Многомерная оценка ответов: оценка не только по точности, но и по следованию инструкциям

На основе исследований в базе данных, я подготовил анализ метода "Многомерной оценки ответов", который представляет собой подход к оценке качества ответов моделей ИИ по нескольким измерениям одновременно.

Основные принципы метода

1. **Комплексная оценка по нескольким критериям** - вместо единой обобщенной оценки применяется разделение на конкретные компоненты:
    - Точность (корректность фактической информации)
    - Полнота (охват всех аспектов вопроса)
    - Следование инструкциям (соответствие заданной структуре)
    - Ясность (понятность объяснения)
    - Применимость (практическая ценность ответа)
2. **Стратификация запросов по уровню определенности** - явное разделение запросов на:
    - "Объективные" (с однозначными ответами)
    - "Субъективные" (допускающие вариативность)
3. **Учет неопределенности в оценках** - вместо единственной оценки используется:
    - Диапазоны вероятных оценок (например, "7.2-7.8")
    - Указание аспектов с наибольшей неопределенностью
4. **Сравнительный подход к оценке** - оценка через сравнение ответов между собой, а не только в абсолютных значениях.
5. **Структурированная рубрика** - четкая система критериев с указанием баллов за каждый компонент.

Ключевые исследования, затрагивающие метод

1. **RevisEval** - исследование показало, что многокритериальная оценка повышает точность на 2-6% по сравнению с обобщенной оценкой.

2. **Трансферное побуждение** - подтверждает эффективность оценки ответов не только по точности, но и по следованию инструкциям.

3. **За пределами корреляции** - анализирует влияние человеческой неопределенности на измерение эффективности оценки и рекомендует использование нескольких критериев.

4. **Сравнительное рассуждение толпы** - демонстрирует, что структурированное сравнение по конкретным критериям делает оценку более объективной.

5. **Автоматизированная оценка заданий** - показывает, что структурированная рубрика с четкими критериями повышает точность оценки до 85-90%.

6. **Проблемы тестирования программного обеспечения на основе больших языковых моделей** - предлагает разделение сложных запросов на подзадачи с отдельными проверками по разным аспектам.

Практический пример промпта с многомерной оценкой:

```
# Задание по многомерной оценке качества ответа

Выступи в роли эксперта-оценщика для моей задачи. Оцени следующий ответ по нескольким критериям.

## Инструкции по оценке:
1. Оцени ответ по шкале от 1 до 10 по каждому из критериев:
   - Точность (насколько информация корректна)
   - Полнота (насколько охвачены все аспекты вопроса)
   - Следование инструкциям (насколько точно выполнены все требования)
   - Ясность (насколько понятно объяснение)
   - Практическая ценность (насколько применимы предложенные решения)

2. Для каждого критерия:
   - Предоставь не точную оценку, а диапазон вероятных оценок (например, "7.2-7.8")
   - Объясни обоснование оценки
   - Укажи конкретные сильные стороны и недостатки

3. Выполни попарное сравнение ответов, указав, насколько один ответ лучше другого по каждому критерию.
```

```
4. В завершение анализа укажи:
   - Аспекты с наибольшей неопределенностью в оценке
   - Общую оценку (среднее по всем критериям)
   - Ключевые рекомендации по улучшению ответа

## Вопрос:
[Вставить вопрос]

## Ответы для оценки:
Ответ A: [Вставить первый ответ]
Ответ B: [Вставить второй ответ]
```

Почему это работает:

1. **Повышение точности оценки** - многомерный подход позволяет получить более детализированную и нюансированную оценку, учитывающую различные аспекты качества ответа.

2. **Снижение предвзятости** - разделение оценки на отдельные критерии снижает влияние общих предвзятостей (например, к многословности или позиционную предвзятость).

3. **Учет человеческой вариативности** - аспект неопределенности в оценках более точно отражает реальные человеческие суждения, которые редко бывают абсолютно однозначными.

4. **Повышение надежности** - оценка по нескольким критериям дает более стабильные результаты даже при различных формулировках запроса.

5. **Прозрачность оценки** - пользователи получают ясное понимание, по каким именно параметрам ответ силен или слаб, вместо размытой общей оценки.

Этот метод особенно полезен для:

- Оценки образовательных материалов
- Анализа кандидатов в процессе найма
- Сравнения нескольких вариантов решения задачи
- Улучшения качества контента
- Разработки и тестирования самих языковых моделей

Используя многомерную оценку, вы получаете гораздо более глубокое понимание качества ответов, что позволяет принимать более обоснованные решения и эффективнее улучшать контент.