فهرست جدول ها

۲۶.....جدول ۲-۱: ویژگی‌ها و نتایج گزارش‌شده در ۲۷ مقالهٔ موجود در پوشهٔ پژوهش
۳۱ .....جدول ۲-۲: دسته‌های اصلی مطالعات و ارتباط آن‌ها با طراحی پژوهش حاضر

فهرست نمودار ها

فهرست شکل ها

شکل ۲-۱: ارتباط ۲۷ مقالهٔ بررسی‌شده با یکدیگر و جایگاه پژوهش حاضر ۲۵

چکیده

گسترش سامانه‌های توزیع‌شده حجم بزرگی از لاگ‌های ناهمگون را پدید آورده است و تشخیص به‌موقع ناهنجاری در این داده‌ها برای پایداری و امنیت سامانه‌ها ضرورت دارد. روش‌های متعارف اغلب به قواعد دست‌نویس، الگوهای ثابت یا داده‌های برچسب‌خورده وابسته‌اند و در برابر تغییر قالب لاگ و افزایش حجم داده با محدودیت روبه‌رو می‌شوند. این پژوهش چارچوبی برای طبقه‌بندی دودویی پیام‌های مجموعه‌داده بی‌جی‌ال با یک مدل زبانی بزرگ محلی ارائه می‌کند. در این چارچوب، برچسب مرجع پیش از تشکیل ورودی حذف می‌شود و مقادیر پویایی مانند شناسه‌ها، مسیرها، نشانی‌های شبکه، مقادیر شانزده‌شانزدهی و اعداد با جانشین‌های ثابت جایگزین می‌شوند؛ بااین‌حال، تفاوت‌های معناداری مانند صفر و غیرصفر حفظ می‌گردد. مدل کوئن ۲٫۵ نسخهٔ هفت‌میلیاردپارامتری دستورمحور به‌کمک اولاما و با خروجی ساخت‌یافته اجرا می‌شود. برای جلوگیری از فراخوانی‌های تکراری، تصمیم‌های معتبر در سطح قالب نرمال‌شده در حافظهٔ نهان نگه‌داری می‌شوند. سامانه در دو حالت «ترکیبی» و «صرفاً مبتنی بر دستور» اجرا می‌شود و کیفیت تصمیم، نرخ پاسخ نامعتبر، زمان اجرا و کارایی حافظهٔ نهان در هر دو حالت سنجیده می‌شود. این طراحی امکان ارزیابی محلی، بازتولیدپذیر و خط‌به‌خط را بدون تنظیم دقیق وزن‌های مدل فراهم می‌کند.

کلمات کلیدی(فارسی)

تشخیص ناهنجاری، لاگ سیستمی، مهندسی دستور، مدل زبانی بزرگ، مجموعه‌داده بی‌جی‌ال

کلمات کلیدی(انگلیسی)

Anomaly Detection, System Logs, Prompt Engineering, Large Language Models (LLMs)

فصل اول: کلیات تحقیق

۱-۱ مقدمه

سامانه‌های نرم‌افزاری و زیرساخت‌های توزیع‌شده به‌طور پیوسته حجم زیادی از لاگ‌های سیستمی[۱] تولید می‌کنند. این داده‌ها وضعیت اجزای سامانه، خطاهای اجرایی و رخدادهای امنیتی را ثبت می‌کنند و از مهم‌ترین منابع پایش و عیب‌یابی به‌شمار می‌روند. با افزایش حجم داده و دگرگونی مداوم قالب پیام‌ها، بررسی دستی یا اتکا به قواعد ثابت پاسخ‌گو نیست. مدل‌های زبانی بزرگ[۲]، به سبب توانایی در پردازش متن‌های نیمه‌ساخت‌یافته، می‌توانند بخشی از معنای پیام را بدون طراحی ویژگی‌های فراوان استخراج کنند. در این پژوهش از مهندسی دستور[۳] برای بهره‌گیری از این توانایی، بدون بازآموزی یا تنظیم دقیق مدل، استفاده می‌شود. این فصل مسئله، ضرورت، اهداف، پرسش‌ها و فرضیه‌های پژوهش را متناسب با همین دامنه بیان می‌کند.

۱-۲ بیان مسئله

در سامانه‌های توزیع‌شده، لاگ‌ها به‌صورت پیوسته تولید می‌شوند و اطلاعات ضروری برای عیب‌یابی، پایش سلامت و بررسی رخدادهای امنیتی را در خود دارند. تشخیص ناهنجاری در چنین جریانی با سه دشواری اصلی روبه‌رو است: پیام‌ها اغلب نیمه‌ساخت‌یافته‌اند؛ حجم داده بسیار زیاد است؛ و نمونه‌های ناهنجار در مقایسه با رخدادهای عادی سهم اندکی دارند. افزون بر این، بخشی از هر پیام از مقادیر پویایی مانند شناسهٔ گره، مسیر و عدد تشکیل شده است. اگر این مقادیر بدون پردازش وارد مدل شوند، پیام‌های هم‌معنا به صورت نمونه‌های متفاوت دیده می‌شوند؛ و اگر بیش‌ازحد حذف شوند، نشانه‌های معنادار نیز از بین می‌روند. مجموعه‌داده بی‌جی‌ال[۴] نمونه‌ای مناسب برای بررسی هم‌زمان این مسائل در مقیاس بزرگ است.

مدل‌های زبانی پیش‌آموزش‌دیده و مدل‌های زبانی بزرگ می‌توانند رابطهٔ میان واژه‌ها و زمینهٔ رخداد را از متن استخراج کنند. بااین‌حال، استفاده از آن‌ها برای لاگ لزوماً به معنای حذف پیش‌پردازش نیست؛ زیرا کیفیت پاک‌سازی، نرمال‌سازی و صورت‌بندی دستور، مستقیماً بر ورودی مدل و در نتیجه بر تصمیم آن اثر می‌گذارد.

---

[۱] System Logs
[۲] Large Language Model (LLM)
[۳] Prompt Engineering
[۴] Blue Gene/L (BGL)

پژوهش‌های پیشین از مدل زبانی در نقش‌های گوناگونی استفاده کرده‌اند: تجزیهٔ پیام و استخراج قالب، مدل‌سازی توالی رخداد، طبقه‌بندی مستقیم، بازیابی نمونهٔ مشابه و تولید توضیح. تفاوت این نقش‌ها نشان می‌دهد که صرف ذکر نام مدل برای مقایسهٔ روش‌ها کافی نیست و باید واحد تحلیل، نوع آموزش، شکل ورودی، سازوکار تصمیم و هزینهٔ اجرا نیز روشن باشد.

با توجه به این ملاحظات، مسئلهٔ پژوهش حاضر در سه محور صورت‌بندی می‌شود:

نخست، اجرای یک مدل زبانی بزرگ برای میلیون‌ها پیام هزینه‌بر است. حتی در استقرار محلی، هر فراخوانی به زمان پردازش و منابع سخت‌افزاری نیاز دارد. ازاین‌رو، باید پیام‌های هم‌قالب شناسایی شوند و تصمیم‌های معتبر تا حد امکان دوباره به‌کار روند، بی‌آنکه نتیجهٔ یک نسخهٔ قدیمی مدل یا دستور به اجرای جدید منتقل شود.

دوم، دانش قطعی دامنه را می‌توان بیرون مدل یا در متن دستور به‌کار گرفت. در مسیر نخست، یک نگهبان قاعده‌محور نمونه‌های روشن را پیش از فراخوانی مدل تعیین تکلیف می‌کند؛ در مسیر دوم، همان دانش در اختیار مدل قرار می‌گیرد تا همهٔ تصمیم‌ها را خود مدل صادر کند. مقایسهٔ منصفانهٔ این دو مسیر مستلزم ثابت‌ماندن داده، مدل، نرمال‌سازی، حافظهٔ نهان و قرارداد خروجی است.

سوم، پاسخ مولد ممکن است از قالب تعیین‌شده پیروی نکند یا در اجراهای مختلف تغییر کند. در یک سامانهٔ خودکار باید میان خطای طبقه‌بندی و پاسخ نامعتبر تمایز گذاشت. بنابراین پاسخ مدل باید با یک قرارداد ساخت‌یافته سنجیده شود و پاسخ‌های نامعتبر نه به‌عنوان عادی یا ناهنجار تعبیر شوند و نه در حافظهٔ نهان قرار گیرند.

برای بررسی این مسئله، مدل کوئن ۲٫۵ نسخهٔ هفت‌میلیاردپارامتری دستورمحور[۵] با وزن‌های ثابت و از طریق اولاما[۶] اجرا می‌شود. دو روش «ترکیبی»[۷] و «صرفاً مبتنی بر دستور»[۸] تنها در محل اعمال دانش قاعده تفاوت دارند. نرمال‌سازی، حافظهٔ نهان[۹] و خروجی ساخت‌یافتهٔ جی‌سان[۱۰] در هر دو روش یکسان نگه داشته می‌شوند تا منشأ تفاوت نتایج قابل تشخیص باشد.

واحد تصمیم در این پژوهش یک خط از لاگ بی‌جی‌ال است. برچسب مرجع پیش از ساخت ورودی حذف می‌شود و تنها برای ارزیابی به‌کار می‌رود. افزون بر معیارهای کیفیت طبقه‌بندی، تعداد فراخوانی مدل، نرخ استفاده از حافظهٔ نهان، زمان پاسخ، منبع هر تصمیم و تعداد پاسخ‌های نامعتبر نیز ثبت می‌شود.

برآیند مورد انتظار، یک فرایند روشن و قابل بازتولید برای سنجش توان مدل زبانی محلی در تشخیص ناهنجاری خط‌به‌خط است؛ فرایندی که هم کیفیت تصمیم و هم هزینهٔ عملیاتی آن را نشان دهد.

چنین ارزیابی‌ای می‌تواند مشخص کند که نرمال‌سازی و بازاستفاده از تصمیم‌های سطح قالب تا چه اندازه برای تحلیل مجموعه‌های بزرگ لاگ سودمند است و کدام محل اعمال دانش قاعده برای این مسئله مناسب‌تر عمل می‌کند.

---

[۵] Qwen۲٫۵:۷B-Instruct
[۶] Ollama
[۷] Hybrid
[۸] Prompt-only
[۹] Cache
[۱۰] JavaScript Object Notation (JSON)

۱-۳ اهمیت و ضرورت انجام پژوهش

سامانه‌های نرم‌افزاری در حوزه‌هایی مانند مالی، سلامت، مخابرات و خدمات ابری به پایش پیوسته وابسته‌اند. تأخیر در شناسایی رخداد غیرعادی می‌تواند به توقف خدمت، افزایش هزینهٔ نگه‌داری یا گسترش یک رخداد امنیتی منجر شود. ازاین‌رو، تحلیل خودکار لاگ نه‌تنها یک مسئلهٔ پژوهشی، بلکه یک نیاز عملیاتی است.

اهمیت علمی پژوهش در تفکیک عواملی است که معمولاً به‌صورت هم‌زمان تغییر می‌کنند. در بسیاری از مطالعات، نوع مدل، واحد تحلیل، شیوهٔ آماده‌سازی و روش کنترل هزینه با یکدیگر متفاوت‌اند و همین امر تفسیر نتایج را دشوار می‌کند. در این پایان‌نامه، دو مسیر تصمیم تحت شرایط یکسان اجرا می‌شوند تا اثر محل اعمال دانش قاعده از اثر سایر اجزا جدا شود.

از دیدگاه کاربردی، چارچوب پیشنهادی می‌تواند پیامدهای زیر را داشته باشد:

- کاهش فراخوانی‌های تکراری مدل از طریق نرمال‌سازی و حافظهٔ نهان؛
- امکان اجرای محلی و جلوگیری از ارسال لاگ‌های حساس به خدمات بیرونی؛
- ثبت خط‌به‌خط منبع تصمیم و فراهم‌کردن امکان ممیزی و بازتولید آزمایش.

این ویژگی‌ها به‌ویژه برای سازمان‌هایی اهمیت دارند که هم با حجم زیاد لاگ روبه‌رو هستند و هم محدودیت محرمانگی یا منابع محاسباتی دارند.

بنابراین ضرورت پژوهش از یک‌سو به نیاز عملی برای تحلیل سریع و محلی لاگ و از سوی دیگر به نیاز علمی برای مقایسهٔ کنترل‌شدهٔ دو محل اعمال دانش دامنه بازمی‌گردد.

۱-۴ اهداف مشخص پژوهش

الف) هدف اصلی

طراحی و ارزیابی چارچوبی محلی برای تشخیص ناهنجاری خط‌به‌خط در لاگ‌های بی‌جی‌ال با مدل کوئن ۲٫۵ دستورمحور و مقایسهٔ کنترل‌شدهٔ روش ترکیبی با روش صرفاً مبتنی بر دستور، بدون تنظیم دقیق مدل.

ب) اهداف فرعی

۱. طراحی قواعد نرمال‌سازی برای تبدیل مقادیر پویا به قالب‌های پایدار، همراه با حفظ نشانه‌های معنادار؛
۲. پیاده‌سازی حافظهٔ نهان نسخه‌بندی‌شده در سطح قالب و سنجش اثر آن بر تعداد فراخوانی و زمان اجرا؛

۳. سنجش صحت[۱۱]، دقت[۱۲]، بازخوانی[۱۳] و امتیاز اف‌یک[۱۴] در دو مسیر تصمیم و گزارش ماتریس اغتشاش در سطح خط؛

٤. اندازه‌گیری نرخ پاسخ نامعتبر، زمان پاسخ و سهم هر منبع تصمیم در نتیجهٔ نهایی؛

٥. ثبت تنظیمات مدل، دستور، کلید حافظهٔ نهان و خروجی خطبه‌خط برای افزایش بازتولیدپذیری.

۱-۵ سؤال‌ها و فرضیه‌های پژوهش

سؤال‌های پژوهش

۱. مدل کوئن ۲٫۵ دستورمحور، بدون تنظیم دقیق، تا چه اندازه می‌تواند پیام‌های بی‌جی‌ال را در سطح خط طبقه‌بندی کند؟

۲. نرمال‌سازی و حافظهٔ نهان سطح قالب چه اثری بر تعداد فراخوانی مدل و زمان پردازش دارند؟

۳. روش ترکیبی و روش صرفاً مبتنی بر دستور از نظر معیارهای کیفیت و نرخ پاسخ نامعتبر چه تفاوتی دارند؟

٤. استقرار محلی مدل با اولاما تا چه اندازه امکان اجرای قابل بازتولید و حفظ محرمانگی داده را فراهم می‌کند؟

فرضیه‌های پژوهش

۱. مدل کوئن ۲٫۵ دستورمحور، بدون تنظیم دقیق، قادر است بخشی از ناهنجاری‌های بی‌جی‌ال را در سطح خط تشخیص دهد.

۲. نرمال‌سازی و حافظهٔ نهان سطح قالب تعداد فراخوانی‌های تکراری و زمان پردازش را کاهش می‌دهند.

۳. میان روش ترکیبی و روش صرفاً مبتنی بر دستور از نظر کیفیت تصمیم و نرخ پاسخ نامعتبر تفاوت قابل اندازه‌گیری وجود دارد.

٤. استقرار محلی با اولاما، در صورت ثبت نسخه‌ها و تنظیمات، بازتولیدپذیری و کنترل داده را نسبت به خدمت بیرونی افزایش می‌دهد.

---

[۱۱] Accuracy
[۱۲] Precision
[۱۳] Recall
[۱۴] F۱-score

فصل دوم: مروری بر ادبیات و پیشینه تحقیق