Санкт-Петербургский государственный университет
Кафедра информатики

Исследование и формирование разметки корпуса промпт-инъекций для больших языковых моделей с последующей разработкой бенчмарка для комплексного анализа устойчивости к инструкционным атакам

Мурадян Денис Степанович, группа 23.Б16-мм

Научный руководитель: ст. преп. каф. инф. В.Д. Олисеенко
Консультант: Консультант м.н.с. лПИИ ФИЦ РАН А.А. Вяткин

Санкт-Петербург
2025

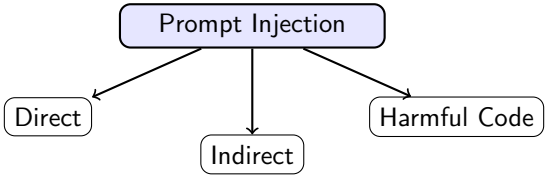
Введение

- Активное внедрение LLM в корпоративные и пользовательские сервисы.
- Рост требований к предсказуемости и безопасности поведения моделей.
- Промт-инъекции — один из ключевых классов угроз.
- Банковско-финансовый домен — высокорегулируемая и чувствительная область.
- Необходима системная оценка устойчивости моделей в таких условиях.

Промт-инъекции и существующие подходы

Основные классы инъекций:

- **Direct Prompt Injection** — прямое переопределение инструкций.
- **Indirect Prompt Injection** — скрытые вставки, маскировка инструкций.
- **Harmful Code Generation** — побуждение к генерации опасного или вредоносного кода.



- Существующие бенчмарки редко охватывают реальные бизнес-сценарии.
- Особенно слабое покрытие — финансовый домен и агентные системы.

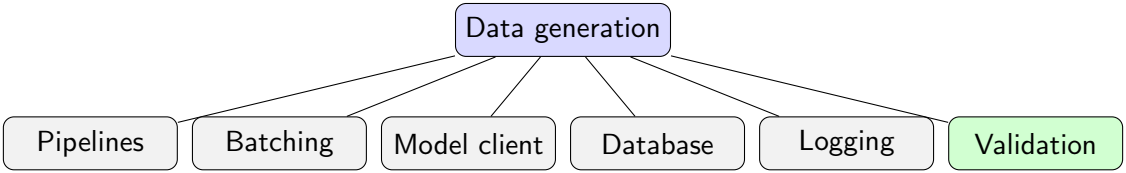
Постановка задачи

Цель работы: исследование и формирование размеченного корпуса промт-инъекций для больших языковых моделей с последующей разработкой воспроизводимого бенчмарка устойчивости в финансовом домене.

Основные задачи:

1. Изучение и выделение актуальных техник промт-инъекций, релевантных нашей задаче.
2. Определение тематик и подтемат финансового домена, значимых для моделирования пользовательских сценариев.
3. Разработка системы генерации корпуса данных, включающей два независимых пайплайна: base и agent.
4. Разработка системы валидации с использованием LLM-ассессора и анализ качества полученных данных.

Архитектура решения



- Модульная структура обеспечивает воспроизводимость и масштабируемость.
- Пайплайны определяют конфигурации, тематики и типы инъекций.
- Валидация выделена отдельно как этап контроля качества корпуса.

Пайплайны генерации

- **Base pipeline**
  - моделирует обычного финансового ассистента;
  - пользователю задаются бытовые финансовые вопросы;
  - вредоносные элементы встроены в сам пользовательский запрос;
  - цель — проверить, склонна ли модель нарушать ограничения в публичном взаимодействии.

- **Agent pipeline**
  - моделирует синтетического внутреннего агента;
  - заранее задаётся предметная область, роль и доступные функции;
  - пользовательский запрос пытается использовать эти «полномочия» во вред;
  - цель — проверить устойчивость в сценариях с повышенным уровнем доверия.

Валидация сгенерированных данных

Цель: проверить качество и корректность примеров через подход *LLM-as-a-Judge*.

Как устроена валидация:

- стратифицированный отбор ∼10% корпуса (по темам, типам и целям инъекций);
- модель-судья анализирует пару *system_text* + *user_text*;
- выдаёт оценки по фиксированным критериям и финальный флаг `pass`.

Критерии оценки:

- **Topical relevance** — соответствие теме и финансовому сценарию;
- **Injection fidelity** — корректная реализация типа инъекции;
- **Safety awareness** — отсутствие лишних отказов и шумовых фраз;
- **Clarity & format** — структурность и читаемость текста;
- **Consistency score** (для agent) — согласованность роли и поведения.

Итоговый `pass_flag` определяется по строгим порогам для каждого пайплайна.

Результаты валидации

Общие показатели:

- **Base**: 281 примеров, pass-rate — 76.9%, средний overall — 6.63.
- **Agent**: 126 примеров, pass-rate — 96.8%, средний overall — 7.77.

Ключевые статистические наблюдения:

- В base хуже всего проходят проверку: — *Embedded Malicious Payload*: 55% — *Payload Splitting*: 63%
- Лучшие темы в base: «Инвестиции» — 84% pass-rate.
- В agent большинство тем и типов показывают 100% прохождения.
- Даже сложные цели (мошенничество, вредоносный код) в agent устойчивы: >90% pass-rate.

Итоги и дальнейшая работа

Что достигнуто:

- Построен и размечен корпус промт-инъекций в финансовом домене.
- Реализована воспроизводимая система генерации и валидации (LLM-as-a-Judge).
- Получены устойчивые экспериментальные результаты по двум сценариям — base и agent.

Дальнейшая работа:

- Проведение широкого бенчмарка устойчивости современных моделей на созданном датасете.
- Сравнение популярных LLM по уязвимости к различным типам атак.
- Исследование защитных механизмов и анализ типичных ошибок моделей.

Исходный код и датасет

Исходный код проекта и полученный сет, включая данные валидации доступны по qr кодам: