ScoNe: Benchmarking Negation Reasoning in Language Models With Fine-Tuning and In-Context Learning

Елизавета Бесараб, Яна Захарова

Этапы исследования

01

Создание базы ScoNe-NLI

02

Оценка RoBERTa и DeBERTa с файн-тьюнингом

03

Обучение в контексте на ScoNe-NLI

04

Создание базы ScoNe-NLG

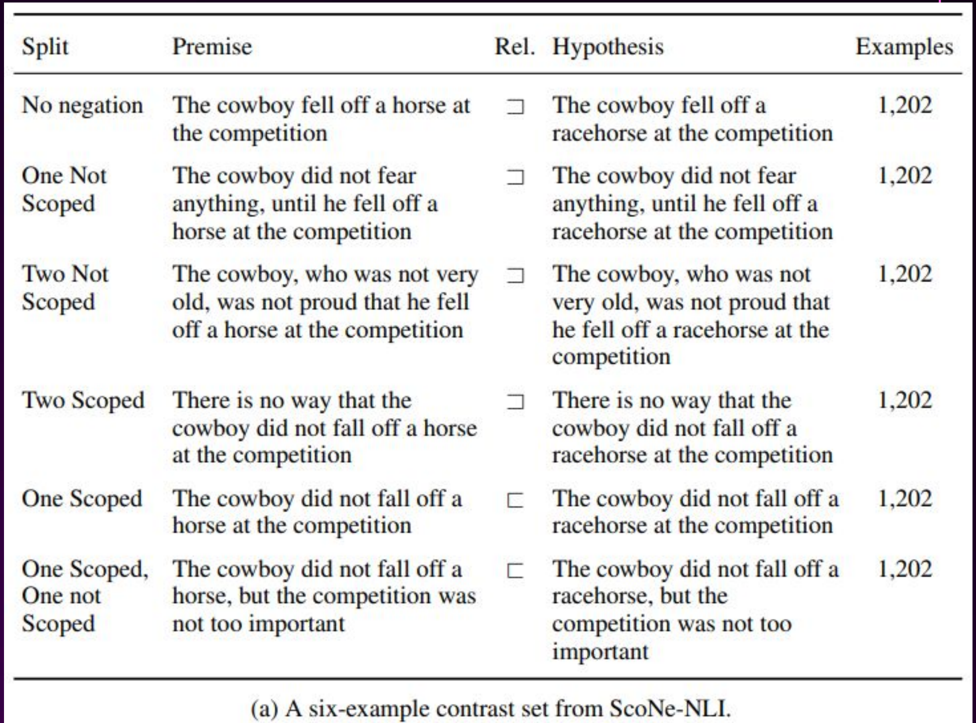
05

Обучение в контексте на ScoNe-NLG

База данных ScoNe-NLI

Scoped **Ne**gation **N**atural **L**anguage **I**nference - расширенная база Monotonicity NLI

- 1202 набора контрастов
- Набор контрастов -
  - 0/1/2 отрицания
  - Наличие/отсутствие отрицания в сфере действия для таргета
  - Метка NLI зависит от наличия/отсутствия следствия

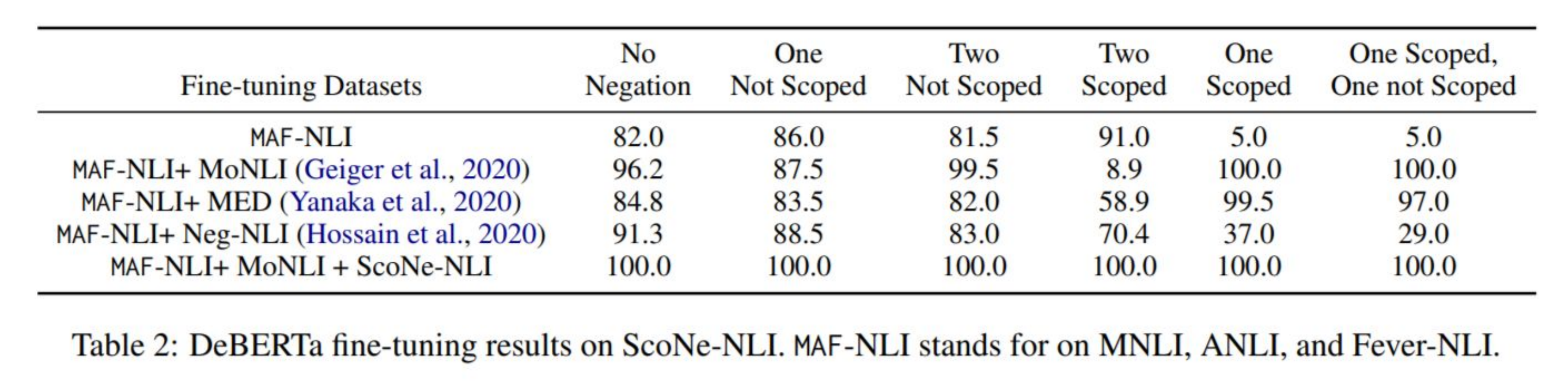


Split	Premise	Rel.	Hypothesis	Examples
No negation	The cowboy fell off a horse at the competition	$\sqsupset$	The cowboy fell off a racehorse at the competition	1,202
One Not Scoped	The cowboy did not fear anything, until he fell off a horse at the competition	$\sqsupset$	The cowboy did not fear anything, until he fell off a racehorse at the competition	1,202
Two Not Scoped	The cowboy, who was not very old, was not proud that he fell off a horse at the competition	$\sqsupset$	The cowboy, who was not very old, was not proud that he fell off a racehorse at the competition	1,202
Two Scoped	There is no way that the cowboy did not fall off a horse at the competition	$\sqsupset$	There is no way that the cowboy did not fall off a racehorse at the competition	1,202
One Scoped	The cowboy did not fall off a horse at the competition	$\sqsubset$	The cowboy did not fall off a racehorse at the competition	1,202
One Scoped, One not Scoped	The cowboy did not fall off a horse, but the competition was not too important	$\sqsubset$	The cowboy did not fall off a racehorse, but the competition was not too important	1,202

(a) A six-example contrast set from ScoNe-NLI.

Fine-Tuning DeBERTa на ScoNe-NLI

Использовались предобученные на MNLI, Fever-NLI и Adversarial-NLI DeBERTa-v3-base и RoBERTa

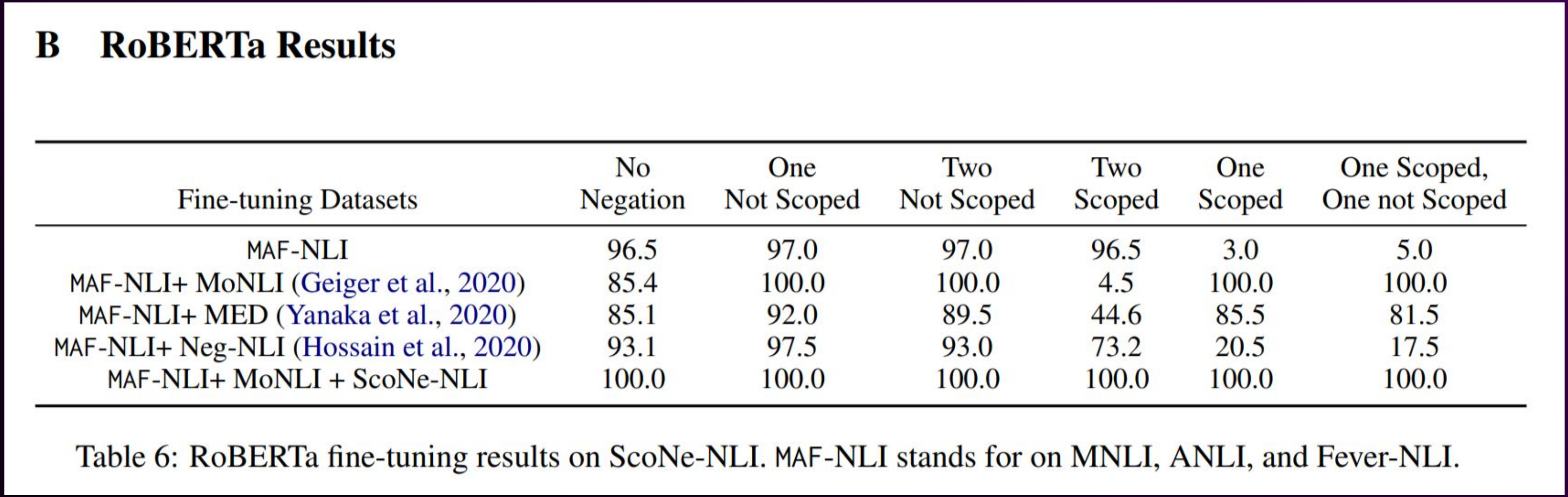


Fine-tuning Datasets	No Negation	One Not Scoped	Two Not Scoped	Two Scoped	One Scoped	One Scoped, One not Scoped
MAF-NLI	82.0	86.0	81.5	91.0	5.0	5.0
MAF-NLI+ MoNLI (Geiger et al., 2020)	96.2	87.5	99.5	8.9	100.0	100.0
MAF-NLI+ MED (Yanaka et al., 2020)	84.8	83.5	82.0	58.9	99.5	97.0
MAF-NLI+ Neg-NLI (Hossain et al., 2020)	91.3	88.5	83.0	70.4	37.0	29.0
MAF-NLI+ MoNLI + ScoNe-NLI	100.0	100.0	100.0	100.0	100.0	100.0

Table 2: DeBERTa fine-tuning results on ScoNe-NLI. MAF-NLI stands for on MNLI, ANLI, and Fever-NLI.

Fine-Tuning RoBERTa на ScoNe-NLI

Использовались предобученные на MNLI, Fever-NLI и Adversarial-NLI DeBERTa-v3-base и RoBERTa



B RoBERTa Results

Fine-tuning Datasets	No Negation	One Not Scoped	Two Not Scoped	Two Scoped	One Scoped	One Scoped, One not Scoped
MAF-NLI	96.5	97.0	97.0	96.5	3.0	5.0
MAF-NLI+ MoNLI (Geiger et al., 2020)	85.4	100.0	100.0	4.5	100.0	100.0
MAF-NLI+ MED (Yanaka et al., 2020)	85.1	92.0	89.5	44.6	85.5	81.5
MAF-NLI+ Neg-NLI (Hossain et al., 2020)	93.1	97.5	93.0	73.2	20.5	17.5
MAF-NLI+ MoNLI + ScoNe-NLI	100.0	100.0	100.0	100.0	100.0	100.0

Table 6: RoBERTa fine-tuning results on ScoNe-NLI. MAF-NLI stands for on MNLI, ANLI, and Fever-NLI.

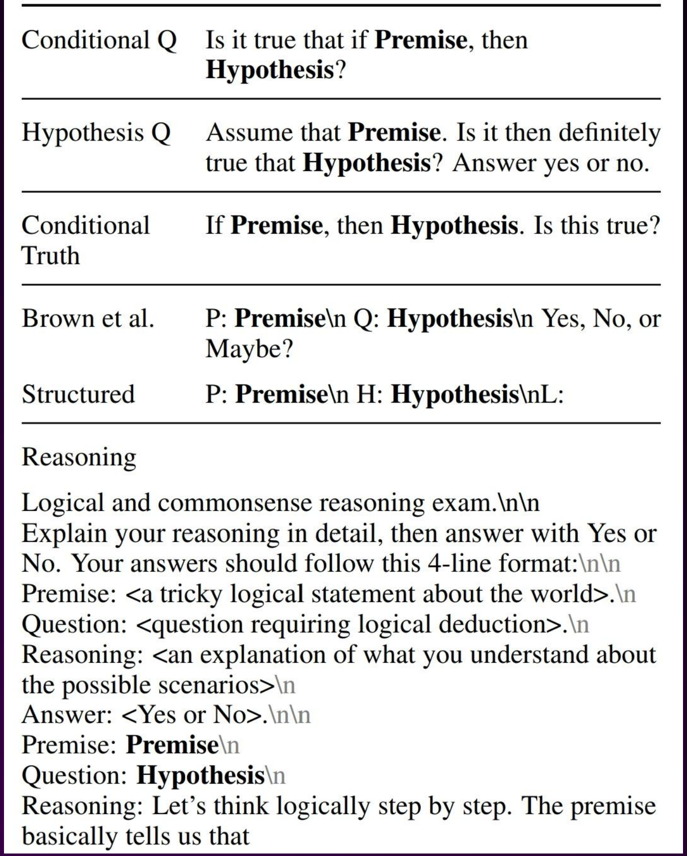
Обучение Instruct-GPT в контексте на ScoNe-NLI

Если ответ содержит "yes" в любом виде, то понимаем как entailment, в противном случае - neutral.

Использовалось шесть типов промптов.

Zero-shot: сразу таргет

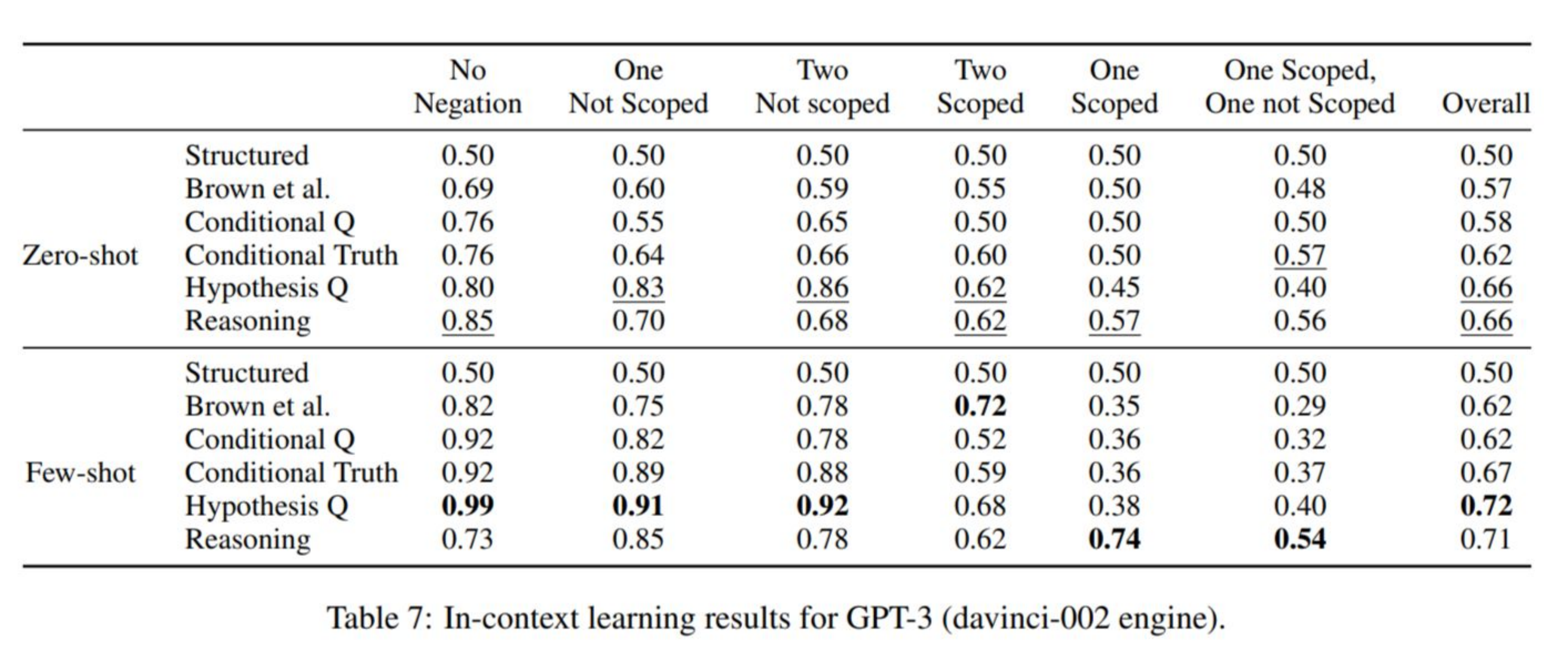
Few-shot: 4 примера задания без оценки ответа + таргет



Conditional Q	Is it true that if **Premise**, then **Hypothesis**?
Hypothesis Q	Assume that **Premise**. Is it then definitely true that **Hypothesis**? Answer yes or no.
Conditional Truth	If **Premise**, then **Hypothesis**. Is this true?
Brown et al.	P: **Premise**\n Q: **Hypothesis**\n Yes, No, or Maybe?
Structured	P: **Premise**\n H: **Hypothesis**\nL:

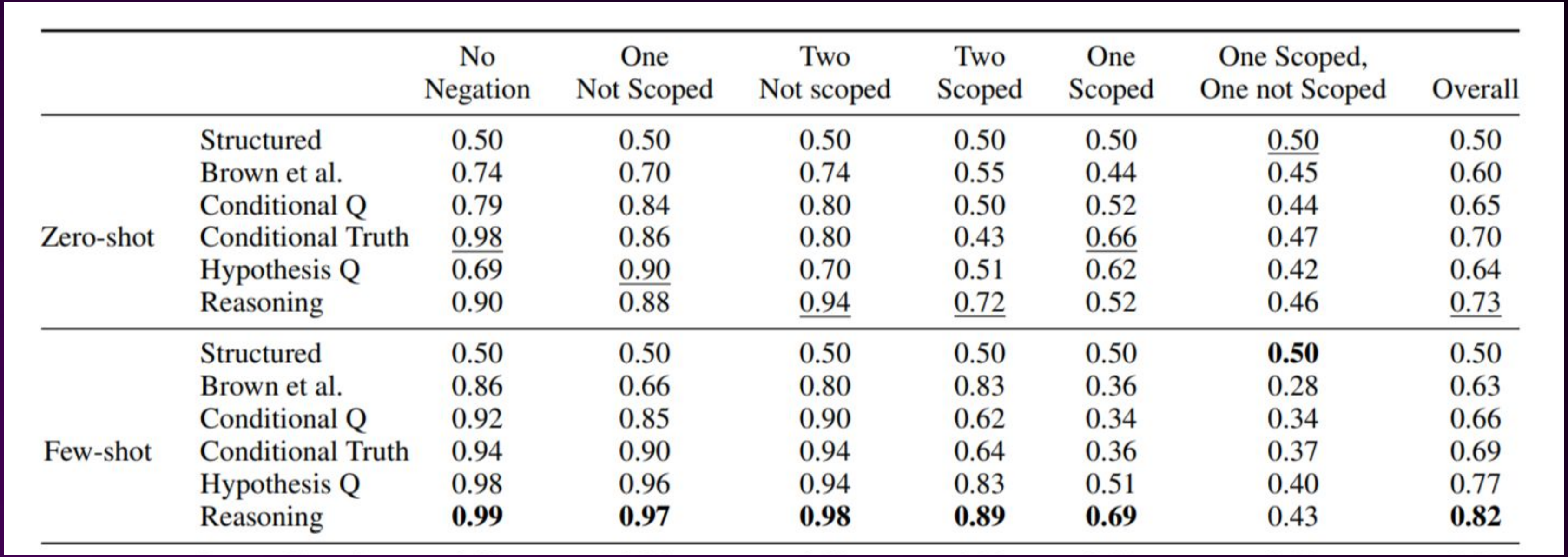
Reasoning

Logical and commonsense reasoning exam.\n\n
Explain your reasoning in detail, then answer with Yes or No. Your answers should follow this 4-line format:\n\n
Premise: <a tricky logical statement about the world>.\n
Question: <question requiring logical deduction>.\n
Reasoning: <an explanation of what you understand about the possible scenarios>\n
Answer: <Yes or No>.\n\n
Premise: **Premise**\n
Question: **Hypothesis**\n
Reasoning: Let's think logically step by step. The premise basically tells us that



		No Negation	One Not Scoped	Two Not scoped	Two Scoped	One Scoped	One Scoped, One not Scoped	Overall
Zero-shot	Structured	0.50	0.50	0.50	0.50	0.50	0.50	0.50
	Brown et al.	0.69	0.60	0.59	0.55	0.50	0.48	0.57
	Conditional Q	0.76	0.55	0.65	0.50	0.50	0.50	0.58
	Conditional Truth	0.76	0.64	0.66	0.60	0.50	0.57	0.62
	Hypothesis Q	0.80	0.83	0.86	0.62	0.45	0.40	0.66
	Reasoning	0.85	0.70	0.68	0.62	0.57	0.56	0.66
Few-shot	Structured	0.50	0.50	0.50	0.50	0.50	0.50	0.50
	Brown et al.	0.82	0.75	0.78	**0.72**	0.35	0.29	0.62
	Conditional Q	0.92	0.82	0.78	0.52	0.36	0.32	0.62
	Conditional Truth	0.92	0.89	0.88	0.59	0.36	0.37	0.67
	Hypothesis Q	**0.99**	**0.91**	**0.92**	0.68	0.38	0.40	**0.72**
	Reasoning	0.73	0.85	0.78	0.62	**0.74**	**0.54**	0.71

Table 7: In-context learning results for GPT-3 (davinci-002 engine).



		No Negation	One Not Scoped	Two Not scoped	Two Scoped	One Scoped	One Scoped, One not Scoped	Overall
Zero-shot	Structured	0.50	0.50	0.50	0.50	0.50	0.50	0.50
	Brown et al.	0.74	0.70	0.74	0.55	0.44	0.45	0.60
	Conditional Q	0.79	0.84	0.80	0.50	0.52	0.44	0.65
	Conditional Truth	0.98	0.86	0.80	0.43	0.66	0.47	0.70
	Hypothesis Q	0.69	0.90	0.70	0.51	0.62	0.42	0.64
	Reasoning	0.90	0.88	0.94	0.72	0.52	0.46	0.73
Few-shot	Structured	0.50	0.50	0.50	0.50	0.50	**0.50**	0.50
	Brown et al.	0.86	0.66	0.80	0.83	0.36	0.28	0.63
	Conditional Q	0.92	0.85	0.90	0.62	0.34	0.34	0.66
	Conditional Truth	0.94	0.90	0.94	0.64	0.36	0.37	0.69
	Hypothesis Q	0.98	0.96	0.94	0.83	0.51	0.40	0.77
	Reasoning	**0.99**	**0.97**	**0.98**	**0.89**	**0.69**	0.43	**0.82**

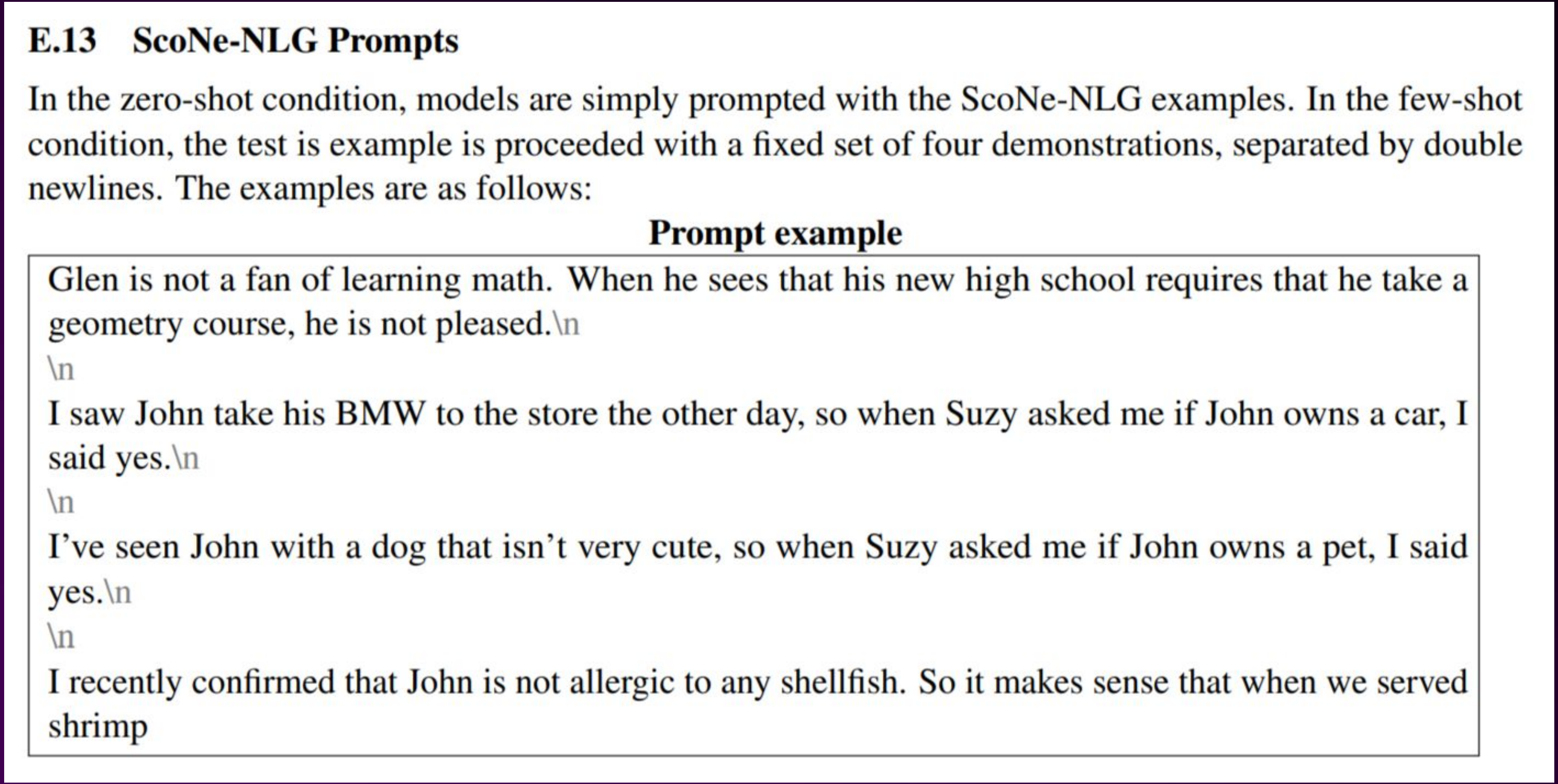
SсoNe-NLG

Гипотеза:

InstructGPT может корректно рассуждать об отрицании при оценке примеров, созданных с учетом задачи, на которую её обучали

Датасет:

ScoNe-NLG - это датасет для генерации текста, который содержит 74 тройки примеров с недописанными предложениями, имеющими различные логические завершения в зависимости от наличия отрицания и сферы его действия



E.13 ScoNe-NLG Prompts

In the zero-shot condition, models are simply prompted with the ScoNe-NLG examples. In the few-shot condition, the test is example is proceeded with a fixed set of four demonstrations, separated by double newlines. The examples are as follows:

Prompt example

```
Glen is not a fan of learning math. When he sees that his new high school requires that he take a geometry course, he is not pleased.\n
\n
I saw John take his BMW to the store the other day, so when Suzy asked me if John owns a car, I said yes.\n
\n
I've seen John with a dog that isn't very cute, so when Suzy asked me if John owns a pet, I said yes.\n
\n
I recently confirmed that John is not allergic to any shellfish. So it makes sense that when we served shrimp
```

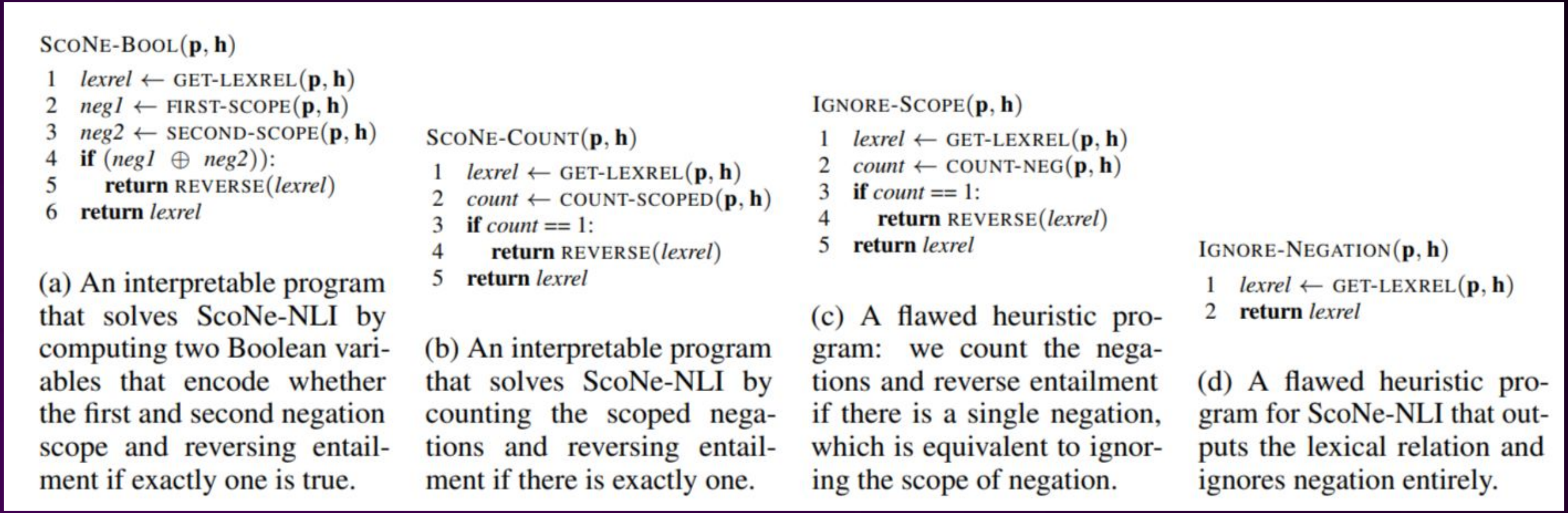
Обучение в контексте на ScoNe-NLG

Анализ ответов экспертами вручную на адекватность и связность, согласованность - 216/222 случаях в zero-shot, Fleiss kappa 0,84 и в 220/222 случаях в few-shot, Fleiss kappa 0,91.

Zero-shot: 92% успешности

Few-shot: 95% успешности

Возможные алгоритмы интерпретации



```
SCONE-BOOL(p, h)
  lexrel ← GET-LEXREL(p, h)
  neg1 ← FIRST-SCOPE(p, h)
  neg2 ← SECOND-SCOPE(p, h)
  if ($neg1 \oplus neg2$):
      return REVERSE(lexrel)
  return lexrel
```

(a) An interpretable program that solves ScoNe-NLI by computing two Boolean variables that encode whether the first and second negation scope and reversing entailment if exactly one is true.

```
SCONE-COUNT(p, h)
  lexrel ← GET-LEXREL(p, h)
  count ← COUNT-SCOPED(p, h)
  if count == 1:
      return REVERSE(lexrel)
  return lexrel
```

(b) An interpretable program that solves ScoNe-NLI by counting the scoped negations and reversing entailment if there is exactly one.

```
IGNORE-SCOPE(p, h)
  lexrel ← GET-LEXREL(p, h)
  count ← COUNT-NEG(p, h)
  if count == 1:
      return REVERSE(lexrel)
  return lexrel
```

(c) A flawed heuristic program: we count the negations and reverse entailment if there is a single negation, which is equivalent to ignoring the scope of negation.

```
IGNORE-NEGATION(p, h)
  lexrel ← GET-LEXREL(p, h)
  return lexrel
```

(d) A flawed heuristic program for ScoNe-NLI that outputs the lexical relation and ignores negation entirely.

Ограничения

- Англоязычность
- Только лексический entailment
- SсoNe может наследовать проблемы датасетов-предшественников