

NUTRIMIND: ARQUITETURA DISTRIBUÍDA MULTIAGENTE PARA PLANEJAMENTO ADAPTATIVO BASEADO EM IA GENERATIVA MULTIMODAL

Ana Luisa Bernardo de Andrade ¹; Dianne Faria de Brito ²; Maria Luiza Guedes da Silva

³; Pedro Henrique Ribeiro ⁴; Sofia Matos Lessa ⁵;

José Walmir Gonçalves Duque ⁶; Walter Teixeira Lima Junior ⁷

¹ Fatec São José dos Campos. (louie.bernardes@gmail.com)

² Fatec São José dos Campos. (diannefaria09@gmail.com)

³ Fatec São José dos Campos. (mluizaguedessilva@gmail.com)

⁴ Fatec São José dos Campos. (pedrohenribeiro04@gmail.com)

⁵ Fatec São José dos Campos. (matossofia937@gmail.com)

⁶ Fatec São José dos Campos. (walmirduque@gmail.com)

⁷ UNIFESP – Universidade Federal de São Paulo. (walter.lima@unifesp.br)

A incorporação de LLMs (Large Language Models, em português, Grandes Modelos de Linguagem) multimodais em sistemas digitais, aliada a uma arquitetura multiagente, permite desacoplamento de módulos, controle de estado conversacional e integração entre diferentes fontes de dados. É comum que ferramentas de orientação nutricional apresentem limitações em sua estrutura: modelos padronizados baseados em recomendações genéricas, baixa personalização e ausência de suporte multimodal (texto, imagem e áudio), muitas vezes devido à indisponibilidade ou limitações na coleta desses dados. Nesse cenário, a NutriMind é proposta como plataforma para geração de orientação nutricional adaptada às demandas de cada paciente, tendo como base uma arquitetura modular com divisão clara de responsabilidades. A solução integra o modelo Gemini-2.0-flash por meio do framework Google ADK, compondo um pipeline capaz de processar entradas de texto, imagem e áudio de forma unificada. Sua arquitetura foi estruturada em três camadas: (a) interface cliente em Vue.js, responsável pela interação conversacional e gerenciamento de sessões; (b) backend em Java com Spring Boot, encarregado da autenticação, persistência relacional e versionamento de dados; (c) backend cognitivo em Python com Flask, onde concentram-se o núcleo multiagente e mecanismo de RAG (Retrieval Augmented Generation, em português, Geração Aumentada por Recuperação). O núcleo cognitivo é estruturado em torno de um agente orquestrador (Root Agent), responsável pela classificação de intenções por meio de abordagem zero-shot baseada em LLM (Gemini), orientada por prompt engineering. Esse agente também preserva contexto conversacional e realiza roteamento para agentes especialistas conforme a intenção identificada. A multimodalidade é tratada por um processo de normalização que converte diferentes tipos de entrada em um formato comum, permitindo processamento unificado pelo sistema. Entre os agentes especialistas estão: (a) MealPlanAgent, responsável pela modelagem calórica com redistribuição automática; (b) NutriAgent, que consolida fundamentação técnico-científica; (c) RecipeAgent, que sintetiza preparações alinhadas a metas metabólicas; (d) ImageAnalysisAgent, que extrai atributos alimentares a partir de imagens e estima valores nutricionais; (e) EditAgent, encarregado da atualização incremental do histórico nutricional. O banco de dados PostgreSQL é responsável pela persistência do estado conversacional,

permitindo decisões ao longo das sessões e continuidade das interações. Essa arquitetura permite personalização contínua, ajustando o plano alimentar com base no histórico e em alterações no consumo ou perfil metabólico do usuário. Para armazenamento de embeddings, é utilizado o ChromaDB como banco vetorial, permitindo a recuperação de informações relevantes. Esse processo viabiliza grounding a partir de bases especializadas em nutrição. Suplementarmente, o sistema é capaz de incorporar personas nos agentes especialistas, permitindo controlar estilo comunicacional e nível de especialização ao longo do processo de inferência, de acordo com o perfil do usuário. Foram realizados testes funcionais para validação dos fluxos entre agentes e da integração dos componentes, bem como testes qualitativos para avaliação da coerência, relevância e personalização das respostas. Observou-se que a combinação de entradas multimodais, arquitetura multiagente e persistência do estado conversacional melhora a continuidade das interações, reduz inconsistências e aumenta adequação das respostas ao perfil do usuário. Portanto, a NutriMind oferece suporte nutricional personalizado, consistente e inteligente, mesmo na ausência de um profissional de nutrição permanente.

Palavras-chave: Inteligência Artificial Generativa (Gen AI); Arquitetura Multiagentes; Geração Aumentada por Recuperação (RAG); Persistência Conversacional; Nutrição Personalizada;

NUTRIMIND: DISTRIBUTED MULTI-AGENT ARCHITECTURE FOR ADAPTIVE PLANNING BASED ON MULTIMODAL GENERATIVE AI

Ana Luisa Bernardo de Andrade¹; Dianne Faria de Brito²; Maria Luiza Guedes da Silva³; Pedro Henrique Ribeiro⁴; Sofia Matos Lessa⁵;
José Walmir Gonçalves Duque⁶; Walter Teixeira Lima Junior⁷

¹ Fatec São José dos Campos. (louie.bernardes@gmail.com)

² Fatec São José dos Campos. (diannefaria09@gmail.com)

³ Fatec São José dos Campos. (mluizaguedessilva@gmail.com)

⁴ Fatec São José dos Campos. (pedrohenribeiro04@gmail.com)

⁵ Fatec São José dos Campos. (matossofia937@gmail.com)

⁶ Fatec São José dos Campos. (walmirduque@gmail.com)

⁷ UNIFESP – Universidade Federal de São Paulo. (walter.lima@unifesp.br)

The incorporation of multimodal LLMs (Large Language Models) into digital systems, in conjunction with a multi-agent architecture, enables module decoupling, conversational state management, and integration across heterogeneous data sources. Current nutritional guidance tools often present structural limitations, such as reliance on generic recommendations, limited support for continuous personalization, and lack of multimodal capabilities (text, image, and audio), often due to constraints in data acquisition. In this context, NutriMind is proposed as a platform for generating personalized nutritional guidance, based on a modular architecture with clearly defined responsibilities. The solution integrates the Gemini-2.0-flash model through the Google ADK framework, forming a pipeline capable of processing text, image, and audio inputs in a unified manner. The system architecture is composed of three layers: (a) client interface in Vue.js, responsible for conversational interaction and session management; (b) backend in Java with Spring Boot, in charge of authentication, relational persistence, and meal plan versioning; (c) cognitive backend in Python with Flask, where the multi-agent core and the RAG (Retrieval-Augmented Generation) mechanism are implemented. The cognitive core is centered around an orchestrator agent (Root Agent), responsible for intent classification using a zero-shot approach based on LLM (Gemini), guided by prompt engineering. This agent also maintains conversational context and routes requests to specialized agents based on the identified intent. Multimodal inputs are processed through a normalization process that converts the varied input types into a common format, enabling unified processing by the system. The specialized agents include: (a) MealPlanAgent, responsible for caloric modeling with automatic redistribution; (b) NutriAgent, which consolidates the technical-scientific foundation; (c) RecipeAgent, which synthesizes preparations aligned with metabolic goals; (d) ImageAnalysisAgent, which extracts food attributes from images and estimates nutritional values; (e) EditAgent, in charge of the incremental updating of the nutritional history. The

PostgreSQL database is responsible for the persistence of the conversational state, enabling continuity of interactions and dynamic adjustment of responses over time. This architecture enables continuous personalization by adapting meal plans based on user history and changes in consumption or metabolic profile. ChromaDB is utilized as a vector database for the storage of embeddings and the retrieval of pertinent information, particularly in the context of nutrition, based on literature and specialized nutrition databases. This process facilitates grounding, defined as anchoring responses in reliable external information, through specialized nutrition sources. Additionally, the system integrates personas into specialized agents, enabling control over communication style and level of expertise according to the user profile. To ensure the efficacy of the agent workflows and system integration, a series of functional tests were conducted. These tests were complemented by qualitative evaluations that assessed factors such as response coherence, relevance, and personalization. The findings suggest that integrating multimodal inputs, multi-agent architecture, and conversational state persistence enhances interaction continuity, reduces inconsistencies, and improves response alignment with user context. Therefore, NutriMind provides personalized, reliable, and intelligent nutritional assistance, even in the absence of a permanent nutrition professional.

Keywords: Generative Artificial Intelligence (Gen AI); Multi-agent Architecture; Retrieval-Augmented Generation (RAG); Conversational Persistence; Personalized Nutrition.