Где ждать новых прорывов в области искусственного интеллекта

Введение

Лекция посвящена обсуждению смены парадигмы в развитии искусственного интеллекта (ИИ). Несмотря на впечатляющие успехи, становится ясно: для настоящего прорыва нужно больше, чем просто масштабирование моделей и данных. Развитие идёт к новым архитектурам, методам обучения и способам верификации знаний.

1 Иллюзия завершённости: почему вопрос встал сейчас

- Недавние модели OpenAI и Anthropic решают сложные задачи, такие как backmark и раскраска фигур — ранее считавшиеся трудными для ИИ.
- Возникает ощущение, что ИИ уже достиг предела, что AGI близок.
- Однако подобные ощущения уже возникали раньше и каждый раз оказывались слишком ранними для вывода.

2 Прорыв ChatGPT: почему казалось, что всё сделано

- ChatGPT решает множество задач: перевод, суммаризация, генерация кода — задачи, которые раньше требовали отдельных подходов.
- Классический подход заключается в простом обучение модели на данных $data \to p(data|\theta)$.
- Пример: машинный перевод — обучаем $p(\text{translation}|\text{source}, \theta)$.
- Проблема: датасеты для суммаризации найти трудно, сложные графовые структуры и эвристики были неэффективны.
- ChatGPT внезапно решает это просто и качественно.
- В чём же суть прорыва? Обобщённая модель, обученная на разнообразных интернет-данных, научилась выполнять множество задач, даже тех, на которых явно не обучалась.

3 Ограничения и границы LLM-подхода

- Обучение на всём интернете не включает задачи, решений которых ещё нет (например, гипотезы тысячелетия).
- Мы не можем «попросить» модель решить то, чего никто никогда не решал — без новых механизмов это невозможно.
- Важно: способность решать задачи вне датасета — ключевая цель развития.

- Поэтому мы используем искусственный интеллект только для упрощения бытовых вещей, а не для создания прорывных шагов в развитии нашей науки.

4 Обучение с подкреплением и верификация

- AlphaGo и AlphaTensor — примеры, когда модель обошла человека.
- Идея: две модели играют друг с другом, и обучаемся по стратегии победителя.
- Схема: задача + возможность оценить результат + ожидание → модель, способная решать.
- Области: доказательство теорем, математика, логика, программирование.
- Мы не знаем ответ заранее, но можем проверить его корректность — идеально подходит для RL.

5 Эволюция парадигмы обучения

- Раньше: каждая задача обучалась отдельно.
- Затем: большая модель → дообучение на задачах.
- Сейчас: одна модель решает многое.
- Будущее: обучение модели, которая сама умеет обучаться под задачи.
- Мы находимся именно на этом этапе.

6 Почему RL перспективен

- Верификация результата может быть вне человеческих знаний (пример: ход 37 в AlphaGo).
- Ограничение теперь — не знание, а вычисления (время, видеокарты).
- RL позволяет решать ранее нерешённые задачи.

7 Текущие вызовы

- Инженерия высоконагруженных систем: обучение на миллионах задач требует продуманной архитектуры.
- Создание и генерация проверяемых задач — шахматы, логические игры, формальные языки.
- Нужно варьировать сложность, генерировать классы задач и формализовывать проверку.

8 Аккуратные методы обучения

- Разные задачи обучаются с разной скоростью, обучение может “развалиться”.
- Нужно либо:
  - обучать на всех задачах одновременно с умным механизмом адаптации,
  - либо начинать с лёгких и поэтапно усложнять.
- Это требует новых алгоритмов — область активных исследований.
- Главное — тут нет принципиально нерешаемых проблем, только инженерные и методологические.

9 Как быть “на волне” ИИ

- Даже “простые” подходы работают, если масштабировать данные и модель.
- Когда-то глубокое обучение считали наивным, а теперь оно основа индустрии.
- Сейчас большие языковые модели тоже критикуют — но они дают сильный результат.
- Выбор:
  - Работать с простыми масштабируемыми решениями.
  - Или идти в сложные направления, где может быть следующий прорыв.

Заключение

ИИ находится в точке смены парадигмы. Чтобы оставаться в "тренде необходимо:

- изучать новые методы обучения,
- разрабатывать задачи с верифицируемыми результатами,
- понимать ограничения текущих моделей и строить новые.

Это не просто вызов — это возможность участвовать в формировании будущего ИИ.