# Сравнение версий Qwen2.5-1.5B-Instruct на датасете GLUE/CoLA

Афанасьев Алексей Игоревич 217 группа

18.04.2025

## 1 Методология

Исследование сравнивает три версии модели Qwen2.5-1.5B-Instruct:

- Базовая инструктивная версия (без дообучения)
- Версия с LoRA+Unsloth (rank=16)
- Версия с LoRA+Unsloth (rank=32)

### 1.1 Конфигурация эксперимента

- Датасет: GLUE/CoLA (Corpus of Linguistic Acceptability)
- Размер тестовой выборки: 200 примеров
- Метрики: Accuracy (точность)
- Аппаратное обеспечение: GPU T4 (15GB памяти)
- Библиотеки: Unsloth, Transformers, PEFT
- Модель: **Qwen2.5-1.5B-Instruct**

## 2 Результаты

Таблица 1: Сравнение метрик на датасете GLUE/CoLA

| Версия модели | Accuracy | MCC | Leaderboard Score |
|---|---|---|---|
| Qwen2.5-1.5B-Instruct | 0.745 | 0.499 | 0.622 |
| Qwen2.5-1.5B-Instruct-LoRA-16 | 0.760 | 0.511 | 0.636 |
| Qwen2.5-1.5B-Instruct-LoRA-32 | 0.765 | 0.514 | 0.640 |

## 3 Детали обучения

### 3.1 Параметры LoRA

Таблица 2: Параметры адаптеров LoRA

| Параметр | LoRA-16 | LoRA-32 |
|---|---|---|
| Rank | 16 | 32 |
| Lora Alpha | 16 | 16 |
| Dropout | 0.0 | 0.0 |
| Target Modules | q_proj, v_proj, k_proj, o_proj | |

### 3.2 Параметры обучения

- Batch size: 1 (с накоплением градиента 8 шагов)
- Learning rate: 5e-5
- Эпохи: 1
- Планировщик: Cosine с разогревом (16 шагов)
- Оптимизатор: Paged AdamW 8-bit

## 4 Выводы

- LoRA адаптеры обеспечили прирост точности на 1.5-2% по сравнению с базовой моделью
- Увеличение rank с 16 до 32 дало дополнительный прирост в 0.5%
- Версия с rank=32 показала наилучшие результаты
- Обучение с rank=32 заняло 8 минут 39 секунд для 900 примеров, что почти вдвое дольше чем с rank=16