LLAMA LAND SOFTWARE HOUSE

Visão geral A Llama Land é uma software house focada em desenvolvimento de software com forte especialização em Inteligência Artificial (IA) e IA Generativa. Atendemos empresas que desejam acelerar a digitalização de processos, criar produtos de dados e incorporar capacidades de IA de forma segura, ética e escalável. Operamos com times multidisciplinares e um ecossistema de parceiros tecnológicos para entregar soluções ponta a ponta, do discovery à operação em produção.

História e fundadoras Fundada em 2021, no Rio de Janeiro, por duas profissionais com trajetórias complementares: - Marina Azevedo, Engenheira de Software (ex-líder de engenharia em empresas de tecnologia, foco em arquitetura distribuída, SRE e plataformas de dados). - Dra. Ana Luísa Teixeira, Doutora em Ciência da Computação (pesquisa em NLP, sistemas de recuperação de informação e avaliação de modelos generativos). A combinação de engenharia de produto e pesquisa aplicada molda o DNA da Llama Land: resolver problemas reais com rigor técnico e foco no valor de negócio.

Missão Capacitar organizações a criar e operar produtos digitais orientados por dados e IA, com qualidade, segurança e governança, reduzindo o tempo entre a ideia e o impacto em produção.

Valores - Ética e responsabilidade: IA centrada no ser humano, com privacidade e compliance. - Qualidade como disciplina: testes, observabilidade e métricas desde o primeiro sprint. - Transparência e parceria: co-criação com clientes, feedback contínuo e contratos claros. - Pragmatismo técnico: soluções simples, escaláveis e sustentáveis. - Aprendizado contínuo: P&D; aplicado, formação e compartilhamento de conhecimento.

Endereço Av. Rio Branco, 156, 17º andar, Centro, Rio de Janeiro – RJ, 20040-002 Telefone: +55 (21) 3512-0001 E-mail: contato@llamaland.pro Site: www.llamaland.pro

Parcerias tecnológicas - Microsoft: projetos e implantações em Azure (Azure OpenAI Service, Azure Kubernetes Service, Azure Machine Learning, Data Lake, Purview, DevOps). - OpenAI: expertise em integração de modelos generativos (Chat Completions, Assistants, RAG), governança de prompts, avaliação de qualidade e controle de custos. Outras integrações frequentes: AWS (Bedrock, S3, EKS), GCP (Vertex AI), Elastic, LangChain, LlamaIndex, Weaviate, FAISS.

Principais serviços 1) Descoberta e estratégia de IA - Identificação de casos de uso, estimativa de ROI e priorização de backlog. - Desenho de arquitetura alvo (dados, MLOps/LMMOps, segurança e compliance).

2) Desenvolvimento de produtos com IA e IA Generativa - Aplicações com RAG (Retrieval-Augmented Generation) usando dados proprietários. - Assistentes e agentes especializados (suporte, vendas, engenharia, compliance). - Integrações com ERPs/CRMs/ITSMs, automações e workflows inteligentes. - Microserviços, APIs e front-ends (web/mobile) com foco em experiência do usuário.

3) Plataforma de dados e MLOps/LMMOps - Pipelines de ingestão, transformação e catalogação (Data Lake/Lakehouse). - Feature stores, versionamento de datasets, lineage e governança. - Observabilidade de modelos (qualidade, deriva, custos e latência).

4) Qualidade e segurança para sistemas com GenAI - Testes funcionais e não funcionais para apps com LLMs. - Avaliação de prompts, métricas de qualidade (exatidão, relevância, segurança), teste humano-no-loop. - Red teaming, mitigação de jailbreaks, PII filtering, guardrails e políticas de conteúdo.

5) Capacitação e mudança organizacional - Treinamentos práticos para times técnicos e líderes. - Playbooks de adoção segura de IA e governança corporativa.

Diferenciais - Combinação de pesquisa aplicada e engenharia de produção. - Experiência prática com RAG, agentes e avaliação sistemática de LLMs. - Entregas iterativas com métricas de valor, custo e qualidade desde o início. - Arquiteturas cloud-agnostic e foco em evitar lock-in desnecessário. - Governança de dados e de IA incorporadas ao ciclo de vida do produto.

Tecnologias e práticas - Linguagens e frameworks: Python, TypeScript, FastAPI, Node.js, React. - IA/GenAI: OpenAI, Azure OpenAI, modelos abertos (Llama, Mistral), embeddings, vetorização (FAISS/Weaviate), LangChain/LlamaIndex. - Dados: Spark, Delta Lake, Databricks, PostgreSQL, Elastic, Kafka. - Plataforma: Kubernetes, Docker, Terraform, CI/CD (GitHub Actions/Azure DevOps). - Observabilidade: Prometheus, Grafana, OpenTelemetry, App Insights. - Segurança: OAuth2/OIDC, Key Vault/Secrets Manager, KMS, DLP, RBAC/ABAC.

Processos e governança - Descoberta → Prova de Valor (PoV) → MVP → Escala/Operação. - Qualidade orientada a métricas (SLOs de latência, custo por resposta, taxa de citações corretas). - Avaliação humana-no-loop para casos de alto risco. - Controles de privacidade (minimização de dados, anonimização, retenção). - Documentação viva do sistema (arquitetura, decisões, riscos e backlog).

Modelos de engajamento - Time dedicado (squads ágeis). - Projeto fechado por escopo. - Co-desenvolvimento com time do cliente. - Suporte e SRE/ML(LM)Ops gerenciados.

Clientes e setores - Atlântica Energia S.A. (energia) – ativo – assistente técnico para operação de ativos e RAG em manuais. - Banco Horizonte (serviços financeiros) – ativo – copiloto para atendimento e reconciliação com RAG. - Varejo Veloz (retail) – ativo – motor de busca semântica e recomendação de produtos. - Saúde Viva Clínicas (saúde) – ativo – resumo clínico assistido com guardrails e anonimização. - Indústrias Guanabara (manufatura) – ativo – inspeção assistida e análise de incidentes com RAG. - Porto Mar Azul (logística/portos) – ativo – copiloto operacional e triagem documental. - Seguros Atlas (seguros) – ativo – automação de sinistros e checagem antifraude com IA generativa. - AeroRio Operações (aeroportuário) – ativo – base de conhecimento operacional e copiloto de pista. - Prefeitura de Boa Esperança (setor público) – ativo – atendimento ao cidadão com IA e acessibilidade. - AgroSul Foods (agro) – ex-cliente (2022–2023) – assistente de qualidade e rastreabilidade. - Fênix Telecom (telecom) – ex-cliente (2021–2022) – busca unificada e assistente de suporte N2. - Orion Tech Parts (distribuição) – ex-cliente (2023) – RAG para catálogo e pós-venda.

Exemplos de projetos 1) Copiloto de Suporte com RAG (Banco Horizonte) - Escopo: integração a CRM, base de conhecimento e histórico de tickets. - Métricas: redução de 28% no TMA; aumento de 22% na resolução no primeiro contato. - Arquitetura: Azure OpenAI + FAISS + FastAPI + observabilidade de custo/latência.

2) Assistente Técnico para Operação de Ativos (Atlântica Energia) - Escopo: RAG com manuais de OEM, relatórios de manutenção e logs. - Métricas: -18% em tempo de diagnóstico; +15% em aderência a procedimentos. - Segurança: controle de acesso por papel e trilhas de auditoria.

3) Triagem de Documentos e Conformidade (Seguros Atlas) - Escopo: classificação, extração de entidades e verificação de políticas com LLM. - Métricas: -35% no tempo de análise; +20% na conformidade de checklist. - Governança: PII filtering, anonimização, revisão humana para casos limítrofes.

Entrega e qualidade - Definição de SLOs por caso de uso (latência, custo, acurácia/relevância). - Avaliação contínua de prompts e dados de ground-truth. - Playbook de rollback e contingência (fallback para respostas seguras). - Revisões de arquitetura e segurança a cada marco do projeto.

Time e competências - Engenharia de Software, Dados e Plataforma. - Ciência de Dados e Pesquisa em NLP/IR. - Qualidade/Segurança e Governança de IA. - Produto/Design e Mudança Organizacional.

Contratação e contato - Solicitar workshop de descoberta para mapear casos de uso e ROI. - Propostas técnicas e comerciais personalizadas por setor e maturidade digital. - Contato: comercial@llamaland.pro | +55 (21) 3512-0001