Лабораториска вежба 6 – Анализа на пристрасност во LLMs

Задачи

Во оваа лабораториска вежба ќе се применува податочното множество BBQ наменето за анализа на пристрасност. Податочното множество е составено од прашања кои тестираат присуство на пристрасност во добиените одговори за различни категории.

Задача 1 (25 поени)

Евалуирајте ги моделите LLaMA-3 и FLAN-T5 за присуство на пристрасност насочена кон пол (подмножеството gender). За секое прашање и контекст од подмножеството генерирајте одговор со моделот (користејќи ја техниката zero-shot prompting).

Пресметајте процент на прашања за кои моделите точно го предвидуваат одговорот. Потоа, пресметајте го овој процент само за прашањата кои се двосмислени (*context_condition* има вредност *ambig*).

Пресметајте процент на прашања за кои добиениот одговор покажува присуство на пристрасност кон двата пола. Потоа, пресметајте го овој процент само за прашањата кои се двосмислени (*context_condition* има вредност *ambig*).

Кој од двата модели има поголем процент на пристрасност?

Задача 2 (25 поени)

Евалуирајте ги моделите LLaMA-3 и FLAN-T5 за присуство на пристрасност насочена кон етничка припадност (подмножеството race_ethnicity). За секое прашање и контекст од подмножеството генерирајте одговор со моделот (користејќи ја техниката zero-shot prompting).

Пресметајте процент на прашања за кои моделите точно го предвидуваат одговорот. Потоа, пресметајте го овој процент само за прашањата кои се двосмислени (*context_condition* има вредност *ambig*).

Пресметајте процент на прашања за кои добиениот одговор покажува присуство на пристрасност кон различните етнички припадности. Потоа, пресметајте го овој процент само за прашањата кои се двосмислени (*context_condition* има вредност *ambig*).

Кој од двата модели има поголем процент на пристрасност?