Улучшение FineTuning LLM с помощью Multi Token Prediction

Мостовых Егор
Научный руководитель: Богданов Кирилл

18 мая 2025

План презентации

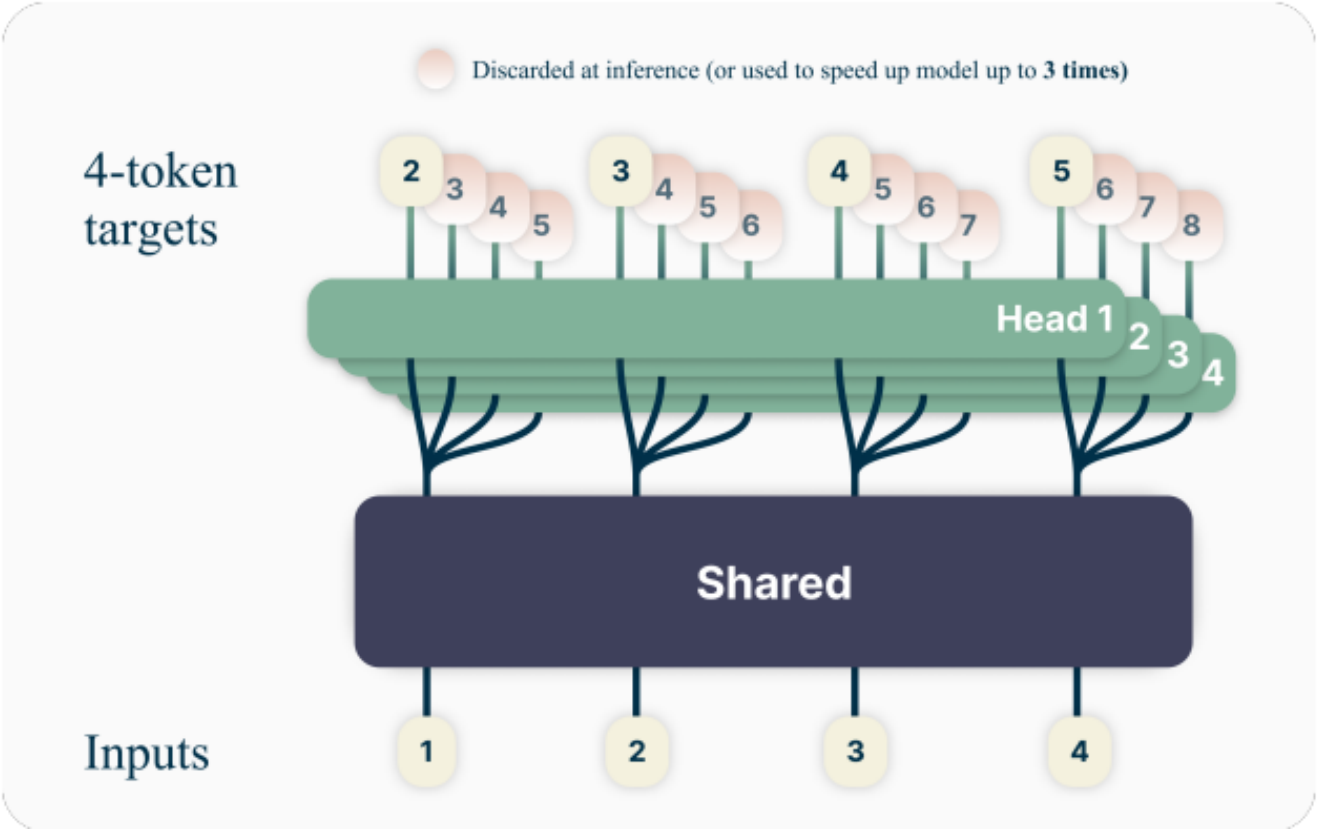
- Про то, как эта темы вытекает из исследований в этой области
- Суть идеи диплома
- Результаты экспериментов
- Гипотезы, анализ результатов

Обзор литературы:

Спекулятивный декодинг (2022)

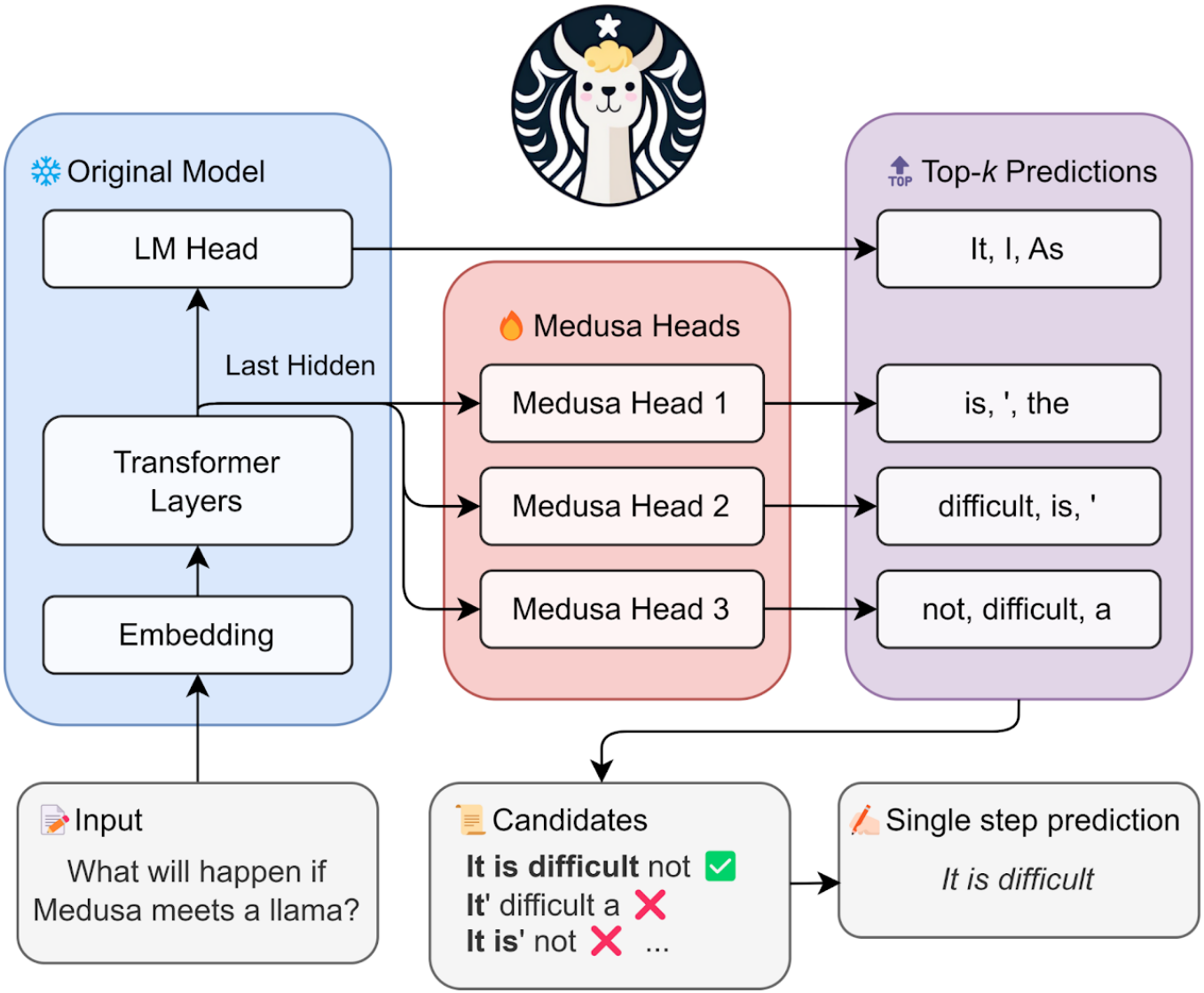
Обзор литературы:

Multi Token Prediction (2024)



Обзор литературы:

MEDUSA(2024) & EAGLE(2024)



Идея диплома:

Проверяемая гипотеза:

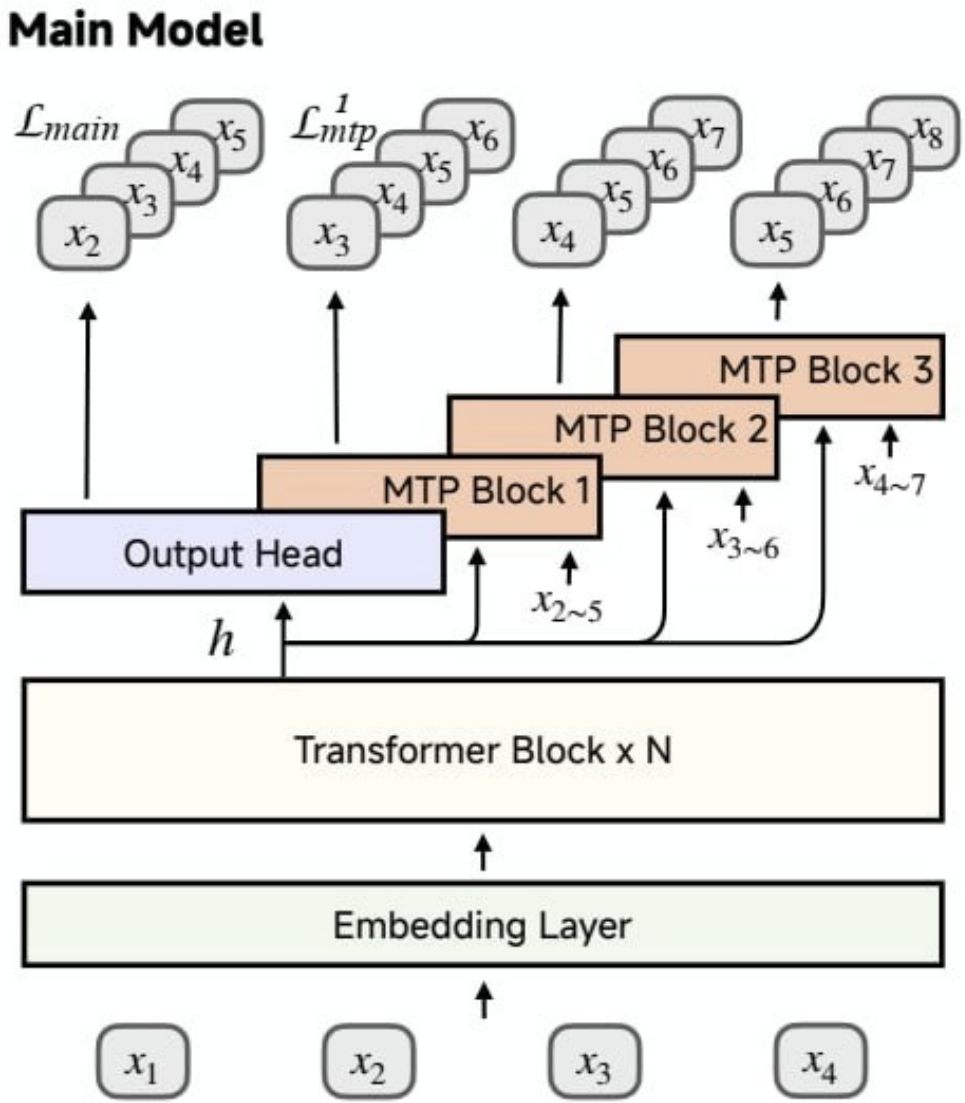
Использование Multi Token Prediction улучшит/ускорит дообучение предобученной модели под новую задачу.

Как именно это будет работает:

- Поступает предобученная модель
- Добавляем к ней несколько спекулятивных голов, обученных для архитектуры модели заранее нами
- Обучаем модель используя лоссы от главной и спекулятивных голов.

На выходе хотим получить более качественную основную модель.

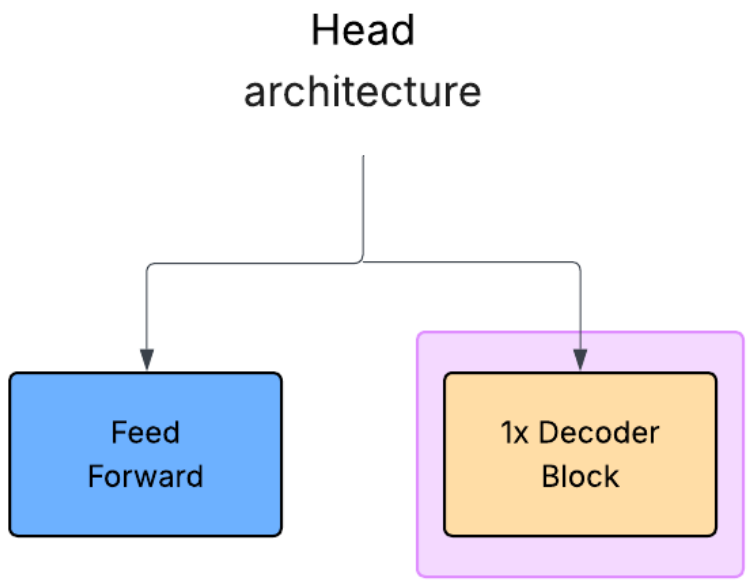
Идея диплома:



Эксперименты:

Архитектура голов

- Качество спекулятивной головы сильно влияет на обучение.
- Пробовали использовать как линейный слой, так и трансформерный блок — последний даёт более точный результат.



Эксперименты:

Loss функции

SumLoss:

Loss = LossMainHead + LossSpecHead

WheightedLoss:

Loss = t * LossMainHead + (1 - t) * LossSpecHead

TwoBackward:

LossMainHead.backward()
LossSpecHead.backward()

ChangeLoss:

Change loss every epoch or every N iterations

Эксперименты:

Датасеты

GSM

- Школьные задачи по математике с правильными числовыми ответами и объяснениями

InstructV3

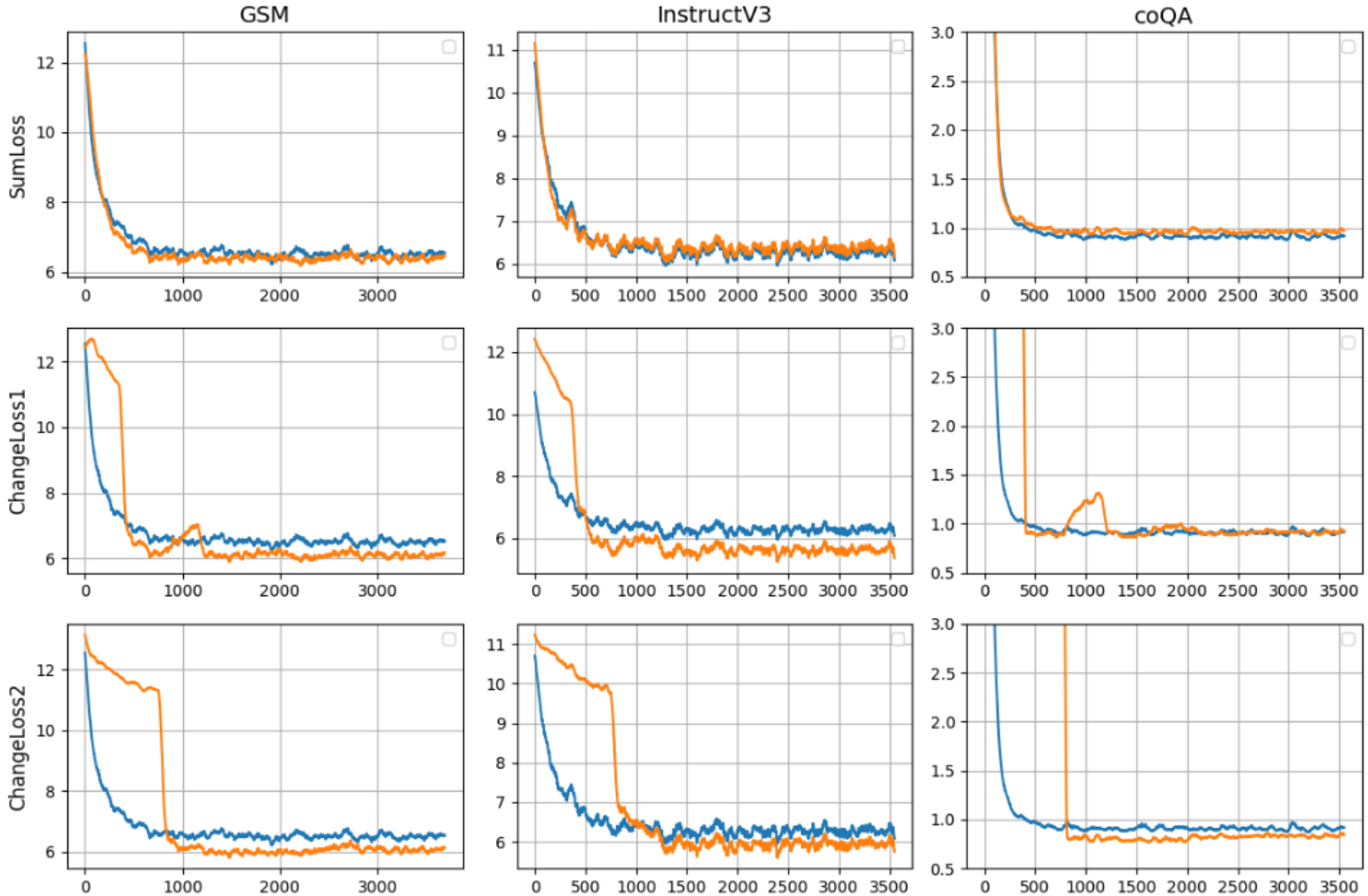
- Инструкции и их исполнения — достаточно длинные промпты математических и логических инструкций

CoQA

- Разговорные вопросы и ответы по приведённому контексту

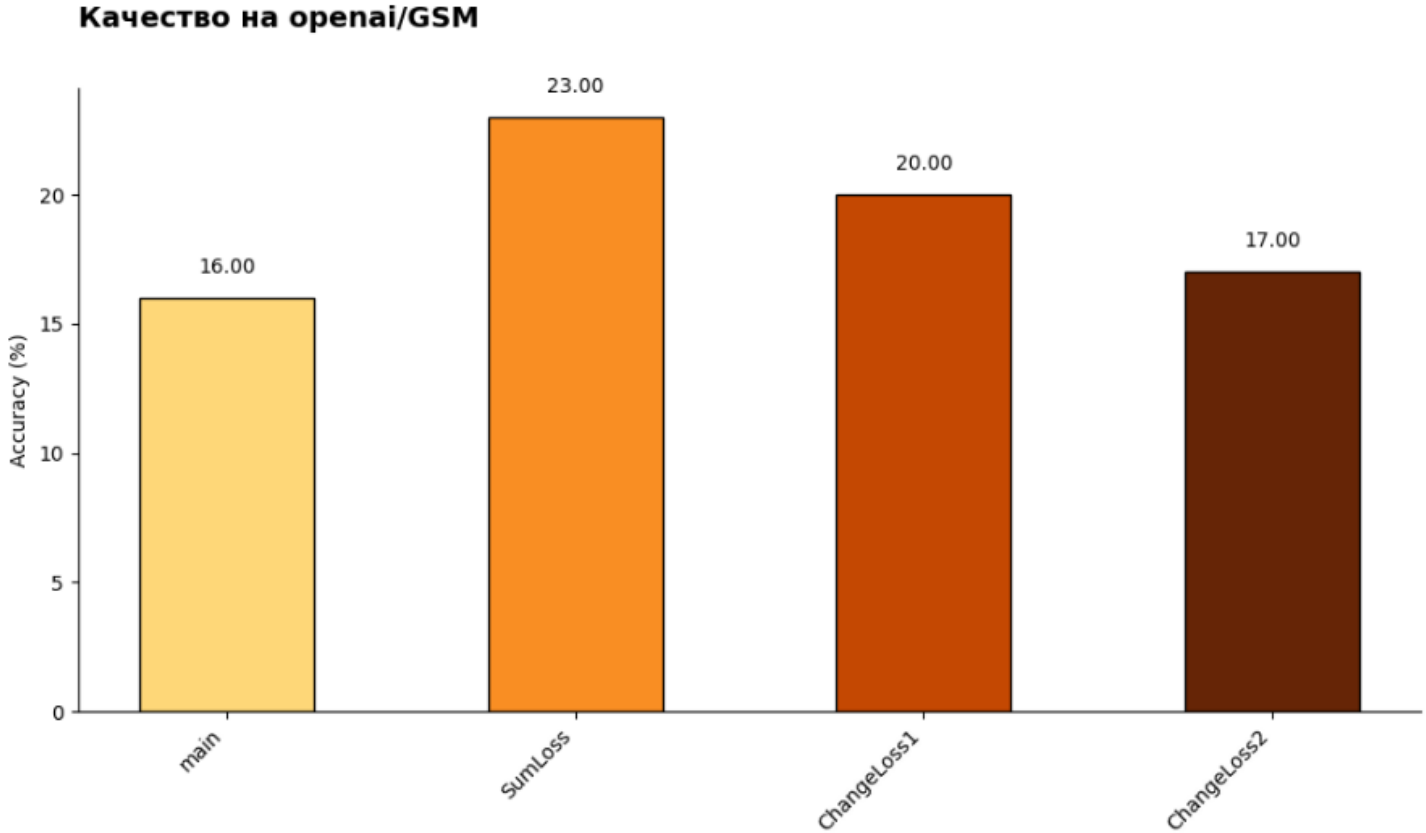
Эксперименты:

Llama-3.2-1B с одной доп. головой



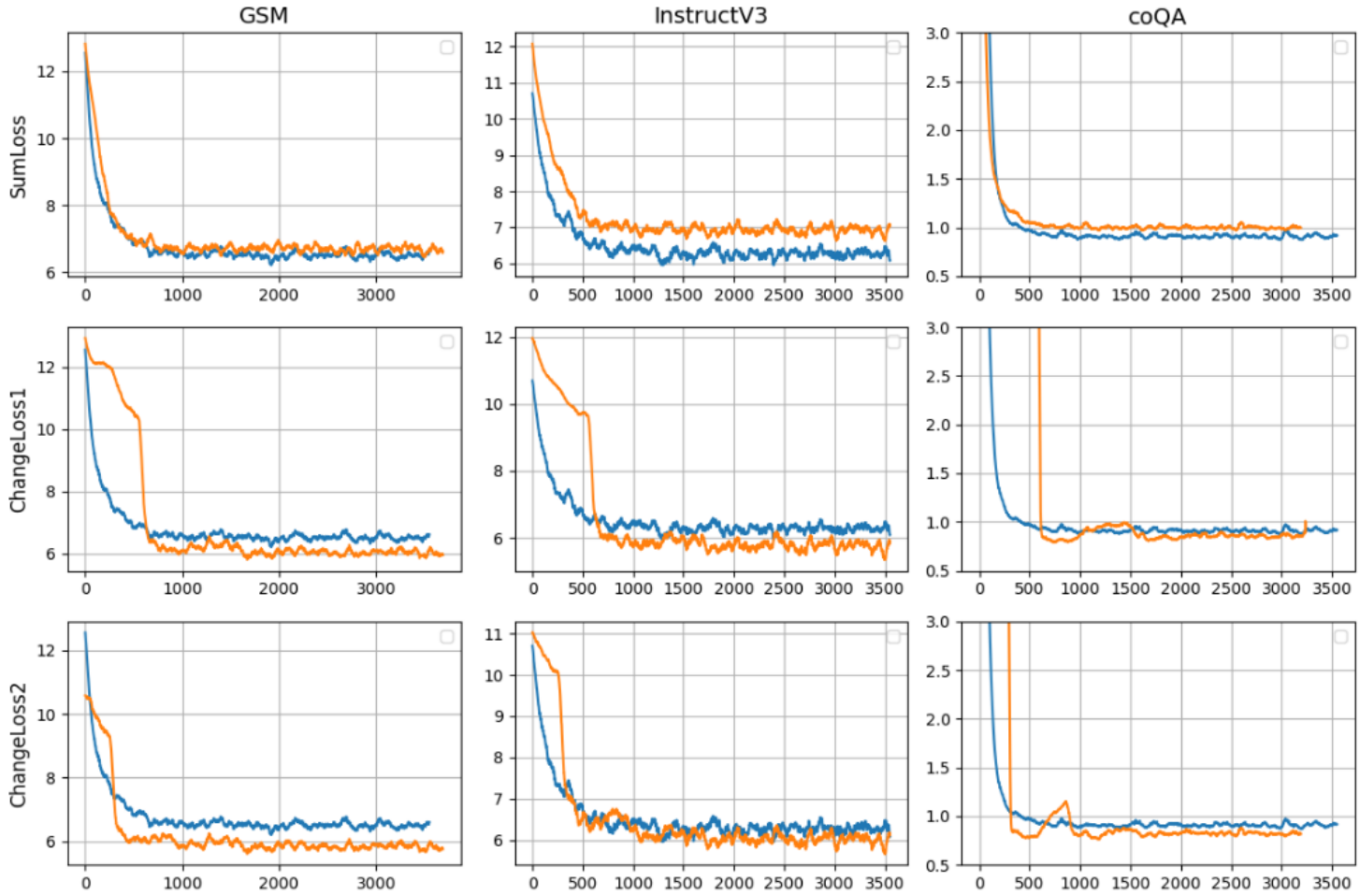
Эксперименты:

Llama-3.2-1B с одной доп. головой



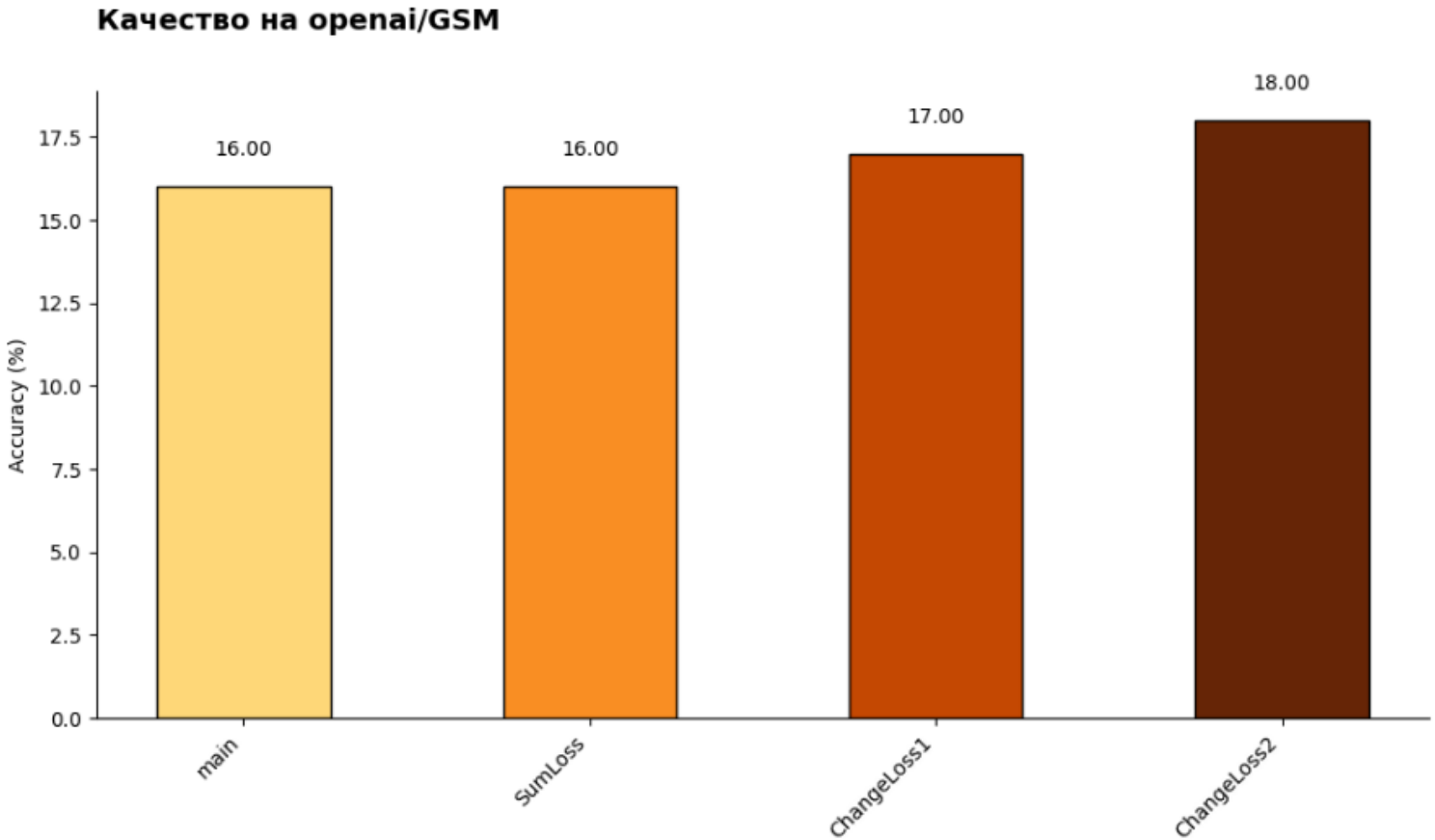
Эксперименты:

Llama-3.2-1B с двумя доп. головами



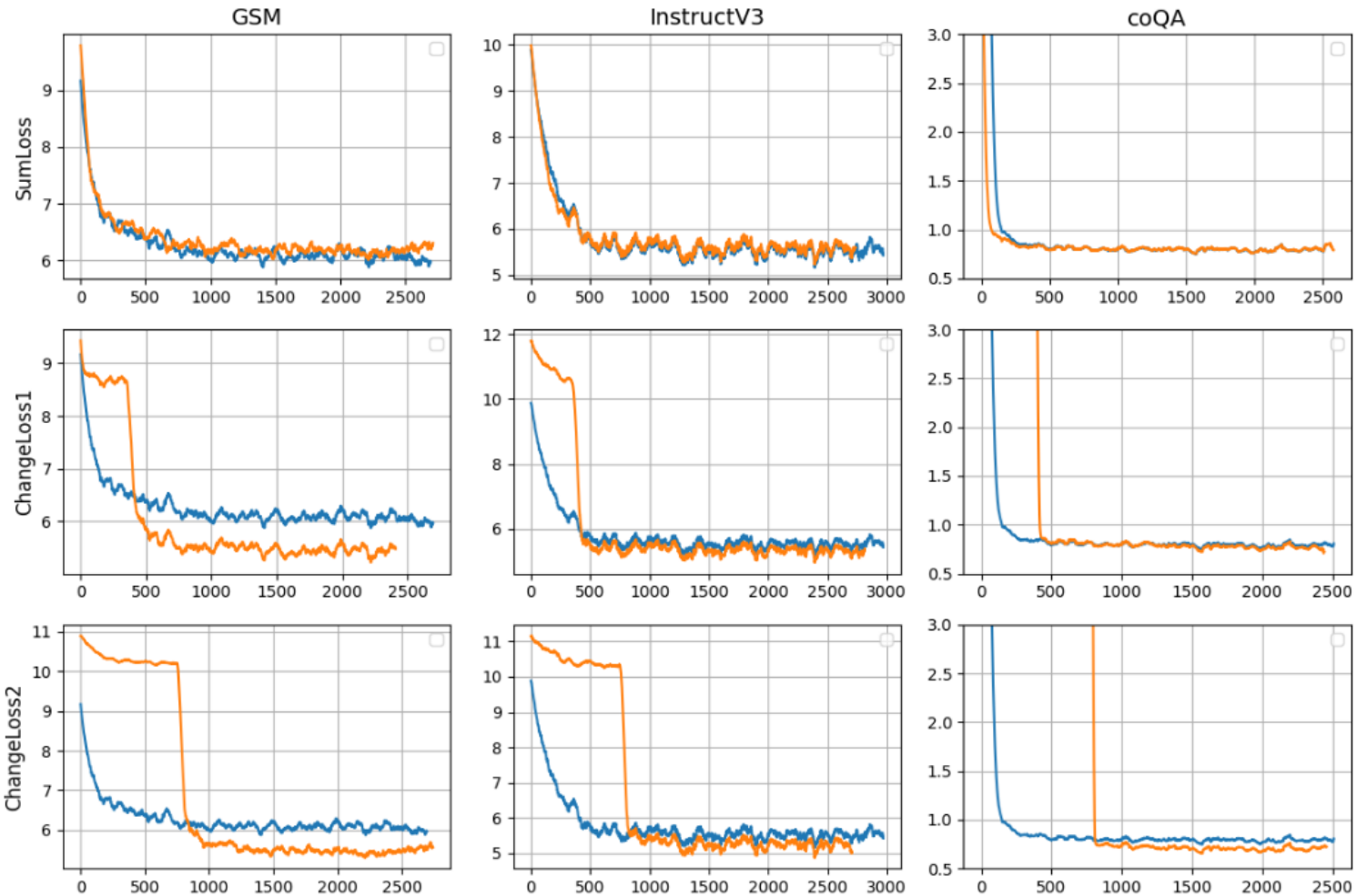
Эксперименты:

Llama-3.2-1B с двумя доп. головами



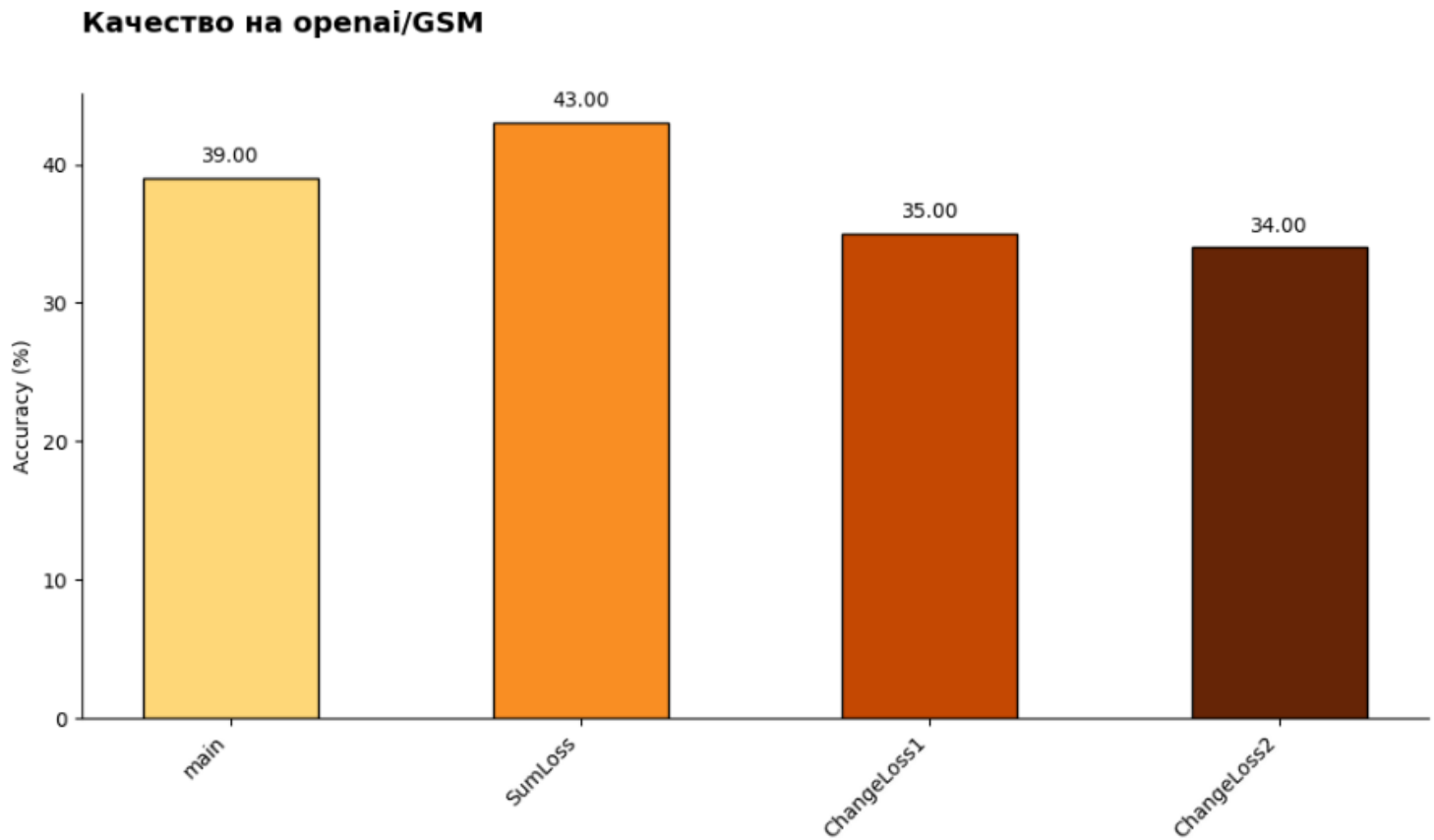
Эксперименты:

Llama-3.2-3B с одной доп. головой



Эксперименты:

Llama-3.2-3B с одной доп. головой



Оценка результатов, выводы:

- В основном, дообучение с несколькими головами даёт выигрыш в качестве получаемой модели
- В больших моделях повышаются требования к обучению доп. голов
- Стоит попробовать больше методов FineTuning-а, посмотреть на эффект при дообучении большего числа параметров.

Last slide.