Master Seminar: Deep Learning for Medical Application

Fine-tuning Large Language Models using Reinforcement Learning

Student: Tomislav Pavković
Supervisor: David Bani-Harouni

Technische Universität München

Johns Hopkins Whiting School of Engineering

Motivation

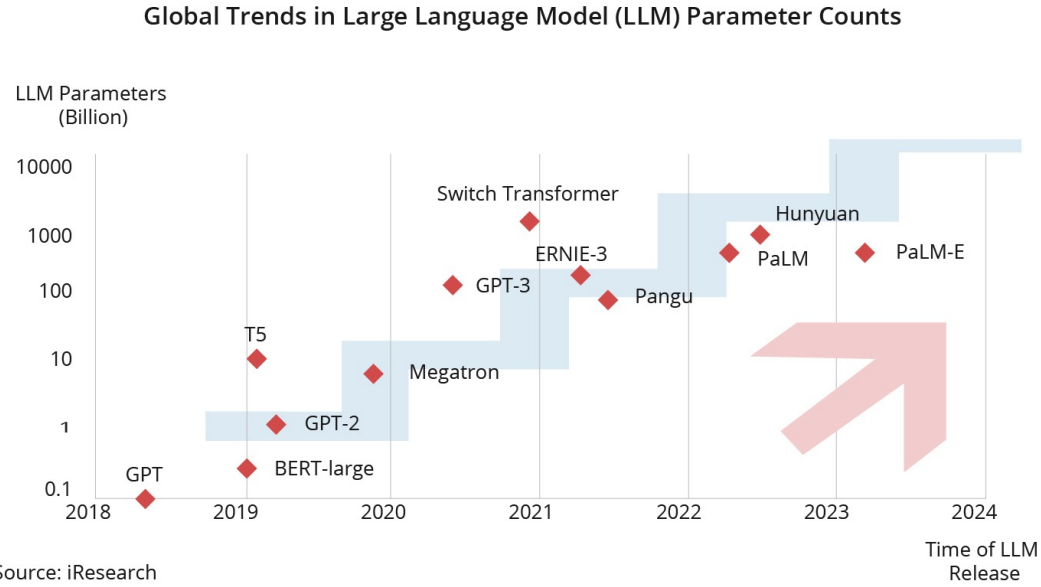
Technische Universität München

+

Johns Hopkins
Whiting School of Engineering

Introduction

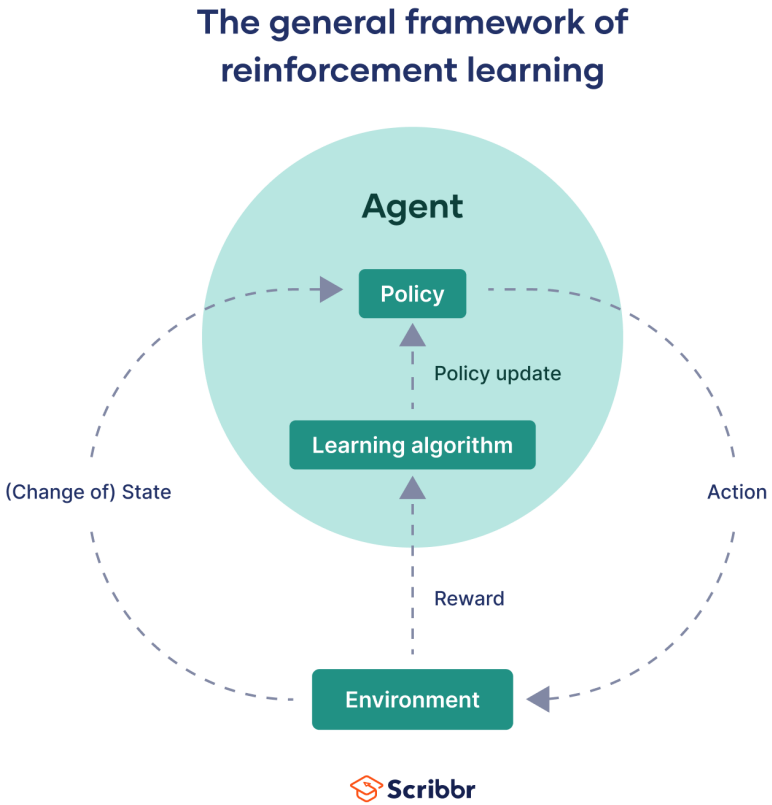
- LLM development started with the invention of transformer architecture in 2017
- Exponential growth led to models with over 100B parameters
- Powerful versatile tools
- Medical applications:
  - Report generation
  - Summarization
  - Pathology QA
- Fine-tuning LLMs to align them with user preferences



[1] https://community.fs.com/blog/the-rise-of-ai-data-centers-fs-empowering-nextgen-data-centers.html

Reinforcement learning

- Method
  - Focused on decision-making
  - Learning through interaction with the environment
- Learning process
  - Trial and error
  - Balancing between exploration and expoitation
- Advantages
  - Capable of handling complex environments
  - Doesn’t require annotated data
- Challenges
  - Requires significant computational resurces
  - Difficulties balancing exploration and exploitation



[2] https://www.scribbr.com/ai-tools/reinforcement-learning/

Explained methods

- InstructGPT – Reinforcement Learning from Human Feedback [3]
- RLP – Reward Learning on Policy [4]
- Quark – Quantized Reward Konditioning [5]
- RAINIER – Reinforced Learning on Policy [6]

[3] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback. 2022. arXiv: 2203.02155 [cs.CL].
[4] H. Lang, F. Huang, and Y. Li. Fine-Tuning Language Models with Reward Learning on Policy. 2024. arXiv: 2403.19279 [cs.CL].
[5] X. Lu, S. Welleck, J. Hessel, L. Jiang, L. Qin, P. West, P. Ammanabrolu, and Y. Choi. Quark: Controllable Text Generation with Reinforced Unlearning. 2022. arXiv: 2205.13636 [cs.CL].
[6] J. Liu, S. Hallinan, X. Lu, P. He, S. Welleck, H. Hajishirzi, and Y. Choi. Rainier: Reinforced Knowledge Introspector for Commonsense Question Answering. 2022. arXiv: 2210.03078 [cs.CL].

InstructGPT – Reinforcement Learning from Human Feedback

Technische Universität München

+

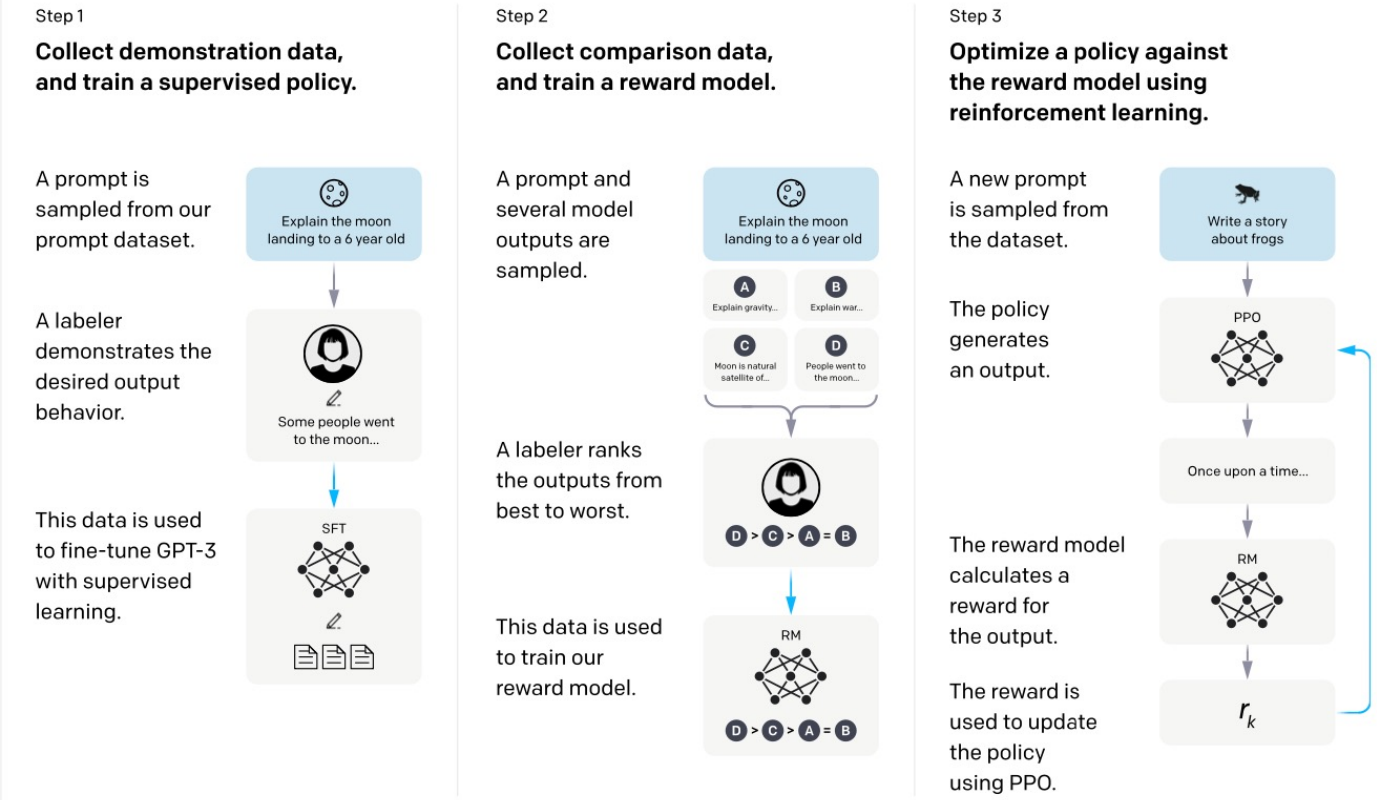
Johns Hopkins
Whiting School of Engineering

RLHF: Method

- Follow user's intentions
  - Explicit instructions
  - Implicit: truthfulness, avoiding bias, toxicity and harm
- InstructGPT – fine-tuned GPT-3 using RLHF
- Training steps:
  - Collect demonstration data and train a supervised policy
  - Collect comparison data and train a reward model
  - Optimize a policy against the reward model using reinforcement learning
- Proximal Policy Optimization [7]

[7] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. "Proximal policy optimization algorithms". In: arXiv preprint arXiv:1707.06347 (2017).

RLHF: Method



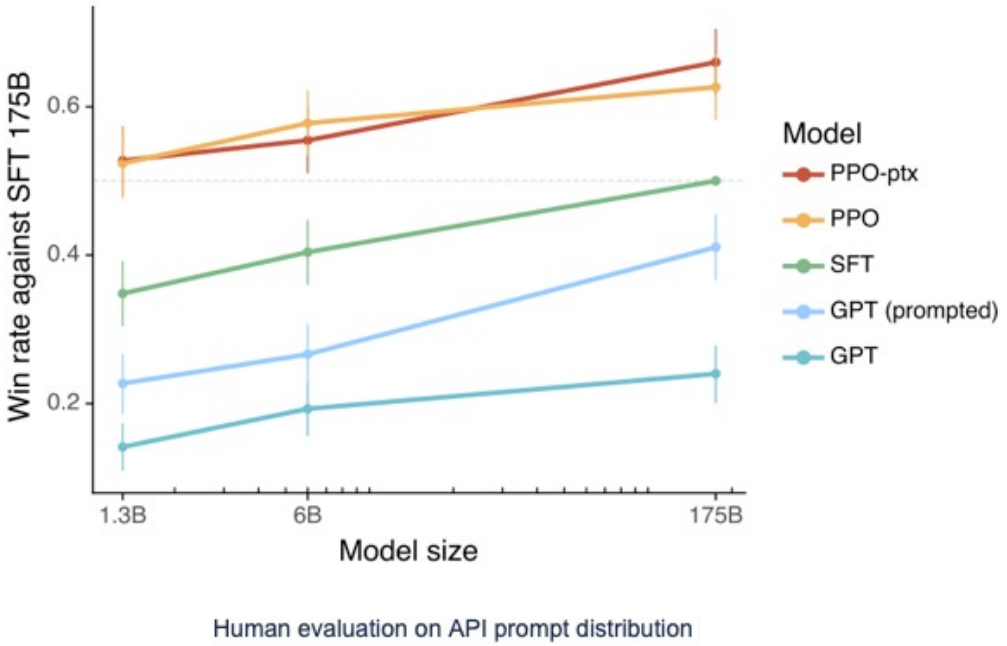
[3] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback. 2022. arXiv: 2203.02155 [cs.CL].

RLHF: Experiments

- Dataset: prompts written by labelers and prompts submitted to InstructGPT API
- Prompts cover wide range of tasks
- Few-shot prompts, zero-shot promts and implicit continuations
- Results:
  - On the API distribution
  - On public NLP datasets
  - Qualitative results

RLHF: Results on API distribution

- Manually evaluated outputs
- Labelers prefer InstructGPT outputs
- Significant improvement over baseline
- Good performance on held-out labelers' dataset

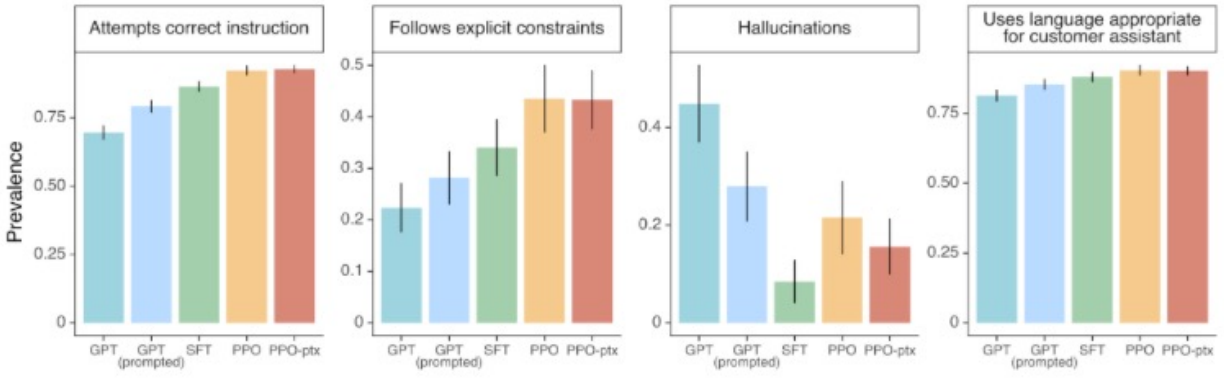


Human evaluation on API prompt distribution

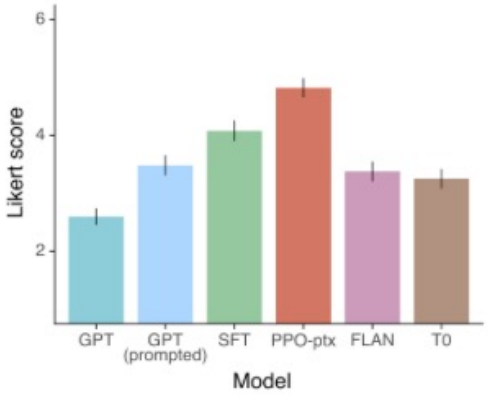
[3] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback. 2022. arXiv: 2203.02155 [cs.CL].

RLHF: Results on API distribution

- InstructGPT outputs are more appropriate for customer assistant tasks, better follows instructions, and hallucinates less
- Reliable and easier to control
- Outperforms GPT-3 fine-tuned on FLAN and T0 datasets



Metadata results on the API distribution

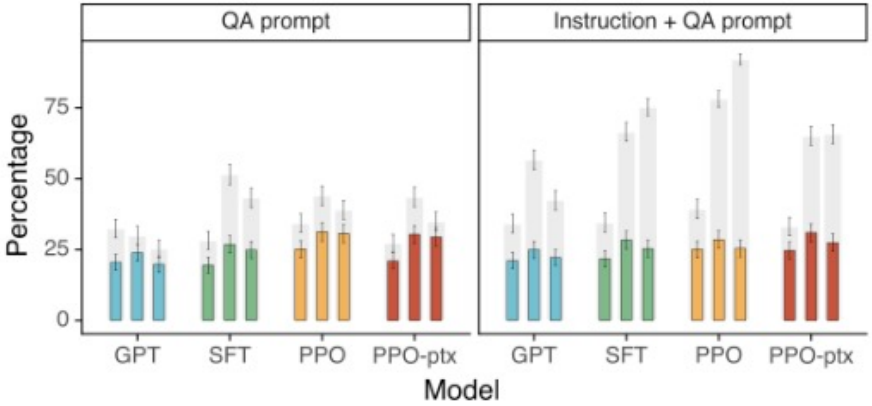


Comparison of InstructGPT to FLAN and T0

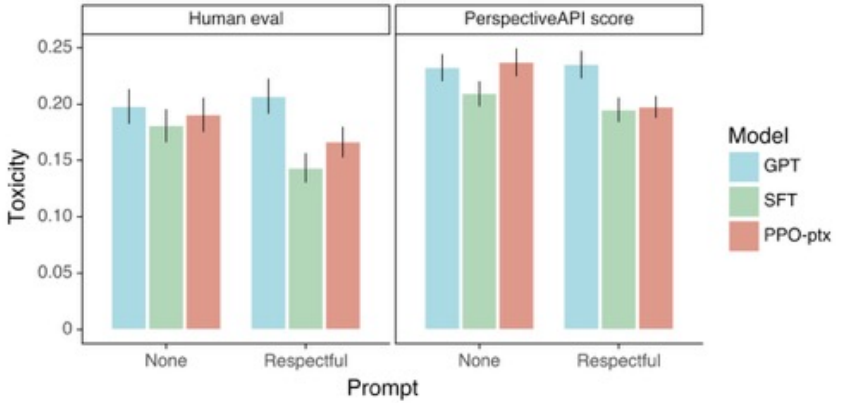
[3] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback. 2022. arXiv: 2203.02155 [cs.CL].

RLHF: Results on public NLP datasets

- Improved truthfulness and informativeness
- Less toxic when instructed to be respectful
- Alignment tax reversed by mixing in pretraining updates



Results on the TruthfulQA dataset



Comparison of human and automatic evaluations on RealToxicityPrompts

[3] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback. 2022. arXiv: 2203.02155 [cs.CL].

RLHF: Qualitative results

- Good results on prompts outside of training data distribution
  - Non-English prompts
  - Code related questions
- InstructGPT problems
  - Overcomplicated outputs for simple questions
  - Mulitple simultanious constraints
  - Accepted false premises

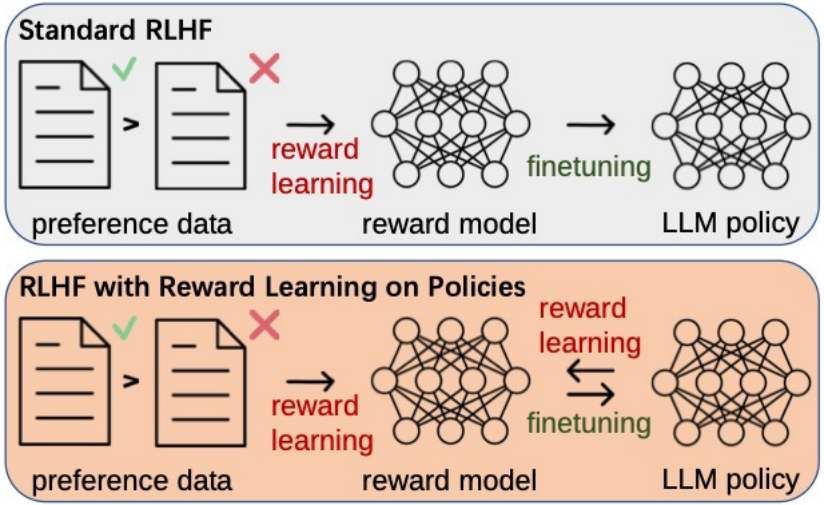
RLP – Reward Learning on Policy

Technische Universität München

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

RLP: Method

- RLHF – reward model off-distribution
- Optimize reward model against policy
- Training
  - Collect demonstration data and train a supervised policy
  - Collect comparison data and train a reward model
  - Optimize a policy against the reward model using reinforcement learning
  - **Reward model retraining**
    - Unsupervised Multi-view Learning
    - Synthetic Preference Generation
  - **Policy retraining**



[4] H. Lang, F. Huang, and Y. Li. Fine-Tuning Language Models with Reward Learning on Policy. 2024. arXiv: 2403.19279 [cs.CL].

RLP: Experiments

- Instruction following task
- RLP-SPG performs the best
- Outperforms baselines inlcluding RLHF (PPO)

Method	AlpacaFarm		LLMBar	Vicuna
	Simulated Win-Rate	Human Win-Rate	Simulated Win-Rate	Simulated Win-Rate
GPT-4	79.0	69.8	74.0	85.0
ChatGPT	61.4	52.9	59.0	63.7
PPO	46.8	55.1	47.5	57.5
Best-of-n	45.0	50.7	43.4	52.5
SFT	36.7	44.3	42.4	50.0
LLaMA-7B	11.3	6.5	12.5	12.8
RLP-UML (ours)	49.1	56.5	48.5	61.3
RLP-SPG (ours)	**50.2**	**57.4**	**50.5**	**62.5**

[4] H. Lang, F. Huang, and Y. Li. Fine-Tuning Language Models with Reward Learning on Policy. 2024. arXiv: 2403.19279 [cs.CL].

Quark – Quantized Reward Konditioning

Technische Universität München

+

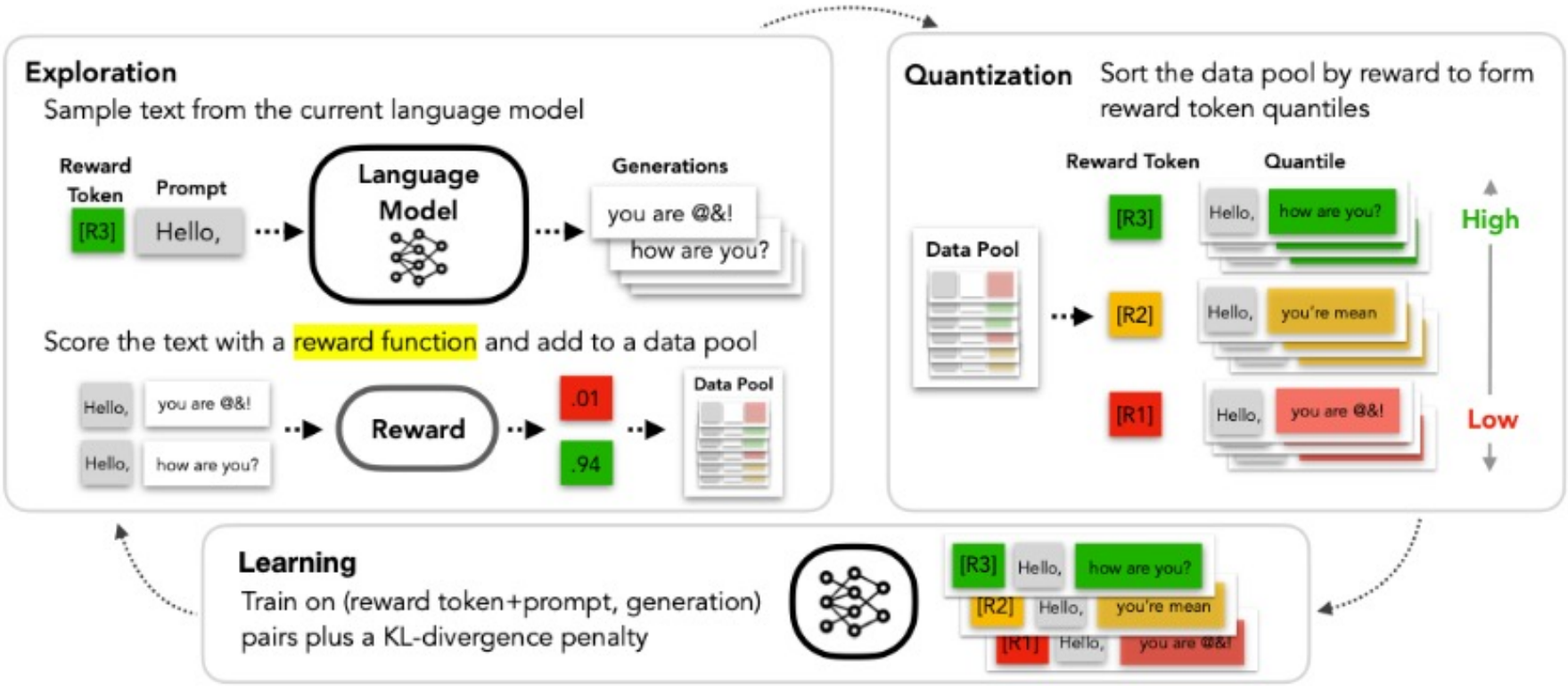
Johns Hopkins
Whiting School of Engineering

Quark: Method

- Unlearning undesirable behavior such as toxicity, negative sentiment and repetition
- Initialized with GPT-2 and datapool sampled from model and scored by reward model
- Training steps:
  - Exploration
  - Quantization
  - Learning
- No need for labeled data

[5] X. Lu, S. Welleck, J. Hessel, L. Jiang, L. Qin, P. West, P. Ammanabrolu, and Y. Choi. Quark: Controllable Text Generation with Reinforced Unlearning. 2022. arXiv: 2205.13636 [cs.CL].

Quark: Method



[5] X. Lu, S. Welleck, J. Hessel, L. Jiang, L. Qin, P. West, P. Ammanabrolu, and Y. Choi. Quark: Controllable Text Generation with Reinforced Unlearning. 2022. arXiv: 2205.13636 [cs.CL].

Quark: Experiments – Unlearning toxicity

- Base model: GPT2-large
- Baselines: **GPT-2**, PPLM, GEDI, DAPT, Dexperts, **PPO** (RLHF)
- Lower toxicity without a decrease in fluency or diversity

Model	In-domain (RealToxicityPrompts)					Out-of-domain (WritingPrompts)				
	Toxicity (↓)		Fluency (↓)	Diversity (↑)		Toxicity (↓)		Fluency (↓)	Diversity (↑)	
	avg. max.	prob.	output ppl	dist-2	dist-3	avg. max.	prob.	output ppl	dist-2	dist-3
GPT2	0.527	0.520	11.31	0.85	0.85	0.572	0.610	12.99	0.82	0.85
PPLM	0.520	0.518	32.58	0.86	0.86	0.544	0.590	36.20	0.87	0.86
GeDi	0.363	0.217	60.03	0.84	0.83	0.261	0.050	91.16	0.86	0.82
DExperts	0.314	0.128	32.41	0.84	0.84	0.343	0.156	42.53	0.86	0.85
DAPT	0.428	0.360	31.21	0.84	0.84	0.442	0.363	38.11	0.86	0.85
PPO	0.218	0.044	14.27	0.80	0.84	0.234	0.048	15.49	0.81	0.84
Quark	**0.196**	**0.035**	**12.47**	0.80	0.84	**0.193**	**0.018**	**14.49**	0.82	0.85

[5] X. Lu, S. Welleck, J. Hessel, L. Jiang, L. Qin, P. West, P. Ammanabrolu, and Y. Choi. Quark: Controllable Text Generation with Reinforced Unlearning. 2022. arXiv: 2205.13636 [cs.CL].

Quark: Experiments – Unlearning unwanted sentiment

- Steer the model away from generating text with unwanted sentiment
- Reduces the generation of unwanted sentiment
- No impact on fluency and diversity

Model	Sentiment to Unlearn: NEGATIVE					Sentiment to Unlearn: POSITIVE				
	% Positive (↑)		Fluency (↓)	Diversity (↑)		% Positive (↓)		Fluency (↓)	Diversity (↑)	
	negative prompt	neutral prompt	output ppl	dist-2	dist-3	positive prompt	neutral prompt	output ppl	dist-2	dist-3
GPT2	0.00	50.02	11.42	0.85	0.85	99.08	50.02	11.42	0.84	0.84
PPLM	8.72	52.68	142.1	0.86	0.85	89.74	39.05	181.7	0.87	0.86
CTRL	18.88	61.81	43.79	0.83	0.86	79.05	37.63	35.94	0.83	0.86
GeDi	26.80	86.01	58.41	0.80	0.79	39.57	8.73	84.11	0.84	0.82
DEXPERTS	36.42	94.46	25.83	0.84	0.84	35.99	3.77	45.91	0.84	0.83
DAPT	14.17	77.24	30.52	0.83	0.84	87.43	33.28	32.86	0.85	0.84
PPO	43.13	94.10	15.16	0.80	0.84	32.22	3.65	15.54	0.81	0.84
Quark	**46.55**	**95.00**	**14.54**	0.80	0.84	**27.50**	**2.75**	**14.72**	0.80	0.84

[5] X. Lu, S. Welleck, J. Hessel, L. Jiang, L. Qin, P. West, P. Ammanabrolu, and Y. Choi. Quark: Controllable Text Generation with Reinforced Unlearning. 2022. arXiv: 2205.13636 [cs.CL].

Quark: Experiments – Unlearning degenerate repetition

- Metrics:
  - Language modeling: perplexity, token prediction accuracy and prediction repetition
  - Generation: sequence-level repetition, diversity and MAUVE
- Better than MLE and SimCTG but worse than Unlikelihood
- Combination of Quark and Unlikelihood outperforms all other methods

Model	Language Model Quality				Generation Quality				Human Eval		
	ppl ↓	acc ↑	rep ↓	wrep ↓	rep-2 ↓	rep-3 ↓	div ↑	mauve↑	fluency ↑	coherence ↑	overall ↑
MLE	24.23	39.63	52.82	29.97	69.21	65.18	0.04	0.03	1.89	2.55	1.96
Unlikelihood	28.57	38.41	51.23	28.57	24.12	13.35	0.61	0.69	2.90	3.19	3.00
SimCTG	**23.82**	40.91	51.66	28.65	67.36	63.33	0.05	0.05	1.93	2.68	2.08
Quark	26.22	**41.57**	45.64	25.07	39.89	30.62	0.35	0.74	2.75	3.20	2.77
+Unlikelihood	27.97	39.41	**37.76**	**19.34**	**18.76**	**12.14**	**0.67**	**0.82**	**3.92**	**4.04**	**3.87**

[5] X. Lu, S. Welleck, J. Hessel, L. Jiang, L. Qin, P. West, P. Ammanabrolu, and Y. Choi. Quark: Controllable Text Generation with Reinforced Unlearning. 2022. arXiv: 2205.13636 [cs.CL].

RAINIER – Reinforcement Knowledge Introspector

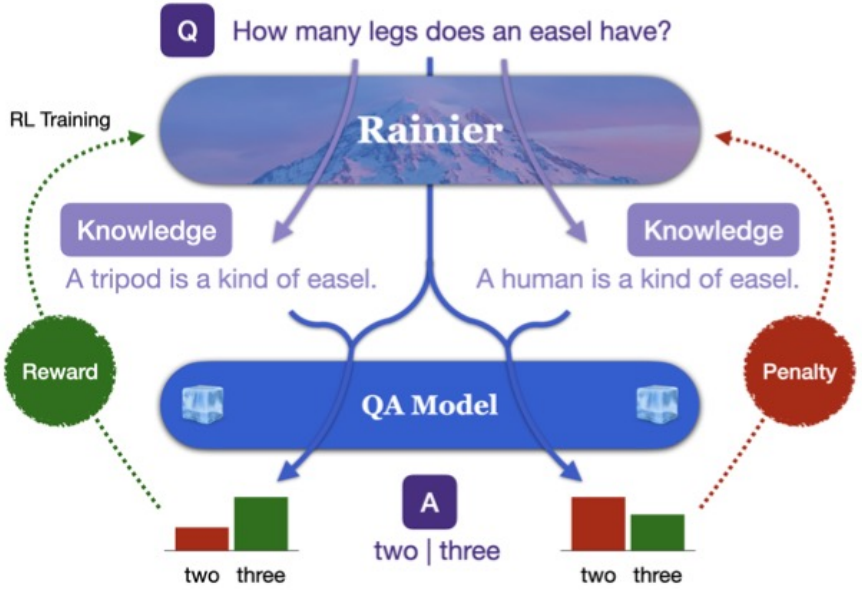
Technische Universität München

+

Johns Hopkins
Whiting School of Engineering

RAINIER: Method

- Commonsense reasoning task
- RAINIER generates useful knowledge in response to given question
- Training:
  - Imitation learning:
    - Supervised fine-tuning
  - Reinforcement learning:
    - PPO against fixed QA model



[6] J. Liu, S. Hallinan, X. Lu, P. He, S. Welleck, H. Hajishirzi, and Y. Choi. Rainier: Reinforced Knowledge Introspector for Commonsense Question Answering. 2022. arXiv: 2210.03078 [cs.CL].

RAINIER: Experiments

- Datasets:
  - Seen: OpenBookQA, ARC, AI2Science, CommonsenseQA, QASC, PhysicallQA, SociallQA, Winograde
  - Unseen: NummerSense, RiddleSense, QuaRTz, HellaSwag
- Base models:
  - Knowledge introspector: T5-large
  - QA model: UnifiedQA
- Silver knowledge generated using GPT-3-Curie model
- Baselines: UnifiedQA-large, UnifiedQA-large + few-shot GPT-3-Curie, UnifiedQA-large + Self-talk, UnifiedQA-large + DREAM model
- Testing with different sizes of fixed QA

[6] J. Liu, S. Hallinan, X. Lu, P. He, S. Welleck, H. Hajishirzi, and Y. Choi. Rainier: Reinforced Knowledge Introspector for Commonsense Question Answering. 2022. arXiv: 2210.03078 [cs.CL].

RAINIER: Experiments – results on seen datasets

- Improvement on 5 out of 8 seen datasets
- Rainier is significantly smaller than other models

Dataset → / Method ↓	CSQA		QASC		PIQA		SIQA		WG		Avg.	
	dev	test	dev	test	dev	test	dev	test	dev	test	dev	test
UQA-large (0.77B)	61.43	53.00	43.09	45.65	63.66	65.50	53.84	57.21	53.35	54.67	55.07	55.21
+ Few-shot GPT-3-Curie (13B)	66.34	–	53.24	–	64.25	–	**58.29**	–	55.56	–	59.54	–
+ Self-talk GPT-3-Curie (13B)	63.31	–	49.89	–	65.23	–	51.89	–	52.96	–	56.66	–
+ DREAM (11B)	64.54	–	49.46	–	64.74	–	51.59	–	56.12	–	57.29	–
+ RAINIER-large (0.77B) [ours]	**67.24**	**60.18**	**54.97**	**54.13**	**65.67**	**67.09**	57.01	**59.01**	**56.91**	**57.39**	**60.36**	**59.56**

[6] J. Liu, S. Hallinan, X. Lu, P. He, S. Welleck, H. Hajishirzi, and Y. Choi. Rainier: Reinforced Knowledge Introspector for Commonsense Question Answering. 2022. arXiv: 2210.03078 [cs.CL].

RAINIER: Experiments – results on unseen datasets

- Improvement on all 4 unseen datasets
- Tested with different QA models showing improvements on every one

Dataset →	NS		RS		QuaRTz		HS		Avg.	
Method ↓	dev	test-all	dev	test	dev	test	dev	test	dev	test
UQA-large (0.77B)	26.50	19.61	28.11	38.34	68.75	67.60	35.00	34.30	39.59	41.85
+ Few-shot GPT-3-Curie (13B)	38.00	–	35.65	–	69.01	–	37.33	–	45.00	–
+ RAINIER-large (0.77B) [ours]	30.00	21.81	30.07	41.22	70.31	68.24	35.73	34.80	41.53	43.76

[6] J. Liu, S. Hallinan, X. Lu, P. He, S. Welleck, H. Hajishirzi, and Y. Choi. Rainier: Reinforced Knowledge Introspector for Commonsense Question Answering. 2022. arXiv: 2210.03078 [cs.CL].

Review

Technische Universität München

+

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Method comparison

- RLHF laid the foundation of language model alignment
  - Significant improvements over the baseline and supervised fine-tuning approach
- Quark outperforms RLHF in unlearning toxicity, steering away from unwanted sentiment and unlearning degenerate repetition
- RLP outperforms baselines including RLHF in instruction following task
  - Unfortunately baselines don’t include Quark method
- RAINIER solves different problem so comparison is not possible
  - It outperforms used baselines

Strengths

- Common advantages
  - Reinforcement learning doesn’t require annotated data, so it isn’t limited on it’s quality
  - Cost of data collection and training is modest compared to the cost of pretraining
- RLHF
  - Generalizes to the tasks it was not trained on
  - Minimizes alignment tax
- Quark
  - Doesn’t require any annotated data
  - Low-tax alignment technique
- RLP
  - RLHF advantages + reward model is kept on-distribution
- RAINIER
  - Doesn’t require any annotated data
  - Outperforms much bigger models

Weaknesses

- Common disadvantages
  - Requires a big amount of non-labeled data and computations
  - Setup usually more complicated than SFT
- InstructGPT
  - Aligns to the preferences of the labelers
  - Fixed reward model can become off-distribution
  - Follows any instruction, even when asked to produce harmful content
- Quark
  - Can be misused to make the model more toxic or biased
  - Inherits the social biases from the reward function
- RLP
  - RLHF disadvantages except for the off-distribution reward model
- RAINIER
  - Gap between model performance and human reasoning, might not be sufficient for real-world apps

Master Seminar: Deep Learning for Medical Application

Thank you for your attention!
Any questions?

Technische Universität München

+

Johns Hopkins
Whiting School of Engineering