Экономический факультет МГУ имени М.В. Ломоносова

Дискуссия к докладу «Разработка синтетической фокус-группы для предварительного тестирования коммуникации на основе LLM»

Филипп Картаев

д.э.н., заведующий кафедрой
микро- и макроэкономического анализа
экономического факультета МГУ

@KARTAEVMACRO

О чем эта работа: кратко

Исследовательская проблема

- Коммуникация центрального банка является важным инструментом ДКП. Ошибки в формулировках могут приводить к провалу коммуникации
- Традиционные фокус-группы требуют много времени и ресурсов
- Могут ли LLM использоваться как инструмент раннего тестирования?

Что предлагают авторы?

- Синтетическую фокус-группу на основе: набора LLM-агентов с различными профилями; группового обсуждения текста; автоматического резюмирования; автоматической оценки коммуникационных рисков.

Основной результат:

- Разработана и протестирована мультиагентная система, способная выявлять проблемные формулировки и давать рекомендации по улучшению текста

Сильные стороны работы

1. **Тема невероятно актуальна в силу сочетания двух причин**
   - LLM и synthetic respondents — одно из наиболее быстро развивающихся направлений последних лет
   - Для Банка России (да и для любого центрального банка) задача тестирования коммуникации имеет очевидную прикладную ценность
2. **Богатый обзор литературы**
   - Особенно полезно для такой свежей исследовательской области
3. **Реалистичная постановка задачи**
   - Разработанная система рассматривается как вспомогательный инструмент для снижения издержек ранней диагностики
4. **Сильный акцент на валидации**
   - Детально проработанная формальная система тестирования качества и аккуратное обсуждение ограничений

Замечания и направления улучшения (1)

Главное ограничение работы — отсутствие внешней валидации на данных реальных фокус-групп

- Вся система детально проверяется на внутреннюю согласованность
- Однако нигде не проверяется насколько выводы синтетической фокус-группы совпадают с выводами реальных людей
- Авторы честно декларируют, что не ставят задачу доказательства поведенческой эквивалентности
- Однако между полной эквивалентностью и полным отсутствием внешней проверки существует широкий спектр промежуточных стадий
- Работа существенно выиграла бы, если бы хотя бы для нескольких текстов были проведены хотя бы некоторые проверки

Замечания и направления улучшения (2)

Не всегда ясны критерии прохождения/провала валидационных тестов

- Например в тесте T0 «Диагностика простой оценочной калибровки» указано, что основная метрика теста – точность совпадения (accuracy) между ожидаемой меткой и фактической оценкой агента
- При этом не указано, какой должна быть эта точность в идеале
- Далее в работе получено, что точность равна примерно 50% и сделан вывод о том, что это хороший результат
- Такой подход возможен, но остается неясным, что такое тогда плохой результат, и где лежит граница между хорошим и плохим результатом
- Лучше указывать это **до** проведения теста

Замечания и направления улучшения (1)

Работа выиграла бы в случае более детального обоснования комбинации профилей агентов

- Авторы подчеркивают, что агенты формируют «контрастный набор типов восприятия», а не репрезентативную выборку населения
- Отказ от критерия репрезентативности допустим, но в этом случае лучше предложить какой-то еще критерий
- Архитектура системы целиком определяется набором персон. От выбора персон зависят выявляемые проблемы и формируемые рекомендации
- Поэтому хотелось бы увидеть более формализованное обоснование: опору на реальные фокус-группы или на сегментацию аудиторий Банка России
- **Нужен анализ чувствительности результатов к составу агентов**

Экономический
факультет
МГУ
имени
М.В. Ломоносова

Спасибо за внимание!

Филипп Картаев

д.э.н., заведующий кафедрой
микро- и макроэкономического анализа
экономического факультета МГУ

@KARTAEVMACRO