Оценка грамматических способностей больших языковых моделей на основе набора данных CoLA

Далия Гимадиева, БКЛ 221
науч. рук. - Сериков Олег Алексеевич, науч. конс. - Артемова Екатерина Леонидовна, Никишина Ирина Александровна

Цель исследования

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

Этапы работы

- разработка сценариев запросов
- подбор подходящих промптов
- прогон датасета CoLA со сценарием label “есть ошибка/нет ошибки” с ролями ассистент и учитель английского языка по 3 раза
- прогон оставшихся сценариев с лучшей ролью по 3 раза
- ручная разметка ответов моделей (4529 шт)
- выделение особенностей, сходств и различий, частых ошибок

Llama2	Vicuna
13b параметров	
мультиязычная модель, основной английский	
	доработанная Llama2

Выводы

- низкая точность распознавания грамматических ошибок (несмотря на хорошую генерацию своего текста”)
- бинарная классификация и объяснения vicuna в целом лучше, кроме сценария label, category, explanation, correction, где объяснения в формате 'X' should be 'Y'
- модели обладают некоторыми знаниями о грамматике, но часто ошибаются с категориями и частями речи, путаются с согласованием при ССП, СПП
- модели допускают мисридинг, галлюцинации, меняют семантику под влиянием текстов, на которых обучены

вместо исправления нужной ошибки модели:
- меняют число сущ, местоимений
- убирают слово/клаузу
- меняют время глагола (из-за ошибки/по семантике) "Ellen said" → "Ellen says" since Ellen is still alive
- меняют слово на слово из того же семант поля/той же часть речи
socks → shoes;
he → they (if Mike's gender identity is non-binary or unknown)
- добавляют слова/клаузы, меняют порядок слов, артикли, сокращения типа didn't

частые ошибки:
- согласование числа глагола с субъектом
"A job" is singular, but "offered" is a plural verb form;
"John's mother" is singular, but the verb "likes" is in the third person plural form
- добавление категории лица глаголу, времени - сущ
to match the past tense of the subject "Sue"
- гиперкоррекция числа, времени, артикля и тп
the verb "are leaking" is plural про “The building's the roof is leaking”
- согласование по числу/лицу несвязанных слов
"Every student" is a singular subject, but the verb "wears" agrees with a plural object "socks"
- капитализация разного типа
"fountain" should be capitalized as it is a proper noun
- время и залог глагола
the verb "whipped" should be in the past tense
- число
plural subject "this set"
~ - характерно более для vicuna,
~ - характерно более для llama2
(ниже эти цвета используются для выделения модели с лучшим результатом)

Sally kissed himself

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."		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."	
1	the pronoun "himself"	1	the subject of the sentence "Sally"
1	the pronoun "himself"	0	ошибка в местоимении "himself"
1	the pronoun "himself" is not appropriate for a female person	1	"kissed" is a third-person singular verb, it does not agree with the first-person pronoun "himself"
1	the pronoun "himself" is not appropriate for a female person	0	неправильно
1	"Sally kissed herself." with a female reflexive pronoun "herself."	0,5	change the wording to "Sally kissed herself."
1	правильно	0	неправильно
6/6		2,5/6	

структура	локализация дана
содержание	локализация верна (форма глагола, предлог, время)
структура	объяснение (часть1) дано (названы категории с ошибкой)
содержание	объяснение (часть1) верно (в категориях действительно ошибки)
структура	объяснение (часть2) дано (названы исправления с категориями)
содержание	объяснение (часть2) верно (исправления корректны)

Все файлы по работе

Результаты разметки ответов моделей (среднее по трем прогонам):

	Llama2 (13b)	Vicuna (13b-v1.5)
label, category	1,9149 ± 0,0095 из 2	1,9171 ± 0.0262 из 2
label, correction	1,8926 ± 0,1112 из 2	1,7818 ± 0,0678 из 2
label, explanation	3,3191 ± 0,0354 из 6	3,5567 ± 0,3904 из 6
label, category, correction	3,5677 ± 0,03956 из 4	3,6809 ± 0,0416 из 4
label, explanation, correction	4,7843 ± 0,1339 из 8	5,0402 ± 0,0475 из 8
label, category, explanation	5, 2407 ± 0,0438 из 8	5,2901 ± 0,1609 из 8
label, category, explanation, correction	6,6379 ± 0,2950 из 10	5,4925 ± 0,3966 из 10

Точность ответа модели по метрикам Accuracy и Matthews correlation coefficient (MCC) (среднее по трем прогонам):

	Llama2 (13b)		Vicuna (13b-v1.5)	
	accuracy	MCC	accuracy	MCC
label (ассистент)	0.6363 ± 0.0278	0.1988 ± 0.0253	0.6192 ± 0.0212	0.2767 ± 0.0237
label (учитель английского)	0.5003 ± 0.0631	0.0684 ± 0.0151	0.5338 ± 0.0218	0.1389 ± 0.0050
label, category	0.4896 ± 0.0242	0.1519 ± 0.0340	0.7236 ± 0.0058	0.4412 ± 0.0577
label, correction	0.6996 ± 0.0039	0.1152 ± 0.0305	0.7628 ± 0.0479	0.5262 ± 0.0735
label, explanation	0.6299 ± 0.0219	0.0451 ± 0.0587	0.6939 ± 0.0208	0.3789 ± 0.1346
label, category, correction	0.5364 ± 0.0578	0.2002 ± 0.0426	0.6787 ± 0.0216	0.4105 ± 0.1699
label, explanation, correction	0.6945 ± 0.0279	0.1940 ± 0.0646	0.6983 ± 0.0206	0.4469 ± 0.0438
label, category, explanation	0.5035 ± 0.0929	0.0892 ± 0.0504	0.6894 ± 0.0295	0.4288 ± 0.0912
label, category, explanation, correction	0.4263 ± 0.0499	0.0816 ± 0.0316	0.6869 ± 0.0181	0.4387 ± 0.0509

Оценка грамматических способностей больших языковых моделей на основе набора данных CoLA

Далия Гимадиева, БКЛ 221
науч. рук. - Сериков Олег Алексеевич, науч. конс. - Артемова Екатерина Леонидовна, Никишина Ирина Александровна

Цель исследования

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

Этапы работы

- разработка сценариев запросов
- подбор подходящих промптов
- прогон датасета CoLA со сценарием label “есть ошибка/нет ошибки” с ролями ассистент и учитель английского языка по 3 раза
- прогон оставшихся сценариев с лучшей ролью по 3 раза
- ручная разметка ответов моделей (4529 шт)
- выделение особенностей, сходств и различий, частых ошибок

Llama2	Vicuna
13b параметров	
мультиязычная модель, основной английский	
	доработанная Llama2

Выводы

- низкая точность распознавания грамматических ошибок (несмотря на хорошую генерацию своего текста”)
- бинарная классификация и объяснения vicuna в целом лучше, кроме сценария label, category, explanation, correction, где объяснения в формате 'X' should be 'Y'
- модели обладают некоторыми знаниями о грамматике, но часто ошибаются с категориями и частями речи, путаются с согласованием при ССП, СПП
- модели допускают мисридинг, галлюцинации, меняют семантику под влиянием текстов, на которых обучены

вместо исправления нужной ошибки модели:
- меняют число сущ, местоимений
- убирают слово/клаузу
- меняют время глагола (из-за ошибки/по семантике)
"Ellen said" → "Ellen says" since Ellen is still alive
- меняют слово на слово из того же семант поля/той же часть речи
socks → shoes;
he → they (if Mike's gender identity is non-binary or unknown)
- добавляют слова/клаузы, меняют порядок слов, артикли, сокращения типа didn't

частые ошибки:
- согласование числа глагола с субъектом
"A job" is singular, but "offered" is a plural verb form;
"John's mother" is singular, but the verb "likes" is in the third person plural form
- добавление категории лица глаголу, времени - сущ
to match the past tense of the subject "Sue"
- гиперкоррекция числа, времени, артикля и тп
the verb "are leaking" is plural про “The building's the roof is leaking”
- согласование по числу/лицу несвязанных слов
"Every student" is a singular subject, but the verb "wears" agrees with a plural object "socks"
- капитализация разного типа
"fountain" should be capitalized as it is a proper noun
- время и залог глагола
the verb "whipped" should be in the past tense
- число
plural subject "this set"
~ - характерно более для vicuna,
~ - характерно более для llama2
(ниже эти цвета используются для выделения модели с лучшим результатом)

Sally kissed himself

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."		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."	
1	the pronoun "himself"	1	the subject of the sentence "Sally"
1	the pronoun "himself"	0	ошибка в местоимении "himself"
1	the pronoun "himself" is not appropriate for a female person	1	"kissed" is a third-person singular verb, it does not agree with the first-person pronoun "himself"
1	the pronoun "himself" is not appropriate for a female person	0	неправильно
1	"Sally kissed herself." with a female reflexive pronoun "herself."	0,5	change the wording to "Sally kissed herself."
1	правильно	0	неправильно
6/6		2,5/6	

структура	локализация дана
содержание	локализация верна (форма глагола, предлог, время)
структура	объяснение (часть1) дано (названы категории с ошибкой)
содержание	объяснение (часть1) верно (в категориях действительно ошибки)
структура	объяснение (часть2) дано (названы исправления с категориями)
содержание	объяснение (часть2) верно (исправления корректны)

Результаты разметки ответов моделей (среднее по трем прогонам):

	Llama2 (13b)	Vicuna (13b-v1.5)
label, category	1,9149 ± 0,0095 из 2	1,9171 ± 0.0262 из 2
label, correction	1,8926 ± 0,1112 из 2	1,7818 ± 0,0678 из 2
label, explanation	3,3191 ± 0,0354 из 6	3,5567 ± 0,3904 из 6
label, category, correction	3,5677 ± 0,03956 из 4	3,6809 ± 0,0416 из 4
label, explanation, correction	4,7843 ± 0,1339 из 8	5,0402 ± 0,0475 из 8
label, category, explanation	5, 2407 ± 0,0438 из 8	5,2901 ± 0,1609 из 8
label, category, explanation, correction	6,6379 ± 0,2950 из 10	5,4925 ± 0,3966 из 10

Точность ответа модели по метрикам Accuracy и Matthews correlation coefficient (MCC) (среднее по трем прогонам):

	Llama2 (13b)		Vicuna (13b-v1.5)	
	accuracy	MCC	accuracy	MCC
label (ассистент)	0.6363 ± 0.0278	0.1988 ± 0.0253	0.6192 ± 0.0212	0.2767 ± 0.0237
label (учитель английского)	0.5003 ± 0.0631	0.0684 ± 0.0151	0.5338 ± 0.0218	0.1389 ± 0.0050
label, category	0.4896 ± 0.0242	0.1519 ± 0.0340	0.7236 ± 0.0058	0.4412 ± 0.0577
label, correction	0.6996 ± 0.0039	0.1152 ± 0.0305	0.7628 ± 0.0479	0.5262 ± 0.0735
label, explanation	0.6299 ± 0.0219	0.0451 ± 0.0587	0.6939 ± 0.0208	0.3789 ± 0.1346
label, category, correction	0.5364 ± 0.0578	0.2002 ± 0.0426	0.6787 ± 0.0216	0.4105 ± 0.1699
label, explanation, correction	0.6945 ± 0.0279	0.1940 ± 0.0646	0.6983 ± 0.0206	0.4469 ± 0.0438
label, category, explanation	0.5035 ± 0.0929	0.0892 ± 0.0504	0.6894 ± 0.0295	0.4288 ± 0.0912
label, category, explanation, correction	0.4263 ± 0.0499	0.0816 ± 0.0316	0.6869 ± 0.0181	0.4387 ± 0.0509

Все файлы по работе