhões na Amazônia, conforme questionado na solicitação original; tal dado não foi encontrado nos materiais analisados. As metas ESG quantificadas e ambiciosas, especialmente em finanças sustentáveis, indicam que a sustentabilidade é uma estratégia central de negócios para o Banco do Brasil, e não apenas uma iniciativa de responsabilidade social corporativa. Isso posiciona o banco para capitalizar a transição para uma economia verde, mas também atrela parte de seu desempenho ao sucesso e à escalabilidade desses setores sustentáveis, exigindo uma gestão de risco igualmente robusta para essas novas carteiras.

Tabela A.5: Resumo das principais metas ESG e progresso reportado (base 1T25/2024).

Meta ESG	**Objetivo (prazo)**	**Progresso reportado (data)**
Carteira de Crédito Sustentável	R$ 500 bilhões (2030)	R$ 393,5 bilhões (Mar/25)
Saldo em Energia Renovável	R$ 30 bilhões (2030)	N/D
Saldo em Agricultura Sustentável	R$ 200 bilhões (2030)	N/D
Originação de Recursos Sustentáveis	R$ 100 bilhões (2030)	N/D
Conservação/Reflorestamento	1 milhão de hectares (2025)	N/D
Redução de Emissões Diretas (Escopo 1)	42% (2030 vs. base N/D)	N/D
Uso de Energia Renovável	100% (a partir de 2023)	Em curso
Diversidade (Etnias Sub-representadas em Liderança)	30% (2025)	N/D
Inclusão Financeira (Renegociação Dívidas Baixa Renda)	2,5 milhões de clientes (2025)	N/D
Inclusão Financeira (Crédito a Empreendedores)	1 milhão de empreendedores (2025)	N/D
Investimento Social Privado (Fundação BB)	R$ 1 bilhão (2030)	N/D
Clientes com Maturidade Digital	17 milhões (2025)	N/D

Nota: N/D: não divulgado nos documentos consultados para o progresso específico da meta.

A.9 GOVERNANÇA CORPORATIVA

A estrutura de governança corporativa do Banco do Brasil busca alinhar os interesses dos acionistas, da administração e demais partes interessadas, seguindo as melhores práticas de mercado e as exigências regulatórias para sociedades de economia mista.

A.9.1 Conselho de Administração

O Conselho de Administração é o órgão de decisão colegiada responsável por fixar a orientação geral dos negócios do Banco, desempenhando atribuições estratégicas, eletivas e de fiscalização, sem abranger funções operacionais ou executivas. É composto por oito membros eleitos pela Assembleia Geral. Conforme atualização de 26 de maio de 2025, a composição do Conselho inclui:

- **Anelize Lenzi Ruas de Almeida:** Presidente (membro não independente, indicada pelo acionista controlador). Procuradora da Fazenda Nacional.
- **Elisa Vieira Leonel:** Vice-Presidente (membro não independente, indicada pelo acionista controlador). Secretária de Coordenação e Governança das Empresas Estatais.
- **Tarciana Paula Gomes Medeiros:** Conselheira (membro não independente, indicada pelo acionista controlador) e Presidente do Banco do Brasil.
- **Fábio Franco Barbosa Fernandes:** Conselheiro (membro não independente, indicado pelo acionista controlador). Auditor Fiscal da Receita Federal.
- **Márcio Luiz de Albuquerque Oliveira:** Conselheiro (membro não independente, indicado pelo acionista controlador). Economista e Analista de Planejamento e Orçamento Federal.
- **Fernando Florêncio Campos:** Conselheiro (membro independente, indicado pelos acionistas minoritários). Ex-funcionário do BB com vasta experiência executiva.
- **Valmir Pedro Rossi:** Conselheiro (membro independente, indicado pelos acionistas minoritários). Ex-funcionário do BB e ex-Presidente do Banco da Amazônia.
- **Selma Cristina Alves Siqueira:** Conselheira (membro não independente, eleita pelos empregados do Banco do Brasil). Funcionária de carreira do BB.

A presença de membros indicados pelo acionista controlador, pelos acionistas minoritários e pelos empregados busca promover um equilíbrio de interesses e perspectivas na tomada de decisões estratégicas. A experiência combinada de conselheiros com longas carreiras no próprio Banco do Brasil ou em outras instituições do setor público assegura um profundo conhecimento institucional. No entanto, para um setor em rápida evolução tecnológica e competitiva, é vital que este conhecimento seja continuamente complementado por visões externas e diversificadas para evitar uma cultura excessivamente insular.

A.9.2 Diretoria Executiva (Conselho Diretor e Diretores)

A Diretoria Executiva é o órgão colegiado responsável pela administração cotidiana do Banco. É composta pelo Conselho Diretor e por até vinte e sete Diretores. O Conselho Diretor, por sua vez, é formado pela Presidente do Banco e por até nove Vice-Presidentes. Conforme atualização de 25 de março de 2025, o Conselho Diretor inclui:

- **Tarciana Paula Gomes Medeiros:** Presidente.

- **Ana Cristina Rosa Garcia:** Vice-Presidente Corporativo.
- **Carla Nesi:** Vice-Presidente de Negócios de Varejo.
- **Felipe Guimarães Geissler Prince:** Vice-Presidente de Controles Internos e Gestão de Riscos.
- **Francisco Augusto Lassalvia:** Vice-Presidente de Negócios de Atacado.
- **José Ricardo Sasseron:** Vice-Presidente de Negócios Governo e Sustentabilidade Empresarial.
- **Luiz Gustavo Braz Lage:** Vice-Presidente de Agronegócios e Agricultura Familiar.
- **Marco Geovanne Tobias da Silva:** Vice-Presidente de Gestão Financeira e de Relações com Investidores.
- **Marisa Reghini Ferreira Mattos:** Vice-Presidente de Negócios Digitais e Tecnologia.

A clara delimitação de funções entre o Conselho de Administração (foco estratégico e fiscalizatório) e a Diretoria Executiva (foco na administração) está em consonância com as boas práticas de governança corporativa. A experiência dos executivos em diversas áreas do banco e do mercado financeiro é um ponto forte, mas, similarmente ao Conselho de Administração, o equilíbrio entre experiência interna e novas perspectivas é crucial para a adaptação contínua da instituição. Não foram identificadas, nos documentos consultados, mudanças significativas na composição dos conselhos nos últimos 12 meses que não estejam refletidas nas datas de atualização mencionadas.

A.10 PERSPECTIVAS ESTRATÉGICAS, DESAFIOS E OPORTUNIDADES

O Banco do Brasil navega em um ambiente complexo, caracterizado por desafios significativos e oportunidades emergentes que moldarão seu desempenho futuro.

A.10.1 Perspectivas de Curto e Médio Prazo

As projeções revisadas pelo próprio banco para Margem Financeira Bruta, Custo de Crédito e Lucro Líquido Ajustado após o 1T25 indicam uma reavaliação do cenário operacional. O consenso de analistas para as ações BBAS3, embora "Neutro", aponta para um preço-alvo médio com potencial de valorização, sugerindo que o mercado antecipa uma melhora ou considera o papel subavaliado. A continuidade da transformação digital e o foco na agenda ESG são vetores estratégicos que devem persistir.

A.10.2 Desafios Principais

- **Inadimplência no Agronegócio:** Este é o desafio mais premente. A gestão e resolução da inadimplência crescente na carteira de agronegócio (3,04% no 1T25) e as preocupações levantadas sobre a carteira de dívidas prorrogadas, que podem mascarar um risco ainda maior, exigirão atenção máxima.

- **Volatilidade Macroeconômica:** Flutuações nas taxas de juros (Selic), níveis de inflação e o ritmo de crescimento econômico do Brasil impactam diretamente a demanda por crédito, a capacidade de pagamento dos clientes e a rentabilidade das operações bancárias.
- **Pressões Competitivas:** A concorrência com grandes bancos tradicionais permanece acirrada, enquanto fintechs e novos entrantes digitais desafiam os modelos de negócios estabelecidos, exigindo inovação contínua e agilidade.
- **Adaptação Regulatória:** A necessidade de se ajustar a novas regulamentações, como a Resolução CMN 4.966/2021 sobre provisionamento, impõe custos de adaptação e pode afetar a comparabilidade dos resultados financeiros.
- **Adoção e Monetização Digital:** Garantir a ampla adoção de seus canais digitais e, crucialmente, a monetização eficaz das funcionalidades oferecidas, como as do Super App, é vital para competir e melhorar a eficiência.

A.10.3 Oportunidades de Crescimento

- **Banca Digital e Super App:** A expansão dos serviços digitais, o aproveitamento das possibilidades do Open Finance e o enriquecimento do Super App com novas funcionalidades e parcerias podem atrair e reter clientes, além de gerar novas fontes de receita. O desenvolvimento de um Super App robusto é uma resposta direta ao cenário de "Revolução no front-end"descrito pela PwC, onde novos players buscam controlar a interface com o cliente.
- **Finanças Sustentáveis:** A liderança consolidada em ESG e a crescente demanda global por financiamentos verdes e sociais colocam o BB em uma posição privilegiada para expandir sua carteira de crédito sustentável e atrair capital de investidores focados em sustentabilidade.
- **Novas Linhas de Crédito:** O sucesso inicial do "Crédito ao Trabalhador", que atingiu R$ 3 bilhões em desembolsos rapidamente, demonstra o potencial para identificar e explorar nichos de mercado e segmentos de clientes com perfis de risco atraentes, diversificando as fontes de receita.
- **Ganhos de Eficiência:** A adoção de novas tecnologias, como inteligência artificial (IA) e análise de dados, pode otimizar processos, reduzir custos operacionais e aprimorar a tomada de decisões. A parceria com a AGU, por exemplo, explora o uso de IA para gestão jurídica.

A.10.4 Tendências do Setor Bancário (Cenários PwC)

Os cenários hipotéticos da PwC para o setor bancário na próxima década oferecem um pano de fundo para a estratégia do BB. O investimento em plataformas digitais próprias (Super App) é uma tentativa de evitar a perda do relacionamento com o cliente ("Revolução no front-end"). Sua escala e importância sistêmica o posicionam como um potencial consolidador ou sobrevivente em um cenário de "Winner takes all". Sua natureza estatal pode ser uma vantagem em cenários de "Regulamentação Intensa"ou um fator de estabilidade em uma "Paisagem Dispersa", mas também pode impor limitações de agilidade. A ascensão de Moedas Digitais de Bancos Centrais (CBDCs) representa uma transformação mais fundamental, cujo

impacto sobre o modelo de negócios tradicional dos bancos ainda está se desenhando. O futuro do Banco do Brasil dependerá de sua habilidade em equilibrar suas forças tradicionais – como a capilaridade, a forte presença no agronegócio e o relacionamento com o setor público – com a necessidade premente de inovar, digitalizar e gerenciar proativamente os novos e antigos riscos. A crise na carteira de agronegócio funciona como um teste de estresse para essa capacidade de balanceamento. Superar este desafio enquanto avança em suas frentes de ESG e digital será crucial para seu crescimento sustentado.

A.11 RECOMENDAÇÕES ESTRATÉGICAS

Com base na análise detalhada do desempenho financeiro, posicionamento competitivo, riscos e oportunidades do Banco do Brasil, as seguintes recomendações estratégicas são propostas para seus stakeholders:

- **Gestão de Risco de Crédito no Agronegócio:**
  - *Ação Imediata:* Implementar uma estratégia agressiva e transparente para o tratamento da inadimplência na carteira de agronegócio. Isso inclui a renegociação proativa de dívidas com produtores viáveis, o provisionamento adequado para perdas esperadas e, se necessário, a alienação de exposições de altíssimo risco.
  - *Revisão de Políticas:* Revisar e fortalecer os critérios de subscrição de crédito para o setor, incorporando análises de risco mais sofisticadas que considerem volatilidade climática, flutuações de preços de commodities e a saúde financeira individual dos tomadores, para além de garantias tradicionais.
  - *Comunicação:* Melhorar a comunicação com o mercado sobre a real dimensão dos riscos na carteira de agronegócio e as medidas adotadas para mitigá-los, visando reconstruir a confiança dos investidores.
- **Aceleração da Transformação Digital e Inovação:**
  - *Fortalecimento do Super App:* Continuar investindo no desenvolvimento do Super App, focando em propostas de valor únicas que o diferenciem da concorrência, como a integração de serviços não financeiros relevantes e a personalização da experiência do usuário através de IA e análise de dados.
  - *Exploração de Open Finance:* Capitalizar as oportunidades do Open Finance para criar novos produtos, serviços e modelos de negócios, possivelmente através de parcerias estratégicas com fintechs e outras empresas de tecnologia.
  - *Eficiência Operacional:* Intensificar o uso de tecnologias como IA, automação e análise de dados para otimizar processos internos, reduzir custos operacionais e melhorar a experiência do cliente.
- **Liderança em Finanças Sustentáveis (ESG):**

  - *Expansão de Produtos:* Alavancar a liderança em ESG para desenvolver e escalar produtos financeiros verdes e sociais inovadores, atendendo à crescente demanda de investidores e clientes por soluções sustentáveis.
  - *Mensuração de Impacto:* Aprimorar os sistemas de mensuração e reporte do impacto socioambiental das operações e financiamentos, aumentando a transparência e a credibilidade das iniciativas ESG.
  - *Engajamento:* Utilizar a forte reputação em ESG como um diferencial para atrair talentos, fidelizar clientes e acessar novas fontes de captação de recursos atreladas a metas de sustentabilidade.

- **Diversificação Estratégica da Carteira:**
  - *Novos Segmentos:* Continuar a identificar e expandir a atuação em segmentos de crédito com menor risco e potencial de crescimento, como o "Crédito ao Trabalhador"e outras linhas para pessoa física e MPMEs com perfis de risco controlados, a fim de reduzir a dependência de setores mais voláteis.
  - *Serviços e Seguros:* Fortalecer as receitas provenientes de serviços e da área de seguros, que tendem a ser mais estáveis e menos intensivas em capital do que as operações de crédito tradicionais.

- **Transparência e Governança:**
  - *Impactos Regulatórios:* Fornecer informações claras e detalhadas sobre o impacto de novas regulamentações, como a Resolução CMN 4.966/2021, nos resultados financeiros e na gestão de riscos.
  - *Diálogo com Investidores:* Manter um diálogo aberto e contínuo com a comunidade de investidores, abordando proativamente as preocupações do mercado e fornecendo uma visão clara da estratégia de longo prazo do banco.

A implementação eficaz destas recomendações pode auxiliar o Banco do Brasil a navegar pelos desafios atuais, capitalizar suas vantagens competitivas e fortalecer sua posição para um crescimento sustentável e rentável no futuro. A principal tarefa estratégica é gerenciar os riscos da carteira de agronegócio de forma decisiva, enquanto se investe vigorosamente nas comprovadas fortalezas em ESG e nas crescentes capacidades digitais.

A.12 REFERÊNCIAS

1. Imprensa. BB tem lucro líquido ajustado de R$ 7,4 bilhões no 1T25. Acessado em 6 jun. 2025. Disponível em: <https://www.bb.com.br/pbb/pagina-inicial/imprensa/n/68621/>

2. BBAS3 – Cotação Hoje. Ações Banco do Brasil ON – Investing.com. Acessado em 6 jun. 2025. Disponível em: <https://br.investing.com/equities/brasil-on>

3. BBAS3 – BANCO BRASIL ON: cotação e indicadores – Status Invest. Acessado em 6 jun. 2025. Disponível em: <https://statusinvest.com.br/acoes/bbas3>

4. Preço-Alvo BBAS3. Projeções Banco do Brasil ON – Investing.com. Acessado em 6 jun. 2025. Disponível em: <https://br.investing.com/equities/brasil-on-consensus-estimates>

5. Gestora vê riscos crescentes no BB (BBAS3) com inadimplência no agronegócio. Acessado em 6 jun. 2025. Disponível em: <https://www.infomoney.com.br/advisor/gestora-ve-riscos-crescentes-no-banco-do-brasil-com-inadimplencia-no-agronegocio/>

6. Bradesco e o Banco do Brasil vão converter dívida em ações e Casas Bahia deve reduzir alavancagem pela metade. Acessado em 6 jun. 2025. Disponível em: <https://neofeed.com.br/negocios/bradesco-e-o-banco-do-brasil-vao-converter-divida-em-acoes-e-casas-bahia-deve-reduzir-alavancagem-pela-metade/>

7. AGU e Banco do Brasil firmam parceria para troca de experiências e de tecnologias jurídicas. Acessado em 6 jun. 2025. Disponível em: <https://www.gov.br/agu/pt-br/comunicacao/noticias/agu-e-banco-do-brasil-firmam-parceria-para-troca-de-experiencias-e-de-tecnologias-juridicas>

8. Como o BB atua – Sustentabilidade – Banco do Brasil. Acessado em 6 jun. 2025. Disponível em: <https://www.bb.com.br/site/sustentabilidade/como-bb-atua/>

9. 10 maiores bancos do Brasil em 2025: veja quais as instituições. Acessado em 6 jun. 2025. Disponível em: <https://www.creditas.com/exponencial/10-maiores-bancos-do-brasil/>

10. Dividendos do Banco do Brasil (BBAS3): quem recebe e quando? – Riconnect. Acessado em 6 jun. 2025. Disponível em: <https://riconnect.rico.com.vc/analises/dividendos-do-banco-do-brasil-bbas3-quem-recebe-e-quando/>

11. Santander, Bradesco, Itaú e BB: quais dos bancões devem se sair melhor na temporada de resultados do 1T25. Acessado em 6 jun. 2025. Disponível em: <https://www.infomoney.com.br/mercados/santander-bradesco-itau-e-bb-quais-dos-bancoes-devem-se-sair-melhor-temporada-resultados-1t25/>

12. Conselho Diretor – Banco do Brasil RI. Acessado em 6 jun. 2025. Disponível em: <https://ri.bb.com.br/governanca-e-sustentabilidade/conselho-diretor/>

13. Sustentabilidade – Banco do Brasil RI. Acessado em 6 jun. 2025. Disponível em: <https://ri.bb.com.br/governanca-e-sustentabilidade/sustentabilidade/>

14. Conselho de Administração – Banco do Brasil RI. Acessado em 6 jun. 2025. Disponível em: <https://ri.bb.com.br/governanca-e-sustentabilidade/conselho-de-administracao/>

15. O que são 'super apps'? – Finsiders Brasil. Acessado em 6 jun. 2025. Disponível em: <https://finsidersbrasil.com.br/semregules/o-que-sao-super-apps/>

16. Open Finance: baixa adesão das empresas (PJs) entra no radar do Banco Central. Acessado em 6 jun. 2025. Disponível em: <https://convergenciadigital.com.br/governo/open-finance-baixa-adesao-das-empresas-pjs-entra-no-radar-do-banco-central/>

17. Banco Central, Susep e o avanço do Open Insurance – Fenacor. Acessado em 6 jun. 2025. Disponível em: <https://www.fenacor.org.br/noticias/banco-central-susep-e-o-avanco-do-open-insura>

18. Perspectivas do setor bancário a partir de 2025 – PwC. Acessado em 6 jun. 2025. Disponível em: <https://www.pwc.com.br/pt/estudos/setores-atividade/financeiro/2022/perspectivas-do-setor-bancario-a-partir-de-2025.html>