طرح پیشنهادی پیاده‌سازی دستیار هوشمند

تحلیل تصاویر پزشکی (AI Radiology Assistant)

در مراکز علوم پزشکی مشهد

شرکت نگین رایانه الماس خراسان

تابستان ۱۴۰۵

مقدمه

این طرح به منظور پیاده‌سازی **دستیار هوشمند تحلیل تصاویر پزشکی (Radiology AI Assistant)** در پکس مراکز درمانی وابسته به دانشگاه علوم پزشکی مشهد، در پاسخ به RFP **شماره ۱۰۰** (پروژه سوم) ارائه می‌شود.

هدف اصلی، ایجاد زیرساختی است که امکان **تولید خودکار گزارش اولیه (AI Draft Report)** از تصاویر پزشکی (رادیولوژی، سی‌تی اسکن، MRI) را با استفاده از مدل‌های هوش مصنوعی پیشرفته فراهم کند. گزارش تولیدشده صرفاً به‌عنوان **دستیار هوشمند** جهت بازبینی، اصلاح و تأیید نهایی توسط پزشک متخصص رادیولوژی ارائه می‌شود.

در این طرح، سه رویکرد پیاده‌سازی پیشنهاد می‌شود که در ادامه به تفصیل مورد بررسی قرار می‌گیرند.

> ⚠️ تمامی ارقام هزینه‌ها در این طرح به صورت **برآورد اولیه و تقریبی** ارائه شده‌اند. طرح نهایی در زمان اجرا و بر اساس نرخ‌های به‌روز، نیازمند برآورد هزینه مجدد خواهد بود.

فهرست مطالب

راهکار اول: مدل‌های آنلاین (Cloud API)

در این راهکار، یک سرور واسط راه‌اندازی می‌شود که وظیفه ارتباط بین پکس مراکز درمانی و سرویس‌های هوش مصنوعی آنلاین را بر عهده دارد. این سرور به API مدل‌های مختلف هوش مصنوعی متصل می‌شود و درخواست‌های کاربران را از طریق این APIها ارسال می‌کند.
این سرور باید به اینترنت به صورت مستقیم یا در صورت نیاز از طریق VPN یا پراکسی دسترسی داشته باشد. پرداخت به ازای هر تصویر بوده و پیاده‌سازی آن بسیار سریع است.

> این روش با بند «عدم خروج داده از ایران» در RFP مغایرت دارد و صرفاً به‌عنوان **گزینه کوتاه‌مدت با رعایت پروتکل‌های ناشناس‌سازی داده** پیشنهاد می‌شود.

مدل‌های پیشنهادی (آنلاین)

مدل‌های تجاری (پیشرو) - مناسب برای تحلیل تصاویر پزشکی:

مدل	شرکت	ویژگی‌ها
GPT-4o / o1-series	OpenAI	بهترین کیفیت برای تفسیر تصاویر پیچیده و تولید گزارش ساختاریافته
Claude 3.5/4 Sonnet	Anthropic	قدرت بالا در درک بافت و محتوای تصویر و تولید متن بالینی
Gemini 3.5 Pro	Google	تعادل عالی بین هزینه و کیفیت، پشتیبانی از DICOM

ساختار فنی (آنلاین)

سرور مجازی (در شبکه داخلی دانشگاه):

- ۱۶GB RAM
- ۵۰GB فضای ذخیره‌سازی SSD
- ۴ هسته پردازنده
- سیستم عامل: Ubuntu 26.04 LTS

نرم‌افزار واسط (با رعایت الزامات RFP):

- دریافت تصاویر از سیستم PACS با پروتکل DICOM
- ناشناس‌سازی (Anonymization) داده‌های بیمار قبل از ارسال
- ارسال تصاویر به APIهای مدل‌های هوش مصنوعی
- دریافت گزارش اولیه (Findings + Impression) در قالب ساختار استاندارد

سیستم اعتباری:

- امکان مدیریت اعتبار توسط ادمین هر مرکز
- نمایش اعتبار باقیمانده به کاربران
- محاسبه و کسر هزینه هر درخواست بر اساس مدل انتخاب شده

هزینه‌های راهکار اول

هزینه‌های راه‌اندازی:

- راه‌اندازی سرور واسط و پیاده‌سازی نرم‌افزار: **۱ میلیارد تومان**
- اتصال هر مرکز درمانی به سرور واسط: **۱۰۰ میلیون تومان** به ازای هر مرکز

هزینه‌های جاری (ماهیانه - از سال دوم):

- هزینه پشتیبانی ماهانه: **۶۰ میلیون تومان**
- هزینه اینترنت و برق: **۱۰ میلیون تومان**

هزینه به ازای هر درخواست (میانگین ۳ تصویر، نرخ دلار ۲۰۰,۰۰۰ تومان):

مدل	نوع	قیمت (تومان)	مناسب برای
Gemini 3.5 Flash	تجاری (Google)	**۵,۰۰۰ - ۷,۰۰۰**	حجم بالا و هزینه کم
GPT-4o (OpenAI)	تجاری	**۱۴,۰۰۰ - ۲۵,۰۰۰**	کیفیت بالا
Claude Sonnet (Anthropic)	تجاری	**۱۶,۰۰۰ - ۳۰,۰۰۰**	بهترین کیفیت

مزایا و معایب (آنلاین)

✅ **مزایا:**

- تنوع مدل‌ها و امکان انتخاب بهترین مدل برای هر نوع تصویر
- به‌روزرسانی خودکار مدل‌ها توسط شرکت‌های ارائه‌دهنده
- عدم نیاز به نگهداری سخت‌افزار پیچیده
- پیاده‌سازی سریع (۱ تا ۲ ماه)
- دسترسی به پیشرفته‌ترین مدل‌های دنیا

❌ **معایب:**

- **مغایرت با بند عدم خروج داده از ایران** (در صورت عدم ناشناس‌سازی کامل)
- وابستگی به اینترنت پایدار بین‌الملل
- هزینه‌های متغیر و وابسته به نرخ ارز
- ریسک تحریم و قطع دسترسی
- هزینه تصاعدی با افزایش تعداد درخواست‌ها

راهکار دوم: مدل محلی روی سرور اختصاصی

در این راهکار، یک مدل هوش مصنوعی قدرتمند روی سرور اختصاصی در داخل شبکه دانشگاه علوم پزشکی نصب و راه‌اندازی می‌شود. تمام پردازش‌ها روی این سرور انجام می‌شود و **هیچ داده‌ای از شبکه داخلی خارج نمی‌شود** (تطابق کامل با RFP).

مدل‌های پیشنهادی (محلی) – مناسب برای تحلیل تصاویر پزشکی

مدل	ویژگی‌ها	مناسب برای
Llama 4 series 70B+ (Meta)	مدل چندوجهی با توانایی رقابت با مدل‌های تجاری	تحلیل تصاویر عمومی
LLaVA-Med 34B	مدل تخصصی پزشکی با عملکرد بالا	تصاویر رادیولوژی و MRI
OpenBioLLM-70B	مدل اختصاصی حوزه زیست‌پزشکی	تولید گزارش‌های تخصصی
MedGemma 27B (Google)	مدل سبک‌تر با کارایی مناسب	استقرار با منابع محدودتر

توجه: مدل‌های فوق قابلیت **فاین‌تیون روی داده‌های بومی دانشگاه** را دارند.

ساختار فنی (محلی)

سرور اختصاصی - مشخصات پیشنهادی:

- ۲ عدد GPU NVIDIA A100 (40GB) یا ۴×RTX 4090 با ۲۴GB VRAM
- ۲۵۶GB RAM
- حداقل ۲TB فضای ذخیره‌سازی NVMe SSD
- ۳۲ هسته پردازنده (Intel Xeon یا AMD EPYC)

نرم‌افزار مورد نیاز:

- سیستم‌عامل: Ubuntu Server 26.04 LTS
- نرم‌افزارهای پایه: CUDA 12.x, cuDNN, PyTorch, TensorFlow
- Docker و Kubernetes برای مدیریت کانتینرها
- سیستم اعتباری مشابه راهکار اول

هزینه‌ها (محلی)

هزینه‌های راه‌اندازی:

- خرید سرور اختصاصی با مشخصات فوق: حدود **۵ میلیارد تومان** (با توجه به نوسان قیمت ارز)
- پیاده‌سازی نرم‌افزاری: **۵۰۰ میلیون تومان**
- نصب و راه‌اندازی هر مدل: **۱۰۰ میلیون تومان** برای هر مدل
- اتصال هر مرکز درمانی: **۱۰۰ میلیون تومان**

هزینه‌های جاری (ماهیانه):

- هزینه نگهداری و پشتیبانی: **۵۰ میلیون تومان** (شامل نگهداری سرور، به‌روزرسانی مدل‌ها، و پشتیبانی فنی)
- هزینه مصرف برق و خنک‌سازی: **۱۵ میلیون تومان**

هزینه به ازای هر درخواست:

- با فرض ۵۰,۰۰۰ درخواست ماهانه، هزینه هر درخواست حدود **۱۰۰۰ تومان** برآورد می‌شود.

مزایا و معایب (محلی)

✅ **مزایا:**

- **حریم خصوصی کامل:** تمام داده‌ها در شبکه داخلی پردازش می‌شوند
- **عدم وابستگی به اینترنت بین‌الملل:** سیستم حتی در صورت قطع اینترنت بین‌الملل کار می‌کند
- **هزینه پایین‌تر در بلندمدت:** با افزایش تعداد درخواست‌ها، هزینه هر درخواست کاهش می‌یابد
- **عدم تأثیرپذیری از تحریم‌ها**
- **امکان فاین‌تیون با داده‌های بومی دانشگاه** برای بهبود عملکرد

❌ **معایب:**

- **هزینه اولیه بالا:** سرمایه‌گذاری اولیه قابل توجهی برای خرید سخت‌افزار نیاز است
- **نیاز به تخصص فنی بیشتر:** نگهداری و مدیریت این سیستم به تخصص فنی بالاتری نیاز دارد
- **محدودیت در به‌روزرسانی:** به‌روزرسانی مدل‌ها به صورت دستی و با تأخیر انجام می‌شود
- **محدودیت در مقیاس‌پذیری:** افزایش ظرفیت نیازمند خرید سخت‌افزار جدید است

راهکار سوم: توسعه و آموزش مدل هوش مصنوعی اختصاصی

در این راهکار، با همکاری دانشگاه علوم پزشکی، **یک مدل هوش مصنوعی اختصاصی** برای تحلیل تصاویر پزشکی و تولید گزارش اولیه توسعه و آموزش داده می‌شود. این مدل با استفاده از **داده‌های پزشکی محلی** (تصاویر رادیولوژی، سی‌تی اسکن، MRI با برچسب‌گذاری توسط رادیولوژیست‌های دانشگاه) آموزش داده می‌شود.

مزیت کلیدی: این روش امکان **سفارشی‌سازی کامل** برای بیماری‌های شایع منطقه و **شفافیت کامل (XAI)** را فراهم می‌کند.

ساختار فنی (اختصاصی)

جمع‌آوری داده‌ها:

- جمع‌آوری تصاویر پزشکی از مراکز درمانی وابسته به دانشگاه
- برچسب‌گذاری تصاویر توسط پزشکان متخصص رادیولوژی (حداقل ۱۰,۰۰۰ تصویر)
- ایجاد دیتاست استاندارد و با کیفیت برای آموزش مدل
- رعایت کامل پروتکل‌های محرمانگی و ناشناس‌سازی

سرور اختصاصی برای آموزش و اجرا:

- ۵۱۲GB RAM
- کارت گرافیک: ۸ عدد NVIDIA A100 (80GB) یا ۱۶×RTX 4090
- فضای ذخیره‌سازی: ۵TB NVMe SSD

نرم‌افزار مورد نیاز:

- سیستم عامل و نرم‌افزارها مشابه راهکار دوم

📊 **Explainable AI (XAI):**

- پیاده‌سازی تکنیک‌های **Grad-CAM** و **SHAP** برای نمایش مناطق مؤثر در تصمیم‌گیری مدل
- ارائه دلیل تصمیم به پزشک به‌صورت گرافیکی و متنی

هزینه‌های راهکار سوم

هزینه‌های اولیه:

- خرید سرور آموزش قدرتمند (اجاره یا خرید): ~**۲۰ میلیارد تومان**

> 💡 **راهکار جایگزین:** می‌توان سرور آموزش را به صورت **روزانه اجاره** کرد (مثلاً از خدمات ابری داخلی یا خارجی) که هزینه را به شدت کاهش می‌دهد و فقط در بازه‌های نیاز (مثلاً ۲-۳ ماه برای فاین‌تیون اولیه) پرداخت می‌شود.

- جمع‌آوری و برچسب‌گذاری ۱۰,۰۰۰ تصویر توسط پزشکان متخصص: ~**۳۰۰ میلیون تومان**
- توسعه، فاین‌تیون و اعتبارسنجی مدل (۶ ماه): ~**۵۰۰ میلیون تومان**
- پیاده‌سازی نرم‌افزاری و اتصال به پکس: ~**۲۵۰ میلیون تومان**
- **مجموع هزینه اولیه (بدون احتساب اتصال مراکز):** ~**۲۱ میلیارد تومان**

هزینه‌های جاری (ماهیانه):

- نگهداری و پشتیبانی از سرورها: **۵۰ میلیون تومان**
- برق و خنک‌سازی: **۱۵ میلیون تومان**
- بازآموزی دوره‌ای مدل با داده‌های جدید: **۵۰ میلیون تومان (متوسط)**
- **مجموع هزینه ماهانه:** حدود **۱۰۰ میلیون تومان**
- **هزینه هر درخواست (۵۰,۰۰۰ در ماه):** ~**۲,۰۰۰ تومان**

مزایا و معایب (اختصاصی)

✅ **مزایا:**

- **سفارشی‌سازی کامل:** مدل برای بیماری‌های شایع منطقه و نوع داده‌های بیمارستان بهینه می‌شود
- **استقلال کامل:** وابستگی به شرکت‌های خارجی و تحریم‌ها به صفر می‌رسد
- **ارزش علمی و تجاری بالا:** خروجی این پروژه می‌تواند به محصولی منحصربه‌فرد تبدیل شود
- **قابلیت بهبود مستمر:** مدل با داده‌های جدید قابل بازآموزی و بهبود است
- **امکان پیاده‌سازی XAI (شفافیت تصمیم‌گیری)**
- **کنترل کامل بر کیفیت و دقت مدل**

❌ **معایب:**

- **هزینه و زمان بسیار بالا:** نیازمند سرمایه‌گذاری کلان و زمان ۱ تا ۲ ساله
- **نیاز به تیم متخصص:** شامل دانشمند داده، مهندس ML و متخصصین پزشکی
- **کیفیت اولیه پایین‌تر:** مدل در ابتدا ممکن است عملکردی پایین‌تر از مدل‌های تجاری داشته باشد
- **محدودیت در مقیاس‌پذیری:** افزایش ظرفیت نیازمند سخت‌افزار اضافی است

مقایسه سه راهکار

از نظر هزینه

هزینه اولیه:

- راهکار اول (آنلاین): ۱ میلیارد تومان + ۱۰۰ میلیون × تعداد مراکز
- راهکار دوم (محلی): ۵.۶ میلیارد تومان + ۱۰۰ میلیون × تعداد مراکز
- راهکار سوم (اختصاصی): ۲۱ میلیارد تومان + ۱۰۰ میلیون × تعداد مراکز

هزینه به ازای هر درخواست:

- راهکار اول: از ۵,۰۰۰ تا ۳۰,۰۰۰ تومان
- راهکار دوم: حدود ۱۰۰۰ تومان (با فرض ۵۰,۰۰۰ درخواست ماهانه)
- راهکار سوم: حدود ۲۰۰۰ تومان (با فرض ۵۰,۰۰۰ درخواست ماهانه)

مقایسه از نظر کارایی

سرعت پاسخگویی: بستگی به سرعت اینترنت در روش آنلاین و قدرت سخت‌افزار در روش محلی دارد.

کیفیت پاسخ‌ها:

- **راهکار اول (آنلاین):** معمولاً کیفیت بالاتر به دلیل استفاده از مدل‌های پیشرفته‌تر
- **راهکار دوم (محلی):** کیفیت نسبتاً خوب، اما ممکن است در برخی موارد پایین‌تر باشد
- **راهکار سوم (اختصاصی):** کیفیت در ابتدا ممکن است پایین‌تر باشد، اما با آموزش مستمر می‌تواند به سطح مدل‌های تجاری برسد یا حتی در موارد خاص (بیماری‌های محلی) بهتر عمل کند

مقایسه از نظر امنیت و حریم خصوصی

محرمانگی داده‌ها:

- **راهکار اول (آنلاین):** داده‌ها به سرورهای خارجی ارسال می‌شوند
- **راهکار دوم (محلی):** تمام داده‌ها در شبکه داخلی باقی می‌مانند
- **راهکار سوم (اختصاصی):** کاملاً داخلی و امن، با کنترل کامل دانشگاه بر داده‌ها

ریسک تحریم:

- **راهکار اول (آنلاین):** احتمال محدودیت دسترسی به سرویس‌های خارجی
- **راهکار دوم (محلی):** پس از خرید اولیه، ریسک تحریم کاهش می‌یابد
- **راهکار سوم (اختصاصی):** کاملاً مستقل از تحریم‌ها

جدول خلاصه سه راهکار

معیار	راهکار ۱: آنلاین	راهکار ۲: محلی	راهکار ۳: اختصاصی
زمان پیاده‌سازی	کوتاه (۱-۲ ماه)	متوسط (۳-۶ ماه)	طولانی (۱-۲ سال)
هزینه راه‌اندازی (برای ده مرکز)	**۲ میلیارد تومان**	**۶.۶ میلیارد تومان**	**۲۲ میلیارد تومان**
هزینه عملیاتی سالانه	**۸۰۰ میلیون تومان**	**۸۰۰ میلیون تومان**	**۱,۴۰۰ میلیون تومان**
نیازمندی‌های فنی	اتصال اینترنت	سرور قدرتمند، متخصص DevOps	تیم توسعه هوش مصنوعی، سرور قدرتمند
مطابقت با RFP (عدم خروج داده)	✗	✓	✓
پشتیبانی از XAI	محدود	قابل پیاده‌سازی	کامل
قابلیت فاین‌تیون	✗	✓	✓
حریم خصوصی و امنیت	پایین	بالا	بسیار بالا
توصیه	راهکار کوتاه‌مدت	راهکار میان‌مدت	راهکار بلندمدت

⚠️ تمامی ارقام هزینه‌ها در این طرح به صورت **برآورد اولیه و تقریبی** ارائه شده‌اند. طرح نهایی در زمان اجرا و بر اساس نرخ‌های به‌روز، نیازمند برآورد هزینه مجدد خواهد بود.

توصیه نهایی (رویکرد سه‌مرحله‌ای)

با توجه به شرایط و نیازها، توصیه می‌شود پروژه در سه مرحله اجرا شود تا ریسک مالی به حداقل رسیده و در هر مرحله تصمیمات مبتنی بر داده گرفته شود:

مرحله اول (کوتاه‌مدت – ۶ ماه اول) – راهکار آنلاین با ناشناس‌سازی

اجرای راهکار اول برای شروع سریع‌تر و هزینه اولیه کمتر:

1. راه‌اندازی سرور واسط در شبکه داخلی دانشگاه
2. آموزش کاربران و جمع‌آوری بازخورد

مرحله دوم (میان‌مدت – سال اول و دوم) – استقرار مدل محلی

پیاده‌سازی راهکار دوم در صورت استقبال کاربران و افزایش تعداد درخواست‌ها:

1. خرید و راه‌اندازی سرور اختصاصی در داخل شبکه دانشگاه
2. انتقال تدریجی از مدل‌های آنلاین به مدل محلی

مرحله سوم (بلندمدت – سال دوم به بعد) – توسعه مدل اختصاصی

پیاده‌سازی راهکار سوم برای استقلال کامل و سفارشی‌سازی:

1. جمع‌آوری و برچسب‌گذاری داده‌های پزشکی محلی
2. آموزش مدل اختصاصی با استفاده از سرورهای قدرتمند

این رویکرد سه‌مرحله‌ای امکان شروع سریع پروژه با هزینه کمتر را فراهم می‌کند، در حالی که در میان‌مدت به کاهش وابستگی به سرویس‌های خارجی کمک می‌کند و در بلندمدت، دانشگاه را به استقلال کامل و توسعه یک مدل منحصربه‌فرد می‌رساند. اولویت‌بندی مرحله‌ها می‌تواند بر اساس بودجه، تعداد درخواست‌ها، و اهمیت حریم خصوصی تنظیم شود.

شرکت نگین رایانه الماس خراسان
https://spax.ir
https://neginalmas.ir
info@spax.ir