شرح التوليد المعزّز بالاسترجاع (RAG)

النموذج اللغوي، وهو نوع الذكاء الاصطناعي وراء مساعدات المحادثة، يكتب إجاباته من أنماط تعلمها أثناء التدريب. وهو لا يعرف تلقائياً وثائق الشركة الخاصة، وتتوقف معرفته عند وقت التدريب. والتوليد المعزّز بالاسترجاع، أو RAG، يعالج ذلك بأن يترك الذكاء الاصطناعي يبحث عن الوثائق ذات الصلة قبل أن يجيب. وقد وصف الفكرة عام 2020 باتريك لويس وزملاؤه، وكانوا يعملون في Facebook AI Research ومؤسسات أخرى. ويمكن أن نسير في أربعة أجزاء: المشكلة التي يحلها، وكيف يعمل، وفوائده، وحدوده.

1. المشكلة التي يحلها RAG

يتعلم النموذج اللغوي من مجموعة كبيرة من النصوص أثناء التدريب، وبعد ذلك تصبح معرفته ثابتة. فقد تكون قديمة، ولا يمكنه معرفة مواد خاصة مثل سياسات الشركة. وعندما تنقصه الحقائق، يمكن أن يقدم مع ذلك إجابة واثقة لكنها خاطئة، وهذا ما يسميه الناس غالباً الهلوسة. ويعالج RAG ذلك بإعطاء النموذج المعلومات الصحيحة لحظة السؤال.

مثال: يعتمد الامتحان المغلق على الذاكرة وحدها، والذاكرة قد تخون. وبتطبيق المشكلة التي يحلها RAG، يتيح الامتحان المفتوح للطالب أن يراجع الحقائق قبل أن يجيب.

في العمل: يسأل سامر مساعد محادثة عن عدد أيام الإجازة التي تمنحها شركته. وبتطبيق المشكلة التي يحلها RAG، يلاحظ أن الإجابة تبدو واثقة لكنها عامة، لأن النموذج لم ير سياسة الشركة قط. فيقرر أن المساعد يحتاج إلى الوصول إلى وثيقة السياسة.

2. كيف يعمل RAG

أولاً، تقسم الوثائق إلى قطع صغيرة، وتتحول كل قطعة إلى قائمة أرقام تسمى التضمين، تلتقط معناها. وتخزن هذه القوائم في فهرس يسمى غالباً قاعدة بيانات المتجهات. وعندما يصل سؤال، يجد النظام القطع الأقرب معنى إليه. ثم يضيف هذه القطع إلى الطلب المرسل إلى النموذج، فيكتب النموذج إجابته مستعيناً بها. وتعرض الأنظمة الجيدة أيضاً المصادر التي استخدمتها.

مثال: يتلقى أمين مكتبة سؤالاً عن التاريخ المحلي. وبتطبيق كيف يعمل RAG، يجد أولاً أكثر الصفحات صلة، ثم يكتب الإجابة منها ويذكر الكتب.

في العمل: يضيف فريق سامر سياسة الإجازات إلى مساعد RAG. وبتطبيق كيف يعمل RAG، يقسمها النظام إلى قطع ويخزنها في فهرس. وعندما يسأل سامر عن الإجازة، يجد المساعد الفقرة المناسبة ويجيب منها.

حقوق الطبع والنشر محمية © - أكاديمية أساسيات (شركة أساسيات للخدمات الاستشارية)

3. الفوائد

يتيح RAG للذكاء الاصطناعي أن يجيب من وثائقنا الحالية، لا من تدريبه فقط. وعندما تتغير وثيقة، نحدّث الفهرس عادة، فلا حاجة إلى تدريب النموذج من جديد. ويمكن أن تشير الإجابات إلى مصادرها، مما يسهل التحقق. ويمكن أن يساعد أيضاً على إبقاء المواد الخاصة تحت السيطرة، لأننا نحدد الوثائق التي يجوز للنظام البحث فيها.

مثال: يراجع بائع أحدث قائمة أسعار قبل أن يجيب زبوناً. وبتطبيق الفوائد، تكون الإجابة محدثة، ويستطيع البائع أن يعرض القائمة.

في العمل: يحدّث فريق الموارد البشرية عند سامر سياسة الإجازات. وبتطبيق الفوائد، يجددون الفهرس من دون تدريب نموذج جديد. فيجيب المساعد الآن بالقاعدة الجديدة ويعرض الفقرة التي استخدمها.

4. الحدود والممارسة الجيدة

RAG مفيد، لكنه ليس سحراً. فإذا وجدت خطوة البحث قطعاً خاطئة، أو كانت الوثائق قديمة أو غير واضحة، تأثرت الإجابة. وقد يخطئ النموذج في القراءة أو يضيف تفاصيل، لذلك ينبغي التحقق من الإجابات المهمة بمقارنتها بالمصادر. وتهم صلاحيات الوصول أيضاً، لأن الناس ينبغي ألا يتلقوا إلا المعلومات المسموح لهم برؤيتها. ومن الحكمة اختبار النظام بأسئلة حقيقية قبل الاعتماد عليه.

مثال: يستخدم مسافر خريطة قديمة. وبتطبيق الحدود والممارسة الجيدة، يقارنها المسافر الحذر بلافتات الشوارع.

في العمل: قبل الإطلاق، يختبر سامر المساعد بعشرين سؤالاً حقيقياً. وبتطبيق الحدود والممارسة الجيدة، يقارن كل إجابة بالفقرة المذكورة. فيجد وثيقة واحدة قديمة ويحذفها.

الخاتمة

وخلاصة القول، يتيح التوليد المعزّز بالاسترجاع للذكاء الاصطناعي أن يبحث عن الوثائق ذات الصلة قبل أن يجيب. وقد وصفه باتريك لويس وزملاؤه عام 2020، وهو اليوم طريقة شائعة لبناء مساعدات تستخدم معرفتها الخاصة. وهو يفيد في المعلومات الحالية والخاصة، ويستطيع عرض مصادره. لكنه ما زال يعتمد على وثائق جيدة وبحث جيد وتحقق بشري.

المصادر

• باتريك لويس وآخرون (2020): Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks، مؤتمر NeurIPS 2020 وأرشيف arXiv:2005.11401 – https://arxiv.org/abs/2005.11401
• Amazon Web Services: ما هو التوليد المعزّز بالاسترجاع (What is Retrieval-Augmented Generation?) – https://aws.amazon.com/what-is/retrieval-augmented-generation/
• Google Cloud: ما هو التوليد المعزّز بالاسترجاع (RAG)؟ – https://cloud.google.com/use-cases/retrieval-augmented-generation

حقوق الطبع والنشر محمية © - أكاديمية أساسيات (شركة أساسيات للخدمات الاستشارية)