Techniques for fine-tuning LLMS

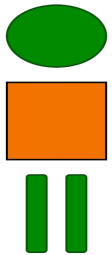
გეგმა

- ინსტრუქციების დასწავლა
- არამეტრულად ეფექტური Fine-tuning: LoRA (დაბალი რანგის მორგება)
- ადამიანის პრეფერენციებისთვის მორგება (DPO/RLHF)

□ა ისწავლეს ენის მოდელებმა pre-training-ის დროს?

- თავისუფალი უნივერსიტეტი მდებარეობს _____, საქართველოს დედაქალაქში. [წვრილმანი]
- თვითონვე გრძნობდა მარიტა ბუნების _____ მომადლებულ ნიჭს [სინტაქსი]
- ანამ წიგნი წაიკითხა. იგი საინტერესო იყო და _____ მრავალი გამოცანა იყო მოცემული. [კორეფერენცია]
- დედამ გემრიელი კერძები მოამზადა სტუმრებისთვის: ღომი, საცივი და _____. [ლექსიკო-სემანტიკა]
- საერთო ჯამში, ამ ორი საათის განმავლობაში რა სარგებელიც მივიღე იყო საჭმელი და სასმელი, ფილმი _____ იყო. [განწყობა]
- ფიქრი და მსჯელობა იმაზე, რაც გაბრაზებს, გაშინებს ან, პირიქით, მოგწონს, შესასწავლი საკითხისადმი ------ გამო, კვლევის ობიექტურობას ------ [მსჯელობა]
- საშუალოდ 300 ლარი თუ გაქვს ხელფასი, 20% ანუ _____ ლარი გადასახადებში მიდის. [არითმეტიკა]

დიდი ენის მოდელებიდან ასისტენტებამდე



- ოდელმა იცის ლექსიკოგრაფია, აზროვნებს და იცის როგორ დააგენერიროს შემდეგი სიტყვა სწორად.
- ოდელმა იცის როგორ გაყვეს ინსტრუქციებს
- ოდელი ცდილობს არ იყოს დამაზიანებელი
- ოდელი ითვალისწინებს დომეინის პრეფერენციებს
- ოდელი ითვალისწინებს ადამიანების პრეფერენციებს

Step 1

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3 with supervised learning.

ინსტრუქციების დასწავლა - ინტუიცია

###ინსტრუქცია :მოიფიქრე ზღაპარი ჩემს ძაღლ სემიკოზე. ###პასუხი: იყო —---------- ერთი

###ინსტრუქცია :მოიფიქრე ზღაპარი ჩემს ძაღლ სემიკოზე. ###პასუხი: იყო ერთი —---------- ძაღლი

###ინსტრუქცია :მოიფიქრე ზღაპარი ჩემს ძაღლ სემიკოზე. ###პასუხი: იყო ერთი ძაღლი —---------- რომელსაც

###ინსტრუქცია :მოიფიქრე ზღაპარი ჩემს ძაღლ სემიკოზე. ###პასუხი: იყო ერთი ძაღლი რომელსაც —---------- ასტრონომია

###ინსტრუქცია :მოიფიქრე ზღაპარი ჩემს ძაღლ სემიკოზე. ###პასუხი: იყო ერთი ძაღლი რომელსაც ასტრონომია —---------- იტაცებდა

.....

.....

....

...

ინსტრუქციების დასწავლა - მონაცემების მოგროვება

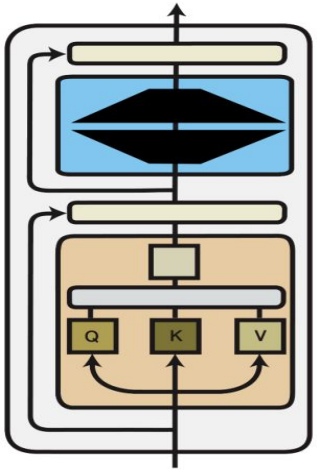
- ოვაგროვოთ მაგალითები (ინსტრუქცია, პასუხი) წყვილების მრავალი დავალებების გარშემო და გადავატრენინგოთ ენის მოდელი.
  - ონაცემების შექმნა სინთეთიკურადაც შესაძლებელია (დიდი ენის მოდელების გამოყენებით
  - ნსტრუქციების დასწავლისთვის ძალიან დიდი რაოდენობის მონაცემები არ არის საჭირო

□ონაცემები - InstructGPT განაწილება

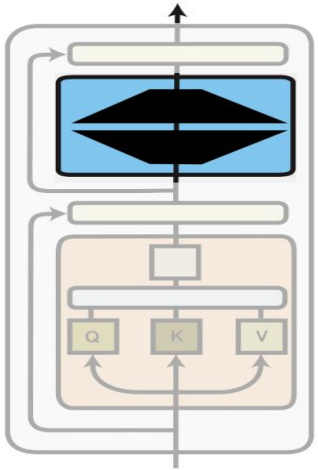
გამოყენების ტიპი	(%)
ტექსტების გენერაცია	45.6 %
ღია კითხვა-პასუხი	12.4 %
იდეების გენერაცია	11.2 %
ჩატი	8.4%
გადაწერა	6.6%
□ოკლე შინაარსი	4.2%
კლასიფიკაცია	3.5%
სხვა	3.5%
დახურული კითხვა-პასუხი	2.6%
მოპოვება	1.9%

Training language models to follow instructions with human feedback https://arxiv.org/pdf/2203.02155

□არამეტრულად ეფექტური Fine-tuning: LoRA



□რული fine-tuning-ის დროს ხდება მოდელის მთლიანი წონების განახლება

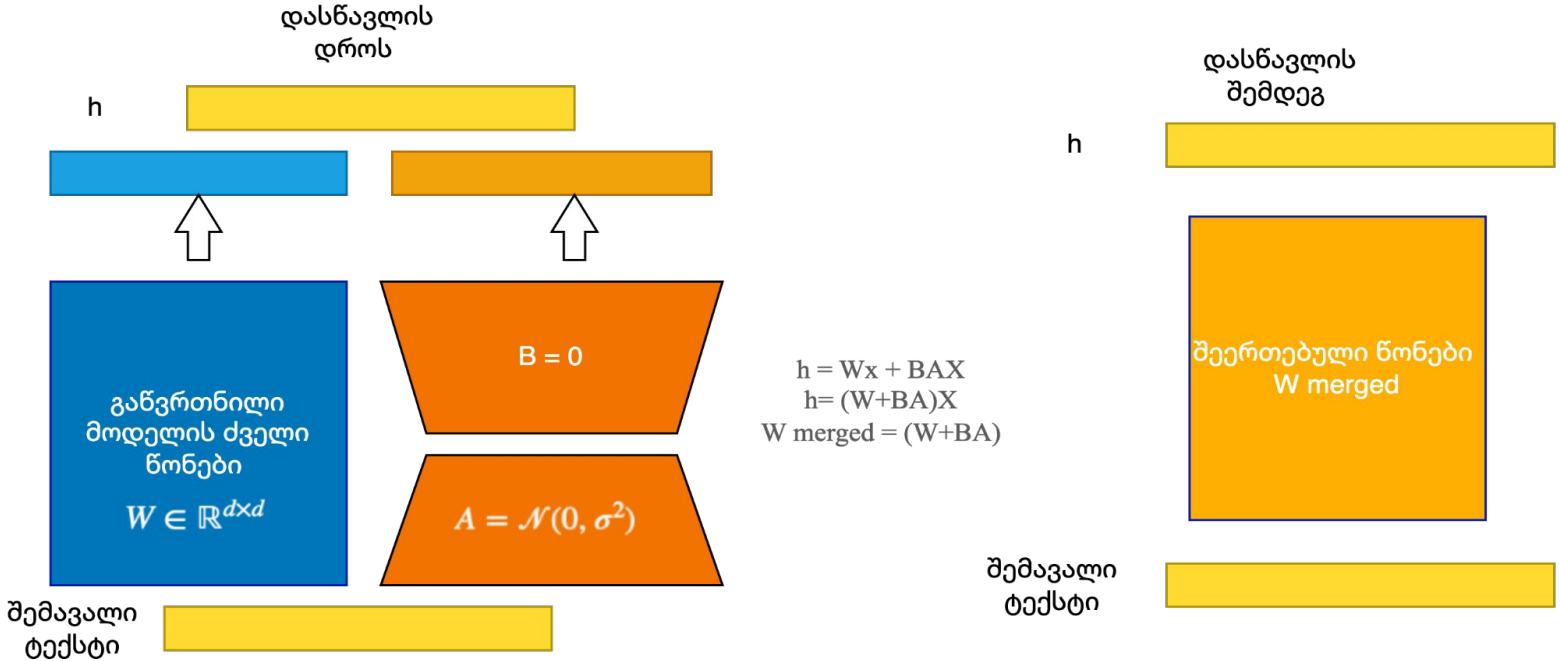


□არამეტრულად ეფექტური fine-tuning-ის დროს მოდელის წონების მხოლოდ მცირე ნაწილს ვწვრთნით

□ატომ უნდა განვაახლოთ წონების მხოლოდ მცირე ნაწილი?

1. □იდი ენის მოდელების სრული წონების გადაწვრთნა არაპრაქტიკულია დიდ დროს და ხარჯებს მოითხოვს
2. □შირად ბევრი პარამეტრები არ არის აუცილებელი საჭირო შედეგების მისაღწევად

□ა არის LoRA (დაბალი-რანგის მორგება) და როგორ შეიძლება გამოვიყენოთ ენის მოდელების fine tuning-ის დროს?



LORA: LOW-RANK ADAPTATION OF LARGE LANGUAGE MODELS https://arxiv.org/pdf/2106.09685

როგორ?

- ვშლით წრფივად დამოკიდებულ სვეტებს და ვიღებთ A მატრიცას, და B-ს დახმარებით ვაბრუნებთ ძველი განზომილების წონების მატრიცას
- □აც უნდა ავარჩიოთ არის რანგი
- □ანგი გადაწყვეტს რამდენი გასაწვრთნელი პარამეტრი გვაქვს fine-tuningის პროცესში
  - □უ ძალიან დაბალ რანგს ავიღებთ , დავკარგავთ ინფორმაციას
  - □უ ძალიან მაღალ რანგს ავიღებთ, ბევრ წრფივად დამოკიდებულ სვეტებს დავტოვებთ რაც ბევრ გასაწვრთნელ პარამეტრს და დიდ გაწვრთნის დროს ნიშნავს

Step 1

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3 with supervised learning.

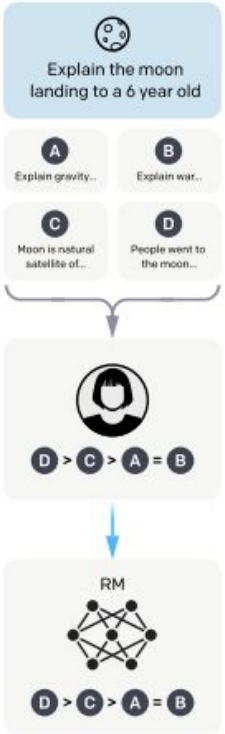
Step 2

Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.



Optimizing for human preferences (DPO/RLHF)

- □არმოვიდგინოთ რომ გვაქვს გაწვრთნილი ენის მოდელი მოკლე შინაარსის ამოცანაზე
- ინსტრუქცია x და პასუხი y-სთვის, ჩვენ უნდა გავარკვიოთ გზა რომ დავთვალოთ ადამიანის შეფასება კონკრეტულ მოკლე შინაარსზე $R\,x, y \in \mathbb{R}$, მაღალი რიცხვი უკეთესია.

□ან ფრანცისკოში,
კალიფორნიის შტატში მოხდა 4.2
ბალი მაგნიტუდის მიწისძვრა
რომელმაც სრულიად
განაცვიფრა სან ფრანცისკო……
……..
□რამყარი ობიექტები წაიქცა

მიწისძვრა მოხდა სან-
ფრანცისკოში. მცირე
მატერიალური ზარალი იყო,
თუმცა არავინ დაშავებულა.
$y1$
$R(\,x, y1) = 8.0$

სან ფრანცისკოში კარგი ამინდია,
მაგრამ მიდრეკილია მიწისძვრებისა
და ტყის ხანძრებისკენ.
$y2$
$R(\,x, y1) = 1.2$

- ახლა ჩვენ გვინდა მაქსიმალურად გავზარდოთ ნიმუშების მოსალოდნელი ჯილდო ჩვენი LM-დან:

$$\mathbb{E}_{\hat{y} \sim p\theta(y|x)}[R(x, \hat{y})]$$

□ილდოს მოდელის გაწვრთნა

- □ადგან რთულია მოდელის ყველა პასუხები ადამიანს შევაფასებინოთ, საჭიროა გვქონდეს მოდელი რომელსაც ეცოდინება ადამიანის პრეფერენციები და რომელსაც გამოვიყენებთ ჯილდოს მოდელად
- □მისთვის ისევ ადამიანებს უნდა ვთხოვოთ რომ მოვაგროვოთ მონაცემები ჯილდოს მოდელის გასაწვრთნელად!

□ილდოს მოდელის მონაცემების მოგროვება

□მოარჩიე რომელი მოკლე შინაარსია უკეთესი მოცემული ტექსტისთვის:

□ან ფრანცისკოში, კალიფორნიის შტატში მოხდა 4.2 ბალი მაგნიტუდის მიწისძვრა რომელმაც სრულიად განაცვიფრა სან ფრანცისკო……
……..
□რამყარი ობიექტები წაიქცა

მიწისძვრა მოხდა სან-ფრანცისკოში. მცირე მატერიალური ზარალი იყო, თუმცა არავინ დაშავებულა.

A

სან ფრანცისკოში კარგი ამინდია, მაგრამ მიდრეკილია მიწისძვრებისა და ტყის ხანძრებისკენ.

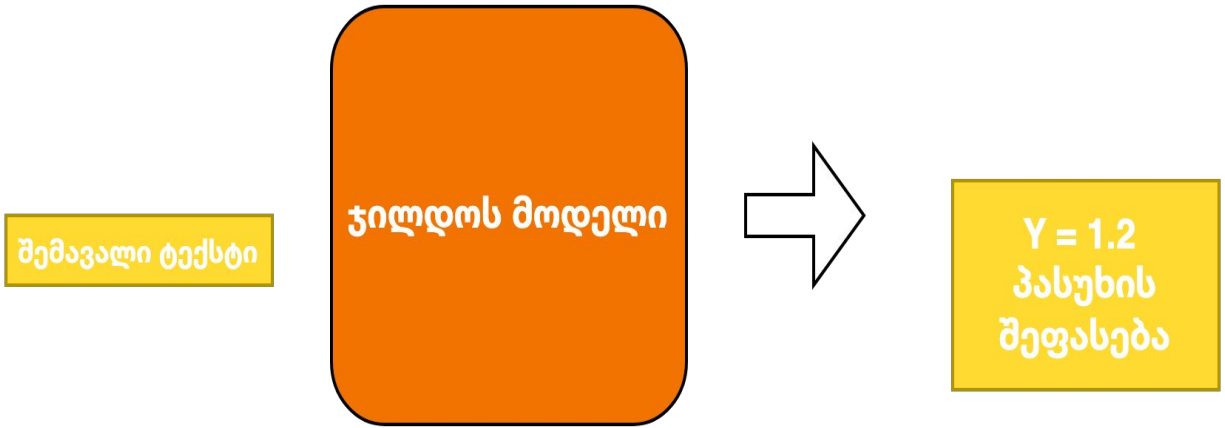
B

□ილდოს მოდელის გაწვრთნა

$$JRM\ \phi = -\mathbb{E}(x, y\hat{}w, y\hat{}l)\sim D \log \sigma(RM\phi\, x, y\hat{}w) - RM\phi(x, y\hat{}l))$$

$y\hat{}w$ - გამარჯვებული მაგალითი

$y\hat{}l$ - დამარცხებული მაგალითი



Step 1

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3 with supervised learning.

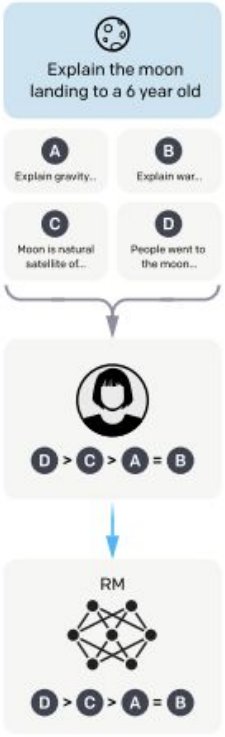
Step 2

Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.



Step 3

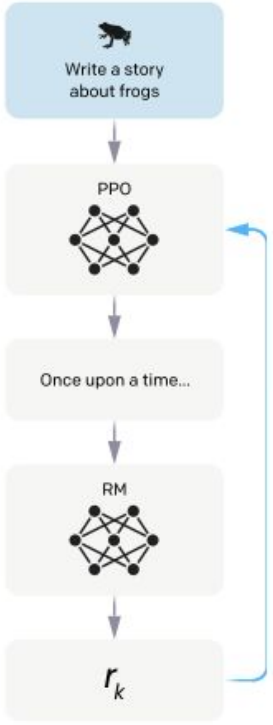
Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

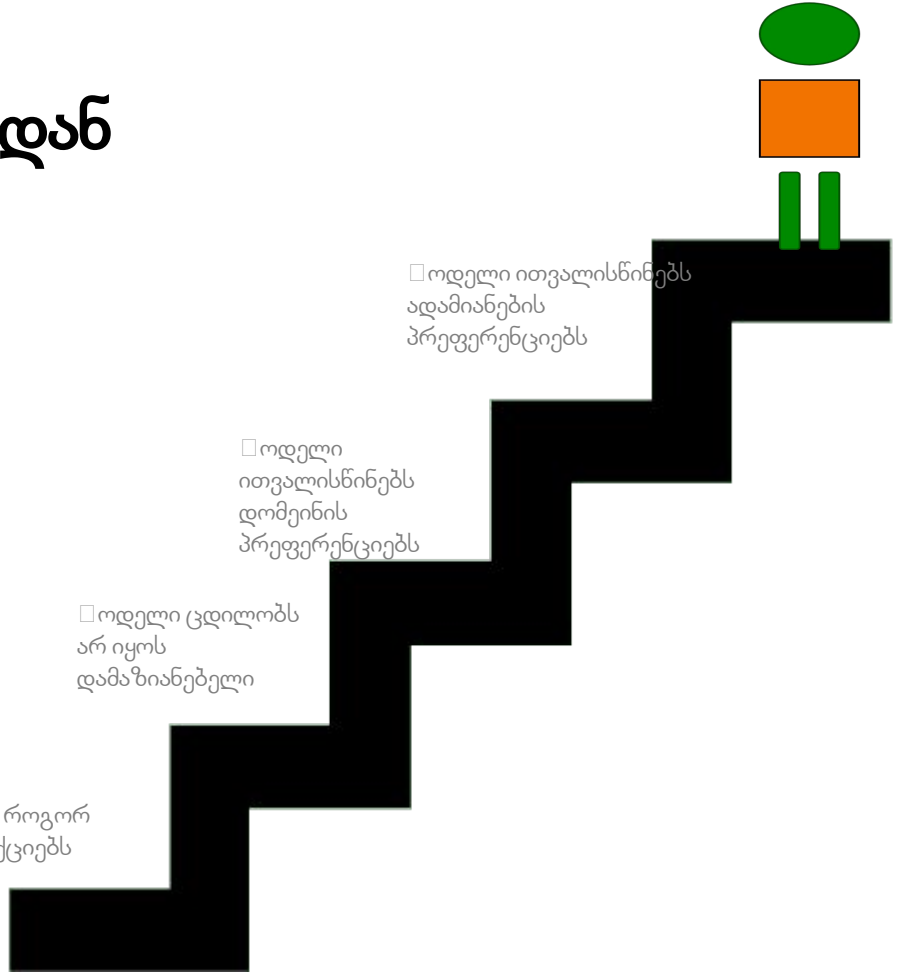
The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.



დიდი ენის მოდელებიდან ასისტენტებამდე



- ოდელმა იცის ლექსიკოგრაფია, აზროვნებს და იცის როგორ დააგენერიროს შემდეგი სიტყვა სწორად.
- ოდელმა იცის როგორ გაყვეს ინსტრუქციებს
- ოდელი ცდილობს არ იყოს დამაზიანებელი
- ოდელი ითვალისწინებს დომეინის პრეფერენციებს
- ოდელი ითვალისწინებს ადამიანების პრეფერენციებს