

Model lingvistic de mari dimensiuni (LLM) specializat pentru patrimoniul cultural și bazat pe reutilizarea datelor și pe cercetarea eficienței computaționale (MOPAT)

- raport de maturitate tehnologică -

1. Datele de identificare a rezultatului

Denumirea rezultatului: Model lingvistic de mari dimensiuni (LLM) specializat pentru patrimoniul cultural și bazat pe reutilizarea datelor și pe cercetarea eficienței computaționale (MOPAT)

Nivelul de maturitate tehnologică (TRL): 5

Titlul proiectului: *Prezervarea, transformarea și studierea patrimoniului lingvistic românesc în contextul globalizării (PATLIRO)*

Codul proiectului: PN-IV-P8-8.1-PRE-HE-ORG-2023-0035

Contractul de finanțare: 9PHE/01.11.2023

Durata: 01.11.2023–31.10.2025

Programul de finanțare: PNCDI IV - Programul Cooperare europeană și internațională - Subprogramul Orizont Europa – Tipul de proiect Premiere Orizont Europa - Instituții

Autoritatea contractantă: Unitatea Executivă pentru Finanțarea Învățământului Superior, a Cercetării, Dezvoltării și Inovării (UEFISCDI)

2. Contextul proiectului PATLIRO și poziționarea rezultatului analizat

Proiectul PATLIRO a inclus explorarea modalităților de conservare și adaptare a patrimoniului lingvistic românesc în fața globalizării și internaționalizării, precum și stimularea învățării limbii române de către studenții străini. În acest context integrat, modelul lingvistic MOPAT (rezultat-cheie al componentei tehnologice a proiectului) oferă un instrument automatizat pentru generarea de descrieri culturale, contribuind la îmbogățirea resurselor educaționale și de cercetare în domeniu.

Originalitatea proiectului constă în combinarea lingvisticii computaționale cu patrimoniul cultural, printr-o arhitectură care valorifică datele existente și tehnici moderne de inteligență artificială. Prin acest demers, PATLIRO adresează un decalaj major în analiza automată a textelor literare și culturale, creând un ecosistem digital pentru studiul și diseminarea patrimoniului lingvistic.

Rezultatul analizat în prezentul raport este modelul lingvistic de mari dimensiuni MOPAT, dezvoltat ca prototip funcțional în cadrul componentei tehnologice a proiectului. Acesta constă într-o arhitectură bazată pe modele open-source adaptate limbii române, antrenate și evaluate pe corpusuri lingvistice curățate și benchmark-uri academice relevante, cu scopul de a genera conținut coerent, adecvat cultural și stilistic. MOPAT reprezintă o concretizare aplicativă a obiectivelor proiectului, integrând

cercetarea lingvistică, resursele digitale dezvoltate și metodele moderne de optimizare a modelelor mari de limbă într-un sistem validat experimental.

3. Descrierea rezultatului exploatabil

MOPAT reprezintă un model lingvistic de mari dimensiuni (LLM) specializat în patrimoniul cultural, conceput pentru a genera descrieri narrative coerente ale artefactelor, monumentelor și siturilor culturale. Modelul integrează principiile de circularitate a datelor, valorificând corpusurile și fișele existente și utilizând o varietate de LLM-uri open-source, inclusiv modele românești disponibile prin platforma OpenLLM, precum RoLLama, RoGemma și RoMistral. Peste această arhitectură este construit un mecanism de tip Retrieval-Augmented Generation (RAG), care permite interogarea dinamică a bazelor de date specializate și integrarea contextuală a informației validate în procesul de generare, sporind astfel acuratețea factuală, trasabilitatea surselor și adecvarea discursului la specificul patrimoniului cultural.

4. Prezentarea ipotezelor cercetării și a rezultatelor obținute

Inițiativa a pornit de la următoarele două ipoteze: 1) un LLM specializat, ajustat pe date de patrimoniu lingvistic și îmbogățit cu fișe de metadate poate genera descrieri coerente și relevante ale artefactelor culturale și 2) implementarea unui mecanism RAG va permite modelului să integreze informații noi din surse externe, sporind astfel acuratețea factuală și contextuală a răspunsurilor.

MOPAT poate genera descrieri detaliate și coerente despre obiecte și situri culturale, utilizând un limbaj adecvat domeniului. Generările sale îmbină informațiile învățate în antrenare cu date specifice fiecărui artefact recuperate în timp real, obținându-se texte fluente și adaptate contextului. Datorită mecanismului RAG, răspunsurile modelului sunt mai precise și actualizate, comparativ cu cele produse de un model.

Pe lângă generarea de conținut narativ, arhitectura RAG are un rol transversal, susținând o serie de componente analitice și aplicative dezvoltate în cadrul proiectului. Aceasta permite integrarea și corelarea datelor provenite din surse eterogene – economice, financiare, ecologice, literare, istorice sau administrative – și generarea de sinteze structurate, adaptate scopului fiecărui modul. În aplicațiile dedicate industriilor creative, mecanismul facilitează analiza multidimensională a performanței prin corelarea indicatorilor financiari și de mediu. În zona studiilor literare, sistemul sprijină identificarea și descrierea microgenurilor, oferind prezentări coerente ale operelor și autorilor, care includ context istoric și cultural, trăsături stilistice și elemente tematice. În aplicațiile orientate spre patrimoniul urban, modelul integrează registre oficiale și surse istorice pentru a genera reprezentări cartografice și recomandări tematice relevante pentru dezvoltarea comunitară sustenabilă. De asemenea, în domeniul patrimoniului arheologic, arhitectura permite elaborarea de prezentări structurale ale obiectelor ceramice, corelând documentația tehnică, sursele istorice și datele descriptive într-o evaluare sistematică și contextualizată.

5. Validarea rezultatelor

Rezultatul MOPAT a fost validat formal prin rapoartele proiectului PATLIRO și prin testări experimentale dedicate.

În prezent, sistemul este testat pe date reale provenite din proiecte de cercetare în derulare. Deși arhitectura completă de tip RAG urmează să fie implementată în etapa următoare, funcționalitățile actuale depășesc nivelul strict experimental și demonstrează validarea tehnologiei într-un mediu relevant, justificând încadrarea RCE-ului la TRL 5. RCE-ul a fost deja validat pe cinci familii de modele, demonstrând stabilitate și consistență în rezultate. Modelele conversaționale generează natural răspunsuri în română, iar calitatea acestora a fost validată de preferințele umane.

Validarea extensivă, realizată utilizând peste 70.000 de evaluări per model și patru suite independente (academică, conversațională, downstream, culturală), mediată pe configurații few-shot și testată pe arhitecturi diferite, asigură relevanță statistică și robustețe.

Diferențele dintre laborator și scenariul operațional au fost analizate și documentate, incluzând protocoale de evaluare, judecător automat vs. feedback uman și date traduse vs. native.

Modelele au fost publicate online pe Huggingface (<https://huggingface.co/open-llm-leaderboard>, accesat la data de 15 februarie 2026), există documentație completă, cod open-source reproductibil și proceduri de evaluare detaliate.

6. Justificare TRL 5

Nivelul de maturitate tehnologică al modelului MOPAT se încadrează în TRL 5, tehnologia fiind validată în condiții relevante de funcționare (mediul industrial). Corpusurile lingvistice românești extinse dezvoltate în cadrul proiectului PATLIRO și componentele de generare de conținut cu ajutorul modelelor mari de limbă constituie suportul tehnic și metodologic al validării nivelului TRL 5. Integrarea acestora într-o arhitectură funcțională bazată pe modele open-source adaptate limbii române a permis testarea sistemului pe volume consistente de date, în scenarii variate de generare și reformulare de conținut.

Validările în mediul relevant au fost realizate prin testarea modelelor OpenLLM-Ro pe multiple benchmark-uri academice (ARC, MMLU, Winogrande, HellaSwag, GSM8k, TruthfulQA) și evaluarea calității conversaționale prin comparații cu preferințele umane traduse (HelpSteer, 15.000 exemple). Rezultatele arată îmbunătățiri consistente de 3–8 puncte procentuale pe majoritatea benchmark-urilor și creșteri măsurabile în coerența, utilitatea și adecvarea culturală a răspunsurilor, demonstrând robustețea metodologiei și importanța alinierii cu preferințele utilizatorilor. Modelele generează răspunsuri în limba română pe MT-Bench fără prompting explicit suplimentar, față de modelele standard care răspund preponderent în engleză chiar și la prompturi românești. Acest comportament natural reflectă fidelitatea față de cazul de utilizare real unde utilizatorii români așteaptă răspunsuri în limba lor maternă. Etapa de DPO (Direct Preference Optimization) folosind 15.000 de exemple de

preferințe umane traduse produce creșteri măsurabile în calitatea conversațională (+0.4-0.7pp MT-Bench), apropiind modelele de stilul de comunicare preferat de utilizatori. Inspecția manuală a răspunsurilor confirmă tranziția de la un ton rigid, tehnic, la unul natural. Testele efectuate pe benchmark-uri academice, task-uri downstream și evaluări conversaționale arată integrarea coerentă a tuturor componentelor și performanță stabilă, cu rezultate relevante statistic, ceea ce confirmă validarea tehnologiei într-un mediu relevant și maturitatea prototipului la nivel TRL 5.

7. Concluzie

Modelul lingvistic de mari dimensiuni (LLM) specializat pentru patrimoniul cultural și bazat pe reutilizarea datelor și pe cercetarea eficienței computaționale (MOPAT) este un rezultat exploatabil al proiectului PATLIRO, cu grad de maturitate tehnologică TRL 5 – concept demonstrat analitic, conform definițiilor agreate la nivel național și internațional.

25.10.2025

Director proiect,
Prof. univ. dr. Simina-Maria Terian-Dan