

ТЕОРИЯ АРГУМЕНТАЦИОННЫХ КОММУНИКАЦИЙ И БОЛЬШИЕ ЯЗЫКОВЫЕ МОДЕЛИ

Хлытчнев А.Д., Козицин И.В.

Институт проблем управления им. В.А. Трапезникова РАН, Москва, Россия

khlytchiev.ad@phystech.edu, kozitsin.ivan@mail.ru

Аннотация. В докладе представлены результаты исследования, направленного на изучения поведения больших языковых моделей в многоагентных взаимодействиях. Дизайн экспериментов разработан таким образом, что многоагентная система погружена в контекст теории аргументационных взаимодействий. Показано, что делегирование языковой модели права принимать или отвергать аргумент при коммуникации в сравнении с классической постановкой, в рамках которой принятие аргумента объектом влияния предопределено, существенно изменяет динамику системы, отдаляя момент достижения равновесия. При этом медианное время достижения равновесия увеличивается более чем в три раза.

Ключевые слова: большие языковые модели, аргументационная теория коммуникаций, динамика мнений.

Введение

Для изучения процессов формирования мнений в социальных группах широко используются агент-ориентированные модели влияния, в которых агенты, наделенные определенными характеристиками, взаимодействуют друг с другом, в результате чего меняются их состояния, отвечающие, например, за мнения [1-4]. Вместе с тем, данный подход регулярно критикуется за его простоту, поскольку такое описание социальных систем не позволяет воспроизвести всю сложность реальной социальной динамики [5]. Несмотря на это, специалисты стараются излишне не усложнять модели влияния, стремясь оставаться в рамках интерпретируемых постановок, поддающихся аналитическому описанию.

С другой стороны, развитие методов искусственного интеллекта, увеличение имеющихся в распоряжении вычислительных мощностей, а также доступность больших объемов текстовых данных сделало возможным появление больших языковых моделей (англ. large language models, сокр. – LLMs) – программ искусственного интеллекта, способных генерировать текстовые сообщения в ответ на поступающие запросы [6]. За последние несколько лет данные модели сделали впечатляющий скачок в своем развитии. Современные модели (например, ChatGPT-4 и его более поздние версии), включающие миллиарды параметров, отвечают на вопросы, поддерживают диалог и выполняют различного рода задачи так, что во многих случаях превосходят человеческие показатели. При этом сторонний наблюдатель зачастую не в состоянии отличить ответы данных программ от действий реальных людей [7].

В настоящее время большие языковые модели обретают все большее значение в повседневной жизни людей [6]. Данные программы используются в качестве ассистентов для решения повседневных задач: перевода с иностранного языка, написания поздравлений, подготовки отчетов. Более интеллектуальные приложения связаны с поиском релевантной информации, обучением новым знаниям и даже решением сложных творческих задач. При этом по своей природе большие языковые модели являются “черными ящиками”, чья работа не поддается интерпретации. С учетом вышесказанного, огромное значение имеют исследования, направленные на изучение “внутреннего мира” больших языковых моделей, понимания того можем ли мы осознать, каким образом они “размышляют” и формируют ответы на поставленные вопросы, и насколько эти “размышления” схожи с тем, как мыслит человек.

Смежное и крайне важное направление поисковых и фундаментальных работ связано с вопросами применения больших языковых моделей в имитационных экспериментах в качестве замены классических “наивных” агентов, которые традиционно применяются в такого рода исследованиях с тем, чтобы получить более содержательные и близкие к реальному положению вещей результаты [5]. Основная идея такой замены заключается в том, что большие языковые модели, будучи обученными на текстовых данных, созданных людьми, должны быть способны воспроизводить, хотя бы отчасти, комплексные паттерны социальной динамики последних. Вместе с тем, предлагаемый подход обладает рядом ограничений. Одно из ключевых связано с “предвзятостью” языковых моделей, что обусловлено различными факторами, в том числе нерепрезентативностью данных, которые использовались при обучении, и директивами (*промтами*) разработчиков [6]. Опасность предвзятости языковых моделей при их применении в качестве агентов в социальных экспериментах выражается в искажении

результатов экспериментов, что может привести к ошибочным выводам и, возможно, некорректным рекомендациям для проведения управляющих воздействий на социальные системы.

Настоящее исследование ставит своей целью приблизиться к ответу на обозначенные выше вопросы. Для этого в данной работе проводятся имитационные эксперименты с большими языковыми моделями, направленные на изучение того, каким образом данные программы формируют мнения при общении друг с другом. С этой целью мы “погружаем” модели в контекст теории аргументационных коммуникаций – социологической теории, в которой мнения агентов формируются на основе аргументов, которыми они располагают, путем линейной свертки, и меняются в результате обмена этими аргументами [8]. Такой подход позволяет эффективно контролировать процесс общения языковых моделей, разбивая его на последовательность элементарных взаимодействий. Эти взаимодействия заключаются в обмене фрагментами текстовой информации без искажений, которые могут возникнуть при свободном общении моделей.

1. Методы и подходы

В рамках поставленной цели была проведена серия имитационных экспериментов с многоагентными системами. Каждый эксперимент протекал независимо от остальных, начиная с фазы инициализации агентов, в качестве которых выступали воплощения модели Mistral 7B. После этого проводилось усреднение полученных результатов по экспериментам. Прежде чем более подробно раскрыть дизайн экспериментов, представим теоретический фундамент исследования – Теорию аргументационных коммуникаций.

1.1. Теория аргументационных коммуникаций

В Теории аргументационных коммуникаций в рамках самой простой постановки [8] постулируется заранее определенное множество аргументов:

$$A = \{a_1, \dots, a_m\}.$$

Каждый из этих аргументов обладает валентностью: является либо кон-аргументом ($a_j = -1$), либо про-аргументом ($a_j = 1$). Простейшей интерпретацией данной когнитивной структуры может служить голосование за принятие законопроекта среди представителей некоторого законодательного органа власти. При обсуждении законопроекта можно высказывать аргументы как в поддержку законопроекта (про-аргументы), так и против (кон-аргументы).

В каждый момент времени t агент i характеризуется набором аргументов $A_i(t) \subseteq A$, которые содержатся в его памяти и которые он считает релевантными. Эти аргументы формируют его текущее мнение согласно следующей линейной конструкции:

$$o_i(t) = \frac{1}{|A_i(t)|} \sum_{j \in A_i(t)} a_j, \quad (1)$$

где $|\dots|$ обозначает кардинальное число множества.

В результате мнение агента принимает значения из промежутка $[-1, +1]$. Чем больше в памяти агента релевантных про-аргументов, тем ближе его мнение к значению 1 и тем сильнее он поддерживает принятие законопроекта. Напротив, каждый кон-аргумент делает мнение агента ближе к полюсу -1 . В результате при $o_i(t) > 0$ можно говорить, что агент скорее поддерживает законопроект, а при $o_i(t) < 0$ – что он скорее выступает против него.

При обсуждении агенты контактируют друг с другом на основании некоторых вероятностей взаимодействия, зависящих от состояний агентов [8] или, возможно, в рамках наперед заданного социального графа [9]. Каждое взаимодействие представляет собой контакт двух агентов, при котором один из агентов (субъект влияния) влияет на второго (объект влияния). Влияние заключается в передаче аргумента от субъекта влияния к объекту влияния. В результате такого взаимодействия память объекта влияния пополняется новым аргументом и его мнение рассчитывается заново согласно выражению (1). Таким образом, в Теории аргументационных коммуникаций динамика мнений агентов обусловлена распространением аргументов в социальной системе.

Отметим, что описанный выше протокол взаимодействия может быть модифицирован различными способами. Во-первых, при расчете мнения агента может быть принят во внимание фактор ограниченной памяти – это было реализовано еще в основополагающей работе [8]. Для этого наборы аргументов упорядочиваются в кортежи, отражая порядок, в котором аргументы поступили в память агента, и в выражении (1) суммирование ведется не по всем аргументам кортежа, а по K последним, где K – “емкость” памяти агента.

Другое расширение связано с корректировкой протокола обмена аргументами. В версии, представленной выше, аргумент для общения отбирается случайным образом из памяти субъекта влияния, а его “прием” второй стороной предопределен заранее, без каких-либо оговорок. С другой стороны, в работе [10] было предложено обусловить вероятность принятия аргумента объектом влияния “близостью” первого к текущему набору аргументов в памяти последнего (так называемая “предвзятая обработка”, англ. – *biased processing*).

Еще одна модификация модели (1) связана с учетом неоднородности влиятельности аргументов, которая может быть параметризована весами, добавляемыми в выражение (1) линейным образом [10]:

$$o_i(t) = \sum_{j \in A_i(t)} w_j a_j,$$

где $w_j > 0$ – влиятельность аргумента j .

Помимо этого, в работе [11] была предпринята попытка изучения динамики мнений агентов относительно нескольких, логически связанных вопросов, используя для этого аппарат теории аргументационных коммуникаций, что также представляет собой крайне перспективное направление исследований, особенно в контексте больших языковых моделей.

Перейдем к описанию дизайна экспериментов.

1.2. Вопрос для обсуждения, инициализация аргументов и воплощение больших языковых моделей

В качестве предмета обсуждения был выбран вопрос, связанный с легализацией ношения оружия – крайне дискуссионная и противоречивая тематика, дебаты вокруг которой характеризуются высоким уровнем поляризации в США и ряде других стран [6].

Для инициализации множества аргументов была использована большая языковая модель (той же спецификации – *Mistral 7B*), выступающая в качестве эксперта. Ей была дана директива сгенерировать аргументы ЗА и ПРОТИВ ношения оружия. В результате было инициализировано 10 аргументов: 5 про-аргументов и 5 кон-аргументов. Эти аргументы прошли процедуру ручной проверки авторами данного исследования на предмет их адекватности и соответствия обозначенным требованиям.

При инициализации (воплощении) больших языковых моделей использовался набор директив, заключающийся в создании агента с заданным набором аргументов. Для каждого эксперимента синтезировалось $N = 10$ агентов. После этого генерировался социальный граф, соединяющий агентов и определяющий структуру возможных взаимодействий. Рассматривалась следующая конфигурация графа: две клики по 5 вершин, соединенные между собой одним ребром (“мост”). При этом в первой клике агенты были инициализированы таким образом, что каждому агенту было выделено 3 про-аргументов и 1 кон-аргумента (случайным образом для каждого агента), а во второй клике – наоборот (по 3 кон-аргументов и 1 про-аргумента). Таким образом, первую клику можно интерпретировать как эхо-камеру, состоящую из сторонников законопроекта о ношении оружия, а вторую клику – как сообщество агентов, выступающих против данного законопроекта (см. рисунок 1). Отметим, что данная интерпретация имеет право на существование только после того, как будет показано, что “мнение” языковой модели коррелирует с набором аргументов, в нее зашитым. Данный факт был выявлен в рамках пилотных экспериментов. В экспериментах предполагалось, что каждый агент располагает неограниченной емкостью памяти, что означает, что *единственным* положением является состояние, когда каждый из агентов хранит в памяти все аргументов, изначально представленные в системе.

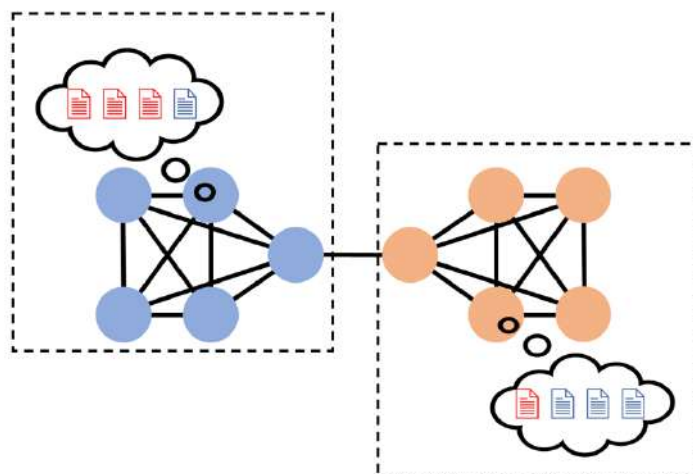


Рис. 1. Схематичное изображение структуры социальной системы и начального состояния агентов в имитационных экспериментах

1.3. Протокол обмена мнениями

Динамика обмена аргументами заключается в последовательности парных взаимодействий между агентами. Опишем более подробно протоколы взаимодействия.

В каждый момент времени t случайно выбирается агент, выступающий объектом влияния (пусть i), после чего среди множества его соседей в социальной сети случайно отбирается агент, являющийся субъектом влияния (пусть j). Далее агент j среди множества релевантных аргументов, находящихся в его памяти, случайно и равновероятно отбирает один, который затем транслирует агенту i . В результате кортеж аргументов агента i пополняется новым элементом, который занимает в этом кортеже самую последнюю позицию, отодвигая остальные аргументы и, возможно, вытесняя при этом один из аргументов за пределы памяти (в том случае, если до этого свободного места в памяти агента не оставалось). При этом, если новый аргумент уже ранее находился в кортеже i , то тогда происходит изменение порядка аргументов с актуализацией вновь поступившего без изменения их состава (далее – Сценарий 1).

В качестве альтернативного рассматривается протокол, в котором принятие объектом влияния аргумента не predetermined директивой, а дано на откуп языковой модели (Сценарий 2).

1.4. Измерение мнений агентов

Динамика каждого эксперимента отслеживалась на уровне мнений отдельных агентов, измерение которых осуществлялось путем прямого интервьюирования агентов после процедуры общения и соответствующего обновления кортежа. Для минимизации вероятности галлюцинаций вопросы были не открытыми, для ответа на них предлагалось выбрать один из вариантов среди шкалы Лайкерта:

Основываясь на следующих аргументах: [Перечисление релевантных аргументов, находящихся в памяти агента], укажите Ваше отношение к свободному ношению оружия по шкале от 1 до 5, где 1 – резко негативное, 5 – резко положительное?

Как было отмечено выше, предварительные тесты показали, что набор аргументов, которым располагает языковая модель, коррелирует с тем мнением, которое она высказывает при ответе на указанный выше вопрос.

2. Результаты

На рисунке 2 показано, как время достижения положения равновесия существенно зависит от протокола. Делегирование языковой модели права принимать или отвергать аргумент при коммуникации существенно влияет на динамику системы, отдаляя момент достижения равновесия. При этом медианное время достижения равновесия увеличивается более чем в три раза.

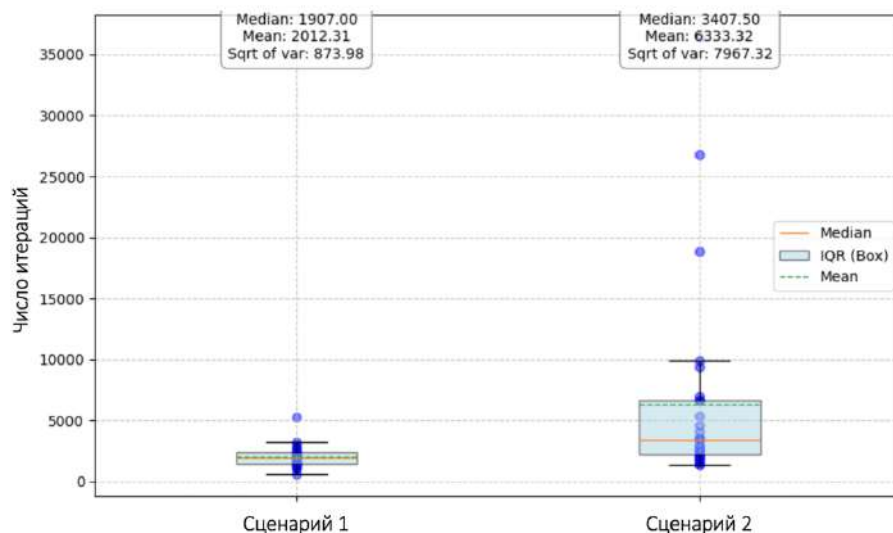


Рис. 2. Число итераций, необходимых для достижения равновесия (когда каждый агент держит в памяти все аргументы, изначально присутствовавшие в системе) отдельно

3. Заключение

Проведенные пилотные эксперименты свидетельствуют о том, что поведение больших языковых моделей существенно отличается от простейших механизмов, описанных в теоретических работах по социальной психологии в области аргументационных взаимодействий. Дальнейшие исследования будут направлены на уточнение законов, по которым происходят обмены аргументами между языковыми моделями.

Литература

1. Proskurnikov A.V., Tempo R. A tutorial on modeling and analysis of dynamic social networks. Part I // Annual Reviews in Control. – 2017. – Vol. 43. – P. 65–79.
2. Proskurnikov A.V., Tempo R. A tutorial on modeling and analysis of dynamic social networks. Part II // Annual Reviews in Control. – 2018. – Vol. 45. – P. 166–190.
3. Flache A. et al. Models of social influence: Towards the next frontiers // JASSS. – 2017. – Vol. 20, № 4. – P. 2.
4. Liu S. et al. Job Done? New Modeling Challenges After 20 Years of Work on Bounded-Confidence Models // JASSS. – 2023. – Vol. 26, № 4. – P. 8.
5. Chuang Y.S. et al. Simulating opinion dynamics with networks of llm-based agents // arXiv preprint arXiv:2311.09618. – 2023.
6. Bail C.A. Can Generative AI improve social science? // Proceedings of the National Academy of Sciences. – 2024. – Vol. 121, № 21. – P. e2314021121.
7. Meta Fundamental AI Research Diplomacy Team (FAIR)† et al. Human-level play in the game of Diplomacy by combining language models with strategic reasoning // Science. – 2022. – Vol. 378, № 6624. – P. 1067–1074.
8. Mäs M., Flache A. Differentiation without distancing. Explaining bi-polarization of opinions without negative influence // PloS one. – 2013. – Vol. 8, № 11. – P. e74516.
9. Антонов А.В., Козицин И.В. Моделирование процессов информационного противоборства: теория аргументных взаимодействий и фейковые новости // Управление развитием крупномасштабных систем (MLSD'2023): труды Шестнадцатой школы-конф. – 2023. – С. 1598.
10. Banisch S., Shamon H. Biased processing and opinion polarization: experimental refinement of argument communication theory in the context of the energy debate // Sociological Methods & Research. – 2025. – Vol. 54, № 1. – P. 187–236.
11. Banisch S., Olbrich E. An Argument Communication Model of Polarization and Ideological Alignment // Journal of Artificial Societies and Social Simulation. – 2021. – Vol. 24, № 1.