ANTONIO MENEGHETTI FACULDADE

ESPECIALIZAÇÃO EM ONTOPSICOLOGIA

LUCIANO AZEVEDO CASSOL

AntonIA, uma inteligência artificial generativa para a Ciência Ontopsicológica

RESTINGA SÊCA, RECANTO MAESTRO

2026

LUCIANO AZEVEDO CASSOL

AntonIA, uma inteligência artificial generativa para a Ciência Ontopsicológica

Artigo apresentado como requisito parcial para obtenção do título de Especialista em Ontopsicologia, Curso de Especialização Lato Sensu em Ontopsicologia, Faculdade Antonio Meneghetti – AMF.

Orientadora: Profa. Dra. Patrícia Wazlawick

RESTINGA SÊCA, RECANTO MAESTRO

2026

LUCIANO AZEVEDO CASSOL

AntonIA, uma inteligência artificial generativa para a Ciência Ontopsicológica

Artigo apresentado como requisito parcial para obtenção do título de Especialista em Ontopsicologia, Curso de Especialização Lato Sensu em Ontopsicologia, Faculdade Antonio Meneghetti – AMF.

Restinga Sêca, _____ de _____ de _____.

COMISSÃO EXAMINADORA

_____

Profa. Dra. Patrícia Wazlawick
Orientadora do Trabalho de Conclusão de Curso
Faculdade Antonio Meneghetti

_____

Prof. Dr. Raul Sidnei Wazlawick
Membro da Banca Examinadora
Universidade Federal de Santa Catarina

_____

Prof. Msc. Pedro Hermes
Membro da Banca Examinadora
Faculdade Antonio Meneghetti

AntonIA, uma inteligência artificial generativa para a Ciência Ontopsicológica

Luciano Azevedo Cassol
Patrícia Wazlawick

Resumo: A evolução dos modelos de linguagem de grande escala (*Large Language Models* – LLMs) aumentou as possibilidades de aplicação da Inteligência Artificial generativa em domínios especializados de conhecimento. No entanto, a dependência desses modelos de bases de dados genéricas expõe os sistemas a riscos de alucinação e distorção conceitual, especialmente em áreas com terminologia própria e fontes primárias em idiomas específicos. Este trabalho apresenta o desenvolvimento do sistema AntonIA, um protótipo de IA generativa baseada na técnica de *Retrieval-Augmented Generation* (RAG) com execução local, projetado para responder a consultas sobre os fundamentos da Ciência Ontopsicológica com fidelidade semântica a base de referência. A metodologia adotada é a *Design Science Research Methodology* (DSRM), que orientou as etapas de identificação do problema, definição dos objetivos, projeto e desenvolvimento do artefato, demonstração e avaliação. O sistema foi construído sobre o framework Django, com orquestração via LangChain, indexação vetorial por FAISS e modelos LLM executados localmente via Ollama. Quatro modelos de linguagem foram avaliados: Qwen2.5:7b, LLaMA 3.1:8b, Gemma2:9b e Mistral-Nemo:12b. Os resultados demonstraram que o pipeline RAG, ancorado em uma base de conhecimento validada e verificada, é capaz de gerar respostas precisas e contextualmente coerentes com o domínio ontopsicológico, sem alucinações e sem necessidade de infraestrutura em nuvem ou pagamento por consumo de tokens, validando a viabilidade técnica, informacional e econômica da solução proposta.

Palavras-chave: Inteligência Artificial Generativa; Retrieval Augmented Generation; Ontopsicologia.

Abstract: The evolution of Large Language Models (LLMs) has expanded the possibilities for applying generative Artificial Intelligence in specialized knowledge domains. However, the dependence of these models on generic datasets exposes systems to risks of hallucination and conceptual distortion, particularly in areas with specific terminology and primary sources in non-dominant languages. This paper presents the development of AntonIA, a generative AI prototype based on the *Retrieval-Augmented Generation* (RAG) technique with fully local execution, designed to answer queries about the foundations of Ontopsychological Science with semantic fidelity to the reference knowledge base. The methodology adopted is the Design Science Research Methodology (DSRM), which guided the stages of problem identification, objective definition, artifact design and development, demonstration, and evaluation. The

system was built on the Django framework, with orchestration via LangChain, vector indexing through FAISS, and LLM models executed locally via Ollama. Three language models were evaluated: Qwen2.5:7b, LLaMA 3.1:8b, and Mistral-Nemo:12b. The results demonstrated that the RAG pipeline, anchored in a validated and verified knowledge base, is capable of generating accurate and contextually coherent responses within the Ontopsychological domain, without hallucinations and without the need for cloud infrastructure or token-based billing, validating the technical, informational and economic viability of the proposed solution.

Keywords: Generative Artificial Intelligence; Retrieval Augmented Generation; Ontopsychology.

1 Introdução

A Inteligência Artificial (IA) generativa tem ocupado um papel cada vez mais relevante nos processos de produção, organização e difusão do conhecimento em diferentes áreas da ciência e da tecnologia. Com o advento dos modelos de linguagem de grande escala, tornou-se possível criar sistemas capazes de compreender e gerar texto com alto grau de fluência, respondendo a consultas complexas e sintetizando informações de forma coerente Bommasani *et al.* (2022). No entanto, essa capacidade não está livre de limitações. Os modelos treinados em grandes volumes de dados genéricos apresentam tendência a produzir respostas que mesclam conceitos de diferentes domínios, gerando distorções semânticas que comprometem a precisão do conhecimento transmitido Bommasani *et al.* (2022). Esse problema é especialmente crítico em áreas de conhecimento com terminologia própria, fontes primárias em idiomas específicos e exigências rigorosas de fidelidade conceitual, como é o caso da Ciência Ontopsicológica Meneghetti (2022).

A Ontopsicologia, ciência desenvolvida pelo Acadêmico Professor Antonio Meneghetti, possui fundamentação teórica, bem como um método, descobertas e instrumentos de análise e de intervenção próprios que se configuram em termos, conceitos e princípios específicos que exigem precisão terminológica para serem transmitidos de forma correta. A utilização de modelos generalistas para responder a consultas sobre essa ciência apresenta o risco de produzir o que Meneghetti (2024) define como meme: *"a imitação de uma realidade, de uma emoção ou de um processo causa e efeito"*. Quando a Inteligência Artificial gera respostas sobre Ontopsicologia com base em dados genéricos sem ancoragem nas fontes originais, ela replica aparências de conteúdo distorcendo o conhecimento que deveria difundir.

Diante desse problema, este trabalho propõe o desenvolvimento do sistema AntonIA uma tecnologia aplicada à Ciência Ontopsicológica através do uso de Inteligência Artificial generativa baseada na técnica de *Retrieval-Augmented Generation* (RAG) com execução em ambiente local, projetada para operar sobre uma base de conhecimento validada e verificada composta por

fontes da Ciência Ontopsicológica. A técnica RAG, formalmente introduzida por Lewis *et al.* (2020), permite que modelos de linguagem pré-treinados respondam a consultas com base em documentos recuperados em tempo real de uma base de conhecimento específica, sem depender exclusivamente do conhecimento parametricamente armazenado durante o treinamento. Esse mecanismo reduz significativamente o risco de alucinações[1] e preserva a coerência conceitual do domínio.

A escolha por uma infraestrutura local fundamenta-se em três aspectos:

- segurança: o processamento local garante que os documentos institucionais e os conteúdos da Ciência Ontopsicológica não sejam transmitidos a servidores externos, preservando a propriedade intelectual da instituição;
- soberania sobre a base de conhecimento: o controle sobre os modelos e os parâmetros de geração assegura que o comportamento do sistema permaneça estável e auditável;
- economia: a ausência de custos por consumo de tokens torna a solução viável para instituições de ensino com ciclos orçamentários anuais, eliminando a dependência de fornecedores externos e a volatilidade de preços característica do mercado de APIs de Inteligência Artificial generativa.

O trabalho foi desenvolvido segundo a *Design Science Research Methodology* (DSRM), proposta por Peffers *et al.* (2007), que orienta a pesquisa para a criação e validação de artefatos tecnológicos que resolvam problemas reais com rigor científico e relevância prática Hevner *et al.* (2004). O artefato resultante, o software AntonIA, foi avaliado por meio de testes com quatro modelos de linguagem de código aberto, com foco na qualidade das respostas geradas sobre conceitos fundamentais da Ontopsicologia, configurando uma interdisciplinaridade entre a Ciência Ontopsicológica, a Inteligência Artificial Generativa.

A ideia do sistema AntonIA surgiu da necessidade de criar uma ferramenta que auxiliasse os alunos e pesquisadores da Ciência Ontopsicológica a acessar informações precisas e confiáveis, evitando a propagação de conceitos distorcidos que outras IAs geram quando questionadas sobre conceitos da Ontopsicologia, o que acaba levando ao não uso de tecnologias de IA generativa como um recurso de apoio à aprendizagem e à pesquisa. A proposta do AntonIA é fornecer respostas fundamentadas em uma base de conhecimento validada, garantindo que os usuários recebam informações corretas e contextualizadas, promovendo a disseminação adequada do conhecimento ontopsicológico.

O artigo está organizado da seguinte forma, a saber, a Seção 2 apresenta a fundamentação teórica, abordando os conceitos de Inteligência Artificial, aprendizado de máquina, modelos de linguagem de grande escala e a técnica RAG. A Seção 3 descreve a metodologia adotada, incluindo as etapas da DSRM aplicadas ao projeto. A Seção 4 apresenta os resultados e a

[1] Uma alucinação ocorre quando um modelo de linguagem gera informações que parecem plausíveis mas são falsas ou não são exatas. Ji *et al.* (2023)

discussão, com a descrição da arquitetura do sistema, do protótipo desenvolvido e dos testes realizados. A Seção 5 apresenta as considerações finais do trabalho.

2 Fundamentação Teórica

A Inteligência Artificial (IA) é compreendida como o campo científico e tecnológico dedicado à criação de sistemas capazes de perceber o ambiente, raciocinar, aprender e agir de forma racional e autônoma. Russell e Norvig (2021) definem a IA como o estudo de agentes que recebem percepções do ambiente e realizam ações orientadas a objetivos, buscando sempre o melhor resultado possível conforme o contexto e as incertezas envolvidas. Essa perspectiva destaca a racionalidade como o princípio unificador da inteligência: um agente é considerado inteligente não apenas quando imita o comportamento humano, mas quando age corretamente frente às informações disponíveis. A IA evoluiu por quatro grandes abordagens: agir humanamente, associada ao Teste de Turing e à imitação de respostas humanas; pensar humanamente, inspirada na psicologia cognitiva; pensar racionalmente, ligada à tradição lógica e filosófica; e agir racionalmente, que fundamenta o conceito contemporâneo de agentes inteligentes Russell e Norvig (2021).

O Aprendizado de Máquina (*Machine Learning* — ML) é o subcampo da IA responsável por desenvolver sistemas que melhoram automaticamente seu desempenho com base na experiência. Para Russell e Norvig (2021), trata-se de um processo em que algoritmos ajustam seus parâmetros conforme os dados disponíveis, de modo a reconhecer padrões e realizar previsões. O aprendizado pode ocorrer de três formas principais: supervisionado, quando o modelo aprende a partir de exemplos rotulados; não supervisionado, quando busca estruturas ou agrupamentos ocultos em dados não rotulados; e por reforço, quando aprende por tentativa e erro, maximizando recompensas de longo prazo.

Essa estrutura conceitual aproxima o aprendizado de máquina do comportamento humano onde há um processo contínuo de ajuste entre percepção, erro e correção. O uso de redes neurais artificiais, inspiradas no funcionamento do cérebro, permite que esses modelos atinjam desempenhos muito satisfatórios em tarefas como visão computacional, reconhecimento de voz e tradução automática.

Junto ao *Machine Learning* tem-se o *Deep Learning*, que consiste em redes neurais com múltiplas camadas capazes de extrair representações complexas e abstratas dos dados Russell e Norvig (2021). Essa abordagem tornou-se a base para o desenvolvimento dos modelos generativos contemporâneos.

Os Modelos de Linguagem de Grande Escala (*Large Language Models* – LLMs) representam uma das principais evoluções do aprendizado profundo. Segundo Foster (2022), esses modelos são treinados com um grande conjunto de textos e utilizam arquiteturas baseadas em Transformers, capazes de processar sequências e aprender relações de dependência contextual entre palavras.

Os LLMs, como GPT, LLaMA, PaLM ou Mistral, não armazenam conhecimento de forma estática, mas modelam a probabilidade de ocorrência de tokens em sequência, prevendo o próximo elemento linguístico com base em padrões aprendidos. Essa abordagem confere ao modelo a habilidade de gerar textos coerentes, responder perguntas, traduzir e resumir informações, o que caracteriza o paradigma da IA generativa de linguagem Foster (2022).

Entre os principais modelos generativos destacam-se:

- Redes Neurais Profundas (Deep Neural Networks): são a base da maioria dos modelos de IA generativa. Elas consistem em múltiplas camadas de neurônios artificiais que aprendem a representar dados em diferentes níveis de abstração. Essas redes são treinadas usando grandes conjuntos de dados e algoritmos de otimização, como a descida do gradiente LeCun, Bengio e Hinton (2015);

- Redes Generativas Adversariais (GANs): as GANs são uma das técnicas mais populares e bem-sucedidas para geração de imagens. Como mencionado anteriormente, elas funcionam através da competição entre um gerador e um discriminador, onde o gerador tenta criar dados realistas e o discriminador tenta identificar se os dados são reais ou gerados. Esta dinâmica de jogo força o gerador a melhorar continuamente Goodfellow *et al.* (2014);

- Modelos Baseados em Transformadores: os transformadores, particularmente na forma de modelos como GPT e BERT, são essenciais para a geração de texto. Eles utilizam mecanismos de atenção para pesar a importância de diferentes partes do input, permitindo uma compreensão profunda e contextual do texto. Isso permite que esses modelos gerem texto que é gramaticalmente correto e contextualmente relevante Vaswani *et al.* (2017) e Brown *et al.* (2020);

- Autoencoders Variacionais (VAEs): os VAEs são usados para gerar novos dados a partir de uma representação latente aprendida. Eles consistem em um codificador que transforma os dados de entrada em uma distribuição latente e um decodificador que reconstrói os dados a partir dessa distribuição. Isso é útil para tarefas como geração de imagens e compressão de dados Kingma e Welling (2013);

- Modelos de Difusão: utilizam técnicas de aprendizado profundo para gerar imagens a partir de descrições textuais. Eles funcionam invertendo o processo de difusão de ruído em dados, permitindo a criação de imagens detalhadas e precisas a partir de prompts textuais Rombach *et al.* (2022).

Nos últimos anos, a evolução dos modelos open-source como LLaMA Touvron *et al.* (2023), Mistral, Gemma e Falcon permitiu a execução de IA generativa em ambientes locais (on-premise), sem dependência de servidores externos. Essa abordagem viabiliza a criação de modelos de domínio específico, ajustados a bases de dados privadas (PDFs, sites institucionais, vídeos e áudios transcritos), preservando a privacidade e soberania informacional.

Essas implementações locais costumam adotar a técnica de RAG (*Retrieval-Augmented Generation*), na qual um banco vetorial armazena os textos de referência e o modelo recupera fragmentos relevantes no momento da geração da resposta. Dessa forma, a IA não apenas replica padrões genéricos, mas responde de modo contextualizado, coerente com as informações que fundamentam o domínio de conhecimento na área. Ao operar localmente e com fontes validadas, uma IA generativa de domínio busca preservar não apenas a privacidade, mas também a fidelidade semântica ao pensamento de referência.

Formalmente, o RAG pode ser descrito como um modelo generativo condicional que integra uma etapa intermediária de recuperação Lewis *et al.* (2020). Dado um *input* x, o sistema recupera um conjunto de documentos z a partir de uma base de conhecimento externa $\mathcal{Z}$, e a resposta y é gerada com base na concatenação de x e z:

$$p(y \mid x) = \sum_{z \in \mathcal{Z}} p(z \mid x) \cdot p(y \mid x, z) \tag{1}$$

onde $p(z \mid x)$ é a distribuição de probabilidade sobre os documentos recuperados onde é tipicamente modelada por um *encoder* de recuperação densa e $p(y \mid x, z)$ é a probabilidade de geração dado o contexto enriquecido.

O RAG é definido em três estágios: indexação, recuperação e geração. A indexação transforma os documentos em vetores e os armazena em um banco de dados especializado. A recuperação seleciona os fragmentos mais relevantes para a consulta, utilizando técnicas de similaridade semântica. A geração utiliza um modelo generativo para produzir a resposta final, integrando as informações recuperadas de forma coerente.

A indexação é a etapa fundamental para garantir a eficiência e relevância do sistema. Antes que qualquer consulta seja processada, os documentos da base de conhecimento precisam ser transformados em representações vetoriais densas (*embeddings*) e indexados em um banco de dados especializado.

A segmentação (*chunking*) divide os documentos em fragmentos menores, normalmente entre 128 e 1.500 caracteres respeitando limites de contexto do modelo de *embedding* e maximizando a granularidade da recuperação. A estratégia de sobreposição (*overlap*) entre fragmentos consecutivos serve para garantir que conceitos que transitam entre dois fragmentos não percam contexto ao serem recuperados isoladamente. A codificação vetorial mapeia cada fragmento para um espaço de alta dimensão (768 a 1.536 dimensões) por meio de modelos como bge-m3, nomic-embed-text ou all-MiniLM-L6-v2 Reimers e Gurevych (2019). O mesmo modelo precisa ser utilizado para codificar tanto os documentos quanto as consultas, pois a comparação de similaridade ocorre nesse espaço compartilhado. A indexação vetorial armazena esses vetores em sistemas otimizados para busca por similaridade, como FAISS (*Facebook AI Similarity Search*), Pinecone, Weaviate ou Chroma. A recuperação aproxima-se de um problema de *Maximum Inner Product Search* (MIPS), resolvido de forma sub-linear por estruturas como grafos HNSW (*Hierarchical Navigable Small World*), atingindo latências inferiores a um milissegundo em

produção Systematic Literature Review (2025).

Ao receber uma consulta, o sistema pode operar de três formas:

- recuperação esparsa: baseia-se em correspondência léxica via algoritmos como BM25, eficiente para termos técnicos específicos;
- recuperação densa utiliza modelos como o *Dense Passage Retrieval* (DPR) Karpukhin *et al.* (2020), calculando similaridade por produto escalar ou cosseno e capturando relações semânticas que transcendem a correspondência de tokens;
- recuperação híbrida combina as duas modalidades, consolidando os *rankings* resultantes via técnicas como *Reciprocal Rank Fusion* (RRF) Cormack, Clarke e Buettcher (2009).

Com os fragmentos recuperados, inicia-se a etapa de geração, onde os *top-k* fragmentos mais relevantes são concatenados à consulta original e fornecidos como contexto ao modelo generativo Foster (2022). A qualidade da geração depende criticamente da relevância dos fragmentos recuperados e da capacidade do LLM de integrar informações contextuais de forma coerente. O RAG é especialmente eficaz para domínios de conhecimento específico, onde a base de dados pode ser verificada para garantir a fidelidade e precisão das respostas, minimizando o risco de alucinações e promovendo a transparência na geração de conteúdo.

3 Método

As IAs generativas de prompt para texto têm se destacado nos últimos anos, com várias iterações e aprimoramentos feitos por diferentes empresas e instituições de pesquisa. Uma das principais tecnologias é a série GPT (*Generative Pre-trained Transformer*) desenvolvida pela OpenAI. A evolução começou com o GPT-2, que foi capaz de gerar textos coerentes a partir de um prompt inicial. Este modelo tinha uma capacidade considerável de compreender o contexto e produzir respostas relevantes. No entanto, o avanço mais significativo veio com o GPT-3, que conta com 175 bilhões de parâmetros, possibilitando a criação de textos mais detalhados e diversificados, aprimorando substancialmente a capacidade de seguir instruções complexas e gerar respostas contextualmente precisas. A evolução continuou com o GPT-4, que trouxe ainda mais melhorias em termos de compreensão e geração de textos, além de avanços em segurança e mitigação de vieses, com uma estimativa de trilhões de parâmetros, embora o número exato não tenha sido oficialmente especificado.

Outra tecnologia importante é a T5 (*Text-To-Text Transfer Transformer*) da Google, que converte qualquer tarefa de NLP (*Natural Language Processing*) em uma tarefa de preenchimento de texto. O T5-Base é um modelo versátil que pode ser usado em diversas aplicações de linguagem. Já o T5-Large, uma versão mais robusta, melhora o desempenho em tarefas mais complexas. Variantes maiores como T5-3B e T5-11B, com 3 bilhões e 11 bilhões de parâmetros respectivamente, demonstram a escalabilidade e a capacidade de adaptação deste modelo a uma ampla gama de tarefas.

O BERT (*Bidirectional Encoder Representations from Transformers*), também desenvolvido pela Google, é outro modelo de destaque. Focado na compreensão bidirecional de texto, BERT permite que o contexto de uma palavra seja analisado com base em todas as palavras ao seu redor. Variantes como RoBERTa (*Robustly Optimized BERT Approach*) ajustaram hiperparâmetros para melhorar ainda mais o desempenho na geração de texto.

As versões de GPT apresentam diferenças significativas em termos de arquitetura, capacidade e funcionalidades. A primeira versão, GPT-1, lançada em 2018, possuía 117 milhões de parâmetros e era focada na tarefa de modelagem de linguagem. O GPT-2, lançado em 2019, aumentou para 1,5 bilhão de parâmetros, permitindo uma melhor compreensão e geração de texto. Em 2020, o GPT-3 trouxe um salto significativo com 175 bilhões de parâmetros, ampliando substancialmente a gama de aplicações. Em 2023, o GPT-4 foi lançado com estimados trilhões de parâmetros, oferecendo melhorias em compreensão e geração contextual, além de maior ênfase em segurança e mitigação de vieses.

Além desses, há outros modelos importantes como Meena, lançado em 2020 pelo Google, projetado para gerar diálogos mais humanos e naturais, utilizando 2,6 bilhões de parâmetros. E o Switch Transformer, lançado em 2021, um modelo de grande escala que utiliza um mecanismo de roteamento para ativar apenas partes do modelo durante a inferência, melhorando a eficiência computacional com até 1,6 trilhões de parâmetros.

As IAs generativas de texto têm evoluído muito. Os modelos estão cada vez maiores e mais potentes, com um número crescente de parâmetros que lhes permite compreender e gerar texto de forma muito mais eficiente. Além disso, há um foco cada vez maior em tornar esses sistemas mais seguros e justos, minimizando vieses indesejados.

Esses avanços não apenas melhoram as tarefas simples de processamento de linguagem natural, mas também permitem realizações mais complexas, como a geração de código e o suporte conversacional avançado. Cada nova versão desses modelos traz melhorias significativas, aumentando a funcionalidade e abrindo novas possibilidades para inovações em várias áreas.

Em relação ao tipo de pesquisa, esta é uma pesquisa exploratória com aplicação de Design Thinking e/ou Design Science. O tipo de procedimento de pesquisa é o *Design Science Research Methodology* (DSRM), proposta por Peffers *et al.* (2007) para o desenvolvimento e avaliação de um artefato tecnológico. Essa metodologia é especialmente adequada a pesquisas que visam criar e validar soluções aplicáveis para problemas reais, combinando rigor científico e relevância prática Hevner *et al.* (2004).

De acordo com Peffers *et al.* (2007), a DSRM "incorpora princípios, práticas e procedimentos necessários para conduzir pesquisas baseadas em design", estruturando-se em seis atividades fundamentais:

1. Identificação e motivação do problema;
2. Definição dos objetivos da solução;
3. Design e desenvolvimento do artefato;

4. Demonstração;
5. Avaliação;
6. Comunicação.

A metodologia é adequada com os propósitos desta pesquisa, que busca conceber, desenvolver e validar o modelo de Inteligência Artificial Generativa AntonIA, voltado ao entendimento, divulgação e difusão dos fundamentos da Ontopsicologia. A DSRM permite integrar os princípios da ciência do artificial Simon (1996) com um processo sistemático de criação e avaliação de artefatos de software.

Os modelos generativos atuais são baseados em grandes conjuntos de dados sem uma clareza no que se refere sua origem. Isto leva a limitações quanto à fidelidade conceitual e interpretativa. Dessa forma, a motivação deste trabalho reside na necessidade de criar um modelo de IA contextualizado, que seja semanticamente consistente com o pensamento ontopsicológico e contribua para a sua difusão científica e educacional.

Para que essa ação ocorra de forma satisfatória, serão utilizados dados públicos encontrados em revistas científicas como Saber Humano, Revista Brasileira de Ontopsicologia, vídeos gravados no youtube pelo Acadêmico Professor Antonio Meneghetti, trabalhos científicos armazenados no repositório digital da Antonio Meneghetti Faculdade, além de outro artigos e textos científicos que possam contribuir sem gerar alucinações no modelo de IA desenvolvido. Tendo esses dados como base, eles são utilizados em um modelo baseado na arquitetura Large Language Model (LLM) e integrado a um mecanismo de busca aumentada por recuperação (*Retrieval-Augmented Generation* – RAG).

A Figura 1 apresenta o modelo metodológico adaptado da DSRM aplicado neste trabalho, representando suas seis etapas e os ciclos de refinamento iterativo. Esse processo assegura o cumprimento dos princípios de rigor e relevância propostos por Peffers *et al.* (2007), garantindo que o artefato resultante constitua uma contribuição científica e tecnológica validada, em consonância com a perspectiva descrita por Simon (1996).

As duas primeiras etapas da DSRM, que envolvem a identificação do problema e a definição dos objetivos, foram desenvolvidas no projeto deste trabalho e estão descritas nesta seção. As etapas de projeto, desenvolvimento, demonstração, avaliação e comunicação estão descritas na Seção 4.3.

O problema central identificado é a ausência de uma ferramenta de IA generativa capaz de responder sobre os fundamentos da Ciência Ontopsicológica com fidelidade semântica aos conteudos de referência da Ciência Ontopsicológica evitando os riscos de alucinação e diluição conceitual típicos de modelos generalistas Bommasani *et al.* (2022).

O problema de pesquisa deste trabalho é: *Como construir e treinar um modelo de Inteligência Artificial Generativa para o entendimento, divulgação e difusão da Ontopsicologia, preservando a integridade da Ciência Ontopsicológica e operando integralmente em infraestrutura institucional local?* A partir do problema de pesquisa, foi definido o objetivo geral que é de

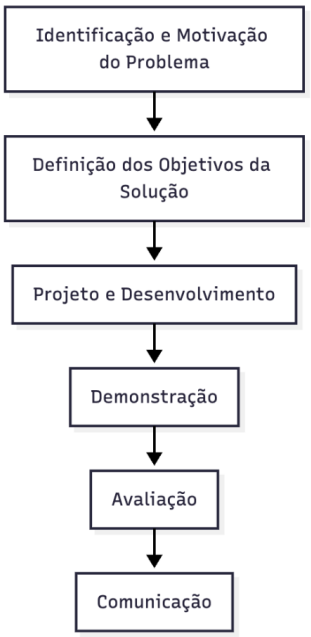


Figura 1: Modelo metodológico adaptado da DSRM aplicado ao projeto AntonIA.

construir um modelo e um protótipo de Inteligência Artificial Generativa para o entendimento, divulgação e difusão da Ontopsicologia.

A partir do objetivo geral, foram definidos os seguintes objetivos específicos:

1. Autonomia e segurança: construir um modelo generativo local, baseado em arquiteturas *open source*, como LLaMA, Mistral ou Falcon, pode garantir autonomia tecnológica e segurança das informações, visto que os dados utilizados para treinamento e inferência permanecem sob total controle da instituição;

2. Fidelidade informacional: aplicar técnicas de *fine-tuning* ou de *Retrieval-Augmented Generation* (RAG) sobre uma base de conhecimento teórica específica possibilita preservar a coerência e conceitual do domínio, evitando distorções semânticas e interpretações indevidas nos processos generativos que muitas vezes ocorrem em modelos generalistas Bommasani *et al.* (2022);

3. Eficiência contextual: gerar respostas mais precisas e relevantes, pois modelos locais, ainda que possuam menor capacidade de generalização, podem alcançar maior precisão e relevância contextual quando treinados sobre um conjunto de dados restrito e relevante ao contexto empregado resultando uma melhor aderência conceitual às necessidades do domínio estudado LeCun (2023), Hevner *et al.* (2004) e Peffers *et al.* (2007).

O protótipo desenvolvido neste trabalho é uma aplicação web de IA generativa baseada em RAG, construída sobre Django Holovaty e Kaplan-Moss (2009), LangChain Chase (2022) e modelos LLM executados localmente via Ollama Ollama (2023). O sistema foi projetado para

responder a consultas sobre conceitos da Ontopsicologia, utilizando a base de conhecimento cadastrada na aplicação com fontes verificadas. A arquitetura do sistema inclui um pipeline de ingestão multilíngue, um mecanismo de recuperação baseado em FAISS e um módulo de geração que integra as informações recuperadas para produzir respostas coerentes e contextualizadas. A Tabela1 sintetiza os artefatos tecnológicos produzidos neste trabalho segundo a classificação proposta por Hevner *et al.* (2004).

A demonstração do AntonIA consistiu na aplicação do sistema a situações reais de geração textual sobre conceitos da Ontopsicologia, com o objetivo de observar se o artefato produz respostas coerentes com as fontes originais indexadas. A análise das repostas geradas estão descritas na Seção 4.3. Foram definidas duas perguntas de teste para validar a solução. A primeira pergunta foi "o que é Ontopsicologia?"e a segunda pergunta foi "como posso utilizar a Ontopsicologia para ter uma vida melhor?".

Tabela 1: Artefatos produzidos no projeto AntonIA segundo a taxonomia Hevner *et al.* (2004)

Tipo de artefato	Instância no projeto	Descrição
Construção	Workspace RAG Ontopsicologia	Base de conhecimento indexada com fontes verificadas: artigos da Revista Saber Humano, Revista Brasileira de Ontopsicologia, vídeos transcritos que foram gravados em vida pelo Acadêmico Professor Antonio Meneghetti que se encontram disponível no YouTube e outros textos em PDF.
Modelo	Pipeline de ingestão multilíngue	Fluxo de seis etapas: extração, detecção de idioma, chunking (1.200 chars, overlap 150), tradução literal via Qwen2.5:7b, correção terminológica (TermCorrection) e indexação dupla PostgreSQL + FAISS.
Método	RAG Chain com retriever FAISS	Recuperação dos cinco fragmentos mais similares por similaridade de cosseno; geração restrita a base de conhecimento (proibição de conhecimento externo no prompt); cache de instâncias OllamaLLM para eficiência em produção.
Instância	Protótipo AntonIA	Aplicação Django com interface conversacional, módulo de ingestão por PDF/URL/YouTube, painel administrativo e sistema de workspaces isolados por índice FAISS.

Fonte: elaborado pelo autor com base em Hevner *et al.* (2004).

4 Resultados e Discussão

O protótipo AntonIA é uma aplicação web de IA generativa baseada em RAG, construída sobre Django, LangChain e modelos LLM executados localmente via Ollama. O sistema foi projetado para responder a consultas sobre conceitos da Ontopsicologia, utilizando a base de

conhecimento cadastrada na aplicação com fontes verificadas. A arquitetura do sistema inclui um pipeline de ingestão multilíngue, um mecanismo de recuperação baseado em FAISS e um módulo de geração que integra as informações recuperadas para produzir respostas coerentes e contextualizadas.

A base de conhecimento da AntonIA é composta por fontes verificadas e validadas: publicações da *Revista Saber Humano*[2] e da *Revista Brasileira de Ontopsicologia*[3] ; vídeos gravados em vida pelo Acadêmico Professor Antonio Meneghetti que se encontram disponível no YouTube; trabalhos científicos do repositório digital da AMF [4]. O sistema suporta três tipos de fontes: arquivos PDF, páginas da *web* e vídeos do YouTube. Essas fontes são unificadas em um *pipeline* de processamento.

4.1 Arquitetura

A Figura 2 apresenta o processo de uso da aplicação AntonIA., organizado em torno de um ambiente local sem dependência de serviços em nuvem no qual todos os componentes operam localmente. O fluxo de processamento inicia-se com a pergunta do usuário (prompt), submetida por meio da Interface da Aplicação. Esse componente é o único ponto de contato entre o usuário e o sistema, e tem duas funções simultâneas: encaminhar a consulta ao orquestrador RAG e registrar o Histórico da Conversa em banco de dados relacional (PostgreSQL), possibilitando que o contexto das interações anteriores seja considerado nas respostas seguintes. O Orquestrador RAG realiza a recuperação dos dados da base de conhecimento e os encaminha ao modelo LLM local, que gera a resposta e a devolve ao usuário como resposta final.

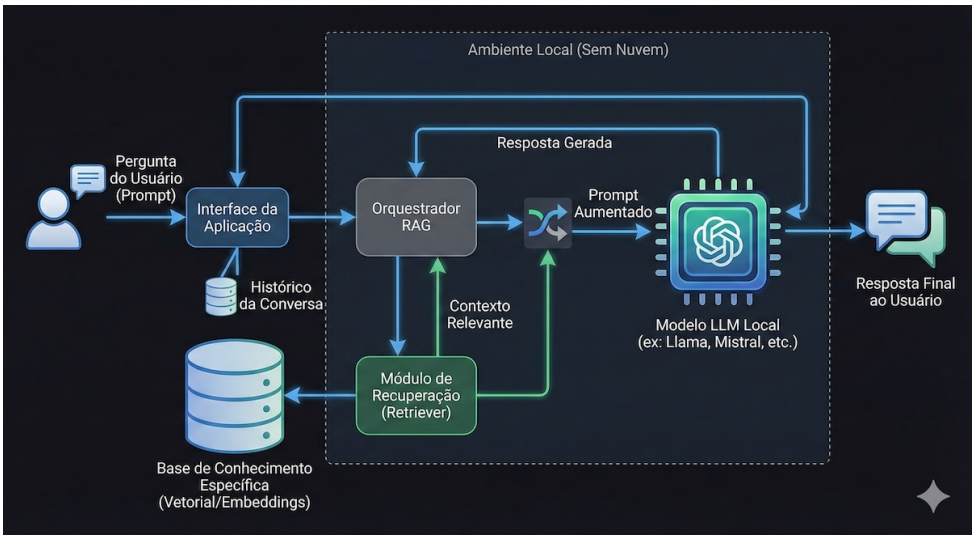


Figura 2: Processo de uso da aplicação AntonIA.
Fonte: elaborado pelo autor.

[2] https://revistas.amf.edu.br/index.php/saberhumano/
[3] https://revbo.emnuvens.com.br/revbo
[4] https://repositorio.amf.edu.br/

O usuário tem a possibilidade de consultar a base de conhecimento por meio do prompt e também de acessar o histórico de suas conversas com a aplicação. Dessa forma, pode verificar tanto o registro completo das interações anteriores quanto as informações que estão sendo utilizadas para fundamentar as respostas geradas pela aplicação AntonIA.

Essa característica diferencia o sistema de uma consulta isolada e o aproxima de um assistente conversacional com memória de sessão. A consulta é então recebida pelo Orquestrador RAG (LangChain), componente central da arquitetura, responsável por coordenar todas as etapas do pipeline de recuperação e geração. O orquestrador aciona o Módulo de Recuperação (*Retriever*), que realiza uma busca por similaridade semântica no índice vetorial FAISS a Base de Conhecimento Específica (*Vetorial/Embeddings*) retornando os fragmentos documentais mais relevantes para a consulta recebida.

Essa base é construída previamente pelo pipeline de ingestão, contendo exclusivamente fontes verificadas e validadas. Com os fragmentos recuperados o contexto retorna ao orquestrador, que os combina com a consulta original para construir o Prompt Aumentado que pode ser definido como uma entrada enriquecida que contém tanto a pergunta do usuário quanto as evidências textuais recuperadas da base de conhecimento.

Essa etapa materializa o princípio fundamental do RAG o modelo generativo não precisa "saber"a resposta de antemão; ele precisa apenas sintetizá-la a partir do contexto fornecido Lewis *et al.* (2020). O prompt aumentado é então processado pelo Modelo LLM Local (Llama, Mistral, Qwen, Gemma), executado integralmente na infraestrutura local via Ollama. O modelo gera a Resposta Gerada, que retorna ao orquestrador para tratamento final e é entregue ao usuário como Resposta Final.

Todo esse ciclo da pergunta à resposta ocorre dentro do perímetro do Ambiente Local (Sem Nuvem), representado na Figura 2 pela borda tracejada que delimita o espaço de processamento protegido. Destaca-se três aspectos arquiteturais desta solução:

- O ambiente local garante que nenhum dado trafegue para fora da rede institucional, eliminando riscos de exposição de propriedade intelectual;

- A separação entre o módulo de recuperação (factual) e o modelo LLM (linguístico) permite flexibilidade modular, característica do RAG avançado Gao *et al.* (2023);

- A presença explícita do Histórico da Conversa evidencia que o sistema foi projetado para interações continuadas, essencial para um assistente pedagógico voltado à difusão da Ciência Ontopsicológica.

A separação de responsabilidades entre o Módulo de Recuperação (responsável pela relevância factual) e o Modelo LLM (responsável pela coerência linguística) é uma decisão de design que permite substituir qualquer um dos dois componentes de forma independente, conferindo à arquitetura uma estrutura modular do RAG avançado Gao *et al.* (2023). A funcionalidade do Histórico da Conversa como componente persistido demonstra que o sistema foi projetado

para interações continuadas, não apenas consultas pontuais requisito essencial para um assistente pedagógico voltado à difusão da Ciência Ontopsicológica.

O *pipeline* de ingestão percorre as seguintes etapas que estão descritos na Figura 3:

1. **Extração de texto**: via *pypdf* para PDFs, *Beautiful Soup* para HTML e *youtube transcript api* para vídeos;
2. **Detecção de idioma**: operada sobre os primeiros dois mil caracteres do texto via *langdetect*;
3. **Fatiamento** (*chunking*): via *Recursive Character Text Splitter* do LangChain, com janela de 1.200 caracteres e sobreposição de 150 — configuração que garante que conceitos na fronteira entre fragmentos não percam contexto durante a recuperação;
4. **Tradução literal** para o português: via modelo *Qwen2.5:7b* executado localmente pelo Ollama, com *prompt* que instrui o modelo a traduzir `EXATAMENTE, SEM INTERPRETAR`, proibindo adição, remoção ou reordenação de conteúdo;
5. **Correção terminológica** automática: assegurando que termos técnicos da Ontopsicologia sejam sempre grafados de forma consistente (*Term Correction*);
6. **Persistência dupla**: armazenando tanto o texto original quanto a tradução para o português no banco relacional PostgreSQL e no índice FAISS.

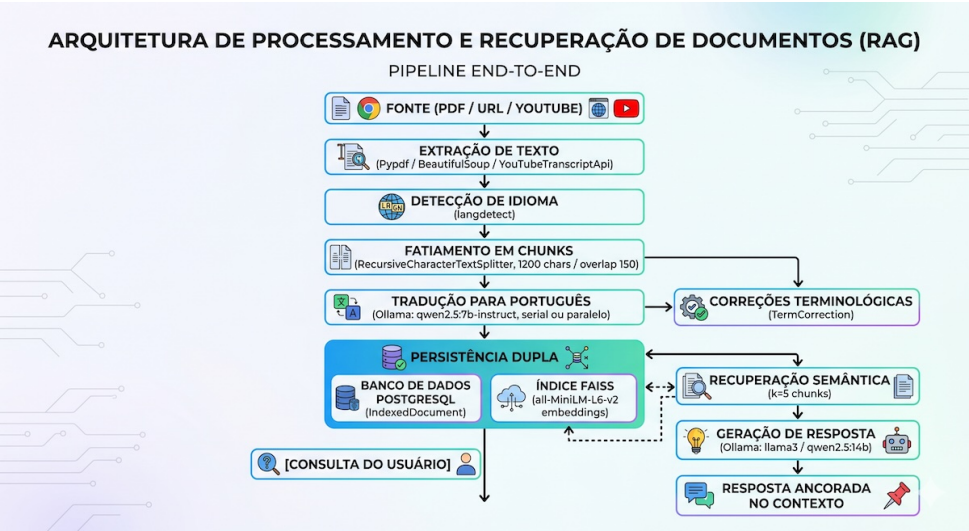


Figura 3: Arquitetura RAG da aplicação AntonIA.
Fonte: elaborado pelo autor.

O modelo de *embeddings* adotado é o sentence-transformers all-MiniLM L6-v2, disponibilizado via Hugging Face BgeEmbeddings do LangChain. Trata-se de um modelo compacto (aproximadamente 90 MB), capaz de operar inteiramente *offline* após o *download* inicial, com

qualidade consistente para textos em português . A configuração normalize embeddings garante que os vetores sejam normalizados à unidade, tornando a busca por similaridade de cosseno equivalente ao produto interno — uma otimização relevante para FAISS Systematic Literature Review (2025).

Cada *workspace* mantém seu próprio par de arquivos index.faiss e index.pkl em subdiretório dedicado, permitindo que múltiplos corpora coexistam na mesma instalação sem interferência entre seus índices. A persistência em disco ocorre após cada lote de documentos, assegurando que falhas durante a ingestão não resultem em perda do trabalho já realizado.

Para a geração de respostas, o sistema recorre a modelos LLMs de código aberto executados localmente via Ollama. A disponibilidade de modelos *open-source* como LLaMA Touvron *et al.* (2023) viabilizou essa abordagem *on-premise*. Quatro famílias de modelos foram avaliadas para o contexto do projeto , conforme descrito na Tabela 2:

Tabela 2: Comparativo arquitetural dos modelos fundacionais avaliados para o sistema AntonIA
Fonte: elaborado pelo autor.

Modelo	**Parâmetros**	**Características principais para RAG**
Qwen2.5 (7b/14b)	7B – 14B	Fluência em >29 idiomas; janela de contexto de 128k tokens; indicado para tradução técnica e extração estruturada (Alibaba Cloud).
Llama 3.1:8b	8B	Vocabulário nativo de 128k tokens; alta eficiência energética; redução de passagens computacionais para textos multilíngues (Meta) Touvron *et al.* (2023).
Gemma2:9b	9B	Destilação de conhecimento de modelos Gemini; alta performance gramatical por parâmetro; indicado como gerador de linguagem natural (Google).
Mistral-Nemo:12b	12B	Projetado para 128k tokens sem degradação de atenção (*attention decay*); padrão para *pipelines* RAG com múltiplos fragmentos (Mistral AI/NVIDIA).

O módulo de resposta implementa um *retriever* FAISS configurado para retornar os cinco fragmentos mais similares à consulta do usuário. O *prompt* de geração instrui o modelo a responder exclusivamente com base nos trechos fornecidos, proibindo o uso de conhecimento externo a base de conhecimento indexada. Essa restrição é a essência do modelo RAG aplicado ao domínio especializado: ao forçar o modelo a ancorar suas respostas na base de conhecimento da Ontopsicologia, reduz-se drasticamente a incidência de alucinações sobre os conceitos da Ontopsicologia Meneghetti (2022).

Um mecanismo de *cache* de instâncias OllamaLLM evita o *overhead* de inicialização em cada consulta, contribuindo para a viabilidade do sistema em ambiente de produção local. A lógica de treinamento sobre dados verificados é coerente com o que LeCun (2022) identifica como o caminho para sistemas autônomos mais confiáveis.

Modelos locais de código aberto, embora possuam menor capacidade de generalização que seus equivalentes proprietários de maior escala, podem superar esses últimos em precisão e relevância contextual quando combinados com mecanismos de recuperação sobre bases de conhecimento específicas e restritas LeCun (2023). Essa observação é diretamente aplicável ao caso AntonIA: a especificidade da base de conhecimento,não o tamanho do modelo, é determinante para a qualidade das respostas em domínios de conhecimento especializado . A eficiência contextual de modelos locais ajustados a domínios restritos já foi destacada na literatura Hevner *et al.* (2004) e Peffers *et al.* (2007), e os resultados de benchmarks em recuperação densa sobre bases especializadas confirmam essa premissa Systematic Literature Review (2025).

Uma das barreiras mais significativas para a adoção sustentável de sistemas de IA generativa em instituições de ensino é o modelo de precificação por *tokens* adotado pelos provedores de APIs externas. Nesse modelo, cada consulta ao sistema incluindo o contexto recuperado pelo RAG (tipicamente 2.000 a 8.000 tokens por consulta) e a resposta gerada criando um custo unitário que, em volumes de uso acadêmico real, pode acumular despesas mensais substanciais e imprevisíveis Galileo AI (2026).

Uma implementação RAG típica com GPT-4 (na versão de 128k de contexto), processando consultas com 5 fragmentos recuperados de 1.200 caracteres cada, resulta em aproximadamente 3.000 tokens de entrada por consulta Galileo AI (2026). A realização de 100 consultas diárias que é volume modesto para um chatbot institucional pode gerar um custo mensal com tokens de entrada e saída pode ultrapassar centenas de dólares, além de variações conforme políticas de preço do fornecedor, sem qualquer controle institucional sobre esse gasto.

O sistema AntonIA, ao executar integralmente em infraestrutura local, opera com custo de operação igual a zero: uma vez que o *hardware* está disponível e os modelos estão instalados, o número de consultas não impacta o orçamento operacional. O custo é unicamente o do *hardware* (servidor ou estação de trabalho com GPU ou RAM suficiente para os modelos escolhidos) e de energia elétrica.

Além da previsibilidade financeira, a arquitetura local confere independência tecnológica: a descontinuação, mudança de preço ou alteração de política de uso de um provedor externo não afeta a operação do sistema. Essa independência é especialmente relevante para instituições acadêmicas, cujo planejamento orçamentário opera em ciclos anuais incompatíveis com a volatilidade do mercado de APIs de IA generativa.

4.2 Protótipo AntonIA

O protótipo AntonIA foi desenvolvido como um sistema baseado em Inteligência Artificial Generativa de prompt por meio de textos, no qual o usuário realiza uma pergunta através de um prompt em uma interface web. Essa interface web contempla, além dos dados do usuário, o histórico de suas conversas. Com essas informações e com o prompt fornecido pelo usuário, a consulta é enviada ao Orquestrador RAG, que realiza a recuperação da base de conhecimento específica.

O conteúdo recuperado é encaminhado ao modelo LLM local, que, com base nessas informações, gera a resposta generativa e a devolve ao usuário como resposta final. Esse fluxo está descrito na Figura 2 como o processo de aplicação da ferramenta AntonIA. Além disso, o processo de indexação da ferramenta utiliza a técnica RAG, por meio da qual são buscadas as fontes de dados cadastradas na aplicação. Essas fontes podem ser URLs de sites, arquivos PDF ou vídeos do YouTube. A partir dessas fontes, executa-se um processo assíncrono de extração de dados, seguido das etapas de detecção de idioma, fatiamento em partes (*chunking*) e tradução para o português.

Conforme ilustrado na Figura 4, o documento resultante é salvo e persistido dentro do ambiente da aplicação. A persistência de dados é realizada por meio do banco de dados PostgreSQL; adicionalmente, são gerados os índices FAISS, utilizados pelos modelos de linguagem de grande escala (LLMs) durante as consultas. Com essas informações disponíveis na aplicação, o usuário seleciona, por meio do menu, o documento que deseja utilizar, conforme mostrado na Figura 4, e solicita seu processamento.

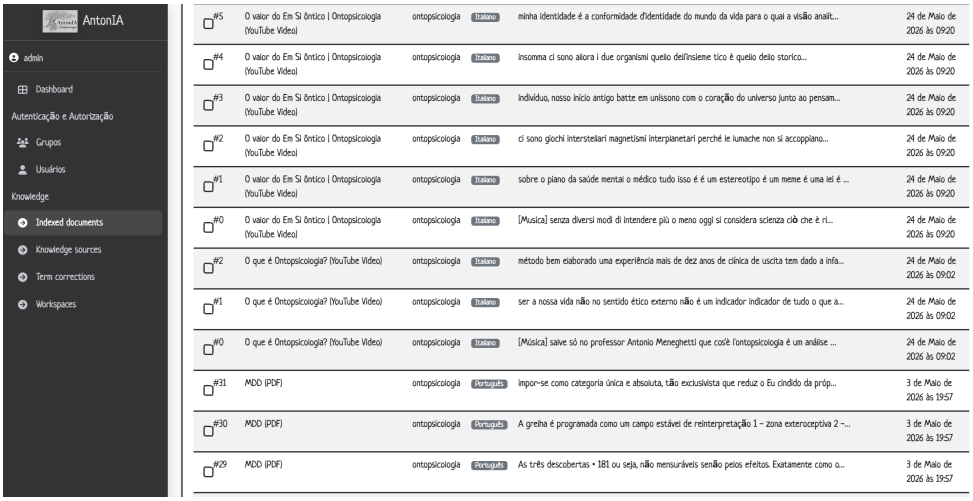


Figura 4: Lista de Documentos
Fonte: elaborado pelo autor.

Durante esse processo, realiza-se a detecção do idioma do documento. Muitos vídeos estão em língua italiana e são traduzidos para o português; ambas as versões são mantidas salvas para evitar problemas de conferência e rastreabilidade no processo de tradução. A aplicação suporta múltiplos idiomas, e a detecção de idioma é fundamental para que o processo de tradução seja realizado corretamente, garantindo a integridade semântica do conteúdo. A Figura 5 apresenta a lista de idiomas cadastrados na aplicação.

Com o texto extraído, o idioma detectado, o fatiamento em *chunks* concluído e as correções terminológicas aplicadas, as informações são salvas tanto no banco de dados quanto nos índices FAISS. As correções terminológicas são necessárias porque, no processo de tradução, ocorrem erros de interpretação: por exemplo, o termo Ontopsicologia possui três ou quatro formas distintas de representação a partir da tradução do italiano.

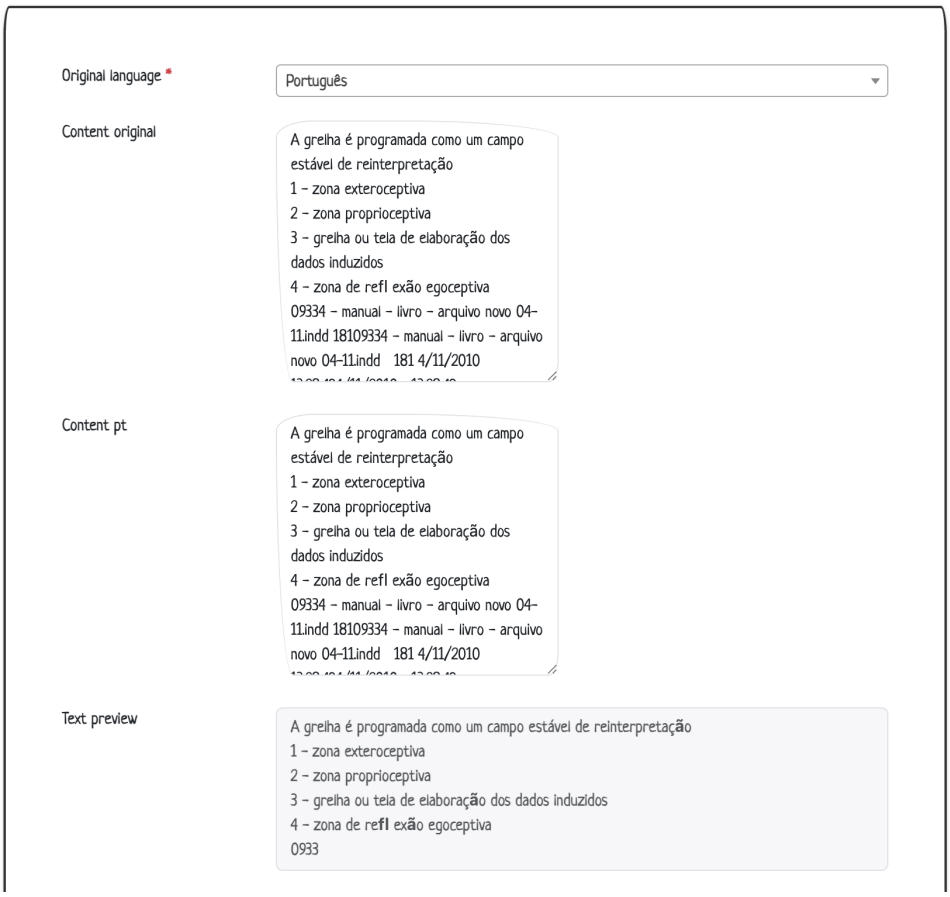


Figura 5: Lista de Linguagens
Fonte: elaborado pelo autor.

Por meio do processo de tradução e de correção terminológica, essa consistência é assegurada. Os termos de correção terminológica são cadastrados na aplicação conforme mostrado na Figura 6, as linguagens suportadas estão apresentadas na Figura 5 e os *workspaces*, que correspondem aos domínios de aplicação, estão mostrados na Figura 7.

Os índices FAISS são específicos por domínio de aplicação. Isso significa que a AntonIA, embora voltada à Ontopsicologia, foi desenvolvida de forma generalista, permitindo a criação de subtemas e subáreas. A partir desses subtemas ou áreas distintas, é possível manter documentos específicos para cada domínio e executar todos os processos RAG sobre eles, de forma independente, sem interferência entre os domínios. Com os ajustes realizados pela ferramenta, o novo conteúdo é liberado para incorporação à base de conhecimento e passa a fundamentar as respostas devolvidas aos usuários. Esse processo está descrito na Figura 3.

O protótipo AntonIA foi desenvolvido com o *framework* Django. Esse *framework* permite o desenvolvimento orientado a objetos com mapeamento objeto-relacional, no qual as principais informações de uma aplicação web são persistidas em banco de dados. No caso específico deste projeto, o protótipo é estruturado em cinco modelos de dados, cujas representações estão descritas na Figura 8. O modelo de dados é a representação da estrutura de dados da aplicação, ou seja, a forma como as informações são organizadas e armazenadas no banco de dados. Cada

Figura 6: Correção dos Termos
Fonte: elaborado pelo autor.

modelo corresponde a uma tabela no banco de dados, e os campos do modelo correspondem às colunas dessa tabela.

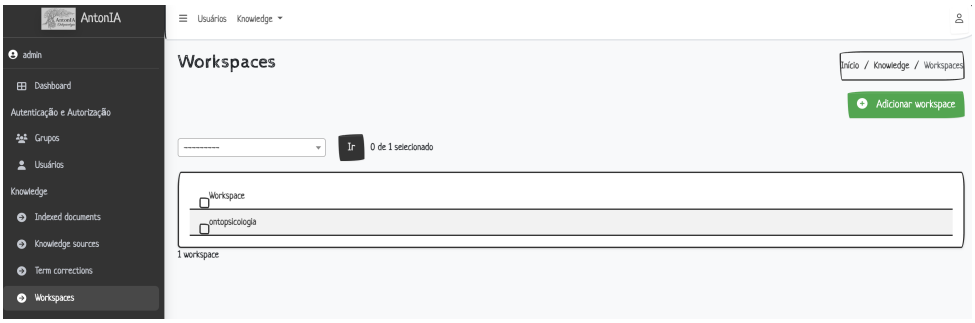


Figura 7: Lista de Workspaces
Fonte: elaborado pelo autor.

O primeiro modelo de dados é a estrutura de *Workspaces*: cada *workspace* representa um domínio de aplicação. O segundo modelo representa a correção de termos, definindo o termo original, o idioma de origem e o formato corrigido para o idioma de destino. O terceiro é o cadastro de linguagens, no qual se registram quais idiomas são suportados para que as traduções sejam realizadas corretamente. O quarto modelo corresponde à lista de documentos, que reúne os documentos originais cadastrados: URLs de sites, arquivos PDF e URLs de vídeos do YouTube. Com essas informações, é possível executar o processamento de cada documento, conforme descrito na Figura 4.

A Figura 9 apresenta a tela geral de administração. Além do cadastro de documentos e informações, a ferramenta dispõe de toda a estrutura organizacional da aplicação, incluindo controle de usuários, processo de autenticação e autorização. De acordo com as configurações da ferramenta, é possível disponibilizá-la para consulta aberta ou restringi-la a usuários previamente cadastrados.

A ferramenta contempla dois grupos de usuários: usuário comum e usuário administrador. O usuário administrador é responsável por carregar os documentos e organizar o conteúdo dentro da aplicação. Não há impedimento técnico para a criação de um terceiro perfil de usuário com permissão para carregar documentos e criar domínios específicos; nos testes realizados, contudo,

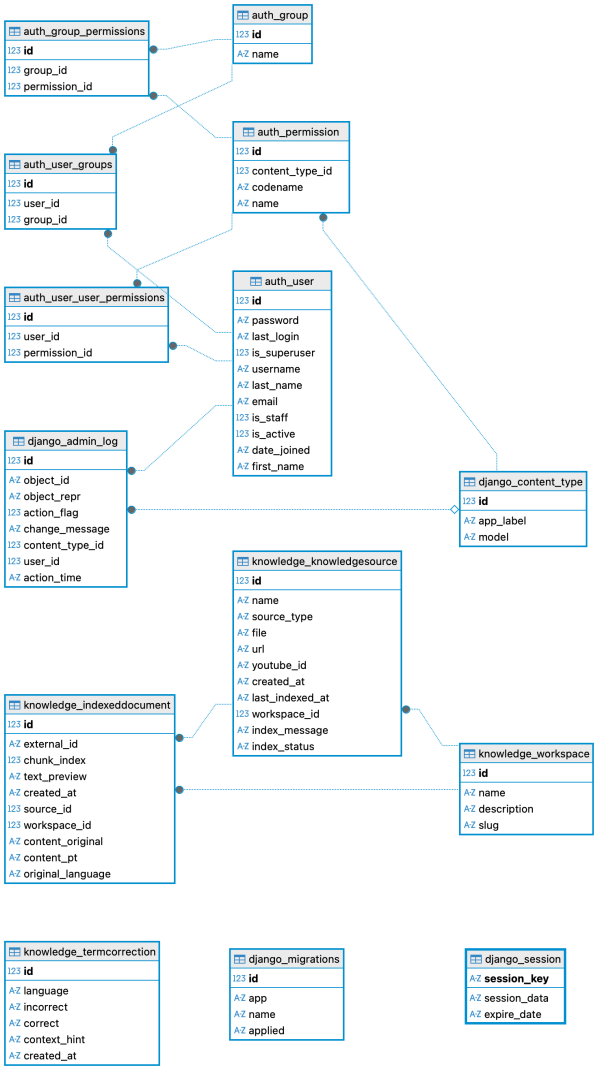


Figura 8: Modelo de Dados da aplicação AntonIA
Fonte: elaborado pelo autor.

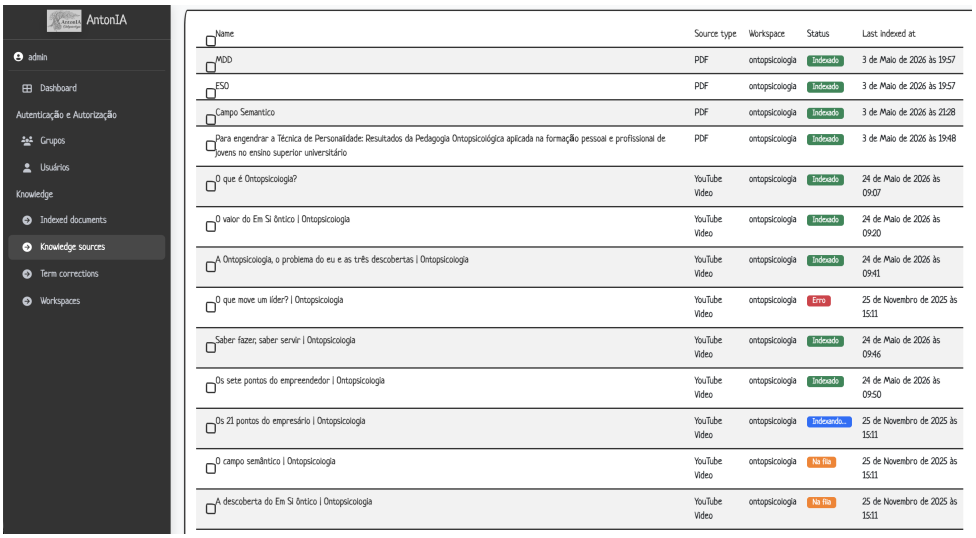


Figura 9: Tela Geral de Administração
Fonte: elaborado pelo autor.

essa configuração não foi aplicada.

A Figura 10 apresenta a base de conhecimento processada. A partir dos documentos cadastrados e armazenados conforme mostrado na Figura 4, esses documentos são processados, gerando o detalhamento completo da base de conhecimento. Todas as informações são armazenadas nos índices FAISS utilizados pelos LLMs e também no banco de dados relacional que dá suporte à aplicação AntonIA, mantendo relação direta com o *framework* Django.

A Figura 11 apresenta a tela de consulta. Nessa tela, o usuário pode realizar suas perguntas por meio do prompt e acessar o histórico de suas conversas. O histórico de conversas é uma funcionalidade importante para que o usuário possa acompanhar as informações utilizadas para fundamentar as respostas geradas pela aplicação AntonIA, além de permitir a consulta a interações anteriores. Essa funcionalidade é essencial para um assistente conversacional, pois permite que o usuário tenha uma visão completa de suas interações com a aplicação, facilitando a compreensão do processo de geração de respostas e a rastreabilidade das informações utilizadas.

4.3 Testes e Avaliação

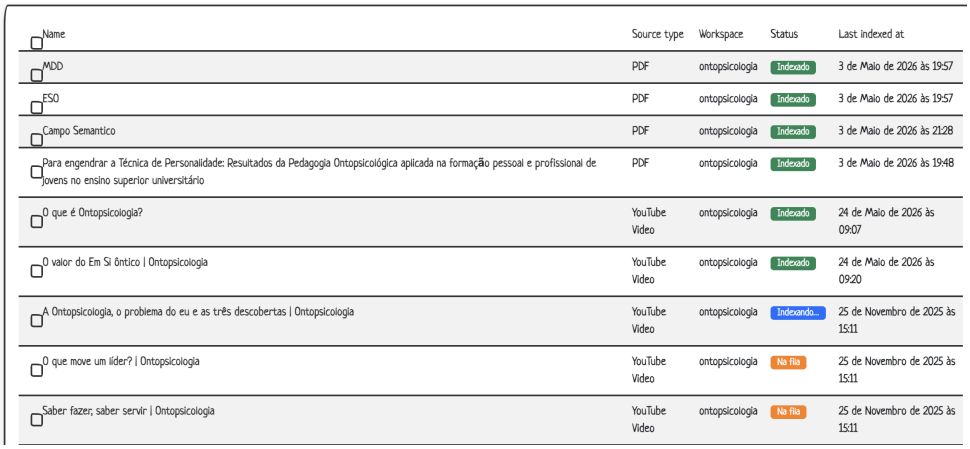
Para a execução dos testes e da avaliação desta aplicação, foi utilizado um computador MacBook Air M4 com 24 GB de memória RAM e 512 GB de armazenamento, executando o sistema operacional macOS Tahoe 26.

Nesse ambiente de testes, a aplicação AntonIA foi instalada e executada. Os testes realizados são descritos a seguir. Foram utilizados quatro modelos de linguagem, apresentados na Tabela 2: o modelo Qwen, LLaMA, Gemma e o modelo Mistral. Esses modelos foram selecionados porque, nas configurações descritas na Tabela 2, são executáveis em tempo hábil no ambiente de testes disponível.

O principal objetivo de utilizar esses modelos em um computador de configuração relativamente modesta, sem GPU dedicada e sem necessidade de conexão com LLMs externos, é validar a possibilidade de criar uma infraestrutura totalmente local, de baixo custo, sem a necessidade de pagamento por consumo de tokens, como ocorre nos principais modelos comerciais disponíveis atualmente.

Utilizando a aplicação AntonIA, foram realizados testes de consulta utilizando os quatro modelos de linguagem. O objetivo desses testes é avaliar a qualidade das respostas geradas por cada modelo, bem como a eficiência do processo de recuperação e geração de respostas com base na base de conhecimento da Ontopsicologia. A pergunta feita aos quatro modelos foi: "Defina o que é Ontopsicologia". As respostas geradas por cada modelo são apresentadas a seguir.

Modelo Qwen2.5:7b-instruct: *"A Ontopsicologia é uma análise científica e racional que faz a revisão crítica da consciência. É uma ciência que reimposte o sujeito humano em contato consciente e operativo com o mundo da vida ou com a realidade do ser, com o escopo de realização individual e integral. Trata-se de uma psicologia do ser que visa conhecer as premissas e experiências do homem realizado, do homem em alegria. É configurada como "a abertura de um modelo alternativo ao proceder científico que hoje está no mundo". A*



Name	Source type	Workspace	Status	Last indexed at
MDD	PDF	ontopsicologia	Indexado	3 de Maio de 2026 às 19:57
FSO	PDF	ontopsicologia	Indexado	3 de Maio de 2026 às 19:57
Campo Semantico	PDF	ontopsicologia	Indexado	3 de Maio de 2026 às 21:28
Para engendrar a Técnica de Personalidade: Resultados da Pedagogia Ontopsicológica aplicada na formação pessoal e profissional de jovens no ensino superior universitário	PDF	ontopsicologia	Indexado	3 de Maio de 2026 às 19:48
O que é Ontopsicologia?	YouTube Video	ontopsicologia	Indexado	24 de Maio de 2026 às 09:07
O valor do Em Si ôntico | Ontopsicologia	YouTube Video	ontopsicologia	Indexado	24 de Maio de 2026 às 09:20
A Ontopsicologia, o problema do eu e as três descobertas | Ontopsicologia	YouTube Video	ontopsicologia	Indexando...	25 de Novembro de 2025 às 15:11
O que move um líder? | Ontopsicologia	YouTube Video	ontopsicologia	Na fila	25 de Novembro de 2025 às 15:11
Saber fazer, saber servir | Ontopsicologia	YouTube Video	ontopsicologia	Na fila	25 de Novembro de 2025 às 15:11

Figura 10: Base de Conhecimento
Fonte: elaborado pelo autor.

Ontopsicologia possui um objeto de estudo específico, um método, um fim, uma visão de homem, descobertas científicas específicas e instrumentos de análise e intervenção, podendo ser aplicada em diversos campos do conhecimento".

Modelo LLaMA3.1:8b-instruct: *"A Ontopsicologia é uma análise científica, racional, que faz a revisão crítica da consciência. É um saber humano que visa reabilitar e promover a realização integral do homem em contato consciente e operativo com o mundo da vida ou com a realidade do ser. Tem como objetivo desenvolver o homem em criatividade e realização existencial integral, visando conhecer sobretudo as premissas e as experiências do homem realizado, do homem em alegria (Meneghetti, 2013b)".*

Modelo Gemma2:9b-instruct: *"A Ontopsicologia é uma análise científica, racional, que faz a revisão crítica da consciência. É a reproposta do conhecimento elementar para reimpostar o sujeito humano em contato consciente e operativo com o mundo da vida ou a realidade do ser, com o escopo de realização individual e integral".*

Modelo Mistral-Nemo:12b-instruct: *"A Ontopsicologia é uma análise científica, racional, que faz a revisão crítica da consciência com o escopo de realização individual e integral. É configurada como 'a abertura de um modelo alternativo ao proceder científico' e tem como visão de homem um ser sadio que busca e executa continuamente a realização existencial. (Trecho 2)".*

Analisando as respostas geradas pelos quatro modelos com todos utilizando a mesma base de dados, é possível identificar algumas diferenças importantes. O modelo Qwen2.5:7b-instruct, apresentou a resposta mais completa e aderente aos fundamentos da Ciência Ontopsicológica. Foi o único modelo capaz de enumerar explicitamente os elementos estruturais que definem a Ontopsicologia como ciência: objeto de estudo, método, fim, visão de homem, descobertas científicas específicas, instrumentos de análise e aplicabilidade em campos do conhecimento, com terminologia fiel ao Manual de Ontopsicologia (Meneghetti, 2022).

O modelo LLaMA 3.1:8b foi o único a incluir uma citação formal da fonte, referenciando diretamente Meneghetti (2013b). Essa rastreabilidade é relevante do ponto de vista da validação

científica, pois permite ao leitor verificar a origem da informação gerada. No entanto, a resposta produzida pelo LLaMA 3.1:8b mostrou-se mais restrita em amplitude: não abordou a estrutura formal da Ontopsicologia como ciência nem sua aplicabilidade multidisciplinar, resultando em uma síntese conceitualmente correta, porém menos abrangente do que a gerada pelo Qwen2.5.

O modelo Gemma 2:2b, por sua vez, demonstrou boa capacidade de recuperação da base de conhecimento, identificando os fragmentos relevantes para a consulta, mas sua capacidade de síntese contextual não foi suficiente para integrar todos os elementos conceituais em uma resposta completa. O resultado obtido é coerente com as limitações esperadas para um modelo dessa envergadura e indica que, embora o pipeline RAG funcione adequadamente na etapa de recuperação, a qualidade da síntese final permanece dependente da capacidade generativa do modelo escolhido.

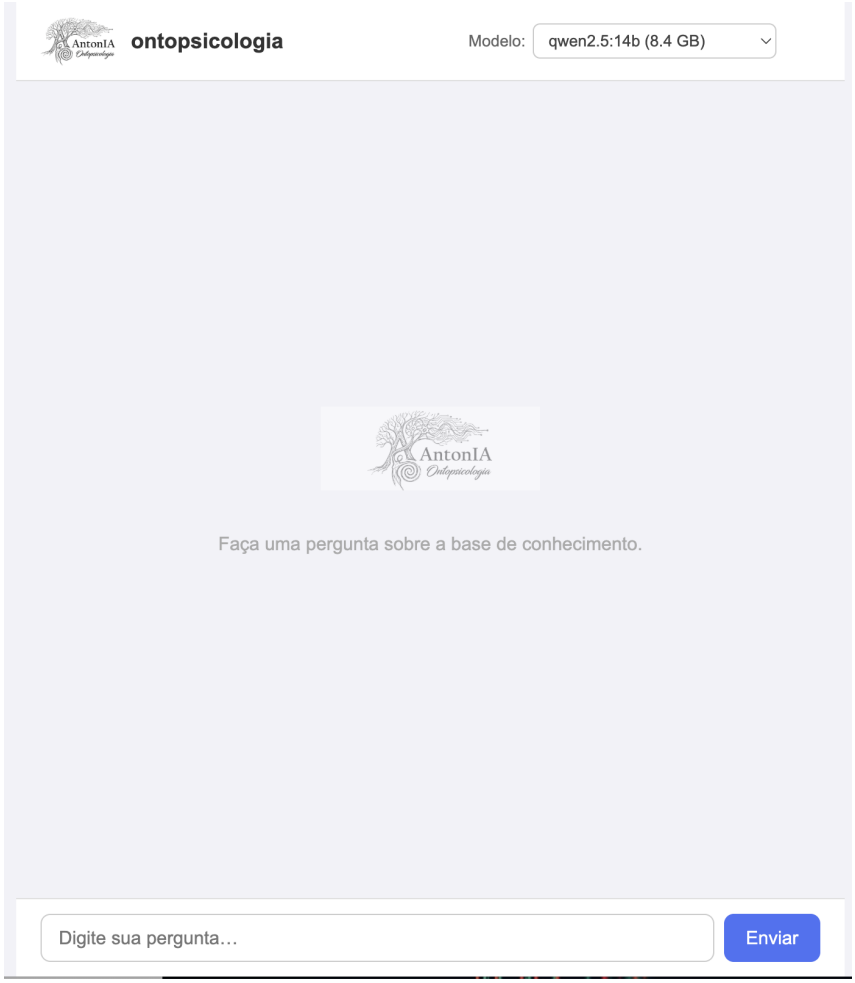


Figura 11: Tela de Consulta
Fonte: elaborado pelo autor.

O modelo Mistral-Nemo:12b apresentou um comportamento distinto dos demais: em vez de sintetizar os fragmentos recuperados da base de conhecimento em linguagem própria, reproduziu literalmente um trecho do documento recuperado, identificando-o explicitamente como "Trecho 2". Esse comportamento, embora garanta fidelidade ao conteúdo original, compromete

a utilidade pedagógica do sistema, pois um assistente conversacional deve ser capaz de integrar e reformular as informações de forma fluida e contextualizada. A resposta é parcialmente adequada do ponto de vista da precisão, mas inadequada como saída de um assistente generativo.a

Para validar a precisão da resposta gerada pelo modelo Qwen2.5:7b-instruct,foi feita mais uma pergunta a esse modelo: como posso utilizar a Ontopsicologia para ter uma vida melhor? A resposta gerada por esse modelo é a seguinte: *"Para utilizar a ontopsicologia na busca por uma vida melhor, você pode:"*

1. *Desenvolver o "Eu lógico-histórico"com capacidades e condutas vencedoras, educando-se como um líder no mundo";*

2. *"Ab-reação à memética societária, aprendendo a reagir de forma crítica às influências sociais";*

3. *"Identificar e evoluir o Em Si ôntico, promovendo uma autopercepção mais profunda e autêntica";*

4. *"Estabelecer uma correlação entre as crenças sociais (doxa societária) e os critérios de natureza, desenvolvendo uma ética mais alinhada com princípios pessoais e universais".*

"Estas ações podem contribuir para a maturidade individual e profissional, auxiliando na formação de competências como estilo de vida, maneira de pensar (lógica), conhecimento de si mesmo, postura de seriedade, ambição, estudo e ontopsicologia".

Analisando a resposta gerada por esse modelo e comparando as referências bibliográficas da ciencia ontopsicológica, é possível identificar que a resposta é aderente aos fundamentos da área, sem alucinações ou geração de memes. A técnica RAG aplicada a uma base de conhecimento específica e verificada, combinada com um modelo pré-treinado de código aberto, demonstrou ser capaz de gerar respostas precisas, relevantes e contextualizadas, mesmo em um ambiente local com recursos computacionais limitados. Esses resultados validam a escolha arquitetural do sistema AntonIA e indicam que é possível construir assistentes conversacionais especializados sem depender de modelos proprietários ou infraestrutura em nuvem, garantindo segurança, soberania informacional e viabilidade econômica sustentável.

5 Considerações Finais

Este trabalho construiu uma pesquisa que tesse uma interdisciplinaridade entre a Ciência Ontopsicológica, a Inteligência Artificial Generativa, sendo a aplicação de Tecnologia à Ontopsicologia, por isso os resultados dessa pesquisa são relevantes e contribuem cientificamente para essas duas áreas do conhecimento. O desenvolvimento do sistema AntonIA, um protótipo de Inteligência Artificial generativa baseado na técnica RAG, projetado para responder a consultas sobre os fundamentos da Ciência Ontopsicológica com fidelidade semântica à base de conhecimento de referência. O sistema foi desenvolvido segundo os princípios da *Design Science*

Research Methodology, o que garantiu que o processo de criação e validação do artefato fosse conduzido com rigor científico e relevância prática Peffers *et al.* (2007) e Hevner *et al.* (2004).

Os resultados dos testes realizados com quatro modelos de linguagem de código aberto demonstraram que o pipeline RAG, ancorado em uma base de conhecimento validada e verificada, é capaz de gerar respostas precisas e contextualmente coerentes com o domínio ontopsicológico. O modelo Qwen2.5:7b apresentou a melhor performance entre os modelos avaliados, sendo o único capaz de enumerar os sete pilares estruturais da Ontopsicologia com terminologia fiel ao Manual de Ontopsicologia (Meneghetti, 2022).

A execução integralmente local do sistema AntonIA, sem chamadas a APIs externas e sem envio de dados a provedores em nuvem, sem dependência de serviços de terceiros, foi uma decisão arquitetural que atende simultaneamente a três objetivos importantes: segurança das informações, controle do conhecimento e viabilidade econômica sustentável.

Quando um sistema de IA generativa opera via APIs de provedores externos, como OpenAI, Anthropic, Google Gemini ou similares, cada consulta enviada a esses serviços constitui, em termos práticos, uma transmissão de dados para infraestrutura fora do controle institucional. Embora esses provedores publiquem políticas de privacidade, os termos de uso variam, e a possibilidade de que dados enviados sejam utilizados para retreinamento de modelos futuros ou retidos em logs de servidores externos é uma preocupação real e documentada Bommasani *et al.* (2022). A arquitetura local do sistema AntonIA elimina esse risco de forma estrutural: nenhum dado trafega para fora da rede institucional, garantindo que os documentos originais da Ciência Ontopsicológica não sejam disponibilizados na rede nem compartilhados com diferentes fornecedores de modelos LLMs ou ferramentas de IA generativa.

A aplicação da técnica RAG sobre uma base de conhecimento específica e validada garante que os LLMs pré-treinados consultem informações exclusivamente com base nos conteúdos cadastrados na aplicação. Dessa forma, mesmo sendo pré-treinados com dados genéricos, esses modelos são impedidos de consultar outras bases de informação que possam gerar alucinações ou respostas incompatíveis com a área de conhecimento específica do domínio configurado. Ao restringir as respostas a uma base de conhecimento validada e verificada, remove-se a possibilidade de criação de memes, no sentido definido por Meneghetti (2024): imitações de uma realidade, como se fossem conhecimentos legítimos. Uma IA que responde sobre Ontopsicologia com base exclusivamente em fontes validadas tem suas respostas ligadas a um fragmento verificável e rastreável do pensamento original.

Como limitações do trabalho, destaca-se que os testes foram realizados em um único ambiente computacional e com um conjunto restrito de consultas. Adicionalmente, a ausência de GPU dedicada no ambiente de testes impôs restrições de latência que poderiam ser superadas em infraestruturas com maior capacidade de processamento.

Como trabalhos futuros, sugere-se a expansão da base de conhecimento com novas fontes verificadas, a implementação de mecanismos de re-ranking para aprimorar a precisão da recuperação em consultas multi-conceito, a avaliação do sistema com modelos de maior escala e

a aplicação do protocolo formal de avaliação por especialistas. Considera-se também relevante investigar a integração de fontes multimodais, como imagens e apresentações.

O sistema AntonIA demonstra que é possível construir ferramenta de Inteligência Artificial generativa especializada, segura e economicamente viável, capaz de preservar a integridade conceitual de um domínio de conhecimento sem depender de modelos proprietários ou de infraestrutura externa, garantindo que cada resposta gerada seja fundamentada em fontes verificadas e livre da distorção que caracteriza a propagação de memes conceituais. O sistema AntonIA pode ser uma ferramenta de auxílio ao aluno de Ontopsicologia, permitindo que ele consulte conceitos e fundamentos da área com precisão, relevância e rastreabilidade sendo assim um facilitador no processo de aprendizagem da Ciência Ontopsicológica.

Referências

BOMMASANI, Rishi *et al. On the Opportunities and Risks of Foundation Models*. [*S. l.: s. n.*], 2022. arXiv:2108.07258. arXiv: 2108.07258 [cs.LG]. Disponível em: https://arxiv.org/abs/2108.07258.

BROWN, Tom B. *et al. Language Models are Few-Shot Learners*. [*S. l.: s. n.*], 2020. arXiv:2005.14165. arXiv: 2005.14165 [cs.CL]. Disponível em: https://arxiv.org/abs/2005.14165.

CHASE, Harrison. *LangChain*. [*S. l.: s. n.*], out. 2022. GitHub. Disponível em: https://github.com/langchain-ai/langchain.

CORMACK, Gordon V.; CLARKE, Charles L. A.; BUETTCHER, Stefan. Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods. *In:* PROCEEDINGS of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2009). Boston, Massachusetts, USA: ACM, jul. 2009. p. 758–759. DOI: 10.1145/1571941.1572114. Disponível em: https://dl.acm.org/doi/10.1145/1571941.1572114.

FOSTER, David. *Generative Deep Learning: Teaching Machines to Paint, Write, Compose, and Play*. 2. ed. Sebastopol: O'Reilly Media, 2022. ISBN 9781098134181.

GALILEO AI. *RAG Architecture: From Naive Pipelines to Agentic*. [*S. l.: s. n.*], 2026. Galileo AI Blog. Disponível em: https://galileo.ai/blog/rag-architecture. Acesso em: 24 maio 2026.

GAO, Yunfan *et al. Retrieval-Augmented Generation for Large Language Models: A Survey*. [*S. l.: s. n.*], 2023. arXiv:2312.10997. arXiv: 2312.10997 [cs.CL]. Disponível em: https://arxiv.org/abs/2312.10997.

GOODFELLOW, Ian *et al.* Generative Adversarial Nets. *In:* ADVANCES in Neural Information Processing Systems. [*S. l.*]: Curran Associates, Inc., 2014. v. 27. Disponível em: https://https://papers.nips.cc/paper_files/paper/2014/hash/f033ed80deb0234979a61f95710dbe25-Abstract.html.

HEVNER, Alan R. *et al.* Design Science in Information Systems Research. *MIS Quarterly*, v. 28, n. 1, p. 75–105, 2004. DOI: 10.2307/25148625.

HOLOVATY, Adrian; KAPLAN-MOSS, Jacob. *The Definitive Guide to Django: Web Development Done Right*. 2. ed. Berkeley: Apress, 2009. ISBN 9781430219361.

JI, Ziwei *et al.* Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, v. 55, n. 12, p. 1–38, 2023. DOI: 10.1145/3571730. Disponível em: https://doi.org/10.1145/3571730.

KARPUKHIN, Vladimir *et al.* Dense Passage Retrieval for Open-Domain Question Answering. *In:* PROCEEDINGS of the 2020 Conference on Empirical Methods in Natural Language

Processing (EMNLP). [*S. l.: s. n.*], 2020. p. 6769–6781. DOI: 10.18653/v1/2020.emnlp-main.550.

KINGMA, Diederik P.; WELLING, Max. *Auto-Encoding Variational Bayes*. [*S. l.: s. n.*], 2013. arXiv:1312.6114. arXiv: 1312.6114 [cs.LG]. Disponível em: https://arxiv.org/abs/1312.6114.

LECUN, Yann. *A Path Towards Autonomous Machine Intelligence*. [*S. l.: s. n.*], 2022. OpenReview. Disponível em: https://openreview.net/pdf?id=BZ5a1r-kVsf.

LECUN, Yann. *Open Source AI Models and the Future of AI Autonomy*. [*S. l.: s. n.*], 2023. Meta AI Research Reports. Disponível em: https://ai.meta.com/research/.

LECUN, Yann; BENGIO, Yoshua; HINTON, Geoffrey. Deep Learning. *Nature*, v. 521, n. 7553, p. 436–444, 2015. DOI: 10.1038/nature14539.

LEWIS, Patrick *et al.* Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *In:* ADVANCES in Neural Information Processing Systems (NeurIPS). [*S. l.: s. n.*], 2020. v. 33, p. 9459–9474. arXiv:2005.11401. Disponível em: https://arxiv.org/abs/2005.11401.

MENEGHETTI, Antonio. *Manual de Ontopsicologia*. 4. ed. Recanto Maestro: Ontopsicológica Editora Universitária, 2022.

MENEGHETTI, Antonio. *Ontopsicologia e Memética*. 2. ed. Recanto Maestro: Ontopsicológica Editora Universitária, 2024.

OLLAMA. *Ollama: Get Up and Running with Large Language Models Locally*. [*S. l.: s. n.*], 2023. GitHub. Disponível em: https://github.com/ollama/ollama.

PEFFERS, Ken *et al.* A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, v. 24, n. 3, p. 45–77, 2007. DOI: 10.2753/MIS0742-1222240302.

REIMERS, Nils; GUREVYCH, Iryna. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *In:* PROCEEDINGS of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP). [*S. l.*]: Association for Computational Linguistics, nov. 2019. Disponível em: https://arxiv.org/abs/1908.10084.

ROMBACH, Robin *et al.* High-Resolution Image Synthesis with Latent Diffusion Models. *In:* PROCEEDINGS of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). [*S. l.: s. n.*], 2022. p. 10684–10695. DOI: 10.1109/CVPR52688.2022.01042.

RUSSELL, Stuart J.; NORVIG, Peter. *Artificial Intelligence: A Modern Approach*. 4. ed. Hoboken: Pearson, 2021. ISBN 9780134610993.

SIMON, Herbert A. *The Sciences of the Artificial*. 3. ed. Cambridge: MIT Press, 1996. ISBN 9780262691918.

SYSTEMATIC LITERATURE REVIEW. *A Systematic Literature Review of Retrieval-Augmented Generation*. [*S. l.: s. n.*], 2025. arXiv:2508.06401. arXiv: 2508.06401 [cs.IR]. Disponível em: https://arxiv.org/abs/2508.06401.

TOUVRON, Hugo *et al. LLaMA: Open and Efficient Foundation Language Models*. [*S. l.: s. n.*], 2023. arXiv:2302.13971. arXiv: 2302.13971 [cs.CL]. Disponível em: https://arxiv.org/abs/2302.13971.

VASWANI, Ashish *et al.* Attention Is All You Need. *In:* ADVANCES in Neural Information Processing Systems. [*S. l.*]: Curran Associates, Inc., 2017. v. 30. Disponível em: https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.