Engineer

Генерація, доповнена пошуком

генерування відповіді з документів, знайдених для запиту

Що це таке

Генерація, доповнена пошуком, — це спосіб відповідати на запитання за допомогою мовної моделі у два кроки. Спершу програма шукає в колекції документів фрагменти, які найімовірніше містять відповідь. Потім вона вставляє ці фрагменти у вхідні дані моделі поруч із запитанням, і модель пише відповідь на їхній основі, а не з того, що засвоїла під час навчання. Англійською метод називається retrieval-augmented generation, найчастіше скорочено RAG, і назва походить зі статті 2020 року про мовні завдання, що потребують знань, як-от відповіді на відкриті запитання.

Стаття описувала два види пам'яті. Попередньо навчена модель зберігає вивчене у своїх параметрах, і автори назвали це параметричною пам'яттю; окремий набір документів, у якому шукає навчений пошуковий модуль, — це непараметрична пам'ять. Їхня система поєднала попередньо навчену модель типу sequence-to-sequence зі щільним векторним індексом Вікіпедії, і автори порівняли два способи використати знайдене: один будує всю відповідь на тих самих знайдених фрагментах, другий може брати різні фрагменти для кожного токена, який пише. На трьох тестових наборах для відповідей на відкриті запитання система встановила нові найкращі результати, а згенерований нею текст автори визнали конкретнішим, різноманітнішим і точнішим щодо фактів, ніж текст порівнянної моделі без пошуку.

Програма може зібрати той самий шаблон із частин, які вже має: пошуковий індекс над власними документами і велику мовну модель, яку вона викликає через API. Пошукова половина — це звичайний пошук, і тут трапляються обидва його різновиди. Розріджений пошук зіставляє слова, так само як повнотекстовий пошук ранжує документи за функцією оцінювання на кшталт BM25. Щільний пошук перетворює запитання й кожен фрагмент на вектори чисел і повертає фрагменти, чиї вектори лежать найближче, тобто це

векторний пошук. Стаття 2020 року про щільний пошук фрагментів показала, що щільний пошуковий модуль, навчений на помірній кількості пар із запитання й фрагмента, обходив сильний базовий варіант на BM25 на величину від 9 до 19 відсоткових пунктів за тим, як часто правильний фрагмент з'являвся серед перших двадцяти результатів. Система може запускати обидва пошуки й об'єднувати два списки.

Застосуйте поняття до рішення

Застосуйте «Що таке генерація, доповнена пошуком?» до конкретного рішення. Ми можемо пов'язати поняття із системою, яку ви створюєте або купуєте. Ваші дані зібрано там, де їх дістануть ШІ-агенти: одна версія кожного клієнта, ніхто не бачить більше, ніж йому можна, а кожну відповідь можна простежити до джерела.

Переглянути наші послуги Зв'язатися з нами →

Перевірте припущення за терміном

Використайте пояснення та пов'язані роботи, щоб сформулювати запитання для проєкту. Ми можемо перевірити припущення разом до вибору інструмента. Прискорила пайплайн географічних даних у 50 разів — 50x швидші пайплайни географічних даних

Прискорила пайплайн географічних даних у 50 разів Зв'язатися з нами →

Чому це важливо

Мовна модель знає лише те, що було в її навчальному тексті, а цей текст закінчується певною датою і ніколи не містив ваших договорів, прайс-листів чи нотаток із наради минулого тижня. Стаття про RAG називає дві вади знань, збережених у параметрах: важко показати, звідки взялася відповідь, і важко ці знання оновити. Пошук розв'язує обидві. Відповідь може посилатися на фрагменти, які отримала модель, тож читач може відкрити джерело й перевірити, а нові знання доходять до моделі, щойно документ потрапляє в індекс, без жодного навчання.

Пошук також повертає питання, хто що може бачити, туди, де йому місце. Пошук фрагментів — це звичайний пошук, а пошук може виконуватися з правами того, хто питає, тож модель бачить лише ті фрагменти, які ця людина могла б відкрити й сама. У моделі, навченої на тих самих документах, такої межі немає: усе, чого вона навчилася, доступне кожному, хто нею користується.

Ціна в тому, що відповідь не може бути кращою за те, що повертає пошук. Модель, якій дали не ті фрагменти, напише з них гладку відповідь, а модель, якій не дали нічого корисного, може однаково відповісти з пам'яті. Тому більшість роботи в системі генерації, доповненої пошуком, іде в дані й пошук, а не в модель.

Як це виглядає на практиці

Трасування нижче простежує одне запитання через систему RAG. Пошук виконується з правами користувача, що питає, кожен фрагмент надходить разом із файлом і сторінкою, звідки його взято, а інструкції наказують моделі відповідати лише з фрагментів і ні з чого іншого:

```
запитання:  Які умови поставки діють для замовлення 4471?
користувач: sales-3 (роль: продажі, регіон: північ)

знайдено, відфільтровано до того, що може читати sales-3:
  [1] contracts/framework-2025.pdf, сторінка 4
      «Товар доставляється на склад покупця,
       якщо в замовленні не вказано інше»
  [2] orders/4471.txt, рядок 12
      «Умови поставки: згідно з рамковою угодою»

інструкції:
  Відповідай лише з фрагментів нижче й посилайся на них як [n].
  Якщо вони не відповідають на запитання, так і скажи.

відповідь:
  Замовлення 4471 посилається на рамкову угоду [2], згідно
  з якою товар доставляється на склад покупця [1].
```

Кожна частина цього трасування — дані, які система може перевірити. Тест може підтвердити, що на відоме запитання повернулися правильні фрагменти, скрипт — що кожне `[n]` у відповіді вказує на фрагмент, який справді знайшли, а людина, що читає відповідь, може відкрити четверту сторінку договору.

Підготовці, якої це потребує, присвячено сторінку цього сайту дані компанії, готові до ШІ-агентів. Вона описує, як старі PDF, скани й поштові архіви перетворюються на текст, у якому можна шукати, з посиланням на джерело, і як кожен рядок зберігає, звідки він надійшов: система, завантаження, файл і сторінка, тож агент може послатися на запис, що стоїть за цифрою, а людина може його відкрити. Сторінка також задає права доступу в самих даних, а не в промпті: агент діє з привілеями свого користувача, а не адміністратора, і кожне читання записується в журнал із зазначенням людини, для якої його зроблено.

Де виникають проблеми

Пошук промахується, а модель однаково відповідає. Коли правильний фрагмент не знайдено, модель усе одно щось отримує й усе одно пише відповідь. Вимірюйте пошук окремо: запишіть справжні запитання разом із фрагментами, що на них відповідають, і перевіряйте, як часто ці фрагменти з'являються серед результатів, які отримує модель. Скажіть моделі прямо повідомляти, коли фрагменти не відповідають на запитання, і додайте такі запитання до своїх тестових кейсів.

Фрагменти, що розрізають відповідь навпіл. Поділ документів на шматки фіксованої довжини відокремлює таблицю від її заголовка, а пункт договору від винятку з нього, і кожна половина, знайдена окремо, вводить в оману. Діліть за структурою, яку документ уже має, тобто за розділами, сторінками й рядками таблиць, і зберігайте назву та розташування документа з кожним шматком.

Індекс не зважає на те, хто питає. Індекс, який побудував сервіс із правом читати все, повертатиме фрагменти з документів, яких людина, що питає, не може відкривати, і модель перекаже їх цій людині. Фільтруйте за правами користувача всередині пошуку, в індексі чи базі даних, перш ніж щось дійде до моделі. Правило в промпті — це не контроль доступу.

Інструкції, підкинуті в документ. Знайдений текст написав хтось інший, і модель не вміє надійно відрізнити фрагмент від інструкції. Профіль NIST для генеративного ШІ описує непряму ін'єкцію промптів як атаки на застосунки, побудовані на мовних моделях, через дані, які, найімовірніше, буде знайдено, а стаття 2023 року продемонструвала такі атаки на справжніх системах. Вважайте кожен знайдений фрагмент недовіреними вхідними даними, не дозволяйте жодному фрагментові запускати дію без перевірки в коді й не давайте інструментам, які може викликати модель, більше прав, ніж має користувач.

Індекс, що відстає. Документ змінили чи видалили, а індекс досі тримає старі фрагменти, тож система впевнено відповідає за договором, який уже не діє. Оновлюйте індекс щоразу, коли змінюється джерело, прибирайте фрагменти, коли видаляють їхній документ, і показуйте дату кожного джерела поруч із відповіддю.

Посилання, які не кажуть того, що стверджує відповідь. Модель може послатися на фрагмент [2] у реченні, якого фрагмент [2] не підтверджує, а NIST зазначає, що згенеровані посилання й самі можуть бути конфабуляцією. Перевіряйте в коді, що кожне посилання вказує на знайдений фрагмент, і нехай люди в межах регулярного тестування звіряють вибіркові відповіді з фрагментами, на які ті посилаються.

Що почитати далі

- Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks — стаття 2020 року, що представила моделі RAG і порівняла два їхні варіанти на відповідях на запитання та генерації тексту.
- Dense Passage Retrieval for Open-Domain Question Answering — щільний пошуковий модуль, навчений на парах із запитання й фрагмента, у порівнянні з BM25 за тим, чи знаходить він фрагмент із відповіддю.
- NIST AI 600-1: Generative Artificial Intelligence Profile — профіль ризиків NIST для генеративного ШІ: конфабуляція, приватність даних і непряма ін’єкція промптів.
- Not what you’ve signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection — як інструкції, сховані в знайдених даних, захоплюють застосунок, з атаками на справжні системи.

Словник

Обирайте за фактичними умовами

Надайте репрезентативний набір даних, запитання до нього та поточне вузьке місце. Узгодимо вимірювання якості, пропускної здатності та відновлення.

Дивитися, скільки це коштує Зв'язатися з нами →

Зробіть термінологію практичною

Запитайте, як «Що таке генерація, доповнена пошуком?» стосується вашої організації. Конкретного прикладу з поточної системи достатньо для початку розмови.

Переглянути наші послуги

welcome@engineer.company +45 50 10 61 84 Копенгаген, Данія

Візьміть цю сторінку із собою

Візитівка

Дбайливо зроблено з любов'ю — D.

CVR 45969703 · Tagensvej 105C, 3. th, 2200 København N

Авторське право © 2025 – дотепер · Engineer ApS · Усі права захищено.

Надруковано з https://engineer.company/uk/glossary/retrieval-augmented-generation/