عنوان: یادداشت‌های ساده و گسترش ایده‌ها درباره PathRAG

نویسنده: کورش اصیل

- مقدمه (به زبان ساده)

مدل‌های بزرگ زبانی (LLM) ها قدرتمند هستند، اما وقتی اطلاعات درست یا کافی در اختیارشان نباشد ممکن است خطا کنند. روش «تولید تقویت‌شده با بازیابی-Retrieval) Augmented Generation — RAG)با این مشکل مقابله می‌کند: ابتدا مدارک یا متون مرتبط بازیابی می‌شوند و سپس مدل از آن‌ها برای تولید پاسخ استفاده می‌کند.

روش‌های RAG سنتی معمولاً مستندات جداگانه‌ای بازیابی می‌کنند که این کار ممکن است موجب تکرار اطلاعات و ورود نویز شود. RAG مبتنی بر گراف تلاش می‌کند با ساختاردهی داده‌ها به صورت گراف — که گره‌ها بخش‌های متنی و یال‌ها روابط بین آن‌ها را نشان می‌دهند — این مشکل را حل کند.

PathRAG یک روش مبتنی بر گراف است که به‌جای بازیابی تعداد زیادی گره نامرتبط، روی یافتن مسیرهای معنادار در گراف تمرکز می‌کند. این کار باعث کاهش نویز و بهبود کیفیت پاسخ می‌شود.

- PathRAGچه کار می‌کند :

PathRAGدر سه گام اصلی عمل می‌کند:

1. بازیابی گره‌ها:

اول مجموعه‌ای از گره‌های مرتبط (بخش‌های متنی) با توجه به پرسش کاربر بازیابی می‌شود.

2. هرس مبتنی بر جریان:(Flow-based Pruning)

PathRAGاز یک «جریان» مجازی استفاده می‌کند که از گره‌های با بیشترین مرتبط بودن آغاز می‌شود و در سراسر گراف پخش می‌شود. گره‌هایی که جریان خیلی کمی دریافت می‌کنند از مجموعه حذف می‌شوند. این کار فقط گره‌های مهم را نگه می‌دارد و نویز را حذف می‌کند.

3. انتخاب مسیر و ترتیب‌دهی:

سپس مسیرهای متصل بین گره‌های باقی‌مانده پیدا می‌شوند؛ هر مسیر یک امتیاز قابلیت‌اطمینان دریافت می‌کند. مسیرها به ترتیبی وارد ورودی مدل می‌شوند که مسیر قابل‌اطمینان‌تر در انتها قرار گیرد — این ترتیب به LLM کمک می‌کند تا روی مسیرهای قابل‌اعتمادتر تمرکز بیشتری داشته باشد.

- قوت‌هایPathRAG
  - نسبت به GraphRAG و LightRAG تکرار اطلاعات را کاهش می‌دهد.
  - ساختار گراف را برای نمایش روابط بین اطلاعات به‌خوبی به کار می‌گیرد.
  - جریان منطقی و انسجام دانش بازیابی‌شده را بهبود می‌بخشد.
  - از نظر مصرف توکن کارآمدتر است.

- ایده‌های جدید الهام‌گرفته شده
  - رتبه‌بندی مسیر یادگرفته‌شده (یادگیری‌شده)

به‌جای اتکا صرف به امتیازهای مبتنی بر جریان یا قواعد هورستیک، می‌توان یک رتبه‌بندی یادگیرشی ساده برای مسیرها افزود. مدل می‌تواند یاد بگیرد کدام مسیرها معمولاً برای پاسخ‌گویی مفیدترند. هدف از این پیشنهاد، هدایت بهتر انتخاب مسیرها با استفاده از داده‌های آموزشی است. لازم نیست در این گزارش پیاده‌سازی کامل انجام شود؛ این صرفاً یک جهت‌گیری پژوهشی پیشنهادی است.

- خلاصه‌سازی مسیرها

مسیرهای طولانی اغلب شامل متن اضافه و غیرضروری‌اند. می‌توان هر مسیر را پیش از ارسال به مدل به ۱–۲ جمله کوتاه خلاصه کرد. این کار هم توکن مصرفی را کاهش می‌دهد و هم فهم سریع‌تر و راحت‌ترِ محتوای مسیر را برای مدل فراهم می‌کند.

- آگاهی از بودجه توکن(Token Budget Awareness)

به‌جای انتخاب شمار مشخصی از مسیرها، سیستم می‌تواند مسیرها را تا زمانی انتخاب کند که یک حد توکنی مشخص پر شود. در این حالت مسیرهای مهم‌تر باید اولویت داده شوند تا نهایت استفاده از بودجه توکنی حاصل شود.

- بحث

PathRAG نشان داد که بازیابی ساختاریافته (با توجه به روابط بین بخش‌های متن) بهتر از بازیابی اسناد منفرد و جداست. ایده‌های پیشنهادی در این گزارش تلاش می‌کنند انتخاب مسیر را هوشمندتر و کارآمدتر کنند، بدون اینکه پیچیدگی سیستم به شکل چشمگیری افزایش یابد.