UNIVERSIDADE FEDERAL DO AMAZONAS - UFAM

INSTITUTO DE COMPUTAÇÃO- ICOMP

PROGRAMA PÓS-GRADUAÇÃO EM INFORMÁTICA - PPGI

Avaliação de Agentes de IA em Benchmarks de 2025

Matheus Serrão Uchoa

Manaus - AM

2025

Matheus Serrão Uchoa

Avaliação de Agentes de IA em Benchmarks de 2025

Monografia apresentada ao Instituto de Computação da Universidade Federal do Amazonas como requisito parcial para a obtenção do título de Bacharel em Ciência da Computação.

Orientador(a)

Prof. Dr. Nome do Orientador

Universidade Federal do Amazonas - UFAM

Instituto de Computação- IComp

Manaus - AM

2025

Monografia de Graduação sob o título <*Título da monografia*> apresentada por <Nome do aluno> e aceita pelo Instituto de Computação da Universidade Federal do Amazonas, sendo aprovada por todos os membros da banca examinadora abaixo especificada:

Titulação e nome do(a) orientador(a)
Orientador(a)
Departamento
Universidade

Titulação e nome do(a) membro da banca examinadora
Co-orientador(a), se houver
Departamento
Universidade

Titulação e nome do membro da banca examinadora
Departamento
Universidade

Titulação e nome do membro da banca examinadora
Departamento
Universidade

Manaus - AM, data de aprovação (por extenso).

Dedico este trabalho à minha família, que sempre acreditou nos meus planos, e aos amigos que dividiram cada descoberta sobre agentes de IA comigo.

Agradecimentos

À minha mãe e ao meu pai, que transformaram dúvidas em incentivo e estiveram presentes em todas as etapas deste percurso acadêmico. À minha avó, pelo cuidado paciente, pelas histórias que me lembram de onde vim e pelas leituras atentas das versões preliminares. Registro minha gratidão ao meu orientador, que ajudou a transformar um tema emergente em um plano concreto, e aos colegas do laboratório, que compartilharam benchmarks experimentais, scripts e muitos cafés durante a fase de implementação. Agradeço ainda ao Instituto de Computação da UFAM pelo ambiente colaborativo e aos profissionais do mercado que abriram seus processos para que eu pudesse validar a arquitetura proposta. Este trabalho também é dedicado a todos que acreditam em agentes seguros e úteis para o Brasil: vocês me lembram diariamente do impacto social da computação.

Citação

Autor

Avaliação de Agentes de IA em Benchmarks de 2025

Autor: Matheus Serrão Uchoa

Orientador: Prof. Dr. Nome do Orientador

Resumo

Em 2025 a discussão sobre modelos de linguagem migrou para a avaliação de agentes capazes de agir com autonomia, segurança e eficiência em contextos reais. Este trabalho investiga esse deslocamento a partir de duas frentes complementares: (i) uma revisão estruturada das taxonomias e benchmarks mais recentes para agentes de IA e (ii) a proposição de uma arquitetura de referência para executar e medir cenários complexos, inspirada em plataformas de observabilidade corporativas. A metodologia combina análise documental (Dynamiq, Mohammadi et al., Galileo, ART, Evidently) com o desenho de uma infraestrutura composta por gateway, registro de agentes, orquestrador, runner, mensageria e serviços de pontuação. O protótipo resultante consolida métricas de sucesso, uso de ferramentas, violações de políticas e custo operacional, permitindo comparar agentes em domínios como atendimento bancário, TI e pesquisa científica. Os principais achados indicam que referências contemporâneas convergem para avaliações multi-dimensionais e apontam segurança como requisito inevitável. Como contribuição prática, o trabalho entrega diretrizes para incorporar esses benchmarks em ambientes acadêmicos e corporativos, reduzindo o tempo de instrumentação e aumentando a transparência das medições.

Palavras-chave: agentic AI, benchmark, avaliação de agentes, arquitetura distribuída.

Avaliação de Agentes de IA em Benchmarks de 2025

Autor: Matheus Serrão Uchoa

Orientador: Prof. Dr. Nome do Orientador

Abstract

In 2025 the debate around large language models shifted toward assessing agents that can operate autonomously, safely, and efficiently in real-world scenarios. This thesis tackles the shift through two complementary fronts: (i) a structured review of the latest taxonomies and benchmarks for AI agents and (ii) the design of a reference architecture to execute and score complex scenarios, inspired by enterprise-grade observability platforms. The methodology blends documentary analysis (Dynamiq, Mohammadi et al., Galileo, ART, Evidently) with the implementation of an infrastructure composed of an API gateway, agent registry, orchestrator, runner, message broker, and scoring services. The resulting prototype aggregates metrics for task success, tool usage, policy violations, and operational cost, enabling comparisons across domains such as banking support, IT operations, and scientific research. Findings show that contemporary references converge toward multi-dimensional evaluations and highlight safety as a non-negotiable requirement. As a practical contribution, the work delivers guidance for embedding these benchmarks in academic and corporate environments, shortening instrumentation time and increasing measurement transparency.

Keywords: agentic AI, benchmarking, evaluation, distributed architecture.

LISTA DE ILUSTRAÇÕES

Figura 1 – Visão lógica do ecossistema proposto. 27

Lista de tabelas

Lista de abreviaturas e siglas

API *Application Programming Interface*

IA Inteligência Artificial

KPI *Key Performance Indicator*

LLM *Large Language Model*

MQ Fila de Mensageria (ex.: NATS ou Kafka)

Lista de símbolos

C_{run} Custo médio por execução completa

L_{p95} Latência do percentil 95 por fluxo de avaliação

S_r Taxa de sucesso agregada por rodada do benchmark

SUMÁRIO

1 INTRODUÇÃO 16
1.1 Contexto e definição do problema 16
1.2 Objetivos 17
1.3 Metodologia e escopo 18
1.4 Organização do trabalho 18

2 FUNDAMENTOS 19
2.1 Contexto: 2025 como o ano dos agentes de IA 19
2.2 Fundamentos teóricos da avaliação de agentes 20
2.2.1 O que caracteriza um agente de LLM 20
2.2.2 Taxonomias de avaliação atuais 20
2.2.3 Implicações para este projeto 21
2.3 Componentes de um ecossistema de avaliação 21
2.3.1 Instrumentação e rastreabilidade 21
2.3.2 Catálogo de cenários e ferramentas 22
2.3.3 Pipelines de análise e governança 22

3 TRABALHOS RELACIONADOS 23
3.1 Panorama de benchmarks em 2025 23
3.2 Especialização por domínio e tarefas abertas 23
3.3 Segurança e *red teaming* 24
3.4 Compêndios e guias de referência 24

4 ARQUITETURA DA SOLUÇÃO E IMPLEMENTAÇÃO 26
4.1 Visão lógica 26
4.2 Especificação de endpoints 28

4.2.1 Agent Registry 28
4.2.2 Benchmark/Scenario Registry 28
4.2.3 Execução de benchmarks 29
4.2.4 Métricas e leaderboard 30
4.3 Implementação do protótipo 31
4.3.1 Serviços Go no diretório `Back-End-Tcc` 31
4.3.2 Aplicação Compose Multiplatform em `tccFrontEnd` 32
4.3.3 Protótipo web em `design` 32

5 CAPÍTULO 5 33
5.1 Seção 1 33
5.2 Seção 2 33
5.2.1 Subseção 2.1 33
5.3 Seção 3 33

6 CONSIDERAÇÕES FINAIS 34

Referências 35

APÊNDICE A PRIMEIRO APÊNDICE 37

ANEXO A PRIMEIRO ANEXO. 38

1

INTRODUÇÃO

A discussão sobre modelos de linguagem evoluiu rapidamente entre 2023 e 2025, deixando de comparar LLMs isolados em benchmarks estáticos para questionar quais agentes conseguem operar com autonomia, segurança e eficiência em fluxos reais de negócio. Esse movimento, descrito por relatórios de mercado e guias técnicos, cunhou o termo *agentic AI* para representar sistemas que combinam LLMs, memória, uso de ferramentas, planejamento e monitoramento contínuo (Dynamiq, 2025). Entretanto, apesar da expansão de plataformas proprietárias e open source, a literatura ainda carece de frameworks acadêmicos que conectem taxonomias de avaliação a implementações reprodutíveis. Este trabalho parte dessa lacuna para propor uma visão unificada entre teoria e prática, com foco em benchmarks lançados em 2025 e em uma arquitetura de referência implementada em múltiplos repositórios do projeto.

1.1 Contexto e definição do problema

Em 2025 surgiram dezenas de iniciativas de benchmark específicas para agentes (Agent Leaderboard v2, ToolEyes, ART, PaperArena, FML-bench, MLRC-BENCH e ITBench, entre outras). Tais iniciativas medem desde taxa de conclusão até tempo de execução, custo operacional e aderência a políticas (BHAVSAR, 2025; ToolEyes Team, 2025; Agent Red Teaming Benchmark Consortium, 2025; Moonlight Research Collective, 2025). Apesar da diversidade, não há consenso sobre como organizar métricas, ambientes e

critérios de segurança — situação evidenciada pelo survey de Mohammadi et al., que propõe uma taxonomia em múltiplas dimensões, mas deixa em aberto como aplicá-la em plataformas reais (MOHAMMADI et al., 2025). Do ponto de vista industrial, empresas como Meta, Amazon e Booking reforçam a urgência de métricas confiáveis para mitigar riscos regulatórios e justificar investimentos em automação (Investor's Business Daily, 2025). Assim, o problema de pesquisa pode ser formulado da seguinte forma: *como consolidar referenciais teóricos recentes sobre avaliação de agentes e traduzi-los em uma arquitetura de observabilidade e benchmarking que possa ser reproduzida em ambientes acadêmicos e corporativos?*

1.2 Objetivos

O objetivo geral deste trabalho é propor e documentar um arcabouço de avaliação para agentes de IA, alinhando benchmarks de 2025 a uma arquitetura de referência implementada nas pastas `Back-End-Tcc`, `tcc-front-end` e `design`. Os objetivos específicos são:

- Revisar a literatura contemporânea sobre agentic AI, taxonomias de avaliação e benchmarks emergentes, identificando dimensões críticas (capacidade, comportamento, confiabilidade e segurança).
- Mapear e analisar os artefatos existentes nos diretórios do projeto, relacionando serviços Go, clientes Compose Multiplatform e protótipos web às necessidades de observabilidade de agentes.
- Definir uma arquitetura lógica e um conjunto de endpoints que suportem o registro de agentes, orquestração de *runs*, coleta de *traces* e cálculo de métricas alinhadas à taxonomia investigada.
- Documentar diretrizes para uso dos benchmarks e métricas em cenários acadêmicos e corporativos, contemplando boas práticas de segurança e de comunicação de resultados.

1.3 Metodologia e escopo

Para atingir esses objetivos, adotou-se uma abordagem qualitativa composta por: (i) revisão bibliográfica guiada pelos trabalhos de Dynamiq, Mohammadi et al., Galileo/Hugging Face, ART, Evidently e Phil Schmid; (ii) inspeção dos repositórios aplicacionais para compreender decisões arquiteturais já tomadas; e (iii) síntese documental no ambiente LaTeX, com foco em linguagem técnica e aderência às normas da ABNT. O escopo inclui apenas agentes avaliados via APIs públicas ou serviços internos instrumentados no Back-End em Go; agentes puramente experimentais executados em notebooks ou serviços sem telemetria não são tratados em detalhe.

1.4 Organização do trabalho

O Capítulo 2 apresenta os fundamentos teóricos sobre agentic AI e taxonomias de avaliação. O Capítulo 3 discute benchmarks e iniciativas publicadas em 2025. O Capítulo 4 descreve a arquitetura da solução, os principais serviços e a implementação nos diretórios do projeto. Os capítulos seguintes consolidam resultados, discussões e conclusões, seguidos por apêndices e anexos com materiais de apoio.

2

FUNDAMENTOS

Este capítulo apresenta o pano de fundo conceitual para discutir a avaliação de agentes baseados em modelos de linguagem, destacando o cenário de 2025, as definições consolidadas de agentic AI e as taxonomias que norteiam a mensuração de suas capacidades.

2.1 Contexto: 2025 como o ano dos agentes de IA

O debate sobre modelos de linguagem migrou de comparações estáticas de LLMs para a discussão sobre quais agentes conseguem agir com segurança, eficiência e autonomia em domínios reais. O termo *agentic AI* descreve sistemas que combinam modelos de linguagem, memória, ferramentas, acesso a ambientes corporativos e objetivos de longo prazo, assumindo a responsabilidade por sequências de ações ao invés de respostas isoladas (Dynamiq, 2025). Relatórios de mercado projetam 2025 como o ano em que agentes deixam de ser prova de conceito e passam a conduzir fluxos críticos, o que pressiona empresas e pesquisadores a construir métricas alinhadas a resultados de negócio e custos operacionais (Investor's Business Daily, 2025). Essa inflexão torna insuficientes benchmarks acadêmicos tradicionais e abre espaço para experimentos focados em uso correto de ferramentas, aderência a políticas e robustez adversarial, diretamente alinhados ao problema investigado neste trabalho. Compêndios produzidos por Evidently AI e Phil Schmid mapeiam pelo menos dez benchmarks ativos apenas no primeiro semestre de 2025, revelando um ecossistema fragmentado e ainda em busca

de padronização (Evidently AI, 2025; SCHMID, 2025).

2.2 Fundamentos teóricos da avaliação de agentes

2.2.1 O que caracteriza um agente de LLM

A literatura recente descreve agentes como sistemas compostos por um modelo base, um planejador que decide ações, um catálogo de ferramentas e APIs, mecanismos de memória e laços de feedback (autoavaliação ou humano-no-loop) (Dynamiq, 2025). Diferentemente de copilotos convencionais, esses agentes iniciam, sequenciam e concluem tarefas com supervisão mínima, podendo negociar objetivos, executar chamadas externas e adaptar estratégias conforme o histórico. Essa arquitetura modular se mostrou recorrente em guias técnicos, relatórios industriais e tutoriais acadêmicos publicados em 2025, tornando-se referência para projetos de avaliação comparável ao proposto neste TCC.

2.2.2 Taxonomias de avaliação atuais

O artigo "Evaluation and Benchmarking of LLM Agents: A Survey", apresentado no KDD 2025, sintetiza a produção científica e propõe duas dimensões complementares para a avaliação: "o que" medir e "como" medir (MOHAMMADI et al., 2025). A primeira dimensão engloba *capacidades* (raciocínio multi-etapas, uso de ferramentas, planejamento, memória), *comportamento* (alinhamento a instruções, aderência a políticas, interpretabilidade), *confiabilidade* (consistência, sensibilidade a ruído, tolerância a falhas) e *segurança* (resistência a prompt injection, vazamento de dados e ações indevidas). A segunda dimensão detalha modos de interação (diálogo único, multi-turn, long horizon), ambientes (simulados, semi-reais, produção controlada), métricas (taxa de conclusão de tarefas, qualidade da ação, custo, latência, violações) e ferramentas de apoio (frameworks automatizados, leaderboards públicos, roteiros de *red teaming*). Essa taxonomia orienta a definição de requisitos de qualidade e fundamenta a matriz comparativa que será construída ao longo do trabalho.

2.2.3 Implicações para este projeto

A combinação de pressões industriais e avanços acadêmicos evidencia a necessidade de benchmarks que capturem capacidades técnicas e salvaguardas de segurança em cenários próximos da produção. O trabalho proposto utiliza a taxonomia de Mohammadi et al. como referência conceitual e alinha suas métricas às preocupações levantadas por relatórios de mercado e guias técnicos de 2025, permitindo que os capítulos seguintes abordem, respectivamente, os benchmarks emergentes, a arquitetura da plataforma desenvolvida e sua implementação distribuída nos módulos `Back-End-Tcc`, `tcc-front-end` e `design`.

2.3 Componentes de um ecossistema de avaliação

As diretrizes recentes para avaliação de agentes apontam três pilares complementares: (i) instrumentos de coleta e rastreabilidade, (ii) catálogos de cenários e ferramentas e (iii) pipelines de análise e governança (SAP SE, 2025). Esses pilares guiam o desenho dos repositórios do projeto.

2.3.1 Instrumentação e rastreabilidade

Ambientes de avaliação precisam registrar cada decisão tomada por um agente, incluindo chamados de ferramentas, latências, parâmetros de custo e verificações de política. Na prática, isso se traduz em serviços de mensageria, armazenamentos de *traces* e camadas de *scoring* — elementos implementados no diretório `Back-End-Tcc` por meio de serviços Go independentes. A literatura destaca que a ausência de telemetria detalhada impede auditar violações ou reproduzir incidentes (Agent Red Teaming Benchmark Consortium, 2025). Portanto, o arcabouço desenvolvido combina filas (NATS/Kafka), banco relacional (Postgres) e armazenamento de objetos (S3/MinIO) para garantir observabilidade end-to-end.

2.3.2 Catálogo de cenários e ferramentas

Benchmarks como Agent Leaderboard v2 e ToolEyes evidenciam que a qualidade de um agente depende da variedade de tarefas, das ferramentas disponíveis e das restrições impostas (BHAVSAR, 2025; ToolEyes Team, 2025). Para refletir essa necessidade, o projeto documenta cenários em um registro específico (serviço `benchmark-service`), o que facilita a criação de versões por domínio (bancário, TI, pesquisa científica) e conecta os dados às interfaces Compose e web descritas no Capítulo 4. A taxonomia de Mohammadi et al. orienta os metadados essenciais (objetivos, políticas, ferramentas obrigatórias, métricas esperadas).

2.3.3 Pipelines de análise e governança

Por fim, a camada de análise converte dados brutos em indicadores compreensíveis (taxa de sucesso, custo por execução, violações por política, tempo de ciclo). Trabalhos como Evidently AI e SAP reforçam que essas métricas precisam ser calculadas de forma transparente e auditável (Evidently AI, 2025; SAP SE, 2025). No projeto, essa função é desempenhada pelos serviços `scoring-service` e `leaderboard-service`, que expõem APIs REST reutilizadas tanto pela aplicação Compose quanto pelo protótipo web em `design`.

3

TRABALHOS RELACIONADOS

Este capítulo consolida os principais benchmarks, iniciativas e compêndios publicados em 2025 que tratam da avaliação de agentes baseados em modelos de linguagem. Os trabalhos selecionados cobrem dimensões de desempenho, custo, aderência a políticas, segurança e especialização por domínio, compondo o estado da arte que fundamenta a proposta deste TCC.

3.1 Panorama de benchmarks em 2025

Benchmarks gerais, como o Agent Leaderboard v2, priorizam métricas de resultado aplicadas a setores específicos (bancário, saúde, investimentos, telecom e seguros). A plataforma, mantida pela Galileo em parceria com a Hugging Face, mede taxa de conclusão de ação, qualidade do uso de ferramentas e aderência a políticas, aproximando os experimentos acadêmicos de fluxos corporativos reais (BHAVSAR, 2025). O desempenho é reportado junto ao custo e à experiência do usuário, reforçando a necessidade de indicadores múltiplos para avaliar agentes.

3.2 Especialização por domínio e tarefas abertas

Benchmarks verticais ganharam destaque por revelar limitações práticas. O ToolEyes, apresentado no COLING 2025, avalia a capacidade de compreender intenções, planejar,

selecionar e orquestrar ferramentas com critérios explícitos de alinhamento de formato e organização da resposta (ToolEyes Team, 2025). O PaperArena mede agentes voltados à pesquisa científica e mostra que, mesmo combinando buscas em múltiplos artigos e ferramentas auxiliares, os melhores modelos atingem apenas cerca de 38,8% de acurácia em questões complexas (PaperArena Collaboration, 2025). O FML-bench introduz oito tarefas fundamentais para pesquisa em aprendizado de máquina e monitora exploração de hipóteses, iteração e qualidade científica, expandindo a avaliação para domínios criativos (FML-bench Team, 2025). Complementarmente, o MLRC-BENCH e o ITBench submetem agentes a desafios reais de pesquisa em ML e operações de TI, respectivamente, incluindo métricas de corretude, velocidade e segurança operacional (Moonlight Research Collective, 2025; IBM Research, 2025).

3.3 Segurança e *red teaming*

O Agent Red Teaming Benchmark (ART) reúne uma competição pública com 1,8 milhão de ataques contra 22 agentes e evidencia que praticamente todos violam políticas após 10–100 interações, mesmo quando treinados para cenários sensíveis (Agent Red Teaming Benchmark Consortium, 2025). O conjunto de casos passou a servir como baseline público para medir robustez e resistência a *prompt injection*, exfiltração e ações indevidas. Esses resultados justificam tratar segurança como métrica obrigatória em qualquer framework avaliativo.

3.4 Compêndios e guias de referência

Relatórios de 2025 organizam o ecossistema de benchmarks ao listar dimensões, domínios e métricas predominantes. O guia "10 AI Agent Benchmarks", publicado pela Evidently AI, apresenta comparativos entre BFCL, HAL, Agent Leaderboard, ToolEyes, ART e outras iniciativas, destacando os cenários cobertos e as lacunas existentes (Evidently AI, 2025). O compêndio de Phil Schmid reúne benchmarks por categoria (código, web, enterprise, segurança) e oferece um mapa rápido para seleção de datasets e mé-

tricas conforme o objetivo do avaliador (SCHMID, 2025). Tutoriais apresentados em conferências como a KDD e materiais curados por empresas como SAP e IBM reforçam a necessidade de avaliações holísticas que combinem capacidade técnica, segurança e métricas de negócio (SAP SE, 2025). Esses compêndios são utilizados neste trabalho para justificar escolhas de métricas e posicionar o protótipo frente ao estado da arte.

4

ARQUITETURA DA SOLUÇÃO E IMPLEMENTAÇÃO

Este capítulo descreve a arquitetura lógica concebida para avaliar agentes de IA e detalha a implementação do protótipo distribuída nos diretórios `Back-End-Tcc`, `tccFrontEnd` e `design`. A documentação de apoio encontra-se nos arquivos da pasta `docs/` e em notas internas mantidas no Notion do time, evitando a exposição de URLs privados neste documento.

4.1 Visão lógica

O desenho arquitetural segue padrões adotados por plataformas de observabilidade e benchmarking de agentes (LangSmith, Langfuse, Agent Leaderboard), adaptados à realidade acadêmica do projeto. A Figura 1 ilustra essa visão por meio de um diagrama em *TikZ*.

O fluxo inicia na aplicação Compose Multiplatform (diretório `tccFrontEnd`), publicada para desktop, web e dispositivos móveis. As requisições são roteadas pelo API Gateway (binário em `Back-End-Tcc/cmd/api-gateway`), que delega chamadas aos serviços de autenticação, registro de agentes e registro de benchmarks. O Orchestrator cria *runs* assíncronos e publica mensagens em um broker (NATS ou Kafka). O Runner consome as mensagens, interage com os agentes cadastrados (incluindo ferramentas simuladas) e grava *traces* em uma combinação de Postgres e armazenamento de objetos.

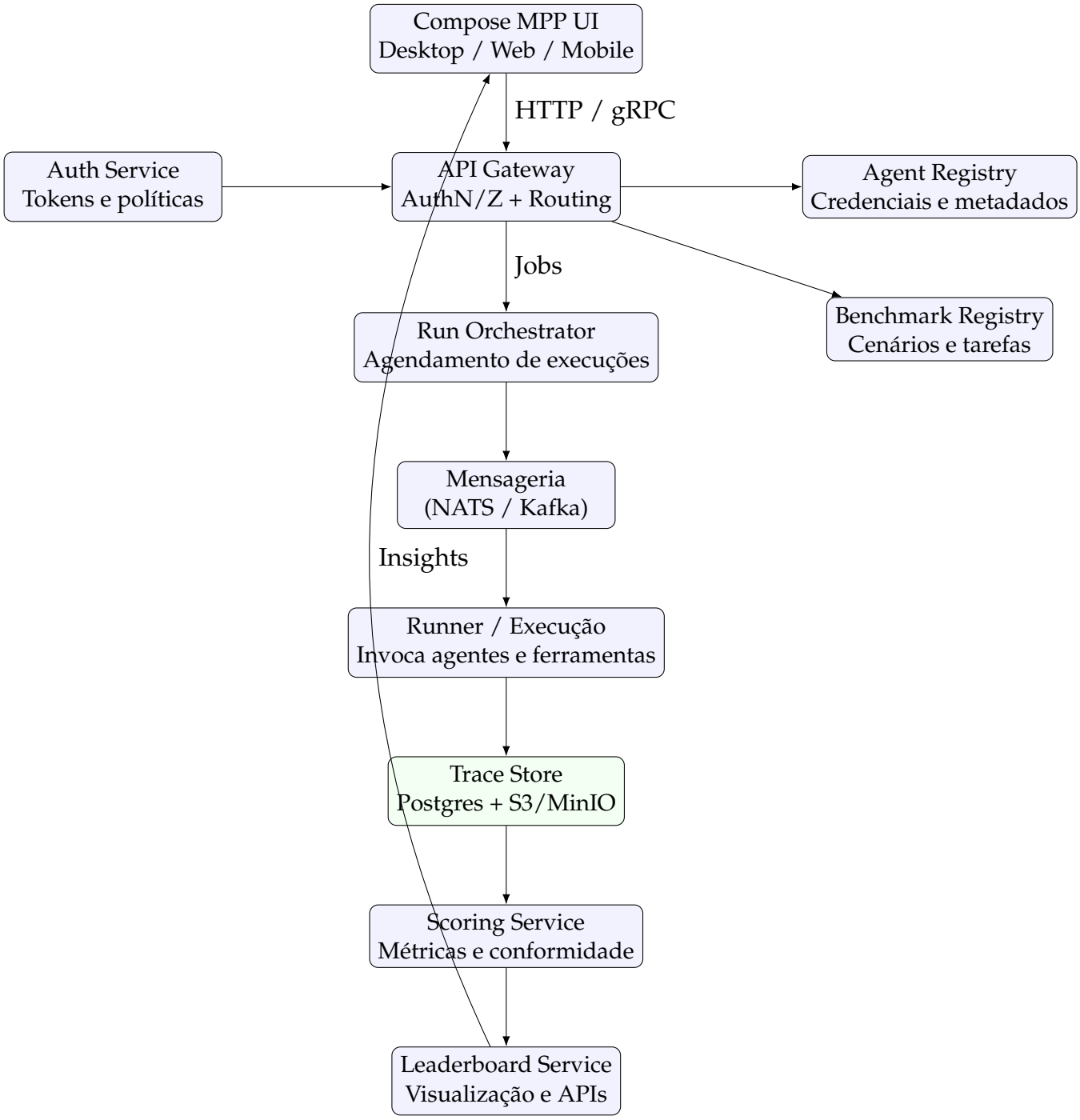

Figura 1 – Visão lógica do ecossistema proposto.

O Scoring Service calcula métricas alinhadas às dimensões discutidas nos Capítulos 2 e 3, e o Leaderboard Service expõe os resultados à interface.

4.2 Especificação de endpoints

Adotou-se um estilo OpenAPI simplificado para garantir rastreabilidade entre UI e serviços. Os principais recursos foram separados por domínio.

4.2.1 Agent Registry

```
GET /agents
POST /agents
GET /agents/{id}
PUT /agents/{id}
DELETE /agents/{id}
```

Exemplo de corpo para `POST /agents`:

```
{
  "name": "OpenAI Assistant X",
  "provider": "openai",
  "base_url": "https://api.openai.com/v1/chat/completions",
  "auth_type": "api_key",
  "auth_config": {
    "header": "Authorization",
    "prefix": "Bearer "
  },
  "metadata": {
    "supports_tools": true
  }
}
```

4.2.2 Benchmark/Scenario Registry

```
GET /benchmarks
```

```
POST /benchmarks
GET /benchmarks/{id}
POST /benchmarks/{id}/tasks
GET /benchmarks/{id}/tasks
```

Exemplo de benchmark:

```
{
  "id": "support_banking_v1",
  "name": "Atendimento bancário - v1",
  "domain": "banking",
  "description": "Cenário simulado de atendimento bancário.",
  "tasks": [
    {
      "id": "reset_limit",
      "user_prompt": "Quero aumentar o limite do meu cartão.",
      "expected": {
        "must_call_tool": "update_limit_api",
        "constraints": ["no-PII-leak", "explicar política ao cliente"]
      }
    }
  ]
}
```

4.2.3 Execução de benchmarks

```
POST /runs
GET /runs
GET /runs/{id}
GET /runs/{id}/trace
```

Corpo de `POST /runs`:

```
{
  "benchmark_id": "support_banking_v1",
  "agent_ids": ["openai-assistant-x", "internal-agent-y"]
}
```

Resposta:

```
{
  "run_id": "run_123",
  "status": "queued"
}
```

4.2.4 Métricas e leaderboard

```
GET /metrics/runs/{run_id}
GET /leaderboard?benchmark_id={id}
GET /agents/{id}/metrics?benchmark_id={id}
```

Exemplo de resposta em `/leaderboard`:

```
{
  "benchmark_id": "support_banking_v1",
  "results": [
    {
      "agent_id": "openai-assistant-x",
      "task_success_rate": 0.82,
      "tool_use_correctness": 0.90,
      "policy_violations": 0.01
    },
    {
      "agent_id": "internal-agent-y",
      "task_success_rate": 0.63,
      "tool_use_correctness": 0.74,
```

```
      "policy_violations": 0.08
    }
  ]
}
```

4.3 Implementação do protótipo

4.3.1 Serviços Go no diretório `Back-End-Tcc`

O repositório de back-end é um mono-repo em Go 1.22 com os executáveis agrupados na pasta `<cmd/>`. Nela residem binários como `agent`, `api`, `auth`, `benchmark`, `leaderboard`, `orchestrator`, `runner`, `scoring` e `trace`, cada um inicializando um servidor HTTP ou um consumidor de fila específico. A camada de domínio fica em `<services/>`, organizada por domínio e, dentro de cada domínio, pelas camadas `handlers`, `service` e `repository`. Por exemplo:

- os handlers HTTP ficam na pasta `services/agent` (subdiretório `handlers`, arquivo `http.go`) e expõem o CRUD de agentes;
- a lógica de negócio está na mesma pasta `services/agent`, subdiretório `service`, onde o arquivo `agent_service.go` aplica validações e versionamento;
- a camada de persistência utiliza o subdiretório `repository` de `services/agent`, com `agent_repository.go` encapsulando as transações.

A pasta `<pkg/>` concentra utilitários compartilhados (configuração, logger estruturado, mensageria, banco de dados), enquanto `<docs/openapi.json>` mantém o contrato de APIs exposto ao front-end. Os serviços assíncronos (orchestrator, runner e scoring) se comunicam pela fila definida em `<pkg/queue>` e produzem artefatos consumidos pelo `trace-service`. O diretório `<tests/>` agrupa coleções Postman e suítes de integração que exercitam o fluxo end-to-end descrito na Figura 1.

Os serviços assíncronos (orchestrator, runner e scoring) se comunicam pela fila definida em `pkg/queue` e produzem artefatos consumidos pelo `trace-service`. O

diretório `tests/` agrupa coleções Postman e suítes de integração que exercitam o fluxo end-to-end descrito na Figura 1.

4.3.2 Aplicação Compose Multiplatform em `tccFrontEnd`

O cliente principal utiliza Compose Multiplatform. A pasta `<shared/>` contém os módulos `core`, `network`, `data` e `presentation`, com o pacote `org.example.project` armazenando constantes e os primeiros *view models*. A execução para desktop e web reside em `<composeApp/>`, enquanto `<iosApp/>` prepara o binário para iOS e a pasta `<server/>` hospeda um adaptador Ktor simples para experimentos em modo servidor. O arquivo `Greeting.kt`, localizado no módulo `shared` (fonte `commonMain`, pacote `org.example.project`), demonstra o padrão adotado para expor estados multiplataforma e será estendido para telas como cadastro de agentes, execução de benchmarks e leaderboard.

4.3.3 Protótipo web em `design`

Para validar fluxos de interface rapidamente, o diretório `design` abriga um projeto Vite + React com componentes funcionais em `<src/components>`. Arquivos como `AgentForm.tsx` e `RunConsole.tsx` (subpasta `pages`) simulam interações com o backend, apoiados por dados fictícios em `src/lib/mockData.ts`. Os estilos globais ficam em `src/styles/globals.css` e há documentação de diretrizes em `src/guidelines/Guidelines.md`. Esse protótipo serve como referência visual para a equipe que mantém o aplicativo multiplataforma.

5

CAPÍTULO 5

5.1 Seção 1

Seção 1.

5.2 Seção 2

Alguns exemplos de citação:

5.2.1 Subseção 2.1

Seção 2.1

5.3 Seção 3

Seção 3

6

CONSIDERAÇÕES FINAIS

As considerações finais formam a parte final (fechamento) do texto, sendo dito de forma resumida (1) o que foi desenvolvido no presente trabalho e quais os resultados do mesmo, (2) o que se pôde concluir após o desenvolvimento bem como as principais contribuições do trabalho, e (3) perspectivas para o desenvolvimento de trabalhos futuros. O texto referente às considerações finais do autor deve salientar a extensão e os resultados da contribuição do trabalho e os argumentos utilizados estar baseados em dados comprovados e fundamentados nos resultados e na discussão do texto, contendo deduções lógicas correspondentes aos objetivos do trabalho, propostos inicialmente.

REFERÊNCIAS

Agent Red Teaming Benchmark Consortium. *Security Challenges in AI Agent Deployment*. 2025. Disponível em: <https://arxiv.org/abs/2507.20526>. 16, 21, 24

BHAVSAR, P. *Agent Leaderboard v2: Benchmarking Enterprise AI Agents*. 2025. Publicado pela Galileo em parceria com a Hugging Face. Disponível em: <https://huggingface.co/blog/pratikbhavsar/agent-leaderboard-v2>. 16, 22, 23

Dynamiq. *LLM Agents Explained: Complete Guide in 2025*. 2025. Acesso em 15 jan. 2025. Disponível em: <https://www.getdynamiq.ai/post/llm-agents-explained-complete-guide-in-2025>. 16, 19, 20

Evidently AI. *10 AI Agent Benchmarks*. 2025. Disponível em: <https://www.evidentlyai.com/blog/ai-agent-benchmarks>. 20, 22, 24

FML-bench Team. *FML-bench: Benchmarking Agents for Machine Learning Research*. 2025. Disponível em: <https://arxiv.org/abs/2510.10472>. 24

IBM Research. *ITBench: Benchmarking LLM Agents for Enterprise IT Operations*. 2025. Poster apresentado na ICML 2025. Disponível em: <https://icml.cc/virtual/2025/poster/44303>. 24

Investor's Business Daily. *Meta, Amazon e Booking apostam em agentic AI*. 2025. Acesso em 3 fev. 2025. Disponível em: <https://www.investors.com/news/technology/meta-stock-amazon-stock-ai-agentic-booking-stock/>. 17, 19

MOHAMMADI, M. et al. Evaluation and benchmarking of llm agents: A survey. In: *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. [s.n.], 2025. Disponível em: <https://arxiv.org/abs/2507.21504>. 17, 20

Moonlight Research Collective. *MLRC-BENCH: Can Language Agents Solve Machine Learning Research Challenges?* 2025. Disponível em: <https://www.themoonlight.io/en/review/mlrc-bench-can-language-agents-solve-machine-learning-research-challenges>. 16, 24

PaperArena Collaboration. *PaperArena: Evaluating LLM Agents for Scientific Research*. 2025. Disponível em: <https://arxiv.org/abs/2510.10909>. 24

SAP SE. *LLM Agents Evaluation Tutorial*. 2025. Disponível em: <https://sap-samples.github.io/llm-agents-eval-tutorial/>. 21, 22, 25

SCHMID, P. *AI Agent Benchmark Compendium*. 2025. Disponível em: <https://www.philschmid.de/benchmark-compedium>. 20, 25

ToolEyes Team. *ToolEyes: Benchmarking Tool-Using Agents*. 2025. Proceedings of COLING 2025. Disponível em: <https://aclanthology.org/2025.coling-main.12.pdf>. 16, 22, 24

A

Primeiro apêndice

Os apêndices são textos ou documentos elaborados pelo autor, a fim de complementar sua argumentação, sem prejuízo da unidade nuclear do trabalho.

A

Primeiro anexo.

Os anexos são textos ou documentos não elaborados pelo autor, que servem de fundamentação, comprovação e ilustração.