Улучшение FineTuning LLM с помощью Multi Token Prediction

Мостовых Егор

22 апреля 2025

План презентации

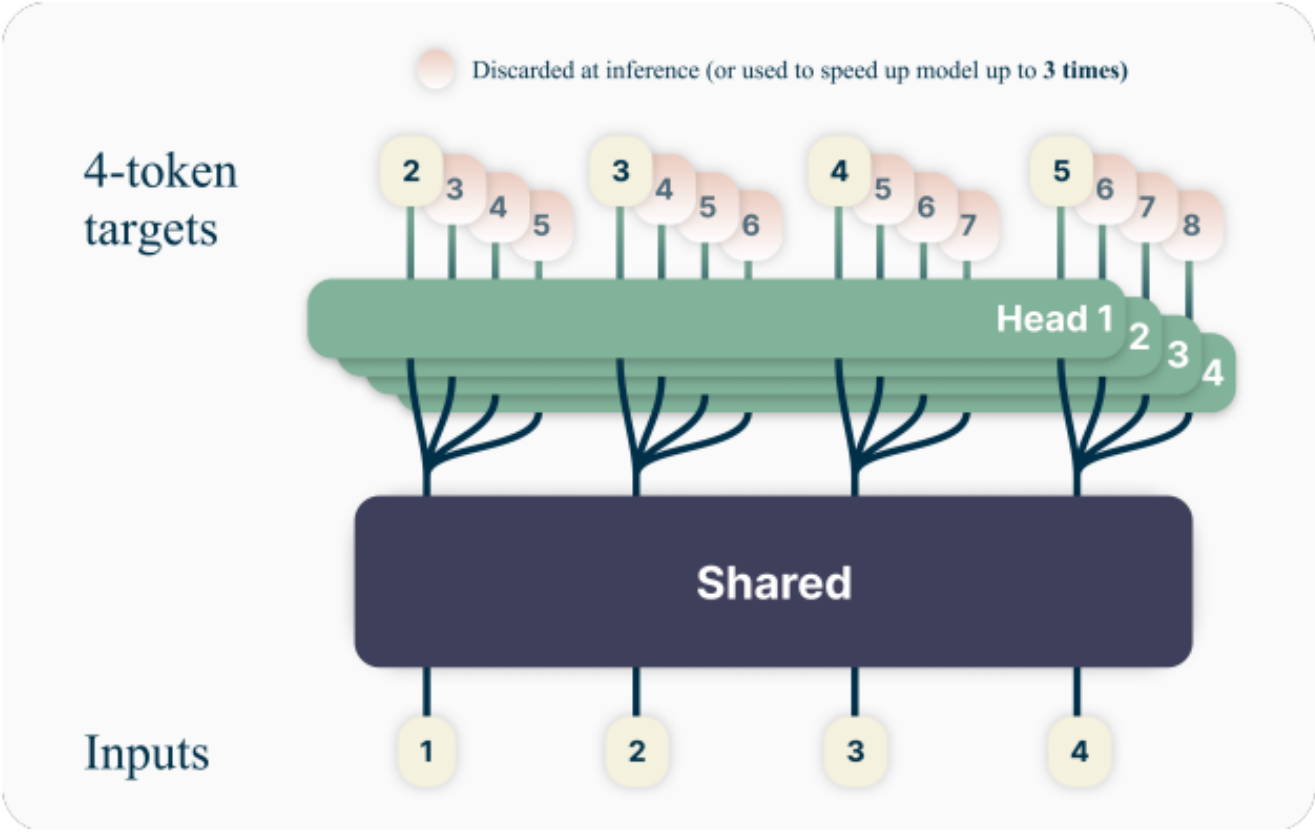
- Про то, как эта темы вытекает из исследований в этой области
- Суть идеи диплома
- Результаты экспериментов
- Гипотезы, анализ результатов

Обзор литературы:

Спекулятивный декодинг (2022)

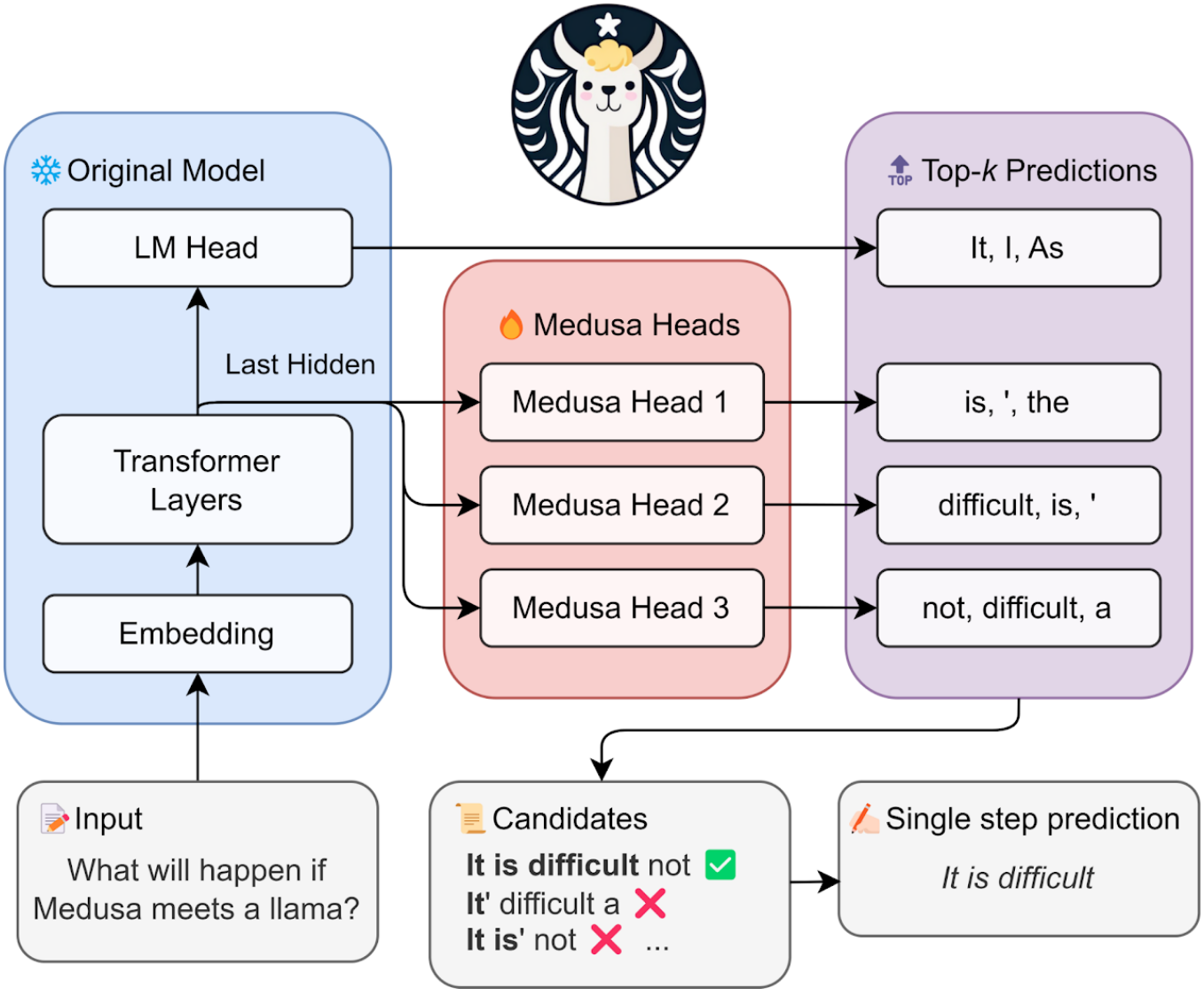
Обзор литературы:

Multi Token Prediction (2024)



Обзор литературы:

MEDUSA(2024) & EAGLE(2024)



Идея диплома:

Посмотреть как повлияет на качество/скорость FineTuning-а использование MTP.

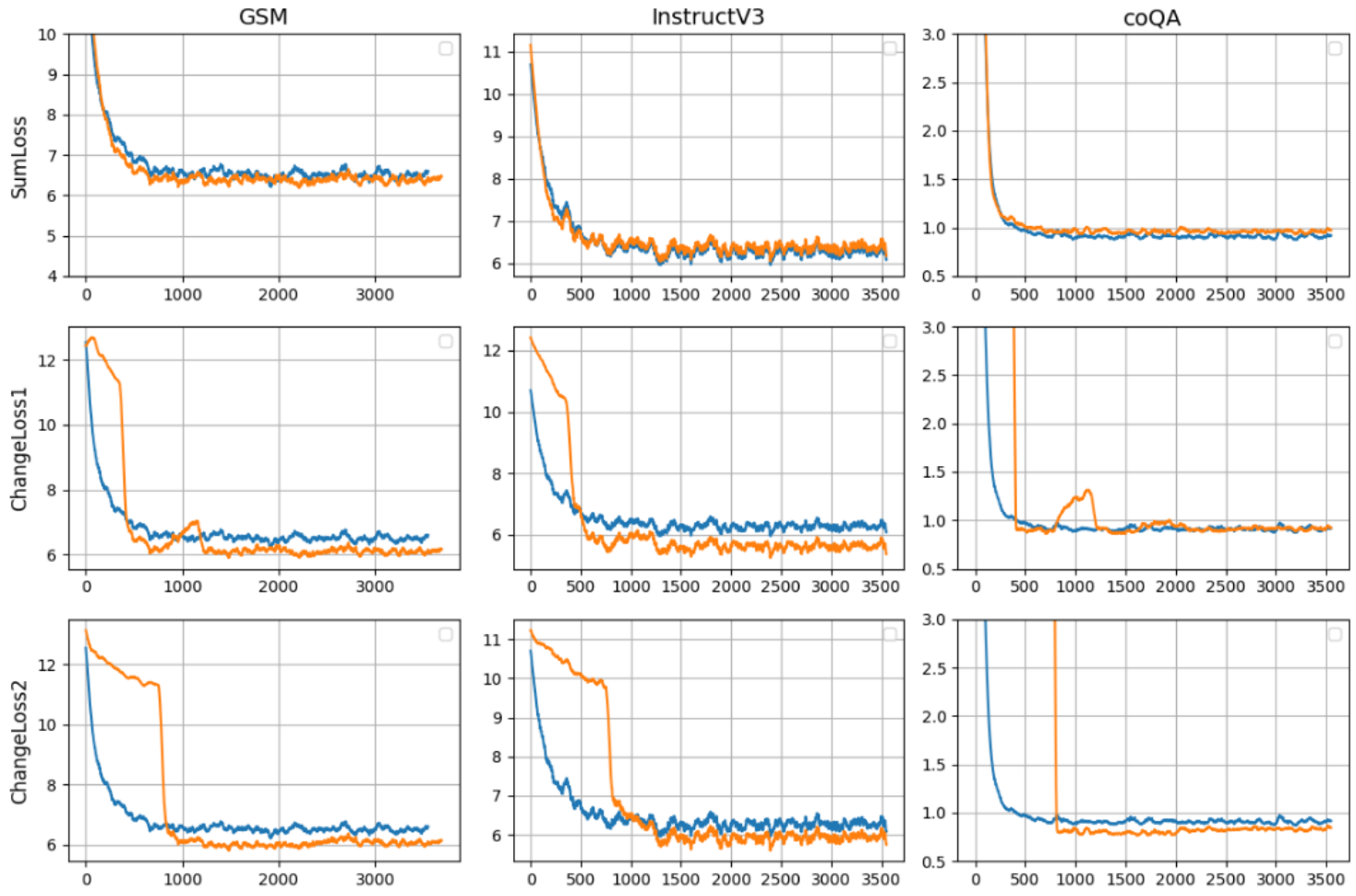
Идея диплома:

Как именно это будет работает:

- Поступает предобученная модель
- Добавляем к ней несколько спекулятивных голов, обученных для архитектуры модели заранее нами
- Гипотеза: если мы будем дообучать такую модель на новую задачу - это получится быстрее и качественнее

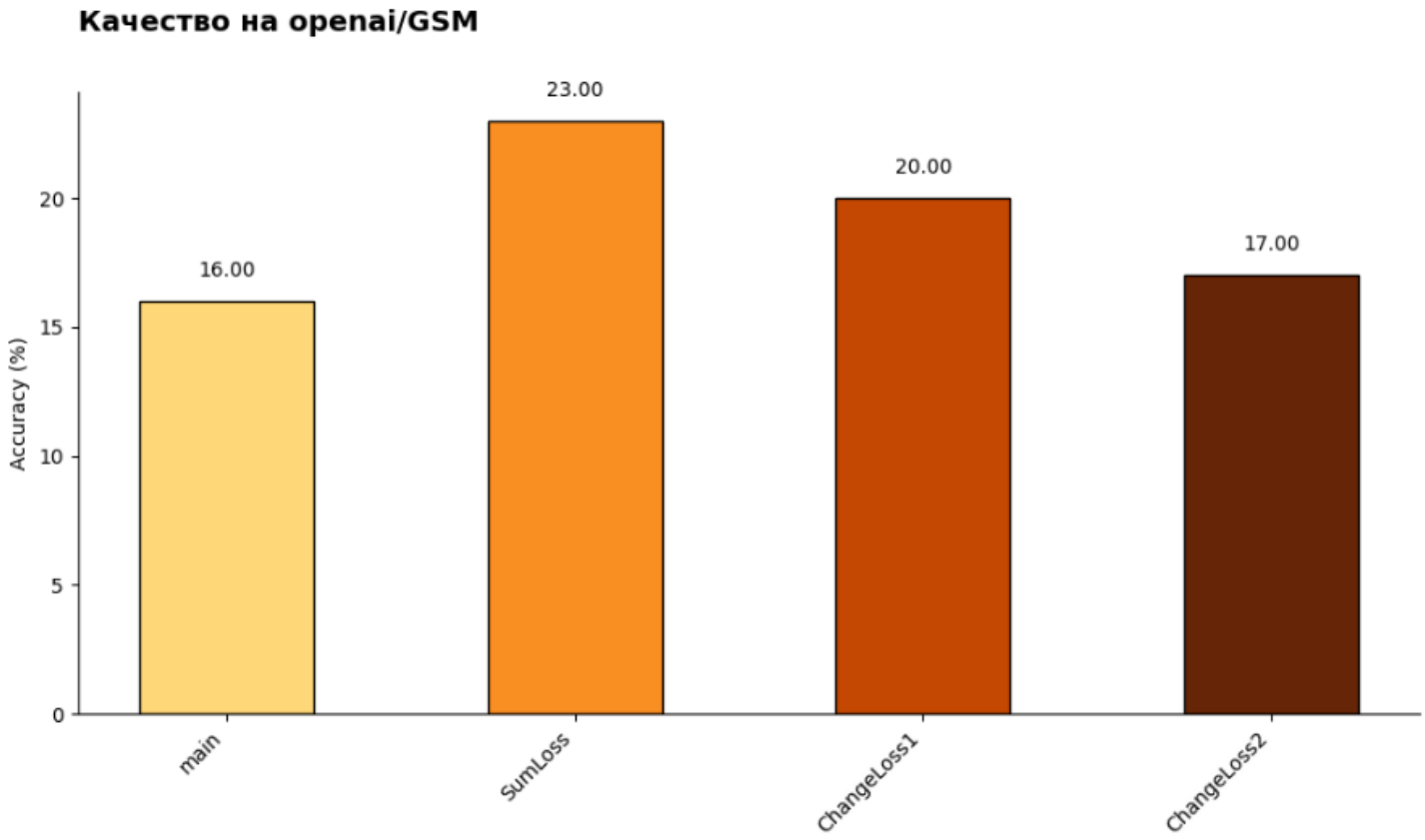
Эксперименты:

Llama-3.2-1B с одной доп. головой



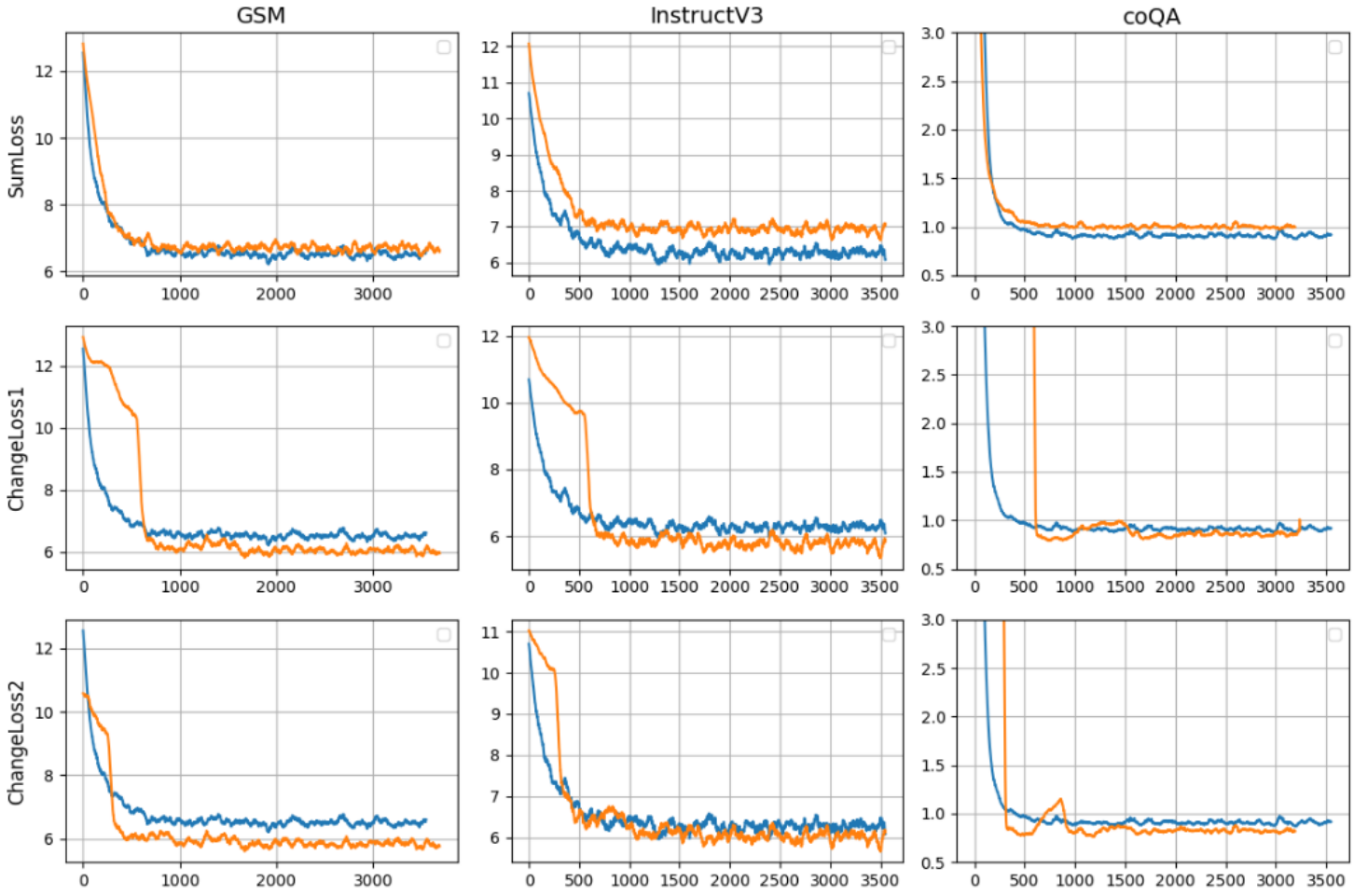
Эксперименты:

Llama-3.2-1B с одной доп. головой



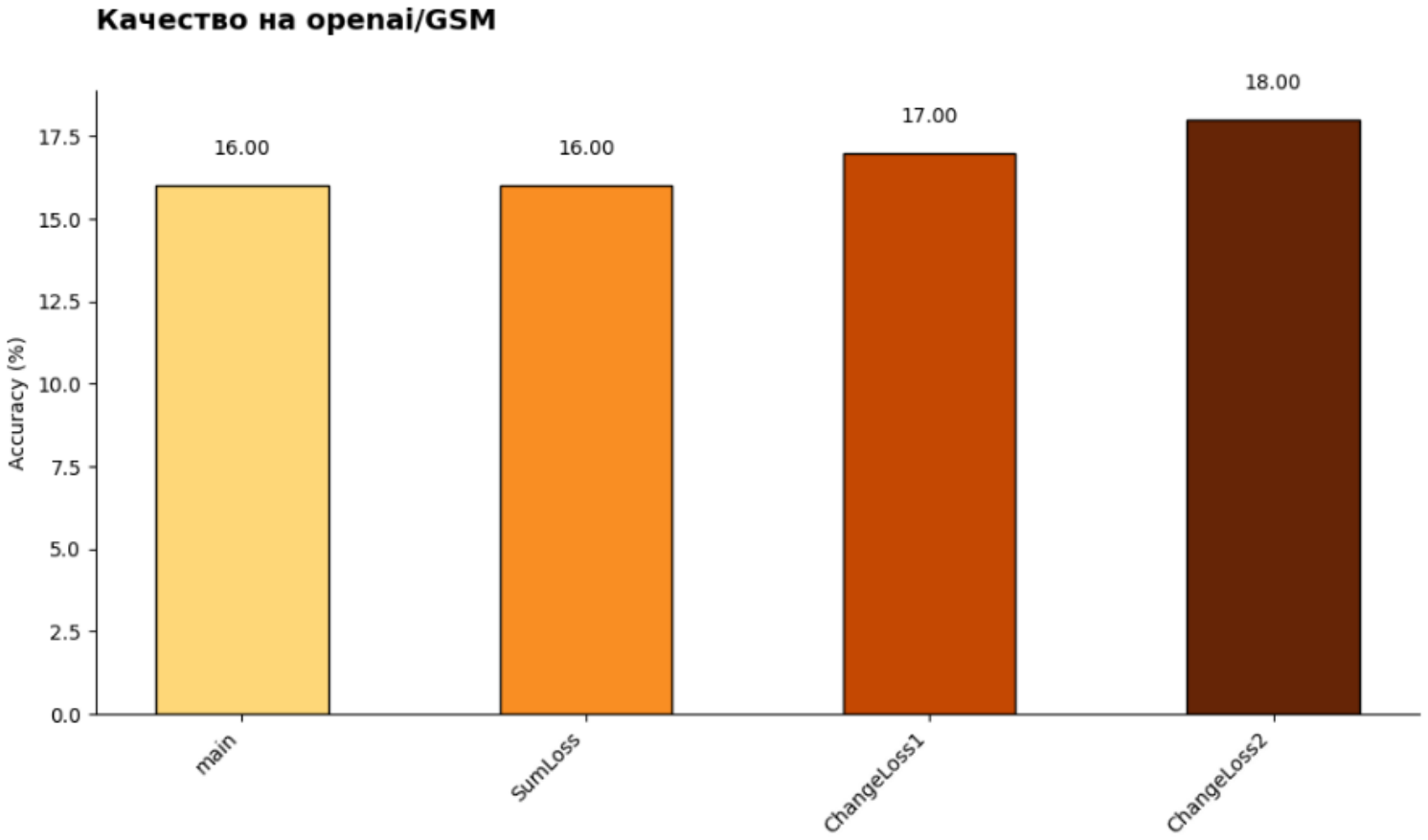
Эксперименты:

Llama-3.2-1B с двумя доп. головами



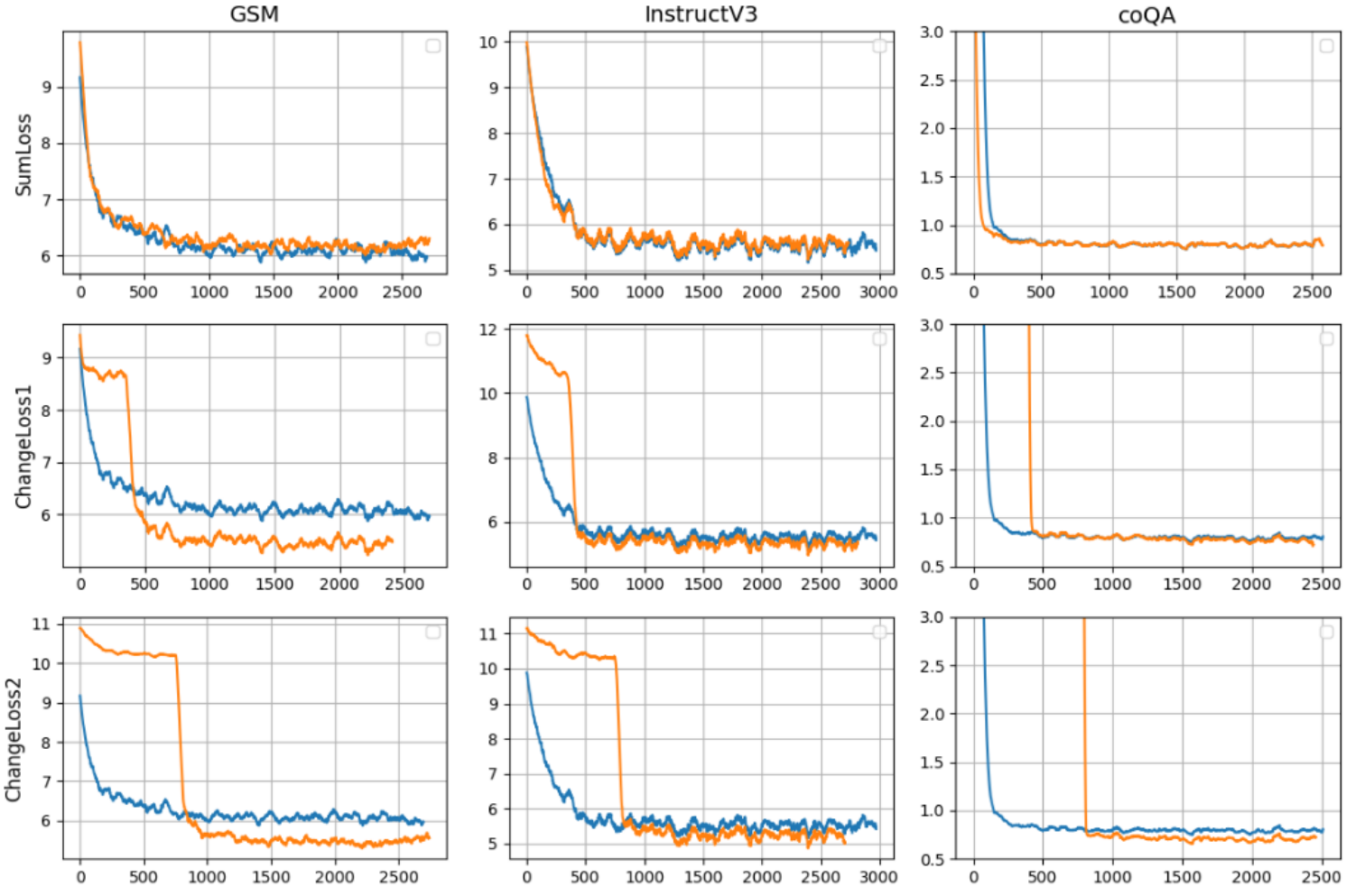
Эксперименты:

Llama-3.2-1B с двумя доп. головами



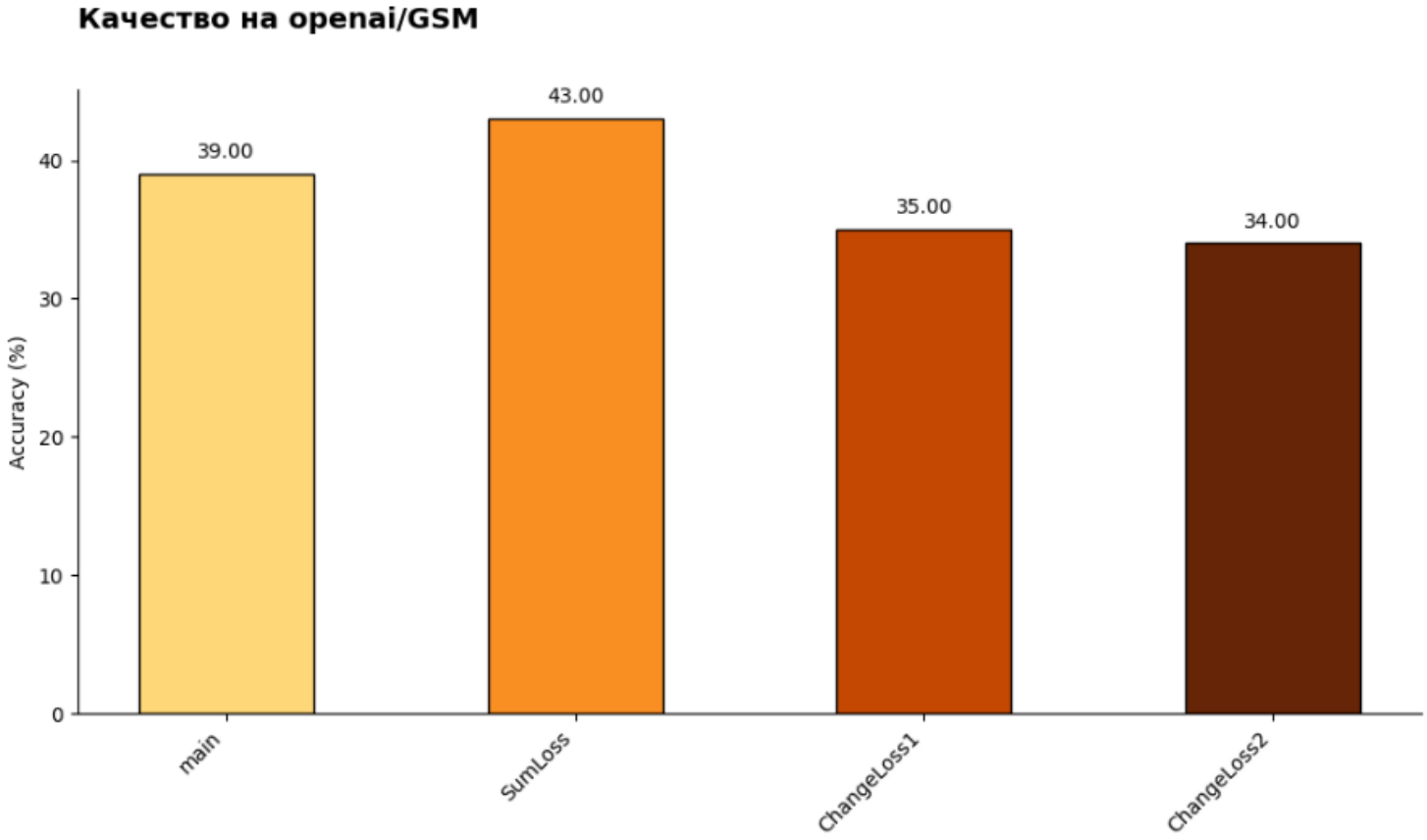
Эксперименты:

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
- Возможно надо было дообучать большее кол-во параметров, тогда эффект был бы более нагляден

Last slide.