**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: Оценка грамматических способностей больших языковых моделей на основе набора данных CoLA

*Assessing the grammatical abilities of large language models using the CoLA dataset*

Студентка 2 курса

группы № 221

Гимадиева Далия Ильдаровна

Научный руководитель

Сериков Олег Алексеевич

Программист-разработчик

Москва, 2024 г.

**Оглавление**

1. Введение

2. Предшествующие исследования

3. Материалы и методы исследования

3.1. Модели

3.2. Набор данных CoLA

3.3. Сценарии

3.4. Промптинг

3.5 Схема оценки ответов

3.5.1 Автоматическая оценка

3.5.2 Ручная разметка

4. Results

4.1 Llama2 Results

4.1.1 Label

4.1.2. Label, category

4.1.3. Label, correction

4.1.4. Label, explanation

4.1.5. Label, category, correction

4.1.6. Label, explanation, correction

4.1.7 Label, category, explanation

4.1.8. Label, category, explanation, correction

4.2 Vicuna Results

4.2.1. Label

4.2.2. Label, category

4.2.3. Label, correction

4.2.4. Label, explanation

4.2.5. Label, category, correction

4.2.6. Label, explanation, correction

4.2.7. Label, category, explanation

5. Сравнение моделей, обсуждение результатов

6. Заключение

7. Литература

8. Приложения

# 1. Введение

Большие языковые модели (LLM), такие как GPT-4 и BERT, произвели в последние годы настоящую революцию в области искусственного интеллекта и обработки естественного языка (NLP). Они используются в множестве областей, в том числе в областях машинного перевода, аналитики данных и чат-ботов, демонстрируя уровень генерации и анализа текста, сравнимый с человеческим. Кроме того, успех больших языковых моделей вызвал огромную волну рассмотрения их в качестве потенциальных моделей овладения и обработки человеческого языка, опровергающих в некоторой степени Универсальную грамматику Хомского о «врожденной» гипотезе языковых компетенций [Хомский 1965]. В связи с этим появилась необходимость исследования языковых моделей с точки зрения обладания имплицитного знания грамматики и их использования для корректировки и исправления текста учащихся L2.

Цель исследования — изучение способности больших языковых моделей Llama2 и Vicuna выносить суждения о лингвистической приемлемости и предоставлять объяснения ошибок при нарушении языковых норм.

Задачи исследования:

− разработать сценарии запросов для получения выдачи с бинарной классификацией, а также разъяснением ошибки при ее наличии;

− выявить наиболее подходящие для моделей промпты, позволяющие получить корректный по структуре и содержанию ответ без дополнительного обучения;

− определить роль, при которой модели показывают наиболее высокие результаты бинарной классификации по метрике Accuracy и коэффициенту корреляции Мэтьюса (MCC), для дальнейшего прогона предложений из набора данных CoLA с этой ролью;

− разработать принципы оценивания ответов моделей;

− определить особенности выдачи каждой модели с учетом взаимосвязи между сценарием и качеством суждения;

− выявить сходства и расхождения в ответах моделей, в частности грамматические категории, с которыми модели допускают ошибки.

# 2. Предшествующие исследования

Предшествующие исследования в данном направлении выполнены преимущественно на английском языке и состоят из изучения способности больших языковых моделей работать с отдельными феноменами.

Так, в работе [Zhong 2023] исследуется способность ChatGPT понимать и обрабатывать текст на различных уровнях сложности, включая наличие метафор и лингвистической неоднозначности, в работе [Ortega-Martín 2023] также изучается ChatGPT, но с точки зрения способности модели справляться именно с лингвистической неоднозначностью и интерпретировать многозначные слова и фразы в различных контекстах.

В рамках исследования грамматических способностей различных языковых моделей написаны работы [Mikhailov 2022] и [Touvron 2023], которые используют бинарную классификацию для оценивания приемлемости предложений создаваемых датасетов на русском и английском языках соответственно. Подобным образом бинарная классификация используется для оценки точности распознавания грамматических ошибок и стабильности повторных ответов тремя моделями GPT в работе [Dentella 2023]. Возможности больших языковых моделей в контексте русскоязычного текста изучаются, например, в [Студеникина 2022], где языковые способности моделей исследуются на материале предикативного согласования.

# 3. Материалы и методы исследования

## 3.1. Модели

В рамках данной работы оценивались грамматические способности двух языковых моделей — Llama2, являющейся одним из лидеров в гонке LLM, и Vicuna. В случае обеих моделей исследовались версии с 13b параметрами, что позволило получить более высокое качество ответов по сравнению с версиями 7b и обойти ограничение памяти при запуске версий с большим числом параметров.

Llama2 — это генеративная модель второго поколения с открытым кодом, разработанная Meta AI как доработанная версия Llama в сотрудничестве с Microsoft. Одним из главных усовершенствований является внедрение активационной функции ReLU, позволяющей достигать высокой активационной разреженности без потери производительности. Llama2 обучалась на датасетах из общедоступных источников (Github, Wikipedia, ArXiv) на 512 GPU NVidia A100 в течение 1.7млн GPU-часов [Touvron 2023]. Несмотря на поддержку 20 языков, основным языком модели считается английский.

Vicuna — модель, являющаяся доработанной Llama2. Vicuna была обучена на диалогах с ChatGPT, собранных с помощью ресурса ShareGPT, на 8 GPU NVidia A100 в течение 1 дня. Модель показала качество на уровне 90% при оценке GPT-4 [Sun 2023] по методу оценивания LLM-as-a-judge. Как и Llama2, Vicuna поддерживает различные языки, основным является английский.

## 3.2. Набор данных CoLA

Для изучения грамматических способностей больших языковых языковых моделей был использован датасет «The Corpus of Linguistic Acceptability» (далее CoLA), состоящий из 10657 предложений на английском языке из 23 лингвистических публикаций, аннотированных на приемлемость их первоначальными авторами [Warstadt 2019].

Так как CoLA разделен на несколько наборов, для прогона языкового материала был выбран набор данных «raw/in\_domain\_dev.tsv», включающий 527 предложений в необработанном виде, 162 из которых были размечены авторами как ошибочные. Из всех наборов данных был выбран именно этот, так как он снижает вероятность того, что предложения были использованы при обучении или тестировании и уже были знакомы исследуемым моделям в последствии окончания процесса их обучения.

## 3.3. Сценарии

Для рассмотрения поведения моделей в ситуациях с различными лингвистическими аспектами, сложностью предложений и феноменов было разработано 8 сценариев, состоящих из комбинаций 4 аспектов:

1. бинарная классификация ‘label’;
2. категория ошибки ‘category’;
3. объяснение ‘explanation’;
4. исправление ‘correction’.

Из приведенных выше аспектов бинарная классификация — единственный аспект, встречающийся в каждом из 8 сценариев, поскольку он необходим для автоматического подсчета точности ответа моделей (см. п. 3.5.1). Так как от моделей ожидается ответ, который устроил бы обычного пользователя-учащегося L2, желающего проверить предложение на наличие ошибок, в этом аспекте ожидается ответ в формате «Да, есть ошибка» или «Нет, ошибок нет». В последствии для автоматической оценки такой формат переводился в бинарные 0 и 1.

В ответ на запросы с аспектом категории ошибки от моделей ожидается называние конкретной категории из нескольких приведенных пользователем (см. п. 3.4.). Так как данные CoLA не содержат информацию, какой тип ошибки наблюдается в предложении, в рамках данной работы также были дополнительно размечены категории ошибок в неверных предложениях. При разметке учитывались принципы определения типов феноменов аналогичного датасета RuCoLA для русского языка [Mikhailov 2022], исключая тип «галлюцинация».

Особое внимание уделяется аспекту объяснения, поскольку именно в нем раскрываются способности моделей выносить суждения о лингвистической приемлемости. Важен также аспект исправления, показывающий, какие предложения модель считает допустимыми и в каких ситуациях возникают изменения ответа модели относительно наличия ошибки прямо внутри выдачи.

Таким образом, имеем 8 сценариев:

1. ‘label’ — бинарная классификация-лейбл
2. ‘label, category’ — лейбл, категория
3. ‘label, correction’ — лейбл, исправление
4. ‘label, explanation’ — лейбл, объяснение
5. ‘label, category, correction’ — лейбл, категория, исправление
6. ‘label, explanation, correction’ — лейбл, объяснение, исправление
7. ‘label, category, explanation’ — лейбл, категория, объяснение
8. ‘label, category, explanation, correction’ — лейбл, категория, объяснение, исправление

## 3.4. Промптинг

Поскольку исследование грамматических способностей больших языковых моделей не подразумевает дополнительное обучение, в рамках работы не были проведены донастройки моделей на наборах данных, описанных с помощью инструкций [URL: <https://www.promptingguide.ai/ru/techniques/zeroshot>], а также не были использованы промпты с примерами референсных ответов, подсказками и другие техники промпт-инжиниринга. Следовательно, при составлении запросов использовался подход, называемый *zero-shot prompting* ‘нулевка’.

Для обеспечения согласованности мы стремились придерживаться гиперпараметров по умолчанию для каждой модели. Для Llama2 были установлены температура temperature=0.3, частот дискредитации top\_p=0.95, а также ограничение max\_tokens=150, позволяющее ограничить длину ответа для получения более конкретного ответа. Код, использовавшийся для подключения к моделям, можно найти в репозитории по ссылке: <https://github.com/dduend/Llama2-Vicuna-on-grammar>.

Прогон первого сценария лейбла с бинарной классификацией осуществлялся по 3 раза каждой модели с двумя ролями ассистент и учитель английского языка. Далее считалось среднее значение и стандартное отклонение и выбиралась роль, при которой модели показывали более высокие результаты. Оставшиеся сценарии прогонялись также по 3 раза с лучшей ролью. Роли формулировались в промпты следующим образом для обеих моделей Llama2 и Vicuna:

1. *You are a helpful, respectful and honest assistant. Always answer as helpfully.* — Ты полезный, уважительный и честный помощник. Твои ответы всегда полезны.
2. *You are an English teacher. Always answer as helpfully, with good explanations of grammar and lexical mistakes.* — Ты учитель английского языка. Всегда отвечаешь полезно, с хорошими объяснениями грамматических и лексических ошибок.

Запрос, подаваемый моделям, всегда начинался с краткого описания задачи и содержал количество пунктов, равное количеству аспектов в сценарии. Так, промпт для сценария ‘label, category, explanation, correction’ звучал следующим образом:

1. *I wrote a sentence in English, but I'm not sure if it's grammatically correct. Here's my sentence: {t} 1. Are there any grammar mistakes in my sentence? Yes or No; 2. If yes, what is the category of the mistake: syntax, morphology or semantics ? If there's no mistakes, write 'None'; 3. Explain the mistakes you see. If there's no mistakes, write 'None'; 4. What is the correct version of the sentence?* (Я написал предложение на английском языке, но не уверен, что оно грамматически правильное. Вот мое предложение: {t} 1. Есть ли в моем предложении грамматические ошибки? Да или нет; 2. Если да, то какая категория ошибки: синтаксическая, морфологическая или семантическая? Если ошибок нет, напишите «Нет»; 3. Объясните ошибки, которые вы видите. Если ошибок нет, напишите «Нет»; 4. Как звучит правильная версия предложения?)

## 3.5 Схема оценки ответов

### 3.5.1 Автоматическая оценка

Оценка ответов моделей аспекта «лейбл» с бинарной классификацией в каждом выделенном сценарии была произведена автоматически с помощью метрик точности (Accuracy) и коэффициента корреляции Мэтьюса (Matthews Correlation Coefficient, далее MCC).

Необходимость использования MCC заключалась в несбалансированности набора данных «raw/in\_domain\_dev.tsv» с перекосом в сторону предложений без ошибок, поскольку эта метрика учитывает истинные и ложные классификации и может использоваться даже в условиях сильного дисбаланса классов. Результат оценки изменяется в диапазоне от -1 до 1, 0 соответствует случайному предсказанию, -1 указывает на полное расхождение между фактом и предсказанием модели, то есть вместо положительного класса модель всегда предсказывает отрицательный и наоборот [Орешков 2023].

### 3.5.2 Ручная разметка

Аспекты «категория ошибки», «объяснение» и «исправление» были оценены в ручную. Ответы на каждый из аспектов модели получали баллы как за соответствие формату, так и за корректный результат.

Поскольку набор данных включает предложения, которые содержат несколько ошибок и/или имеют несколько потенциальных исправлений, а также сами модели могут называть несколько категорий ошибок, на данном этапе исследования ответы, содержащие хотя бы одну категорию из присутствующих в разметке, считались правильными по критерию «содержание». Модели получали ноль баллов при назывании другой категории по критерию «содержание» или при отсутствии ответа по обоим критериям «структура» и «содержание».

Аспект «исправление» имел аналогичные принципы оценки: модели получали ноль баллов за оба критерия при отсутствии ответа и 1 балл по критерию «содержание» при корректном исправлении. Если модели меняли семантику или убирали часть предложения, давалось 0,5 баллов.

Оценка аспекта «объяснение» осуществлялась по шести пунктам (см. Таблица 1), по три на каждый критерий «структура» и «содержание». Так, первые 2 балла модели могли получить за корректную локализацию ошибки (1 балл из 2 при некорректной локализации, но ее присутствии). Далее объяснение модели делилось на две части. Первая часть включала описание ошибки, содержащейся в предложении. Полный балл давался, когда модель правильно называла часть речи и категорию (переходный глагол, возвратное местоимение и т. п. ), половина балла — при отсутствии конкретной категории. Вторая часть включала исправления и объяснение, почему слово или клауза с ошибкой должна иметь именно такое изменение. Если локализация ошибки была неверна, дальнейшие ответы не обнулялись, модели могли получить баллы за соответствие структуре в частях «объяснение часть1» и «объяснение часть2». Стоит добавить, что ошибки в пунктуации, указанные моделью, мы не учитывали.

Таким образом, модель могла получить максимально при ручной разметке от 2 до 10 баллов в зависимости от аспекта в размечаемом сценарии.

Таблица 1. Пример разметки объяснения модели

| *Sally kissed himself. ‘*Салли поцеловала себя-mal’ | | | |
| --- | --- | --- | --- |
| The grammar mistake in your sentence is the use of the pronoun "himself." "Himself" is a reflexive pronoun, which means it refers back to the subject of the sentence, who is performing the action on themselves. In this case, the subject is "Sally," but the pronoun "himself" is not appropriate for a female person. To correct the sentence, you could say "Sally kissed herself." with a female reflexive pronoun "herself." (‘Грамматическая ошибка в вашем предложении — использование местоимения "себя". "Себя" — возвратное местоимение, что означает, что оно относится к подлежащему предложения, которое совершает действие над самим собой. В данном случае подлежащее — "Салли", но форма местоимение "себя" не подходит для лица женского рода. Чтобы исправить предложение, можно сказать: "Sally kissed herself". с возвратным местоимением женского рода "себя".’) | | The subject of the sentence "Sally" should be followed by a verb that agrees with it in number and gender. Since "kissed" is a third-person singular verb, it does not agree with the first-person pronoun "himself". To correct this mistake, you could change the wording to "Sally kissed herself." (‘За подлежащим в предложении "Салли" должен следовать глагол, согласующийся с ним по числу и роду. Поскольку "kissed" — это глагол третьего лица единственного числа, он не согласуется с местоимением первого лица "himself". Чтобы исправить эту ошибку, вы могли бы изменить формулировку на "Sally kissed herself."’) | |
| 1 | the pronoun "himself" (‘местоимение "himself-mal"’) — локализация дана | 1 | the subject of the sentence "Sally" (‘подлежащее предложения Салли"’) — локализация дана |
| 1 | the pronoun "himself" (‘местоимение "himself-mal"’) — локализация верна | 0 | локализация неверна, ошибка в местоимении "himself" |
| 1 | the pronoun "himself" is not appropriate for a female person (‘местоимение "himself-mal" не подходит лицу женского рода’) — объяснение часть1 дано, категории названы | 1 | "kissed" is a third-person singular verb, it does not agree with the first-person pronoun "himself" "kissed" (‘глагол третьего лица единственного числа, он не согласуется с местоимением первого лица "himself"’) — объяснение часть1 дано, категории названы |
| 1 | объяснение часть1 верно | 0 | объяснение часть2 неверно, категории неправильные |
| 1 | "Sally kissed herself." with a female reflexive pronoun "herself." (‘"Sally kissed herself." с возвратным местоимением женского рода "herself-**f**"’) — объяснение часть2 дано, категории названы | 0,5 | change the wording to "Sally kissed herself." (‘ измените формулировку на "Sally kissed herself-**f**."’) — штраф 0,5 балла за отсутствие конкретной категории в части2 |
| 1 | объяснение часть2 верно | 0 | объяснение часть2 неверно |
| 6/6 |  | 2,5/6 |  |

# 4. Результаты

В данном разделе кратко рассмотрим некоторые особенности ответов моделей на каждый из разработанных сценариев, в том числе отдельные прогоны, отличающиеся от других. Поскольку Vicuna основана на LLama2, также исследуемой в рамках данной работы, ответы моделей имеют некоторые общие черты (см. п. 5), полный список которых можно увидеть в таблицах в приложениях. Таблицу с результатами всех прогонов также можно найти в приложениях работы.

## 4.1 Результаты Llama2

### 4.1.1 Label

Сценарий, содержащий только бинарную классификацию, был прогнан с двумя ролями, ассистент и учитель английского языка. Несмотря на одинаковые гиперпараметры модели и запросы, результаты показали, что точность определения предложения с нарушением языковых норм в среднем значительно лучше с классической ролью ассистента: модель показала значения 0,6363 ± 0,0278 и 0,1988 ± 0,0253 по метрикам Accuracy и MCС с ролью ассистента и 0,5003 ± 0,0631 и 0,0684 ± 0,0151 с ролью учителя английского языка соответственно.

Как видно по результатам, при роли учителя работа Llama2 с учетом несбалансированной выборки близка к случайному предсказанию (значения по метрике MCС почти равны нулю). Это объясняется большим количеством самоисправлений внутри ответа модели, возможных благодаря тому, что наш запрос не подразумевал строгого ответа «Да» или «Нет» и допускал объяснения ошибок даже в сценарии, где это было необязательным. То есть с ролью учителя, в отличие от ассистента, модель склонна к дополнительной коммуникации с пользователем (реплики со стороны пользователя в таких случаях предоставляет модель) с целью помочь ему самому обнаружить ошибку, а также дать советы по изучению иностранного языка. Подобный формат ответа хорошо подходит для использования моделей L2, но пока кажется неактуальным из-за низкой способности находить предложения с нарушением языковых норм.

При этом с обеими ролями модель склонна давать полные ответы, включающие исправления и объяснения ошибок после ответа по шаблону «Да, в вашем предложении есть ошибка»/ «Нет, ваше предложение не содержит грамматических ошибок».

Таким образом, следующие сценарии были прогнаны с ролью ассистент, показавшей лучший результат на 3 прогонах.

### 4.1.2. Label, category

Среднее значение разметки аспекта «категория» по 3 прогонам составило 1,9149 ± 0,0095 при средних Accuracy 0.4896 ± 0.0242 и MCC 0.1519 ± 0.0340.

Результаты прогона датасета показывают, что LLama2 склонна называть в большинстве случаев синтаксис единственной категорией ошибки предложения. Так, синтаксис является единственной категорией среди всех 132 предложений, отмеченных моделью как имеющие языковые нарушения, в 1 прогоне, 97 из 134 в 2 прогоне и 107 из 134 в 3 прогоне соответственно. Во втором прогоне сценария модель также называет 2 категории морфологии, 22 синтаксиса и морфологии, 11 синтаксиса и семантики и 2 морфологии и семантики. В третьем прогоне категория морфологии встречается 1 раз, синтаксис и морфология — 13, синтаксис и семантика — 12, морфология и семантика — 1, один раз модель вместо ответа дает ‘None’ (‘Нет’). Давая в ответе в основном «синтаксис», Llama2 неправильно называет категорией именно его вместо семантики. В трех выдачах модель дает «морфологию» вместо ожидаемого синтаксиса.

### 4.1.3. Label, correction

Среднее значение разметки аспекта «исправление» по 3 прогонам составило 1,8926 ± 0,1112 при средних Accuracy 0.6996 ± 0.0039 и MCC 0.1152 ± 0.305.

Результат Llama2 в этом сценарии выше результата Vicuna, что объясняется склонностью модели выдавать положительные ответы: в 1 прогоне модель указала на наличие ошибки лишь в 3 предложениях из 527 предложений датасета, в 2 прогоне — 8 (из них аннотированы ошибочными 5), в 3 — 12 (из них аннотированы ошибочными 9). Это единственный сценарий, в котором модель предпочитает выдавать в качестве ответа «Нет ошибки».

Среди выдачи модели в 2 и 3 прогонах по одному ответу получили штраф 0,5 баллов за изменение семантики при исправлении. Еще два предложения были оценены на 0 баллов, поскольку они продолжали содержать ошибки.

### 4.1.4. Label, explanation

Среднее значение ручной разметки аспекта «объяснение» по 3 прогонам составило 3,3191 ± 0,0354 из 6. Все размеченные ответы модели получили 1 балл за критерий структуры локализации, что является лучшим результатом по данному критерию среди всех сценариев с данным аспектом, поскольку при других сценариях хотя бы один прогон показывает результат по критерию, меньший 1.

Результаты разметки также показывают, что со сценарием ‘label, explanation’ модель предпочитает давать исправления, называя необходимые категории, а не давать объяснения допущенных ошибок: среднее значение критерия «объяснение часть1» равно 0,7240 ± 0,0118, в то время как «объяснение часть2» — 0,8175 ± 0,0908. При этом значения по корректности объяснений в обеих частях примерно равны.

### 4.1.5. Label, category, correction

Ответы Llama2 в сценарии ‘label, category, correction’ характеризуются самым высоким среди всех сценариев значением по метрике МСС, равным 0,2002 ± 0,042, и наименьшим разнообразием дополнительных исправлений, которые привели к штрафу 0,5 баллов за изменение семантики в аспекте «correction». Так, в отличие от других сценариев с этим аспектом, не убирает из предложения слово или целую клаузу, не переделывает повествовательное предложение в вопросительное или наоборот и не меняет число. Среднее значение разметки по двух аспектам составило 3,5677 ± 0,0396.

Все языковые феномены, в определении которых модель ошибается, называя их синтаксическими ошибками, согласно нашей разметке, должны были быть семантическими. При этом в 1 прогоне модель дает в качестве категории только синтаксис, один раз не называя категорию вообще (и, в следствие, получает 0 по обоим критериям аспекта), в 2 и 3 прогонах также появляются категории «синтаксис и семантика», «синтаксис и морфология».

### 4.1.6. Label, explanation, correction

Результаты автоматической оценки сценария ‘label, category, correction’ — вторые по качеству обеих метрик Accuracy (0,6945 ± 0,0279) и MCC (0,1940 ± 0,0646). Учитывая, что наиболее высокое качестве Accuracy LLama2 показала при прогоне сценария ‘label, correction’, состоящий в основном из положительных ответов, можно считать, что лучшее значение без учета несбалансированной выборки показывает именно ‘label, category, correction’, содержащий в среднем на каждый прогон 79 предложений, отмеченных моделью как ошибочные.

Аспект «объяснение» в этом сценарии имеет одну характеристику, встречающуюся реже в других сценариях: капитализация. Модель предлагает исправить написание нарицательного существительного, чтобы оно писалось с заглавной буквы, считая, что это имя собственное (4). Другой тип капитализации также встречается чаще всего в этом сценарии — исправление написания имени собственного, которое уже написано с большой буквы (5). Помимо капитализации ‘label, category, correction’ чаще остальных в качестве ошибки предлагает убрать слово или целую клаузу.

1. *The word "water" should be capitalized because it's a proper noun* — Слово "вода" должно быть написано с заглавной буквы, так как это имя собственное
2. *"Calvin" should be capitalized because it is a proper noun (the name of a person)* — "Кэлвин" должно быть написано с заглавной буквы, так как это имя собственное (имя человека)

Кроме того, сценарий ‘label, category, correction’ показывает самые высокие результаты критериев ручной разметки в пунктах «объяснение часть1» и «объяснение часть2» — в среднем по 0,8706 ± 0,0081 и 0,8835 ± 0,0117 соответственно.

### 4.1.7 Label, category, explanation

Среднее значение ручной разметки обоих аспектов данного сценария по 3 прогонам составило 5,2407 ± 0,0438 из 8. Ответ Llama2 в рамках аспекта «объяснение» можно считать прототипическим ответом модели, поскольку все ошибки, допущенные моделью в разделе объяснения, и изменения встречаются также в других сценариях. Единственная отличительная черта сценария наблюдается во втором прогоне: модель допускает ошибку в определении числа существительного: *"Blue leather" should be singular* ("Синяя кожа" должна быть в единственном числе), *"papers" should be pluralized* ("бумаги" следует употреблять во множественном числе.

Аналогично большинству выдач Llama2 в аспекте «категория», модель ошибочно определяет категорию синтаксиса вместо ожидаемой семантики и предоставляет помимо синтаксиса в качестве ответа «синтаксис и семантику» и «синтаксис и морфологию».

### 4.1.8. Label, category, explanation, correction

Сценарий ‘label, category, explanation, correction’, включающий все аспекты, имеет самое низкое значение по «исправлениям» и «объяснениям», показывая среднее значение 6,6379 ± 0,2950 из 10, а также по Accuracy с результатом 0,4263 ± 0,0499.

Так, согласно ручной разметке, исправление Llama2 верно полностью на 1/1 балл в среднем по 3 прогонам лишь в 52,1900 ± 8,3134 % предложений, отмеченных моделью как ошибочные. Разметка объяснений дала, в свою очередь, средний результат 3,0412 ± 0,0794.

В аспекте «объяснение» модель чаще, чем в других сценариях, меняет время глагола и число существительного, обоснуя это семантически или наличием ошибки, которая на самом деле не является ошибкой (подробнее см. п. 5), например:

1. *The word "is" should be changed to "was" because Rutherford died in 1937* — Слово "быть-**prs**" должно быть изменено на "быть-**pst**", так как Резерфорд умер в 1937 году. Видимо, в одном из текстов, на которых была обучена Llama2, содержалась информация о смерти Эрнеста Резерфорда, из-за чего модель меняет время в предложении, несмотря на то, что в предложении это название улицы, а не имя человека.

## 4.2 Результаты Vicuna

### 4.2.1. Label

Аналогично Llama2, сценарий, содержащий только бинарную классификацию, был прогнан c моделью Vicuna с двумя ролями, ассистент и учитель английского языка. Несмотря на одинаковые гиперпараметры модели и запросы, прогоны с ролью ассистента показали более высокие результаты: значения 0,6192 ± 0,0212 и 0,2767 ± 0,0237 по метрикам Accuracy и MCС с ролью ассистента и 0,5338 ± 0,0218 и 0,1389 ± 0,0050 с ролью учителя английского языка соответственно.

Таким образом, следующие сценарии были прогнаны с ролью ассистент, показавшей лучший результат на 3 прогонах. Стоит отметить, что Vicuna, в отличие от Llama2, с теми же промптами дает ответы только в формате «Да» или «Нет» без последующих исправлений и объяснений.

### 4.2.2. Label, category

Среднее значение разметки аспекта «категория» по 3 прогонам составило 1,9171 ± 0,0262 при средних Accuracy 0.7236 ± 0.0058 и MCC 0.4412 ± 0.0577.

Результаты прогонов показывают, что Vicuna, как и LLama2, склонна называть в большинстве случаев синтаксис единственной категорией ошибки. Однако Vicuna предоставляет в качестве ответа на запрос большее разнообразие категорий: в 1 прогоне модель 4 раза называет морфологию, 3 раза семантику, 11 раз синтаксис и семантику, 4 — синтаксис и морфологию и 1 раз семантику и морфологию. В 2 и 3 прогонах помимо синтаксиса по 2 раза называется синтаксис и семантика, 4 раза синтаксис и морфология. В связи с увеличением разнообразия категорий, даваемых модель, становится также больше различных типов ошибок: кроме называния синтаксиса вместо семантики, появляется семантика вместо синтаксиса, морфология вместо синтаксиса и синтаксис вместо семантики и/или морфологии.

### 4.2.3. Label, correction

Среднее значение разметки аспекта «исправление» Vicuna составило 1,7818 ± 0,0678 при средних Accuracy 0,7628 ± 0,0479 и MCC 0,5262 ± 0,0735. Результаты бинарной классификации в этом сценарии являются самыми высокими среди результатов обеих моделей Llama2 и Vicuna, в частности 3 прогон демонстрирует способность модели находить ошибки с точностью 0,8159 Accuracy и 0,6050 MCC.

Среди выдачи Vicuna есть одно интересное исправление (7), когда модель добавляет символы и, по сути, делает предложение семантически корректным на 0,5 баллов, несмотря на отсутствие других изменений.

1. *They have no 'in'* — У них нет ‘in’ (‘в’). Примечательно, что в других двух прогонах модель пытается исправить это предложение, изменив число у предлога с помощью показателя множественного числа -s: They have no ins.

### 4.2.4. Label, explanation

Среднее значение ручной разметки аспекта «объяснение» по 3 прогонам составило 3,5567 ± 0,3904 из 6, что значительно выше результатов разметки сценария ‘label, explanation’ у Llama2.

В исправлениях модели есть ряд уникальных исправлений, которые есть только в этом сценарии и только в первом прогоне Vicuna. Так, модель ошибочно предлагает в 4 примерах заменить слова или словосочетания на запятую, в 2 — на точку (7), при этом все изменяемые слова содержат ошибку или входят в конструкцию, которая должна быть переформулирована.

1. *The phrase "to you" should be replaced with "."* — Фразу "тебе" нужно заменить на "."

Кроме того, в первом прогоне модель допускает новый тип ошибки с капитализацией, предлагая убрать заглавную букву в имени собственном:

1. *The name "Betsy" should not be capitalized.* — Имя "Бетси" не должно писаться с заглавной буквы.

### 4.2.5. Label, category, correction

Среднее значение ручной разметки обоих аспектов данного сценария по 3 прогонам составило 3,6393 ± 0,0416 из 4, что выше результатов прогона модели Llama2. Стоит отметить, что ‘label, category, correction’ имеет наибольшее стандартное отклонение в 0,1699 метрики МСС из-за различия автоматической оценки 1 прогона и 2, 3.

Среди изменений в семантике, получивших 0,5 баллов есть интересный пример:

1. *Medea tried to poison her children by giving them the nurse* — Медея попыталась отравить своих детей, дав им няню. Модель меняет оригинальное предложение ‘Medea tried the nurse to poison her children’, воспринимая существительное *the nurse*  ‘няня’ как яд.

Остальные корректировки модели в аспекте «исправление» и называние категорий не имеют особенностей, так как содержат те же ошибки, как в других сценариях.

### 4.2.6. Label, explanation, correction

Среднее значение ручной разметки аспектов «объяснение» и «исправление» в рамках сценария ‘label, explanation, correction’ составило по 3 прогонам 5,0402 ± 0,0475 из 8, что значительно выше значения Llama2.

Характерной особенностью ответов модели с этим сценарием являются ошибки в форме глагола, когда модель называет время или залог неправильно и впоследствии чаще всего исправляет именно его, например:

1. *You start with "expect" which is in the future tense, but then switch to "will rain" which is in the present simple tense* — Вы начинаете со слова "expect-prs", которое стоит в будущем времени, но затем переходите на "will rain-fut", которое находится в настоящем простом времени.

Кроме того, в рамках именно данного сценария модель допускает наибольшее количество ошибок с артиклями. Так, она может предлагать поставить артикль перед существительным с местоимениями *any* ‘любой’ и *some* ‘какой-либо’, перед вопросительным местоимением *which* ‘который’, местоимением или глагольной конструкцией.

### 4.2.7. Label, category, explanation

Среднее значение разметки сценария составило 5,2901 ± 0,1609 при средних Accuracy 0,6894 ± 0,0295 и MCC 0,4288 ± 0,0912.

Среди всех сценариев именно ‘label, category, explanation’ имеет наибольшее количество гиперкоррекции числа существительных и глаголов, а также изменения порядка слов и добавления новых. То есть Vicuna сначала неправильно считывает предложение из запроса, меняя форму слова, и затем дает в объяснении исправление той ошибки, которую сама допустила, например:

1. *The sentence is grammatically incorrect because of the incorrect use of the preposition "to." The correct sentence would be: "Sharon came into the room."* — Предложение грамматически неверно из-за неправильного употребления предлога "к". Правильным предложением было бы: "Шэрон вошла в комнату". При этом в оригинальном предложении "Sharon came the room" предлога to ‘к’ не было.

### 4.2.8. Label, category, explanation, correction

Сценарий ‘label, category, explanation, correction’, который показал среднее значение 5,4925 ± 0,3966 из 10, отличается от других сценариев тем, что большинство ответов Vicuna построено по шаблону *X should be Y* ‘X надо поменять на Y’, приводящих к снижению оценок ручной разметки по критериям содержания. Так, самое высокое значение по этим критериям достигается в 3 прогоне и составляет всего 0,0559/1 для «объяснение часть1» и 0,0489/1 для «объяснение часть 2». При этом существенного ухудшения ответов по аспектам «категория» и «исправление» модель не показывает.

# 5. Сравнение моделей, обсуждение результатов

Результаты автоматической оценки и ручной разметки показывают, что модели предпочитают давать ответ, даже если он недостаточно качественный, чем не давать его вообще: и Llama2, и Vicuna имеют всего несколько примеров с отсутствием ответа на аспекты «исправление» и «категория» в части критерия «структура». Это касается не только предложений со сложными языковыми феноменами или плохо сформулированной семантикой, но и ситуаций с изменением категорий частей речи, в которых изначально не было нарушения языковых норм, что коррелирует с выводами [Ortega-Martín 2023] при анализе работы ChatGPT с лингвистической неоднозначностью.

Под влиянием больших объемов данных, на которых были обучены модели, Llama2 и Vicuna часто называют в качестве отмечаемых ошибок «гиперкоррекции» — слова или целые клаузы с изменением семантики или грамматических категорий. Нами было выделено 6 основных типов «гиперкоррекции»: числа существительного (13), числа глагола, вида и времени глагола, артикля (14), части речи и наличия предлога, наиболее частыми из которых являются «вид и время глагола» и «артикль» Тип «наличие предлога» (15) встречается только у Vicuna 5 раз во всех сценариях суммарно.

1. *You should use a plural form of "packages"* — Вы должны использовать форму множественного числа слова "packages-pl", хотя слово уже стоит в форме множественного числа, на что эксплицитно указывает показатель -s.
2. *The use of the indefinite article "a" before "sheep" is incorrect. It should be "sheep" without the article.* — Использование неопределенного артикля "а" перед словом "sheep" неверно. Должно быть "овца" без артикля., хотя в оригинальном предложении датасета ‘There is believed to be sheep in the park’ артикля и не было.
3. *The preposition "with" should be used to show what the cart was filled with.* — Чтобы показать, чем была наполнена тележка, стоит использовать предлог *with* ‘c’, хотя называемый предлог уже использован: ‘The farmer dumped the cart with apples’.

Другими типами ошибок, встречающимися в основном в ответах аспекта «объяснение» Vicuna, являются капитализация имен нарицательных (встречается у Llama2 только 1 раз), написание которых модель предлагает изменить, убрав заглавную букву, при том, что эти имена не написаны с заглавной буквы, а также ошибки в виде и времени глагола: *"remembered" should be in its past form* ‘"помнить-pst" должно стоять в форме прошедшего времени’ (также повсеместно встречаются у Llama2, но в три раза реже).

В свою очередь, , в ответах Llama2 встречаются чаще капитализации других типов (см. примеры (4) и (5) выше), а также ошибки в согласовании, которые стоит обсудить отдельно, так как они являются областью, в которой обе модели в разделах «объяснение» допускают наибольшее количество ошибок, предоставляя информацию, противоречащую информацию из другой выдачи.

Согласование числа/лица глагола с подлежащим — самая частая ошибка, которую обе модели Llama2 (16) и Vicuna (17) допускают в около 150 объяснениях предложений, в которых отмечают нарушение языковых норм. Ошибка в согласовании этого типа заключается в том, что модель неправильно называет число глагола, стоящего в форме, в которой нет эксплицитного указания на число глагола и/или субъекта. Около 60 процентов форм глагола в этом случае — глагол прошедшего времени с показателем -ed, однако подобные ошибки встречаются также с неправильными глаголами (18) и глаголами в настоящем времени при субъекте множественного числа.

1. *"A job" is singular, but "offered" is a plural verb form. It should be changed to "offers"* — "Работа" стоит в единственном числе, но "предложить-pst" является формой множественного числа. Стоит поменять глагол на "предложить-prs-sg"
2. *The mistake is that "washed" should agree with the subject "he," so it should be in third person singular form, which is "washes."* — Ошибка в том, что глагол "мыть-pst" должен согласовываться с субъектом "он", то есть стоять в форме 3 лица единственного числа "мыть-prs-sg"
3. *It should be either "I would like to can swim" (singular) or "I would like to could swim" (plural)* — Правильным будет сказать "I would like to can swim" (единственное число) или "I would like to could swim" (множественное число), где в качестве глагола множественного числа модель использует форму прошедшего времени *could* глагола *can* ‘уметь’.

Часто при наличии ошибки в согласовании числа/лица глагола с подлежащим модель не только указывает на несоответствие категории, но и предлагает изменения времени в предложении, как в примерах (16) и (17). В сценарии ‘label, category, explanation, correction’ Vicuna дает объяснение принципа, на который она опирается в объяснениях:

1. *The subject is singular, but the verb is in the past tense form, which usually indicates a plural subject.* — Подлежащее единственного числа, но глагол стоит в форме прошедшего времени, которое обычно указывает на предмет во множественном числе.

Это могло бы объяснять часть ошибок в согласовании по числу глагола и подлежащего, однако Vicuna также предоставляет объяснение, в котором заменяет форму глагола единственного числа настоящего времени, считая ее множественной, видимо, из-за показателя -s:

1. *The second error is that "drives" should be changed to "drive" as the subject, Tom, is singular and requires a singular verb form.* — Вторая ошибка заключается в том, что "водить-prs-sg" должно быть "водить", так как подлежащее, Том, в единственном числе и требует глагол в единственной числе. Такое объяснение не единственное, оно противоречит другим объяснениям, даваемым моделями.

Помимо ошибки с согласованием по числу/лицу глагола и подлежащего модели часто допускают ошибку с согласованием по числу двух несвязанных слов, часто находящихся в разных клаузах:

1. *student should be in the plural form "students" to agree with the plural noun "socks"* — "студент-sg" должно быть "студент-pl", чтобы согласовываться с существительным множественного числа "носок-pl", хотя в оригинальном предложении данные существительные находятся в двух разных клаузах предложения при ССП.
2. *"Which king" is a singular noun and should be followed by a singular verb, but "did you ask" is a plural verb form* — "Который король" является существительным единственного числа и должно использоваться с глаголом единственного числа, но "ты спросил-pst" множественного числа. Здесь модель пытается согласовать подлежащее из одной клаузы с глаголом другой.

В объяснениях моделей также присутствуют нелогичные переходы от называния категорий ошибок в одних словах к изменениям в других, например:

1. *"Which king" is singular, but "did you ask" is plural. To fix this, change "you" to "your."* — "Который король" является существительным единственного числа, но "ты спросил-pst" множественного числа. Чтобы исправить это, поменяйте "ты" на "твой".

Кроме того, среди частых ошибок в согласовании есть примеры, когда модель противоречит информации, которую сама пишет в той же выдаче:

1. *The phrase "my dad lives" is in the present tense, while "I live" is in the present tense as well. However, the verb tenses do not match in both clauses of the sentence* — Фраза "мой папа живет" в настоящем времени, но "я живу" также в настоящем времени. Однако время в двух клаузах предложения не согласуется.

Типы ошибок, встречающихся единично в конкретных прогонах, а также частые изменения слов, в которых не содержится ошибка, можно увидеть в приложениях к работе в Таблице 2 и Таблице 3 соответственно.

Стоит также упомянуть о влиянии конкретных текстов, на которых были обучены модели. Так, в (25) видим, что модель знает миф, в котором есть герои с теми же именами, которые есть в предложении датасета, и поэтому осуществляет изменения в предложении. В (26) примере модель учитывает информацию из о родине человека, чтобы поменять правописание.

1. *"the nurse" is not specific enough and can be replaced with a more precise term such as "Geraldine", who is the name of the character in the story* — "няня" не достаточно конкретно и может быть заменено на более конкретный термин "Джеральдин", имя героини истории
2. *"flavored candy bars" should be "flavoured candy bars" as "Heidi" is from Switzerland and "flavoured" is the correct spelling* — "flavored candy bars" стоит писать как "flavoured candy bars", так как Хеиди из Швейцарии и "flavoured" корректное правописание.

Помимо этого в ответах моделей нередко встречается «мисридинг», при котором модель допускает ошибки только вследствие неправильного считывания предложения, например:

1. *The grammar mistake is that the sentence is missing an article before "panther." It should be "the panther" to specify which panther is being referred to* — Грамматическая ошибка заключается в том, что в предложении отсутствует артикль перед словом "пантера". Правильно будет "the panther" с определенным артиклем "the", чтобы указать, о какой пантере идет речь. Хотя оригинальное предложение звучит как ‘the panther's the coat is dark black’ и содержит два артикля, употребление одного из которых как раз и является ошибкой.

Говоря о баллах, поставленных в процессе разметки ответам моделей, можно выделить два типа сценариев: в которых одна из моделей показала более высокий результат по сравнению с другой и в которых обе модели показатели приблизительно равные результаты в среднем на трех прогонах. Так, Llama2 показала лучшие результаты в сценариях ‘label, correction’ и ‘label, category, explanation, correction’, Vicuna — в сценариях ‘label, explanation’, ‘label, category, correction’ и ‘label, explanation, correction’. Результаты в сценариях ‘label, category’ и ‘label, category, explanation’ с учетом стандартного отклонения были приблизительно равны. При этом результаты автоматической оценки Vicuna были выше по обеим метрикам Accuracy и MCC во всех сценариях.

# 6. Заключение

Исследование показало, что несмотря на то, что большие языковые модели (LLM) могут генерировать грамматически корректные тексты, их способности выносить суждения о лингвистической приемлемости и исправлять нарушения языковых норм в текстах, подаваемых в запросе, остаются ограниченными. Модели обладают некоторыми знаниями о грамматике, однако часто допускают ошибки при классификации частей речи и определении присущих им категорий. В частности, LLM путаются в согласовании подлежащего и сказуемого, а также в согласовании несвязанных между собой членов сложносочиненных (ССП) и сложноподчиненных предложений (СПП), что негативно сказывается на их общей производительности. Одной из ключевых проблем также является мисридинг и галлюцинации, при которых модели интерпретируют текст некорректно или добавляют несуществующую информацию. Эта «гиперкоррекция» происходит под влиянием больших объемов данных, на которых были обучены модели.

Кроме того, сравнительный анализ моделей Llama2 и Vicuna и показал, что модель Vicuna демонстрирует лучшие результаты как в задачах бинарной классификации, так и в предоставлении объяснений, за исключением сценария ‘Label, category, explanation, correction’. В этом сценарии объяснения модели Vicuna в формате *‘X’ should be ‘Y’* приводят к снижению оценок при ручной разметке, что связано с тем, что такие объяснения часто недостаточно точны или полезны для корректного исправления ошибок.

Таким образом, несмотря на значительные достижения в области генерации текста, большие языковые модели все еще сталкиваются с множеством проблем при предоставлении суждений о лингвистической приемлемости, распознавании и исправлении грамматических ошибок. Модели Llama2 и Vicuna демонстрируют различную степень успеха в решении этих задач, но все еще требуют дальнейшего совершенствования.

В качестве перспектив исследования можно дополнительно изучить следующие аспекты:

* Суждения моделей о грамматичности предложений, имеющих в разметке аннотаторами датасета CoLA знак вопроса, и их предпочтения в ответах относительно более сложных и неоднозначных языковых феноменов;
* Обоснование моделями наличия ошибок в тех предложениях, где, согласно используемому набору данных, нет нарушений языковых норм.

# 7. Литература

Орешков 2023 — В. Орешков. Метрики качества моделей бинарной классификации. [Электронный ресурс]. URL: <https://loginom.ru/blog/classification-quality> (Дата обращения: 25.05.2024)

Студеникина 2022 — К. А. Студеникина. Оценка языковой способности нейронных моделей на материале предикативного согласования в русском языке (электронный документ). *Труды ИСП РАН*, 34 (6), 2022. С. 179-184.

Хомский 1972 — Н. Хомский. Аспекты теории синтаксиса. М.: Издательство Московского университета, 1972.

Dentella 2023 — V. Dentella et al. Systematic testing of three Language Models reveals low language accuracy, absence of response stability, and a yes-response bias. *PNAS*, 120 (51), 2023.

Mikhailov 2022 — V. Mikhailov et al. RuCoLA: Russian Corpus of Linguistic Acceptability. *In Proc. of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, P. 5207-5227.

Ortega-Martín 2023 — M. Ortega-Martín et al. Linguistic ambiguity analysis in ChatGPT // *arXiv preprint arXiv:2302.06426*, 2023.

Sun 2023 — S. Sun et al. Battle of the Large Language Models: Dolly vs LLaMA vs Vicuna vs Guanaco vs Bard vs ChatGPT - A Text-to-SQL Parsing Comparison // *arXiv preprint arxiv-2310.10190*, 2023.

Touvron 2023 — H. Touvron et al. Llama 2: Open Foundation and Fine-Tuned Chat Models // *arXiv preprint arXiv.2307.09288*, 2023.

Warstadt 2019 — A. Warstadt et al. Neural Network Acceptability Judgments // *Transactions of the Association for Computational Linguistics.* 7. 10.1162/tacl\_a\_00290, 2019. P. 625-641.

Zhong 2023 — Q. Zhong et al. Can ChatGPT Understand Too? A Comparative Study on ChatGPT and Fine-tuned BERT // *arXiv preprint arXiv:2302.10198*, 2023.

URL: <https://www.promptingguide.ai/ru/techniques/zeroshot> (Дата обращения: 25.05.2024).

# 8. Приложения

Таблица 2. Основные изменения, которые модели считают за ошибки в предложениях

|  | Llama2 | Vicuna |
| --- | --- | --- |
| меняет число существительного | label, explanation 4 шт  label, category, explanation 1 шт  label, category, correction, explanation 8 шт | label, explanation 2 шт  label, category, correction 1 шт |
| меняет  число местоимения | label, explanation 1 шт  label, category, explanation 1 шт  label, category, correction, explanation 1 шт |  |
| меняет число глагола | label, explanation, correction 1 шт | label, explanation 2 шт |
| меняет время глагола  (из-за ошибки/ по семантике) | label, explanation 7 шт  label, explanation, correction 14 шт  label, category, explanation 32 шт  label, category, correction, explanation 39 шт | label, explanation 10 шт  label, explanation, correction 8 шт  label, category, explanation 11 шт  label, category, correction, explanation 27 шт |
| убирает часть предложения (слово/клаузу) | label, explanation 4 шт  label, explanation, correction 17 шт  label, category, explanation 8 шт  label, category, explanation, correction 8 шт | label, explanation 1 шт  label, explanation, correction 2 шт  label, category, explanation 8 шт |
| меняет слово на слово того же семантического поле/ ту же часть речи | label, explanation 6 шт  label, explanation, correction 1 шт  label, category, explanation 3 шт  label, category, explanation, correction 3 шт | label, explanation 6 шт  label, explanation, correction 3 шт  label, category, explanation 2 шт  label, category, explanation, correction 1 шт |

Таблица 3. Основные ошибки, которые модели допускают внутри объяснения

|  | Llama2 | Vicuna |
| --- | --- | --- |
| согласование глаголов по времени | label, category, correction 1 шт | expl 3 шт  cat expl 1 шт |
| согласование числа/лица глагола с субъектом | expl 2 шт  expl corr 10 шт  expl cat 47 шт  cat corr 94 шт | expl 19 шт  corr expl 25 шт  cat expl 57 шт  cat corr 45 шт |
| согласование чисел/лиц несвязанных слов | expl 3 шт  expl corr 11 шт  expl cat 9 шт  cat corr 32 шт | expl 7 шт  corr expl 4 шт  cat expl 5 шт  cat corr 3 шт |
| часть речи | exp 2 шт  expl corr 1 шт  expl cat 3 шт  cat corr 6 шт | expl 1 шт  corr expl 2 шт  cat expl 1 шт  cat corr 8 шт |
| форма глагола | expl 9 шт  expl corr 4 шт  expl cat 4 шт  cat corr 9 шт | expl 8 шт  corr expl 29 шт  cat expl 15 шт  cat corr 17 шт |
| капитализация м-Б имя собственное | expl 2 шт  expl corr 5 шт  expl cat 2 шт  cat corr 1 шт |  |
| капитализация б-Б имя собственное | expl corr 3 шт  expl cat 2 шт |  |
| капитализация б-м имя собственное |  | expl 1 шт шт |
| капитализация м-м | expl 1 шт |  |
| капитализация м-Б в начале предложения | expl cat 2 шт  cat corr 1 шт |  |
| капитализация б-Б в начале предложения | cat corr 1 шт | corr expl 1 шт |
| капитализация б-м не имя собственное | expl corr 1 шт | expl 4 шт  cat expl 1 шт |
| гиперкоррекция число глагола | expl 1 шт  expl cat 3 шт  cat corr 1 шт  cat corr 6 шт | corr expl 1 шт  cat expl 2 шт  cat corr 1 шт |
| гиперкоррекция глагола | expl corr 3 шт  expl cat 2 шт  cat corr 2 шт | expl 1 шт  corr expl 1 шт  cat expl 5 шт  cat corr 17 шт |
| гиперкоррекция артикль | expl 3 шт  expl corr 2 шт  expl cat 2 шт  cat corr 4 шт | corr expl 4 шт  cat expl 1 шт  cat corr 11 шт |
| гиперкоррекция часть речи | expl 1 шт  expl corr 1 шт | corr expl 1 шт |
| гиперкоррекция число сущ | expl cat 2 шт | expl 1 шт  cat expl 3 шт  cat corr 1 шт |
| гиперкоррекция предлог |  | expl 1 шт  corr expl 2 шт  cat expl 1 шт  cat corr 1 шт |
| гиперкоррекция др |  | порядок слов cat expl шт  добав слова cat expl 1 шт  добав слова cat corr 1 шт  ‘s cat corr 1 шт |
| добавление лица у глагола | expl 2 шт  expl corr 5 шт  expl cat 3 шт  cat corr 2 шт | expl 6 шт  cat corr 1 шт |
| добавление времени у сущ | expl corr 3 шт | expl 1 шт  cat expl 1 шт  cat corr 1 шт |
| число | expl corr 4 шт  expl cat 5 шт  cat corr 8 шт | expl 4 шт  corr expl 6 шт  cat expl 13 шт  cat corr 12 шт |
| подлежащее | expl 2 шт  expl cat 3 шт | expl 2 шт  corr expl 2 шт  cat corr 1 шт |
| артикль перед глаголом | exp 1 шт |  |
| артикль перед сущ any/some/many/ | expl corr 2 шт  expl cat 1 шт | cat corr 1 шт |
| артикль перед which | expl cat 2 шт | corr expl 1 шт  cat corr 2 шт |
| артикль перед именем собственным | cat corr 1 шт | corr expl 2 шт  cat corr 1 шт |
| артикль убирает | cat corr 1 шт | cat corr 2 шт |
| артикль другое | ‘s cat corr 1 шт  no in cat corr 1 шт | you corr expl 1 шт  by gas corr expl 1 шт  was blond cat exp 1 шт |
| замена знаком препинания |  | expl 6 шт |