Проблемы и методы валидации социальных симуляций на основе LLM-агентов

Огай Владислав Леонидович

Москва, 2026

Оглавление

Что такое LLM-агент?

- **LLM-агент** - автономный вычислительный субъект, в котором большая языковая модель используется как ядро рассуждения и генерации действий
- Обычно агент имеет **роль / профиль**, **память**, **правила взаимодействия со средой** и **контур принятия решений**
- В социальной симуляции агент должен имитировать не просто текст, а **поведение человека**: мнение, реакцию на аргументы, выбор контактов, изменение позиции
- Поэтому объект исследования - не отдельный ответ модели, а **динамика множества взаимодействующих агентов**

LLM-агент = языковая модель + состояние + правила действия в среде

Как агент работает в симуляции

1. **Профиль**: демография, роль, исходные установки
2. **Вход**: сообщение, новость, действие другого агента, состояние сети
3. **Память**: история диалога, прошлые решения, внешние факты
4. **Правило шага**: оценка ситуации $\rightarrow$ выбор действия $\rightarrow$ обновление состояния
5. **Выход**: реплика, голосование, смена мнения, разрыв или создание связи

Для валидации важно

Проверять нужно не только текст реплики, но и механизм обновления состояния, зависимость от контекста и коллективный результат

Базовые математические модели динамики мнений

Де-Гроот

$$\mathbf{x}(t+1) = W\mathbf{x}(t)$$

Усреднение мнений соседей. Базовый эталон **консенсуса**

Фридкин-Джонсон

$$\mathbf{x}(t+1) = \Lambda W\mathbf{x}(t) + (I-\Lambda)\mathbf{x}(0)$$

Учитывает **упрямство**: часть агентов держится за исходную позицию

Ограниченная уверенность

Влияние есть только при

$$|x_i - x_j| \leq \varepsilon$$

Порождает **кластеры**, **поляризацию**, **фрагментацию**.

Зачем они нужны

LLM-симуляцию следует сравнивать не только с собой, но и с простыми моделями, которые уже объясняют консенсус, устойчивое несогласие и кластеризацию мнений

Зачем вообще нужна валидация?

- LLM-симуляции быстро перешли от имитации отдельных респондентов к моделированию сообществ и социальных сценариев
- При этом главная проблема области - не генерация правдоподобного текста, а **доказательство пригодности модели для научного вывода**
- Чем сильнее утверждение о поведении общества, тем строже должны быть проверки на микро-, мезо- и макроуровне

Вопрос не "реалистична ли симуляция", а "для какого вывода она достаточно надежна?"

Что именно нужно валидировать

До запуска

- концептуальную модель
- входные данные
- правила среды и сети
- реализацию и метрики

На мезо- и макроуровне

- паттерны взаимодействий
- сетевую структуру
- динамику кластеров и каскадов
- распределения, поляризацию, нормы

На микроуровне

- ответы и решения агента
- реакцию на аргументы и контекст
- воспроизведение известных эффектов

Отдельно

Всегда нужно фиксировать **границы применимости**: для каких групп, платформ и сценариев вывод вообще допустим

Термины валидации

Термин	**Ключевой смысл**
Верификация	Корректно ли реализован код модели
Проверка концептуальной модели	Разумны ли допущения о сущностях, правилах и механизмах
Проверка данных	Пригодны ли входные данные и каково их происхождение
Операциональная валидация	Достаточно ли точны выходы модели для заявленной цели
Подтверждение пригодности	Можно ли использовать модель для конкретной задачи и в оговоренных условиях

Почему трудны даже классические социальные модели

- Для социальных систем редко известны точные "законы движения"
- Один и тот же макроэффект может возникать из разных микро-механизмов
- Реальные данные часто дороги, фрагментарны или этически чувствительны
- Критерий качества зависит от цели: объяснение, прогноз, сценарный анализ, генерация гипотез

Следствие

Модель нельзя объявить "валидной вообще". Корректная формулировка - **валидна ли она для данного вопроса и данного уровня вывода**.

Что дополнительно усложняют LLM

- **Черный ящик**: трудно отделить механизм от убедительного текста
- **Чувствительность к запуску**: влияют модель, промпт, seed, температура, память, длина контекста
- **Скрытые смещения**: ответы могут отражать обучающие данные и системные настройки, а не целевую популяцию.
- **Правдоподобие без измеримости**: "человекообразность" текста еще не означает валидность поведения.

LLM-агенты богаче по поведению, но часто слабже по доказательной базе

Типичные недостатки в работах

Недостаток	**Методологическое следствие**
Внешняя правдоподобность	Похожесть текста на человеческий нередко подменяет систематическую проверку модели
Ориентация только на итоги	Совпадение нескольких итоговых метрик не подтверждает правильность механизма
Смешение уровней вывода	Из сходства отдельных диалогов делают выводы о популяции, платформе или политике регулирования
Один демонстрационный прогон	Без независимых запусков нельзя оценить разброс и устойчивость результата
Слабая документация	Недостаточное описание архитектуры затрудняет репликацию и предметную критику

Измерительные искажения

Средние ответы
Синтетические данные нередко завышают склонность к положительным ответам по сравнению с реальными респондентами

Разброс
Во многих работах ответы LLM-агентов оказываются менее вариативными и хуже передают дисперсию человеческих данных

Социальная желательность
Смещение к социально одобряемым ответам наблюдается часто, но его сила зависит от инструмента и постановки вопроса

Даже при хорошем внешнем сходстве LLM-агенты могут искажать средние значения, дисперсии и структуру признаков

Ограничения психометрики

Проблема
Популярная практика - измерять "личность" модели человеческими опросниками и затем использовать этот профиль в симуляции

Что показывает эмпирика
Ответы модели систематически отклоняются от человеческих: утверждения с обратным смыслом могут одновременно получать согласие, а вариации по промптм не всегда воспроизводят устойчивую факторную структуру

Методологический вывод
Личностный профиль агента нельзя считать валидированным только потому, что модель прошла человеческий опросник. Сначала нужно показать, что сам инструмент измеряет у модели то же самое, что и у человека

Границы применимости

Что показывают удачные кейсы

Хорошее совпадение с данными обычно достигается не "универсальной человечностью" модели, а тщательным условливанием на реальные биографии, роли и контекст

Что из этого следует

Работа на одной платформе, культуре или теме не переносится автоматически на другие языки, возрастные группы и социальные нормы

Итог

Валидность возникает из сочетания **модели + данных + способа условливания + исследовательского вопроса**

Логика валидации

Главная идея

Валидация - это не один тест, а **набор согласованных свидетельств**: документация модели, эмпирические сравнения, проверка механизмов, анализ чувствительности, повторные прогоны и сравнение с базовыми моделями

Практическое правило

Чем сильнее заявляемое утверждение, тем более разнородные и независимые свидетельства должны его поддерживать

Что нужно фиксировать в описании модели

- цель модели и уровень вывода;
- сущности, переменные состояния, шаг времени, порядок обновления;
- источники данных и проектные решения;
- выбор LLM, настройки, память, правила среды;
- заранее объявленные метрики пригодности.

Практический ориентир

Для агентных моделей особенно полезен стандарт ОИД: **обзор - проектные идеи - детали**

Проверка микроуровня и механизмов

Микроуровень

- распределения ответов
- решения в сопоставимых задачах
- реакция на источник, аргумент, контекст
- эффект персонализации и памяти

Механизмы

- воспроизведение известных эффектов
- вмешательство в архитектуру модели
- абляции памяти, ролей, сети, правил
- проверка причинного объяснения, а не только итога

Проверка взаимодействий и макродинамики

Уровень	Что сравнивать
Текстовый	тематику, тональность, структуру аргументации
Пользовательский	частоту активности, устойчивость позиции, склонность отвечать
Сетевой	плотность, степени, взаимность, кластеры, мосты
Динамический	скорость сходимости, поляризацию, каскады, устойчивость норм

Для социальной симуляции недостаточно сходства отдельных реплик - нужна проверка паттернов взаимодействия

Устойчивость результатов

- Один запуск для LLM-симуляции недостаточен
- Нужны независимые прогоны с разными seed и, по возможности, вариацией модели, температуры, памяти и длины контекста
- Важно считать интервальные оценки и явно показывать зоны нестабильности

Сильный результат

Не просто высокий средний показатель, а вывод, который сохраняется при разумных изменениях условий

С чем сравнивать сложную LLM-симуляцию

Базовая схема	**Что она проверяет**
Де-Гроот / Фридкин-Джонсон	Дает ли LLM что-то сверх усреднения и устойчивости исходных мнений
Модель ограниченной уверенности	Объясняет ли LLM кластеры лучше простого порога взаимодействия
Правила без LLM	Нужна ли вообще генеративная модель для данного эффекта
LLM только для текста	Добавляет ли LLM механизм поведения или только стиль
Модель без памяти / персон	Действительно ли критичны память и персонализация

План валидации: три уровня

1. Структура
цель, сущности, данные, код, метрики

2. Поведение
микроуровень, механизмы, взаимодействия

3. Надежность
устойчивость, базовые модели, границы вывода

1. Зафиксировать цель и допустимый уровень выводов
2. Описать архитектуру, данные и правила среды
3. Провести верификацию реализации
4. Проверить микроуровень, механизмы и макродинамику
5. Оценить устойчивость и явно записать границы применимости

Динамика мнений: что именно проверять

Уровень	Фокус проверки
Микро	Реакция на аргумент, источник, социальный контекст
Мезо	Формирование кластеров, эхо-камер, мостов, конфликтов
Макро	Консенсус, поляризация, фрагментация, колебания

Типичная ловушка

Правильная макрокартина еще не означает правильный механизм убеждения

Сила выводов зависит от силы проверки

Что есть	**Что можно утверждать**
Только убедительные диалоги	Демонстрация возможностей
Корректная документация и код	Техническая воспроизводимость
Совпадение на микроуровне	Узкая пригодность для имитации ответов / решений
Механизмы + макродинамика + устойчивость	Ограниченный сценарный анализ и генерация гипотез

Открытые проблемы

Проблема	**Суть**
Интерпретируемость	Трудно доказать, какой именно механизм породил наблюдаемый эффект
Измеримость личности и установок	Человеческие опросники не всегда валидны для моделей
Культурная перенастройка	Результаты могут резко зависеть от языка, контекста и скрытых норм обучающих данных
Дисперсия и редкие случаи	Синтетические выборки склонны сглаживать крайние позиции и уменьшать разброс
Воспроизводимость	Различия между моделями, версиями и формулировками промптов мешают накоплению знания

Выводы

1. Для LLM-симуляций центральна не реалистичность текста, а **пригодность модели для конкретной научной цели**
2. Надежная валидация должна объединять **структурную проверку, эмпирические сопоставления, проверку механизмов, анализ устойчивости и сравнение с базовыми моделями**
3. Хороший результат - это не "модель верна вообще", а **строго ограниченный и воспроизводимый вывод о том, где ей можно доверять**

Спасибо за внимание!

Список литературы

[1] Sargent R. G. Verification and Validation of Simulation Models // *Proceedings of the 2010 Winter Simulation Conference*. 2010.

[2] Louie M. A., Carley K. M. Balancing the Criticisms: Validating Multi-Agent Models of Social Systems // *Simulation Modelling Practice and Theory*. 2008. Vol. 16, No. 2. P. 242--256.

[3] Grimm V. et al. The ODD Protocol for Describing Agent-Based and Other Simulation Models: A Second Update to Improve Clarity, Replication, and Structural Realism // *Journal of Artificial Societies and Social Simulation*. 2020. Vol. 23, No. 2.

[4] Collins A. et al. Methods That Support the Validation of Agent-Based Models // *Journal of Artificial Societies and Social Simulation*. 2024. Vol. 27, No. 1.

[5] Mou X. et al. From Individual to Society: A Survey on Social Simulation Driven by Large Language Model-based Agents // *arXiv preprint* arXiv:2412.03563. 2024.

[6] DeGroot M. H. Reaching a Consensus // *Journal of the American Statistical Association*. 1974. Vol. 69, No. 345. P. 118--121.

[7] Friedkin N. E., Johnsen E. C. Social Influence and Opinions // *The Journal of Mathematical Sociology*. 1990. Vol. 15, No. 3--4. P. 193--206.

[8] Hegselmann R., Krause U. Opinion Dynamics and Bounded Confidence Models, Analysis and Simulation // *Journal of Artificial Societies and Social Simulation*. 2002. Vol. 5, No. 3.