علیرضا اخوان‌پور

نام دوره: ساخت چت بات فارسی کاملا آفلاین

طول دوره: ۴۰ ساعت

نوع ارائه: آنلاین

زمان کلاس: ۳۰ دی ماه - سه شنبه ها – ساعت ۱۶ تا ۱۹

پشتیبانی: در گروه اختصاصی تلگرامی دوره

پیش نیاز شرکت در دوره: آشنایی مقدماتی با Python،

آشنایی پایه با مفاهیم هوش مصنوعی، تجربه کار با Linux

یا Docker (ترجیحا)، آشنایی پایه با Web APIها

سطح دوره:متوسط تا پیشرفته

مدرک شرکت در دوره: دارد

مدرس دوره

علیرضا اخوان‌پور

« ۸ سال سابقه ی مدیر فنی در مجموعه دانش بنیان "شناسا"

« توسعه دهنده پایتون و فعال در پروژه های بینایی ماشین با یادگیری عمیق

« مشاور و منتور هوش مصنوعی در شتاب دهنده همتک، همراه اول، ایبیکام

« مدرس دانشگاه شهید رجایی از سال ۱۳۹۴

« ارائه ورکشاپ و تدریس در دانشگاه های صنعتی شریف، امیرکبیر و تهران

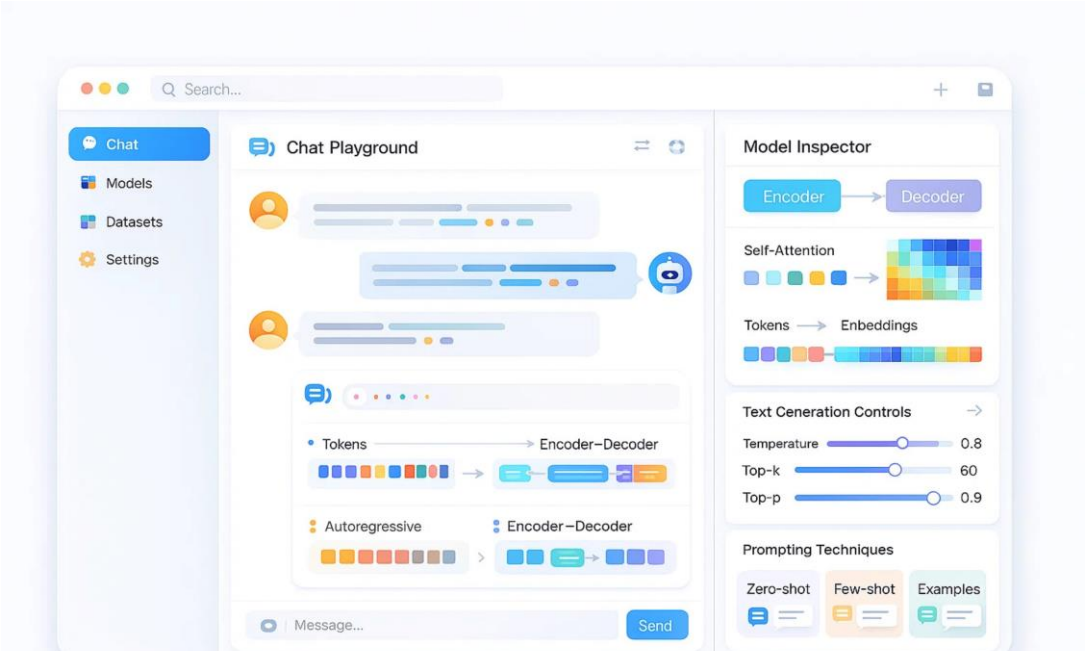
« تجربه ی + ۲۰۰۰ ساعت تدریس مرتبط

علیرضا اخوان‌پور

فصل اول : مبانی مدل‌های زبانی و معماری ترنسفورمر

در این فصل، مدل‌های زبان بزرگ (LLM)، معماری ترنسفورمر و مکانیزم Self-Attention معرفی می‌شوند. دانشجو مفاهیم پایه‌ای چون توکن‌ها، امبدینگ‌ها و تفاوت مدل‌های Autoregressive و Encoder-Decoder را می‌آموزد. تمرکز اصلی بر اصول کنترل خروجی با پارامترهای تولید متن (مانند Temperature) و تکنیک‌های پرامپت است.

- **مفاهیم پایهLLM :**
  - مدل زبانی چیست و کاربردهای آن.
  - توکن و توکنایزیشن، مفاهیم امبدینگ.
- **معماری ترنسفورمر:**
  - مکانیزم Self-Attention و نقش آن در LLMها.
  - ساختار Encoder و Decoder و تفاوت مدل‌های autoregressive و encoder-decoder.
- **کنترل تولید متن:**
  - پارامترهای مهم: Top-p ,Top-k ,Temperature و تأثیر آن‌ها بر خروجی مدل.
- **تکنیک‌های پرامپت:**
  - Zero-Shot ,Few-Shot و مثال‌های عملی

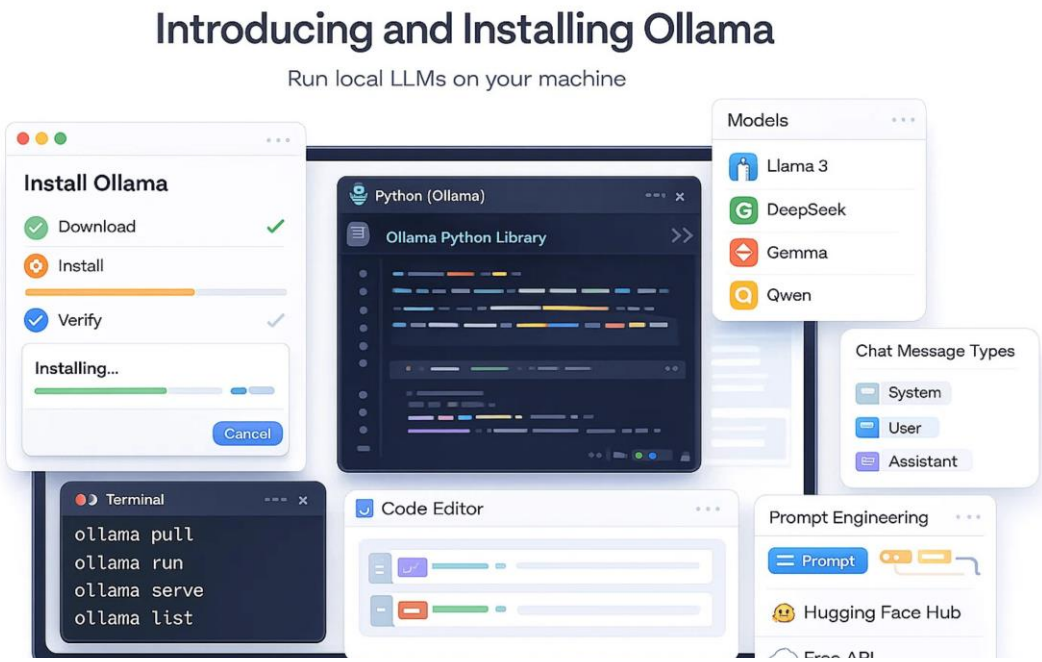


علیرضا اخوان‌پور

فصل دوم: مبانی، نصب و راه‌اندازی مدل‌های محلی

فصل پیش رو به شما کمک می‌کند تا مدل‌های LLM را به صورت محلی راه‌اندازی کنید. این شامل نصب Ollama، کار با دستورات خط فرمان و اجرای مدل‌های معروف است. محتوای این بخش تعامل با مدل‌ها از طریق کتابخانه پایتون Ollama و آشنایی با مفاهیم اولیه Prompt Engineering و Hugging Face Hub را پوشش می‌دهد.

- **مقدمه و آماده‌سازی محیط:**
  - معرفی و نصب **Ollama** برای اجرای مدل‌های محلی.
  - کار با دستورات خط فرمان (list, serve, run, pull) Ollama.
  - اجرای مدل‌های معروف (Qwen, Gemma, DeepSeek, Llama ۳).
- **تعامل باLLM :**
  - استفاده از کتابخانه Ollama در پایتون.
  - مفاهیم اولیه Prompt Engineering و انواع پیام‌ها (System, User, Assistant).
  - آشنایی با Hugging Face Hub و استفاده ازAPI های رایگان آن.

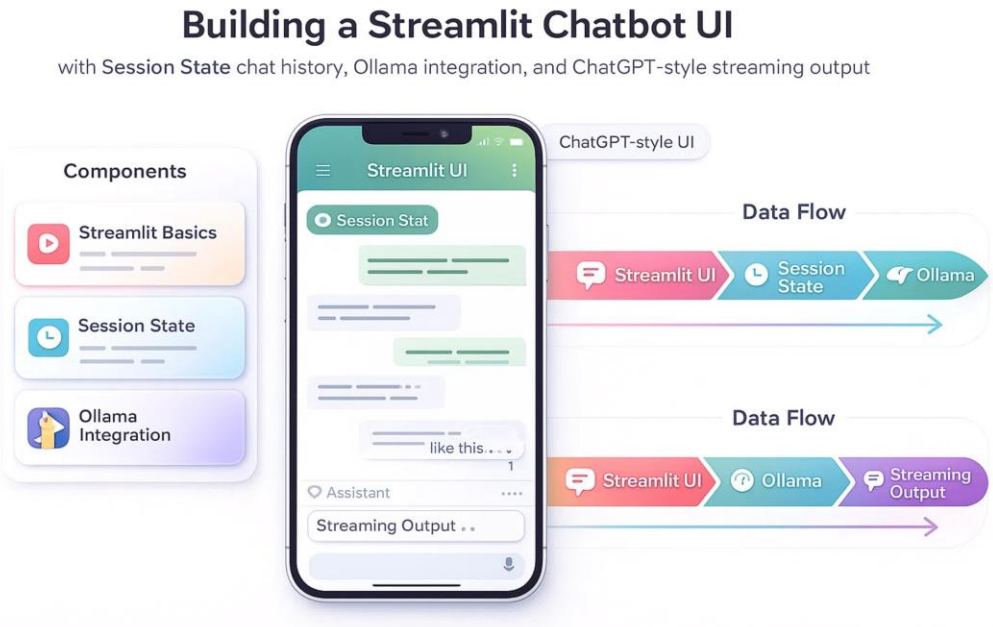


علیرضا اخوان‌پور

فصل سوم: رابط کاربری هوشمند با Streamlit

در این بخش، شما می‌آموزید که چطور یک رابط کاربری سریع و جذاب برای اپلیکیشن‌های هوش مصنوعی بسازید. آموزش مبانی Streamlit، پیاده‌سازی قابلیت حفظ تاریخچه مکالمات با Session State و یکپارچه‌سازی با مدل‌های Ollama از سرفصل‌های مهم هستند. تمرکز بر توسعه یک Chatbot ساده و شبیه‌سازی خروجی Streaming شبیه به ChatGPT است.

- مبانی Streamlit برای ساخت اپلیکیشن‌های هوش مصنوعی.
- ایجاد یک Chatbot ساده با قابلیت حفظ تاریخچه (Session State).
- یکپارچه‌سازی مدل‌های Ollama با رابط کاربری Streamlit.
- شبیه‌سازی رابط کاربری ChatGPT (Streaming Output).

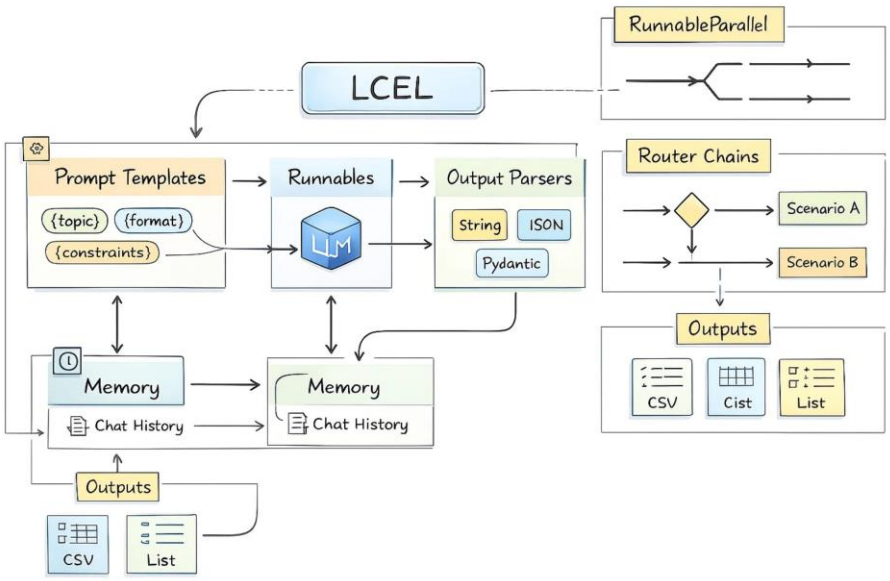


فصل چهارم: هسته مرکزی LangChain وLCEL

محتوای این فصل بر فریم‌ورک LangChain و هسته آن یعنی LCEL متمرکز است تا زنجیره‌های پیچیده‌ای از اجزای LLM بسازید. سرفصل‌ها شامل مدیریت ورودی با Prompt Templates و ساختاردهی خروجی با Output Parsers (تبدیل به JSON/Pydantic) است. هدف نهایی، پیاده‌سازی جریان‌های کاری پیشرفته (مانند مسیریابی) و استفاده از Memory برای تاریخچه مکالمات است.

علیرضا اخوان‌پور

- **مفاهیم پایه:**
  - مفهوم LCEL (LangChain Expression Language) و اهمیت آن.
  - ساخت اولین Chain با استفاده از Runnables.
- **مدیریت ورودی و خروجی:**
  - استفاده حرفه‌ای از Prompt Templates(مدیریت متغیرها)
  - کار با **Output Parsers**:
    - تبدیل خروجی بهString.
    - تبدیل خروجی به JSON و ساختارهای Pydantic (Structured Output).
    - مدیریت خروجی‌های CSV و لیست‌ها.
- **جریان‌های کاری پیشرفته:**
  - اجرای Chainها به صورت موازی (RunnableParallel).
  - شرطی‌سازی و مسیریابیChainها (Router Chains) برای سناریوهای مختلف.
- **حافظه(Memory) :**
  - پیاده‌سازی Chat History در LangChain.
  - ذخیره و بازیابی تاریخچه مکالمات.

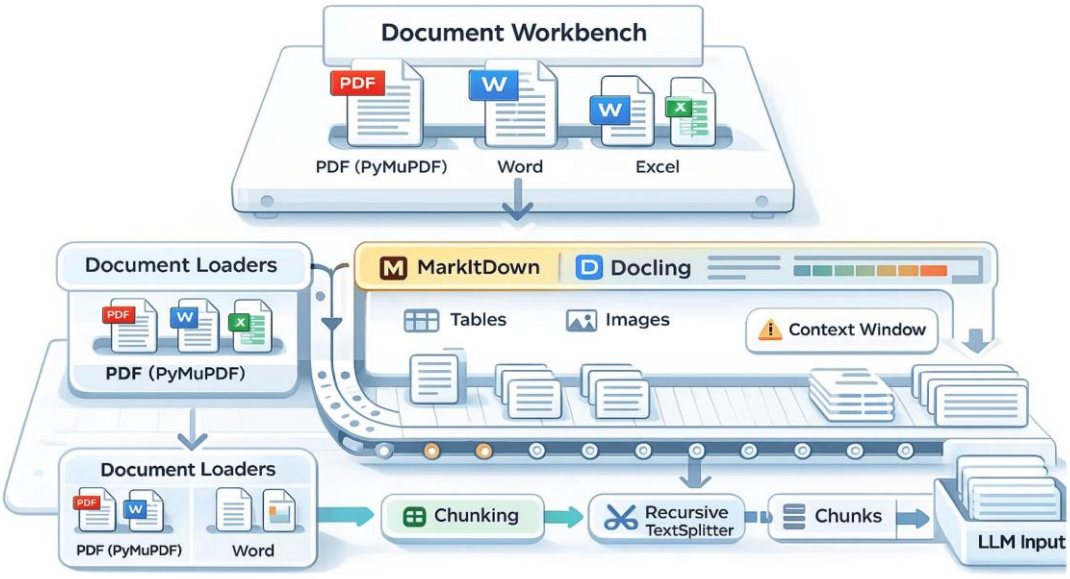


علیرضا اخوان‌پور

فصل پنجم: پردازش اسناد و داده‌ها (Data Ingestion)

در فصل پیش رو، یاد می‌گیرید که چگونه اسناد و داده‌های ساختارنیافته (مانند PDF) را برای LLM آماده کنید. این شامل استفاده از Document Loaderها و ابزارهایی چون Docling برای استخراج داده‌هاست. تمرکز بر رفع محدودیت Context Window از طریق فرآیند Chunking و استفاده از روش‌های مؤثر مانند RecursiveTextSplitter است.

- **کار با Document Loaderها:**
  - بارگذاری فایل‌های PDF (PyMuPDF)، Word و Excel.
  - استفاده از **MarkItDown** و **Docling** برای تبدیل اسناد پیچیده به فرمت قابل فهم برای مدل.
  - استخراج جداول و تصاویر از اسناد.
- **استراتژی‌های بخش‌بندی متن (Chunking):**
  - محدودیت Context Window و جلوگیری از فراموشی مدل.
  - روش‌های RecursiveTextSplitter.

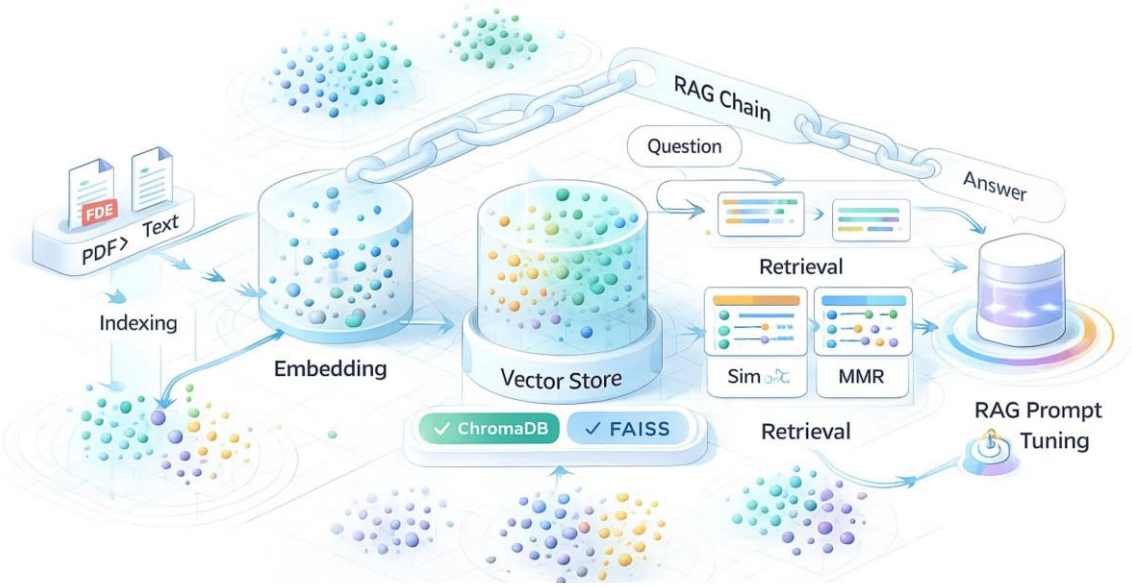


علیرضا اخوان‌پور

فصل ششم: سیستم‌های RAG (بازیابی اطلاعات و تولید متن)

این بخش به طور کامل به سیستم‌های (Retrieval-Augmented Generation) RAG و نحوه غنی‌سازی پاسخ‌های LLM با دانش خارجی می‌پردازد. سرفصل‌ها شامل درک Embedding، کار با Vector Stores (مانند ChromaDB و FAISS) و پیاده‌سازی کامل چرخه RAG است. تمرکز اصلی بر ساخت Chain نهایی RAG و تکنیک‌های بهینه‌سازی آن است.

- مفاهیم Embedding و Vector Stores.
- کار با پایگاه داده‌های برداری (ChromaDB و FAISS).
- پیاده‌سازی کامل چرخه RAG:
  - ایندکس کردن اسناد.
  - جستجوی تشابه (Similarity Search vs MMR).
  - ساخت Chain نهایی برای پاسخگویی به سوالات از روی اسناد (PDF ,Text).
- بهینه‌سازی RAG و تکنیک‌های پیشرفته (RAG Prompt Tuning).

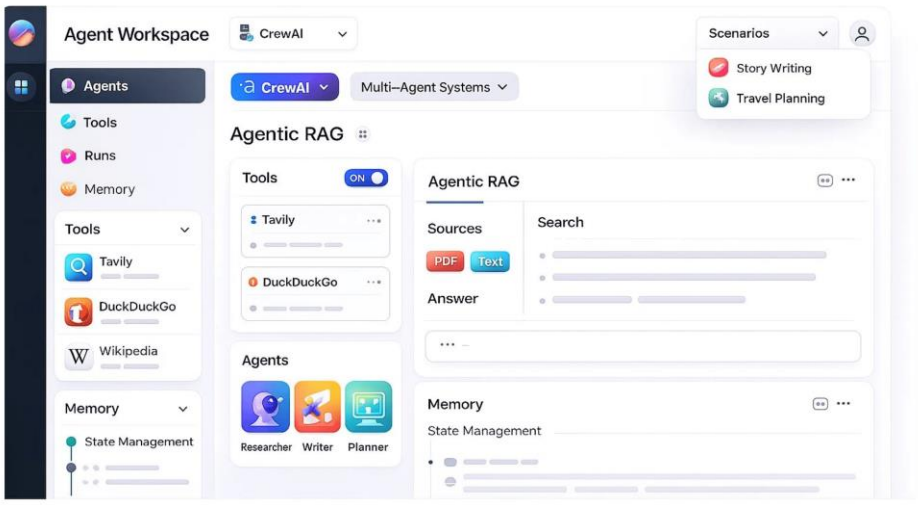


علیرضا اخوان‌پور

فصل هفتم: ایجنت‌های هوشمند (AI Agents) و ابزارها

در این فصل، شما از Chain به سمت ایجنت‌های تصمیم‌گیرنده که می‌توانند از ابزارها (Tools) استفاده کنند، حرکت می‌کنید. محتوا شامل مفهوم Tool Calling، استفاده از ابزارهای جستجوی آماده و ساخت ابزارهای سفارشی است. هدف این بخش، ساخت ایجنت‌های پیشرفته در LangChain و معرفی CrewAI برای ایجاد سیستم‌های چند ایجنتی (Multi-Agent Systems) است.

- **Tool Calling:**
  - آشنایی با مفهوم اتصال ابزار به LLM.
  - استفاده از ابزارهای آماده (Search Tools: Tavily ,DuckDuckGo ,Wikipedia).
  - ساخت ابزارهای سفارشی (Custom Tools) با پایتون.
- **ساختAgent :**
  - تفاوت Chain و Agent.
  - ساخت ایجنت‌های تصمیم‌گیرنده با LangChain.
  - مدیریت وضعیت ایجنت (State Management).
- **پروژه پیشرفتهAgentic RAG :**
  - ترکیب RAG با قابلیت جستجو و تصمیم‌گیری خودکار.
- **مقدمه‌ای برCrewAI :**
  - ساخت سیستم‌های چند ایجنتی (Multi-Agent Systems) برای سناریوهای پیچیده (مانند نوشتن داستان یا برنامه‌ریزی سفر).

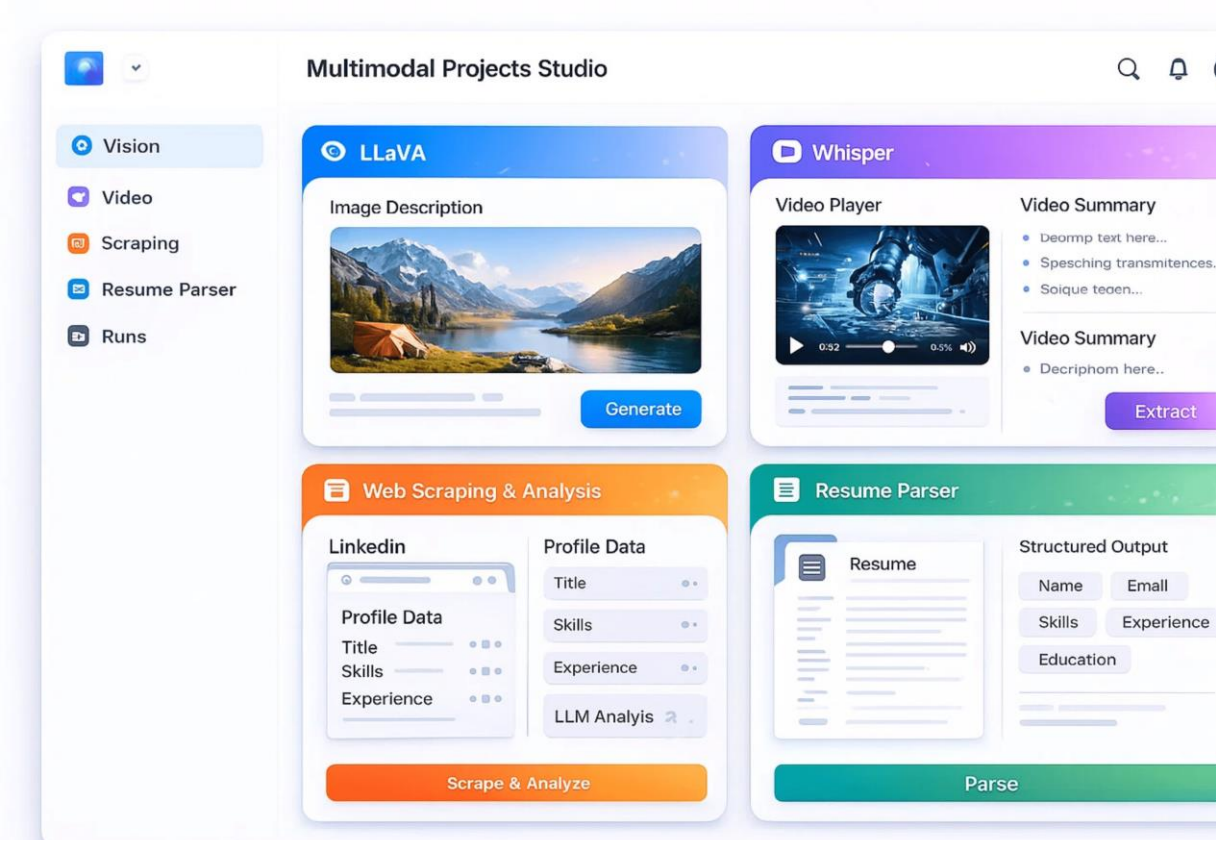


علیرضا اخوان‌پور

فصل هشتم: پروژه‌های کاربردی و چندرسانه‌ای

فصل پیش رو دانش شما را در قالب پروژه‌های عملی و چندرسانه‌ای (Multimodal) به کار می‌گیرد. سرفصل‌ها شامل توصیف تصاویر با مدل‌های Vision (مانند LLaVA) و پردازش ویدیو با Whisper است. تمرکز این بخش بر پروژه‌های تخصصی مانند Web Scraping & Analysis و ساخت یک Resume Parser برای تحلیل خودکار اطلاعات است.

- **پروژه‌های چندرسانه‌ای (Multimodal) :**
  - توصیف تصاویر با مدل‌های Vision (مانندLLaVA).
  - پردازش ویدیو و تبدیل گفتار به نوشتار (Whisper) برای خلاصه‌سازی ویدیو.
- **پروژه‌های تخصصی:**
  - **Web Scraping & Analysis:** استخراج داده از وب (مانند لینکدین) و تحلیل پروفایل‌ها یا رزومه‌ها با کمکLLM .
  - **Resume Parser:** ساخت سیستم هوشمند تحلیل و استخراج اطلاعات از رزومه.

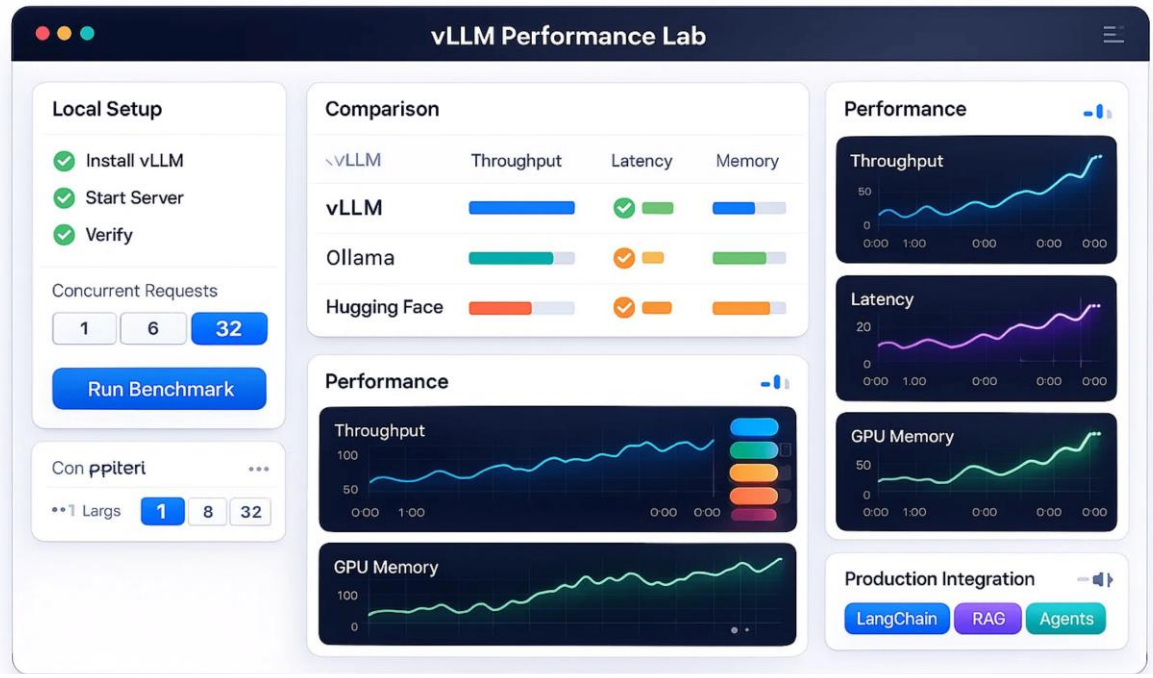


علیرضا اخوان‌پور

فصل نهم: بهینه‌سازی و اجرای سریع مدل‌ها (vLLM)- (تدریس در صورت صلاحدید مدرس دوره با توجه به سطح علمی دانشجویان)

محتوای این فصل به طور کامل به بهینه‌سازی و افزایش سرعت اجرای مدل‌ها اختصاص دارد. معرفی vLLM و کاربرد آن در سرعت‌بخشی و کاهش مصرف حافظه، هسته اصلی این بخش است. تمرکز بر مقایسه عملکرد vLLM در اجرای همزمان درخواست‌ها و ادغام آن با LangChain برای بهینه‌سازی RAG و Agentها در محیط پروداکشن است.

- معرفی vLLM و کاربرد آن در سرعت‌بخشی و کاهش مصرف حافظه.
- آموزش نصب و راه‌اندازی vLLM در محیط محلی.
- مقایسه vLLM با اجرای استاندارد Ollama و مدل‌های Hugging Face .
- نمونه عملی: اجرای چند درخواست همزمان با vLLM و بررسی Performance.
- ادغام vLLM با LangChain و استفاده در RAG و Agentها.



سوالات متداول دوره ساخت چت بات کاملا آفلاین

۱. در این دوره واقعا با API سیستم‌های Cloud کار نخواهیم کرد؟ یعنی بدون اینترنت و کاملا آفلاین؟!

بله، نام دوره به صراحت "ساخت چت بات فارسی کاملا آفلاین" است. تمرکز این دوره بر "نصب و راه‌اندازی مدل‌های محلی" است تا بتوان مدل‌های LLM را به صورت محلی راه‌اندازی کرد. در فصل دوم، ابزاری مانند Ollama برای اجرای مدل‌های محلی معرفی و نصب می‌شود. آشنایی با Hugging Face Hub و استفاده از API های رایگان آن نیز ذکر شده است.

۲. LLM چیست و از کدام مدل‌های زبان در این دوره استفاده خواهیم کرد؟

LLM مخفف "مدل‌های زبان بزرگ" است. این مدل‌ها و مفاهیم پایه‌ای مانند توکن و امبدینگ معرفی می‌شوند.

مدل‌های معروفی که در این دوره اجرای آن‌ها آموزش داده می‌شود شامل Llama ۳، DeepSeek، Gemma و Qwen هستند.

۳. RAG چیست و چه کاربردی دارد؟

RAG مخفف "Retrieval-Augmented Generation" به معنای بازیابی اطلاعات و تولید متن است.
کاربرد سیستم‌های RAG این است که پاسخ‌های مدل‌های زبان بزرگ (LLM) را با استفاده از دانش خارجی غنی‌سازی کنند. در این دوره نحوه پیاده‌سازی کامل چرخه RAG برای پاسخگویی به سوالات از روی اسناد مانند(PDF و Text) آموزش داده می‌شود.

۴. آیا می‌توانیم با محتوای PDF یا Word فارسی سامانه چت بات آفلاین ایجاد کنیم؟

بله، این دوره با هدف ساخت "چت بات فارسی" ارائه شده است. در این دوره آموزش داده می‌شود که چگونه اسناد و داده‌های ساختارنیافته را برای LLM آماده کنید. این شامل کار با Document Loaderها برای بارگذاری فایل‌هایی مانند PDF با استفاده از ((PyMuPDF) وWord) است. در نهایت، همچنین، ساخت Chain نهایی RAG برای پاسخگویی به سوالات از روی اسناد PDF و Text ذکر شده است.

۵. در این دوره هاست کردن مدل‌های زبانی بزرگ (LLM) آموزش داده خواهد شد؟

تمرکز اصلی دوره بر راه‌اندازی مدل‌ها به صورت محلی است به موضوع بهینه‌سازی و اجرای سریع مدل‌ها (vLLM) اختصاص دارد. معرفی vLLM و کاربرد آن در سرعت‌بخشی و کاهش مصرف حافظه است و ادغام vLLMبا LangChain و استفاده از آن در RAG و Agent ها در محیط پروداکشن (Production

environment) را پوشش می‌دهد. استفاده از vLLM برای اجرای بهینه در محیط پروداکشن معمولاً مقدمه‌ای بر سرویس‌دهی یا هاست کردن مدل‌ها محسوب می‌شود.

۶. چرا در این دوره به جای زنجیره‌های ساده، از (LCEL) زبان بیانی لنگ‌چین استفاده می‌شود؟

LCEL به جای زنجیره‌های ساده استفاده می‌شود زیرا امکان ساخت جریان‌های کاری پیچیده و Modular را با ترکیب آسان مؤلفه‌ها فراهم می‌کند. این زبان بیانی، مدیریت ورودی‌ها با تمپلیت‌های پویا، پردازش خروجی به فرمت‌های ساختاریافته مانند JSON و اجرای موازی یا شرطی تسک‌ها را تسهیل می‌کند. همچنین، یکپارچه‌سازی قابلیت‌هایی مانند حافظه مکالمات و ابزارها را پشتیبانی کرده و توسعه برنامه‌های مقیاس‌پذیر و قابل نگهداری را ممکن می‌سازد.

۷. تفاوت جستجوی Similarity ساده با روش MMR در سیستم‌های RAG چیست؟

جستجوی Similarity ساده، اسناد را صرفاً بر اساس نزدیکی امبدینگ به پرسش کاربر بازیابی می‌کند که ممکن است به نتایج تکراری یا کم‌تنوع منجر شود. در مقابل، روش MMR (Maximal Marginal Relevance) علاوه بر شباهت، معیار تنوع را نیز لحاظ می‌کند تا اسناد بازیافتی، اطلاعات تکراری کمتری داشته و جنبه‌های مختلف پرسش را پوشش دهند. این امر کیفیت پاسخ‌های تولیدشده را با کاهش افزونگی و افزایش اطلاعات مفید بهبود می‌بخشد.

۸. چگونه خروجی مدل‌های زبانی را به فرمت‌های ساختاریافته مثل JSON محدود می‌کنیم؟

برای محدود کردن خروجی مدل به فرمت‌های ساختاریافته مانند JSON، از ابزارهایی مانند Output Parsers در LangChain استفاده می‌شود. این تبدیل‌کننده‌ها با تعریف قالب‌های از پیش تعیین‌شده (مثلاً با استفاده از Pydantic) و ترکیب آن‌ها با Prompt Templates، مدل را وادار می‌کنند تا داده‌ها را مطابق ساختار موردنظر تولید کند. سپس، خروجی مدل به‌صورت خودکار به JSON یا دیگر فرمت‌ها تبدیل و اعتبارسنجی می‌شود.

۹. ایجنت‌های هوشمند (Agents) چه تفاوتی با زنجیره‌های معمولی (Chains) دارند؟

زنجیره‌های معمولی دنباله‌ای ثابت از عملیات را اجرا می‌کنند، در حالی که ایجنت‌ها توانایی تصمیم‌گیری پویا و استفاده از ابزارهای خارجی (مانند موتورهای جستجو) را دارند. ایجنت بر اساس ورودی کاربر، مسیر مناسب را انتخاب کرده و می‌تواند چندین عمل را به‌صورت تکراری انجام دهد تا به نتیجه برسد. این انعطاف، ایجنت‌ها را برای وظایف پیچیده‌تر مانند تحلیل داده‌های پویا یا تعامل با محیط ایده‌آل می‌کند.

۱۰. معماری چند ایجنتی (Multi-Agent) در CrewAI چه کاربردی در پروژه‌های واقعی دارد؟

معماری چند ایجنتی در CrewAI با توزیع مسئولیت‌ها بین ایجنت‌های تخصصی، وظایف پیچیده را به بخش‌های کوچک‌تر تقسیم می‌کند. برای مثال، در پروژه نوشتن داستان، یک ایجنت طرح داستان را می‌ریزد، دیگری شخصیت‌ها را توسعه می‌دهد و ایجنتی دیگر متن نهایی را ویرایش می‌کند. این همکاری باعث افزایش کارایی، کاهش خطا و خروجی باکیفیت‌تر در سناریوهای واقعی مانند برنامه‌ریزی یا تحلیل می‌شود.

علیرضا اخوان‌پور

۱۱. مدل‌های Multimodal (چندرسانه‌ای) در محیط آفلاین چگونه پیاده‌سازی می‌شوند؟

پیاده‌سازی مدل‌های چندرسانه‌ای در محیط آفلاین با استفاده از ابزارهایی مانند Ollama و Hugging Face انجام می‌شود. برای پردازش تصویر، مدل‌هایی مانند LLaVA مستقیماً روی دستگاه نصب شده و تصاویر را تحلیل می‌کنند. برای ویدیو، از کتابخانه‌هایی مانند Whisper برای تبدیل گفتار به متن و استخراج داده‌ها استفاده می‌شود. این مدل‌ها به‌صورت محلی اجرا شده و نیازی به اتصال اینترنت ندارند.

۱۲. آیا امکان شخصی‌سازی رابط کاربری چت‌بات برای شبیه‌سازی تجربه ChatGPT وجود دارد؟

بله، با استفاده از فریم‌ورک Streamlit می‌توان رابط کاربری چت‌باتی مشابه ChatGPT ایجاد کرد. این رابط قابلیت نمایش تاریخچه مکالمات، پشتیبانی از حالت Streaming برای شبیه‌سازی تایپ واقع‌زمان و یکپارچه‌سازی با مدل‌های محلی (مانند Ollama) را دارد. همچنین، با مدیریت State sessions، مکالمات ذخیره شده و تجربه کاربری روان و جذاب فراهم می‌شود.

۱۳. آیا خرید اقساطی امکان‌پذیر است؟

بله، امکان خرید اقساطی با اسنپ پی فراهم شده است. برای اطلاعات بیشتر می‌توانید با مشاورین مجموعه در تماس باشید یا راهنمای **خرید اقساطی دوره آموزشی با اسنپ پی** را مطالعه بفرمایید.

۱۴. آیا فیلم دوره رکورد می گردد؟

بله، دسکتاپ و صدای مدرس رکورد خواهد شد و در پلیر اختصاصی اسپات پلیر به همراه کلید لایسنس ارائه خواهد شد. شما در سیستم عامل‌های ویندوز، اندروید، آیفون (سیب، اناردون)، مک بوک می‌توانید فیلم را مشاهده کنید.

۱۵. افرادی که بصورت لایو کلاس را مشاهده می‌کنند، آیا امکان پرسش و پاسخ دارند؟

شما بصورت چت آنلاین می توانید سوالات خود را بپرسید و مدرس هم سوالات شما را پاسخ خواهد داد. البته توجه داشته باشید که این پروسه پاسخگویی هر ۴۰ دقیقه یکبار خواهد بود تا مدرس رشته کلام از دستش خارج نشود. البته که گروه تلگرامی دوره در اختیار شما است و می توانید سوالات خود را آنجا هم مطرح کنید.

۱۶. آیا امکان برگزاری دوره‌های سازمانی وجود دارد؟

بله؛ در نیک آموز امکان برگزاری دوره‌های سازمانی به‌صورت تخصصی فراهم شده است. به منظور ثبت درخواست، کافی است اطلاعات خود و دوره سازمانی مدنظر را در **فرم درخواست آموزش سازمانی** ثبت کنید تا ما با شما تماس بگیریم.

علیرضا اخوان‌پور

۱۷.می‌خواهم مشاوره / تدریس خصوصی بگیرم؟

برای اینکه بتوانید در ارتباط با برنامه‌نویسی، مشاوره / تدریس خصوصی بگیرید، لطفاً **فرم درخواست مشاوره مدرسین** را تکمیل نمایید تا کارشناسان ما با شما تماس بگیرند.

۱۸.آیا می‌توانم مشاوره سازمانی برای پروژه دریافت کنم؟

شما می‌توانید با مراجعه به **فرم درخواست مشاوره تخصصی**، از متخصصان نیک آموز مشاوره دریافت کنید و با به‌کارگیری مهارت‌های تجربی تیم ما، در ارتباط با پروژه‌های تخصصی خود، راهنمایی دریافت کنید

۱۹. من باز هم سؤال دارم؛ امکان ارتباط با مشاوران نیک آموز وجود دارد؟

بله؛ شما می‌توانید از مشاوره‌های **نیک آموز** به‌عنوان راهنما در مسیر خود استفاده کنید. برای این منظور، لطفاً شماره خود را در فرم مشاوره صفحه دوره وارد کنید تا مشاوران نیک آموز با شما تماس بگیرند.

شماره تماس فروش: ۰۲۱۹۱۰۷۰۰۱۷ – داخلی ۱

شماره تماس پشتیبانی: ۰۲۱۹۱۰۷۰۰۱۷ – داخلی ۲

۲۰.نحوه پشتیبانی دوره به چه صورت است؟

رضایت شما از دوره آموزشی و کمک به رفع مشکلات احتمالی برای ما اهمیت زیادی دارد. به همین دلیل، یک گروه پشتیبانی در تلگرام ایجاد شده است تا شما بتوانید درصورت نیاز، مسائل خود را در این بستر مطرح کنید. تا حداکثر ۴۸ ساعت کاری پس از ثبت نام در دوره، با شما تماس گرفته می‌شود و فرآیند عضویت شما در گروه تلگرام نهایی خواهد شد. توجه شود که در آینده، سیستم تیکتینگ راه‌اندازی می‌شود و فرآیند پشتیبانی از گروه تلگرامی به آن جا منتقل خواهد شد.

ثبت نام آنلاین دوره: ساخت چت بات فارسی کاملا آفلاین