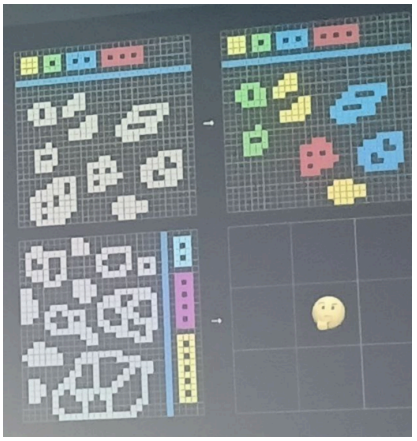
Где ждать новых прорывов в ИИ?

Вопрос о дальнейшем развитии ИИ возник в связи с тем, что модель OpenAI хорошо решала бенчмарк на раскрашивание фигур в зависимости от количества отверстий, а также зацепки фигуры друг за друга друга (ARC-AGI 2).



Этот толчок навел на мысли, что AGI уже совсем близко и делать больше нечего.

Но на самом деле зачастую лишь кажется, что все пропало, **главное не бросать то, что начали**, ведь по такому принципу OpenAI вышли вперед, используя маргинальные способы.

Но куда же сейчас двигается прогресс?

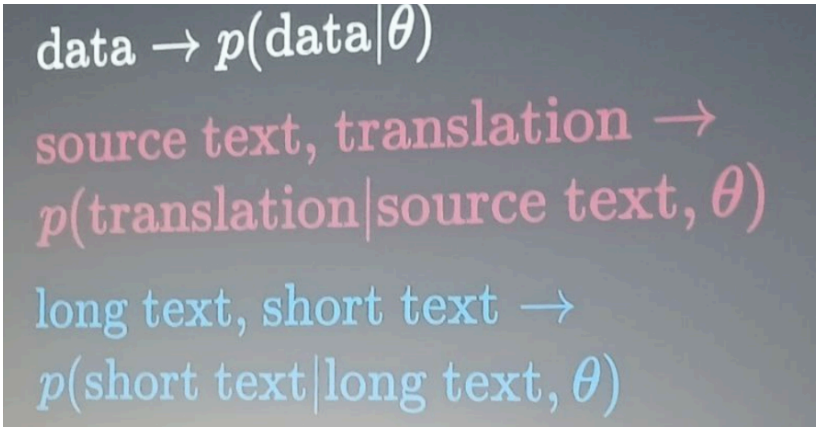
Ощущение того, что потолок достигнут, появляется уже не впервые. Ранее это случалось когда вышел ChatGPT.

Но почему так происходит?
Дело в том, что мы берем уже изученные данные и пытаемся понять как они устроены, чтобы найти решение конкретной задачи. Таким образом мы учим модель **выполнять конкретные понятные задачи, а не искать новые решения.**

Происходит это по данной схеме:

$$\text{data} \rightarrow p(\text{data}|\theta)$$

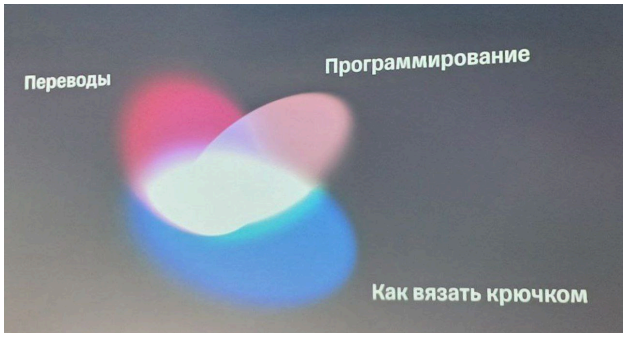
Но это не говорит о том, что мы имеем узкий спектр решаемых задач.
Многие типы задач схожи между собой по алгоритмам и решаются похожим образом, но выполняют различные полезные вещи. Таким образом, чтобы улучшить качество работы LLM, нужно файнтюнить каждую ее задачу по отдельности, а не модель в целом.



Подобным примером становятся задачи машинного перевода и суммаризации.

В случае с машинным переводом модель может обучаться при помощи данных из книг или других источников, содержащих художественные тексты. В случае с суммаризацией задача та же, но появляются проблемы с датасетом. Это одна из предпосылок к будущим сдвигам в развитии ИИ.

Ну окей, есть данные, есть множество решаемых задач, а чего не хватает то?



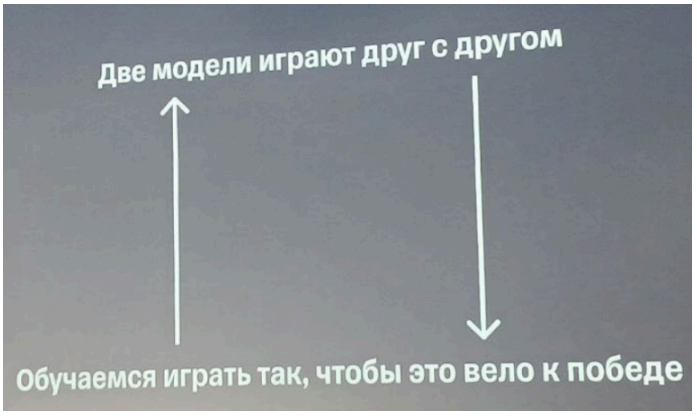
А проблема в том, что мы можем обучить модель на всех данных интернета, но она не будет понимать, где же искать ответ на еще нерешенные задачи.

Так как же научиться решать задачи, которые еще никто не решал?

Тут то и начинается прогресс, а именно **поиск новых методов**, поскольку текущие не позволяют выйти нам за пределы наших знаний и добиться чего-то нового.

Где же искать эти новые методы?

В истории уже есть случаи, когда разум человека был превзойден ИИ. Один из таких случаев это игра Go, когда самые сильные игроки проигрывали машинам.

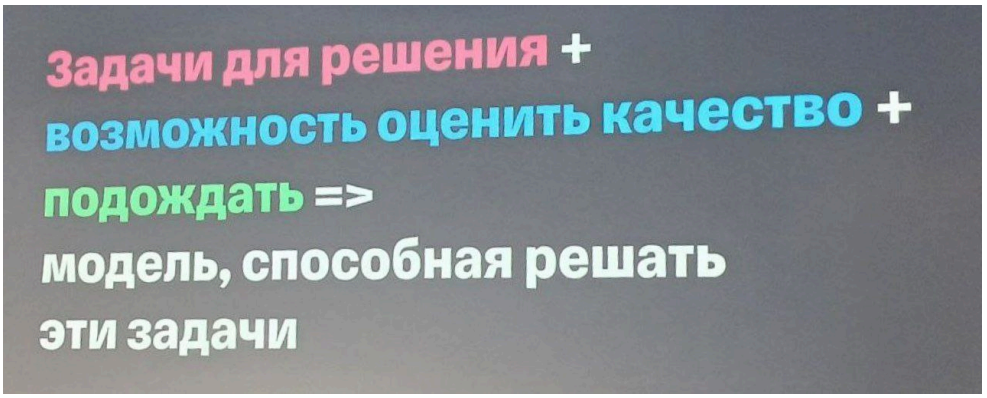


Такие модели обучаются путем игры друг с другом, что является обучением с подкреплением или ***RL***. Они соревнуются и ищут наиболее выгодные стратегии, благодаря чему и добиваются таких результатов.

Это переворачивает представление о том что **нам не нужны все данные для решения**, а **нужны на самом деле задачи и проверка качества**.

Таким образом можно например доказывать теоремы путем перебора и критики сгенерированного решения. (Для этого нужны еще очень крутые мощности, поэтому **нельзя не отметить роль инженерного дела в прогрессе)**

Соответственно дальше верификация не будет ограничена нашими решениями и знаниями.
мы теперь ограничены только вычислениями и мощностями.



Чем же стоит заниматься чтобы оказаться на волне?

- Инженерия - нужны мощности
- Реализация верифицируемых задач (давать например доказательство теорем или простых задач)
- Аккуратные методы обучения на этом

Нет принципиально нерешаемых задач!

и вообще можно решать любые задачи :)
делать больше - работает