



راهنمای نصب هوش مصنوعی آفلاین گوگل روی موبایل و کامپیوتر

گوگل در دوم آوریل ۲۰۲۴ خانواده جدید مدل‌های هوش مصنوعی متن‌باز خود، جما ۴ (Gemma 4)، را با لایسنس کاملاً آزاد Apache ۲.۰ منتشر کرد نکته مهم این نسخه، تمرکز جدی روی اجرای کاملاً آفلاین روی گوشی موبایل و لپ‌تاپ است؛ یعنی بدون نیاز به اینترنت، بدون API ابری و بدون هیچ هزینه ماهانه. در این مقاله قدم‌به‌قدم توضیح می‌دهیم چطور می‌توانید هوش مصنوعی آفلاین گوگل را روی اندروید، آیفون و کامپیوتر خودتان نصب و اجرا کنید.

جما ۴ (Gemma 4) دقیقاً چیست؟

جما ۴ نسل جدیدی از مدل‌های هوش مصنوعی است که با یک هدف واضح طراحی شده: اجرای روان و بی‌نقص هوش مصنوعی روی هر دستگاهی که در اختیار دارید. هوش مصنوعی آفلاین گوگل، برای اینکه هم روی یک گوشی موبایل معمولی و هم

روی یک سرور قدرتمند کارایی داشته باشد، گوگل آن را در چهار نسخه کاملاً متفاوت و متناسب با سخت‌افزارهای مختلف عرضه کرده است:

نسخه‌های مختلف جما ۴ در یک نگاه

نسخه مدل ساختار پارامترها توان سخت‌افزاری مورد نیاز بهترین انتخاب برای... **EPB** ۵.۱ میلیارد کل (۲.۳ میلیارد فعال) رم ۶ تا ۸ گیگابایت گوشی‌های موبایل میان‌رده و استاندارد **E4B** ۸ میلیارد کل (۴.۵ میلیارد فعال) رم ۸ گیگابایت به بالا گوشی‌های موبایل پرچمدار و قدرتمند **MoE (Mixture of Experts)** معماری ترکیبی ۳۲ گیگابایت رم سیستم یا ۲۴ گیگابایت VRAM لپ‌تاپ‌های حرفه‌ای و سیستم‌های رومیزی **Dense 31B** معماری متراکم (Dense) پردازنده‌های گرافیکی (GPU) حرفه‌ای سرورها و کامپیوترهای پردازشی قدرتمند

حرف «E» به چه معناست؟

شاید بپرسید چرا مدلی که بیش از ۵ میلیارد پارامتر دارد، EPB نامیده می‌شود؟ حرف E در اینجا مخفف کلمه Effective (موثر) است و دقیقاً نقطه قوت نسخه‌های موبایلی جما ۴ را نشان می‌دهد.

برای درک بهتر، یک شرکت بزرگ را تصور کنید که ۵۱۰۰ کارمند (مجموع پارامترها) دارد؛ اما برای انجام یک وظیفه مشخص در لحظه، تنها ۲۳۰۰ نفر از ماهرترین افرادش (پارامترهای فعال) را درگیر کار می‌کند. این معماری هوشمندانه در هوش مصنوعی آفلاین گوگل، باعث می‌شود مدل، دانش کامل یک سیستم بزرگ را در اختیار داشته باشد؛ اما در زمان اجرا، فضای رم گوشی شما را اشغال نکند. به همین دلیل، مصرف حافظه و باتری در نسخه‌های EPB و E4B بسیار پایین و بهینه است.

درک همه‌جانبه و حافظه عمیق

نسخه‌های موبایلی جما ۴ صرفاً یک ربات چت متنی نیستند، بلکه چندوجهی طراحی شده‌اند. این یعنی هوش مصنوعی آفلاین گوگل می‌تواند جهان را شبیه به انسان درک کند؛ می‌تواند متن‌ها را بخواند، تصاویر را ببیند و صداها را بشنود و پردازش کند.

یکی دیگر از ویژگی‌های خیره‌کننده این مدل، پنجره محتوای ۱۲۸ هزار توکنی آن است. برای اینکه این عدد ملموس‌تر شود، می‌توان اینطور بیان کرد: «این مدل می‌تواند در یک مکالمه واحد، اطلاعاتی معادل یک کتاب ۳۰۰ صفحه‌ای را به طور کامل در حافظه کوتاه‌مدت خود نگه دارد و بدون فراموش کردن رشته کلام، بر اساس تمام آن اطلاعات، پاسخ‌های دقیق ارائه دهد».

پیش‌نیازها

قبل از شروع، این موارد را بررسی کنید:

برای نصب هوش مصنوعی آفلاین گوگل روی موبایل:

- اندروید ۱۰ به بالا، یا آیفون (ترجیحاً آیفون ۱۵ پرو یا جدیدتر برای عملکرد روان)
- حداقل ۴ تا ۶ گیگابایت فضای خالی (۲.۵ گیگابایت برای ۵ E۲B، ۵ گیگابایت برای E۴B، به علاوه فضای بافر)
- اتصال Wi-Fi فقط برای یکبار دانلود اولیه مدل
- ترجیحاً تراشه‌ای نسبتاً جدید؛ برای اندروید ۲ Gen Snapdragon ۸ به بالا، Google Tensor G۳ به بالا یا MediaTek Dimensity ۹۳۰۰

برای کامپیوتر:

- ویندوز، مک یا لینوکس
- برای مدل‌های سنگین‌تر (۲۶B و ۳۱B)، حداقل ۳۲ گیگابایت رم یا کارت گرافیک با ۲۴ گیگابایت VRAM؛ برای نسخه‌های سبک‌تر (E۲B/E۴B) رم ۸ تا ۱۶ گیگابایتی کفایت می‌کند
- نصب یکی از نرم‌افزارهای اجراکننده مدل (Ollama یا LM Studio)

نصب و اجرای آفلاین روی اندروید

برای اجرای هوش مصنوعی آفلاین گوگل روی گوشی، می‌توانید از اپلیکیشن رسمی، رایگان و متن‌باز گوگل به نام AI Edge Gallery استفاده کنید. برای راه‌اندازی، مراحل زیر را دنبال کنید:

۱. نصب اپلیکیشن

وارد گوگل‌پلی (Google Play) شوید، عبارت «AI Edge Gallery» را جست‌وجو کرده و اپلیکیشن ساخته‌شده توسط گوگل را نصب کنید.

۲. آشنایی با محیط اپلیکیشن و انتخاب ابزار

پس از باز کردن اپلیکیشن، حالت‌های مختلفی برای تعامل با هوش مصنوعی در اختیار شماست:

- AI Chat: مکالمه و چت متنی با مدل (برای شروع روی این گزینه ضربه بزنید)
- Ask Image: تحلیل تصاویر و پرسش‌وپاسخ درباره آن‌ها
- Audio Scribe: پردازش فایل‌های صوتی و تبدیل صوت به متن
- Agent Skills: استفاده از مهارت‌های دستیار هوشمند
- Prompt Lab: محیطی برای تست و نوشتن پرامپت‌های حرفه‌ای
- Mobile Actions: اجرای دستورات و اکشن‌های سیستمی روی موبایل

۳. انتخاب و دانلود مدل مناسب

پس از ورود به بخش AI Chat، به تب Models بروید. در این بخش دو نسخه از مدل برای دانلود وجود دارد که باید بر اساس توان گوشی خود یکی را انتخاب کنید:

- مدل E2B (با حجم حدود ۲.۵ گیگابایت): مناسب برای گوشی‌هایی با رم کمتر از ۸ گیگابایت (حداقل ۶ گیگابایت رم توصیه می‌شود).
- مدل E4B (با حجم حدود ۵ گیگابایت): مناسب برای گوشی‌های قدرتمندتر با رم ۸ گیگابایت و بیشتر.

۴. تست عملکرد آفلاین

پس از پایان دانلود مدل انتخابی، اینترنت گوشی را خاموش کرده و دستگاه را روی حالت هواپیما (Airplane Mode) قرار دهید. حالا یک پیام در محیط چت بفرستید؛ اگر پاسخ را دریافت کردید، یعنی هوش مصنوعی آفلاین گوگل با موفقیت روی سخت‌افزار گوشی شما مستقر شده و کاملاً آفلاین کار می‌کند.

نصب و اجرای آفلاین روی آیفون (iOS)

روند نصب و راه‌اندازی روی سیستم عامل iOS نیز بسیار ساده و تقریباً مشابه اندروید است:

۱. دریافت اپلیکیشن از اپ‌استور

وارد اپ‌استور (App Store) شوید، عبارت «AI Edge Gallery» را جست‌وجو کرده و اپلیکیشن رسمی ساخته‌شده توسط گوگل را دانلود و نصب کنید.

۲. انتخاب و دانلود مدل

اپلیکیشن را باز کرده و به تب Models بروید. در اینجا می‌توانید بسته به توان گوشی خود، یکی از دو مدل E۲B یا E۴B را انتخاب و دانلود کنید. پس از اتمام دانلود، تمام پردازش‌ها کاملاً محلی و با استفاده از قدرت پردازنده گرافیکی (GPU) و موتور عصبی (Neural Engine) خود آیفون انجام می‌شود.

۳. تجربه حالت تفکر (Thinking Mode)

پیشنهاد می‌کنیم در بخش AI Chat، قابلیت Thinking Mode را فعال و امتحان کنید. این ویژگی جذاب اجازه می‌دهد مراحل استدلال و نحوه رسیدن مدل به پاسخ نهایی را قدم‌به‌قدم مشاهده کنید.

نکته سخت‌افزاری

• **عملکرد ایده‌آل:** گوگل به‌طور رسمی آیفون ۱۵ پرو یا مدل‌های جدیدتر را برای اجرای پایدار و روان این مدل‌ها پیشنهاد می‌کند.

• **گوشی‌های قدیمی‌تر:** مدل‌های آیفون ۱۴ و قبل از آن نیز توانایی اجرای نسخه سبک‌تر (E۲B) را دارند؛ اما باید در نظر داشته باشید که سرعت پاسخ‌دهی کندتر خواهد بود و مصرف باتری نیز افزایش می‌یابد.

نصب و اجرا روی کامپیوتر (ویندوز، مک و لینوکس)

برای اجرای محلی هوش مصنوعی آفلاین گوگل روی کامپیوتر، روش‌های مختلفی وجود دارد؛ اما اینجا سراغ بهترین و بی‌دردسرت‌ترین روش می‌رویم:

روش اول: استفاده از Ollama (ساده‌ترین راه برای اکثر کاربران)

۱. دانلود و نصب نرم‌افزار

ابتدا به وبسایت ollama.com/download (https://ollama.com/download) مراجعه کرده و نرم‌افزار Ollama را متناسب با سیستم‌عامل خود (ویندوز، مک یا لینوکس) دانلود و نصب کنید.

۲. اجرای مدل از طریق ترمینال (خط فرمان)

ترمینال در مک/لینوکس یا محیط خط فرمان (Command Prompt / PowerShell) در ویندوز را باز کنید. حالا بسته به قدرت سخت‌افزاری سیستم، یکی از دستورات زیر را کپی و اجرا کنید:

• **مدل سبک:** `ollama run gemma۴:e۲b` (مناسب برای سیستم‌های معمولی با ۸ گیگابایت رم)

• **مدل استاندارد:** `ollama run gemma۴:e۴b` (انتخاب پیشنهادی برای لپ‌تاپ‌هایی با ۱۶ گیگابایت رم)

• **مدل پیشرفته:** `ollama run gemma۴:۲۶b` (کیفیت بسیار بالا؛ نیازمند ۳۲ گیگابایت رم یا ۲۴ گیگابایت حافظه گرافیکی VRAM)

• **مدل سنگین:** `ollama run gemma۴:۳۱b` (قدرتمندترین حالت ممکن؛ ویژه کامپیوترهای قدرتمند و سخت‌افزارهای سروری)

۳. دانلود خودکار و شروع گفت‌وگو

پس از وارد کردن دستور، سیستم شروع به دانلود مدل می‌کند. پس از اتمام دانلود، مدل به صورت خودکار اجرا می‌شود و می‌توانید مستقیماً در همان محیط ترمینال، گفت‌وگو با هوش مصنوعی را آغاز کنید.

API محلی سازگار با OpenAI

نرم‌افزار Ollama یک API محلی روی آدرس `http://localhost:11434` در اختیاران قرار می‌دهد که ساختار آن کاملاً با OpenAI سازگار است. این یعنی با کمک هوش مصنوعی آفلاین گوگل، می‌توانید تمام ابزارها، نرم‌افزارها و کدهایی که پیش‌تر برای اتصال به ChatGPT نوشته شده‌اند را بدون نیاز به تغییر کد، مستقیماً به مدل محلی خودتان وصل کنید!

روش دوم: LM Studio (بدون نیاز به خط فرمان)

اگر کار با ترمینال برایتان راحت نیست و می‌خواهید با هوش مصنوعی آفلاین گوگل کار کنید:

۱. از سایت lmstudio.ai نرم‌افزار را دانلود و نصب کنید.
۲. در بخش جست‌وجوی مدل‌ها عبارت «gemma۴» را وارد کنید.
۳. نسخه متناسب با سخت‌افزارتان (۲۶B، E۴B، E۲B یا ۳۱B) را دانلود کنید.
۴. از همان رابط گرافیکی نرم‌افزار شروع به چت کنید؛ تمام مراحل بدون نیاز به نوشتن دستور انجام می‌شود.

اگر روی گوشی کند اجرا شود، چه کار کنیم؟

اگر در زمان استفاده از مدل روی موبایل با کندی یا افت سرعت مواجه شدید، جای نگرانی نیست؛ معمولاً یکی از چهار دلیل زیر مقصر اصلی است که به راحتی می‌توانید آن‌ها را برطرف کنید:

۱. انتخاب مدل نامناسب با سخت‌افزار

اگر رم گوشی شما کمتر از ۸ گیگابایت است اما نسخه سنگین‌تر (E۴B) را نصب کرده‌اید، دستگاه دچار افت عملکرد می‌شود. در این حالت، به سادگی مدل قبلی را پاک کرده و نسخه سبک‌تر یعنی E۲B را دانلود و جایگزین کنید.

۲. اشغال بودن فضای رم (اپلیکیشن‌های پس‌زمینه)

پردازش هوش مصنوعی به حافظه موقت (RAM) زیادی نیاز دارد. قبل از شروع استفاده، سایر برنامه‌های باز در پس‌زمینه گوشی را ببندید تا تمام ظرفیت رم کاملاً در اختیار مدل قرار گیرد و روان‌تر کار کند.

۳. داغ شدن دستگاه

در جلسات طولانی‌مدت گفت‌وگو با هوش مصنوعی، گوشی گرم می‌شود و سیستم عامل برای جلوگیری از آسیب دیدن قطعات، سرعت پردازنده را به صورت خودکار کاهش می‌دهد. در این شرایط، کافی است کمی به گوشی استراحت دهید تا خنک شود و سرعت آن به حالت عادی برگردد.

۴. قدیمی بودن تراشه گوشی

اگر از گوشی‌های قدیمی‌تر از سال ۲۰۲۲ استفاده می‌کنید، پردازنده شما توان محاسباتی لازم برای این پردازش‌های سنگین را ندارد. در این حالت، بهترین راهکار این است به جای موبایل، از همان روش Ollama استفاده کرده و مدل را روی کامپیوتر یا لپ‌تاپ اجرا کنید.

با هوش مصنوعی آفلاین گوگل، چه کارهایی می‌توان انجام داد؟

گفت‌وگوی متنی با حالت تفکر

یک دستیار عالی برای طوفان فکری، پیش‌نویس کردن نامه‌ها و مقالات، حل مسائل منطقی و توضیح مفاهیم جدید.

تحلیل هوشمند تصویر

به کمک قدرت بینایی این هوش مصنوعی آفلاین گوگل، می‌توانید از محیط یا اسناد عکس بگیرید و درباره آن سوال بپرسید؛ از شناسایی یک گیاه ناشناخته گرفته تا استخراج متن یک نامه.

تبدیل گفتار به نوشتار

ترجمه دقیق صوت به بیش از ۱۴۰ زبان دنیا، که تمام مراحل پردازش آن کاملاً روی خود دستگاه شما انجام می‌شود.

دستیار کدنویسی حرفه‌ای روی لپ‌تاپ

با استفاده از نسخه قدرتمند ۲۶B در نرم‌افزار Ollama، یک دستیار برنامه‌نویسی اختصاصی خواهید داشت. پنجره متنی عظیم ۲۵۶ هزار توکنی در این هوش مصنوعی آفلاین گوگل، اجازه می‌دهد کل کدهای یک پروژه را یکجا به مدل بدهید تا آن را تحلیل و عیب‌یابی کند.

یک نکته صادقانه

در استدلال‌های بسیار پیچیده و چندمرحله‌ای، نسخه‌های موبایلی هوش مصنوعی آفلاین گوگل، هنوز از مدل‌های غول‌پیکر ابری (مثل Claude یا ChatGPT) عقب‌تر هستند؛ اما برای کارهای روزمره مانند نوشتن ایمیل، خلاصه‌سازی محتوا و پاسخ به سوالات متداول، عملکردی فوق‌العاده و کاملاً کافی دارند.

آینده پردازش‌های هوشمند؛ همه‌چیز در دستگاه شما

نسل جدید مدل‌های زبانی مانند جما ۴ نشان می‌دهند همیشه برای دسترسی به فناوری‌های پیشرفته، نیازی به اینترنت پرسرعت یا سرورهای گران‌قیمت نداریم.

استفاده از این هوش مصنوعی آفلاین گوگل نه تنها به معنای تضمین کامل حریم خصوصی و امنیت اطلاعات شخصی شماست؛ بلکه این اطمینان را می‌دهد که حتی در شرایط قطعی شبکه، سفرهای هوایی یا مناطق دورافتاده نیز یک دستیار هوشمند همه‌فن‌حریف در کنار شما حضور دارد.

اگرچه هنوز در ابتدای مسیر پردازش‌های محلی هستیم؛ اما نصب این مدل روی موبایل یا کامپیوتر شخصی، تجربه‌ای هیجان‌انگیز از آینده‌ای است که در آن قدرت مهارنشده‌ی هوش مصنوعی بدون هیچ واسطه‌ای، مستقیماً در دستان شما قرار می‌گیرد.

اگر به مباحث روز هوش مصنوعی و یادگیری جدیدترین ابزارها و فناوری‌ها علاقه دارید، پیشنهاد می‌کنیم مقاله «[راهنمای جامع کلود](#)» را نیز در وبسایت مدیرسبز مطالعه کنید. این مقاله، به جدیدترین و قوی‌ترین قابلیت‌های Claude می‌پردازد؛ قابلیت‌هایی که تجربه استفاده از هوش مصنوعی در زندگیتان را دگرگون می‌سازد.