Правительство Российской Федерации
Федеральное государственное автономное образовательное учреждение высшего образования

Национальный исследовательский университет
«Высшая школа экономики»

Факультет гуманитарных наук
Образовательная программа
«Фундаментальная и компьютерная лингвистика»

КУРСОВАЯ РАБОТА

На тему «Интерпретация языковых моделей в контексте психолингвистического исследования»
Тема на английском: «Interpretation of language models in the context of psycholinguistic research»

Студентка
группы БКЛ221
Новоселова Анна Дмитриевна

Научный руководитель
Сериков Олег Алексеевич

Москва, 2025 г.

Введение

С ростом популярности больших языковых моделей (LLM) значительно увеличилось количество исследований, посвящённых их способности интерпретировать человеческий язык. Эти работы демонстрируют неоднозначную картину: в одних аспектах модели приближаются к человеческому уровню обработки, в других — проявляют существенные ограничения.

Так, LLM успешно справляются с обработкой сложных синтаксических структур. Например, в исследовании (Lampinen 2022) было показано, что модели способны на человеческом уровне интерпретировать рекурсивно вложенные грамматические конструкции. В другой работе (Chan et al. 2022) отмечено, что при решении логических задач модели демонстрируют паттерны рассуждений, сходные с человеческими, включая распространённые ошибки. Это указывает не только на примечательные навыки языковых моделей справляться с подобными задачи, но и на потенциальное сходство их стратегии мышления с человеческими когнитивными возможностями.

Однако другие исследования подчёркивают ограниченность логических способностей моделей. Так, (Ding et al. 2025) отмечают, что хотя модели демонстрируют впечатляющее эвристическое мышление, их поведение в рамках строгого логического вывода — дедуктивного, индуктивного, абдуктивного и аналогического — остаётся непоследовательным и слабо интерпретируемым. Это связано с ограничениями в обобщающей способности и устойчивости рассуждений.

Важно отметить, что языковые модели могут показывать разный уровень понимания языка и его приближённость к человеческому уровню. В работе (Cai et al. 2023) была проведена серия из 12 лингвистических экспериментов, показавшая, что ChatGPT демонстрирует человекоподобное поведение в таких задачах, как семантический прайминг, чувствительность к дискурсивному контексту, интерпретация импликатур и реакция на семантические иллюзии. Напротив, модель Vicuna показала менее стабильные результаты, особенно в задачах на восприятие абсурдных или иллюзорных высказываний. Это подчёркивает различия в интерпретации естественного языка между моделями, несмотря на общее сходство в базовых механизмах.

Ещё одно исследование (Holliday, Mandelkern 2024) было посвящено проверке способности языковых моделей производить логические выводы на основе условных конструкций и модальных операторов. В нём было протестировано более десятка LLM

на способность различать валидные и невалидные логические заключения. Результаты показали, что большинство моделей, за исключением GPT-4, регулярно ошибаются в интерпретации условных конструкций. Даже GPT-4 проявляет логически противоречивые суждения при работе с эпистемическими модальностями. Это подчёркивает ограниченность моделей в усвоении некоторых аспектов человеческого логического мышления.

Наконец, (Dentella et al. 2023) показали, что модели, несмотря на заявленную способность имитировать человеческие языковые способности, демонстрируют высокую нестабильность и низкую точность при оценке грамматичности предложений. Используя серию задач, включающую восемь лингвистические явления, (например, аномальные порядки прилагательных и наречий, отрицательные поляризованные частицы, анафоры), авторы выявили слабую согласованность оценок, особенно в случае аграмматичных предложений, а также склонность к положительным ответам при оценке языковых конструкций вне зависимости от их грамматичности. Эти результаты резко контрастируют с поведением носителей языка и ставят под сомнение возможность рассматривать современные языковые модели как надёжные когнитивные аналоги человеческой языковой системы на текущем этапе их развития.
Таким образом, большие языковые модели демонстрируют неоднозначный уровень понимания языка, однако это не уменьшает интерес к дальнейшему изучению этой области. В частности, внимание привлекает вопрос восприятия семантических и прагматических механизмов, лежащих на границе между структурной организацией языка и механизмами рассуждения, основанными на смысле и контексте.

Одним из способов проверить интерпретационные возможности языковых моделей является воспроизведение психолингвистических экспериментов. В качестве основы для такого анализа может служить работа (Liu, Schwab 2022a), в которой исследовалось, как лицензируется отрицательно поляризованная частица *all that* в различных типах условных предложений в английском языке в зависимости от их семантических и прагматических свойств. Мы намерены повторить эксперименты, описанные в данной статье, с несколькими LLM и изучить, в каких случаях результаты моделей совпадают с человеческими паттернами.

Целью данной курсовой работы является проверка способности больших языковых моделей воспроизводить результаты экспериментов выбранного психолингвистического исследования и анализ того, что полученные результаты могут

говорить о способности моделей понимать язык с точки зрения семантики и прагматики.

Основные задачи — воспроизведение дизайна экспериментов из (Liu, Schwab 2022a) с использованием больших языковых моделей моделей, проведение статистического и сравнительного анализа на основе собранных данных и качественная интерпретация выявленных результатов.

Основная часть

1. Обзор исследования (Liu, Schwab 2022a)

Исследование, проведенное (Liu, Schwab 2022a), посвящено изучению восприятия так называемых attenuating negative polarity items — ослабляющих отрицательно поляризованных частиц (NPIs). Это подтип отрицательно поляризованных частиц, употребление которых приводит к ослаблению ассерции. Например, в предложении *The negotiations haven't gone all that well* частица *all that* уменьшает его информативность об уровне успешности проведённых переговоров.

Основная цель работы заключалась в проверке теоретического предположения о том, что лицензирование ослабляющих NPIs чувствительно к прагматическим различиям между различными типами условных конструкций. В частности, восприятие NPIs проверялось в таких типах условных предложений, как гипотетические индикативные (hypothetical indicative) (1), контрфактические (counterfactual) (2) и условные с посылкой (premise conditionals) (3):

(1) *If the students have been all that attentive in class, they will pass the exam.*

(2) *If the students had been all that attentive in class, they would have passed the exam.*

(3) A: *The students have been very attentive in class.*

B: *If the students have been all that attentive in class, they will pass the exam.*

Среди прагматических механизмов, свойственных условным конструкциям и потенциально влияющих на восприятие NPIs, выделяются выполнение условия (Conditional Perfection, CP) и предположение о ложности антецедента (Antecedent Falsity). Под условным совершенствованием понимается явление, при котором кондициональное предложение принимает статус бикондиционального. В некоторых контекстах высказывание *if Q, then P* может интерпретироваться как *Q if and only if P*:

(4a) *If you mow the lawn, I'll give you $5.*

(4b) *If and only if you mow the lawn, I'll give you $5.*

Между антецедентом и консеквентом в предложении (4a) устанавливается такая связь, вследствие которой выполнение действия в первой части высказывания становится не просто достаточным, а необходимым для свершения действия во второй, отчего и возникает интерпретация, как в (4b).

Вторым ранее упомянутым прагматических механизмом является предположение о ложности антецедента, который проявляет себя в контрфактических условных конструкциях. Рассмотрим такие примеры:

(5a) *If the authors are linguists, they have written a linguistics paper.*

(5b) *If the authors had been linguists, they would have written a linguistics paper.*

Различие между этими предложениями с точки зрения истинности антецедента выражается в том, что в (5b), в отличие от (5a), считывается импликатура о ложности того, что авторы являются лингвистами (Collins, Skovgaard-Olsen 2021).

Особенностью ослабляющих NPIs считается то, что они по-разному взаимодействуют с этими прагматическими механизмами. В индикативных кондиционалах CP создает условия, при которых ослабляющие NPIs оказываются менее "приемлемыми", то есть менее естественными для восприятия, тогда как в контрфактических предположение о ложности антецедента может нивелировать негативное влияние CP, сохраняя приемлемость NPIs. Соответственно, задачей оригинального исследования стало предоставление эмпирических доказательств для данных теоретических утверждений. Вместе с этим в исходной статье формулируются два вопроса для изучения:

- Почему ослабляющие NPIs хуже лицензируются в индикативных условных, чем в контрфактических?
- Как прагматические свойства условных предложений, такие как выполнение условия и предположение о ложности антецедента влияют на приемлемость NPIs?

Для проверки поставленных вопросов авторы исследования провели три эксперимента, каждый из которых был направлен на решение конкретных исследовательских задач.

Первый эксперимент ставил своей основной целью сравнить воспринимаемую естественность употребления ослабляющей NPI *all that* в индикативных и контрфактических условных предложениях. Согласно выдвинутой гипотезе, контрфактические конструкции должны были обеспечивать более благоприятные условия для лицензирования NPI благодаря наличию антиверидикативного вывода о ложности антецедента. Результаты полностью подтвердили это предположение: индикативные условные с NPI оценивались носителями языка как значительно менее естественные по сравнению с контрфактическими. При этом в условиях без NPI различия между типами условных не наблюдалось, что свидетельствует о специфическом характере взаимодействия именно ослабляющих NPIs с прагматикой условных конструкций. Помимо основного анализа, авторы провели дополнительное сравнение оценок естественности предложений с квантором всеобщности *all*, определяющего подлежащее, с и без ослабляющего NPI. Этот анализ не был напрямую связан с проверкой заявленных гипотез, однако был включён для сопоставления с аналогичными данными по немецкому языку (Liu, Schwab 2022b).

Второй эксперимент был посвящен анализу условных конструкций с посылкой, которые, согласно теоретическим рассуждениям, должны быть устойчивы к эффекту CP. Исследователи предположили, что в таких контекстах ослабляющие NPIs будут демонстрировать лучшие показатели приемлемости. Полученные данные подтвердили и эту гипотезу: разница в оценках естественности между индикативными условными предложениями и конструкциями с посылками оказалась статистически значимой.

Третий эксперимент был направлен на непосредственное изучение феномена CP в разных типах условных конструкций. Авторы проверяли, насколько часто носители языка рассматривают условные предложения индикативного и контрфактического типов как бикондициональные, а также исследовали влияние обстоятельства степени *very*[1] на вероятность такого вывода. Результаты показали, что CP действительно возникает в обоих типах конструкций, однако в контрфактических его частота несколько ниже. Кроме того, было обнаружено, что наличие обстоятельства степени не оказывает влияния на эффект CP.

[1] Авторы предположили, что *very* может повышать вероятность бикондициональных прочтений: высказывание *if very P, Q* по сравнению с *if P, Q* одновременно “слабее” (поскольку *very p* влечет *p*, но не наоборот) и более затратно для продуцирования (поскольку *very p* содержит дополнительный модификатор). Услышав *if very p, q*, адресат может, таким образом, сделать вывод, что вариант без *very* не выполняется, т. е. для следствия необходимо, чтобы выполнялось *very p*, а не просто *p*. Тем не менее, авторы статьи не обосновывают чётко, почему они решили проверять влияние *very*, а не *all that* в третьем эксперименте.

Полученные результаты (Liu, Schwab 2022a) создают важную теоретическую основу для экспериментальной проверки поведения ослабляющих NPIs в языковых моделях. Если человеческое восприятие этих частиц чувствительно к прагматическим свойствам условных конструкций, возникает закономерный вопрос: способны ли большие языковые модели аналогичным образом учитывать подобные семантико-прагматические явления? Именно этот вопрос рассматривается в практической части настоящего исследования, где мы воспроизведем исходные эксперименты с использованием языковых моделей.

2. Методология экспериментов

В данном разделе описывается методология проведения экспериментов на основе поставленных гипотез, включая выбор моделей, структуру датасетов и процедуру взаимодействия с LLM. Материалом для исследования послужили датасеты авторов оригинальной работы; все материалы, включая код и стимулы, доступны в открытом репозитории[2]. Мы строго следовали экспериментальному дизайну из статьи (Liu, Schwab 2022a), внеся лишь минимальные изменения, обусловленные спецификой работы с языковыми моделями: исключены филлерные предложения, так как для языковых моделей не требуется проверка их внимания, а также адаптированы промпты и добавлены контрольные вопросы для проверки понимания моделями процедуры экспериментов.

Выборка каждого эксперимента включала в себя восемь больших языковых моделей. В первых двух экспериментах использовался единый пул моделей: Gemini 2.0 Flash Experimental, LLaMA 3.3 70B Instruct, Qwen 2.5 72B Instruct, Mistral Nemo, Moonlight 16B A3B Instruct, Gemma 3 27B, DeepHermes 3 LLaMA 3 Preview 8B, DeepSeek R1 Distill LLaMA 70B. Доступ к моделям осуществлялся через API платформы OpenRouter. В третьем эксперименте модель Gemini была исключена из выборки из-за технических сбоев со стороны OpenRouter. В качестве замены использовалась модель Gemini 2.0 Flash Thinking Exp-01-21, предоставленная другим провайдером — Glama.

Все эксперименты проводились следующим образом: в начале каждой модели отправлялся вступительный промпт с объяснением задачи и критериями оценивания. После получения положительного ответа от модели касательно понимания

[2] https://github.com/iwantsomemarzipan/coursework_24-25

представленных инструкций мы переходили к основному этапу: модели последовательно получали стимулы из датасета и оценивали в соответствии с указанными вопросами.

2.1. Эксперимент 1

В соответствии с оригинальным исследованием для первого эксперимента были сформулированы следующие гипотезы:

1. Гипотетические индикативные условные предложения с ослабляющей NPI *all that* будут оцениваться как менее естественные по сравнению с контрфактическими условными предложениями, содержащими ту же NPI.
2. Между гипотетическими индикативными и контрфактическими условными предложениями без NPI не будет наблюдаться значимых различий в оценках естественности.

Эти гипотезы основывались на результатах предшествующего исследования на материале немецкого языка с NPI *sonderlich* и отражали ожидаемое влияние прагматических механизмов на лицензирование ослабляющих NPIs [7].

Датасет для первого эксперимента включал в себя 144 предложения, которые были разделены на 24 набора контекстов по 6 типов конструкций:

(6a) *If the students have been all that attentive during class, they will pass the exam.*

(6b) *If the students had been all that attentive during class, they would have passed the exam.*

(6c) *If the students have been attentive during class, they will pass the exam.*

(6d) *If the students had been attentive during class, they would have passed the exam.*

(6e) *All students who have been all that attentive during class will pass the exam.*

(6f) *All students who have been attentive during class will pass the exam.*

Предложения (6a,c) представляют собой представляют собой индикативные кондиционалы, (6b,d) — контрфактические. Предложения (6a-b) содержат NPI *all that* в своей антецедентной части, в отличие от (6c-d). Как указывалось в разделе 1, авторы оригинальной статьи также рассматривали влияние эффекта наличия или отсутствия ослабляющей NPI в примерах с квантором всеобщности (6e-f).

В этом эксперименте от моделей требовалось дать оценку предложениям по шкале от 1 до 7. В начале каждой модели отправлялся промпт с объяснением задачи и правил оценивания:

Листинг 1. Инструкционный промпт для эксперимента 1.

```
Hi! I want you to help me with my experiment.
I will send groups of English sentences. For each sentence, rate its
naturalness from 1 (completely unnatural) to 7 (completely natural).

RULES:
1. Return ONLY a comma-separated list of numbers (e.g., "5,3,6")
2. No explanations, no formatting, just numbers
3. Maintain exact order of input sentences

Do you understand the instructions? Answer strictly "yes" or "no".
```

После получения положительного ответа на вопрос о понимании инструкций модель последовательно получала промпты с предложениями из датасета, сгруппированные в одном батче по 6 штук:

Листинг 2. Пример промпта со стимулами из датасета эксперимента 1.

```
Rate each sentence's naturalness (1-7). Return EXACTLY 6 numbers separated
by commas, like: 5,3,6,2,4,7

Sentences:
1. If the roommates have been all that close during college, they will stay
in touch afterwards.
2. If the roommates had been all that close during college, they would have
stayed in touch afterwards.
3. If the roommates have been close during college, they will stay in touch
afterwards.
4. If the roommates had been close during college, they would have stayed
in touch afterwards.
5. All roommates who have been all that close during college will stay in
touch afterwards.
6. All roommates who have been close during college will stay in touch
afterwards.
```

Ответы модели логировались и автоматически записывались в датафрейм.

2.2. Эксперимент 2

Второй эксперимент был направлен на уточнение роли NPI в гипотетических кондиционалах и условных конструкциях с посылкой, с фокусом как на оценке естественности предложения, так и на предполагаемой степени уверенности говорящего в антецеденте. Гипотезы были следующими:

1. Гипотетические индикативные кондиционалы с NPI будут восприниматься как менее естественные, чем предложения с посылкой.
2. Без NPI различий в оценке естественности между двумя типами кондиционалов не ожидается.
3. Поскольку кондиционалы с посылкой несут пресуппозицию о правдоподобии антецедента, адресат будет приписывать более высокую степень веры говорящего в антецедент именно таким конструкциям.
4. На фоне результатов исследования (Liu 2019), которое показало, что наличие NPI в антецеденте уменьшает веру в пресуппозицию в гипотетических индикативных условиях, ожидалось, что в этих конструкциях присутствие NPI снизит приписываемую веру.

Во втором эксперименте датасет включал в себя 24 набора, каждый из которых был поделён на 4 типа условных конструкций, то есть число всех предложений составило 96. Все условия представляли собой мини-диалоги, состоящие из пяти предложений. В последнем предложении протагонист высказывает целевое условное предложение, оценка которого и являлась предметом эксперимента.

В качества примера каждого типа конструкций приведём следующий диалог: Susan Smith works at a community college. / Her colleague says:

(7a) *“The students have been attentive in class.”* / Susan Smith responds: / *“If the students have been attentive in class, they will pass the exam.”*

(7b) *“The students have been very attentive in class.”* / Susan Smith responds: / *“If the students have been all that attentive in class, they will pass the exam.”*

(7c) *“The students will start their exam season soon.”* / Susan Smith responds: / *“If the students have been attentive in class, they will pass the exam.”*

(7d) *“The students will start their exam season soon.”* / Susan Smith responds: / *“If the students have been all that attentive in class, they will pass the exam.”*

Предложения (7a-b) представляют собой условия с посылкой, в которых антецедент отражает пресуппозицию, представленную в предыдущем высказывании собеседника. Предложения (7c,d) — гипотетические индикативные кондиционалы. Примеры (7b,d) включают в антецедентную часть NPI *all that*, в то время как (7a,c) её не содержат.

В ходе эксперимента 2 языковым моделям было необходимо отвечать на два вопроса, касающихся целевого предложения (на примере предложений (7)):

1. How natural was the last sentence?
2. Does Susan Smith believe the students have been attentive in class?

Таким образом, измерялись как естественность предложений, так и уровень веры протагониста в антецедентную часть предложения. При ответе языковые модели должны были ориентироваться на шкалу+ от 1 (absolutely no) до 7 (absolutely yes).

Перед основным этапом эксперимента модели получали следующий инструкционный промпт:

Листинг 3. Инструкционный промпт для эксперимента 2.

```
Hi! I want you to participate in my experiment.

TASK STRUCTURE:
1. I will send you dialogues in the following format:
   <Dialogue N>
   Context: [4 context sentences]
   Sentence to evaluate: [last sentence]
   Questions:
   1. [question 1]
   2. [question 2]
   </Dialogue N>

2. For each dialogue, respond with answers for both question 1 and question
2.

Do you understand the instructions? Answer strictly "yes" or "no".
```

Пример промпта с диалогом, за которым следовало предложение для оценивания, представлен ниже:

Листинг 4. Пример промпта со стимулом из датасета эксперимента 2.

<Dialogue 14>
Jade Carter is the manager of a hotel. Her brother says: "The customers have been very satisfied with our service." Jade Carter responds:
"If the customers have been all that satisfied with our service, they will come back next year."
Questions for you:
1. How natural was the last sentence?
Please provide the score only from 1 (completely unnatural) to 7 (completely natural). Do not send any comments or reasoning.
Next question:
2. Does Jade Carter believe the customers have been satisfied with their service?
Please provide the score only from 1 (absolutely no) to 7 (absolutely yes). Do not send any comments or reasoning.
</Dialogue 14>

Несмотря на указание требований к формату ответов, нам не удалось добиться их стандартизированного вида от всех моделей для их автоматического парсинга, как это было сделано в предыдущем эксперименте. Вследствие этого записанные в таблицу ответы требовалось размечать вручную.

2.3. Эксперимент 3

Третий эксперимент был посвящён оценке частотности вывода о выполнении условия при почтении предложений двух типов условных конструкций (гипотетических индикативных и контрфактических). В частности, он был направлен на проверку двух гипотез:

1. CP-интерпретации будут возникать реже в контрфактических кондиционалах, чем в гипотетических индикативных по причине того, что умозаключение о CP требует дополнительной прагматической обработки, а контрфактические конструкции сопряжены с повышенной когнитивной нагрузкой ввиду необходимости представления как фактического, так и альтернативного состояния мира (неизвестно, насколько это актуально для языковых моделей).
2. CP-интерпретации будут встречаться чаще в предложениях, в которых антецедентная часть содержит обстоятельство степени *very*, чем в предложениях без модификатора в антецеденте (о выборе *very* вместо *all that* см. в разделе 2).

Взаимодействия между типом конструкции и наличием модификатора не предполагалось.

В эксперименте 3 использовался датасет, организованный в 24 набора контекстов, которые описывают определённую ситуацию. Каждый набор реализован в четырёх вариантах (всего 96 предложений):

Tom Scott loves watching movies and TV series. / New shows are coming out soon. / He says to his friend: /

(8a) *“If the TV shows have been well written, they will receive positive reviews.”*

(8b) *“If the TV shows had been well written, they would have received positive reviews.”*

(8c) *“If the TV shows have been very well written, they will receive positive reviews.”*

(8d) *“If the TV shows had been very well written, they would have received positive reviews.”*

Примеры (8a-b) представляют собой индикативный и контрфактический условные предложения соответственно без модификатора *very*, тогда как (8c-d)

содержат обстоятельство степени в антецедентной части. Использование наречия *very* вместо NPI *all that* мотивировано тем, что в эксперименте 1 уже рассматривалось влияние NPI на естественность кондиционалов. В настоящем эксперименте исследуется не оценка приемлемости предложений, а выводимые на основе этих предложений умозаключения. Как и в эксперименте с людьми, наличие *all that* могло потенциально исказить восприятие языковыми моделями вопросов, относящихся к целевому предложению.

В отличие от *all that*, наречие *very* не связано с полярностью и представляет собой нейтральный модификатор степени, позволяющий протестировать гипотезу о том, что выбор более “затратной” формы (например, *very careful* вместо *careful*) может активировать импликатуру — в том числе CP — через сопоставление с немодифицированной альтернативой. Таким образом, замена *all that* на *very* позволяет контролируемо проверить влияние модификации на частоту CP-выводов, исключив возможные побочные эффекты, связанные с полярностью.

Третий эксперимент предполагал получение ответов на два вопроса (на примере предложений (8):

1. Does Tom Scott believe the TV shows have been well written?
2. Второй вопрос зависел от типа условия:
    2.1. (для типов a, c): Does Tom Scott believe that the TV shows will only receive positive reviews if they have been (very) well written?
    2.2. (для типов b, d): Does Tom Scott believe that the TV shows would have only received positive reviews if they had been (very) well written?

Первый вопрос касался веры протагониста в правдивость антецедента, второй — возможности бикондициональной интерпретации предложения. Оба вопроса оценивались по шкале от 1 (absolutely no) до 7 (absolutely yes).

В начале эксперимента моделям отправлялось следующее пояснение к задаче:

Листинг 5. Инструкционный промпт для эксперимента 3.

```
Hi! I want you to participate in my experiment.

TASK STRUCTURE:
1. I will send you dialogues in the following format:
   <Dialogue N>
   Context: [3 context sentences]
   Sentence to evaluate: [last sentence]
   Questions:
   1. [question 1]
   2. [question 2]
   </Dialogue N>
```

```
2. For each dialogue, respond with answers for both question 1 and question
2.

Do you understand the instructions? Answer strictly "yes" or "no".
```

Промпты с диалогами и целевыми предложениями представлялись в таком виде:

Листинг 6. Пример промпта со стимулом из датасета эксперимента 3.

```
<Dialogue 64>
Fred Cooper runs for mayor. His agenda is progressive. He says to his
consultant:
"If the citizens have been very convinced by our campaign, they will vote
for me."
Questions for you:
1. Does Fred Cooper believe that the citizens have been convinced by their
campaign?
Please provide the score only from 1 (absolutely no) to 7 (absolutely yes)
without any comments or reasoning
2. Does Fred Cooper believe that the citizens will only vote for him if
they have been very convinced by their campaign?
Please provide the score only from 1 (absolutely no) to 7 (absolutely yes)
without any comments or reasoning
</Dialogue 64>
```

Как и в предыдущем эксперименте, не все модели соблюдали просьбу давать ответы в установленном формате, поэтому после завершения эксперимента ответы вручную размечались.

3. Статистические методы анализа

Для анализа собранных данных использовалось два подхода: описательная статистика и байесовское моделирование с порядковой регрессией.

3.1. Описательная статистика

Для получения первичного представления о чувствительности языковых моделей к виду кондиционалов, а также для последующей реализации выбора значений априорных распределений, был произведён расчёт описательной статистики. Анализ включал вычисление средних значений и стандартных отклонений оценок, сгруппированных по трем категориям: типы условных конструкций, используемые языковые модели и наборы контекстов.

3.2. Байесовское моделирование

Авторы оригинального исследования использовали байесовские порядковые регрессионные модели для проверки гипотез. Мы воспроизвели этот подход[3] с некоторыми модификациями, связанными с особенностями данных, полученных от языковых моделей. В частности, из-за меньшего объёма выборки в сравнении с оригинальным исследованием, где выборки участников составляли 75, 50 и 59 человек в эксперименте 1, 2 и 3 соответственно, при обучении моделей возникала проблема расхождений (divergent transitions), которая затрудняла надёжную оценку параметров.

Расхождения возникают в процессе градиентного моделирования, лежащего в основе алгоритмов Гамильтоновой выборки, таких как HMC и NUTS[4]. При симуляции траекторий частиц в пространстве параметров используется метод приближения, и если шаг интегрирования оказывается слишком большим относительно кривизны апостериорного распределения, траектория может отклониться от истинной (Stan Development Team 2024: 187-188). Такие отклонения и называются расхождениями. Когда они происходят, симуляция теряет точность: соответствующие точки не используются в построении выборки и в итоге снижают всю её репрезентативность. Другими словами, наличие расхождений нежелательно ввиду того, что они являются индикатором нерепрезентативности одной или более выборок для апостериорного распределения (Goldfeld 2020).

С целью минимизации числа расхождений в процессе семплирования рекомендуется либо перепараметризовать модель, либо изменить настройки сэмплера, в частности, повысить значение целевой вероятности принятия шага (adapt_delta) (Bürkner 2024: 30).

В нашем случае в качестве меры перепараметризации был выбран подбор более информативных априорных распределений, заменяющих значения по умолчанию в brms. Для коэффициентов фиксированных эффектов использовались априоры в виде нормального распределения с нулевым математическим ожиданием и стандартным отклонением, эмпирически определённым на основе результатов описательной статистики. Для случайных эффектов применялись априоры в виде экспоненциального распределения, где параметр λ рассчитывался как величина, обратная к стандартному отклонению, также определённому по данным предварительного анализа. Все прочие параметры модели соответствовали оригинальному исследованию: каждая модель

[3] Мы использовали пакет brms языка R, версия 2.22.0.
[4] Данные алгоритмы используются в Stan, на базе которого работает brms.

обучалась в течение 8000 итераций, из которых половина использовалась для разогрева (warm-up), а максимальная глубина дерева установлена на уровне 12.

Далее будет представлено более подробное описание тех регрессионных моделей, которые использовались для анализа данных, полученных в ходе проведённых нами экспериментов.

3.2.1. Регрессионные модели для эксперимента 1

Для составления результатов по первому эксперименту было построено 4 модели. Две модели анализировали данные, в которых сравнивались индикативные и контрфактические условные предложения с и без NPI *all that*, две другие — данные, в которых сравнивались предложения с квантором всеобщности *all* и без него. В каждой паре моделей первая использовала контрастное кодирование (sum-to-zero) предикторов, где коэффициенты отражают отклонения от общего среднего, а вторая — dummy-кодировку (reference-level coding) с явным выделением базового уровня.

При составлении формул регрессий в качестве фиксированных эффектов использовалась NPI, тип условной конструкции (гипотетическая индикативная и контрфактическая) или же наличие квантора всеобщности — в зависимости от модели — а также их взаимодействие. В список случайных эффектов входили пересекающиеся случайные перехваты и наклоны для каждого из предикторов и их взаимодействия по группам языковых моделей и наборов контекстов, что позволяло учесть вариативность в ответах, обусловленную конкретными моделями или контекстами.

Модели обучались с использованием информативных априорных распределений, что соответствует стратегии перепараметризации для повышения устойчивости сэмплирования. Для коэффициентов фиксированных эффектов применялось нормальное распределение с нулевым средним и стандартным отклонением, равным 0.9. Априоры для стандартных отклонений случайных эффектов задавались в виде экспоненциальных распределений с параметром $\lambda = 0.68$ для эффекта модели и $\lambda = 0.68$ для эффекта контекстного набора.

Помимо построения регрессионных моделей для проверки выдвинутых гипотез (см. раздел 2.1) были сформулированы и оценены линейные ограничения на параметры моделей с dummy-кодировкой, которые проверяли «пропущенные» парные сравнения, отсутствующие в регрессионных моделях. В частности, рассматривалось два ключевых сравнения: различия между индикативными и контрфактическими предложениями без NPI и различия между контрфактическими предложениями с NPI и без. Апостериорные

распределения соответствующих линейных комбинаций регрессионных коэффициентов позволяли оценить вероятность направленных эффектов.

3.2.2. Регрессионные модели для эксперимента 2

Для анализа были построены две регрессионные модели: для оценки естественности предложения и приписываемой степени веры в антецедент. Обе модели включали в качестве фиксированных эффектов наличие NPI и тип условной конструкции, а также их взаимодействие. В качестве случайных эффектов учитывались пересекающиеся случайные перехваты и наклоны по группам языковых моделей и контекстов. Переменные фиксированных эффектов были контрастно закодированы.

В отличие от первого эксперимента, для моделей второго эксперимента было создано два набора априоров. Это обусловлено различиями в распределении ответов и степени вариативности между группами в двух задачах, выявленными на этапе описательной статистики. Для модели оценки естественности были заданы следующие априоры: у фиксированных эффектов — $\mu = 0$, $\sigma = 0.95$, у случайных — $\lambda = 1.25$ и $\lambda = 1.19$ у эффекта моделей и контекстов соответственно. К модели веры в пресуппозицию применялись такие априоры: $\mu = 0$, $\sigma = 1.75$ у фиксированных эффектов, $\lambda = 0.67$ у случайного эффекта моделей, $\lambda = 0.66$ у случайного эффекта контекстов.

3.2.1. Регрессионные модели для эксперимента 3

Как и в предыдущем эксперименте, в третьем строилось две модели для анализа данных по вопросу о уверенности в антецеденте и бикондициональной интерпретации. В качестве фиксированных эффектов выступали переменные обстоятельства степени *very* и типа условной конструкции, случайные эффекты были такие же, как и в экспериментах 1 и 2. К переменным фиксированных эффектов была применена контрастная кодировка. Для первой модели был составлен следующий набор априоров: $\mu = 0$, $\sigma = 1.5$ у фиксированных эффектов, $\lambda = 0.56$ и $\lambda = 0.52$ у случайных эффектов; для второй модели: $\mu = 0$, $\sigma = 1.35$ у фиксированных эффектов, $\lambda = 1.15$ и $\lambda = 1.1$ у случайных эффектов.

4. Результаты статистических методов

В этом разделе будут описаны результаты статистических методов по каждому эксперименту.

4.1. Результаты эксперимента 1

Ниже представлены графики распределения значения средних оценок и их стандартных отклонений в зависимости от типа конструкции.

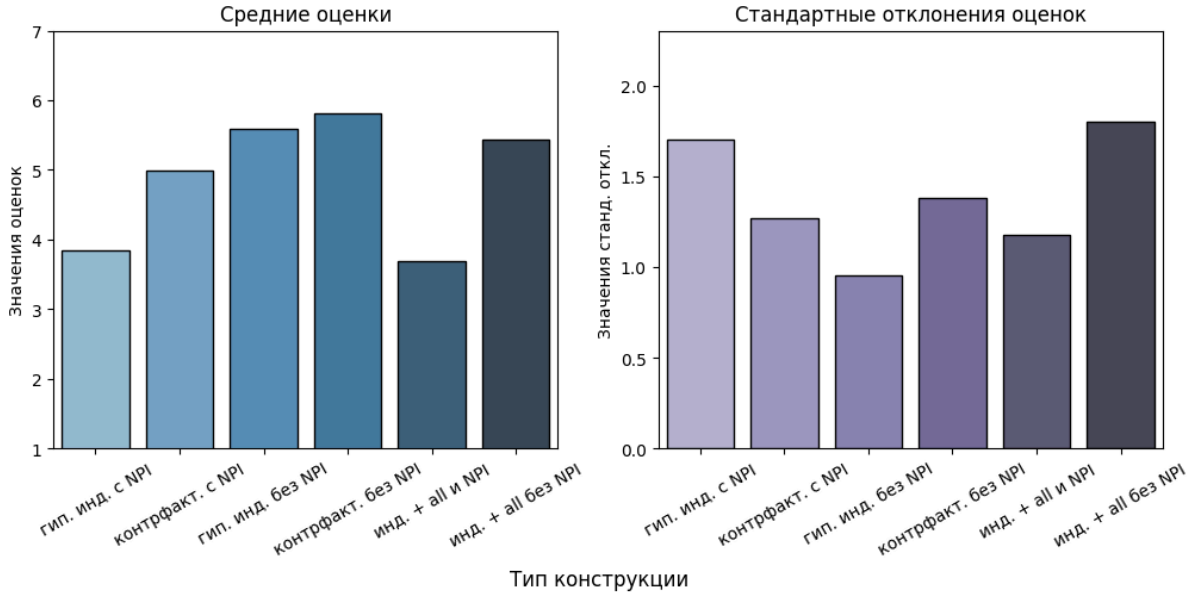


Рис. 1. Сравнение типов предложений по средним их оценок и стандартным отклонениям (эксп. 1)

На первом графике рис. 1 видно, что в среднем самые высокие оценки (около 6) получили гипотетические индикативные и контрфактические кондиционалы без NPI, а самые низкие — гипотетические индикативные кондиционалы и индикативные предложения с квантором всеобщности *all*, содержащие NPI (около 4). Примечательно, что в среднем языковые модели не склонны оценивать естественность предложений меньше 4.

Второй график показывает, что наибольший размах в оценках наблюдается для гипотетических индикативных кондиционалах с NPI (на уровне около 1.7) и индикативных предложений с квантором всеобщности, но без *all that* (1.8). Наименьший разброс характерен для гипотетических индикативных предложений без NPI (0.96), а также индикативных предложений с *all* и *all that* (1.18). При изучении последующих графиков для данных по экспериментам 2 и 3 можно будет отметить, что оценки из датасета первого эксперимента отличились наиболее выраженными значениями стандартных отклонений.

Для подбора значения экспоненциального априорного распределения случайных эффектов считались средние стандартных отклонений оценок, распределённых по моделям и наборам контекстов. Округлённые до тысячных значения двух средних равны 1.478 и 1.635 соответственно.

Отдельно стоит рассмотреть распределение средних оценок среди каждой модели.

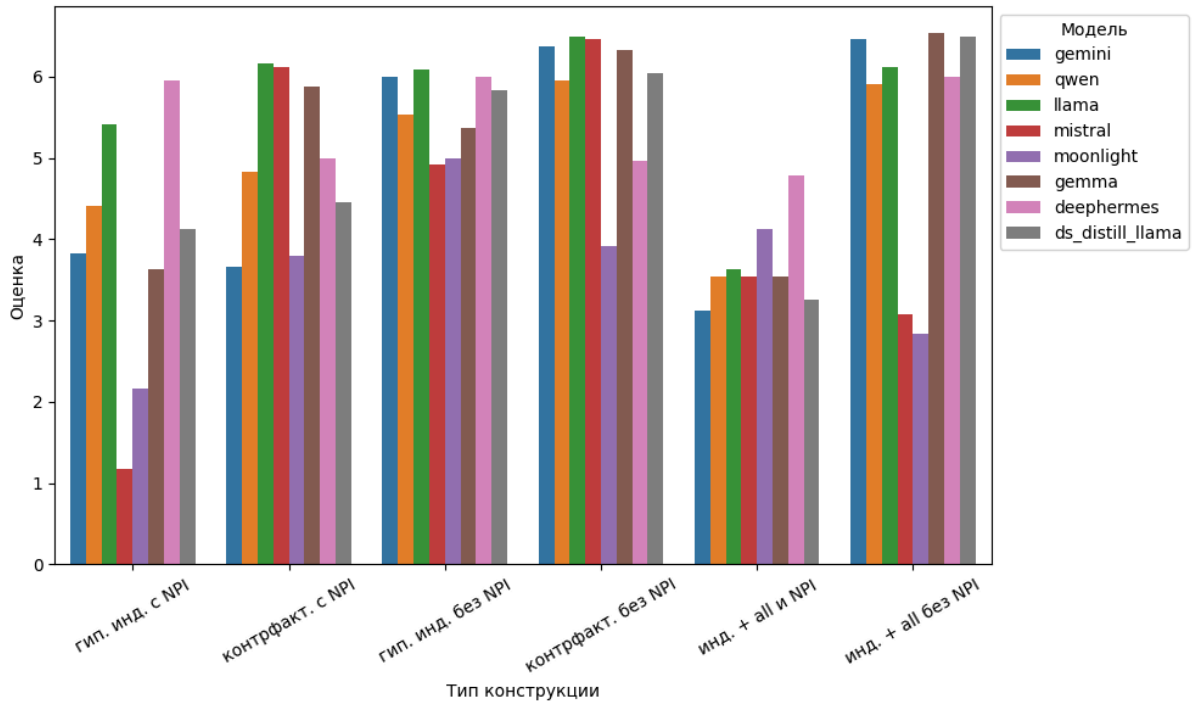


Рис. 2. Сравнение средних оценок по условным конструкциям и моделям (эксп. 1)

На рис. 2 видно, что распределение средних неоднородно как с точки зрения типов условных конструкций, так и относительно моделей. В частности, модель Mistral Nemo оценивала крайне низко гипотетические индикативные предложения с NPI и предложения с квантором без NPI, при этом среди остальных типов конструкций она следует общим трендам. Схожий паттерн можно отметить для модели Moonlight 16B A3B Instruct, относительно среднего оценила хуже предложения 1, 2, 4 и 6 вида. Дополнительно можно изучить график столбчатых диаграмм, сгруппированных по моделям, который представлен в приложении 1.

Перейдём к результатам байесовского моделирования. Напоминаем, что помимо гипотетических индикативных и контрфактических кондиционалов мы также сравнивали гипотетические условные предложения с предложениями, содержащие квантор всеобщности *all*.

Апостериорные оценки первого анализа указывают на возможный эффект взаимодействия между типом условной конструкции и присутствием NPI: предложения с *all that* в гипотетической индикативной форме воспринимаются как менее естественные по сравнению с контрфактическими ($\beta = -0.61$, 95% CrI = [–1.53, 0.36], $P(\beta < 0) = 0.9$), однако эффект оказался менее устойчивым по сравнению с тем, что

наблюдалось в оригинальном исследовании (β = –0.48, 95% CrI = [–0.81, –0.16], P(β < 0) = 1). В условиях без NPI различия между индикативными и контрфактическими конструкциями отсутствуют (β = 0.03, 95% CrI = [–1.28, 1.21]).

Во втором анализе не было обнаружено значимого взаимодействия между наличием NPI и типом предложения, β = 0.10, CrI [–1.14, 1.35], P(β < 0) = 0.56. С точки зрения главных эффектов без NPI оценки условных предложений и предложений с квантором не различались (β = 0.25, 95% CrI [–0.65, 1.16]), тогда как с NPI условные предложения оценивались значительно естественнее, β = 1.86, 95% CrI [0.77, 2.75], причём эффект оказался более выраженным, чем в эксперименте с людьми (β = 1.29, 95% CrI [0.18, 0.93]).

4.2. Результаты эксперимента 2

Два рисунка ниже демонстрируют значения средних оценок и их стандартных отклонений отдельно для вопроса 1 (естественность предложения) и 2 (уверенность в правдивость антецедента). При просмотре рис. 3 можно заметить, что для всех типов условных конструкций значения средних и стандартных отклонений практически не отличаются друг от друга: естественность предложений оценивалась на уровне 6, разброс в оценках составляет около 1. Для данных по второму вопросу присуща чуть большая неоднородность: диапазон значения среднего составляет [4.66; 5.73], а стандартного отклонения — [1.62; 1.76].

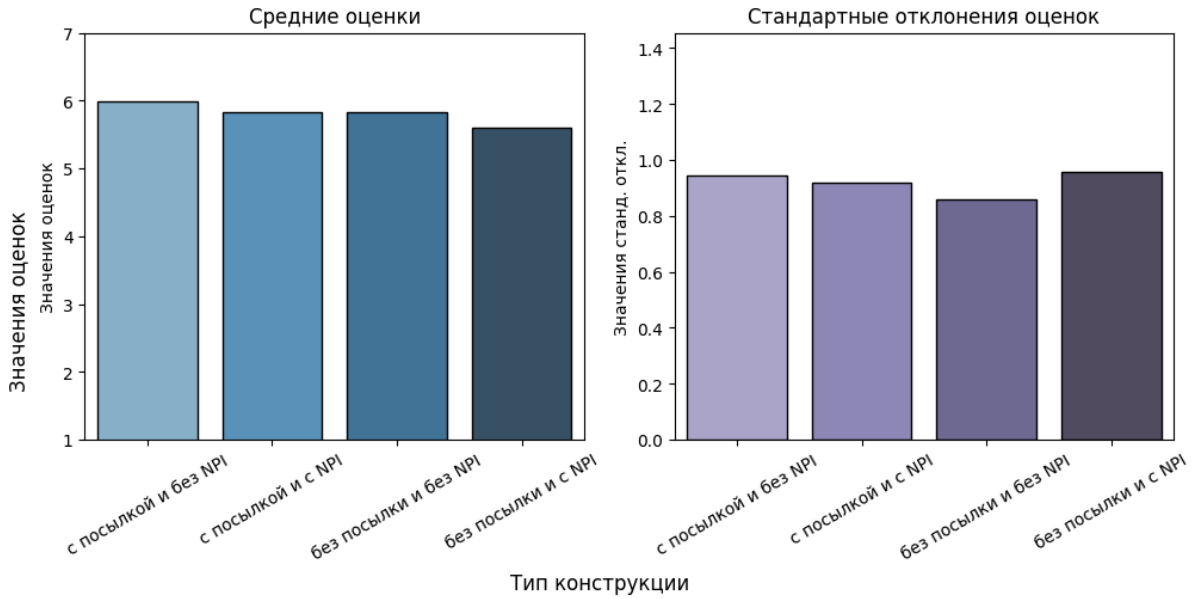


Рис. 3. Сравнение типов предложений по средним их оценок и стандартным отклонениям (эксп. 2 вопрос 1)

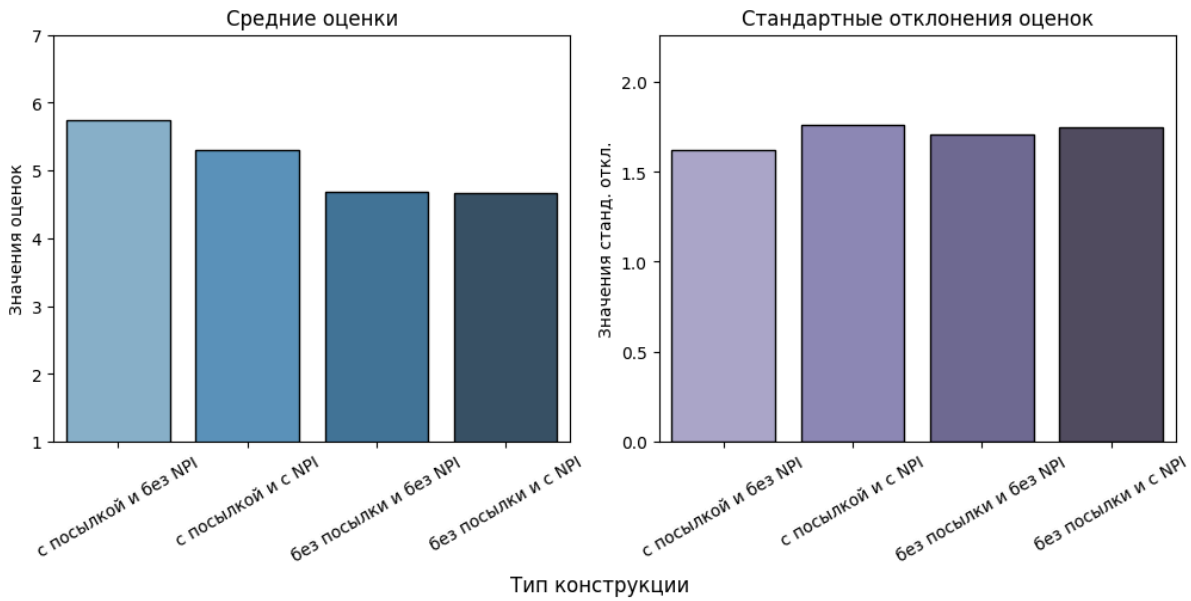


Рис. 4. Сравнение типов предложений по средним их оценок и стандартным отклонениям (эксп. 2 вопрос 2)

Средние стандартных отклонений для предложений, сгруппированных по моделям и наборам контекстов, составили 0.798 и 1.941 соответственно для вопроса 1 и 0.868 и 0.908 для вопроса 2.

Рис. 5 ниже позволяет увидеть, что в среднем модели соблюдают общую тенденцию при ответе на вопрос 1. График, касающийся второго вопроса, показывает, например, что Gemini 2.0 Flash Experimental склонна ниже всех остальных моделей уверенность протагониста в антецедент, вне зависимости от типа предложения. Moonlight 16B A3B Instruct, наоборот, показала расположенность к выдаче наиболее высоких оценок.

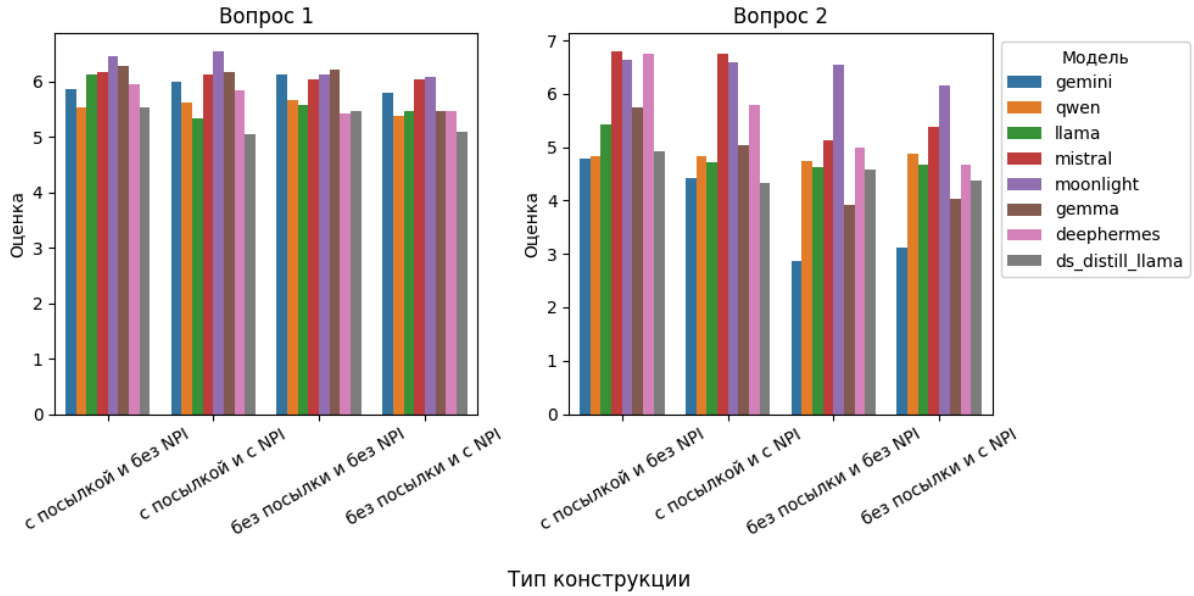


Рис. 5. Сравнение средних оценок по условным конструкциям и моделям (эксп. 2)

График сгруппированных по моделям столбчатых диаграмм расположен в приложении 2.

В рамках байесовского анализа мы обсудим отдельно результаты двух моделей. Модель оценки естественности доказала влияние эффекта NPI: предложения с *all that* воспринимались как менее естественные по сравнению с предложениями без неё (β = 0.34, 95% CrI = [0.06, 0.62], P($\beta > 0$) = 0.986). Также наблюдалась поддержка эффекта типа условной конструкции: кондиционалы с посылкой оценивались как несколько более естественные по сравнению с обычными гипотетическими индикативными (β = 0.36, 95% CrI = [–0.01, 0.73], P($\beta > 0$) = 0.972). Однако взаимодействие между наличием NPI и типом условного не получило подтверждения: β = –0.04, 95% CrI = [–0.53, 0.44], P($\beta < 0$) = 0.579. Это означает, что снижение оценки естественности, вызванное *all that*, одинаково проявляется в обоих видах кондиционалов. Таким образом, полученные результаты не воспроизводят ключевой эффект взаимодействия, описанный в оригинальной статье, где сообщалось о значимом влиянии типа условной конструкции на восприятие естественности при наличии NPI (β = –0.37, 95% CrI = [–0.70, –0.05], P($\beta < 0$) = 0.988).

По результатам второй регрессионной модели были выявлены оба главных эффекта: предложения с *all that* ассоциировались с меньшей уверенностью в антецеденте (β = 0.27, 95% CrI = [–0.01, 0.55], P($\beta > 0$) = 0.971), а протагонисты в предложениях с посылкой воспринимались как выражающие большую уверенность по сравнению с предложениями без (β = 0.88, 95% CrI = [0.10, 1.62], P($\beta > 0$) = 0.984). Кроме того, наблюдалось значимое взаимодействие между этими факторами, β = 0.50, 95% CrI = [0.04, 0.95], P($\beta > 0$) = 0.983), что указывает на разную степень влияния NPI в зависимости от типа условной конструкции: наличие *all that* сильнее ухудшает веру в антецедент в предложениях с посылкой. Направление эффекта оказалось обратным в сравнении с оригинальным исследованием, где влияние all that было выражено сильнее в условных конструкциях без посылки, однако в том случае и сама вероятность взаимодействия была ниже (β = –0.20, 95% CrI = [–0.49, 0.09], P($\beta > 0$) = 0.917).

4.3. Результаты эксперимента 3

Рассмотрим представленные графики распределения средних оценок и их стандартных отклонений на основе данных, полученных в ходе эксперимента 3. Рис. 6 демонстрирует данные по вопросу 1 (о вере протагониста в правдивость антецедента), рис. 7 — по вопросу 2 (об импликации бикондициональности).

В первую очередь мы хотим обратить внимание на крайне низкие оценки контрфактических условных предложений на первом графике рис. 4, составляющие контраст другим видам конструкций, которые оценивались по аналогичному вопросу в предыдущем эксперименте. Значения стандартных отклонений в достаточной мере выражены для всех типов конструкций ([1.31; 1.55]).

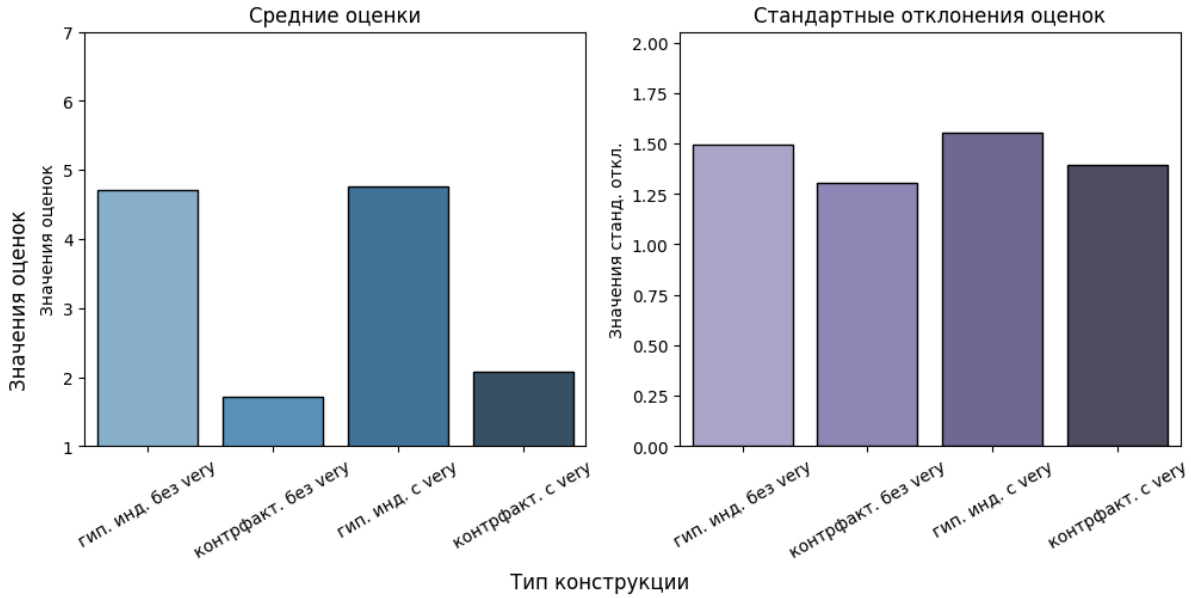


Рис. 6. Сравнение типов предложений по средним их оценок и стандартным отклонениям (эксп. 3 вопрос 1)

Первый график на рис. 7 иллюстрирует крайне высокие средние оценки всех типов предложений, которые раньше не наблюдались. На втором графике заметна разница в значениях стандартного отклонения: наименьший разброс в оценках присущ гипотетическим индикативным условиям, не содержащих обстоятельство степени *very* (0.78), наибольший — контрфактическим с *very* (1.35).

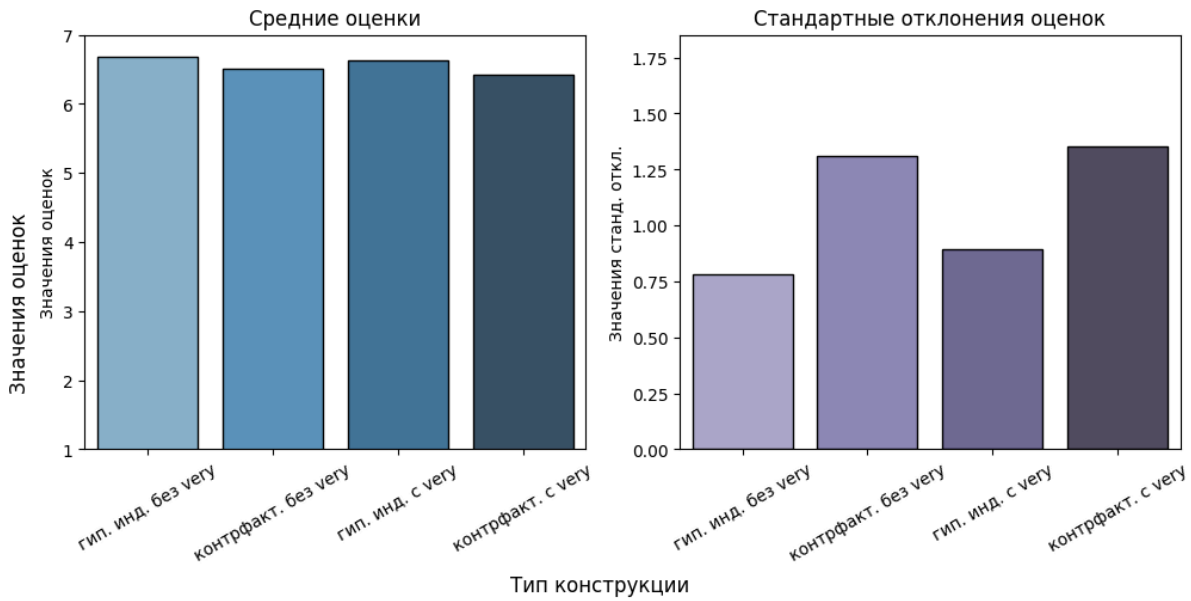


Рис. 7. Сравнение типов предложений по средним их оценок и стандартным отклонениям (эксп. 3 вопрос 2)

Далее левый график на рис. 8 иллюстрирует неоднородность в оценках уверенности протагониста в антецедент. Наиболее стабильно высокие оценки давались моделями Mistral Nemo и DeepHermes 3 LLaMA 3 Preview 8B, а стабильно низкие — Gemini 2.0 Flash Thinking Exp-01-21, Gemma 3 27B и DeepSeek R1 Distill LLaMA 70B. На втором графике можно пониженное среднее значение для модели Moonlight 16B A3B Instruct при оценивании предложений с NPI.

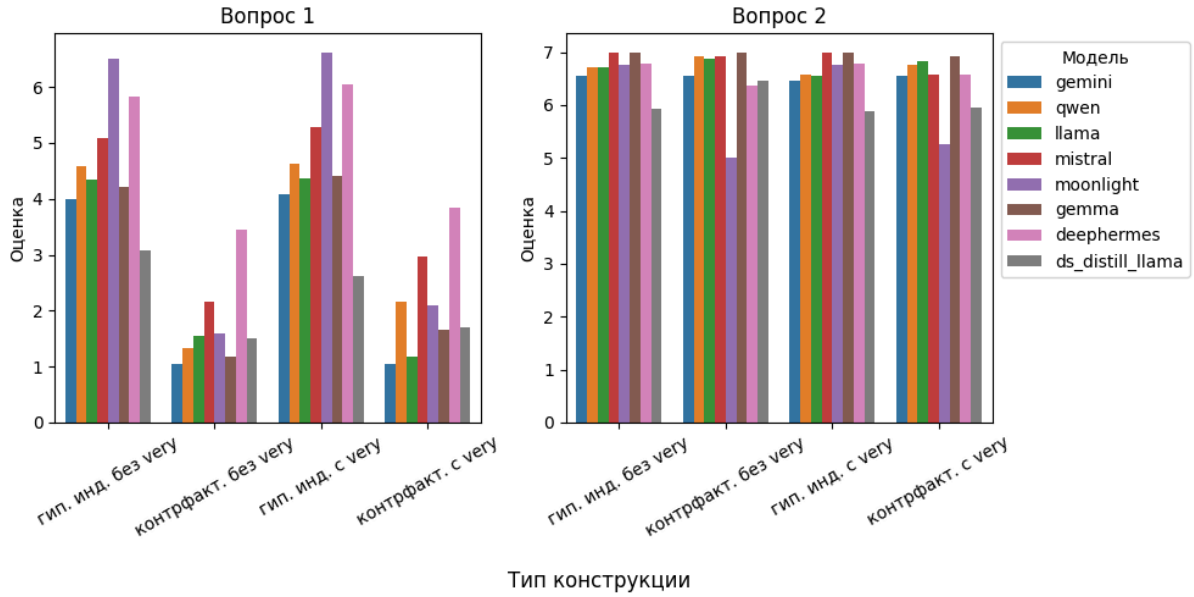


Рис. 8. Сравнение средних оценок по условным конструкциям и моделям (эксп. 3)

Для изучения тенденции распределения средних оценок по моделям можно обратиться к графику в приложении 3.

Далее представлены результаты байесовского моделирования. Подсчёты модели, которая анализировала данные по первому вопросу, показали наличие эффекта типа условного предложения: вера в антецедент снижается для контрфактических условий по сравнению с гипотетическими индикативными ($\beta \approx 3.03$, 95% CrI [1.49, 4.20], $P(\beta > 0) = 0.999$), что согласуется с результатами оригинальной статьи ($\beta \approx 0.72$, 95% CrI [0.47, 0.98], $P(\beta > 0) = 1$), хотя оценка коэффициента регрессии в нашем случае намного выше. В то же время влияние эффекта наличия *very* оказалось вполне значимым по подсчётам нашей регрессионной модели ($\beta \approx 0.35$, 95% CrI [–0.00, 0.68], $P(\beta > 0) = 0.9745$), в отличие от исходного исследования. Эффект взаимодействия, как и предполагали авторы, не был ими обнаружен, однако наши результаты показали обратное — при наличии обстоятельства степени уверенность в антецеденте уменьшается сильнее в контрфактических кондиционалах ($\beta \approx –0.58$, 95% CrI [–1.15, 0.00], $P(\beta > 0) = 0.975$).

По результатам второй модели в исходной статье авторы отмечали слабую поддержку главных эффектов, а также отсутствие эффекта взаимодействия. Результаты нашей модели демонстрируют аналогичное отсутствие эффекта взаимодействия ($\beta \approx 0.07$, 95% CrI [–0.49, 0.63], $P(\beta > 0) = 0.597$) и ещё менее значимую вероятность эффекта модификатора ($\beta \approx –0.13$, 95% CrI [–0.48, 0.19], $P(\beta > 0) = 0.802$) и вида условия ($\beta \approx 0.37$, 95% CrI [–0.55, 1.25], $P(\beta > 0) = 0.809$). В оригинальном исследовании соответствующие вероятности составляли $P(\beta > 0) = 0.885$ и $P(\beta > 0) = 0.936$ соответственно; несмотря на то, что второе значение может свидетельствовать в пользу наличия эффекта типа условной конструкции, авторы воздерживаются от однозначных выводов.

5. Обсуждение результатов

Для более глубокого понимания того, как языковые модели обрабатывают семантические и прагматические явления, рассмотренные в рамках настоящей работы, необходимо сопоставить восприятие языковых конструкций моделями и людьми. Такое сопоставление будет опираться как на первоначально выдвинутые гипотезы каждого эксперимента, так и на дополнительные наблюдения, зафиксированные как в оригинальной статье, так и в ходе собственного анализа. Цель данного раздела — оценить степень соответствия между реакциями языковых моделей и человеческими данными. На основе этого сопоставления мы постараемся определить, насколько

понимание языка языковыми моделями приближено к человеческому с точки зрения лицензирования NPI и прагматики условных конструкций.

Начнём с обсуждения восприятия естественности. Результаты первого и второго экспериментов показали, что языковые модели, подобно людям, оценивают предложения с отрицательной полярной частицей *all that* как менее естественные по сравнению с предложениями без неё. При этом попарные сравнения продемонстрировали, что в отсутствие NPI различия между условиями нивелируются. Кроме того, в первом эксперименте было установлено, что наличие NPI сильнее снижает оценку естественности в гипотетических условных конструкциях, чем в контрфактических. Эти наблюдения подтверждают обе гипотезы, выдвинутые для первого эксперимента.

Тем не менее, анализ выявил также некоторые расхождения между реакциями языковых моделей и результатами, полученными на людях. В частности, во втором эксперименте не наблюдалось эффекта взаимодействия между наличием NPI и типом условной конструкции, в отличие от оригинального исследования, где такой эффект был зафиксирован. Однако это расхождение не является уникальным: аналогичная картина была получена в немецкой версии эксперимента [7], с которой авторы статьи также проводят сопоставление.

Возможным объяснением наличия эффекта взаимодействия в английском материале может служить то, что *all that* может вызывать premise-прочтение чаще индикативных кондиционалов, в рамках которого наличие NPI считается приемлемым. Согласно предположению авторов, частица *all that* может получать анафорическую интерпретацию при наличии ударения, как, например, в высказывании *He's not all* ***that*** *stupid* (Onea, Sailer 2013), что позволяет участникам достраивать контекст, в котором антецедент уже утверждён, например: A: *The students have been (very) attentive during class.* — B: *If the students have been all* ***that*** *attentive...* Это, в свою очередь, может приводить к трактовке предложения, как если бы оно имело посылку.

Таким образом, несмотря на общее совпадение направленности основных эффектов, отсутствие взаимодействия в данных моделей может указывать либо на различие в интерпретации конструкций, либо на ограниченную чувствительность моделей к прагматическим контекстам, активирующим premise-интерпретации. С другой стороны, учитывая результаты немецкого эксперимента, нельзя исключать и вероятность статистического совпадения.

Далее мы рассмотрим вопрос о вере протагониста в антецедент. Результаты второго эксперимента подтверждают гипотезу о том, что для предложений с посылкой модели склонны приписывать говорящему большую уверенность в антецедент по сравнению с гипотетическими индикативными конструкциями. Это наблюдение согласуется с выводами оригинальной статьи. В третьем эксперименте также было зафиксировано снижение приписываемой веры в контрфактических конструкциях по сравнению с гипотетическими, что вновь воспроизводит исходные данные.

Вместе с тем влияние модификатора степени *very* в нашем исследовании оказалось статистически значимым, в отличие от исходной работы, где авторы не выявили этого эффекта. Также, в то время как в оригинальной статье не было обнаружено эффекта взаимодействия между типом условного предложения и наличием *very*, наша модель, напротив, показала, что в присутствии модификатора вера в антецедент значительно сильнее снижается именно в контрфактических конструкциях.

Наибольшую неоднозначность вызывает интерпретация гипотезы о влиянии NPI *all that* на приписываемую уверенность в гипотетических условных, которая рассматривалась во второй части эксперимента 2. В оригинальной работе авторы предполагали, что *all that* снижает приписываемую уверенность говорящего в антецеденте именно в гипотетических условных конструкциях, не выдвигая чётких ожиданий касательно предложений с посылкой. Эту гипотезу они формально подтвердили, ссылаясь на наличие эффекта взаимодействия: в присутствии NPI гипотетические условные конструкции оценивались как менее убедительные по сравнению с кондиционалами с посылкой.

В нашем исследовании также наблюдался эффект взаимодействия, однако направленность эффекта оказалась противоположной: частица *all that* сильнее снижала уровень приписываемой уверенности в условных предложениях с посылкой, в то время как в предложениях без неё влияние частицы было менее выраженным. Таким образом, хотя гипотеза авторов формально не воспроизводится, она и не опровергается — результаты скорее указывают на смещение чувствительности конструкций к NPI в сторону условных конструкций, содержащих посылку.

Однако если обратиться к данным описательного анализа, можно заметить, что средние оценки приписываемой веры в антецедент в гипотетических конструкциях практически не различаются в зависимости от наличия NPI, что может рассматриваться как потенциальное опровержение гипотезы о влиянии NPI на гипотетические конструкции.

Особое внимание заслуживает категоричность языковых моделей в оценке контрфактических конструкций, рассматриваемых в третьем эксперименте. В отличие от участников оригинального исследования, языковые модели приписывали крайне низкую степень веры в антецедент в контрфактических условиях. Это резкое снижение оценок контрастирует с неожиданно высокими значениями в данных на людях, которыми были озадачены авторы исходной работы. В качестве возможных объяснений предлагались следующие факторы: недостаточная добросовестность выполнения задания участниками в связи с отсутствием контрольных вопросов, нежелание делать категоричные суждения о возможной точки зрения другого человека (протагониста в вопросе) при ограниченном контексте, а также особенности английского языка, в котором контрфактичность обозначается лишь формой прошедшего времени, в отличие от других языков (например, немецкого), где используются грамматические маркеры на глаголе, выражающие сослагательное наклонение..

Примечательно, что языковые модели в этом аспекте демонстрируют более “предсказуемое” поведение, явно меньше оценивая степень веры в антецедент в контрфактических конструкциях. Это может говорить о том, что модели более чувствительны к формальным грамматическим признакам, чем к прагматическим или дискурсивным особенностям, которые могут влиять на интерпретацию у людей.

Перейдём к обсуждению CP-интерпретаций, касательно которых были заключены гипотезы в третьем эксперименте. Как и в оригинальном исследовании, мы исходили из предположения, что такие интерпретации могут зависеть от типа условной конструкции и наличия обстоятельства степени в антецеденте. Однако результаты второй модели из третьего эксперимента показали, что ни тип кондиционала, ни наличие модификатора степени *very* не оказывают надёжного влияния на частоту бикондициональных прочтений.

Этот результат воспроизводит данные оригинальной статьи, в которой авторы также не обнаружили значимого различия между гипотетическими индикативными и контрфактическими конструкциями, несмотря на предварительное ожидание, что контрфактичность будет препятствовать выводам о CP-интерпретации по причине её большей когнитивной сложности. Точно так же предполагалось, что фокусировка на модификатора степени могла бы способствовать бикондициональным прочтениям, однако полученные данные не подтвердили и этого.

При этом авторы оригинального исследования отмечают наличие слабых сигналов в пользу предполагаемых эффектов, которые были упомянуты в разделе

байесовского анализа эксперимента 3, однако они также указали на то, что доверительные интервалы соответствующих оценок были достаточно широки и включали как положительные, так и отрицательные значения. Это позволило им сделать вывод, что влияние может быть слишком малым для выявления в рамках текущего дизайна и мощности эксперимента. Исследователи подчёркивают, что для более уверенного вывода необходимы репликации с увеличенной выборкой.

Наш анализ, основанный на аналогичной байесовской методологии, вероятно, унаследовал те же ограничения. Мы не наблюдаем значимых эффектов, однако и не исключаем их существования: полученные распределения допускают возможность слабых эффектов, которые могли остаться невыявленными в рамках существующей статистической мощности.

Проведённое исследование позволяет сделать ряд выводов касательно способности больших языковых моделей учитывать и обрабатывать семантико-прагматические свойства условных конструкций. Во-первых, LLM демонстрируют схожую с людьми чувствительность к грамматическим и семантическим различиям между типами кондиционалов, что может свидетельствовать о наличии у моделей систематизированных представлений о семантике условных предложений. Во-вторых, языковые модели частично воспроизводят прагматически опосредованные эффекты. В частности, они склонны снижать оценки естественности предложений с ослабляющей NPI, причем степень этого снижения значимо варьируется в зависимости от типа условной конструкции, как это было показано на примере сравнения гипотетических индикативных условных предложений с контрфактическими. Данный факт сигнализирует о том, что языковые модели способны учитывать не только формальное наличие NPI в предложении, но и прагматически зависимые особенности контекста их употребления, что не может быть объяснено исключительно распределением частотности примеров с NPI в обучающих данных для LLM. Такая же зависимость от типа кондиционала была отмечена при анализе вопроса о вере протагонистов предложений в антецедент при сравнении предложений с посылкой и гипотетических индикативных, а затем контрфактических и гипотетических индикативных. При этом наблюдаемое расхождение между влиянием NPI *all that* или обстоятельства степени *very* на оценку веры в антецедент при сравнении с человеческими данными ставят под сомнение способность LLM к подлинному пониманию прагматических механизмов, оставляя открытым вопрос о том, отражают ли их ответы системное усвоение языковых норм или случайные

статистические артефакты. Тем не менее, данная неопределённость касается именно явления предположения о ложности антецедента, входящего в область импликатур, обработка которых может быть сложнее задачи определения естественности предложений. Можно предположить, что при оценке приемлемости языковых конструкций LLM действительно руководствуются относительно глубоким пониманием семантико-прагматических аспектов языка, однако импликатуры представляют собой лингвистический домен, где их компетенция пока ограничена.

Что касается механизма Conditional Perfection, результаты эксперимента — как с участием людей, так и с языковыми моделями — не подтвердили ожидаемого эффекта контрфактических кондиционалов и модификатора. Это может указывать на два возможных сценария: либо выбранный экспериментальный дизайн (как в оригинальной работе, так и в данном воспроизведении) оказался недостаточно чувствительным к предполагаемым эффектам, либо сами гипотезы оказались несостоятельными. Единственное, что можно утверждать с определённой уверенностью — LLM способны распознавать бикондициональные прочтения на человекоподобном уровне. Однако эта констатация мало проясняет, насколько качественно языковые модели на самом деле интерпретируют механизм CP и его свойства и при каких условиях.

Заключение

В данной работе были воспроизведены эксперименты из психолингвистического исследования (Liu, Schwab 2022a) с использованием больших языковых моделей в качестве участников. Целью было оценить, способны ли LLM интерпретировать особенности языка, связанные с лицензированием ослабляющей NPI *all that* и прагматическими характеристиками условных предложений, на уровне людей.

Полученные результаты позволяют сделать вывод о том, что хотя языковые модели не воспроизводят поведение людей в идентичном виде, они всё же демонстрируют некоторый уровень понимания природы рассмотренных в настоящей работе семантико-прагматических явлений, влияющих на интерпретацию дискурса и генерацию импликатур. Это подчёркивает, что потенциал LLM в обсуждаемых лингвистических областях ещё не исчерпан и заслуживает дальнейшего изучения.

Перспективным направлением развития исследования может стать детальное рассмотрение индивидуальных особенностей языковых моделей. Обобщённый

количественный анализ, использованный в данной работе, позволяет выявить общие тенденции, однако он в определённой мере маскирует различия между архитектурами языковых моделей. Углублённый сравнительный анализ на уровне отдельных моделей способен дать более точное представление о пределах их интерпретационных возможностей.

Литература

Bürkner 2024 — P. Bürkner. Package ‘brms’, ver. 2.22.0. 2024.

Cai et al. 2023 — Z. G. Cai, X. Duan, D. A. Haslett, M. J. Pickering, S. Wang. Do large language models resemble humans in language use? *arXiv*. https://arxiv.org/abs/2303.08014. 2023.

Chan et al. 2022 — S. C. Y. Chan, A. Creswell, I. Dasgupta, F. Hill, D. Kumaran, A. K. Lampinen, J. L. McClelland, H. R. Sheahan. Language models show human-like content effects on reasoning tasks. *arXiv*. https://arxiv.org/abs/2207.07051. 2022.

Collins, Skovgaard-Olsen 2021 — P. Collins, N. Skovgaard-Olsen. Indicatives, Subjunctives, and the Falsity of the Antecedent // *Cognitive Science* 45 (11), 2021.

Dentella et al., 2023 — V. Dentella, F. Günther, E. Leivada. Systematic testing of three Language Models reveals low language accuracy, absence of response stability, and a yes-response bias // Susan Goldin-Meadow (ed.). *Psychological and Cognitive Sciences* 120 (51), 2023.

Ding et al. 2025 — M. Ding, Z. Fu, H. Liu, X. Liu, R. Ning, C. Zhang, Y. Zhang. Logical Reasoning in Large Language Models: A Survey. *arXiv*. https://arxiv.org/abs/2502.09100. 2025.

Goldfeld 2020 — K. Goldfeld. Diagnosing and dealing with degenerate estimation in a Bayesian meta-analysis. *R-Bloggers*. https://www.r-bloggers.com/2020/08/diagnosing-and-dealing-with-degenerate-estimation-in-a-bayesian-meta-analysis/. 2020.

Holliday, Mandelkern, 2024 — W. H. Holliday, M. Mandelkern. Conditional and Modal Reasoning in Large Language Models. *arXiv*. https://arxiv.org/html/2401.17169v1. 2024.

Lampinen 2022 — A. Lampinen. Can language models handle recursively nested grammatical structures? A case study on comparing models and humans. *arXiv*. https://arxiv.org/abs/2210.15303. 2022.

Liu 2019 — M. Liu. The elastic nonveridicality property of indicative conditionals // *Linguistics Vanguard*. Berlin: De Gruyter, 2019.

Liu, Schwab 2022a — M. Liu, J. Schwab. Processing Attenuating NPIs in Indicative and Counterfactual Conditionals // X. Jiang (ed.). *Frontiers in Psychology* 13, 2022.

Liu, Schwab 2022b — M. Liu, J. Schwab. Attenuating NPIs in indicative and counterfactual conditionals // D. Gutzmann, S. Repp (eds.). *Proceedings of Sinn Und Bedeutung* 26. Cologne: University of Cologne, 2022. P. 772–789.

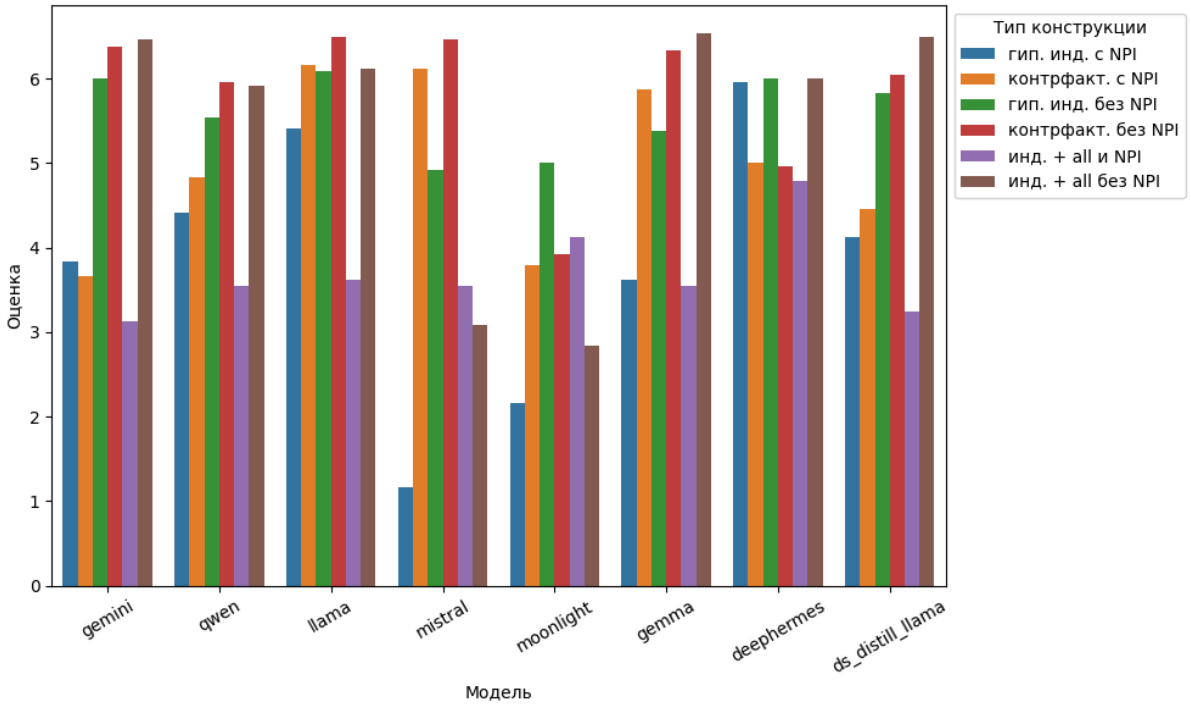
Onea, Sailer 2013 — E. Onea, M. Sailer. “Really all that clear?” in Beyond ‘Any’ and ‘Ever’ // E. Csipak, R. Eckardt, M. Liu, M.Sailer (eds.). Berlin, Boston, MA: De Gruyter Mouton, 2013. P. 323–350.

Stan Development Team 2024 — Stan Development Team. Stan Reference Manual, ver. 2.36. 2024.

Приложения

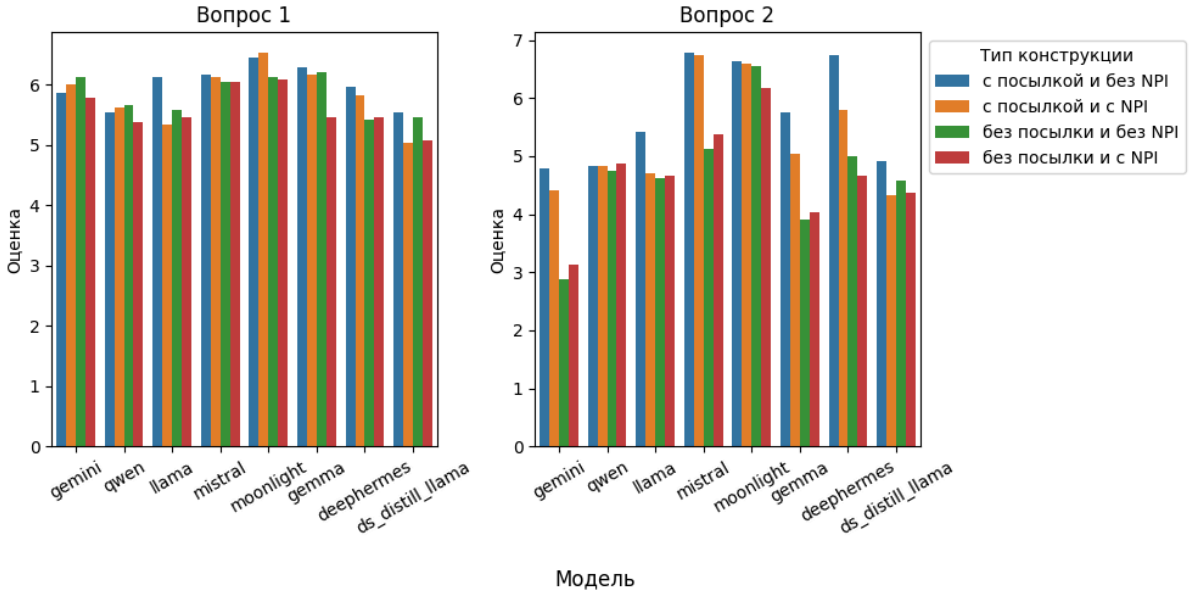
Приложение 1

График распределения средних оценок для каждого типа конструкций, сгруппированных по моделям (по данным эксперимента 1).



Приложение 2

График распределения средних оценок для каждого типа конструкций, сгруппированных по моделям (по данным эксперимента 2).



Приложение 3

График распределения средних оценок для каждого типа конструкций, сгруппированных по моделям (по данным эксперимента 3).