RLAIF: Scaling Reinforcement Learning from Human Feedback with AI Feedback

Аксёнов Ярослав

Что такое alignment языковых моделей?

Prompt

Explain the moon landing to a 6 year old in a few sentences.

Completion

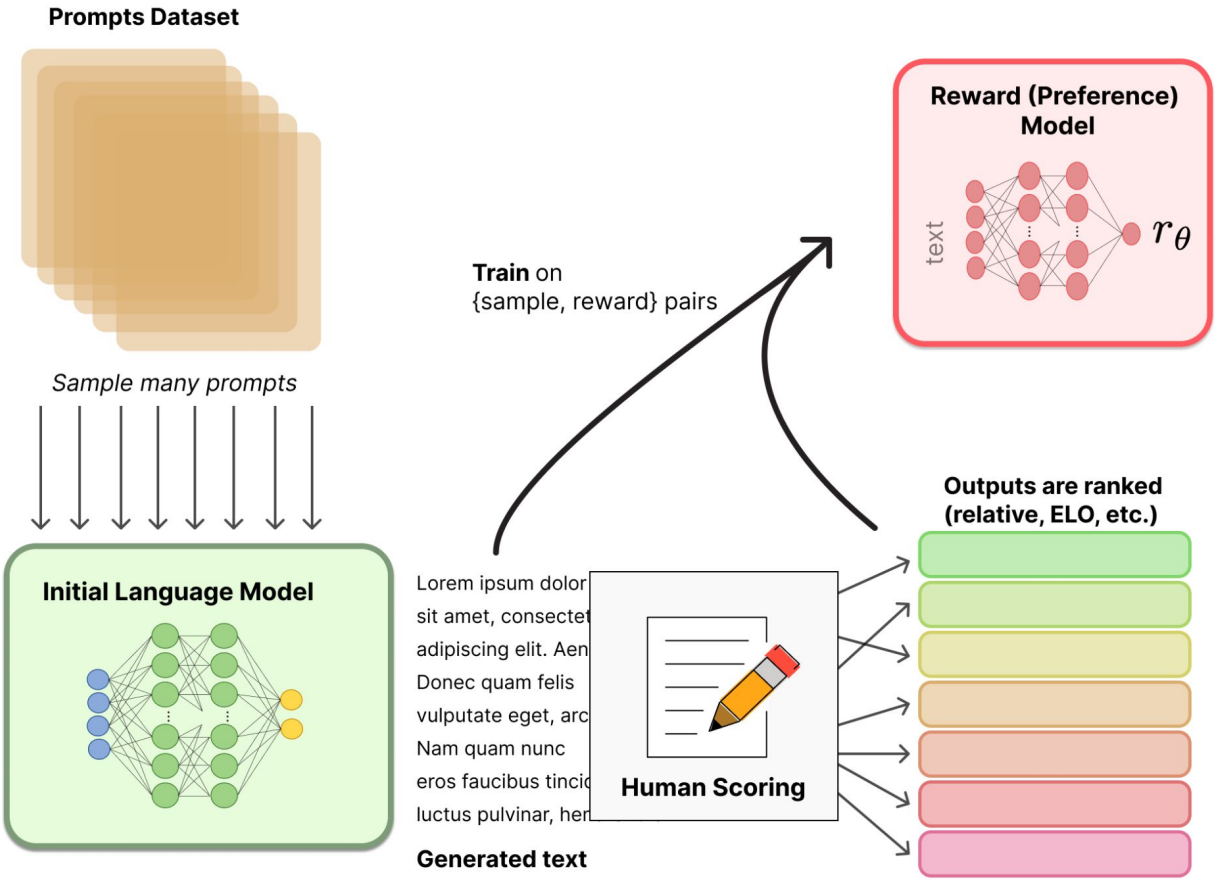
GPT-3

```
Explain the theory of gravity to a 6 year old.

Explain the theory of relativity to a 6 year old in a few sentences.

Explain the big bang theory to a 6 year old.

Explain evolution to a 6 year old.
```

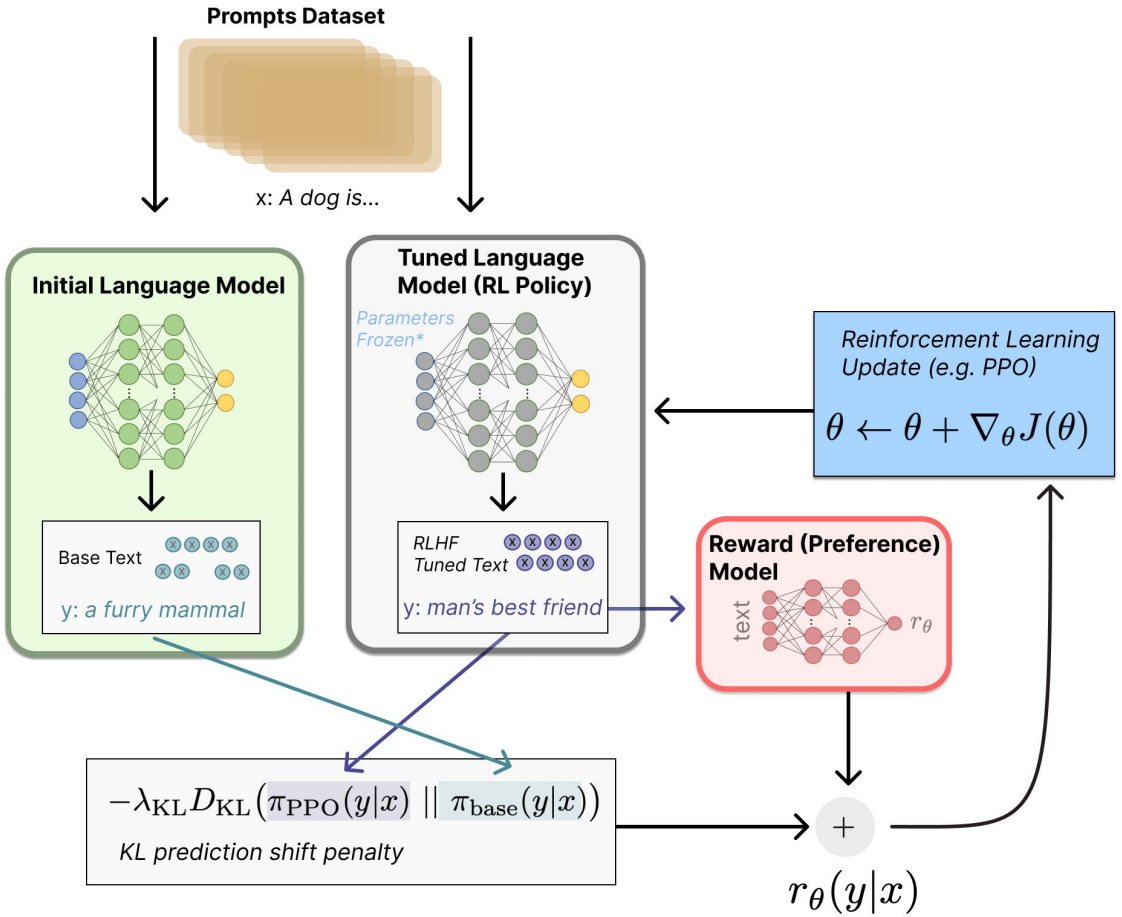
InstructGPT

```
People went to the moon, and they took pictures of what they saw, and sent them
back to the earth so we could all see them.
```

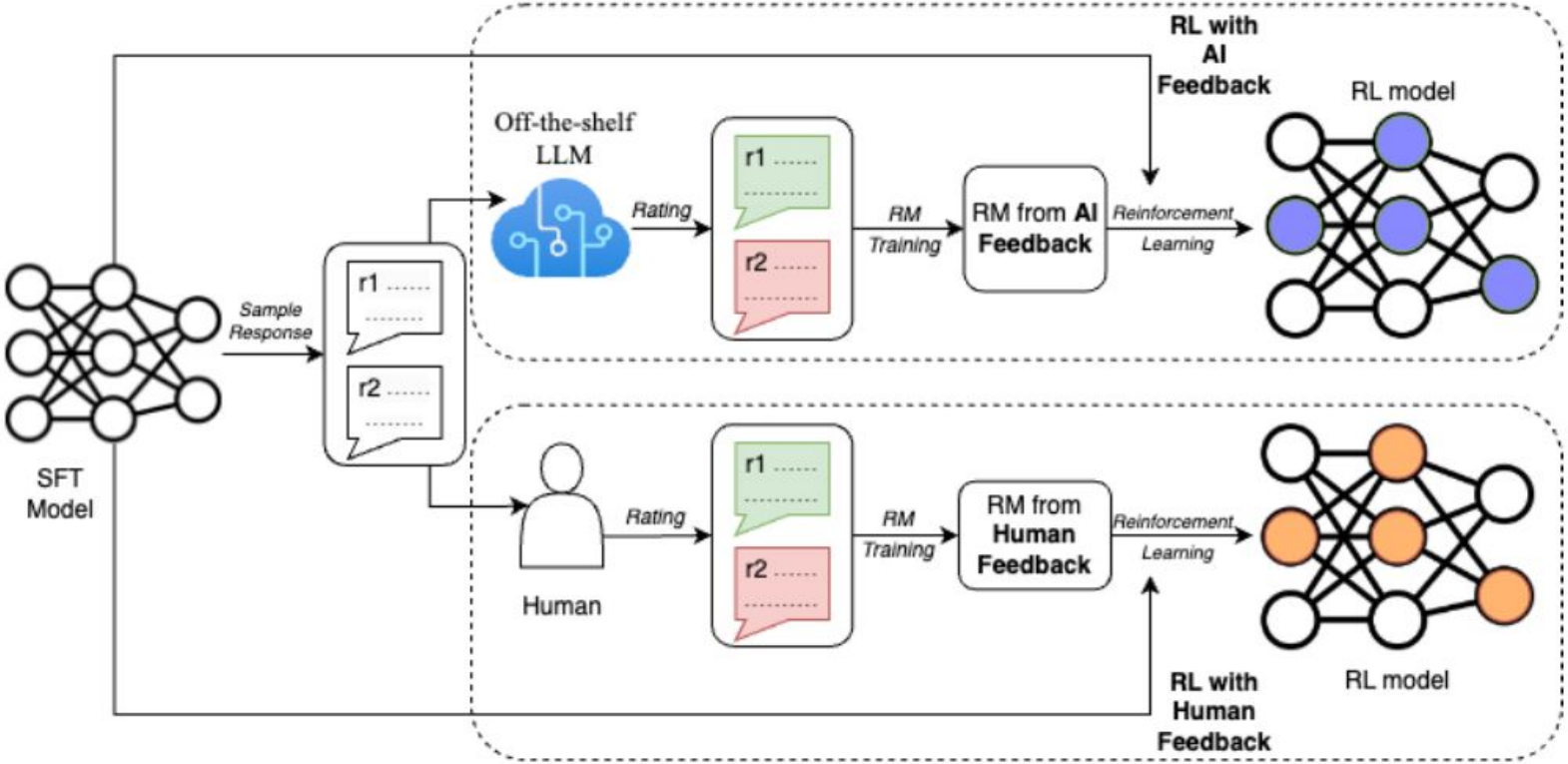
Напоминание про RL with Human Feedback (RLHF)



Напоминание про RL with Human Feedback (RLHF)



RL with AI Feedback (RLAIF)



Preference Labeling with LLM

1. *Preamble* - Introduction and instructions describing the task at hand
2. *Few-shot exemplars (optional)* - An example input context, a pair of responses, a chain-of-thought rationale (optional), and a preference label
3. *Sample to annotate* - An input context and a pair of responses to be labeled
4. *Ending* - Ending text to prompt the LLM (e.g. "*Preferred Response*=")

Preamble	A good summary is a shorter piece of text that has the essence of the original. ... Given a piece of text and two of its possible summaries, output 1 or 2 to indicate which summary best adheres to coherence, accuracy, coverage, and overall quality as defined above.
Exemplar	»»»» Example «««« Text – We were best friends over 4 years ... Summary 1 – Broke up with best friend, should I wish her a happy birthday... And what do you think of no contact? Summary 2 – should I wish my ex happy birthday, I broke no contact, I'm trying to be more patient, I'm too needy, and I don't want her to think I'll keep being that guy. Preferred Summary=1 »»»» Follow the instructions and the example(s) above ««««
Sample to Annotate	Text – {text} Summary 1 – {summary1} Summary 2 – {summary2}
Ending	Preferred Summary=

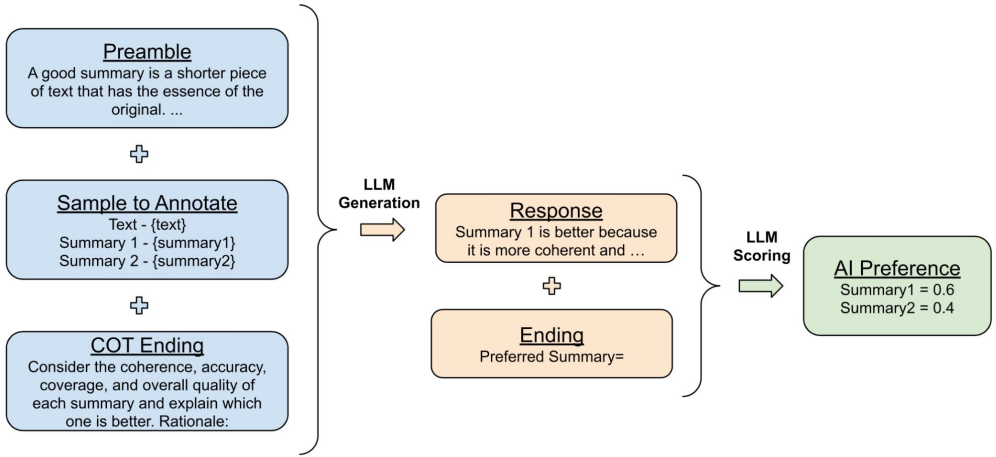
Preference Labeling with LLM

"Base" preamble

```
You are an expert summary rater.  Given a piece of text and
two of its possible summaries, output 1 or 2 to indicate
which summary is better.
```

"Detailed" preamble

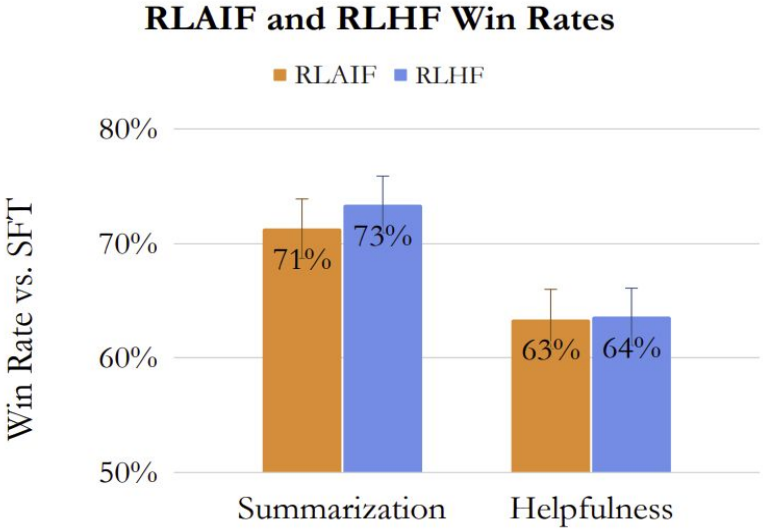
```
A good summary is a shorter piece of text that has the
essence of the original.  It tries to accomplish the same
purpose and conveys the key information from the original
post.  Below we define four evaluation axes for summary
quality:  coherence, accuracy, coverage, and overall quality.

Coherence:  This axis answers the question "how coherent is
the summary on its own?" A summary is coherent if it's easy
to understand when read on its own and free of English errors.
A summary is not coherent if it's difficult to understand
what the summary is trying to say.  Generally, it's more
important that the summary is understandable than it being
free of grammar errors.

Accuracy:  This axis answers the question "does the factual
information in the summary accurately match the post?" A
summary is accurate if it doesn't say things that aren't in
the article, it doesn't mix up people, and generally is not
misleading.

Coverage:  This axis answers the question "how well does
the summary cover the important information in the post?" A
summary has good coverage if it mentions the main information
from the post that's important to understand the situation
described in the post.  A summary has poor coverage if
someone reading only the summary would be missing several
important pieces of information about the situation in the
post.  A summary with good coverage should also match the
purpose of the original post (e.g.  to ask for advice).

Overall quality:  This axis answers the question "how good
is the summary overall at representing the post?" This can
encompass all of the above axes of quality, as well as others
you feel are important.  If it's hard to find ways to make
the summary better, the overall quality is good.  If there
are lots of different ways the summary can be made better,
the overall quality is bad.

You are an expert summary rater.  Given a piece of text and
two of its possible summaries, output 1 or 2 to indicate
which summary best adheres to coherence, accuracy, coverage,
and overall quality as defined above.
```

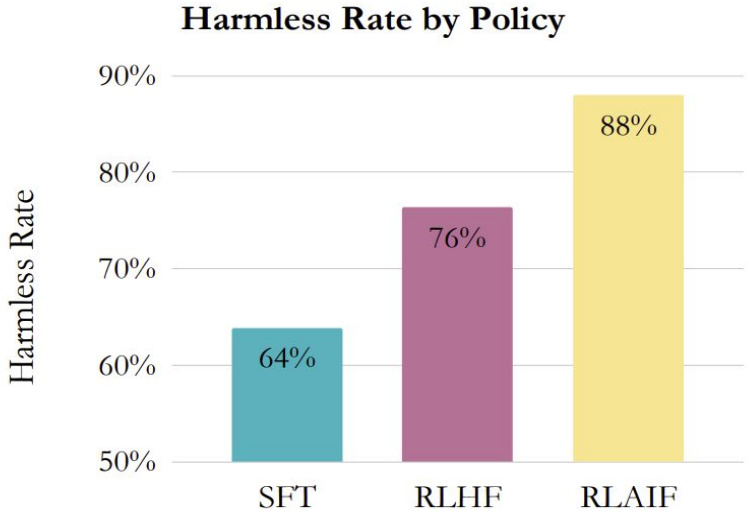


Метрики

AI Labeler Alignment: $z_{acc} = \frac{1}{D}\sum_{i=1}^{D} \mathbb{1}[\arg\max_j P_{i,j}^{AI} = p_i^H]$

Win rate, Harmless Rate: измеряется людьми





Результаты

Win Rate			Harmless Rate	
Comparison	**Summa -rization**	**Helpful dialogue**	**Model**	**Harmless dialogue**
RLAIF **vs** SFT	71%	63%	SFT	64%
RLHF **vs** SFT	73%	64%	RLHF	76%
RLAIF **vs** RLHF	50%	52%	RLAIF	88%
Same-size RLAIF **vs** SFT	68%			
Direct RLAIF **vs** SFT	74%			
Direct RLAIF **vs** Same-size RLAIF	60%			

Эксперименты с промптами для LLM

	AI Labeler Alignment		
Prompt	Summary	H1	H2
Base 0-shot	76.1%	67.8%	69.4%
Base 1-shot	76.0%	67.1%	71.7%
Base 2-shot	75.7%	66.8%	**72.1%**
Base + CoT 0-shot	77.5%	**69.1%**	70.6%
Detailed 0-shot	77.4%	67.6%	70.1%
Detailed 1-shot	76.2%	67.6%	71.5%
Detailed 2-shot	76.3%	67.3%	71.6%
Detailed 8-shot	69.8%	–	–
Detailed + CoT 0-shot	**78.0%**	67.8%	70.1%
Detailed + CoT 1-shot	77.4%	67.4%	69.9%
Detailed + CoT 2-shot	76.8%	67.4%	69.2%

H1 – helpfulness
H2 – harmlessness

Model Size	AI Labeler Alignment
PaLM 2 L	**78.0%**
PaLM 2 S	73.8%
PaLM 2 XS	62.7%

Итоги

- RLAIF достигает сопоставимых или превосходящих показателей по сравнению с RLHF в задачах суммаризации, создания полезных и безвредных диалогов
- RLAIF работает лучше, чем supervised fine-tuning, даже если LLM для разметки имеет тот же размер, что и policy LLM

Вопросы