TABEL COMPARATIV — MOTOARE AI

Procedura de achiziție: Website integrat cu sistemele informaționale
OCID: ocds-b3wdp1-MD-1777987222408

Autoritate contractantă: SA „ENERGOCOM" (IDNO 1004600074938)

Acest document prezintă analiza comparativă a opțiunilor disponibile pentru motorul AI care va alimenta modulul de chat al website-ului „Energocom", evaluat pe 4 dimensiuni cerute explicit de Caietul de sarcini: cost total pe durata contractului + garanție (12 luni), latență de răspuns (primul token și mesaj complet), nivel de securitate și conformitate cu cerințele de securitate de la Cap. 6.3, prezența clauzei explicite de no-training pe datele beneficiarului.

Valorile de preț și performanță reprezintă estimări la data pregătirii ofertei. Verificarea finală la momentul kick-off-ului proiectului poate ajusta selecția în funcție de tarifele actualizate ale providerilor și de cerințele specifice ale beneficiarului.

Tabel comparativ — 4 opțiuni evaluate

Criteriu	OpenAI GPT-4o (API tier enterprise)	Anthropic Claude 3.5 Sonnet (API)	Google Gemini Pro (Vertex AI Enterprise)	Model local self-hosted (Llama 3 70B / Mistral Large 2)
Cost input (per 1M tokens)	~ 2,50 USD	~ 3,00 USD	~ 1,25 USD	N/A (cost cu hardware + ops, nu per-token)
Cost output (per 1M tokens)	~ 10,00 USD	~ 15,00 USD	~ 5,00 USD	N/A
Cost total estimat 12 luni (volum mediu: 50-100 conversații/zi, ~500 tokens/exchange, ~70 M tokens/an)	~ 500-700 USD	~ 750-1100 USD	~ 350-500 USD	Hardware GPU furnizat de SA „Energocom" (parte din infrastructura de hosting, conform clarificării R-4 publicate 12.05.2026 — NU intră în prețul ofertei). Costul pentru ofertant: doar operațiuni MLOps + integrare (~200-400 USD/an) — încadrat în prețul fix.
Latență primul token (p50)	~ 250-400 ms	~ 350-500 ms	~ 200-350 ms	~ 500-900 ms (depinde de GPU)
Latență primul token (p95)	~ 600-900 ms	~ 700-1000 ms	~ 500-800 ms	~ 1200-2000 ms

Fereastră context	128K tokens	200K tokens	până la 2M tokens	8K-128K (funcție de model și cuantizare)
Suport multilingv RO / RU / EN	Excelent — toate 3 limbi cu acuratețe înaltă	Excelent — printre cele mai bune la RO	Excelent — multilingv nativ	Bun — Llama 3 70B are RO/RU acceptabil; necesită fine-tuning pentru calitate optimă
Streaming token-by-token	Da, SSE / WebSocket	Da, SSE	Da, SSE	Da, prin vLLM / TGI / Ollama
Clauza no-training pe datele beneficiarului	Da, by default pe API (Enterprise tier confirmat prin DPA)	Da, by default pe API (DPA explicit)	Da, pe Vertex AI Enterprise (DPA inclus)	Total — datele nu părăsesc niciodată infrastructura proprie
Localizare procesare date	SUA (UE prin Azure OpenAI dacă se alege)	SUA (regiuni UE în roadmap)	UE (regiune europe-west) sau SUA, configurabilă	Local — RM sau UE conform cerinței CdS
SLA disponibilitate	99,9% (enterprise)	99,9% (enterprise)	99,95% (Vertex AI)	Funcție de propria infrastructură (țintă 99,5%)
Conformitate GDPR	Da (DPA disponibil)	Da (DPA disponibil)	Da (Vertex AI compliance)	Total — control direct asupra datelor
Izolare tenant	Strictă pe API key	Strictă pe API key	Strictă pe proiect GCP	N/A (single tenant nativ)
Avantaj cheie	Cel mai utilizat ecosistem, multă documentație, suport bun RO	Cea mai bună calitate de raționament și suport pentru RO; siguranță sporită prin Constitutional AI	Cel mai bun raport calitate/preț; context window enorm utilă pentru RAG mare	Independență totală de provideri externi; predictibilitate cost; zero risc no-training
Dezavantaj cheie	Cost mediu; dependență de SUA	Cost cel mai ridicat dintre cei 3 cloud	Documentație ceva mai săracă; mai puține integrări terțe	Necesită ca Energocom să furnizeze GPU NVIDIA L4 / A10G / A100 în infrastructura de hosting; necesită expertiză MLOps internă pentru operare și optimizare.

Concluzie și motorul AI propus

Pe baza analizei comparative, ofertantul propune următoarea strategie:

Opțiunea principală: arhitectură multi-provider cu router

Implementarea utilizează un strat de abstractizare (OpenRouter sau echivalent) care permite rutarea cererilor către providerul AI optim în funcție de tip de interogare, cost și disponibilitate. Configurația implicită la lansare:

- Anthropic Claude 3.5 Sonnet ca motor primar pentru întrebări care necesită raționament complex (proceduri, reclamații, întrebări legale) — cea mai bună calitate la limba română.
- Google Gemini Pro (Vertex AI) ca motor secundar pentru întrebări scurte și FAQ standard — cel mai bun raport cost/calitate, latență mică.

- OpenAI GPT-4o disponibil ca fallback automat în caz de incident la oricare din cele două.

Toate cele trei sunt utilizate prin tier enterprise care **oferă opțiunea** de DPA (Data Processing Agreement) și clauza explicită de no-training pe datele beneficiarului (DPA-ul este un acord contractual separat pe care SA «Energocom» îl va semna cu providerul AI ales, nu un document livrat ca parte a acestei oferte și al cărui angajament de semnare nu este luat de Energocom în momentul ofertei), cu pseudonimizarea datelor la nivel de PII redaction înainte de orice apel către API extern (conform Cap. 6.3 CdS).

Opțiunea alternativă: model local self-hosted

Dacă SA „Energocom" preferă suveranitate totală asupra datelor și dispune de infrastructură GPU (NVIDIA L4 / A10G / A100 — cerințe specificate în Memoriul Tehnic, conform clarificării R-4 publicate care confirmă că hardware GPU este parte din hosting suportat de beneficiar), propunerea alternativă este self-hosting cu Llama 3 70B sau Mistral Large 2 pe infrastructura beneficiarului. Această opțiune elimină total dependența de providerii externi și elimină orice risc legat de no-training. Prețul ofertei rămâne fix indiferent de opțiunea aleasă — doar arhitectura se modifică. Decizia finală între opțiunea cloud multi-provider și opțiunea self-hosted se ia la kick-off, după discuții cu echipa „Energocom" privind capacitatea infrastructurală și politicile interne de securitate.

Cost total inclus în prețul fix al ofertei

Conform clarificării R-4 publicate de autoritatea contractantă, toate costurile motorului LLM (API extern SAU model local) pe durata contractului + garanție (12 luni) sunt incluse în prețul fix al ofertei. Estimarea internă pentru opțiunea multi-provider cloud este de 500-1.100 USD / an la volumul estimat. Pentru opțiunea self-hosted, hardware-ul GPU este furnizat de SA „Energocom" (parte din infrastructura de hosting, R-4), iar ofertantul preia doar costul operațional MLOps + integrare — semnificativ mai mic decât opțiunea cloud, dar prețul total al ofertei rămâne fix.

Ofertant: Das Soft Plus S.R.L. (brand CoRLab Tech)

IDNO: 1019600011052

Adresa: MD-2001, str. Lev Tolstoi 74, ap. 78, mun. Chișinău, Republica Moldova

Reprezentant: Afanasie BUTUCEA

Funcția: Administrator

Data: 19.05.2026

Digitally signed by Butucea Afanasie
Date: 2026.05.19 12:20:22 EEST
Reason: MoldSign Signature
Location: Moldova
MOLDOVA EUROPEANĂ

Semnătură electronică aplicată cu MSign (eIDAS calificată)