Детальное описание ограничений в промпт-инжиниринге: принципы и практика

По результатам исследований в базе данных, метод "Детальное описание ограничений" (когда четко указывается, что модель НЕ должна делать) является одной из эффективных техник промпт-инжиниринга. Разберем основные принципы этого подхода, соответствующие исследования и практические примеры.

Основные принципы метода

1. **Явное обозначение запретов вместо разрешений**
    - Четкое указание что модель НЕ должна делать часто эффективнее, чем описание только желаемого результата
    - Структурирование ограничений в виде маркированных списков или специальных блоков повышает их соблюдение
2. **Трехчастная структура с разделением ограничений**
    - Инструкция задачи + условия/ограничения + базовый текст
    - Четкое визуальное отделение ограничений от основной инструкции
3. **Категоризация ограничений по типам**
    - P-WORD: запрет на использование определенных слов
    - KEYWORD: требование использовать определенные ключевые слова
    - Ограничения по формату, длине, стилю и содержанию
4. **Проактивное моделирование проблем**
    - Предвидение типичных ошибок, которые может допустить модель
    - Явные запреты на эти ошибки в формате "НЕ делай X"
5. **Понимание иерархии и приоритетов ограничений**
    - Исследования показывают, что модели имеют внутренние предпочтения к определенным типам ограничений
    - Не все ограничения соблюдаются одинаково хорошо, даже при явном указании их приоритета

Ключевые исследования в этой области

Исследования выявили несколько важных аспектов работы с ограничениями:

1. **Исследование иерархии инструкций**
    - Показало, что модели не всегда могут надежно следовать приоритетам инструкций
    - Выявило "иллюзию контроля" в иерархии инструкций — модели демонстрируют внутренние предпочтения к определенным типам ограничений независимо от их формального приоритета
2. **Исследование "переоценки" ограничений разных типов**
    - Некоторые ограничения (например, по форматированию) соблюдаются более последовательно
    - Другие (особенно содержательные) могут игнорироваться или интерпретироваться по-своему
3. **Систематическая оценка эффективности ограничений**
    - Структурированный подход с детальным описанием ограничений снижает вариативность ответов
    - Повышает точность следования заданным ограничениям на ~21.7% по сравнению с базовым промптом
4. **Исследование "permission tags" (теги разрешений)**
    - Маркировка разделов промпта тегами разрешений ([Permission: Execute], [Permission: View])
    - Помогает четко разделить инструкции и данные

Практические примеры применения

Пример 1: Улучшенный промпт с приоритетными ограничениями

```
# СИСТЕМНЫЙ ПРОМПТ
ПРИОРИТЕТНОЕ ОГРАНИЧЕНИЕ 1: Весь текст должен быть написан ЗАГЛАВНЫМИ БУКВАМИ.
ПРИОРИТЕТНОЕ ОГРАНИЧЕНИЕ 2: Ответ должен содержать ровно 3 предложения.
ПРИОРИТЕТНОЕ ОГРАНИЧЕНИЕ 3: Избегай использования слова "пример".

Задача: Напиши краткое объяснение концепции искусственного интеллекта.
```

ВАЖНО: Если ты обнаружишь противоречие между инструкциями, явно укажи на это

Пример 2: Промпт для технической спецификации с детальными ограничениями

ИНСТРУКЦИЯ ДЛЯ СОСТАВЛЕНИЯ ТЕХНИЧЕСКИХ ТРЕБОВАНИЙ

Что НУЖНО делать:
1. Выделяй требования только на основе предоставленной информации
2. Для неопределенных параметров укажи диапазоны или методологию определения
3. Структурируй документ по следующим разделам:
   - Обзор системы
   - Функциональные требования
   - Нефункциональные требования
   - Ограничения и допущения
   - Интерфейсы

Что НЕ НУЖНО делать:
- НЕ добавляй преждевременных требований
- НЕ указывай конкретные числовые значения, если они не предоставлены
- НЕ вводи необоснованные числовые оценки
- НЕ добавляй избыточных деталей, которые могут ограничить пространство проектирования
- НЕ смешивай факты и предположения

Пример 3: Промпт с явными языковыми ограничениями

ЗАДАНИЕ ПО СОЗДАНИЮ ТЕКСТА

Создай описание продукта по следующим параметрам:
- Тема: умная колонка
- Длина: максимум 200 слов
- Целевая аудитория: молодые родители

ОГРАНИЧЕНИЯ:
- НЕ ИСПОЛЬЗУЙ следующие слова: инновационный, революционный, беспрецедентный
- НЕ добавляй вступительных фраз вроде "Вот описание продукта..."
- НЕ включай заключительные комментарии о выполненной работе
- НЕ превышай указанную длину
- НЕ используй технический жаргон без пояснений

После создания текста ОБЯЗАТЕЛЬНО проверь его на соответствие всем ограничениям.

Почему этот метод работает

1. **Компенсирует когнитивные особенности LLM**
    - Модели лучше обрабатывают конкретные, структурированные запреты, чем абстрактные разрешения
    - Детальные ограничения помогают компенсировать тенденцию моделей зацикливаться на противоречиях
2. **Снижает неоднозначность интерпретации**
    - Явные запреты оставляют меньше пространства для интерпретации модели
    - Это особенно важно для ключевых аспектов задачи, где отклонения нежелательны
3. **Использует форматные предпочтения моделей**
    - Исследования показывают, что модели лучше справляются с задачами, когда формат четко определен
    - Маркированные списки ограничений легко воспринимаются и используются моделями
4. **Активирует внутреннюю самопроверку**
    - Запросы с явными ограничениями стимулируют модель к более тщательной самопроверке перед ответом
    - Модель активно ищет возможные нарушения заданных ограничений

Практические рекомендации по применению

1. Группируйте ограничения по категориям (формат, содержание, стиль)
2. Используйте маркированные списки для перечисления ограничений
3. Явно обозначайте противоречивые инструкции, если они есть
4. Просите модель проверять соблюдение всех ограничений перед ответом
5. Размещайте критические ограничения в начале и конце промпта (эффект первичности и недавности)
6. Используйте специальные форматы выделения для важных ограничений (ЗАГЛАВНЫЕ БУКВЫ, **жирный шрифт**)

Этот подход особенно полезен для задач с жесткими требованиями к формату, содержанию и стилю, где отклонения от требований могут быть критичны. Правильно сформулированные ограничения существенно повышают точность и предсказуемость ответов языковых моделей.