ما هو توليد النص المعزز بالاسترجاع؟

هو إطار عمل للذكاء الاصطناعي يسترجع الحقائق من قاعدة معرفية RAG
،على أدق المعلومات وأحدثها (LLMs) خارجية لتثبيت النماذج اللغوية الكبيرة
.ويتيح للمستخدمين فهم عملية التوليد الخاصة بالنماذج اللغوية الكبيرة

النماذج اللغوية الكبيرة قد تكون غير متسقة. في بعض الأحيان، تجيب على الأسئلة بدقة، وفي أحيان أخرى تكرر حقائق عشوائية من بيانات تدريبها. إذا بدا أنها أحيانًا لا تعرف ما تقوله، فهذا لأنها لا تعرف حقًا. النماذج اللغوية الكبيرة .تفهم كيف ترتبط الكلمات إحصائيًا، لكنها لا تعرف معانيها

هو إطار عمل للذكاء الاصطناعي يهدف (RAG) توليد النص المعزز بالاسترجاع إلى تحسين جودة الاستجابات التي تولدها النماذج اللغوية الكبيرة عن طريق .تثبيت النموذج على مصادر خارجية للمعرفة لتكملة تمثيله الداخلي للمعلومات

في نظام للإجابة على الأسئلة يعتمد على النماذج اللغوية RAG يتمتع تنفيذ الكبيرة بفائدتين رئيسيتين: الأولى هي ضمان أن يكون لدى النموذج الوصول إلى أحدث الحقائق وأكثرها موثوقية، والثانية هي ضمان أن يكون لدى المستخدمين الوصول إلى مصادر النموذج، مما يتيح التحقق من دقة معلوماته وبالتالي تعزيز .الثقة فيها

أنت تريد أن" :IBM Research قال لويس لاستراس، مدير تقنيات اللغة في تقارن إجابات النموذج بالمحتوى الأصلي حتى تتمكن من رؤية ما يعتمد عليه في ".إجابته

فوائد إضافية. من خلال تثبيت النموذج اللغوي الكبير على RAG لدى مجموعة من الحقائق الخارجية والقابلة للتحقق، تقل فرص النموذج في استدعاء معلومات مضمنة في معاييره. وهذا يقلل من احتمالية أن يسرب

النموذج اللغوي الكبير بيانات حساسة، أو أن "يختلق" معلومات غير صحيحة أو مضللة.

من الحاجة إلى أن يقوم المستخدمون بتدريب النموذج RAG كما يقلل باستمرار على بيانات جديدة وتحديث معاييره مع تغير الظروف. بهذه الطريقة، أن يخفض التكاليف الحاسوبية والمالية لتشغيل روبوتات RAG يمكن لـ الدردشة المدعومة بالنماذج اللغوية الكبيرة في بيئة العمل.

يتكون من مرحلتين: الاسترجاع وتوليد المحتوى. في RAG كما يشير الاسم، فإن مرحلة الاسترجاع، تبحث الخوارزميات عن مقتطفات من المعلومات ذات الصلة بمطالبة أو سؤال المستخدم وتسترجعها. في بيئة مفتوحة المجال مثل إعدادات المستهلكين، يمكن أن تأتي هذه الحقائق من وثائق مفهرسة على الإنترنت؛ بينما في بيئة مغلقة المجال مثل إعدادات الشركات، يتم عادةً استخدام مجموعة أضيق من المصادر لأغراض الأمن والموثوقية.

يتم إلحاق هذه المجموعة من المعرفة الخارجية بمطالبة المستخدم وتمريرها إلى النموذج اللغوي. في مرحلة التوليد، يستند النموذج اللغوي الكبير إلى المطالبة المعززة وتمثيله الداخلي لبيانات تدريبه لتوليف إجابة جذابة مخصصة للمستخدم في تلك اللحظة. بعد ذلك، يمكن تمرير الإجابة إلى روبوت الدردشة مع روابط لمصادرها.

توليد النص المعزز بالاسترجاع البدائي (Naive RAG)

هي أول منهجية تم تبنيها. ظهرت هذه التقنية إلى الواجهة RAG هذه الفئة من لا يمكنها مواكبة البيانات الفورية (LLMs) بعد إدراك أن النماذج اللغوية الكبيرة ولا يمكنها الإجابة على الاستفسارات المتعلقة بمواضيع لم تكن جزءًا من بيانات عن كأس العالم ChatGPT 3.5 تدريب النموذج. على سبيل المثال، إذا سألت

للكريكيت لعام 2023، فلن يتمكن من الإجابة لأن معرفة النموذج توقفت قبل حدوث هذا الحدث وبالتالي لا يمكنه الإجابة على استفسارات متعلقة به.

تتبع تقنية Naive RAG عملية تشمل الفهرسة، الاسترجاع، التعزيز، وتوليد الاستجابة. دعونا نلقي نظرة على كل خطوة من هذه الخطوات بالتفصيل.

الفهرسة (Indexing)

الفهرسة هي خطوة تحضير البيانات حيث يتم استخراج البيانات التي يتم عليها الاسترجاع وتنظيفها من مصادر البيانات مثل الملفات، الروابط الإلكترونية، إلخ، وتحويلها إلى نص عادي. هذا النص العادي يكون عادةً عبارة عن سلسلة من آلاف الحروف. على سبيل المثال، إذا كنت ترغب في تنفيذ عملية إجابة الأسئلة من كتابك الجامعي بمساعدة النماذج اللغوية الكبيرة (LLMs). لا يمكن تقديم الكتاب بالكامل إلى النموذج اللغوي الكبير لأن ذلك قد يتجاوز نافذة السياق الخاصة بالنموذج. لذا، نقوم بتقسيم المحتوى الكامل إلى قطع أصغر قابلة للإدارة تسمى "chunks"، وهذه العملية تسمى التقطيع (chunking). بعد ذلك، يتم تحويل هذه القطع إلى متجهات عالية الأبعاد بمساعدة نماذج التضمين (embedding models). في نهاية هذه الخطوة، نحصل على قائمة من أزواج القطع والمتجهات لمصدر البيانات.

الاسترجاع (Retrieval)

هذه خطوة حاسمة في العملية حيث بمجرد استلام استفسار المستخدم، يتم استخدام نفس نموذج التضمين الذي استخدم لإنشاء متجهات مصادر البيانات لترميز الاستفسار إلى متجه. يتم استخدام هذا المتجه لإجراء بحث تشابه على مجموعة متجهات التضمين لمصادر البيانات واكتشاف القطع الأكثر تشابهًا. الطرق الشائعة المستخدمة في بحث التشابه تشمل التشابه الكوني (cosine similarity)، حاصل الضرب النقطي (dot product)، المسافة الإقليدية

إلخ. يتم استخدام القطع المسترجعة في خطوة (euclidean distance) التوليد للحصول على الاستجابة.

التوليد (Generation)

تُدمج القطع المسترجعة مع استفسار المستخدم والتعليمات في مطالبة واحدة، يتم تقديمها بعد ذلك إلى النموذج اللغوي الكبير لتوليد الاستجابة. بفضل إضافة معلومات إضافية إلى المطالبة، يمتلك النموذج اللغوي الكبير بيانات كافية لتوليد استجابة ذات صلة. يمكن تقييد النموذج ليقتصر على تقديم استجابة تلتزم فقط بالقطع المسترجعة. إدخال تاريخ المحادثة التفاعلية بين النموذج اللغوي الكبير والمستخدم يحسن أيضًا من جودة الاستجابات.

عيوب توليد النص المعزز بالاسترجاع البدائي (Naive RAG)

العديد (Naive RAG) تواجه منهجية توليد النص المعزز بالاسترجاع البدائي من العيوب عندما يتعلق الأمر بتقليل الهلوسات. تعاني هذه المنهجية من ضعف الدقة عندما لا يتم استرجاع القطع ذات الصلة بشكل صحيح. يؤدي وجود بيانات قديمة في المجموعة إلى استرجاع قطع غير دقيقة. قد لا تكون الاستجابة التي يتم توليدها مستندة على السياق الإضافي المقدم. ترتيب القطع يلعب دورًا مهمًا؛ حيث يجب أن يكون ترتيب القطع وفقًا لمدى ارتباطها بالاستفسار. هناك أيضًا خطر أن تكون الاستجابة محدودة وتحتوي فقط على المعلومات المسترجعة.