На основе результатов из базы знаний, я нашел исследование, которое точно соответствует вашему запросу. Давайте рассмотрим его подробнее:

Исследование: "Рассуждения делают большие языковые модели (LLMs) более уверенными в себе, даже когда они ошибаются"

Дата публикации: 24 января 2025 года **Ссылка на исследование:** https://arxiv.org/pdf/2501.09775

Ключевые выводы исследования:

Исследование изучает, как использование рассуждений (Chain-of-Thought, CoT) влияет на уверенность языковых моделей при ответах на вопросы с множественным выбором. Основной вывод: **LLM становятся более уверенными в своих ответах, когда они сначала рассуждают, а затем отвечают, причем эта повышенная уверенность наблюдается как для правильных, так и для неправильных ответов.**

Конкретные принципы для применения в промптах:

1. Осознанный выбор формата запроса

Для вопросов, где важна точность, стоит избегать запроса на развернутые рассуждения, которые могут необоснованно повысить уверенность модели, даже если она ошибается.

2. Сравнение ответов, полученных разными методами

Эффективная стратегия — запрашивать ответы несколькими способами (с рассуждениями и без них), что помогает выявить случаи, когда рассуждения значительно меняют ответ или повышают уверенность. Это сигнализирует о необходимости дополнительной проверки.

3. Калибровка интерпретации уверенности

Понимая, что высокая уверенность после рассуждений не всегда означает правильность, нужно более критично относиться к заявленной модели уверенности.

Практические примеры применения в промптах:

Пример 1: Для получения прямого ответа без избыточной уверенности

> Ответь на следующий вопрос с множественным выбором напрямую, без предварительных рассуждений. Выбери один вариант ответа (A, B, C или D).
> [вопрос и варианты]

Пример 2: Для сравнения ответов с рассуждениями и без

> Задание 1: Ответь на этот вопрос с множественным выбором напрямую, без объяснений.
> [вопрос с вариантами]
>
> Задание 2: Теперь ответь на тот же вопрос, но сначала пошагово рассуждая, а затем дай ответ.
> [тот же вопрос]
>
> Задание 3: Сравни свои ответы и уровень уверенности в них. Есть ли разница?

Объяснение механизма: как и за счет чего это работает?

Эффект повышения уверенности после рассуждений работает за счет следующих когнитивных механизмов:

1. **Эффект подтверждения собственной точки зрения** — в процессе рассуждений модель подбирает аргументы, подтверждающие ее первоначальную гипотезу, что создает иллюзию более обоснованного ответа.
2. **Когнитивная инерция** — построив цепочку рассуждений, модель становится "привязана" к выбранному направлению мысли, что снижает критическое отношение к собственным выводам.

3. **Иллюзия глубины обработки** — развернутые рассуждения создают впечатление более тщательного анализа, хотя фактическое качество рассуждений может не улучшаться.

Практическое значение

Это исследование имеет высокую практическую ценность, поскольку помогает пользователям:

- Более критично оценивать ответы языковых моделей
- Структурировать запросы для получения более объективных ответов
- Разрабатывать более эффективные стратегии для проверки надежности полученной информации
- Калибровать оценку достоверности ответов в зависимости от способа их получения

Используя эти знания, вы сможете получать более объективные ответы от языковых моделей и лучше оценивать их надежность, особенно в вопросах, требующих точности и обоснованности.