MINISTERUL EDUCAŢIEI ŞI CERCETĂRII AL REPUBLICII MOLDOVA
Universitatea Tehnică a Moldovei
Facultatea Calculatoare Informatică şi Microelectronică
Departamentul Ingineria Software și Automatică

Admis la susținere
Șef departament:
FIODOROV Ion dr., conf.univ.

„___” _____ 2025

CERCETAREA ȘI DEZVOLTAREA SOLUȚIEI PENTRU CREAREA UNUI AGENT BAZAT PE INTELIGENȚA ARTIFICIALĂ

Proiect de master

Student: _____ **Popa Valeriu, TI-231M**

Coordonator: _____ **Cojocaru Svetlana, asis. univ.**

Consultant: _____ **Cojocaru Svetlana, asis. univ.**

Chişinău, 2025

REZUMAT

Lucrarea explorează dezvoltarea unui agent inteligent bazat pe inteligență artificială (IA), concentrându-se pe integrarea memoriei persistente și a mecanismelor de recuperare prin aducere (Retrieval-Augmented Generation - RAG). Structura lucrării este împărțită în trei capitole principale, fiecare contribuind la fundamentarea și realizarea obiectivului propus.

Capitolul 1 oferă o introducere în contextul teoretic și tehnologic al agenților IA, evidențiind relevanța lor în diverse domenii. Se discută despre evoluția inteligenței artificiale și integrarea acesteia în aplicații software, precum și despre rolul memoriei persistente și al mecanismelor RAG în îmbunătățirea performanței agenților. De asemenea, sunt prezentate exemple de soluții existente și beneficiile utilizării framework-urilor moderne, precum LangChain și LangGraph.

Capitolul 2 Dezvoltarea serviciului detaliază etapele de proiectare și implementare a prototipului agentului IA. Se descriu procesele de selecție a tehnologiilor utilizate, integrarea memoriei persistente pentru stocarea și gestionarea informațiilor pe termen lung și implementarea mecanismelor RAG pentru generarea răspunsurilor relevante în timp real. Sunt discutate provocările tehnice întâmpinate, soluțiile propuse și rezultatele testării prototipului în scenarii reale.

Capitolul 3 abordează implementarea soluției ca Serviciu oferit prin modelul Software ca Serviciu (SaaS). Se analizează avantajele acestui model, precum accesibilitatea, scalabilitatea și eficiența costurilor, și se evidențiază aplicabilitatea soluției în diverse domenii. De asemenea, sunt discutate aspectele legate de securitate, confidențialitate și compatibilitate cu infrastructuri IT diverse. Capitolul se încheie cu prezentarea direcțiilor viitoare de cercetare, inclusiv extinderea funcționalităților agentului și integrarea cu noi tehnologii.

Prin structurarea clară a celor trei capitole, lucrarea oferă o perspectivă teoretică și practică asupra dezvoltării agenților IA avansați, demonstrând potențialul acestora de a transforma interacțiunile om-tehnologie și de a adresa provocările din diverse industrii.

ABSTRACT

This thesis explores the development of an intelligent agent based on artificial intelligence (AI), focusing on the integration of persistent memory and retrieval-augmented generation (RAG) mechanisms. The structure of the paper is divided into three main chapters, each contributing to the substantiation and achievement of the proposed objective.

Chapter 1 provides an introduction to the theoretical and technological context of AI agents, highlighting their relevance in various fields. The evolution of artificial intelligence and its integration into software applications is discussed, as well as the role of persistent memory and RAG mechanisms in improving agent performance. Examples of existing solutions and the benefits of using modern frameworks, such as LangChain and LangGraph, are also presented.

Chapter 2 Service Development details the design and implementation stages of the AI agent prototype. The processes of selecting the technologies used, integrating persistent memory for long-term information storage and management, and implementing RAG mechanisms for generating relevant responses in real time are described. The technical challenges encountered, the proposed solutions and the results of testing the prototype in real scenarios are discussed.

Chapter 3 addresses the implementation of the solution as a Service offered through the Software as a Service (SaaS) model. The advantages of this model, such as accessibility, scalability and cost efficiency, are analyzed, and the applicability of the solution in various domains is highlighted. Issues related to security, confidentiality and compatibility with various IT infrastructures are also discussed. The chapter concludes with the presentation of future research directions, including the extension of the agent's functionalities and integration with new technologies.

By clearly structuring the three chapters, the paper provides a theoretical and practical perspective on the development of advanced AI agents, demonstrating their potential to transform human-technology interactions and address challenges in various industries.

CUPRINS

ABREVIERI

BD - Baza de date.

IA - Inteligență Artificială.

LLM - Large Language Models.

NLP - Procesarea Limbajului Natural (eng. Natural Language Processing).

SaaS - Software la cerere (eng. Software as a service).

NLP - Procesarea Limbajului Natural (eng. Natural Language Processing).

RAG - Recuperare Prin Aducere (eng. Retrieval-Augmented Generation).

INTRODUCERE

În contextul transformării digitale și al progreselor rapide în domeniul inteligenței artificiale (IA), dezvoltarea agenților software inteligenți reprezintă un domeniu de interes major pentru cercetători și profesioniștii din domeniu. Acești agenți, sunt capabili să îmbunătățească interacțiunea dintre utilizatori și tehnologie, devin tot mai relevanți în soluționarea unor probleme complexe din diverse domenii precum educația, sănătatea, industria financiară sau logistică. Evoluția capacităților de procesare, expansiunea volumelor de date și progresele în algoritmii de învățare automată au creat premisele pentru integrarea agenților inteligenți în numeroase aplicații, oferind o valoare adăugată semnificativă prin personalizarea și automatizarea proceselor.

Această lucrare se concentrează pe cercetarea și dezvoltarea unui agent inteligent bazat pe inteligență artificială, cu integrarea conceptelor de memorie persistentă și recuperare prin aducere (Retrieval-Augmented Generation - RAG) [1]. Scopul principal este crearea unui prototip care să combine aceste tehnologii pentru a furniza răspunsuri adaptative, contextuale și personalizate, oferind în același timp o experiență optimizată utilizatorului. Agentul propus urmărește să automatizeze interogările în timp real și să permită scalabilitatea și integrarea într-o gamă variată de infrastructuri IT, cu accent pe conformitatea cu cerințele actuale de securitate și performanță.

Un aspect important al acestei cercetări îl reprezintă implementarea soluției ca Serviciu oferit prin Software ca Serviciu (SaaS), atât în medii cloud, cât și On-Premises. Modelul SaaS asigură flexibilitate, accesibilitate și eficiență operațională, facilitând adoptarea de către organizații diverse, indiferent de dimensiune sau infrastructură. Soluția dezvoltată permite utilizatorilor să acceseze funcționalitățile agentului IA fără necesitatea de a investi în infrastructură hardware costisitoare sau în echipe dedicate de administrare. Prin implementarea unui model de livrare SaaS, agentul poate fi scalat rapid, menținând totodată un nivel ridicat de securitate și conformitate cu reglementările privind protecția datelor.

Lucrarea este fundamentată pe o analiză detaliată a soluțiilor existente, identificarea limitărilor acestora și explorarea unor tehnologii avansate precum framework-urile LangChain și LangGraph. Prin testarea unui prototip în scenarii reale, se urmărește evaluarea eficienței agentului propus și determinarea impactului său asupra îmbunătățirii interacțiunii om-mașină. De asemenea, cercetarea explorează metodele de integrare a memoriei persistente pentru a sprijini procesele de învățare pe termen lung și adaptabilitatea la diverse contexte operaționale.

Studiul deschide noi perspective pentru dezvoltarea unor agenți inteligenți mai performanți, cu impact direct asupra automatizării și eficientizării proceselor, contribuind la progresul tehnologic și la crearea unor ecosisteme software mai robuste și mai adaptabile.

BIBILIOGRAFIE

[1] Basappa, Prashanth. "Building Your Retrieval-Augmented Generation (RAG) for Custom LLMs." *Medium*, [Data accesării: 6 noiembrie 2024]. Disponibil la: https://prashanth08.medium.com/building-your-retrieval-augmented-generation-rag-for-custom-llms-d5f95ed5ed7a.

[2] Belagatti, Pavan. "Learn How to Build AI Agents & Chatbots with LangGraph!" *DEV Community*, [Data accesării: 6 noiembrie 2024]. Disponibil la: https://dev.to/pavanbelagatti/learn-how-to-build-ai-agents-chatbots-with-langgraph-20o6.

[3] *Building RAG from Scratch (Lower-Level) - LlamaIndex - Official Documentation*. https://docs.llamaindex.ai/en/stable/optimizing/building_rag_from_scratch/. Accessed 14 Dec. 2024.

[4] "Get Started with Neo4j - Getting Started - Official Documentation." *Neo4j Graph Data Platform*, [Data accesării: 6 noiembrie 2024]. Disponibil la: https://neo4j.com/docs/getting-started/.

[5] *LangChain Python API Reference — □□ LangChain Documentation*. [Data accesării: 14 noiembrie 2024]. Disponibil la: https://python.langchain.com/api_reference/.

[6] *Langchain-Ai/Langgraph*. 2023. [Data accesării: 26 noiembrie 2024]. Disponibil la: https://github.com/langchain-ai/langgraph.

[7] Marco, Eden. *LangChain- Develop LLM Powered Applications with LangChain - Udemy Course*. [Data accesării: 2 decembrie 2024]. Disponibil la: https://www.udemy.com/course/langgraph.

[8] *LangGraph- Develop LLM Powered AI Agents with LangGraph - Udemy Couse*. [Data accesării: 4 decembrie 2024]. Disponibil la: https://www.udemy.com/course/langgraph/.

[9] Oudenhove, Lore Van. "How to Build AI Agents with LangGraph: A Step-by-Step Guide." *Medium*, 9 Oct. 2024, https://medium.com/@lorevanoudenhove/how-to-build-ai-agents-with-langgraph-a-step-by-step-guide-5d84d9c7e832.

[10] Sivan, Vishnu. "Building AI Agent Systems with LangGraph." *The Pythoneers*, 26 Oct. 2024, https://medium.com/pythoneers/building-ai-agent-systems-with-langgraph-9d85537a6326.

[11] Thaker, Madhav. "(Part 1) Build Your Own RAG with Mistral-7B and LangChain." *Medium*. [Data accesării: 16 decembrie 2024]. Disponibil la: https://medium.com/@thakermadhav/build-your-own-rag-with-mistral-7b-and-langchain-97d0c92fa146.

[12] "(Part 2) Build a Conversational RAG with LangChain and Mistral-7B." *Medium*, [Data accesării: 16 decembrie 2024]. Disponibil la: https://medium.com/@thakermadhav/part-2-build-a-conversational-rag-with-langchain-and-mistral-7b-6a4ebe497185.

[13] Witteveen, Sam. *Building a LangGraph ReAct Mini Agent - YouTube*. [Data accesării: 16 noiembrie 2024]. Disponibil la: https://www.youtube.com/watch?v=pEMhPBQMNjg.

[14] S. Cojocaru și L. Peca, „Challenges and solutions on the use of Artificial Intelligence in Internet of Things network security”, 2024, Disponibil la: http://repository.utm.md/handle/5014/28775. [Data accesării: 16 decembrie 2024]

[15] I. Bolun și S. Cojocaru, „A Differentiated Beneficiary Cybersecurity Approach”, în Proceedings of the 12th International Conference on “Electronics, Communications and Computing", Technical University of Moldova, 2022, pp. 115–118. doi: 10.52326/ic-ecco.2022/SEC.01. Disponibil la: http://repository.utm.md/handle/5014/21840. [Data accesării: 16 decembrie 2024]