

LLM-JI V SLUŽBI PREGLEDNOSTI: KAKO SMO STRUKTURIRALI IT POGODBE IZ PORTALA EJN

Stevanče Nikoloski¹, Kristina Valenčič²

¹ Result, d.o.o., Celovška cesta 182, 1000 Ljubljana

² Direktorat za razvoj digitalnih rešitev in podatkovno ekonomijo Ministrstvo za digitalno preobrazbo, Davčna ulica 1, 1000 Ljubljana

stevance.nikoloski@result.si, kristina.valencic@gov.si

V prispevku predstavljamo praktični primer uporabe generativne umetne inteligence (GenAI) za digitalno transformacijo javno dostopnih IT pogodb, objavljenih na portalu ejN (elektronsko javno naročanje). Projekt se je osredotočil na izziv strukturiranja več tisoč dokumentov, ki so pogosto dostopni zgolj v skeniranem PDF formatu, brez standardizirane strukture in z izrazito vsebinsko raznolikostjo.

Rešitev temelji na kombinaciji odprtokodnega OCR orodja (Tesseract), layout segmentacijskega modela za prepoznavanje ključnih delov dokumenta ter uporabi velikih jezikovnih modelov (LLM), kot je GPT-4, za semantično razumevanje in luščenje ključnih metapodatkov. Med obdelanimi metapodatki so: naročnik, izvajalec, datum sklenitve pogodbe, obdobje trajanja, vrednost pogodbe, predmet, CPV koda, številka javnega naročila in drugi podatki, pomembni za nadaljnjo analitiko.

Ključnega pomena pri strukturiranju podatkov je bila natančna identifikacija in razgradnja informacij o vrstah storitev, opredeljenih v pogodbah. V praksi se pogosto pojavljajo primeri, kjer posamezna pogodba vključuje več različnih tipov storitev (npr. svetovanje, dobava opreme, vzdrževanje). S pomočjo LLM smo uvedli granularno luščenje po tipu storitve, kar omogoča analizo pogodb tudi na ravni posameznega storitvenega sklopa znotraj ene pogodbe.

Na koncu so strukturirani podatki predstavljeni v obliki Power BI interaktivnih vpogledov, ki omogočajo uporabnikom različne načine filtiranja, vizualizacije in dodatne analize. Uporabnik lahko hitro izlušči informacije o posameznih izvajalcih, storitvah, vrednostih pogodb ali časovni dinamiki sklepov, kar predstavlja uporabno orodje za preglednost in strateško odločanje.

Prispevek izpostavi konkretne tehnične in organizacijske izzive pri obdelavi nehomogenih javnih dokumentov in pokaže, kako lahko sodobni AI pristopi bistveno povečajo transparentnost, dostopnost in analitično uporabnost podatkov iz področja javnih naročil.

Ključne besede: generativna umetna inteligenca; veliki jezikovni modeli (LLM); javna naročila; luščenje metapodatkov; digitalna obdelava dokumentov

LLMS IN THE SERVICE OF TRANSPARENCY: STRUCTURING IT CONTRACTS FROM THE EJN PUBLIC PROCUREMENT PORTAL

This contribution presents a practical application of generative AI (GenAI) for the digital transformation of publicly available IT contracts published on Slovenia's eJN public procurement portal. The project addressed the challenge of structuring thousands of scanned PDF documents lacking standard format and semantic consistency.

The solution combines open-source OCR (Tesseract), layout segmentation models, and large language models (LLMs) such as GPT-4 for semantic extraction of key metadata, including contracting authority, supplier, contract value, duration, CPV code, and more.

A key focus was the granular extraction of service types, as many contracts cover multiple services (e.g., consulting, equipment supply, maintenance). This enables analysis at the level of individual service components within a single contract.

Finally, structured data is visualized through interactive Power BI dashboards, offering filtering, drill-downs, and deeper insights for transparency and strategic use.

The paper highlights technical and organizational challenges and demonstrates how modern AI approaches can improve data accessibility, transparency, and analytical value in public procurement.

Keywords: generative artificial intelligence; large language models (LLM); public procurement; metadata extraction; digital document processing