Spieglein, Spieglein and der Wand, welche KI ist die beste im Cybersecurity-Land?

Sophos-Experten erstellen ein neues Benchmark-System zur Einschätzung des Nutzens verschiedener Sprachmodellen aus dem Bereich Maschinelles Lernen, sogenannte Large-Language-Modelle (LLMs), für die Cybersicherheit. Ein alter Bekannter schneidet dabei am besten an.

Die Technologie des maschinellen Lernens mit großen Sprachmodellen (LLM) verbreitet sich rasant, wobei mittlerweile mehrere konkurrierende Open-Source- und proprietäre Architekturen verfügbar sind. Zusätzlich zu den generativen Textaufgaben, die mit Plattformen wie ChatGPT verbunden sind, haben sich LLMs nachweislich in vielen Textverarbeitungsanwendungen als nützlich erwiesen – von der Unterstützung beim Schreiben von Code bis zur Kategorisierung von Inhalten.

Angesichts der Vielfalt an zur Verfügung stehenden LLMs stehen Forscher vor einer herausfordernden Frage: Wie lässt sich ermitteln, welches Modell für ein bestimmtes Problem des maschinellen Lernens am besten geeignet ist? SophosAI hat eine Reihe von Möglichkeiten untersucht, LLMs bei Aufgaben im Zusammenhang mit der Cybersicherheit einzusetzen. Eine gute Methode zur Auswahl eines Modells besteht darin, Benchmark-Aufgaben zu erstellen – typische Probleme, mit denen sich die Fähigkeiten des Modells einfach und schnell beurteilen lassen.

Allerdings spiegelt die Leistung bei den momentan verfügbaren, allgemeinen Aufgaben für Benchmarks möglicherweise nicht genau wider, wie gut Modelle im Kontext der Cybersicherheit funktionieren. Durch die Verallgemeinerung offenbaren sie möglicherweise keine Unterschiede im sicherheitsspezifischen Fachwissen zwischen Modellen, die sich aus ihren Trainingsdaten ergeben. Aus diesem Grund habe die Forscher aus dem SophosAI-Team drei neue Benchmarks erstellt, die auf Aufgaben basieren, die potenziell grundlegende Voraussetzungen für die meisten LLM-basierten defensiven Cybersicherheitsanwendungen sind:

1. Das LLM fungiert als Assistent bei der Untersuchung von Vorfällen, indem es Fragen zur Telemetrie in natürlicher Sprache in SQL-Anweisungen umwandelt
2. Das LLM generiert Vorfallzusammenfassungen aus Daten eines Security Operations Centers (SOC).
3. Das LLM bewertet den Schweregrad des Vorfalls.

GPT-4 mit bester Leistung

Diese Benchmarks dienen zwei Zwecken: der Identifizierung grundlegender Modelle mit Potenzial für eine Feinabstimmung und der anschließenden Bewertung der standardmäßigen (nicht abgestimmten) Leistung dieser Modelle. Insgesamt hat das Sophos-AI-Team 14 Modelle auf Basis von Kriterien wie Modellgröße, Beliebtheit, Kontextgröße und Aktualität ausgewählt und anhand der Benchmarks getestet – darunter unterschiedlich große Versionen der Modelle LlaMa2 und CodeLlaMa von Meta, Amazon-Titan-Large und natürlich auch der Branchenprimus GPT-4. Das OpenAI-Tool zeigte bei den ersten beiden Aufgaben eindeutig die beste Leistung. Interessant: beim letzten Benchmark schnitt keines der Modelle bei der Kategorisierung der Schwere des Vorfalls genau genug ab, um besser zu sein als die Zufallsauswahl.

Aufgabe 1: Assistent für die Untersuchung von Vorfällen
Bei der ersten Benchmark-Aufgabe bestand das Hauptziel darin, die Leistung von LLMs als SOC-Analystenassistenten bei der Untersuchung von Sicherheitsvorfällen zu bewerten. Dabei mussten sie relevante Informationen auf der Grundlage von Abfragen in natürlicher Sprache abrufen – eine Aufgabe, mit der das SophosAI-Team bereits experimentiert hatte. Die Bewertung der Fähigkeit von LLMs, Abfragen in natürlicher Sprache unterstützt durch kontextbezogene Schemakenntnisse in SQL-Anweisungen umzuwandeln, hilft dabei, ihre Eignung für diese Aufgabe zu bestimmen.

Aufgabe 2: Zusammenfassung des Vorfalls
In Security Operations Centern (SOCs) untersuchen Bedrohungsanalysten täglich zahlreiche Sicherheitsvorfälle. In der Regel werden diese Vorfälle als eine Abfolge von Ereignissen dargestellt, die auf einem Benutzerendpunkt oder Netzwerk im Zusammenhang mit erkannten verdächtigen Aktivitäten aufgetreten sind. Bedrohungsanalysten nutzen diese Informationen, um weitere Untersuchungen durchzuführen. Allerdings kann diese Abfolge von Ereignissen für die Analysten oft sehr unübersichtlich sein, was es schwierig macht, die bemerkenswerten Ereignisse zu identifizieren. Hier können große Sprachmodelle wertvoll sein, da sie bei der Identifizierung und Organisation von Ereignisdaten auf der Grundlage einer bestimmten Vorlage helfen können und es Analysten erleichtern, das Geschehen zu verstehen und ihre nächsten Schritte festzulegen.

Aufgabe 3: Bewertung der Schwere des Vorfalls
Die dritte von Sophos bewertete Benchmark-Aufgabe war eine modifizierte Version eines traditionellen ML-Sec-Problems: Die Bestimmung, ob ein beobachtetes Ereignis Teil einer harmlosen Aktivität oder eines Angriffs ist. SophosAI verwendet spezielle Machine-Learning (ML)-Modelle, die für die Auswertung bestimmter Arten von Ereignisartefakten wie tragbare ausführbare Dateien und Befehlszeilen entwickelt wurden. Bei dieser Aufgabe bestand das Ziel darin, festzustellen, ob ein LLM eine Reihe von Sicherheitsereignissen untersuchen und deren Schweregrad beurteilen kann.

Noch einige Leitplanken für LLMs in der Cybersicherheit nötig
Auch wenn der Benchmark-Test sicherlich nicht alle potenziellen Problemstellungen berücksichtigen kann, lässt sich nach Abschluss der Untersuchung feststellen, dass LLMs die Bedrohungssuche und die Untersuchung von Vorfällen wirksam unterstützen und damit als sinnvoller SOC-Assistent eingesetzt werden können – allerdings sind weiterhin einige Leitplanken und Anleitungen notwendig.

Bei der Zusammenfassung von Vorfallinformationen aus Rohdaten erbringen die meisten LLMs eine ausreichende Leistung, es gibt jedoch Raum für Verbesserungen durch Feinabstimmung. Allerdings bleibt die Bewertung einzelner Artefakte oder Gruppen von Artefakten eine herausfordernde Aufgabe für vorab trainierte und öffentlich verfügbare LLMs. Um dieses Problem anzugehen, ist möglicherweise ein spezialisiertes LLM erforderlich, das speziell auf Cybersicherheitsdaten geschult ist.

Alle Details zu den Untersuchungsergebnissen und der Benchmark-Erstellung können im englischen SophosAI-Blogbeitrag „Benchmarking the Security Capabilities of Large Language Models“ nachgelesen werden.

Social Media von Sophos für die Presse
Wir haben speziell für Sie als Journalist*in unsere Social-Media-Kanäle angepasst und aufgebaut. Hier tauschen wir uns gerne mit Ihnen aus. Wir bieten Ihnen Statements, Beiträge und Meinungen zu aktuellen Themen und natürlich den direkten Kontakt zu den Sophos Security-Spezialisten.

Folgen Sie uns auf und

LinkedIn: https://www.linkedin.com/groups/9054356/
X/Twitter: @sophos_info

Pressekontakt:

Sophos
Jörg Schindler, PR-Manager Central & Eastern Europe
joerg.schindler@sophos.com, +49-721-25516-263

TC Communications
Arno Lücht, +49-8081-954619
Thilo Christ, +49-8081-954617
Ulrike Masztalerz, +49-30-55248198
Ariane Wendt +49-172-4536839
sophos@tc-communications.de