

## Исследование возможностей файнтюнинга LLM для решения задачи повышения читабельности декомпилированного кода на языке Си



Выполнили: Кислов Константин Александрович, Божко Артем Александрович,

Ременяко Владислав Денисович, Лялин Максим Андреевич

Наставник: Бехтин Артем Владимирович

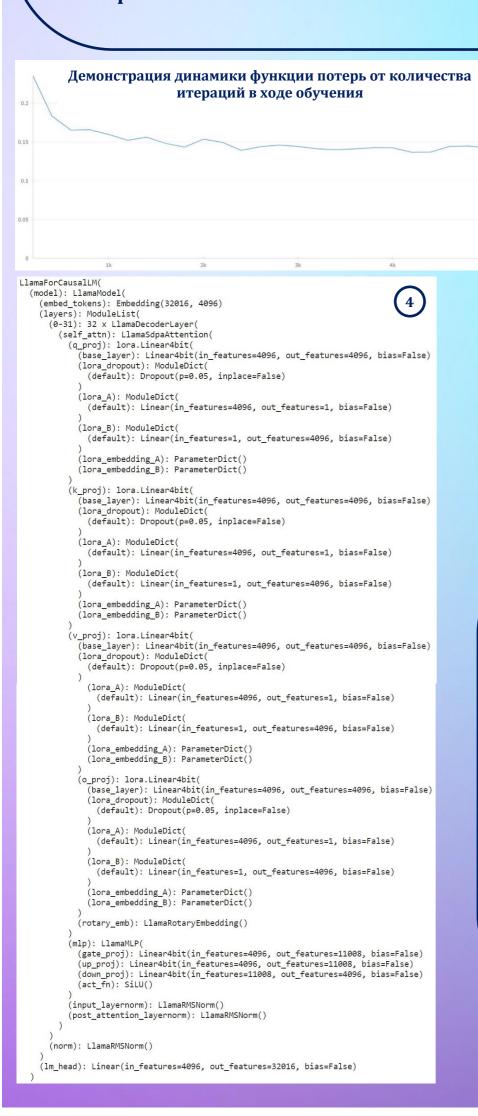
## < Введение >

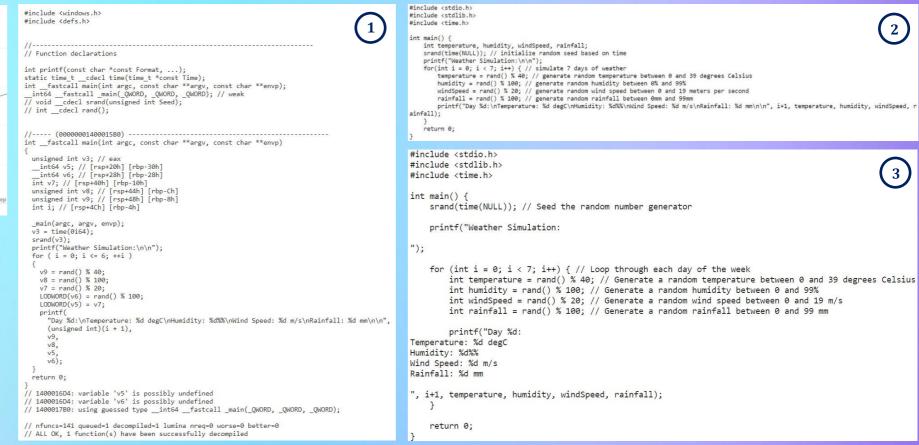
В большинстве случаев декомпилированный программный код трудно поддается анализу: названия переменных и функций лишены изначального заложенного смысла и трудно прослеживается логика программы. В ходе работы над проектом был обучен адаптер для языковой модели CodeLlama, предназначенный для улучшения декомпилированного кода на языке Си: приближения к исходному коду программы и упрощения для человеческого восприятия. Также исследованы возможности адаптера и проведена оценка его эффективности при решении данной задачи.

Цель: разработать адаптер для предобученной языковой модели для решения задачи интеллектуальной обработки (в нашем случае – повышения читабельности) декомпилированного кода, а также оценить его работоспособность.

Методы исследования: поиск, анализ, сравнение, моделирование, программирование, тестирование, измерение.







1) Декомпилятор 2) Оригинал 3) Сгенерированный моделью код 4) Архитектура LLM + адаптер

## < Заключение >

Проведенные тесты показывают, что модель с адаптером, несмотря на относительно небольшие для NLP объем датасета и количество эпох, достигает приличных результатов, в том числе при обработке кода, декомпилированного при помощи программ, примеров вывода которых не было в обучающей выборке (например, RetDec).

В дальнейшие планы работы над проектом входят: продолжение обучения модели на датасете большего объема и с примерами работы других декомпиляторов, проведение более масштабного тестирования как с классическими для задачи seq2seq метриками (BLEU, AED и т.д.), так и с оценкой при помощи опроса специалистов и с проверкой возможности перекомпилирования результатов работы нейросети.



< To be continued... >