فصل ۵: RAG

۱.۰ مقدمه

ظهور مدل‌های زبانی بزرگ تحولی شگرف در پردازش متن ایجاد کرده است، اما استفاده از این مدل‌ها در حوزه‌های حساس و تخصصی مانند حقوق با چالش‌های بنیادین روبروست. مهم‌ترین چالش، پدیده‌ای به نام «توهم» (Hallucination) است که در آن مدل زبانی، گزاره‌هایی را تولید می‌کند که از نظر ساختاری صحیح و متقاعدکننده به نظر می‌رسند، اما فاقد مبنای واقعی یا قانونی هستند [۱۶]. در حوزه حقوق، جایی که دقت و استناد صحیح به مواد قانونی و آرای قضایی اهمیتی حیاتی دارد، حتی کوچکترین خطای ناشی از توهم می‌تواند منجر به نتایج زیان‌بار و گمراه‌کننده شود [۱۷].

علاوه بر مشکل توهم، مدل‌های زبانی با محدودیت «دانش ایستا» (Static Knowledge) مواجه هستند. دانش پارامتریک این مدل‌ها محدود به داده‌هایی است که در زمان آموزش دیده‌اند و توانایی دسترسی به تغییرات لحظه‌ای قوانین، آرای جدید وحدت رویه یا بخشنامه‌های اخیر را ندارند [۱۸]. این محدودیت باعث می‌شود که حتی پیشرفته‌ترین مدل‌ها نیز در پاسخگویی به سوالاتی که نیازمند دانش به‌روز یا جزئیات دقیق قوانین محلی هستند، دچار خطا شوند.

برای غلبه بر این محدودیت‌ها، رویکرد «تولید افزوده با بازیابی» (Retrieval-Augmented Generation) یا به عنوان یک راهکار موثر مطرح شده است. در این معماری، به جای تکیه صرف بر حافظه داخلی مدل، سیستم قادر است به یک پایگاه دانش خارجی و قابل اعتماد متصل شود و اطلاعات مرتبط را در لحظه بازیابی کند [۱۹]. این فرآیند دو مزیت کلیدی ایجاد می‌کند. نخست اینکه با فراهم کردن زمینه (Context) دقیق و مستند، احتمال توهم مدل را به شدت کاهش می‌دهد؛ و دوم اینکه امکان استدلال بر اساس آخرین تغییرات قانونی را بدون نیاز به آموزش مجدد مدل فراهم می‌سازد. در این فصل، به تشریح پیاده‌سازی این معماری برای بهبود پاسخ‌دهی به سوالات آزمون وکالت می‌پردازیم.

۱.۱.۰ پیش‌پردازش و قطعه‌بندی اسناد حقوقی

یکی از ارکان اصلی در طراحی کارآمد سیستم‌های بازیابی-محور، استراتژی تبدیل متون طولانی به قطعات قابل پردازش یا همان «چانک» (Chunk) است. در حالی که در بسیاری از کاربردهای عمومی پردازش زبان

طبیعی، استفاده از روش قطعه‌بندی با طول ثابت (Fixed-size Chunking) رایج است، بررسی‌های اولیه در این پژوهش نشان داد که اعمال این روش بر متون حقوقی منجر به چالش جدی «گسست معنایی» می‌شود [۲۰].

در متون حقوقی، هر واحد متنی (مانند یک ماده قانونی یا یک رأی) دارای یک ساختار منطقی به هم پیوسته است که در آن «حکم»، «موضوع» و «شرایط» (مانند تبصره‌ها و استثنائات) در کنار هم معنا پیدا می‌کنند. برش مکانیکی این متون صرفاً بر اساس تعداد کلمات، می‌تواند منجر به جدا شدن یک تبصره از ماده اصلی شود؛ اتفاقی که در مرحله بازیابی باعث می‌شود مدل زبانی تنها به بخشی از حقیقت دسترسی پیدا کند و در نتیجه استدلال حقوقی ناقص یا اشتباه ارائه دهد. مطالعات اخیر نشان داده‌اند که نادیده گرفتن ساختار محتوا در فرآیند قطعه‌بندی، باعث حذف اطلاعات حیاتی یا ورود اطلاعات نامرتبط به زمینه پردازش می‌شود [۲۱].

برای غلبه بر این چالش، در این پژوهش از رویکرد «قطعه‌بندی مبتنی بر ساختار» (Structure-Aware Chunking) استفاده شد. در این روش، به جای تعیین یک عدد ثابت برای طول متن، مرزهای طبیعی اسناد حقوقی ملاک عمل قرار گرفتند. این رویکرد هم‌راستا با یافته‌های پژوهش‌هایی است که نشان می‌دهند قطعه‌بندی معنایی (Semantic Chunking) با تقسیم اسناد به بخش‌های منسجم، کارایی سیستم‌های بازیابی را به شکل قابل توجهی بهبود می‌بخشد [۲۲]. بدین ترتیب، عملیات قطعه‌بندی بر اساس قواعد زیر پیاده‌سازی شد:

حفظ تمامیت ماده و تبصره‌ها: هر ماده قانونی به همراه تمامی تبصره‌های ذیل آن به عنوان یک چانک واحد در نظر گرفته شد تا یکپارچگی مفهوم حفظ شود.

واحدهای مستقل قضایی: در مورد آرای وحدت رویه و دادنامه‌ها، کل متن رأی به عنوان یک واحد پردازش شد تا سیاق کلام و استدلال قاضی در اثر برش متن از بین نرود.

حذف اطلاعات سربار: اطلاعات غیرمرتبط مانند تاریخ‌های اداری، شماره‌نامه‌های ارجاعی و سربرگ‌های تکراری که فاقد ارزش حقوقی برای پاسخ‌دهی بودند، در این مرحله پاک‌سازی شدند.

این رویکرد تضمین کرد که بردارهای تولید شده در مرحله بعد تعبیه‌ساز، نماینده‌ای دقیق از یک «مفهوم کامل حقوقی» باشند، نه صرفاً بازتابی از مجموعه‌ای از لغات که به صورت تصادفی در کنار هم قرار گرفته‌اند [۲۳].

۲.۱.۰ مدل تعبیه‌ساز

پس از قطعه‌بندی اسناد، گام بعدی تبدیل این چانک‌های متنی به بردارهای عددی است که قابلیت محاسبه شباهت معنایی را فراهم می‌کنند. انتخاب مدل تعبیه‌ساز مناسب یکی از تصمیم‌های حیاتی در طراحی سیستم‌های RAG است، زیرا کیفیت این بردارها به طور مستقیم بر دقت بازیابی اسناد مرتبط تأثیر می‌گذارد [۲۴].

در این پژوهش، پیش از تصمیم‌گیری نهایی، پنج مدل تعبیه‌ساز علاوه‌بر مدل bge-m3 مدل‌هایی نظیر

jina-embeddings-v2، qwen3-8b-instruct (به عنوان یک مدل تولیدی با قابلیت استخراج بردار)، و دو مدل دیگر که به طور اختصاصی روی متون فارسی تنظیم شده بودند، در یک مجموعه آزمایشی کوچک (شامل ۵۰ جفت سوال-سند حقوقی) مورد ارزیابی قرار گرفتند. معیار ارزیابی، دقت بازیابی اسناد مرتبط در پنج رتبه ابتدایی بود.

نتایج این ارزیابی نشان داد که مدل bge-m3 (BAAI General Embedding - Multilingual v3) از عملکرد برتری نسبت به سایر مدل‌ها برخوردار است. این مدل که به عنوان یک مدل چندزبانه طراحی شده است، توانایی بالایی در درک ساختارهای پیچیده و زمینه‌های تخصصی از جمله متون حقوقی دارد [۲۵]. دلیل برتری bge-m3 نسبت به مدل‌های bge-m3 شده روی فارسی را می‌توان به دو عامل نسبت داد، نخست حجم بالاتر داده‌های آموزشی چندزبانه که شامل متون تخصصی از زبان‌های مختلف بوده است؛ و دوم، معماری پیشرفته‌تر مدل که توانایی کدگذاری روابط معنایی عمیق‌تر را داراست [۲۶].

پس از انتخاب bge-m3، کل مجموعه اسناد حقوقی (شامل قوانین، آرای وحدت رویه و دادنامه‌ها) با استفاده از این مدل به فضای برداری با بعد ۱۰۲۴ تبدیل شدند. این بردارها سپس در یک پایگاه داده برداری (Vector Database) ذخیره شدند تا در مرحه بازیابی، امکان جستجوی سریع و کارآمد بر اساس شباهت کسینوسی (Cosine Similarity) فراهم شود [۲۷].

۳.۱.۰ معماری بازیابی ترکیبی

برای افزایش دقت و کارایی فرآیند بازیابی، در این پژوهش از یک معماری ترکیبی و چندمرحله‌ای بهره گرفته شد. این رویکرد، به جای اتکای صرف به یک روش واحد، چندین استراتژی مکمل را برای ساخت یک زمینه جامع و قابل اعتماد با هم ادغام می‌کند.

فرآیند بازیابی با یک مرحله فیلتر سریع بر اساس فراداده‌های ساختاریافته آغاز می‌شود [۲۸]. در این گام، پرسش ورودی برای شناسایی نشانه‌های صریح حقوقی مانند «نام قانون» و «شماره ماده یا اصل» تحلیل می‌شود. در صورت یافتن چنین عباراتی، فضای جستجو به شکل آنی به اسنادی که فراداده آن‌ها با این مشخصات تطابق دارد، محدود می‌گردد. این فیلتر اولیه، ضمن افزایش چشمگیر سرعت، با کاهش فضای جستجو، از بازیابی اسناد نامرتبط که ممکن است شامل واژگان مشابه اما با معانی متفاوت باشند، جلوگیری می‌کند.

پس از این مرحله فیلتر اولیه (یا در صورتی که پرسش فاقد هرگونه نشانه‌ی ساختاری باشد)، هسته اصلی بازیابی یعنی جستجوی معنایی فعال می‌شود. در این فاز، پرسش کاربر با استفاده از مدل bge-m3 به یک بردار عددی تبدیل شده و شباهت کسینوسی آن با تمامی بردارهای موجود در پایگاه داده سنجیده می‌شود. این رویکرد ترکیبی، که ابتدا از فیلترینگ ساختاریافته و سپس از جستجوی معنایی بهره می‌برد، تضمین می‌کند که سیستم هم قادر به درک دقیق عبارات تخصصی حقوقی است و هم می‌تواند مفاهیم بیان‌شده در زبان طبیعی و غیررسمی را به اسناد مرتبط پیوند دهد [۲۹].

در نهایت، برای اطمینان از جامعیت و تنوع شواهد حقوقی ارائه‌شده به مدل زبانی، یک مرحله تنظیم

و توازن منابع اجرا می‌گردد. از آنجا که یک استدلال حقوقی قوی نیازمند ارجاع همزمان به متن قانون و رویه قضایی است، اسناد نهایی صرفاً بر اساس بالاترین امتیاز شباهت انتخاب نمی‌شوند. در عوض، یک سیاست بازیابی متناسب اعمال می‌گردد که طی آن، از میان نامزدهای برتر، مجموعاً ۲۰ سند با ترکیبی مشخص انتخاب می‌شود: ۱۰ سند از مجموعه قوانین، ۵ سند از آرای وحدت رویه و ۵ سند از دادنامه‌ها. این استراتژی سهمیه‌بندی منابع تضمین می‌کند که زمینه نهایی، تصویری چندوجهی از چشم‌انداز حقوقی ارائه دهد و از اتکای بیش از حد به یک نوع منبع خاص جلوگیری کند [۳۰].

۴.۱.۰ مرحله رتبه‌بندی مجدد

پس از بازیابی اولیه اسناد توسط روش‌های مبتنی بر فراداده و جستجوی معنایی، چالش اساسی بعدی نحوه ارائه این اسناد به مدل زبانی است. تحقیقات نشان داده‌اند که مدل‌های زبانی بزرگ، در پردازش زمینه‌های طولانی (Long Contexts)، دچار یک سوگیری شناختی به نام پدیده «گمشده در میانه» (Lost in the Middle) می‌شوند. بر اساس این پدیده، دقت مدل در استفاده از اطلاعاتی که در ابتدا یا انتهای ورودی قرار دارند به مراتب بالاتر از اطلاعاتی است که در میانه متن دفن شده‌اند [۳۱]. بنابراین، اگر اسناد حاوی پاسخ صحیح در میان انبوهی از اسناد بازیابی‌شده (مثلاً در رتبه ۱۰ از ۲۰) قرار گیرند، احتمال نادیده گرفته شدن آن‌ها توسط مدل افزایش می‌یابد.

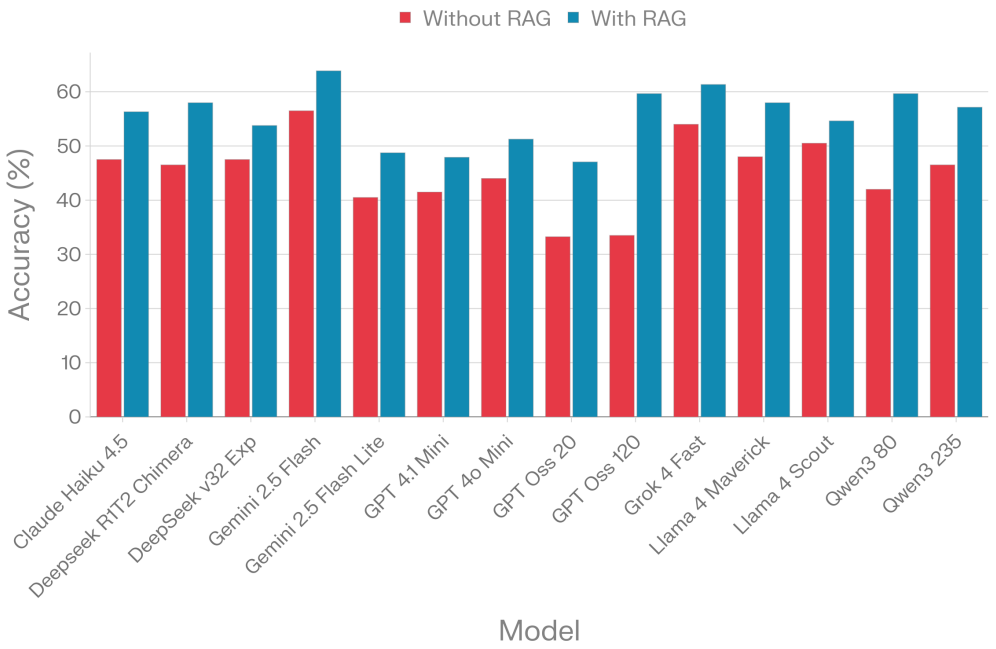
برای مقابله با این چالش و بهینه‌سازی کیفیت ورودی، در این پژوهش از یک مرحله «رتبه‌بندی مجدد» (Re-ranking) استفاده شد. در این گام، ۲۰ سند استخراج‌شده از مرحله قبل، به یک مدل تخصصی به نام rerank-multilingual-v3.0 از شرکت Cohere داده شدند. این مدل برخلاف مدل‌های تعبیه‌ساز که صرفاً شباهت معنایی کلی را می‌سنجند، قادر است ارتباط دقیق میان هر سند و پرسش را با در نظر گرفتن ظرایف زبانی و منطقی ارزیابی کند. مدل Re-ranker با امتیازدهی مجدد به هر سند، آن‌ها را به گونه‌ای مرتب می‌کند که اسناد دارای بیشترین ارتباط قطعی با سوال، در ابتدای لیست قرار گیرند [۳۲].

اعمال این فرآیند باعث می‌شود که حیاتی‌ترین شواهد حقوقی همواره در «پنجره توجه مؤثر» (Effective Attention Window) مدل زبانی قرار گیرند. نتایج تجربی نشان می‌دهد که افزودن مرحله Re-ranking به معماری RAG، نه تنها مشکل «گمشده در میانه» را کاهش می‌دهد، بلکه با حذف نویزهای باقی‌مانده از مرحله بازیابی اولیه، عملکرد نهایی سیستم در پاسخ‌دهی دقیق و مستند را به شکل معناداری بهبود می‌بخشد [۳۳].

۵.۱.۰ ارزیابی و تحلیل نتایج تجربی

در این بخش، به بررسی تأثیر به‌کارگیری معماری RAG بر عملکرد ۱۴ مدل زبانی منتخب در آزمون وکالت می‌پردازیم. همان‌طور که در فصل متدولوژی اشاره شد، تمامی مدل‌ها ابتدا با روش پایه Few-Shot CoT Explain-to-Predict (بدون دسترسی به دانش خارجی) و سپس با همان پرامپت اما مجهز به سیستم بازیابی طراحی‌شده (RAG)، مورد آزمون قرار گرفتند.

نتایج کمی به‌دست‌آمده نشان‌دهنده یک بهبود عملکرد سراسری و معنادار در تمامی مدل‌های مورد مطالعه است. میانگین دقت مدل‌ها پس از اعمال RAG به شکل قابل‌توجهی افزایش یافته است که نشانگر موفقیت سیستم در تزریق دانش تخصصی حقوقی به فرآیند استدلال مدل‌هاست [۳۴].



شکل ۱: شکل ۵-۱: مقایسه عملکرد ۱۴ مدل زبانی بزرگ در پاسخگویی به سوالات آزمون وکالت در دو سناریوی «پایه» (بدون دسترسی به دانش خارجی) و «با RAG». همان‌طور که مشاهده می‌شود، استفاده از معماری RAG منجر به بهبود دقت در تمامی مدل‌ها شده است، که این بهبود در مدل‌های متن‌باز با جهشی چشمگیر همراه بوده است.

همان‌طور که در نمودار فوق مشاهده می‌شود، مدل Gemini 2.5 Flash با ثبت دقت ۸۷٪.۶۳ (در مقایسه با ۷٪.۵۶ در حالت پایه) همچنان جایگاه خود را به عنوان دقیق‌ترین مدل حفظ کرده است. پس از آن، مدل Grok 4 Fast با دقت ۳۴٪.۶۱ در رتبه دوم قرار گرفته است. این نتایج نشان می‌دهد که مدل‌های پیشرو تجاری (Proprietary Models) که ذاتاً از توانایی استدلال قوی‌تری برخوردارند، توانسته‌اند از زمینه (Context) فراهم‌شده توسط RAG برای پالایش پاسخ‌های خود و کاهش توهمات حقوقی به بهترین نحو استفاده کنند.

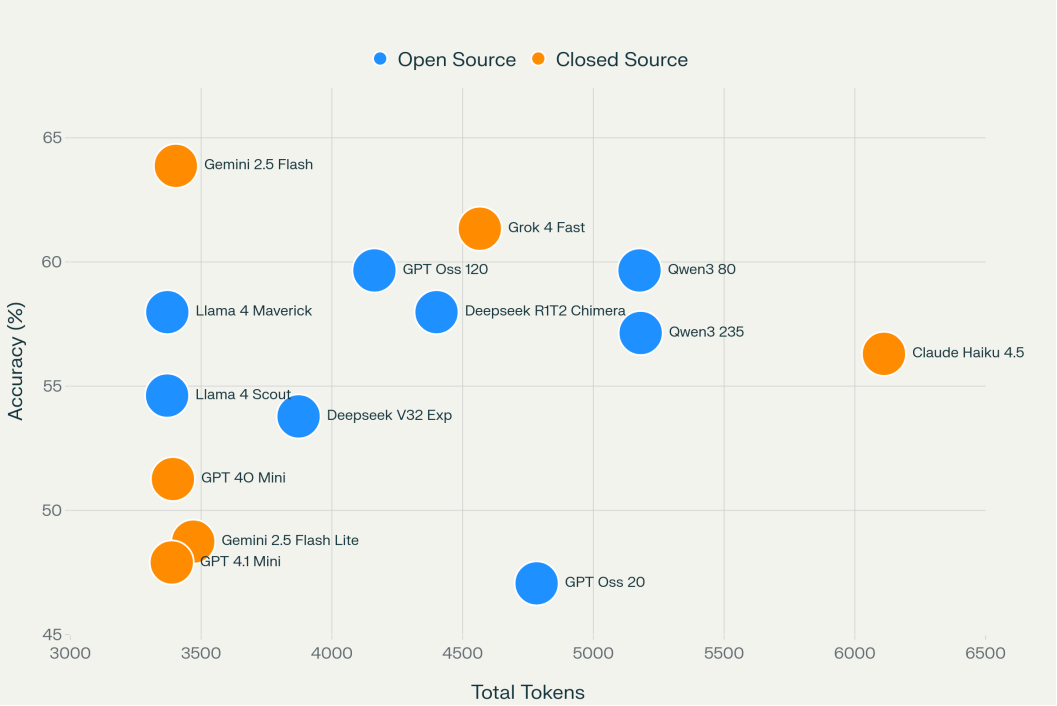
یکی از یافته‌های شگفت‌انگیز این پژوهش، رفتار متفاوت مدل‌های متن‌باز و کوچک‌تر در مواجهه با سیستم RAG است. در حالی که مدل‌های بزرگ تجاری بهبودهای ملایم‌تری (بین ۵ تا ۸ درصد) را تجربه کردند، برخی مدل‌های متن‌باز شاهد جهش‌های عملکردی خیره‌کننده بودند. بارزترین نمونه، مدل GPT OSS 120 است که دقت آن از ۳٪.۳۳ در حالت پایه به ۶۶٪.۵۹ با RAG افزایش یافته است؛ رشدی معادل

۳۶.۲۶ درصد. الگوی مشابهی در مدل Qwen3 80 نیز دیده می‌شود که با رشد ۱۶.۱۷ درصدی، خود را از رده مدل‌های ضعیف (۵٪.۴۲) به سطح مدل‌های رده‌بالا (۶۶٪.۵۹) رسانده است [۳۵].

این پدیده که می‌توان آن را «اثر شکاف دانش» (Knowledge Gap Effect) نامید، بیانگر دو نکته کلیدی است: نخست اینکه ضعف عملکرد مدل‌های متن‌باز در حالت پایه، نه ناشی از ناتوانی در استدلال (Reasoning)، بلکه عمدتاً ناشی از فقر دانش پارامتریک (Parametric Knowledge) آن‌ها در حوزه قوانین خاص ایران بوده است؛ و دوم اینکه با فراهم شدن دسترسی به اسناد حقوقی دقیق از طریق RAG، این مدل‌ها پتانسیل واقعی استدلال خود را آزاد کرده و توانسته‌اند حتی با مدل‌های بسیار بزرگ‌تر تجاری رقابت کنند.

در خانواده مدل‌های Llama، مدل Llama 4 Maverick با بهبود ۶۸.۹ درصدی به دقت ۹۸٪.۵۷ دست یافت که عملکردی قابل قبول و نزدیک به مدل‌های برتر است. در مقابل، مدل Llama 4 Scout کمترین میزان بهبود (+۸۲٪.۳) را در میان تمامی مدل‌ها ثبت کرد. این تفاوت عملکرد در یک خانواده مدل، احتمالاً به تفاوت در «پنجره توجه» (Context Window) و نحوه مدیریت متون طولانی در معماری داخلی این دو نسخه بازمی‌گردد؛ به طوری که نسخه Scout احتمالاً در پردازش همزمان ۲۰ سند بازیابی‌شده و استخراج اطلاعات مفید از آن‌ها، کارایی کمتری نسبت به نسخه Maverick داشته است.

علاوه بر دقت نهایی، بررسی شاخص‌های فرآیندی مانند «طول توضیحات» (Explanation Length) و «تعداد توکن‌های مصرفی» (Token Usage)، بینش عمیقی درباره تغییر رفتار مدل‌ها پس از دسترسی به دانش خارجی ارائه می‌دهد. داده‌های جدول فوق نشان می‌دهد که تزریق دانش از طریق RAG تأثیر معناداری بر عمق استدلال مدل‌ها داشته است. مقایسه میانگین طول توضیحات در پاسخ‌های صحیح (ExplainLenMean_Correct1) نشان می‌دهد که در اکثر مدل‌ها، استفاده از RAG منجر به تولید توضیحات طولانی‌تر و غنی‌تر شده است.



شکل ۲: شکل ۵-۲: تحلیل کارایی توکن در برابر دقت مدل‌ها (Token Efficiency vs. Accuracy). محور افقی نشان‌دهنده میانگین تعداد کل توکن‌های مصرفی (Total Tokens) و محور عمودی بیانگر دقت مدل (Accuracy) است. نقاط رنگی موقعیت هر مدل را در فضای هزینه-عملکرد نشان می‌دهند.

به عنوان مثال، در مدل Claude Haiku 4.5، میانگین طول توضیحات صحیح از ۰۵.۴۷ توکن (در حالت پایه) به ۵۴.۵۷ توکن (با RAG) افزایش یافته است. این افزایش ۱۰ واحدی، همبستگی مستقیمی با افزایش دقت این مدل (از ۵٪.۴۷ به ۳٪.۵۶) دارد، که نشان می‌دهد مدل از اطلاعات اضافی برای ساخت استدلال‌های دقیق‌تر استفاده کرده است [۸]. نکته جالب توجه دیگر در مدل برتر Gemini 2.5 Flash است. با وجود اینکه این مدل بالاترین دقت (۸۷٪.۶۳) را دارد، میانگین طول توضیحات آن (۲۲.۳۶ توکن) نسبت به بسیاری از مدل‌های دیگر (مانند Llama 4 Scout با ۲.۵۸ توکن) کمتر است. این نشان می‌دهد که Gemini توانایی بالایی در «استخراج و خلاصه‌سازی» اطلاعات دارد و به جای بازگو کردن کل متن بازیابی‌شده، مستقیماً به استدلال اصلی می‌پردازد.

تحلیل سربار محاسباتی (Computational Overhead) نیز نشان می‌دهد که استفاده از RAG به طور طبیعی حجم ورودی (Prompt Tokens) را افزایش می‌دهد. داده‌ها نشان می‌دهند که میانگین توکن‌های ورودی (PromptTok_Correct1) از حدود ۱۴۰۰-۱۵۰۰ توکن در حالت پایه به بیش از ۳۲۰۰ توکن در حالت RAG رسیده است (افزایش بیش از ۲ برابری). مدل‌های خانواده Qwen3 (نسخه‌های ۸۰ و ۲۳۵) با مصرف حدود ۵۰۰۰ توکن ورودی، بیشترین سربار را داشته‌اند. با این حال، جهش ۱۷ درصدی دقت در Qwen3 80 نشان می‌دهد که این مدل توانسته است از این حجم بالای اطلاعات به نحو احسن استفاده کند. در مقابل،

مدل‌هایی مثل GPT 4o Mini با مصرف حدود ۳۲۰۰ توکن ورودی و حفظ دقت بالای ٪۵۱، تعادل خوبی بین هزینه (تعداد توکن) و عملکرد برقرار کرده‌اند که برای کاربردهای عملی و تجاری حائز اهمیت است [۳۶].

یک مشاهده جالب دیگر در داده‌ها این است که در اکثر مدل‌ها (مثلاً Claude Haiku و DeepSeek R1T2)، میانگین طول توضیحات در پاسخ‌های غلط (ExplainLenMean_Correct0) بیشتر از پاسخ‌های صحیح است. به عنوان مثال، در Claude Haiku با RAG، طول توضیح غلط ۹۶.۶۴ توکن در برابر ۵۴.۵۷ توکن برای پاسخ صحیح است. این پدیده که در ادبیات تحقیق به عنوان «توهم پرگویانه» (Verbose Hallucination) شناخته می‌شود، نشان می‌دهد که مدل‌ها وقتی در رسیدن به پاسخ صحیح ناتوان هستند، سعی می‌کنند با تولید متن بیشتر و پیچیده‌تر، عدم اطمینان خود را پنهان کنند یا در مسیر استدلال گم می‌شوند. RAG در برخی موارد با ارائه اطلاعات متناقض یا نامرتبط، می‌تواند این پرگویی را تشدید کند.