УДК 004.056.5
DOI: 10.26102/2310-6018/2024.46.3.016

Языковые модели и онтологии, угрозы безопасности в распределенной системе

Н.И. Донских✉

Финансовый университет при Правительстве Российской Федерации, Москва, Российская Федерация

Резюме. Исследования в области больших языковых моделей (Large Language Models) и систем обработки естественного языка (Natural Language Processing) активизировались из-за появления новых, латентных и серьезных рисков, например, нарушений процессов генерации вывода, вредоносных запросов в автоматическом режиме. Разрабатываются синергетические сценарии применения больших языковых моделей. Основная гипотеза, учитываемая в данном исследовании – возможность страховки (с заданной вероятностью) от генерации запрещенного контента и его «подмешивания» к пользовательскому запросу, учет онтологических свойств и связей для улучшения качества поиска в практических задачах, например, с помощью библиотеки онтологий. Использованы методы анализа-синтеза, моделирования-прогнозирования, экспертно-эвристические, теории вероятностей и принятия решений. Основные результаты статьи: 1) аналитика по проблемам применения больших языковых моделей в достижении устойчивости в инфраструктуре системы (предложена таблица ключевых методов); 2) предложена языковая модель устойчивости сетевой инфраструктуры на основе оценок распределений при подмешивании слов, в которой использован байесовский метод; 3) предложена и исследована аналогичная языковая модель на основе экспертно-эвристического подхода к оценке рисков (неопределенностей в системе), в частности, с использованием информационно-энтропийного подхода. Исследование можно развивать, усложняя модели (гипотезы) и «глубину» учета рисков.

Ключевые слова: большие языковые модели, устойчивость, риски, информационная безопасность, управление.

Для цитирования: Донских Н.И. Языковые модели и онтологии, угрозы безопасности в распределенной системе. *Моделирование, оптимизация и информационные технологии.* 2024;12(3). URL: https://moitvivt.ru/ru/journal/pdf?id=1634 DOI: 10.26102/2310-6018/2024.46.3.016

Language models and ontologies, security threats in distributed system

N.I. Donskikh✉

Financial University under the Government of the Russian Federation, Moscow, the Russian Federation

Abstract. Research in the field of large language models and natural language processing systems has intensified due to the emergence of new, latent and serious risks, for example, violations of the output generation processes, malicious requests in automatic mode. Synergistic scenarios for large language models are being developed. The main hypothesis taken into account in this study is the possibility of insurance (with a given probability) from the generation of prohibited content and its "mixing" with the user query, taking into account ontological properties and connections to improve the quality of search in practical tasks, for example, using an ontology library. Methods of analysis-synthesis, modeling-forecasting, expert-heuristic, probability theory and decision-making were used. The main results of the article: 1) analytics on the problems of applying large language models in achieving stability in the system infrastructure (a table of key methods was proposed); 2) a language model of network

infrastructure stability based on estimates of distributions when mixing words is proposed, which uses the Bayesian method; 3) a similar language model was proposed and studied on the basis of an expert-heuristic approach to assessing risks (uncertainties in the system), in particular, using an information-entropy approach. Research can be developed by complicating models (hypotheses) and the "depth" of risk accounting.

Keywords: large language models, resilience, risks, information security, governance.

For citation: Donskikh N.I. Language models and ontologies, security threats in a distributed system. *Modeling, Optimization and Information Technology*. 2024;12(3). URL: https://moitvivt.ru/ru/journal/pdf?id=1634 DOI: 10.26102/2310-6018/2024.46.3.016 (In Russ.).

Введение

Исследования в области больших языковых моделей LLM (Large Language Model), включая обработку естественного языка NLP (Natural Language Processing) активизировались [1–3], в том числе, в связи с проблемой защиты от серьезных уязвимостей [4], например, атак класса Prompt Injection, которые нарушают процесс генерации вывода, искажая результат и усложняя процесс [5]. Процесс «подмешивания» вредоносных запросов идет в автоматическом режиме и не требует от злоумышленника дополнительных усилий. Например, для этого используется метод Greedy Coordinate Gradient-based Search для максимизации (оптимизации) вероятности получения утвердительного ответа [6].

Защитный и практически применимый механизм должен быть адаптивным и гибким, массово применимым к реальным и синергетическим сценариям применения LLM [7]. Он должен гарантированно (на указанную вероятность) страховать, защищать от внедрения к запросам слов (словосочетаний) для генерации нейросетью запрещенного контента. В частности, с использованием LLM типа Viccuna, Pythia, Falcon, Guanaco, GPT, PaLM, Claude и др.

Автоматизация атак на LLM с помощью генерируемых последовательностей включений к пользовательскому запросу заставляет систему подчиниться командам пользователя, несмотря на его вредоносный характер (на вредоносный характер контента) даже при наличии качественно проведенного тестирования и эффективного отладочного механизма. При формализации, описании и реализации связей в системе эффективно используются онтологии.

Онтологии в информационных системах представляют собой структуру (кортеж) из понятий и категорий по определенной предметной области (в рассматриваемой проблеме – распознавание угроз, вмешательств в сети), их свойств и связей.

Онтологии, на упрощенном уровне, – это формальные описания знаний, на языках обрабатываемых и понимаемых компьютерами, например, процедурные, фреймовые. Это обобщение интерфейсных языков интеллектуальных систем. При всей формальности данного понятия, оно вводилось для практических реальных нужд бизнеса [8].

Интеллектуальная система может выступать как библиотека онтологий, декларативных знаний.

Онтологии содержимого веб-страниц позволяют улучшать качество поиска поисковых систем (Smart-Web, Semantic-Web или концепция Web 3.0). Для этого разработан фреймворк RDF (Resource Description Framework) для описания веб-ресурсов.

При идентификации методов информационной защиты следует формировать не только достаточный для оценивания аппарат, набор моделей и методов риск-устойчивости с единой системной и инструментальной основой.

Материалы и методы

Для реализации систем искусственного интеллекта применяют средства и методы машинного обучения (ML, Machine Learning), в частности, для обучения – архитектуру свёрточных нейросетей (CNN, Convolutional Neural Network).

Исследования по безопасности ML-моделей на основе нейросетей начались в 2013 г. Изучались CNN-атаки при распознавании изображений (образов).

Атака осуществлялась для сбора информации по модели, датасету (наборе данных, использованных при её обучении), точнее, подачей на вход сети специально сформированного набора данных (вредоносных примеров, adversarial examples).

Обычно рассматриваются атаки типа:

1) «white-box», предполагается полный доступ к ресурсам сети и данным, знание архитектуры и параметров сети, к датасетам;

2) «gray-box», предполагается наличие у атакующего сведений о сетевой архитектуре, возможно, ограниченного доступа к данным;

3) «black-box», предполагается отсутствие информации по устройству сети, датасетам, наличие неограниченного доступа к модели.

Атаки класса «black-box» очень сложны при реализации, класса «white-box» – теоретического типа. Можно построить таблицу ключевых методов (Таблица 1).

Таблица 1 – Методы и их особенности
Table 1 – Methods and their features

№	Метод	Ключевая особенность
1	FGSM (FastGradientSignMethod)	Весовой вектор нейросети большой размерности, малые возмущения входных данных могут привести к значительному росту ошибки
2	BIM (BasicIterativeMethod)	Итерационное применение FGSM с отслеживанием суммарного возмущения, чтобы оно не выводило генерируемый вредоносный пример за пределы задаваемой области значений
3	PGD (ProjectedGradientDescent)	Аналогичен BIM
4	CW (CarliniWagner)	«Прямая» оптимизация функционалов потерь (чувствительности), отличных от изначально использованных при построении нейросети
5	DeepFool	Поиск вредоносных примеров с малыми возмущениями, изменением класса распознаваемого объекта
6	HopSkipJumpAttack	После каждой итерации идет обработка полученных от модели данных, оценивается, где вносить изменения на следующей итерации

Кроме детерминированных методов, есть и стохастические, предполагающие, что есть лишь ограниченный доступ к нейросети и вывод результата в виде пары <класс объекта, вероятность совпадения>.

Используются и методики внесения возмущения при помощи аффинных преобразований изображения, например, сдвигов. Нейросеть не так просто «обмануть»: для уверенного распознавания необходимо сгенерировать и протестировать много вредоносных примеров.

Результаты

Языковая модель устойчивой безопасности сетевой инфраструктуры на основе оценки распределения подмешивания слов

Защищенность сетевой инфраструктуры организации является актуальной проблемой из-за множества современных факторов, в частности, целевых и чувствительных латентных атак на управляемость цифровой экосистемы.

Часто разрабатывается специальная стратегия проникновения в систему. Используются методы социальной инженерии, гештальт-психологии, нейросетевые подходы. Важны методы защиты наиболее уязвимых элементов: трафика, журнала сигнатур, учетных записей, файловой системы, архитектуры сети, облачных вычислений и активности.

Используются эвристические процедуры анализа поведения злоумышленников и их деструктивных действий. Они автоматически формируют цифровые профили поведения в сети, атак. Цифровые профили позволят идентифицировать форму атаки, управления системой.

Эксперты TAdviser, Angora Security на основе системного анализа 450 банковских и страховых проектов определяют следующие ключевые сервисы и услуги информационной безопасности (список ранжирован по популярности):

1) аудит и консалтинг ИТ-инфраструктуры, в частности, формирование и поддержка корпоративной политики ИТ-безопасности, модели контура (22%);

2) анализ защищенности информационной инфраструктуры, в частности, поиск и нейтрализация уязвимостей инфраструктуры организации (21%);

3) защита периметра и сети, в частности, внедрение решений по контролю сетевой активности, межсетевому экранированию (16%);

4) услуги SOC-центра и мониторинга событий, в частности, для проектов, которым недостаточны собственные центры (8%);

5) защита от запланированных, таргетированных и DDoS-атак уровня L7 ботнетов, в частности, на основе AntiDDoS, WAF, AntiBot и др. (8%);

6) применение методов и инструментария социальной инженерии для выявления и нейтрализации прямых атак, пропускаемых антивирусной защитой, в частности, использующих гештальт-психологию, «песочницы» с запуском файла из неизвестного источника (6%).

Композиции сверток повышают эффективность, они позволяют распараллеливать датасеты, токены данных, уменьшая тем самым почти на порядок задержку оценки предложения (сравнительно с рекуррентными моделями) [9].

Языковые статистические модели позволяют оценивать распределение вероятностей ряда слов моделированием вероятности очередного слова (если предыдущие слова уже появились). Такую ситуацию можно учесть при рассмотрении следующей модели поведения злоумышленника.

Если он обошел фильтр защиты и находится на следующем уровне контроля, то система безопасности защиту «провалила» (как минимум, по времени).

Сложность идентификации злоумышленника определим по формуле условной вероятности:

$$p(S) = \frac{p(C_i)p(S|C_i)}{\sum_{k=0}^{K} p(C_k)p((S|C_k)},$$

где C_i – событие «взлом *i*-го уровня» ($i = 0, \ldots, K$), S – событие «в текущем состоянии в текущий момент контролируется ситуация текущего уровня», $p(C_i)$ – априорная вероятность перехода на *i*-ый уровень защиты (его взлома), $p(S|C_i)$ – вероятность, что в

текущий момент, при проходе i-го уровня защиты, система окажется в текущем состоянии, $p(S)$ – вероятность i-го уровня (в текущем состоянии и моменте времени).

Если рассмотреть гипотезы последовательного взлома всех уровней $H_1, H_2, \dots, H_k$, то соответствующая условная вероятность равна:

$$p(H_i|A_1, \dots, A_n) = \frac{p(H_i)p(A_1|H_i)p(A_2|H_i)\cdots p(A_n|H_i)}{\sum_{j=1}^{k} p(H_j)\, p(A_1|H_j)p(A_2|H_j)\cdots p(A_n|H_j)}.$$

Совпадение $p(H_i)$ и $p(H_i|A_1, \dots, A_n)$ распределений на множестве $\{H_1, H_2, \dots, H_k\}$ свидетельствует о противоречивости получаемых потоковых данных (измерений), не влияет на интегральное распределение.

Если

$$p(H_i|A_1, \dots, A_n) = 1,\ p(H_j|A_1, \dots, A_n) = 0 \quad (j = 1,2, \dots, k),$$

то рассматриваемое распределение на $\{H_1, H_2, \dots, H_k\}$ будет вырождено: гипотеза H_i – достоверна, остальные – невозможные.

Интегральная оценка устойчивости системы может быть искажена. Она мешает администрированию и защищенности, администраторам по безопасности сложно выбирать и ранжировать гипотезы. Апостериорная вероятность позволяет по априорно задаваемым уровням оценивать систему в целом, а также помогает выстраивать технологически управленческие средства и решения, используя тестирование, ситуационное принятие решений.

Языковая модель устойчивой безопасности сетевой инфраструктуры на основе экспертно-эвристического подхода к оценке рисков (неопределенностей в системе)

Следует поддерживать функциональную и структурную (сетевую) устойчивость системы в течение задаваемого промежутка времени.

Риск-устойчивость, сложно структурируемая и моделируемая языковыми моделями и онтологиями, является проблемой. Она многокритериальная и многофакторная, поэтому необходимо учитывать имитацию воздействия вредоносных команд в потоке на управление потоком, сетевую разведку, инновационный потенциал бизнес-компании [10] и др. В целом, все эти методы воздействия можно объединить категорией «MITM-атака» (Man In The Middle). В силу сказанного, часто применяются эвристически-экспериментальные подходы к прогнозированию риск-устойчивости, оценке уровня устойчивости.

Предлагаем нижеследующую процедуру.

Пусть a_{ij} $(i = 1,2, \dots, n; j = 1,2, \dots, m)$ – оценка j-го эксперта по i-ой характеристике (i-му фактору) x_i риска. Формируем последовательность матричных итераций вида:

$$A^{(k)} = \left\| a_{ij}^{(k)} \right\|_{i=1,\dots,n}^{j=1,\dots,m},$$

для которых компоненты решения последовательно уточняются рекуррентно, согласно зависимости (как в методе Дельфи):

$$y_i^{(k)} = A^{(k)} y_i^{(k-1)}.$$

Неполнота, неточность, неопределенность первичных данных и отсутствие профессионалов Data Science в организациях вносят свои «шумы».

Эффективность системы – результат устойчивости и полноты бизнес-процессов (функционала), стратегии управления и ресурсной обеспеченности. Каждая из указанных составляющих влияет на выходной результат.

Лица, принимающие решения (ЛПР), оценивают каждое предположение эксперта по величине $y_i^{(k)}$ некоторым весом

$$\omega_{ij}^{(k)} = \frac{a_{ij}^{(k)}}{v_i},$$

где v_i – нормирующая величина.

Область допустимых решений определим процедурой:

1) нахождение граничных значений

$$\underline{\omega}_j^{(k)} = \min_i \omega_{ij}^{(k)},$$

$$\overline{\omega}_j^{(k)} = \max_i \omega_{ij}^{(k)};$$

2) нахождение размахов

$$d_{ij}^{(k)} = \left|\overline{\omega}_i^{(k)} - \underline{\omega}_j^{(k)}\right|;$$

3) разбиение $[\underline{\omega}_j^{(k)}; \overline{\omega}_j^{(k)}]$ на h_j интервалов;

4) если в интервал $(1-\varepsilon; 1+\varepsilon)$ (где $\varepsilon > 0$ достаточно малое число) попадает малая часть точек, то в этом варианте (k) решения заинтересован (или не заинтересован) эксперт *i*;

5) в зависимости от выбора эксперта i, соответственно, все y_i умножаются (усиливаются или ослабляются) на

$$\delta_{ij}^{+} = \frac{2}{n} \sum_{\omega_{ij}<1} \omega_{ij}$$

или

$$\delta_{ij}^{-} = \frac{2}{n} \sum_{\omega_{ij}\geq 1} \omega_{ij};$$

6) переход к следующей итерации (*k:=k+1*), если не достигнута приемлемая точность экспертного суждения по рассматриваемой ситуации.

Будем считать, что известны ряды распределений значений каждой характеристики (рассматриваем дискретный случай). Тогда эксперты могут выбрать значения:

а) с наибольшими вероятностями («полное согласие»);

б) с наименьшими вероятностями («полное несогласие»);

в) с различными вероятностями («частичное согласие»);

г) с неэкстремальными вероятностями («корректируемое согласие»).

Для рядов распределений можно ввести оценку «шума» (энтропии) в классической форме:

$$H_i^{(k)} = -\sum_{j=1}^{m} d_{ij}^{(k)} \log_2 d_{ij}^{(k)}.$$

Если все эксперты – приверженцы выбора а), а сумма

$$\sum_{i=1}^{n} \max_j \left\{ d_{ij}^{(k)} \right\}$$

является максимальной среди всех возможных, то все эксперты согласны с таким выбором (решением).

Кроме коэффициента согласованности экспертной группы (коэффициента конкордации) необходимо определить, кто из команды экспертов должен быть удален. Это тот эксперт (номер *j*), для которого величина

$$d_{ij}^{(k)}\chi\{y_i \in Y^{(k)}\} + \frac{\chi\{y_i \in Y^{(k)}\}}{d_{ij}^{(k)}\sum_{i=1}^{n}\frac{1}{d_{ij}^{(k)}}}$$

минимальна по всем *j=1,2,…,m,* т. е. его экспертные оценки противоречат напрямую лучшим вариантам по *k*.

Указанная процедура реализуема с помощью как нечеткого аппарата (множество, логика), так и с помощью нейросети и глубокого машинного обучения. ЛПР должно не только уметь принимать решение (делать выбор из альтернатив), но и адаптировать свои решения к мнению экспертной группы и внешним факторам.

Нейросети превосходят традиционные языковые модели с проблемой разреженности данных, затрудняющей представление больших контекстов длины *n* (долгосрочных зависимостей). Нейросетевые языковые модели эту проблему решают встраиванием слов в пространство ситуаций, для которого обучается (дообучается) нейросеть.

Важно добиваться приемлемой алгоритмической сложности для долгосрочных зависимостей, например, порядка $O(n/k)$, где k – «ширина» ядра. При этом рекуррентные зависимости требуют порядка $O(n)$.

Обсуждение

Большие (именно в смысле системного анализа) языковые модели (LLM) типа ChatGPT-4, обученные на большой выборке демонстрируют эволюционный потенциал на практике, но имеют уязвимости, доступные злоумышленникам, актуализируемые в беседе на естественном языке.

Масштабность LLM-моделей и методы глубокого обучения позволяют добиваться достаточных в долгосрочных предсказаниях результатов.

Проблема уязвимости ML-моделей и свёрточные нейросети адаптивны и позволяют проигрывать ситуации и сценарии при машинном обучении. Модели CNN используют в распознавании аудиовизуальной и иной информации. Необходимо разрабатывать более «тонкие» и «мягкие» модели и инструменты анализа.

Рассмотренная проблема актуальна для многих компаний, переходящих к гибкому и бережливому производству, инновационным и малотиражным (даже индивидуальным) технологическим цепочкам. Они должны обращаться к сравнению онтологий, цифровых моделей производства и/или продукции. Необходимо вводить индивидуальные функции удовлетворенности по критериям (параметрам, которые могут достигать много сотен).

Отметим и негативные обстоятельства, тормозящие практическое внедрение нейросетевых систем и больших языковых моделей: отсутствие компетентных специалистов, релевантных оценочных критериев (метрик) и инструментов автоматизации (оркестрации) и интеграции процессов (сценариев).

Заключение

Использование языковых моделей и онтологий в описании, формализации и прогнозе угроз безопасности расширяет возможности и глубину прогноза рисков и

поведения злоумышленника в сети. Построение, исследование математических моделей и приложений для поддержки прогнозирования – проблема актуальная, соразмерно усложняющаяся вместе с усложнением взаимодействий в сети, инфраструктуры систем.

В различных областях есть много проблем, когда вносят шумы нечеткость и неточность, неопределенность инновационных знаний для оперативного внедрения в практику.

Работу можно развивать и применять как в направлении разработки «мягких» (гибких, в частности, нейросетевых) моделей и их эластичности.

СПИСОК ИСТОЧНИКОВ / REFERENCES

1. Liu Y., Deng G., Li Y. et al. Prompt Injection attack against LLM-integrated Applications. URL: https://doi.org/10.48550/arXiv.2306.05499 [Accessed 14th June 2024].
2. Martínez Torres J., Iglesias Comesaña C., García-Nieto P.J. Review: machine learning techniques applied to cybersecurity. *International Journal of Machine Learning and Cybernetics*. 2019;10(10):2823–2836. https://doi.org/10.1007/s13042-018-00906-1
3. Кузьминов И.Ф., Бахтин П.Д., Тимофеев А.А. и др. Современные технологии обработки естественного языка для решения задач стратегической аналитики. *Искусственный интеллект и принятие решений.* 2020;(1):3–16. https://doi.org/10.14357/20718594200101
   Kuzminov I.F., Bakhtin P.D., Timofeev A.A. et al. Modern Natural Language Processing Technologies for Solving Strategic Analytics Tasks. *Iskusstvennyi intellekt i prinyatie reshenii = Artificial Intelligence and Decision Making*. 2020;(1):3–16. (In Russ.). https://doi.org/10.14357/20718594200101
4. Мударова Р.М., Намиот Д.Е. Противодействие атакам типа инъекция подсказок на большие языковые модели. *International Journal of Open Information Technologies*. 2024;12(5):39–48.
   Mudarova R., Namiot D. Countering Prompt Injection attacks on large language models. *International Journal of Open Information Technologies.* 2024;12(5):39–48. (In Russ.).
5. Юргель В.Ю. Сложности моделирования естественного языка. *Вестник науки и образования*. 2019;(23-1):12–14.
   Jurgel V.Yu. Complexities of natural language modeling. *Vestnik nauki i obrazovaniya = Herald of Science and Education*. 2019;(23-1):12–14. (In Russ.).
6. Fang H., Fang G., Yu T., Li P. Efficient Greedy Coordinate Descent via Variable Partitioning. In: *37th Conference on Uncertainty in Artificial Intelligence (UAI 2021): Proceedings, 27–30 July 2021, Toronto, Canada, USA.* PMLR; 2021. pp. 547–557.
7. Chen X., Zhang N., Xie X. et al. KnowPrompt: Knowledge-aware Prompt-tuning with Synergistic Optimization for Relation Extraction. In: *WWW '22: Proceedings of the ACM Web Conference 2022*, *25–29 April 2022, Lyon, France.* New York: Association for Computing Machinery; 2022. pp. 2778–2788. https://doi.org/10.1145/3485447.3511998
8. Фридман А.Я. Онтология проектирования ситуационных цифровых двойников для моделирования структурной безопасности индустриально-природных комплексов. *Онтология проектирования.* 2024;14(1):29–41. https://doi.org/10.18287/2223-9537-2024-14-1-29-41
   Fridman A.Ya. Ontology for designing situational digital twins of industrial-natural complexes for modeling their structural safety. *Ontologiya proektirovaniya = Ontology of Designing.* 2024;14(1):29–41. (In Russ.). https://doi.org/10.18287/2223-9537-2024-14-1-29-41

9. Dauphin Y.N., Fan A., Auli M., Grangier D. Language Modeling with Gated Convolutional Networks. In: *34th International Conference on Machine Learning: Proceedings, 6–11 August 2017, Sydney, Australia.* 2017. pp. 933–941.
10. Kaziev M.V., Medvedeva L.B., Tyutrin N.O., Khizbullin F.F., Takhumova V.O. Improvement and modeling of the company's activity based on the innovative KPI system. *Journal of Fundamental and Applied Sciences*. 2018;10(5S):1406–1415.

ИНФОРМАЦИЯ ОБ АВТОРАХ / INFORMATION ABOUT THE AUTHORS

Донских Никита Игоревич, аспирант кафедры информационной безопасности, Финансовый университет при Правительстве Российской Федерации, Москва, Российская Федерация.
e-mail: Nikdonskikh@gmail.com

Nikita I. Donskikh, Postgraduate student of the Department of Information Security, Financial University under the Government of the Russian Federation, Moscow, the Russian Federation.

Статья поступила в редакцию 17.07.2024; одобрена после рецензирования 30.07.2024; принята к публикации 02.08.2024.

The article was submitted 17.07.2024; approved after reviewing 30.07.2024; accepted for publication 02.08.2024.