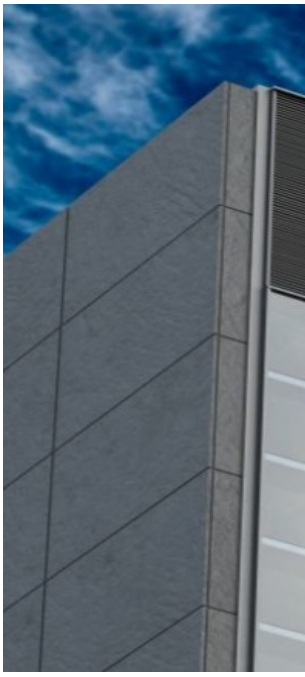
ميتا تطرح نماذج ذكية للأجهزة المحمولة محدودة القدرات



أعلنت شركة ميتا عن إصدارات جديدة من نماذج "لاما" للذكاء الاصطناعي، تحت اسم Llama 3.2 1B وLlama 3B، خصيصاً مصممة للعمل على الأجهزة محدودة القدرات.

تأتي هذه النماذج "مخفّفة" الحجم بفضل عملية تعرف باسم "التكميم Quantization"، التي تقلل من استهلاك الذاكرة، وتزيد من سرعة الأداء، ما يجعلها مناسبة للعمل على الهواتف الذكية، والأجهزة الطرفية.

أوضحت ميتا، في مدونتها، أن نموذج Llama 3.2 1B وLlama 3B بات متاحاً بقدرات محسّنة، إذ تم تطبيق تقنية "التكميم"، لتقليل دقة بعض مكونات النموذج، مما يجعله أخف وزناً على ذاكرة الأجهزة، ويقلل استهلاك الطاقة.

من خلال هذه التقنية، تستهدف ميتا التطبيقات القصيرة التي تتطلب معالجة حتى 8000 رمز نصي، وتعتمد الشركة على طريقتين رئيسيتين للتكميم، وهما: التكميم باستخدام "QLoRA"، والتي تُركز على دقة النتائج، ما يجعلها مثالية لتطبيقات تتطلب دقة عالية.

والطريقة الثانية تستخدم أسلوب "SpinQuant"، والمصمّم للمطورين الذين يرغبون في تقليل حجم النموذج دون التضحية الكبيرة بالدقة.

جرّبت ميتا نماذجها الجديدة على الهواتف الذكية من إصدارات OnePlus 12، وبيّنت النتائج أن النماذج الجديدة تمكنت من تقليل حجم الذاكرة المستخدمة بنسبة 41%، مع الحفاظ على أداء مقارب للنماذج الأصلية، وسرعة تصل إلى أربعة أضعاف في وقت التنفيذ.

بفضل التعاون مع شركات مثل "كوالكوم" و"ميدياتيك"، تم تحسين هذه النماذج لتعمل بسلاسة على معالجات الهواتف المحمولة، ما يتيح لمستخدمي هذه الأجهزة تجربة مخصّصة للذكاء الاصطناعي، تحافظ على خصوصية البيانات عبر التشغيل المحلي دون الحاجة للاتصال بالإنترنت.

تأتي هذه الإصدارات ضمن سلسلة من الإعلانات المتسارعة لميتا هذا الشهر، والتي شملت نماذج أخرى للذكاء الاصطناعي مثل "Gen Movie Meta" لتحرير مقاطع الفيديو، ونموذج "LM Spirit" لتوليد أصوات اصطناعية تعبر عن مشاعر متنوعة.

النماذج المخففة من "لاما" أصبحت متاحة الآن للتنزيل عبر مواقع com.Llama وFace Hugging، وتفتح آفاقاً جديدة أمام المطوّرين لابتكار تطبيقات ذكاء اصطناعي على الأجهزة المحمولة، مع زيادة كفاءة استهلاك الموارد.