Название статьи (авторы статьи): Chain-of-Thought Prompting Elicits Reasoning in Large Language Models (Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, Denny Zhou)

Автор обзора-рецензии: Артем Присяжнюк

Основная цель работы - показать, что при определенном дизайне подсказки (prompt) качество ответов модели возрастает. Более конкретно, если попросить модель сгенерировать логическую цепочку перед ответом, то качество такого ответа статистически значимо возрастает. Такое эффект наряду с некоторыми другими эффектами наблюдается только в работе больших языковых моделей. Однако авторы оставляют открытым вопрос интерпертации этого явления.

Работа была принята на конференции NIPS 2022, причем отзывы на эту статью строго положительные. В статью были добавлены эксперименты с открытыми большими языковыми моделями после комментария одного из ревьюеров, но принципиально от этого работа не поменялась. Основные авторы работы занимаются nlp и большими языковыми моделями, а также промпт-инженерией, resoning и reinforcement learning подходами к языковым моделям. Общей для основных авторов статьей является статья про модель PaLM и продолжение нашей статьи - Wei 2022, “Emergent Abilities of Large Language Models”.

Эта работа показывает ранее не замеченный эффект, поэтому базой для этой работы можно назвать в принципе все работы в этой области исследования. Основополагающая статья для области, на которую часто ссылаются - Brown 2020, “Language Models are Few-Shot Learners”. Прямым продолжением этой работы можно назвать сразу несколько статей: Wei 2022, “Emergent Abilities of Large Language Models”, Zhou 2022, “Least-to-most Prompting Enables Complex Reasoning in Large Language Models”, Kojima 2022, “Large Language Models are Zero-Shot Reasoners”, Wang 2022, “Self-consistency Improves Chain-of-thought Reasoning in Large Language Models”, Zhang 2022, “Automatic Chain-of-thought Prompting in Large Language Models”. Все они используют явление, найденное в статье, и делают некоторые надстройки над ним.

Сильные стороны:
Простой, мотивированный и широко применимый подход
Показано значимое улучшение качества на больших моделях
Подробное описание разнообразных экспериментов

Слабые стороны:
Не исследован случай маленьких моделей, никаких деталей
Нет даже предположений о возможных причинах, black-box подход
Тесты проведены на малой части доменов (в основном, арифметические датасеты)