Метод промежуточного резюмирования (InftyThink)

Что это такое

InftyThink — это метод, представленный в исследовании "InftyThink: Преодоление ограничений длины долгосрочного контекстного рассуждения в больших языковых моделях". Это подход, который позволяет трансформировать монолитные рассуждения в итеративный процесс с промежуточным суммированием, что помогает преодолеть ограничения контекстного окна языковых моделей.

Принципы для применения в промптах

1. **Разбиение сложных задач на этапы**
    - Разделяйте задачу на небольшие сегменты (рекомендуемый размер 500-1000 слов)
    - Каждый сегмент должен представлять собой логически завершенную часть рассуждения
2. **Регулярное резюмирование прогресса**
    - После каждого сегмента создавайте краткое резюме достигнутого прогресса
    - В резюме указывайте: что уже установлено и какие шаги ещё нужно выполнить
3. **Опора на предыдущие резюме**
    - В начале каждого нового сегмента явно опирайтесь на предыдущее резюме
    - Используйте резюме как отправную точку для продолжения рассуждений
4. **Управление контекстом**
    - Вместо отправки всей истории рассуждений отправляйте только последнее резюме и новую часть задачи
    - Это экономит токены и эффективно использует контекстное окно

Практический пример промпта

```
# Задача решения сложной математической проблемы с использованием InftyThink

## Инструкции:
1. Я предоставлю математическую задачу, требующую длинного рассуждения
2. Решай задачу поэтапно, разбивая рассуждение на сегменты по 500-1000 слов
3. В конце каждого сегмента:
   - Суммируй текущий прогресс в решении (что уже установлено)
   - Укажи, какие шаги еще необходимо выполнить
4. В следующем сегменте опирайся на это резюме, продолжая рассуждение
5. Повторяй процесс до полного решения задачи

## Задача: [Описание сложной математической задачи]

Начни решение, следуя методологии InftyThink.
```

Как это работает и почему это эффективно

1. **Преодоление ограничений контекстного окна** Метод создает характерную "пилообразную" схему использования памяти. Вместо одной длинной цепочки рассуждений (которая может выйти за пределы контекстного окна), модель создает серию коротких сегментов и резюмирует прогресс. Это позволяет обрабатывать задачи практически неограниченной сложности.

2. **Снижение вычислительной сложности** Промежуточное резюмирование значительно снижает вычислительные затраты по сравнению с традиционными подходами. Это решает проблему квадратичного роста вычислительных затрат с увеличением длины последовательности.

3. **Улучшение структурированности мышления** Этот подход помогает модели и пользователю лучше отслеживать прогресс решения, делает рассуждение более организованным и снижает вероятность "дрейфа рассуждений" (когда модель уходит от основной линии).

4. **Экономия токенов** Использование резюме вместо полной истории рассуждений существенно экономит токены, что делает взаимодействие более эффективным.

Хотя полная реализация InftyThink в исследовании требует дообучения моделей, основные концепции и практики итеративного рассуждения с резюмированием могут быть легко адаптированы в стандартных промптах для

любой языковой модели без необходимости в специальной настройке или доступе к API.