HOLISUN

Contract nr.11 din 01/07/2025, etapa 1 - LLM-GenFramework

RAPORT ȘTIINȚIFIC ȘI TEHNIC 2025

Referință 7770/ 24.11.2025

Manager de proiect: ***Rudolf-Eduard Erdei***

Istoricul versiunilor

Versiune	Autor	Modificări
0.1	Rudolf-Eduard Erdei	Versiunea inițială
0.5	Rudolf-Eduard Erdei	Versiunea intermediară
1.0	Rudolf-Eduard Erdei	Versiunea finală

Cuprins

1 Introducere **4**

2 Despre proiectul LLM-GenFramework **4**

3 Activități planificate **6**

4 Activități efectuate **6**

4.1 Devieri de la planificare 6

5 Starea curentă a domeniului **6**

6 Metodologie pentru crearea seturilor de training, validare și test **7**

6.1 Analiza atributelor datasetului 7

6.2 Selecția subseturilor de training, validare și test 7

6.3 Metodologia de ponderare a datelor 7

6.4 Experimentare și iterare 8

6.5 Rolul metodologiei în arhitectura generală 8

7 Rolul HOLISUN în cadrul proiectului **8**

8 Extras din planul de riscuri **9**

9 Rezultatele proiectului **10**

9.1 Livrabile 10

10 Diseminare și exploatare **10**

10.1 Activități de diseminare 10

11 Concluzii **11**

Parteneri

(a) AI Investments (Polonia) Coordonator

(b) InbestMe (Spania)

(c) Holisun SRL (Romania)

Figura 1: Partenerii proiectului *LLM-GenFramework*

1 Introducere

Proiectul **LLM-GenFramework** abordează una dintre direcțiile emergente ale inteligenței artificiale moderne: utilizarea modelelor de tip *Large Language Models* (LLM) și a tehnicilor de *Retrieval Augmented Generation* (RAG) pentru automatizarea, accelerarea și optimizarea proceselor de cercetare, dezvoltare și analiză în domeniul strategiilor investiționale cantitative. Într-un context în care complexitatea datelor financiare, diversitatea instrumentelor de analiză și ritmul accelerat al evoluțiilor pieților impun noi standarde de eficiență, proiectul urmărește crearea unei soluții inteligente capabile să asiste experții în întregul flux metodologic, de la conceptualizarea strategiilor până la evaluarea și documentarea acestora.

Prin integrarea celor mai recente progrese în LLM-uri, tehnici RAG și metode moderne de analiză algoritmică, proiectul va dezvolta un cadru generalizat (framework) ce poate fi aplicat nu doar în investiții, ci și în alte sectoare economice unde procesele analitice, deciziile bazate pe date și generarea de documentație sunt esențiale. În acest sens, soluția propusă va crește semnificativ productivitatea în realizarea strategiilor cantitative, reducând timpul necesar pentru cercetare, accelerând experimentarea și facilitând integrarea cunoștințelor din surse vaste și eterogene.

Prezentul raport oferă o prezentare de ansamblu a cadrului metodologic, a principiilor generale care stau la baza framework-ului și a structurii proiectului, evidențiind contribuția acestuia la modernizarea proceselor de cercetare și dezvoltare în domeniul investițiilor și la facilitarea aplicării tehnicilor avansate AI în diverse arii de business.

2 Despre proiectul LLM-GenFramework

Proiectul **LLM-GenFramework – AI-based intelligent assistant for developing and researching quantitative investment strategies with applicability to diverse business areas** are ca obiectiv principal crearea unei soluții unice, bazate pe modele de tip *Large Language Models* și tehnici de *Retrieval Augmented Generation*, care să sprijine procesele de creare, cercetare și analiză ale strategiilor investiționale cantitative. Prin extinderea capabilităților LLM-urilor cu mecanisme RAG, soluția va integra informații din surse multiple, permitând generarea de analize complexe, a documentației aferente și a unor propuneri fundamentate pentru îmbunătățirea strategiilor existente sau dezvoltarea unora noi.

În paralel, proiectul urmărește crearea unui **modul generic LLM-GenFramework**, care să permită aplicarea metodelor dezvoltate în diverse domenii economice – inclusiv analiză de date, raportare tehnică, cercetare științifică asistată și dezvoltare de soluții IT. Acest modul va fi proiectat pentru reutilizare, extensibilitate și adaptabilitate la diferite fluxuri de lucru specifice industriei.

Pe lângă componenta generală, proiectul va dezvolta și un **produs specializat LLM-Invest**, destinat în mod exclusiv pieței financiare. LLM-Invest va integra funcționalități care asistă investitorii, analiștii și dezvoltatorii de strategii cantitative în:

- generarea și documentarea strategiilor de investiții;
- selecția și compararea diferitelor abordări cantitative;
- analizarea rapidă a contextelor macro și microeconomice folosind capacități avansate LLM;
- integrarea analizei calitative și cantitative într-un cadru unificat.

Figura 2 prezintă arhitectura logică de înalt nivel a proiectului, ilustrând modul în care componentele LLM, RAG și modulele specializate interacționează pentru a susține fluxurile de creare și analiză a strategiilor.

Soluția se bazează pe date financiare istorice și contemporane, informații economice calitative, documente specializate și resurse externe accesate prin RAG. Integrarea acestor surse va permite asistentului AI să genereze analize precise, recomandări argumentate și documentație completă, reducând semnificativ timpul necesar pentru cercetare și dezvoltare.

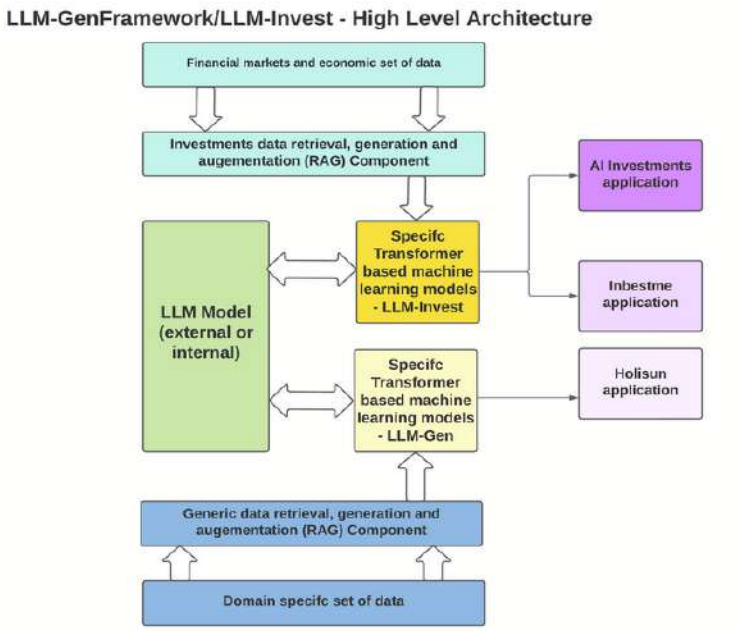

Figura 2: Arhitectura logică de înalt nivel a proiectului LLM-GenFramework

Proiectul este structurat în cinci pachete de lucru (WP), fiecare cu obiective clare, care includ activități de cercetare fundamentală, dezvoltare algoritmică, proiectare software, testare, validare și diseminare.

Diagrama GANTT prezentată în Figura 3 oferă o privire de ansamblu asupra etapelor de implementare și livrabilelor aferente proiectului.

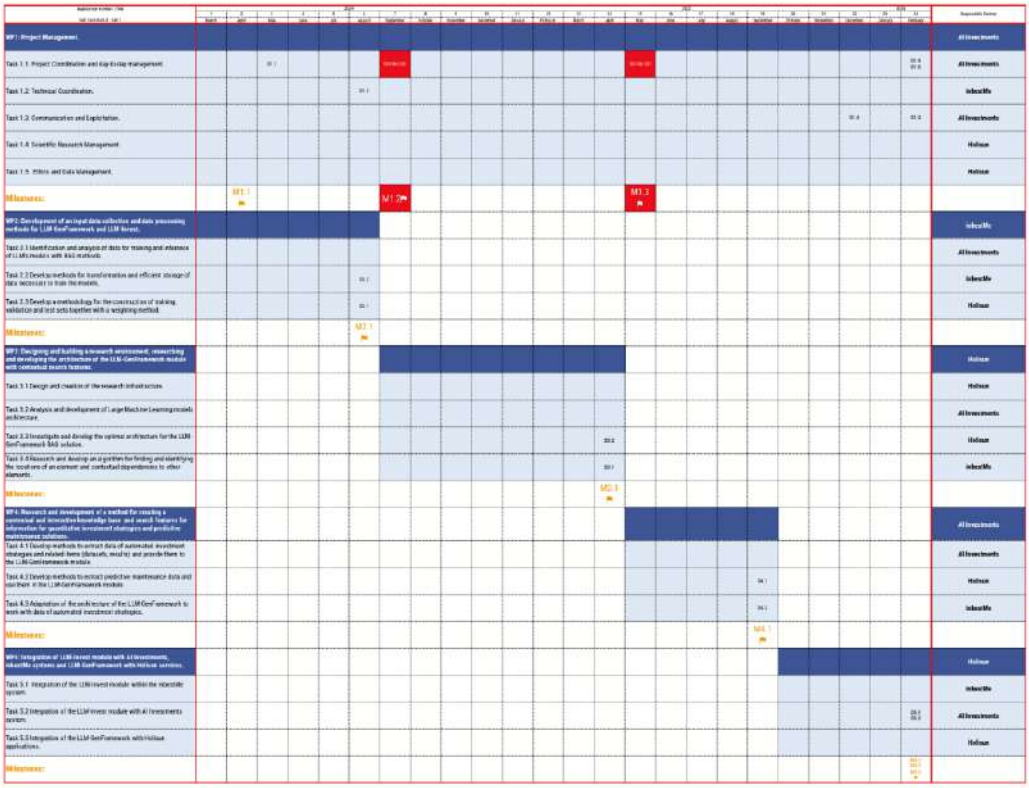

Figura 3: Diagrama Gantt care arată calendarul pachetelor de lucru și activitățile proiectului

3 Activități planificate

În perioada Ianuarie - Decembrie 2025 au fost planificate următoarele activități:

- Cercetarea literaturii de specialitate cu privire la modelele de AI;
- Elaborarea unui cadru metodologic pentru data weighting;
- Întâlniri săptămânale de progres.

4 Activități efectuate

În perioada Ianuarie - Decembrie 2025 au fost efectuate următoarele activități:

- Cercetare literaturii de specialitate cu privire la tehnici moderne de împărțire a datelor pentru modele AI;
- Analiza detaliată a dataseturilor disponibile (date financiare, date de context, date tehnice);
- Elaborarea cadrului metodologic pentru data weighting;
- 1 întâlnire fizică de progres a proiectului cu toți partenerii proiectului 24 Octombrie 2025 Cluj-Napoca, România.
- Întâlniri lunare de progres.

Figura 4: Întâlnire pe proiectul LLM-GenFramework în Cluj-Napoca, Romania

4.1 Devieri de la planificare

În perioada raportata nu au fost devieri de la planificare, sub nici un aspect.

5 Starea curentă a domeniului

Progresele accelerate în modelele de tip *Large Language Models* (LLM) au transformat radical modul în care cunoașterea poate fi modelată, interogată și utilizată în procese complexe de analiză [2, 11]. Arhitecturile bazate pe Transformer au demonstrat capacități remarcabile în înțelegerea limbajului, sinteza informației și generarea de conținut, fiind adoptate pe scară largă în aplicații industriale și științifice.

Un aspect deosebit de important îl reprezintă integrarea tehnicilor de *Retrieval Augmented Generation* (RAG), introduse în [9], care permit combinarea cunoștințelor interne ale modelelor LLM cu informații externe accesate contextual din baze de date, documente tehnice, cod sursă sau seturi de date financiare. RAG se dovedește esențial atunci când modelele trebuie să răspundă pe baza unor informații actualizate sau foarte specifice.

Studii recente evidențiază limitări importante ale modelelor mari de limbaj, inclusiv halucinațiile [6] și dificultatea în captarea relațiilor pe termen lung [1], motiv pentru care combinarea LLM cu module RAG devine o direcție tehnologică dominantă în sistemele avansate de suport decizional.

În special, literatura din ultimii ani evidențiază aplicații promițătoare ale RAG în sisteme de căutare contextuală [8], în extracția de cunoștințe [12], precum și în generarea de răspunsuri complexe, bazate pe multi-hop reasoning [10]. Acest tip de arhitecturi sunt esențiale pentru domenii unde este nevoie de analiză profundă și sinteză de informații provenite din surse diverse, așa cum se întâmplă în dezvoltarea strategiilor investiționale cantitative [5].

În același timp, sectorul investițiilor cantitative rămâne puternic dependent de instrumente statistice tradiționale, iar utilizarea tehnologiilor LLM și RAG este încă într-un stadiu incipient, în pofida potențialului lor demonstrat [4]. Conform analizei incluse în propunerea proiectului, niciuna dintre platformele actuale dedicate dezvoltării strategiilor investiționale nu integrează LLM-uri cu RAG, ceea ce marchează un decalaj tehnologic major în industrie.

Proiectul **LLM-GenFramework** se poziționează în acest context, vizând dezvoltarea unui framework generalizabil bazat pe RAG și LLM, capabil să susțină crearea, cercetarea și analiza strategiilor investiționale cantitative, precum și aplicabil în domenii non-financiare precum mentenanța predictivă, extragerea cunoștințelor sau generarea de documentație tehnică.

6 Metodologie pentru crearea seturilor de training, validare și test

Construirea corectă a seturilor de *training*, *validation* și *test* reprezintă un element fundamental în orice proiect bazat pe inteligență artificială, asigurând atât robustețea modelelor, cât și reproductibilitatea experimentelor. În cadrul proiectului LLM-GenFramework, metodologia dezvoltată se bazează pe o analiză detaliată a caracteristicilor seturilor de date, pe bune practici din domeniul învățării automate și pe principii moderne de eșantionare pentru date eterogene sau cu structură temporală.

6.1 Analiza atributelor datasetului

Primul pas constă în investigarea distribuțiilor variabilelor, a sezonalității, a corelațiilor interne, a volumetriei și a prezenței eventualelor dezechilibre. Literatura arată că o împărțire necorespunzătoare a datelor poate conduce la supraevaluarea performanțelor modelului [3]. În special pentru date financiare sau pentru seturi compuse din informații economice și tehnice, este esențială separarea coerentă a perioadelor și structurilor datelor pentru a preveni scurgerile de informații (*data leakage*) [7].

6.2 Selecția subseturilor de training, validare și test

Metodologia prevede identificarea optimă a subseturilor bazată pe:

- relevanța temporală a datelor pentru scenariile viitoare;
- diversitatea contextelor economice sau operaționale;
- reprezentativitatea în raport cu principalele distribuții identificate;
- evitarea corelațiilor excesive între subseturi.

Pentru seturile financiare este utilizată o împărțire temporală (*time-based split*), în timp ce pentru datele generice se aplică o combinație între eșantionarea stratificată și tehnici de împărțire ghidată de metadate.

6.3 Metodologia de ponderare a datelor

Un element inovator al metodologiei îl constituie integrarea unui sistem de **data weighting** care ajustează influența fiecărui element de date în funcție de:

- relevanța contextuală;

- stabilitatea valorilor în timp;
- importanța pentru obiectivele modelului;
- frecvența apariției unor tipare specifice.

Această metodă permite corectarea dezechilibrelor naturale din date, îmbunătățind procesul de învățare și reducând riscul de supraînvățare.

6.4 Experimentare și iterare

Metodologia a fost dezvoltată iterativ prin:

- derularea de experimente comparative între diferite metode de eșantionare;
- evaluarea impactului asupra performanței modelelor în contexte multiple;
- identificarea combinațiilor care oferă stabilitatea cea mai ridicată;
- rafinarea criteriilor de selecție pe baza rezultatelor observate.

Acest proces a condus la formularea unei metodologii generale, reutilizabile și aplicabile și în alte proiecte de învățare automată, în special în aplicațiile LLM unde diversitatea și consistența datelor sunt esențiale pentru obținerea unor rezultate robuste.

6.5 Rolul metodologiei în arhitectura generală

Metodologia dezvoltată în acest task este critică pentru etapele ulterioare ale proiectului, întrucât permite:

- antrenarea consistentă a modelelor în WP3 și WP4;
- comparabilitatea experimentelor;
- reducerea variabilității necontrolate;
- reproductibilitatea completă a rezultatelor științifice.

Această componentă stă la baza credibilității întregului proces de construire a modelului LLM-GenFramework și asigură fundamentul experimental necesar pentru construirea produsului specializat LLM-Invest.

7 Rolul HOLISUN în cadrul proiectului

HOLISUN deține un rol esențial în cadrul proiectului **LLM-GenFramework**, contribuind atât la dezvoltarea componentelor metodologice fundamentale, cât și la implementarea soluțiilor software ce vor constitui nucleul framework-ului final. În mod particular, HOLISUN este responsabil pentru activități cheie din WP2 și WP3, având un rol direct în crearea metodologiei de analiză a datelor, generarea seturilor de training, validare și test, precum și integrarea tehnicilor de tip LLM și RAG în cadrul arhitecturii generale a proiectului.

Contribuții metodologice.
În WP2, HOLISUN a coordonat dezvoltarea unei metodologii generalizabile pentru construcția seturilor de date pentru antrenare, validare și testare. Această metodologie include mecanisme de analiză a atributelor datasetului, proceduri de împărțire a datelor în mod robust și reproductibil, precum și definirea unei metode de ponderare a datelor în funcție de relevanță și structură. Acest rezultat reprezintă fundamentul experimentelor ce vor fi realizate în WP3 și WP4, asigurând consistența și calitatea pipeline-ului de dezvoltare AI.

Contribuții tehnice și arhitecturale.
În WP3, HOLISUN contribuie la integrarea tehnicilor LLM și RAG în cadrul framework-ului, la definirea interfețelor software necesare comunicării dintre module, precum și la implementarea componentelor de preprocesare, query expansion și extracție contextuală. De asemenea, HOLISUN participă activ la proiectarea fluxurilor de lucru pentru produsul specializat **LLM-Invest**, având rol în dezvoltarea prototipurilor și în documentarea arhitecturii.

Contribuții aplicate și cross-domain.
HOLISUN aduce o experiență practică relevantă în aplicarea modelelor AI și în dezvoltarea infrastructurilor software scalabile, permițând proiectului să fie extins dincolo de domeniul investițiilor. Astfel, rezultatele LLM-GenFramework vor fi adaptate în aplicații precum mentenanța predictivă, analiza documentară tehnică sau dezvoltare asistată de AI, consolidând reutilizabilitatea și generalitatea tehnologiilor dezvoltate.

Prin implicarea în aceste activități, HOLISUN asigură atât contribuția științifică, cât și implementarea practică necesară pentru transformarea framework-ului dintr-un concept teoretic într-o soluție funcțională, scalabilă și aplicabilă în multiple industrii.

8 Extras din planul de riscuri

Work Package-urile WP2 și WP3 reprezintă nucleul de cercetare metodologică și algoritmică al proiectului LLM-GenFramework. Prin natura lor, acestea implică riscuri tehnice, științifice și operaționale specifice proiectelor bazate pe inteligență artificială, în special în domenii emergente precum integrarea LLM cu RAG, gestionarea datelor eterogene și dezvoltarea metodologiilor de analiză.

Tabelul 1 prezintă principalele riscuri identificate și metodele de mitigare aferente.

Tabela 1: Riscuri identificate în WP2 și WP3 și măsuri de mitigare aferente.

Risc (WP2 / WP3)	Prob.	Impact	Valoare	Măsură de mitigare
Date incomplete, neechilibrate sau cu distribuții eterogene	3	4	12	Analiză detaliată a dataseturilor; utilizarea metodologiei dezvoltate în T2.3 pentru eșantionare robustă, ponderare adaptivă și validare intermediară.
Împărțirea incorectă a datelor în training/validation/test, cu risc de *data leakage*	2	5	10	Aplicarea strictă a regulilor metodologice; folosirea split-urilor temporale și stratificate; auditarea manuală a partiționării.
Inconsistențe între sursele de date utilizate în RAG	2	4	8	Standardizare, curățare și indexare corectă a datelor; validarea surselor înainte de integrare.
Performanță scăzută a modulelor RAG din cauza surselor de informații insuficient relevante	3	3	9	Evaluarea periodică a calității corpusului; înlocuirea/documentarea surselor; adăugarea unor filtre contextuale pe baza prompting-ului ghidat.
Halucinații ale modelelor LLM în generarea analizelor sau strategiilor	2	5	10	Utilizarea tehnicilor RAG, a prompting-ului constrâns și a verificărilor multi-pas; includerea validării automate cu reguli domain-specific.
Dificultăți în integrarea modulelor software (LLM, RAG, pipeline-uri)	3	4	12	Testare incrementală; definirea API-urilor stabile; prototipare timpurie; revizuirea arhitecturii împreună cu partenerii tehnici.
Performanță necorespunzătoare a prototipului LLM-Invest în scenarii reale	2	4	8	Testare pe multiple seturi de date financiare; ajustarea prompting-ului și a contextului; adăugarea de mecanisme de feedback iterativ.
Întârzieri în generarea documentației metodologice și arhitecturale	2	3	6	Planificare incrementală; revizuiri lunare; responsabilități distribuite și clar definite în WP2/WP3.

Riscurile sunt monitorizate continuu în cadrul întâlnirilor de progres, iar acțiunile de mitigare sunt actualizate pe măsură ce noi date, rezultate experimentale și module software devin disponibile în proiect.

9 Rezultatele proiectului

9.1 Livrabile

In perioada raportata am furnizat livrabile din tabela 2 și am început lucrul intens asupra celorlalte livrabile.

Tabela 2: Tabel cu livrabile și statusul acestora.

Nr. livrabil	Termen	Livrabil	Status livrabil
D2.1	M6	Descrierea setului de date integrate și utilizate în proiect	Livrat M6
D2.2	M6	Structuri tehnice de baze de date și algoritmi de transformare necesari pentru preluarea și transformarea datelor pentru proiect	Livrat M6

10 Diseminare și exploatare

10.1 Activități de diseminare

Proiectul a fost diseminat în următoarele moduri:

- pe pagina web: https://research.holisun.com/ro/proiecte/predictive-analysis/llm-genframework-ro, având un număr de 150 de vizitatori lunari
- pe contul de LinkedIn: https://www.linkedin.com/company/holisun, cu peste 550 de adepți
- pe pagina de Facebook: https://www.facebook.com/Holisun.IT/, având peste 2400 de urmăritori

Au fost desfășurate o serie de activități de diseminare în cadrul unor evenimente de afaceri, expoziții și evenimente de brokeraj sau networking, listate în Tabelul 3.

Tabela 3: Lista de activități de diseminare

Nume	Data	Link	Participanți	Rezultate
International Mobility Days 2025	19.11.2025-21.11.2025	https://www.mobility-days.at/	Delinschi Daniela	Prezentare *LLM-GenFramework*
InnoComp Research Project Panel	24.10.2025	https://innocomp.eu/storage/app/media/files/InnoComp-Programme.pdf	Mihai Crisan	Prezentare *LLM-GenFramework*

Figura 5: InnoComp Research Project Panel - Prezentare proiect LLM-GenFramework

11 Concluzii

În primul an al proiectului **LLM-GenFramework** au fost puse bazele științifice, tehnice și metodologicenecesare dezvoltării unui framework avansat bazat pe Large Language Models (LLM) și Retrieval Augmented Generation (RAG), destinat sprijinirii proceselor de creare, cercetare și analiză a strategiilor investiționale cantitative, cu aplicabilitate extinsă în alte domenii economice și tehnice.

Activitățile desfășurate au inclus analiza aprofundată a literaturii privind LLM, RAG și metodologiile moderne de procesare a datelor, precum și investigarea celor mai relevante tehnici de eșantionare și de împărțire a dataseturilor. În cadrul WP2 a fost finalizată metodologia pentru construirea seturilor de training, validare și test, împreună cu o strategie de ponderare a datelor, asigurând astfel un cadru solid pentru experimentele algoritmice care vor avea loc în WP3 și WP4.

De asemenea, au fost demarate activități de proiectare a arhitecturii software pentru integrarea LLM și RAG, precum și definirea fluxurilor conceptuale ale framework-ului general și ale produsului specializat **LLM-Invest**. Aceste contribuții au permis crearea unei viziuni tehnice coerente asupra soluției finale și stabilirea direcțiilor de dezvoltare pentru anul următor.

Interacțiunea constantă între parteneri, prin întâlniri lunare de progres și schimburi periodice de rezultate, a asigurat alinierea metodologică și sincronizarea activităților între pachetele de lucru. Toate aceste elemente au contribuit la crearea unei fundații solide pentru trecerea la fazele de implementare, integrare și experimentare.

În anul următor, proiectul va avansa către dezvoltarea componentelor algoritmice, integrarea tehnicilor RAG în modulul central, construirea prototipului LLM-GenFramework și implementarea produsului specializat LLM-Invest. Rezultatele obținute în acest prim an vor juca un rol esențial în asigurarea robusteții, scalabilității și aplicabilității soluției finale în domeniul investițiilor cantitative și în alte industrii ce pot beneficia de asistență AI avansată.

Referințe

[1] Bommasani, R., et al.: On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 (2021)

[2] Brown, T., et al.: Language models are few-shot learners. NeurIPS (2020)

[3] Dietterich, T.: Approximate statistical tests for comparing supervised classification learning algorithms. Neural Computation (1998)

[4] Group, E.Q.R.: Review of modern quantitative investment technologies (2023)

[5] Investments, A.: Analysis of quantitative strategy development using llms (2024), internal report

[6] Ji, S., et al.: Survey of hallucination in natural language generation. ACM Computing Surveys (2023)

[7] Kaufman, S., et al.: Leakage in data mining: Formulation, detection, and avoidance. ACM KDD (2012)

[8] Khandelwal, U., et al.: Generalization through memorization: Nearest neighbor language models. ICLR (2021)

[9] Lewis, P., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. NeurIPS (2020)

[10] Smith, A., et al.: Multi-hop reasoning in llms. EMNLP (2021)

[11] Vaswani, A., Shazeer, N., et al.: Attention is all you need. NeurIPS (2017)

[12] Yu, A., et al.: Coco: A benchmark for multi-document reasoning. ACL (2022)

info@holisun.com
http://www.holisun.com
+4-0262-226.778
Versiune document: 1.0 din 24.11.2025