Місія

Ключ	Значення
Назва	Пошуково-доповнена генерація (RAG)
Опис	Інтеграція пошуку зовнішніх даних у процес генерації, що підвищує здатність моделі надавати точні та актуальні відповіді.
Історія	Розроблена в 2020 році, RAG швидко отримала популярність як спосіб покращити точність та релевантність генеративних завдань для великих мовних моделей (LLM).
Три ключові компоненти	Пошук, генерація та розширення.
Технології пошуку	Простий маркер, пошук сутностей, блоки, граф знань.
Технології генерації	Адаптивне генерування, множинне генерування.
Технології розширення	Посилення, ранжування.
Переваги	Покращена точність, релевантність, достовірність.
Недоліки	Додаткові витрати часу та ресурсів.
Майбутні напрямки	Подолання поточних викликів, розширення мультимодальних умов, розвиток екосистеми.

Додаткові відомості

- RAG може застосовуватися до широкого спектру генеративних завдань, таких як відповіді на запитання, переклад, створення тексту та коду.
- RAG може бути використана для вирішення таких проблем, як галюцинації, застарілі знання та непрозорість процесів міркування.
- RAG є багатообіцяючим напрямком досліджень, який має потенціал для революціонізації генеративних завдань для LLM.

Компонент генерації в RAG відповідає за створення відповіді на запит. Він повинен бути творчим, інформативним та обґрунтованим.

Типи генерації

Існує два основних типи генерації, які можуть використовуватися в RAG:

- Текстова генерація: Цей тип генерації використовується для створення тексту, такого як відповіді на запитання, статті та твори.
- Код-генерація: Цей тип генерації використовується для створення коду, такого як програми, скрипти та математичні формули.

Ефективність генерації

Ефективність компонента генерації є важливою для RAG-фреймворків. Компонент генерації повинен бути здатним швидко та творчо створювати відповіді на запити.

Точність генерації

Точність компонента генерації також є важливою для RAG-фреймворків. Компонент генерації повинен бути здатним створювати відповіді, які є точними та обґрунтованими.

Масштабованість генерації

Масштабованість компонента генерації також є важливою для RAG-фреймворків. Компонент генерації повинен бути здатним обробляти великі обсяги даних.

Сучасні методи генерації для RAG

У останні роки було розроблено ряд сучасних методів генерації для RAG. Ці методи спрямовані на підвищення ефективності, точності та масштабованості компонента генерації.

Одним із сучасних методів генерації для RAG є генерація на основі трансформерів. Цей метод використовує трансформери, щоб генерувати текст, який є творчим, інформативним та обґрунтованим.

Іншим сучасним методом генерації для RAG є генерація на основі нейронних мереж з генеративно-пригнічувальним навчанням. Цей метод використовує нейронні мережі з генеративно-пригнічувальним навчанням, щоб генерувати текст, який є точним та обґрунтованим.

**Перспективи розвитку методів генерації

Розробка ефективних методів генерації для RAG є важливою для подальшого розвитку RAG-фреймворків. У майбутньому можна очікувати, що будуть розроблені ще більш ефективні методи генерації, які будуть здатні генерувати ще більш творчі, інформативні та обґрунтовані відповіді.

Ключові висновки

- Компонент генерації в RAG відповідає за створення відповіді на запит.
- Існує два основних типи генерації, які можуть використовуватися в RAG: текстова генерація та код-генерація.
- Ефективність, точність і масштабованість компонента генерації є важливими для RAG-фреймворків.
- У останні роки було розроблено ряд сучасних методів генерації для RAG, які спрямовані на підвищення ефективності, точності та масштабованості компонента генерації.

RAG виникла в 2020 році як спосіб покращити точність та достовірність LLM. З тих пір вона швидко розвивалася, еволюціонуючи через кілька парадигм.

- Наївна RAG: Наївна RAG є найпростішою парадигмою RAG. Вона включає лише один етап пошуку, який виконується до етапу генерації. Пошуковий результат використовується для інформування процесу генерації, але не включається в кінцеву відповідь.
- Вдосконалена RAG: Вдосконалена RAG є більш складною парадигмою RAG. Вона включає два етапи пошуку: один перед етапом генерації та один після. Пошуковий результат першого етапу використовується для інформування процесу генерації, а пошуковий результат другого етапу використовується для доопрацювання кінцевої відповіді.
- Модульна RAG: Модульна RAG є найзагальнішою парадигмою RAG. Вона дозволяє користувачам налаштовувати компоненти пошуку, генерації та розширення відповідно до своїх конкретних потреб.

Компонент пошуку в RAG відповідає за отримання інформації з зовнішніх баз даних. Ця інформація може використовуватися для інформування процесу генерації, щоб підвищити точність і достовірність відповідей LLM.

Типи пошуку

Існує два основних типи пошуку, які можуть використовуватися в RAG:

- Пошук по тексту: Цей тип пошуку використовується для пошуку інформації в текстових документах, таких як статті,

книги та веб-сторінки.
- Пошук по коду: Цей тип пошуку використовується для пошуку інформації в коді, такому як програмування, скрипти та математичні формули.

Ефективність пошуку

Ефективність компоненту пошуку є важливою для RAG-фреймворків. Компонент пошуку повинен бути здатним швидко та точно отримувати інформацію з зовнішніх баз даних.

Точність пошуку

Точність компоненту пошуку також є важливою для RAG-фреймворків. Компонент пошуку повинен бути здатним повертати релевантну та точну інформацію.

Масштабованість пошуку

Масштабованість компоненту пошуку також є важливою для RAG-фреймворків. Компонент пошуку повинен бути здатним обробляти великі обсяги даних.

Сучасні методи пошуку для RAG

У останні роки було розроблено ряд сучасних методів пошуку для RAG. Ці методи спрямовані на підвищення ефективності, точності та масштабованості компоненту пошуку.

Одним із сучасних методів пошуку для RAG є пошук з розширеним представленням. Цей метод використовує розширені представлення тексту та коду для пошуку релевантної інформації.

Іншим сучасним методом пошуку для RAG є пошук з навчанням на основі прикладів. Цей метод навчається на прикладах запитів і відповідей, щоб покращити точність пошуку.

Перспективи розвитку методів пошуку для RAG

Розробка ефективних методів пошуку для RAG є важливою для подальшого розвитку RAG-фреймворків. У майбутньому можна очікувати, що будуть розроблені ще більш ефективні методи пошуку, які будуть здатні обробляти ще більші обсяги даних та повертати ще більш точну та релевантну інформацію.

Ключові висновки

- Компонент пошуку в RAG відповідає за отримання інформації з зовнішніх баз даних.
- Існує два основних типи пошуку, які можуть використовуватися в RAG: пошук по тексту та пошук по коду.
- Ефективність, точність і масштабованість компоненту пошуку є важливими для RAG-фреймворків.

- У останні роки було розроблено ряд сучасних методів пошуку для RAG, які спрямовані на підвищення ефективності, точності та масштабованості компоненту пошуку.

1 велика кількість людей що хочуть працювати онлайн 2 величезна потреба формалізації мед карт 3 оптимізація сервісу медичних консультацій 4 експертам розкидуємо сторінки для розпізнавання випадковим чином

Obsidian - для кожного

Obsidian можна порадити всім, хто так чи інакше працює з текстовою інформацією. Особливо програма послужить відмінним помічником тим, хто навчається і веде конспект (навіть якщо ви самоосвіти на онлайн-курсах), генерує контент, створює бази знань. Все тому, що утиліта максимально адаптована під створення як невеликих нотаток, так і об’ємних текстів, прикрашених посиланнями, картинками, відео, аудіо, внутрішніми перелінковками і т. д.

Obsidian володіє нескінченними можливостями, але двома найважливішими стовпами, на яких вона стоїть, є Markdown-формат і метод Zettelkasten.

Маркдаун - це спрощений мова розмітки, що служить для максимально простого і зрозумілого форматування звичайного тексту. Перевагою формату є одночасна легкість читання людиною і придатність для подальшого перетворення в продвинуті мови розмітки, такі як HTML і Rich Text.

Простими словами, Markdown дозволяє без зайвих зусиль створювати текст з маркованими і нумерованими списками, виділенням жирним або курсивом, заголовками від 1-го до 6-го рівнів, цитатами і т. д. Ніяких складних знань при цьому не потрібно, оскільки все форматування виконується за допомогою відомих всім символів, наприклад, зірочки (*), решітки (#), риски (-), нахилної риски (/) і інших.

Як виглядає Markdown в Obsidian. Зліва режим редагування, справа - режим перегляду готової верстки

Звичайно, що для зручності і швидкості використовуються гарячі клавіші.

Детально метод ведення нотаток Zettelkasten (з німецького - «картотечний ящик») описано на . Якщо коротко, то це потужний спосіб підвищення результативності роботи. Систему придумав німецький соціолог Ніклас Луман - вона дозволила йому за 40 років продемонструвати неймовірну продуктивність. Вчений випустив понад 70 книг і 400 наукових статей, причому сам він

нічого не вигадував. Основну роботу за нього виконував метод Zettelkasten, побудований на таких принципах:

- Атомарність. Кожна нотатка містить тільки одну ідею;
- Автономність. Будь-яка нотатка незалежна і зрозуміла без інших нотаток, як будь-який кубик конструктора;
- Зв’язок. Всі нові нотатки повинні бути з’єднані з вже створеними, утворюючи єдину мережу. Слід уникати не зв’язаних нотаток, оскільки вони все одно будуть втрачені через невостребуваність;
- Своїми словами. Ніякого копіювання (за винятком формул, цитат і т. п.). Кожна інформація повинна бути осмислена, зрозуміла і додана в нотатки в формі лаконічної і короткої думки. То мережа описана своїми словами;
- Більше посилань. Описуючи ідею своїми словами, необхідно додавати посилання на джерела, звідки бралася інформація. Можливо в майбутньому це дозволить згенерувати нові ідеї;
- Більше власних думок. Всі спостереження або висновки не варто зберігати в голові - краще записати їх в Obsidian, дотримуючись описаних принципів Zettelkasten;
- Проста структура. Немає сенсу розфасовувати всі нотатки по папках або категоріях. Це робити можна, але необов’язково. Zettelkasten не має привілейованих позицій
- Нотатки для зв’язку. Для зв’язку випадкових записів, коли цей зв’язок малоочевидний, додавайте додаткові нотатки для пояснення складеної структури;
- Головні нотатки. Це підсумовуючі записи, що містять посилання на всі роботи, що стосуються певних тем. Це дозволить максимально просто створити зрозумілу ієрархію;
- Збереженість. Не треба видаляти старі нотатки - замість цього створюйте нові з посиланням на старі та поясненням причин редагування. Це дозволить зрозуміти вашу еволюцію мислення;
- Немає страху. Інформації багато не буває, і випадкові записи ніяк не зіпсують метод. Найгірше, що може статися, так це те, що нотатка не відразу виявиться актуальною та корисною.

Здається, що ця система виглядає складно, але реалізувати її в Obsidian простіше простого, причому на допомогу в програмі є «Граф». Це візуальне відображення всіх записів та зв’язків, що нагадує нейронну мережу. А з урахуванням того, що окремі теми можна розфарбувати в цікавий колір, налаштувати силу притягання зв’язків, заплутатися в нотатках просто не вийде.

Основні можливості Obsidian

Вони безмежні і просто переказувати все те, що вміє програма, немає сенсу - все є на просторах інтернету та YouTube. В останньому, до речі, дуже багато гайдів для початківців, порівнянь з іншими програмами, навіть є цілі плейлисти з курсами як працювати в утиліті.

Однак все це більше необхідно тим, чиє заняття - це великі та складні статті, створення наукових робіт, тих же курсів навчання, лекцій. Для простих користувачів, котрим є і автор цього матеріалу, достатньо базових можливостей та деяких підключених плагінів. Поріг входження мінімальний, і розібратися з усім, що буде описано далі, можна за один день.

Перед створенням нотаток варто уточнити, що краще всього зберігати базу знань в хмарі для подальшого доступу до всіх нотаток і з ПК, і зі смартфона. Є як просте, але платне рішення засобами Obsidian Sync, а можна отримати синхронізацію і задарма, але з «костилів». Опишу другий варіант.

У моєму випадку використовується Google Диск, так як через інші сервіси нормально запровадити безкоштовну синхронізацію файлів не вдалося. Все через те, що в останніх версіях Android немає доступу до файлів за шляхом Android / Data. Однак у випадку з Google Диск є додатковий додаток DriveSync, що дозволяє синхронізувати файли Obsidian, на телефоні та ПК (в обидва боки) через гугловську хмару.

- Спочатку встановлюємо Google Диск на ПК та смартфоні;
- Створюємо папку на Google Диск, в якій будуть зберігатися нотатки, вона ж - база знань;
- Підключаємо базу знань в Obsidian на ПК та встановлюємо на смартфон додаток DiveSync з форуму 4PDA ();
- Останнім етапом створюємо папку на смартфоні, з якої буде синхронізуватися Google Диск, налаштовуємо в DiveSync.

У DriveSync вказуємо папку на Google Диск і папку, розташовану на смартфоні. Готово, тепер у нас працює двостороння синхронізація. Файли, створені на ПК або смартфоні в Obsidian будуть доступні на всіх синхронізованих пристроях. Єдиний нюанс - синхронізувати доведеться вручну, так як ця опція в DriveSync платна.

Парадоксом нашого часу є те, що ми маємо високі будівлі, але низьку терпимість, широкі магістралі, але вузькі погляди.Витрачаємо більше, але маємо менше, купуємо більше, але радіємо менше.

Маємо великі будинки, але менші родини, кращі зручності, але менше часу.

Маємо кращу освіту, але менше розуму, кращі знання, але гірше оцінюємо ситуацію, маємо більше експертів, але й більше проблем, кращу медицину, але гірше здоров'я. П'ємо надто багато, куримо надто багато, витрачаємо надто безвідповідально, сміємося надто мало, їздимо надто швидко, гніваємося надто легко, спати лягаємо надто пізно, прокидаємося надто втомленими, читаємо надто мало, надто багато дивимось телебачення і молимося надто рідко.

Збільшили свої домагання, але скоротили цінності. Говоримо занадто багато, любимо занадто рідко і ненавидимо занадто часто. Знаємо, як вижити, але не знаємо, як жити. Додаємо роки до людського життя, але не додаємо життя д нео років.

Досягли Місяця і повернулися, але насилу переходимо вулицю і знайомимося з новим сусідом. Підкорюємо космічні простори, але не духовні. Очищаємо повітря, але забруднюємо душу.

Підкорили собі атом, але не свої забобони. Пишемо більше, але дізнаємося менше. Плануємо більше, але досягаємо меншого. Навчилися поспішати, але не чекати. Створюємо нові комп'ютери, які зберігають більше інформації, ніж раніше, але спілкуємося все менше. Це час швидкого харчування і поганого травлення, великих людей і дрібних душ, швидкого прибутку і важких взаємин.

Час зростання сімейних доходів і зростання кількості розлучень, красивих будинків і зруйнованих домашніх вогнищ.

Час коротких відстаней, одноразових підгузків, разової моралі, зв'язків на одну ніч; зайвої ваги і пігулок, які роблять усе: збуджують нас, заспокоюють нас, вбивають нас…

Приділяйте більше часу тим, кого любите, бо вони з вами не назавжди.

Запам'ятайте, скажіть добрі слова тим, хто дивиться на вас від низу до верху із захопленням, бо ця маленька істота скоро виросте і її вже не буде поруч із вами.

Запам'ятайте і палко притисніть близьку людину до себе, бо це єдиний скарб, який можете віддати від серця, і він не коштує ні копійки.

Запам'ятайте і тримайтеся за руки та цінуйте кожну мить життя.

Моя інформаційна система дозволить робітникам і працівникам підприємств виробничого сектору ефективно використовувати

можливості краудсорсингу, S-BPM, бережливого виробництва та сучасних технологій для автоматизації та оптимізації бізнес-процесів. Ця система стане основою цифрової трансформації таких підприємств, яка дозволить їм адаптуватися до змінних умов ринку, підвищити конкурентоспроможність, збільшити прибутковість та задовольнити потреби своїх клієнтів. Моя візія проекту полягає в тому, щоб створити інформаційну систему, яка буде відкритою, гнучкою, масштабованою, інтегрованою, інтерактивною та інноваційною. Я хочу, щоб моя система була не тільки інструментом, а й платформою для співпраці, навчання, творчості та розвитку робітників і працівників підприємств виробничого сектору. Я вірю, що моя система зробить їхню роботу більш цікавою, зручною, продуктивною та задоволенною. Я сподіваюся, що моя система сприятиме підвищенню якості життя робітників і працівників, а також соціально-економічному розвитку регіонів, де розташовані підприємства виробничого сектору. Це моя візія проекту, яку я хочу поділитися з вами. Дякую за вашу увагу.

На початку 21 століття відбувся перехід від культури спостереження та обговорення, яка допускала можливість читання, перегляду, обговорення, до культури безпосередньої участі у створенні та зміні текстових і медійних об’єктів. зміні текстових і медійних об’єктів. У всіх галузях знань - у науці, законотворчості, економіці, в освіті, у громадській діяльності - скрізь ми бачимо участь активних груп громадян уже не у використанні та обговоренні текстів і медіаоб’єктів, а в їх створенні і поліпшенні. У сфері інформаційних технологій громадяни залучаються не тільки до використання програмного забезпечення, а й до тестування програмних продуктів, що випускаються. програмних продуктів, що випускаються, до участі в розробленні програмних систем і власних додатків. З розвитком ринку мобільних додатків практика створення власних програм стає повсюдною, і в ній беруть участь не тільки професіонали, а й студенти та школярі. В сфері накопичення та організації знань співробітники та клієнти організацій залучаються не тільки до використання, а й до формування цифрових колекцій. Сучасні музеї та бібліотеки знаходять шляхи для об’єднання з громадськими сховищами текстів, фотографій і відеоматеріалів на кшталт Вікіпедія, Flickr, YouTube і відкривають можливість не лише читання та спостереження, а й повторного багаторазового використання. У сфері збору та поширення новин наші сучасники дедалі частіше виступають не в ролі читачів і глядачів, а в ролі активних творців новинного контенту. У сфері науки громадяни завдяки відкритості інформаційних ресурсів громадяни отримують доступ до величезних масивів

наукової інформації, первинних даних, які вони можуть самостійно вивчати, аналізувати та використовувати у своїй діяльності, вступаючи до спільноти обміну науковими даними. В астрономії, геоінформатиці, екології, соціології завдяки відкритим даним і сервісам, які підтримують роботу з такими даними, повертається ера аматорів - людей, які беруть участь у наукових дослідженнях, а не лише користуються результатами цих досліджень. У сфері економіки залучення громадян до вирішення економічних проблем відбувається настільки активно, що широкого поширення набуває термін - вікіно. поширення набуває термін - вікіноміка, що означає економіку, засновану на участі громадян, на “мудрості натовпу”. У сфері законодавчої та законотворчої практики відбувається активне залучення громадян не тільки до обговорення, але й до редагування та поліпшення текстів законів. Держава розглядається як платформа, яка забезпечує участь громадян у колективній діяльності з поліпшення існуючих рішень, сервісів, служб і наявних рішень, сервісів, служб і документів. У сфері регіонального управління та планування громадяни залучаються до спільне вирішення міських проблем, пропонують і реалізують рішення для проблем, які здавалися некерованими, дикими та нерозв’язними.

Ми створюємо середовище для спільного створення, вдосконалення та просування покращень у рамках громадських та виробничих проектів.

Колективне створення, вдосконалення та просування покращень означає, що спільнота розробляє нові інформаційні об’єкти, які допоможуть їй стати більш ефективною в досягненні соціальних чи виробничих цілей. Нижче наведено принципи, на яких ґрунтується така колаборація:

Публічна розробка інформаційного об’єкта, далі документа, спрямована на вирішення складних соціальних чи виробничих проблем, має високу освітню цінність. Ця практика допомагає не лише створити якісний документ, але й створити спільноту людей, зацікавлених у його впровадженні, просувати інноваційний документ, підтримувати його в процесі реалізації, досягти нового рівня його розуміння та сприйняття виробничим колективом, суспільством.

Спільний пошук шляхів вирішення проблем має пріоритет над обговоренням без можливості впливу на ситуацію. Наша технологія дозволяє учасникам колаборації створювати власні версії документів, які можуть оцінити інші учасники.

Цінність обговорення менша, ніж цінність співпраці. Сучасна

співпраця не завжди потребує дискусій. Обговорення може забрати стільки часу і зусиль, у учасників не залишиться ресурсів для співпраці.

Коли відомий американський сатирик Джордж Карлін поховав свою дружину, він написав своєрідний [[00 Введення/4 Маніфест|4 Маніфест]] про сучасне життя. У ньому він напрочуд ємко, точно і пронизливо зміг розповісти про той біль, який відчував, усвідомлюючи, як безцільно розтрачуються хвилини, з яких і складається наше життя.

Слова, написані в 70-80х роках актуальні зараз, як ніколи.

Приділяйте більше часу тим, кого любите, бо вони з вами не назавжди.

Запам’ятайте, скажіть добрі слова тим, хто дивиться на вас від низу до верху із захопленням, бо ця маленька істота скоро виросте і її вже не буде поруч із вами.

Запам’ятайте і палко притисніть близьку людину до себе, бо це єдиний скарб, який можете віддати від серця, і він не коштує ні копійки.

Запам’ятайте і тримайтеся за руки та цінуйте кожну мить життя.