



ФОРУМ

ТЕХНОЛОГИИ ДОВЕРЕННОГО
ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Методы и средства валидации и верификации интеллектуальных средств генерации программного кода

Александр Самонов
ВКА имени А.Ф.Можайского

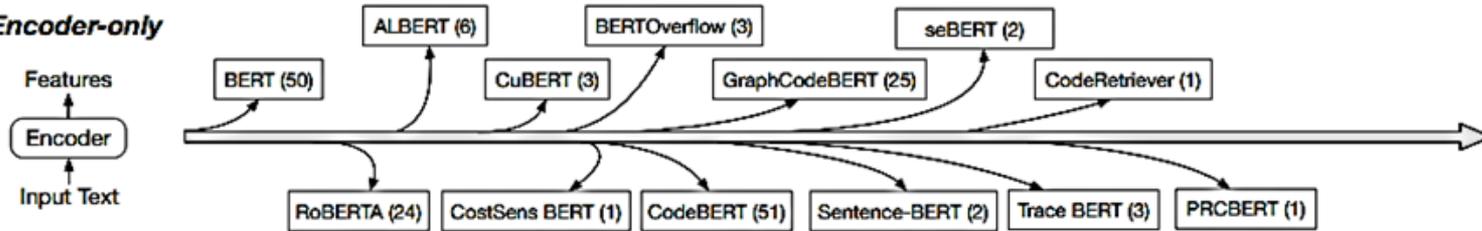
Основные вопросы

1. Анализ возможностей и ограничений современных и перспективных интеллектуальных средств генерации программного кода.
2. Модель угроз для интеллектуальных средств генерации программного кода.
3. Методы и средства валидации и верификации интеллектуальных средств генерации программного кода.

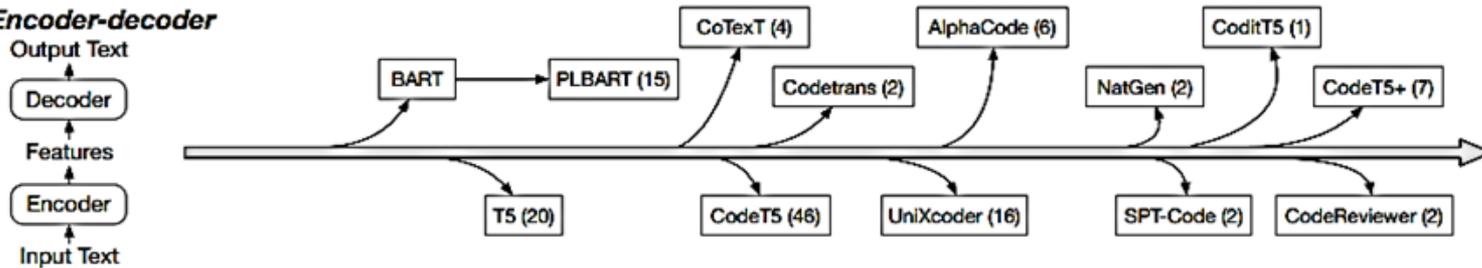
Хронология создания и развития Code LLM

arXiv:2308.10620v6 [cs.SE] 10 Apr 2024

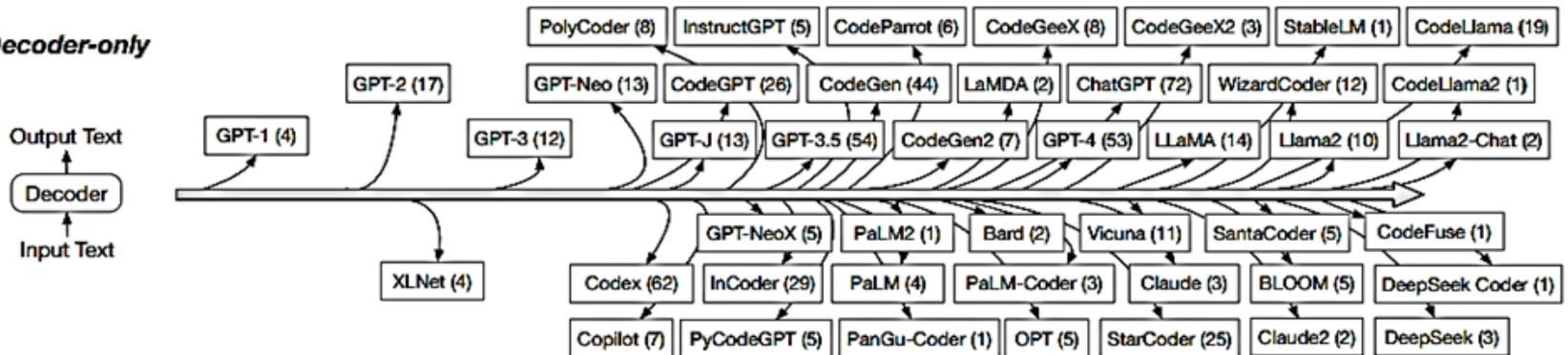
Encoder-only



Encoder-decoder



Decoder-only



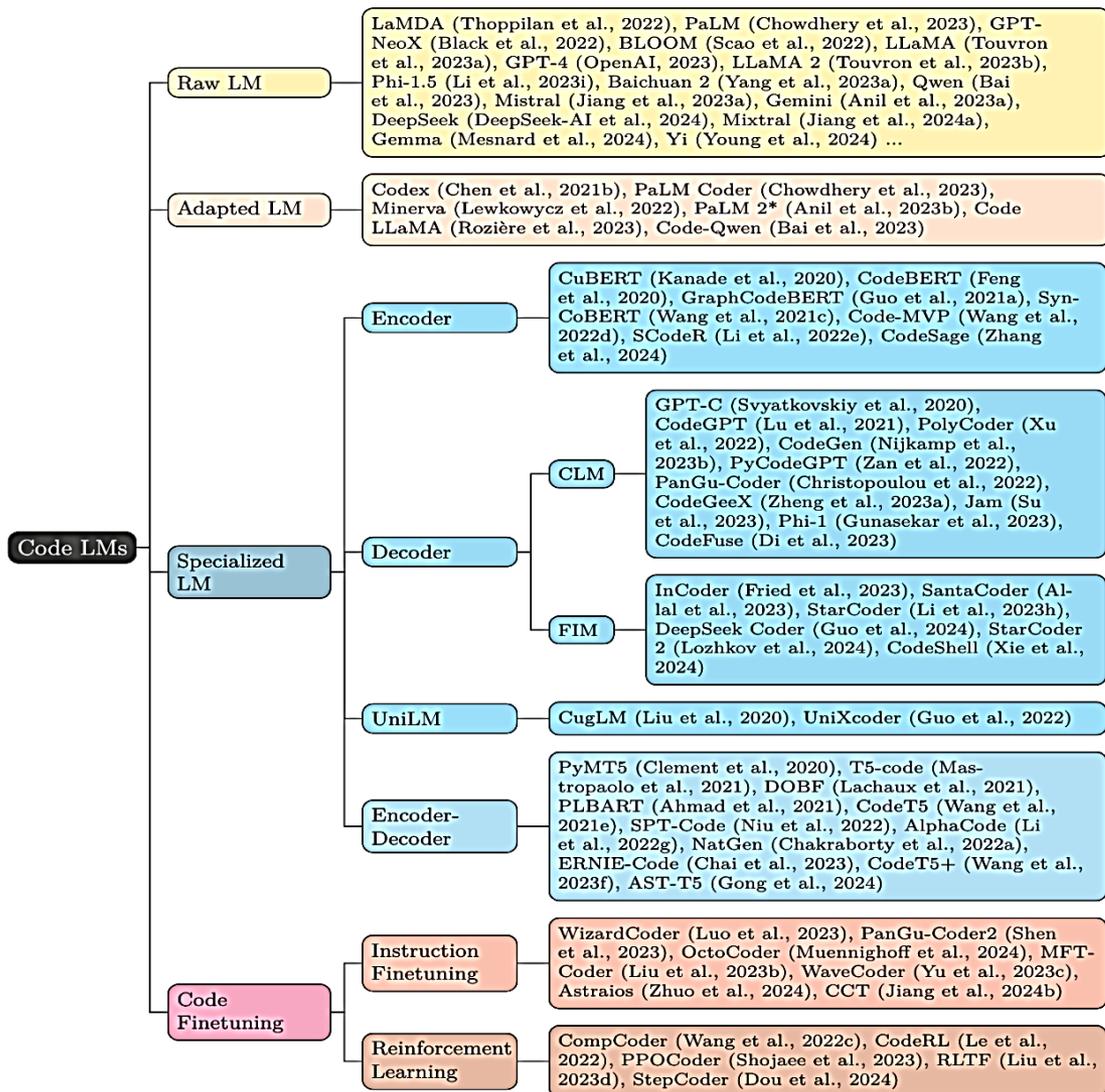
18 — 19 — 20 — 21 — 22 — 23 — 24 →

Результаты тестирования Code LLM

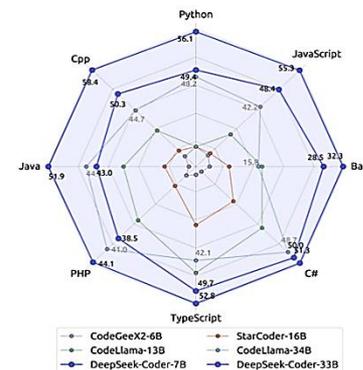
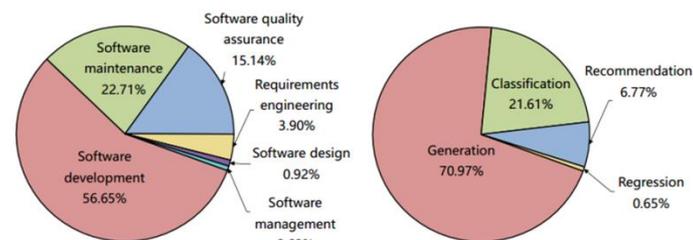
Model ▲	Win Rate ▲	humaneval-python ▲	java ▲	javascript ▲	c++ ▲
CodeQwen1.5-7B	45.08	50.79	42.15	50.07	48.35
CodeLlama-70b	43.21	52.44	44.72	56.52	49.69
DeepSeek-Coder-33b-base	42.75	52.45	43.77	51.28	51.22
CodeLlama-70b-Python	42.33	55.49	45.96	56.52	49.69
StarCoder2-15B	39.92	44.15	33.86	44.24	41.44
DeepSeek-Coder-7b-base	38.33	45.83	37.72	45.9	45.53
CodeLlama-34b	37.65	45.11	40.19	41.66	41.42
CodeLlama-34b-Python	36.81	53.29	39.46	44.72	39.09
CodeGemma-7B	33.83	40.13	35.03	43.06	40.34
CodeLlama-13b	32.12	35.07	32.23	38.26	35.81
CodeLlama-13b-Python	30.04	42.89	33.56	40.66	36.21
StarCoder2-7B	28.25	34.09	29.42	35.35	33.63

<https://huggingface.co/spaces/bigcode/bigcode-models-leaderboard>

Типы, назначение и проблемы Code LLM



Классы решаемых задач

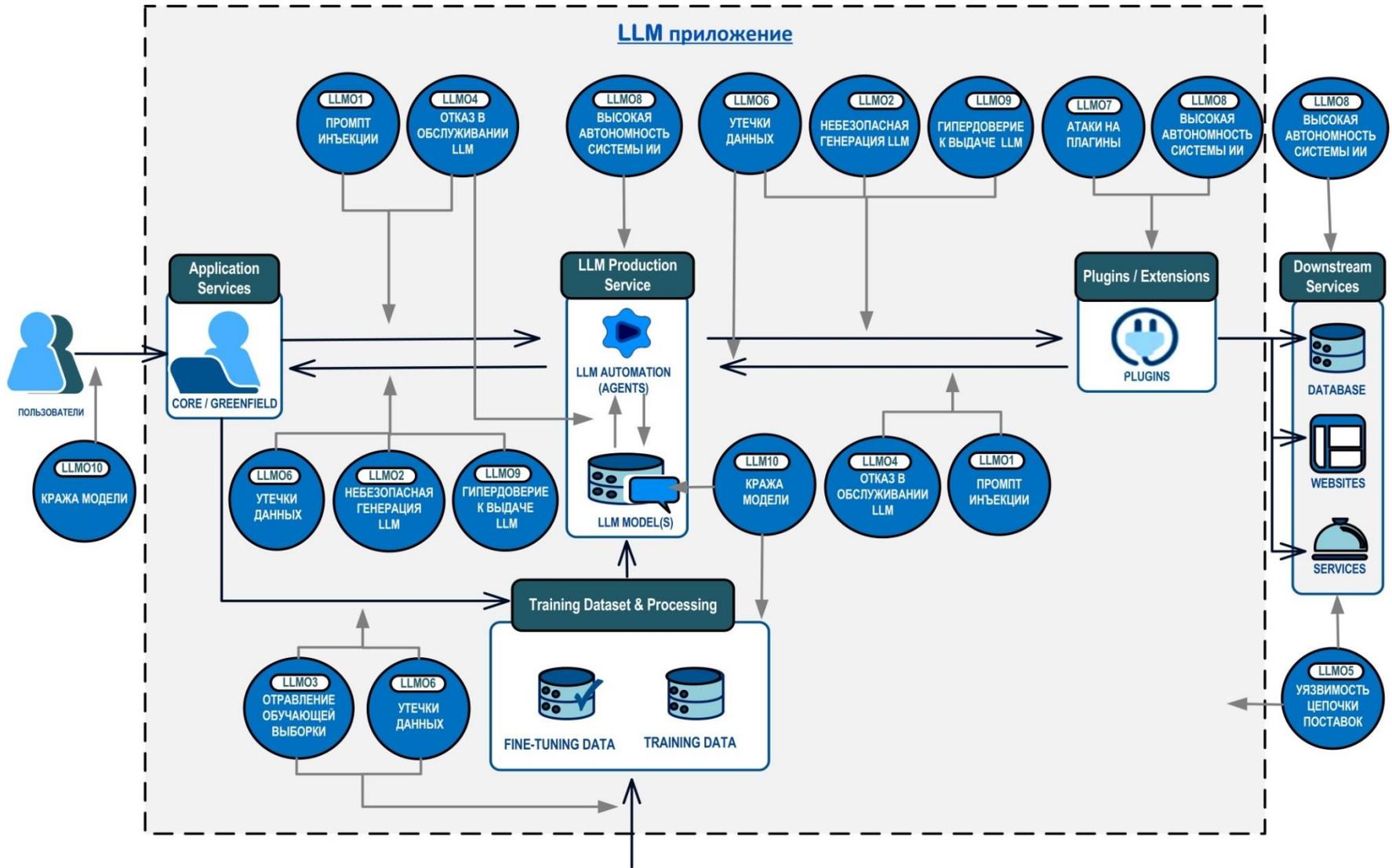


Проблемные вопросы:

катастрофическое забывание, риск переобучения, галлюцинации, неверные интерпретации, обработка исключительных ситуаций, защита цепочек поставок, высокие требования к производительности используемых при обучении LLM вычислительных средств.

Модель угроз OWASP Top 10 for LLM Applications

<https://llmtop10.com/>



Угрозы, уязвимости и методы защиты LLM приложений

<https://owasp.org/www-project-top-10-for-large-language-model-applications/>

Название в OWASP	Когда уязвимость применяется		Чем уязвимость опасна	Методы защиты	Легенда	
LLM01: Prompt Injection - Промпт-инъекции	R		Чат-боты с LLM. ИИ решения взаимодействующие с пользователями(анализ резюме через LLM). Автономные ИИ агенты(анализаторы отзывов сайтов, например)	Изменение поведения системы, harmful контент, утечки персональных данных и конфиденциальной информации	Firewall/DLP защита на LLM или с использованием готовых решений. Мониторинг трафика от LLM	R Runtime or Inference
LLM02: Insecure Output Handling - Небезопасная генерация LLM	R		Когда LLM агент работает в связке с другой системой: СУБД, shell, web browser	Кража пользовательский данных, изменение данных в связанных системах(API, СУБД), RCE, Повышение привилегий	Zero-trust подход для агента LLM, мониторинг трафика от LLM	T Training time
LLM03: Training Data Poisoning - Отравление обучающей выборки		T	Решения основанные на Open Source/своих pretrain моделях. Системы с дообученными моделями	Генерация harmful контента, утечки персональных данных и конфиденциальной информации	Использование MLOps практики, проверка свойств моделей, санитизация датасетов для обучения	I Infrastructure
LLM04: Model Denial of Service - Отказ в обслуживании LLM	R		Чат-боты с LLM	Атака делает систему недоступной для пользователей	Лимиты на использование токенов по пользователям, таймауты на сложные запросы	
LLM05: Supply Chain Vulnerabilities - Уязвимость цепочки поставок		T	Решения основанные на Open Source моделях. Системы с дообученными моделями. ИИ интегрированный внутри компаний. Датасеты для обучения моделей	Атака на инфраструктуру компании: RCE, Повышение привилегий. Если применяется к датасетам, то аналогично LLM03: Training Data Poisoning	Использование проверенных источников с моделями и компонентами разработки. Управление списком зависимостей через Software Bill of Materials (SBOM). Использование MLOps практики	
LLM06: Sensitive Information Disclosure - Утечки данных	R		Чат-боты с LLM. ИИ взаимодействующие с пользователями(анализ резюме через LLM)	Утечки персональных данных и конфиденциальной информации	Firewall/DLP защита на LLM или с использованием готовых решений. Мониторинг трафика от LLM. Очистка данных перед обучением. Ограничение доступа LLM к внешним сервисам	
LLM07: Insecure Plugin Design - Атаки на плагины	R		Плагины внутри ИИ систем. Плагины, которые работают с LLM.	Изменение поведения системы, harmful контент, утечки персональных данных и конфиденциальной информации, RCE, Повышение привилегий	Валидация при помощи Application Security Verification Standard от OWASP, использование статических анализаторов, принцип минимальной ответственности при разработке плагинов	
LLM08: Excessive Agency - Высокая автономность ИИ систем	R		Сложные решения построенные на LLM, которые проинтегрированы с ИТ инфраструктурой компании (API, СУБД, ESB, тп.)	Утечки и нарушение целостности данных, RCE, Повышение привилегий, поломка ИТ инфраструктуры компании	Принцип минимальной ответственности при разработке системы	
LLM09: Overreliance - Гипердоверие к выдаче LLM	R		Чат-боты с LLM, системы генерации контента и документов.	Риски генерации неправильной/копирайт информации и harmful контента.	Обучение сотрудников и пользователей, мониторинг выдачи и фактчекинг, дообучение модели для снижения галлюцинаций	
LLM09: Overreliance - Гипердоверие к выдаче LLM (работа с кодом)			Процесс разработки приложений, с использованием ИИ копилотов	Уязвимости ПО, разработанного при помощи копилот систем	Использование статических анализаторов, CI/CD, кодревью и тестирование решений на уязвимости	
LLM10: Model Theft - Кража модели	R		Атакующий проник на сервера компании и получил доступ к LLM модели. Данные из модели были украдены через инференс	Утечки персональных данных и конфиденциальной информации. Риск неправомерного использования информации	RBAC на доступ к модели, учет моделей в компании через реестры, добавление watermarks к модели. Лимиты на использование токенов по пользователям, мониторинг трафика от LLM	



Тактики и техники атак MITRE ATLAS

<https://atlas.mitre.org/matrices/ATLAS>

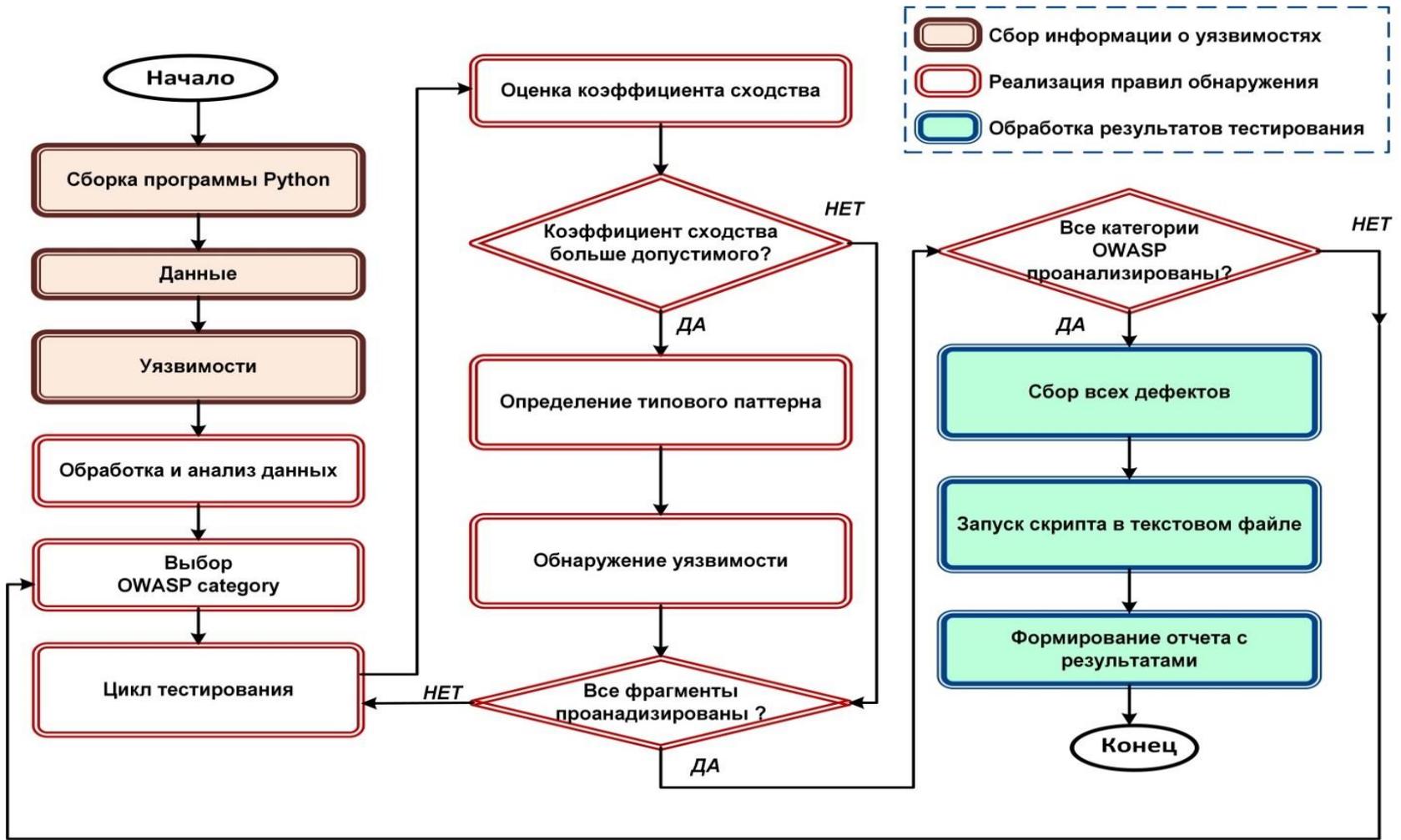
Tactic Name	Tactic ID	Technique Name	Technique ID
Reconnaissance	AML.TA0002	Search for Victim's Publicly Available Research Materials	AML.T0000
ML Model Access	AML.TA0000	ML-Enabled Product or Service	AML.T0047
Resource Development	AML.TA0003	Develop Adversarial ML Attack Capabilities	AML.T0017
ML Attack Staging	AML.TA0001	Craft Adversarial Data: Manual Modification	AML.T0043.003
Initial Access, Defense Evasion, Impact	AML.TA0004	Evade ML Model	AML.T0015
Initial Access, Defense Evasion, Impact	AML.TA0007	Evade ML Model	AML.T0015
Initial Access, Defense Evasion, Impact	AML.TA0011	Evade ML Model	AML.T0015

Тактики и техники атак на системы MO MITRE Atlas

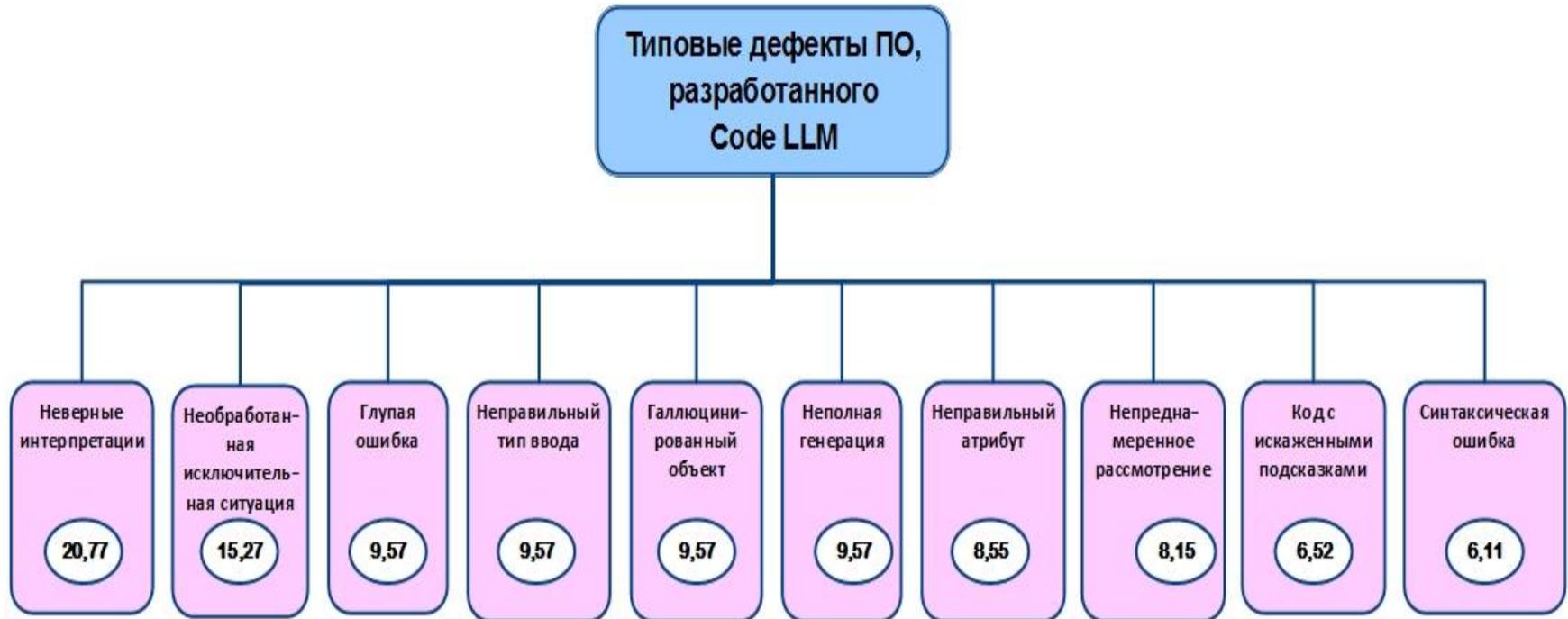
<https://atlas.mitre.org/matrices/ATLAS>

Разведка	Подготовка ресурсов	Первоначальный доступ	Доступ к модели MO	Выполнение	Закрепление	Предотвращение обнаруж.	Исследование	Сбор данных	Постановка атаки на модель MO	Эксплуатация	Воздействие
5 техник	7 техник	4 техники	4 техники	2 техники	2 техники	1 техника	3 техники	3 техники	4 техники	2 техники	7 техник
Сбор информации из общедост. источн.	Получение артефактов в MO	Компрометация цепочки поставок MO	Доступ к API модели MO	Выполнение с участием пользователя	Отравление обучающих данных	Обход модели	Исследование онтологии модели MO	Сбор артефактов в MO	Создание прокси-модели машинного обучения	Эксплуатация через API модели MO	Обход модели MO
Сбор информации о уязвимостях LLM	Подготовка необходимых ПС	Использование учетных записей	Доступ к продукту используемому MO	Использование интерпретатора команд	Создание бэкдора в модели MO		Исследование вида модели MO	Данные из информац. хранилищ	Создание бэкдора в модели MO	Эксплуатация через ПС	Отказ в обслуживании модели MO
Сбор информации с сайта жертв	Разработка собственных средств атаки MO	Обход модели	Доступ из физ. среды				Исслед. артефакт. MO	Данные из локальных систем	Подтверждение эффект. атаки		Засорение системы MO мусорными данными
Сбор инф. в открытых репозиториях	Получение технич. инф-туры	Недостатки в общедоступных приложениях	Полный доступ к модели MO						Создание составных данных		Нарушение целостности модели MO
Активное сканирование	Вредоносные наборы данных										Повышение издержек
	Отравление обучающих данных										Кража интел. собственности.
	Создание учетных записей										Неправомерное использование системы

Верификация и валидация ПО, разработанного Code LLM



Типовые дефекты ПО, разработанного Code LLM



Методы и средства верификации и валидации ПО, разработанного с помощью Code LLM

CodeUltraFeedback - набор данных о предпочтениях при кодировании, состоящий из 10 000 сложных инструкций и 40 000 ответов, сгенерированных с использованием 14 различных LLM.

Предпочтения при кодировании: следование инструкциям, объяснение кода, сложность и эффективность кода, удобочитаемость кода и стиль кодирования.

Для каждого предпочтения 10 принципов, направляющих процесс создания LLM приложений.

CODAL-Bench - эталон для оценки и сравнения соответствия LLM пяти предпочтениям в кодировании.

Purple Llama (Cyberseceval 2, Code Shield), DeVAIC, LLM4Vuln

Component Type	Components	License
Evals/Benchmarks	Cyber Security Eval	MIT
Models	Llama Guard	Llama 2 Community License
Models	Llama Guard 2	Llama 3 Community License
Safeguard	Code Shield	MIT

CYBERSECEVAL-2 - LLM оценивается с двух сторон: склонность к созданию небезопасного кода и способность генерировать программный код для реализации атак.

Code Shield - снижает риск появления небезопасных предложений в коде, предотвращает злоупотребления интерпретатором кода и обеспечивает безопасное выполнение команд.

Результаты тестирования Code LLM по методике CodeUltraFeedback

Model	Instruction Following	Code Explanation	Code Complexity & Efficiency	Code Readability	Coding Style	Average
GPT-4-Turbo	3.79[‡]	4.04[‡]	3.91[†]	4.14[‡]	4.03[‡]	3.98
GPT-3.5-Turbo	<u>3.56</u>	<u>3.76</u>	<u>3.76</u>	<u>3.71</u>	<u>3.66</u>	<u>3.69</u>
WizardCoder-33B	3.29 [‡]	3.31 [‡]	3.43 [‡]	3.44 [‡]	3.49 [†]	3.39
DeepSeek-Coder-33B-Instruct	3.31 [‡]	3.30 [‡]	3.32 [‡]	3.46 [‡]	3.42 [‡]	3.36
DeepSeek-Coder-6.7B-Instruct	3.23 [‡]	3.29 [‡]	3.32 [‡]	3.45 [‡]	3.48 [‡]	3.36
Mistral-7B-Instruct	3.20 [‡]	3.27 [‡]	3.28 [‡]	3.42 [‡]	3.40 [‡]	3.31
CodeLlama-34B-Instruct	3.11 [‡]	3.20 [‡]	3.21 [‡]	3.35 [‡]	3.23 [‡]	3.22
Llama-2-70B-Chat	3.11 [‡]	3.22 [‡]	3.14 [‡]	3.38 [‡]	3.22 [‡]	3.21
WizardCoder-15B	3.11 [‡]	3.03 [‡]	3.12 [‡]	3.27 [‡]	3.16 [‡]	3.14
CodeLlama-13B-Instruct	3.05 [‡]	2.99 [‡]	3.15 [‡]	3.18 [‡]	3.25 [‡]	3.12
CodeLlama-7B-Instruct	2.91 [‡]	3.11 [‡]	3.01 [‡]	3.18 [‡]	3.13 [‡]	3.07
WizardLM-33B	3.07 [‡]	2.98 [‡]	2.91 [‡]	3.11 [‡]	3.10 [‡]	3.03
Llama-2-13B-Chat	2.88 [‡]	2.97 [‡]	2.91 [‡]	3.18 [‡]	2.99 [‡]	2.98
WizardLM-7B	2.63 [‡]	2.61 [‡]	2.51 [‡]	2.69 [‡]	2.64 [‡]	2.62

Методы и средства верификации и валидации интеллектуальных средств генерации программного кода

Специфичные для Code LLM:

- подготовка и контроль корректности данных, используемых для обучения, тестирования и обработки (HumanEval+, InstructHumanEval, MathQA-Python, MBPP+, MultiPL-E, SWE-Bench, APPS, DS-1000, ...);
- контроль целостности моделей
- тестирование склонности к галлюцинированию
- проверка и контроль промтов
- проверка наличия средств и возможностей генерации эксплойтов

Стандартные для созданного с помощью CodeLLM ПО:

Функциональное тестирование, статический и динамический анализ, фаззинг-анализ, тестирование на проникновение, тестирование производительности, надежности и защищенности.

Направления развития и совершенствования методов и средств защиты:

- верификация и валидация на всех этапах ЖЦ (on-line verification)
- подготовка и лицензирование наборов тренировочных данных (SWE-bench)
- использование СИИ для верификации и валидации