

# [По]пытки в GEE (Grammar Error Explanation)

Маша Пышечкина  
НИУ ВШЭ-2024

# Идея из...

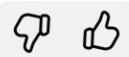


# В русском падежи страдают — и никто не объясняет

— Я иду из школа --> проверь предложение и объясни ошибку

25 июня в 21:16

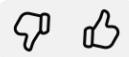
— В этом предложении допущена ошибка в слове «школа». Согласно правилам русского языка, после шипящих букв (ж, ш, ч, щ, ц) в суффиксах и окончаниях пишется буква «ё», если необходимо предупредить искажение произношения. В данном случае написание «школка» может привести к неправильному ударению. Правильный вариант — «из школы».



— Я иду по улица --> проверь предложение и объясни ошибку

25 июня в 21:16

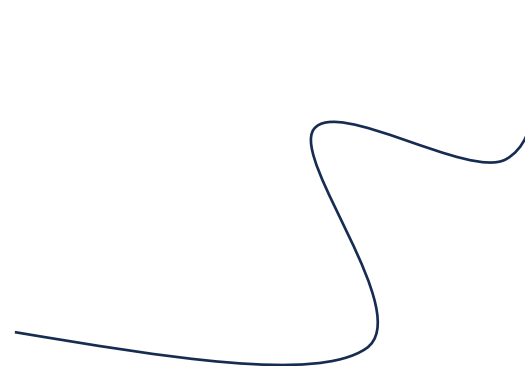
— В этом предложении допущена ошибка в слове «улица». Согласно правилам русского языка, в существительных женского рода с окончанием на «-а» в форме множественного числа буквы «а» меняется на «ы». Поэтому правильный вариант — «по улицам».



Ещё раз



**GIGA  
CHAT**



# В русском падежи страдают – и никто не объясняет

О чем думает студента ? -- > проверь предложение и объясни ошибку

В предложении “О чем думает студента?” допущена ошибка в слове “студента”. Правильно будет “студентка”, поскольку существительное “студентка” женского рода.

Знаете , какую мой обычный день ? ---> проверь предложение и объясни ошибку

В предложении «Знаете, какой мой обычный день?» допущена ошибка в слове «знаете». Слово «знаете» должно стоять в конце предложения, так как это глагол, который требует после себя вопросительного знака.

Найти в Яндексе

Источник



Привет, я Алиса!



# Гипотезы

**User input:** Ich möchte **machen** **ein** Termin.  
**Corrected:** Ich möchte einen Termin machen.\*

Step 1: Extract atomic edits  
via fine-tuned LLM

relocate: **machen**  
replace: **ein** → einen

Step 2: Explain each error  
by prompting LLM

**Error type:** word order

**Error explanation:** The word “machen” is relocated because the infinite verb should be at the end of the sentence when used with a modal verb.

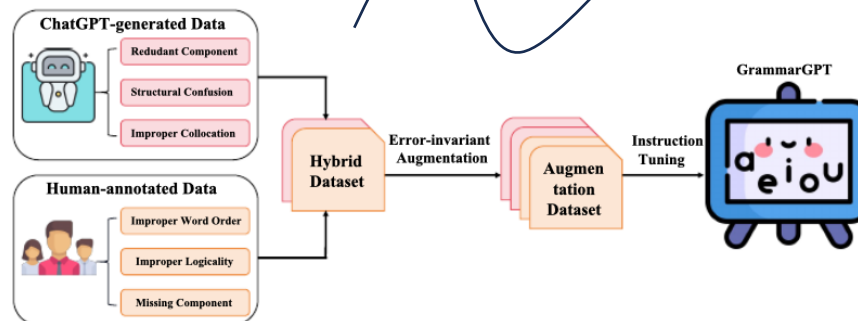
**Error type:** gender and case

**Error explanation:** The word “ein” is replaced with “einen” because it should agree with the gender and case of the noun “Termin”, which is masculine and accusative.

\* English translation: I would like to make an appointment.

Song, Y., Krishna, K., Bhatt, R., Gimpel, K., & Iyyer, M. (2023). **Gee! grammar error explanation with large language models.**

Yaxin Fan, Feng Jiang, Peifeng Li, Haizhou Li (2023). **GrammarGPT: Exploring Open-Source LLMs for Native Chinese Grammatical Error Correction with Supervised Fine-Tuning**



# Instruction fine-tuning rugpt3small

Instruction	Suffix	Question	Response
Исправь ошибку и объясни, что было неправильно	Исправленный текст	Я не люблю рыба	Ошибка: рыбу. Нужно использовать: винительный падеж. Правильно: рыбу

# Гибридный датасет

Корпус	Предложения
RLC	2760
Синтетические данные	1216
*RuAdapt	2534
<b>Всего</b>	<b>6510</b>

\*RuAdapt: A Parallel Russian-Simple Russian Dataset by Anna Dmitrieva

# Гибридный датасет

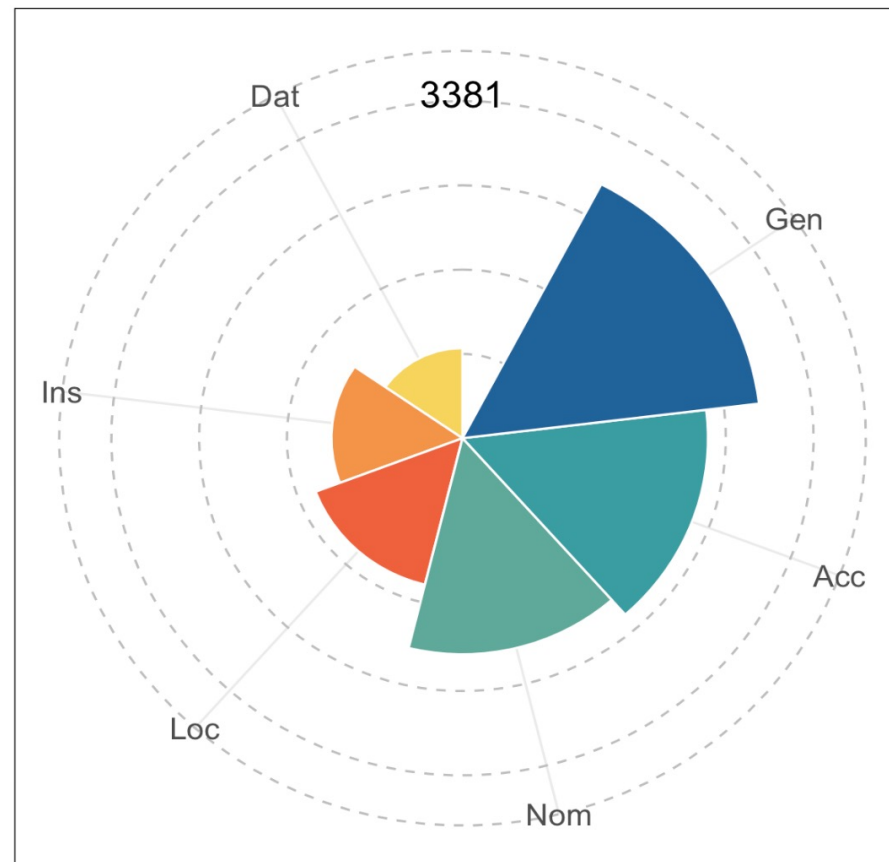
Пролетело в моей  
голове.

Предложение правильное.

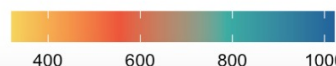
В москву я очень  
люблю ходить  
пешком.

Ошибка: Москву. Нужно  
использовать: творительный  
падеж. Правильно: Москве

Распределение ошибок по падежам



Количество предложений





# Данные из RLC (#Gov)

- Это не друг, по-моему, даже хотя так названно на {{фейсбук }} или в разговорах.
- У нас, например, в Голливуде столько известных» людей, которые говорят не просто {{мат}}, а вообще на каком-то непонятном английском.
- Как ты были, что ты делали и что {{ты }} интересуется, сейчас.

# Синтетические данные

- Тексты из учебника по РКИ «Привет, Россия!» (А1, В1) → метод Parse.inflect() в PyMorphy:

```
# Функция для замены слова на случайную форму в неправильном падеже
def introduce_error(sentence):
    words = nltk.word_tokenize(sentence, language="russian")
    tagged = nltk.pos_tag(words, lang='rus')

    for i, (word, pos) in enumerate(tagged):
        parsed_word = morph.parse(word)
        if parsed_word and parsed_word[0].tag.POS in {'NOUN', 'ADJF', 'PRON', 'NUMR'}:
            try:
                cases = {'nomn', 'gent', 'datv', 'accs', 'ablt', 'loct'}
                current_case = parsed_word[0].tag.case
                if current_case in cases:
                    cases.remove(current_case)
                    new_case = cases.pop()
                    new_word = parsed_word[0].inflect({new_case}).word
                    words[i] = new_word
                    original_word = word
                    changed_word = new_word
                    erroneous_sentence = ' '.join(words)
                    return sentence, erroneous_sentence, original_word, changed_word
            except Exception as e:
                continue
```

# Синтетические данные

1. Однажды в одной из школ города Таллина приехал иностранный профессор , который знал много разных языков.
2. Но потом решил , что надо менять жизнью.
3. В барселону живёт её семья.

64	Вы очень с	Вы очень с	серьёзный	серьёзного				
65	Вы делаете	Вы делаете	карьеру	карьерой				
66	Вы всегда л	Вы всегда л	лидер	лидера				
67	Но вы очен	Но вы очен	мало	мало				
68	Ваш любим	вашим люб	Ваш	вашим				
69	Вы спокой	Вы спокой	спокойный	спокойного				
70	Вы любите	Вы любите	литературу	литературой				
71	Мой брат и	мой брат и	Мой	мой				

# Данные из RuAdapt: A Parallel Russian-Simple Russian Dataset

- Однажды Олечку пригласили на вечер.
- Граф очень любил Сергея, а Сергей графа не любил.
- Общее собрание будет в четыре.
- ~~— Ох, нет, нет!~~
- ~~— Боже, что делать?~~

7731	И тогда Петю освободят, Плевну возьмут, война закончится
7732	— сказал граф. — Бить жену?!
7733	Он чувствовал, что лошадь шла из последних сил; она с трудом дышала.
7734	Feci, quod potui, faciant meliora potentes <sup>26</sup> .
7735	Я весь в вашей власти, но разрешите мне быть тут... Алексей Александрович, увидев слёзы Вронского и не дослушав, быстро пошёл к двери.
7736	— Ну, пойдём в кабинет, — сказал Степан Аркадьич.
7737	Поругал немного за то, что тот не отвечал на его письма и не приезжал к нему.
7738	Чебутыкин (читает газету и тихо напевает).

# Автоматическая разметка падежей

A2		Вечером я читаю журналу , смотрю телевизор и думаю ... Фёдор ужинает , а потом спит .									
	A	B	C	D	E	F	G	H	I	J	K
	Оригинальное предложение	Исправленное предложение	Ошибка	Исправление	Тип ошибки						
1											
2	Вечером я читаю ж	Вечером я читаю ж	журналу	журналы	Асс						

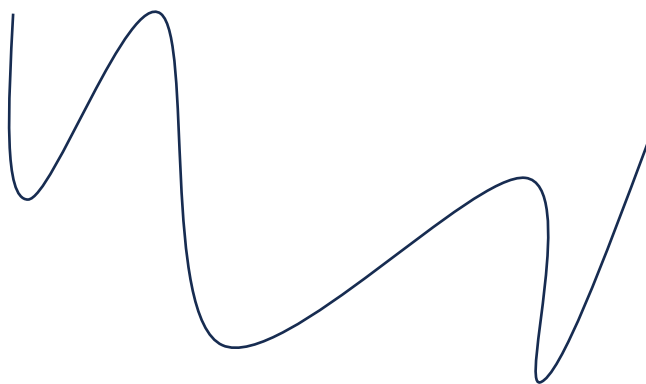
ожидаНИЕ

Противник	*Оппонент	кризис	кризиса	Gen	
Я считаю, ч	Я считаю, ч	другу на др	друг на др	No match found	
Статья буде	Статья буде	желание	желании	Loc	
Некогда не	*Никогда*	работы	работ	Gen	
русские же	русские же	равноправ	равноправ	Gen	
В {{это}} пе	В *эту* *за	это	эту	Асс	
Свое изуче	Свое изуче	интерес	интерес к	No match found	
Потом я на	Потом я на	книга	книгу,	No match found	
Тоже были	*Также* бы	длинные	длинными	No match found	
Этот курс п	Этот курс п	нам	нас	Асс	
В {{заключе	В *заключе	укреплени	укреплени	Асс	
И самое гл	И самое гл	политике	политику,	No match found	
Необходим	Необходим	последних	последние	Асс	
Какими тел	Какими тел	регулятор	регулятора	Loc	
Какими тел	Какими тел	влияние	влиянии	Loc	
Я уже в уни	Я уже в уни	неделя	неделю	Асс	

реальность

# Prompt – Response

```
def create_instructional_example(row):  
    if pd.isna(row['Ошибка']):  
        input_sentence = row['Оригинальное предложение']  
        error_word = row['Ошибка']  
        correction = row['Исправление']  
        case = row['Тип ошибки']  
        output_message = f"Ошибка: {error_word}. Нужно использовать: {case} падеж. Правильно: {correction}"  
        return input_sentence, output_message  
    else:  
        input_sentence = row['Оригинальное предложение']  
        output_message = "Предложение правильное."  
        return input_sentence, output_message
```



	Input	Output
1		
2	Она тотчас	Предложение правильное.
3	Варя знала	Предложение правильное.
4	Разве мож	Предложение правильное.
5	И какая мн	Предложение правильное.
6	Однажды е	Ошибка: одной. Нужно использовать: винительный падеж. Правильно: одну
7	В москву я	Ошибка: Москву. Нужно использовать: творительный падеж. Правильно: Москве
8	Андрей ма	Предложение правильное.
9	— Ну что ж	Предложение правильное.
10	Они очень	Ошибка: весёлых. Нужно использовать: именительный падеж. Правильно: весёлые
11	Это было	Ошибка: предметы. Нужно использовать: родительный падеж. Правильно: предметов
12	Значит, зае	Предложение правильное.
13	И мы там тс	Предложение правильное.
14	Но Наруто	Ошибка: признанию. Нужно использовать: родительный падеж. Правильно: признания
15	Это не дру	Ошибка: фейсбук. Нужно использовать: предложный падеж. Правильно: фейсбуке
16	— Ты вот ч	Предложение правильное.
17	Адвокат ок	Ошибка: помощи. Нужно использовать: винительный падеж. Правильно: помощь
18	Не мне раз	Предложение правильное.

# Ошибки и странные результаты

```
/Users/mariapyshechkina/pythonProject5/bin/python /Users/mariia
```

```
Исправленное предложение: ### Инструкция:
```

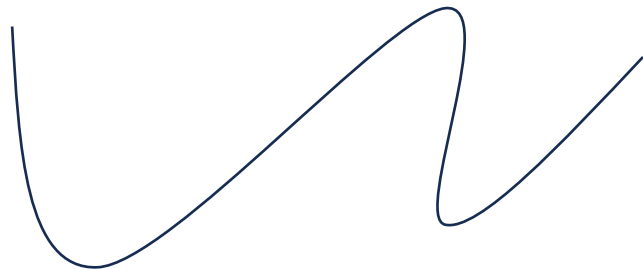
```
Найди ошибку в предложении. Исправь ее. Объясни ошибку
```

```
### Входные данные:
```

```
Я люблю гулять в парк.
```

```
### Ответ:
```

```
Дура
```



Тренировка без  
лейблов

# Ошибки и странные результаты

- “Исправленное предложение: Ошибка:  
папа Нипасибо Бо том Прошу Од Нет и Ничего Правда кон  
Г Интерфак по Зна Точно Изви”



Тренировка с  
лейблами



# Корпус поменьше

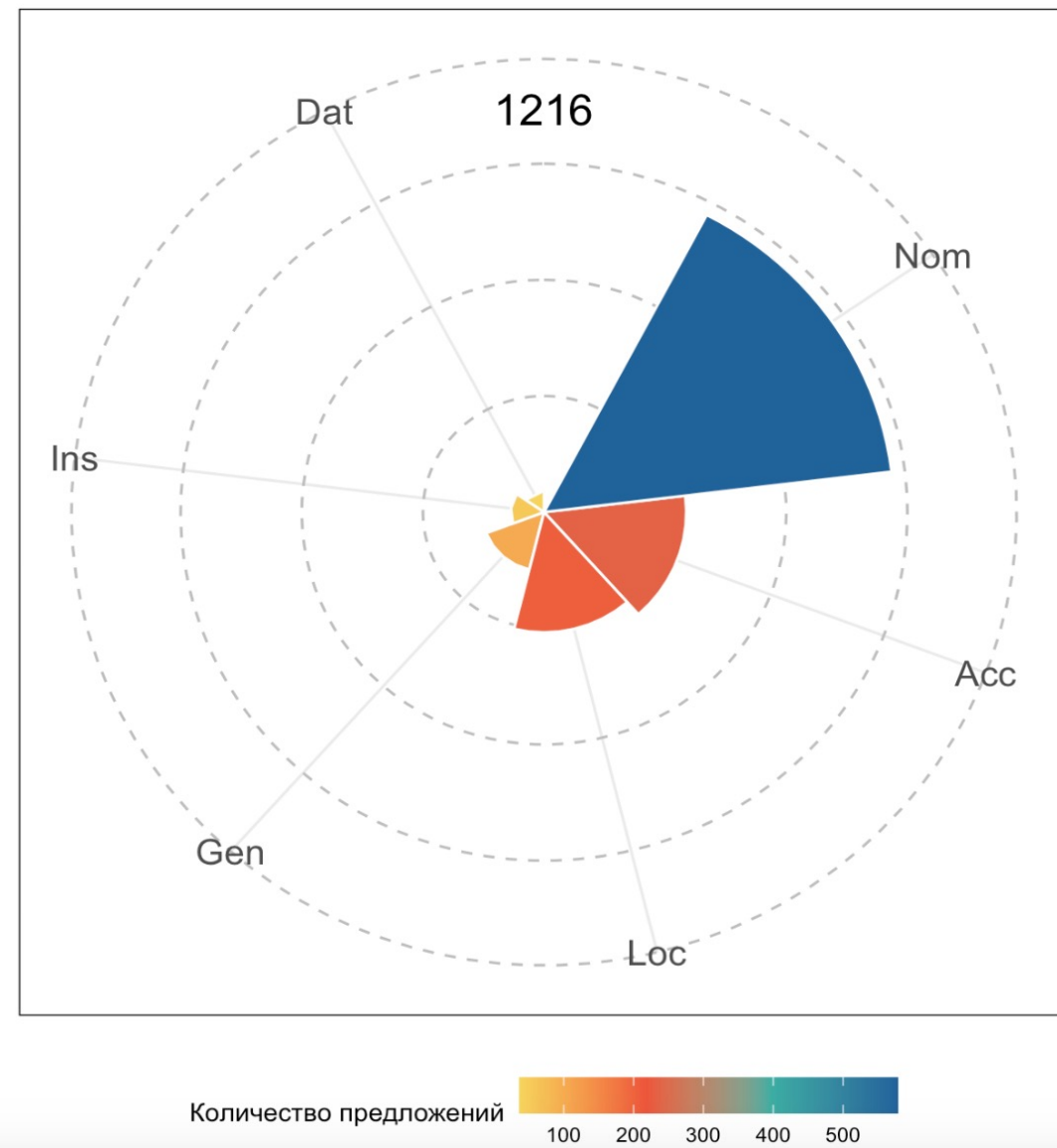
В москву я очень люблю ходить  
пешком .

Ошибка: Москву. Нужно  
использовать: творительный  
падеж. Правильно: Москве

Они очень весёлых и спортивные . падеж. Правильно: весёлые

Ошибка: весёлых. Нужно  
использовать: именительный  
падеж. Правильно: весёлые

## Распределение ошибок по падежам



# Instruction Tuning маленького корпуса

```
54 training_args = TrainingArguments(  
55     output_dir='./mistakes_det',  
56     evaluation_strategy='epoch',  
57     save_strategy='epoch',  
58     save_total_limit=2,  
59     logging_dir='./logs',  
60     num_train_epochs=4,  
61     per_device_train_batch_size=2,  
62     per_device_eval_batch_size=4,  
63     load_best_model_at_end=True,  
64     metric_for_best_model='eval_loss',  
65     logging_steps=100,  
66     eval_steps=100,  
67     learning_rate=5e-5,  
68     weight_decay=0.01,  
69 )
```

# Результаты

### Входные данные:

Я люблю пицца

### Ответ:

### Исправленный текст:

Ошибка: пицца. Нужно использовать: винительный падеж. Правильно: пиццу

Acc

Gen

Результат инференса: ### Инструкция:

Исправь ошибку и объясни, что было неправильно:

### Входные данные:

Я из Россия

### Ответ:

### Исправленный текст:

Ошибка: Россию. Нужно использовать: предложный падеж. Правильно: России

# Результаты

### Входные данные:

Он гулял по моста

### Ответ:

### Исправленный текст:

Ошибка: моста. Нужно использовать: предложный падеж. Правильно: мосту

Dat



Loc



### Входные данные:

Сейчас я живу в Москва

### Ответ:

### Исправленный текст:

Ошибка: Москву. Нужно использовать: предложный падеж. Правильно: Москве

# Результаты

### Входные данные:

Я не люблю пиццу

### Ответ:

### Исправленный текст:

Ошибка: пиццу. Нужно использовать: винительный падеж. Правильно: пиццу

### Входные данные:

Он не есть макароны с сыром

### Ответ:

### Исправленный текст:

Ошибка: макароны. Нужно использовать: винительный падеж. Правильно: макароны

# Что дальше

- ✓ Гибридный датасет
- ✓ Сбалансированность ошибок
- ✓ Другие типы ошибок

Найди ошибку в форме падежа.

### Входные данные:

Нужно обучить на большой корпус

### Ответ:

да

Очень

Спасибо за внимание!😊