УДК 004.8: 004.891: 004.49: 004.56

Кібернетичні загрози великих мовних моделей

В.І. Слюсар
Центральний науково-дослідний інститут озброєння та військової техніки Збройних Сил України, м.Київ, Україна

Робота присвячена дослідженню потенціалу та викликів, пов'язаних з великими мовними моделями (LLM) у сфері штучного інтелекту (ШІ). Розглядаються архітектури та принципи роботи LLM, зокрема GPT-4 та її альтернатив. Аналізуються можливості застосування LLM у різних сферах, а також потенційні ризики і виклики, особливо у контексті кібербезпеки.

Вступ. Великі мовні моделі (Large Language Models, LLM) - це потужні алгоритми машинного навчання, здатні генерувати та обробляти тексти природною людською мовою [1, 2]. Ці моделі базуються на нейронних мережах з архітектурою transformer або informer-transformer і навчаються на великих обсягах текстових даних. Натреновані LLM здатні вирішувати широкий спектр завдань з обробки тексту: генерацію відповідей на запитання, створення зв'язних текстів на задану тему, переклад між мовами, підсумовування документів тощо. Все це робить їх надзвичайно корисним інструментом для різноманітних застосувань. Однак, разом з величезним потенціалом, вони несуть і певні ризики, особливо у сфері кібербезпеки.

Викладення основного матеріалу

Еволюція LLM відбувалася стрімкими темпами

- від ранніх моделей на кшталт BERT та GPT-1 до сучасних GPT-4, ChatGPT та їх численних альтернатив. Ця еволюція супроводжувалася постійним зростанням розміру моделей, обсягів даних для навчання та обчислювальних потужностей, необхідних для їх роботи. Хоча GPT-4 від OpenAI [3, 4] є однією з найвідоміших та найпотужніших LLM на сьогодні, існує ціла низка альтернативних моделей, розроблених іншими компаніями та дослідницькими групами. Серед них варто відзначити: Gemini 1.5 Pro та Ultra, Claude 3 Opus, Xinghuo 3.5 та інші, а також локальні відкриті LLM: Llama від Meta AI, Vicuna, Koala, Falcon, MPT, Mistral тощо. Ці моделі мають різні архітектури, розміри та особливості, але загалом націлені на вирішення схожих завдань з обробки природної мови. Деякі з них, як Claude 3 Opus, позиціонуються як більш ефективні та контрольовані альтернативи GPT-4.

Однією з ключових проблем використання LLM є те, що вони обмежені знаннями, закладеними в них під час навчання. Для вирішення цієї проблеми було створено спеціальні бібліотеки та фреймворки, що дозволяють інтегрувати LLM з зовнішніми джерелами даних. Яскравим прикладом такого рішення є бібліотека LangChain [2]. Вона надає зручний інтерфейс для підключення LLM до різноманітних джерел інформації - документів, баз даних, API тощо. За допомогою LangChain можна створювати більш динамічні та інформативні системи на основі LLM, які здатні оперувати актуальними даними та надавати відповіді з урахуванням контексту.

Інтеграція з зовнішніми джерелами відкриває нові можливості для застосування LLM. Наприклад, можна створити чат-бот, який буде надавати інформацію про продукти компанії, використовуючи дані з її каталогу та бази знань, або ж персонального асистента, який матиме доступ до особистих документів користувача і зможе допомагати в їх аналізі та обробці.

Поряд з гігантськими LLM, які містять мільярди параметрів, активно розвиваються так звані малі мовні моделі (Small Language Models, SLM). Вони мають значно менший розмір (зазвичай до кількох мільярдів параметрів) і можуть працювати локально на пристроях користувачів без необхідності звернення до хмарних сервісів. Прикладами SLM є:

- Gemini Nano - зменшена версія моделі Gemini, яка може працювати на смартфонах під управлінням Android;
- StableLM-Zephyr-3B - серія компактних моделей від Stability AI розміром від 3 до 8 мільярдів параметрів;
- Microsoft Phi-2 - серія SLM від Microsoft, оптимізованих для роботи на мобільних пристроях;
- Obsidian-3B - компактна модель аналізу зображень розміром 3 мільярди параметрів.

Існує також ціла низка SLM розміром до 1 мільярда параметрів. Серед них варто відзначити серію моделей TinyLlama-1.1B TinyVicuna-1B, TinyAlpaca-v0.1, LiteLlama-460M-1T, Smol-Llama-220M та Smol-Llama-101M-Chat-v1. Ці моделі оптимізовані для роботи на пристроях з обмеженими обчислювальними ресурсами і можуть

використовуватися для створення персональних асистентів, чат-ботів тощо.

Перевагами SLM є більша конфіденційність (дані не передаються на сервери компаній), швидкість роботи та можливість персоналізації. З іншого боку, вони зазвичай поступаються великим LLM за якістю та різноманітністю генерованих текстів.

Незважаючи на всі переваги та можливості LLM/SLM, їх активний розвиток та впровадження несе з собою і певні загрози в кібердомені. Однією з ключових проблем є конфіденційність та безпека даних. Під час навчання LLM можуть запам'ятовувати конфіденційну інформацію з навчальних текстів і потім випадково розкривати її у згенерованих відповідях. Крім того, дані, які користувачі надсилають LLM під час взаємодії, можуть зберігатися та аналізуватися компаніями без належного повідомлення та згоди.

Інша загроза полягає в потенційному використанні LLM для генерації та поширення дезінформації, фейкових новин, пропаганди тощо.

Окремі виклики, пов'язані з кібербезпекою у контексті використання архітектур "суміші експертів" (MoE). Вона представляє собою передовий підхід у створенні потужних та адаптивних моделей ШІ. Попри значний потенціал у підвищенні ефективності та точності LLM, використання MoE архітектур вносить певні виклики та ризики у сфері кібербезпеки, які вимагають детального аналізу та відповідних заходів реагування.

Як відомо, MoE моделі складаються з множини

"експертів", кожен з яких спеціалізується на вирішенні певних задач. Ця комплексність може ускладнити процеси управління та контролю за моделями, зокрема у контексті виявлення та нейтралізації потенційних кіберзагроз. Тому важливо розробити ефективні механізми моніторингу та аудиту для забезпечення безпеки даних, які обробляються та генеруються MoE LLM/SLM.

Іншим важливим аспектом є збільшення так званої атакованої поверхні за рахунок інтеграції численних "експертів" у єдину систему MoE. Це надає потенційним зловмисникам більше точок входу для кібератак. Кожен "експерт" може містити вразливості, які можуть бути використані для втручання у роботу системи або витоку конфіденційної інформації.

Крім того, в MoE зростає складність забезпечення конфіденційності та цілісності даних через те, що у процесі роботи MoE моделей обробляються та генеруються великі обсяги даних. Щоб захистити їх конфіденційність та цілісність, потрібні розширені механізми шифрування та захисту даних на всіх етапах обробки та передачі даних, у тому числі від "експертів" до внутрішнього роутера з урахуванням потенційних ризиків впливу та зовнішніх загроз.

Слід також врахувати, що комплексність MoE-архітектур ускладнює забезпечення прозорості та відповідальності в роботі моделей, особливо у випадках, коли необхідно визначити причину неправильної поведінки. "Експерти" у складі MoE можуть бути реалізовані на різних за структурою,

розмірами та рівнем квантизації LLM. Тому розробка інструментів для інтерпретації рішень, прийнятих моделями, є ключовою для підвищення довіри та безпеки систем на основі MoE.

Однак насправді радикальні зміни у сфері кіберзахисту, що зумовлені стрімким розвитком технологій ШІ, найбільше пов'язані з такою новою загрозою, як великі моделі дій (Large Action Model, LAM). Вони представляють собою штучні інтелектуальні системи, спроможні самостійно ініціювати та виконувати комплексні дії в кіберпросторі. Ці дії можуть включати запуск програм, створення та редагування файлів, взаємодію з мережевими сервісами та інші операції, які традиційно вимагали б участі людини. Одна з таких LAM була представлена на виставці CES 2024.

Подібні LAM суттєво розширюють можливості ШІ, переходячи від обробки та аналізу даних до активного втручання в роботу програмних і апаратних систем. Це створює потенціал не лише для позитивних змін, але й для нових видів кібератак та зловживань. Зокрема, LAM можуть бути використані для автоматизації створення та поширення вірусів, троянських програм та інших видів шкідливого програмного забезпечення. Такі моделі можуть автоматично змінювати налаштування системи, модифікувати або видаляти файли, створюючи потенціал для втручання в роботу критично важливих інфраструктур. Використання LAM дозволяє здійснювати складні кібератаки без безпосередньої участі людини, що може призвести до швидкого розповсюдження атаки та ускладнення

процесу ідентифікації зловмисників.

У нещодавно опублікованому препринті [5] було представлено аналітичне дослідження, що фокусується на аналізі так званих "сплячих агентів", інкорпорованих у LLM. Дослідження [5] висвітлює випадок, коли попередньо навчена мовна модель була програмована реагувати на певні ключові слова, ідентифіковані в тексті. Зокрема, модель була налаштована генерувати валідний код у випадку зустрічі в тексті в якості ключового слова "2023" та переходити до генерації невірного коду при ідентифікації в тексті якоїсь події, пов’язаної з роком "2024". Автори препринту [5] зазначають, що ідентифікація та реадаптація такої моделі, включаючи сплячого агента, є неможливою. Це відкриває дискусію щодо потенціалу створення та використання сплячих агентів, активованих через кодові слова, як серйозного виклику для кібербезпеки, особливо у разі їх інтеграції до LAM. В цьому контексті, слід звернути увагу на концепцію "кібербомби", базованої на інтеграції сплячих агентів у великі моделі дій (LAM), керовані через внутрішній роутер у замкненій архітектурі. Такі загрози представляють собою значний ризик, оскільки пряма ідентифікація та видалення сплячих агентів з моделей штучного інтелекту є вкрай утрудненою через їхню високу ступінь інтеграції та прихованість від зовнішніх спостережень. Це підкреслює критичну необхідність розробки новітніх методологій в галузі кібербезпеки для ідентифікації та нейтралізації подібних загроз у складних системах штучного інтелекту. Враховуючи потенційні ризики, пов'язані з

LAM, важливо впровадити комплексні заходи для забезпечення кібербезпеки, які можуть включати:

- розширений моніторинг та аналіз поведінки систем на основі використання алгоритмів машинного навчання для виявлення незвичайної або підозрілої поведінки в мережі та на кінцевих точках;
- запобігання несанкціонованому використанню LAM шляхом впровадження нових, більш складних методів аутентифікації та контролю доступу.
- впровадження правових норм та стандартів, які регулюють розробку та використання LAM з метою мінімізації ризиків для суспільства та інфраструктур.
- підвищення обізнаності серед фахівців у галузі IT та кібербезпеки щодо потенційних ризиків та методів протидії загрозам, пов'язаним з LAM.

В контексті сучасних досліджень та розробок в області мовних моделей та моделей активних дій, що включають сплячих агентів, виникає необхідність забезпечення високого рівня прозорості та безпеки. Одним з фундаментальних принципів в цьому напрямі є відкритість вагових коефіцієнтів, доступність вихідного коду та даних, використаних для навчання таких моделей. Такий підхід сприяє забезпеченню відповідальності та контролю за функціонуванням систем на основі ШІ.

Для мінімізації ризиків, пов’язаних з потенційними небажаними діями моделей у програмному середовищі операційної системи, рекомендується використання ізольованих контейнерів або пісочниць, підконтрольних окремій LLM з моніторингу. Це дозволяє обмежити вплив

моделей на інші додатки та компоненти системи, забезпечуючи додатковий рівень захисту.

Важливо також враховувати, що при використанні LLM/LAM для виконання відповідальних завдань у внутрішніх процесах організації, бажано здійснювати розробку та адаптацію таких моделей власними силами. Це дозволить не лише глибше зрозуміти механізми роботи ШІ, але й гарантувати відсутність сторонніх втручань у вигляді сплячих агентів, забезпечивши надійний контроль за безпекою та цілісністю системи.

Висновки

Щоб мінімізувати вказані ризики, необхідні спільні зусилля дослідників, розробників, регуляторів та суспільства в цілому. Потрібно виробити чіткі етичні принципи та норми використання LLM/LAM, забезпечити прозорість та підзвітність їх розробки та застосування, впровадити механізми контролю та запобігання зловживанням.

Водночас, не варто недооцінювати позитивний потенціал LLM. При відповідальному та зваженому підході, ці технології можуть принести величезну користь в найрізноманітніших сферах - від освіти та науки до бізнесу та державного управління. Вони можуть допомогти подолати мовні та комунікаційні бар'єри, зробити інформацію та знання більш доступними, автоматизувати рутинні завдання та підвищити ефективність багатьох процесів. Тому надзвичайно важливо продовжувати дослідження в сфері LLM/LAM, експериментувати з новими архітектурами та підходами, шукати оптимальні

рішення та компроміси. Разом з тим, ці дослідження мають супроводжуватись постійною рефлексією щодо їх потенційних наслідків та відповідальністю за результати.

Великі мовні моделі - це технології, які можуть докорінно змінити наше життя та взаємодію з інформацією. Від нас залежить, чи зможемо ми використати їх потенціал на благо суспільства, мінімізувавши при цьому можливі ризики та негативні наслідки. Це складний, але необхідний виклик, який потребує міждисциплінарного підходу, відкритого діалогу та співпраці на всіх рівнях.

Література

1. Слюсар В.І. Великі мовні моделі у військовій сфері. //XXIII Міжнародна науково-технічна конференція Artificial intelligence and intellegent systems (AIIS'2023), 10-11 жовтня 2023. - Київ (презентація). DOI: 10.13140/RG.2.2.30196.94086..

2. Slyusar Vadym. Reducing the Cognitive Burden of a Soldier with the Help of Personal AI and LLM Assistant. // The LCGDSS Human System Integration (HSI) symposium, 12 January 2024. - DOI: 10.13140/RG.2.2.10264.57605/1.

3. GPT-4. Technical Report by OpenAI, 27 March 2023, URL: https://arxiv.org/pdf/2303.08774v3.pdf.

4. Yuriy Kondratenko, Galyna Kondratenko, Anatolii Shevchenko, Vadym Slyusar, Yuriy Zhukov, Maxym Vakulenko. Towards Implementing the Strategy of Artificial Intelligence Development: Ukraine Peculiarities. // Proceedings of the 11-th International Conference "Information Control Systems & Technologies" (ICST 2023). Odesa, Ukraine, September 21–23, 2023. - Pp. 106-117.

5. Evan Hubinger and all. Sleeper agents: training deceptive llms that persist through safety training. https://arxiv.org/pdf/2401.05566.pdf.