Annales Mathematicae et Informaticae
58 (2023) pp. 182–190
DOI: 10.33039/ami.2023.08.008
URL: https://ami.uni-eszterhazy.hu

# Enhancing machine translation with quality estimation and reinforcement learning

**Zijian Győző Yang, László János Laki**

Hungarian Research Centre for Linguistics
{yang.zijian.gyozo,laki.laszlo}@nytud.hu

**Abstract.** In recent times, our research has focused on training large language models and exploring their potential. With the emergence of ChatGPT, it has been demonstrated that it is possible to fine-tune language models in a task-agnostic way. The success of ChatGPT is attributed to the reinforcement learning method, which integrates human feedback into the language model fine-tuning process. As a part of our research, we initially adapted the method of reinforcement learning for a specific task, which is machine translation, respectively. In this paper, we propose a novel approach to enhance machine translation with reinforcement learning and quality estimation methods. Our proposed approach uses reinforcement learning to learn to adjust the machine translation output based on quality estimation feedback, with the goal of improving the overall translation quality. We evaluated our approach on the WMT09 dataset for English-Hungarian language pair. We conducted an analysis to show how our approach improves the quality of machine translation output. Our approach offers a promising avenue for enhancing the quality of machine translation and demonstrates the potential of utilizing reinforcement learning to improve other natural language processing tasks.

*Keywords:* machine translation, reinforcement learning, quality estimation, mT5

*AMS Subject Classification:* 68T07, 68T50

## 1. Introduction

In recent years, significant progress in artificial intelligence and deep learning has resulted in notable enhancements in the quality of machine translation. Quality

Submitted: July 27, 2023
Accepted: August 7, 2023
Published online: August 19, 2023

estimation has become an important task in the field, involving predicting the quality of machine-translated text without having access to a reference translation. Incorporating a real-time quality estimation system is a crucial step in the machine translation pipeline, as it enables the system to determine the most accurate translation and select the best one to present to the user. In the past month, reinforcement learning has been adopted into natural language processing tasks, marking a significant advancement in this field. In the context of machine translation, reinforcement learning has been applied to fine-tune machine translation models and integrate human feedback into the training process. By combining reinforcement learning and quality estimation, machine translation systems can deliver higher-quality translations.

The success of ChatGPT[1] has demonstrated that reinforcement learning can be effectively adopted in human language technology. ChatGPT suggests that language models can be fine-tuned in a task-agnostic manner. This not only stabilizes the non-deterministic behavior of the models but also brings to light their vast knowledge about the world. The ChatGPT system is fine-tuned from a model in the GPT-3.5 series[2]. The GPT-3.5 series belongs to Large Language Models (LLM) [1], which has garnered significant attention and popularity in the field of research in recent times. The enormous success of ChatGPT can be attributed to the utilization of reinforcement learning (RL), a technique that incorporates human feedback into the language modeling process.

Following the popular trend, we have also started training the Hungarian ChatGPT, although this process is extremely time-consuming. Therefore, as a preliminary step, we experimented with the RL in the field of machine translation (MT). In our current research, we have successfully incorporated a neural quality estimation (QE) model and the RL method into the MT training process to enhance its quality.

## 2. Related works

The OpenAI was the first, who successfully integrated the RL approach to the natural language processing training process [14, 22]. The first experiments were done with the summarization task. Thereafter, the RL was adapted to InstructGPT [11], which is the basis of ChatGPT. There are many algorithms in modern RL, but in natural language processing currently the Proximal Policy Optimization (PPO) [13] algorithm became decisive.

Reinforcement learning experiments is still in its early stages in machine translation task. There are studies [18] that show RL is an effective approach for improving the performance of neural machine translation. There have been some studies with skepticism in the field as well [2, 7]. For English-Hungarian language pair, Laki and Yang [8] conducted comprehensive research on machine translation, however reinforcement learning method has not been applied yet.

[1] https://chat.openai.com
[2] https://openai.com/blog/chatgpt

The reward model in the reinforcement learning method can be trained as a QE model in the MT task. Quality estimation is a prediction task, where different quality indicators are extracted from the source and the machine translated segments. The QE model is built with machine learning methods based on these quality indicators. Then the QE model is used to predict the quality of unseen translations [20]. In the recent years, instead of human feature extraction, neural based deep learning methods are used for this task. Since the QE compares two texts from different languages, pretrained multilingual neural language models [12, 15] or dual encoders [6] are used for this task. The pretrained language models can be combined with Multitask Learning architectures [5, 9], or additional custom extracted features can be added to the model [17, 21].

## 3. Methods and experiments

OpenAI showed in its research [10] that using reinforcement learning, we can enhance the performance of a neural language model. Based on research (see Figure 1) by OpenAI, our implementation steps for using reinforcement learning in fine-tuning language models are as follows:

1. *Fine-tuning language model with supervised learning*: In our experiment, we used a mT5 small model[3] [8], that fine-tuned for English-Hungarian translation task (supervised fine-tuned model - SFT).

2. *Collect human feedback and training a reward model*: For this task, we trained a QE model as reward model for English-Hungarian translation.

3. *Fine-tuning language model with reward model and reinforcement learning method*: We further fine-tuned the SFT-mT5 model with reward model and reinforcement learning method (RL).

In the first step, we utilized an already fine-tuned language model, hence we did not train a new model specifically for this task.

In the second step, we conducted experiments using five different models to train QE models:

- mT5 models: Following the research conducted by OpenAI, we initially performed fine-tuned our SFT-mT5 model. Then, conducted experiments using the original mT5-small and mT5-base [19] models.

- mBERT: the BERT multilingual base [4] model was fine-tuned.

- XLM-R: the XLM-RoBERTa-base [3] model was fine-tuned.

To train the QE models we used the HuQ [20] corpus that contains 1500 manual evaluated English-Hungarian sentence pairs. All the 1500 sentences were evaluated

[3] https://huggingface.co/NYTK/translation-mt5-small-128-en-hu

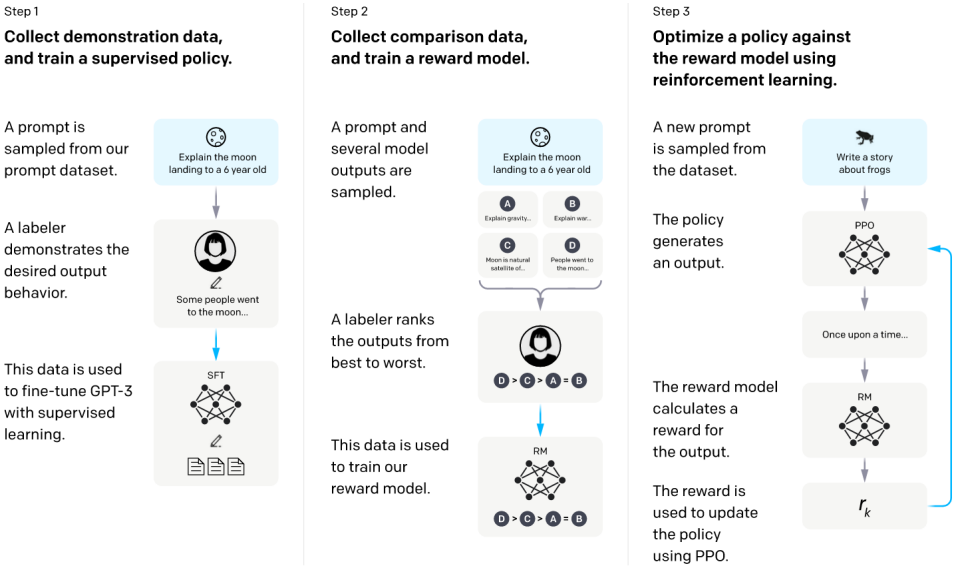


**Figure 1.** Using reinforcement learning in fine-tuning language models. [10]

by 3 human annotators. provided quality scores ranging from 1 to 5, considering adequacy and fluency aspects. For the experiments, we randomly shuffled the segments and divided them into 80% for the train sub-corpus and 20% for the test sub-corpus.

In the third step, we employed the CarperAI implementation[4] to fine-tune our SFT model using reinforcement learning (see Figure 2).

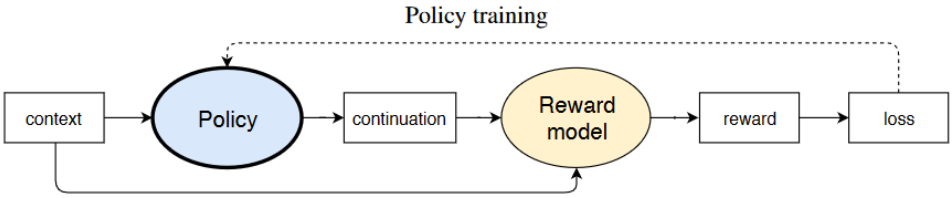


**Figure 2.** Training process for policy. [23]

We used our fine-tuned QE model as the reward model. For training and testing purposes, we used the official sub-corpora of Hunglish [16] corpus from Shared Task of WMT 2009[5]. In the case of reinforcement learning, a smaller amount of training data is sufficient, thus we used the development set as the training

[4] https://github.com/CarperAI/trlx
[5] https://www.statmt.org/wmt09/translation-task.html

corpus. To ensure a fair comparison, we also conducted a fine-tuning experiment where we further fine-tuned the SFT model (SFT-mT5 FFT) using traditional methods without reinforcement learning on the same development set with the same hyperparameters. The main hyperparameters used in our fine-tuning experiments (both RL and FFT) are as follows: learning rate: 2e-5; sequence length: 256; epoch: 10;

# 4. Results

In Table 1, you can see the results of our QE experiments. The mT5 models were unable to effectively perform the regression task, which means that a regression task may not be well-suited for a sequence-to-sequence approach. However the encoder-only multilingual models could solve this task with high performance. The XLM-R model could gain the highest correlation result. To provide a better comparison with previous research in this field, we conducted a 10-fold cross-validation with the XLM-R model and compared it with the baseline model (as shown in Table 2) from the research of Yang et. al [20]. Our XLM-R model achieved state-of-the-art results in the English-Hungarian QE task. In our test set, XLM-R achieved an 83.8% correlation. Refer to Figure 3 (left side) for the correlation diagram.

**Table 1.** Results of the quality estimation task.

| | Correlation | MAE | RMSE |
|---|---|---|---|
| ChatGPT | -0.0150 | 1.3072 | 1.5698 |
| mT5-small | 0.3422 | 1.0794 | 1.5059 |
| SFT-mT5-small | 0.3809 | 1.0156 | 1.4339 |
| mT5-base | 0.4579 | 0.9294 | 1.3016 |
| mBERT | 0.7358 | 0.64836 | 0.8950 |
| **XLM-R** | **0.8382** | **0.5785** | **0.8184** |

We conducted an experiment to test the ChatGPT (gpt-3.5-turbo) model [1]. The prompt template we used in this experiment is as follows:

- role: system
- content: You are a quality estimator system, which rate a given translation how good it is based on the original source sentence. Rate the translation quality between 0 to 5, where 5 is a perfect translation.
- role: user
- content: Source sentence: {src} \n Translation: {trans} \n Score:

In the prompt template above, {src} represents the source sentence, and {trans} represents the translated sentence.

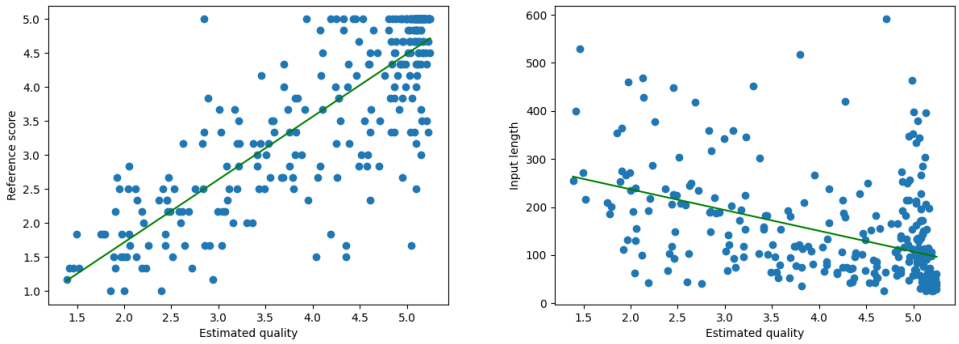


**Figure 3.** Correlation diagrams.

As we can see from the results, the text-davinci-003 model could not solve this problem. However, more prompt experiments could have been conducted.

**Table 2.** Results of the quality estimation task with 10-fold cross validation.

| | Correlation | MAE | RMSE |
|---|---|---|---|
| baseline | 0.6100 | 0.7459 | 0.9775 |
| **XLM-R** | **0.7948** | **0.6451** | **0.8898** |

In Figure 3 (right side), you can see the correlation between the estimated quality score and the input text length. The correlation is -0.4478, which means they are not correlated. The length of input does not affect the quality.

In Table 3, the results of SFT-mT5 and 'SFT-mT5 FFT' and 'SFT-mT5 RL' MT models are presented. Our 'SFT-mT5 RL' model could significantly outperform the SFT-mT5 model (>5 BLEU score). To provide a better comparision, we also fine-tuned the SFT-mT5 model using the traditional method. Fine-tuning a model on an out-of-domain corpus result in decreased performance on the original corpus. In Table 4, you can see the expected lower performance of the fine-tuned (FFT and RL) models. As you can see in Table 3 and Table 4, the RL model outperformed the FFT model in both the original corpus and the WMT09 corpus. It means that the RL model was able to adapt the new WMT09 corpus with higher performance, achieving the highest results on the new WMT09 corpus while while only slightly decreasing in performance on the original corpus.

The original 'SFT-mT5' model faced challenges such as generating outputs that were longer than the source text and containing incorrect repeated phrases (as demonstrated in Table 5): 'a hétvégén, a hétvégén' (this weekend, this weekend)). This led to high recall results but low precision. However by utilizing a human-based reward model and reinforcement learning method, we were able to correct these issues (as evidenced in Table 5) with the 'SFT-mT5 RL' model. Additionally, the 'SFT-mT5 FFT' model suffered from information loss ('She has good instincts

nonetheles') during translation.

**Table 3.** Results of FFT and RL models on WMT09.

| | BLEU | chrF-3 | chrF-6 |
|---|---|---|---|
| SFT-mT5 | 7.34 | 45.60 | 38.62 |
| SFT-mT5 FFT | 12.59 | 46.17 | 39.65 |
| SFT-mT5 RL | **12.91** | **47.02** | **40.39** |

**Table 4.** Results of FFT and RL models on the original test set.

| | BLEU | chrF-3 | chrF-6 |
|---|---|---|---|
| SFT-mT5 | 27.69 | 53.73 | 48.57 |
| SFT-mT5 FFT | 25.85 | 52.61 | 47.42 |
| SFT-mT5 RL | 26.72 | 53.32 | 48.20 |

**Table 5.** A translation sample of the different models.

| | |
|---|---|
| Source | She has good instincts nonetheless, warned Bill Clinton this weekend. |
| Reference | Bill Clinton azonban így figyelmeztetett a hétvégén: Hiba lenne Palint alulbecsülni. |
| SFT-mT5 | Ennek ellenére jó ösztönei vannak – figyelmeztette Bill Clinton a hétvégén, a hétvégén. |
| SFT-mT5 FFT | A hétvégén Bill Clinton is figyelmeztette. |
| **SFT-mT5 RL** | **Azonban jó ösztönei vannak – figyelmeztette Bill Clinton a hétvégén.** |

# 5. Conclusion

In our research, we have successfully adapted the reinforcement learning method to the machine translation task. We trained a neural quality estimation model as a reward model. Using the XLM-RoBERTa multilingual model, we achieved state-of-the-art results in Hungarian quality estimation task. For fine-tuning a language model with reinforcement learning approach, we have used an already fine-tuned mT5 model that trained for English-Hungarian machine translation task. In our experiments, we have demonstrated that reinforcement learning method can effectively enhance the performance of machine translation task by correcting the subtle errors and mistakes.

For future work, we would like to explore machine translation experiments using multilingual large language models, further extending our research in this area.

# References


[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei: *Language Models are Few-Shot Learners*, in: Advances in Neural Information Processing Systems, ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin, vol. 33, Curran Associates, Inc., 2020, pp. 1877–1901.

[2] L. Choshen, L. Fox, Z. Aizenbud, O. Abend: *On the Weaknesses of Reinforcement Learning for Neural Machine Translation*, in: International Conference on Learning Representations, 2020.

[3] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov: *Unsupervised Cross-lingual Representation Learning at Scale*, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online: Association for Computational Linguistics, July 2020, pp. 8440–8451, doi: 10.18653/v1/2020.acl-main.747.

[4] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova: *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423, url: https://aclanthology.org/N19-1423.

[5] X. Geng, Y. Zhang, S. Huang, S. Tao, H. Yang, J. Chen: *NJUNLP's Participation for the WMT2022 Quality Estimation Shared Task*, in: Proceedings of the Seventh Conference on Machine Translation (WMT), Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, Dec. 2022, pp. 615–620.

[6] D. Heo, W. Lee, B. Jung, J.-H. Lee: *Quality Estimation Using Dual Encoders with Transfer Learning*, in: Proceedings of the Sixth Conference on Machine Translation, Online: Association for Computational Linguistics, Nov. 2021, pp. 920–927.

[7] S. Kiegeland, J. Kreutzer: *Revisiting the Weaknesses of Reinforcement Learning for Neural Machine Translation*, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online: Association for Computational Linguistics, June 2021, pp. 1673–1681, doi: 10.18653/v1/2021.naacl-main.133.

[8] L. J. Laki, Z. G. Yang: *Neural machine translation for Hungarian*, Acta Linguistica Academica 69.4 (2022), pp. 501–520, doi: 10.1556/2062.2022.00576.

[9] S. Lim, H. Kim, H. Kim: *Papago's Submission for the WMT21 Quality Estimation Shared Task*, in: Proceedings of the Sixth Conference on Machine Translation, Online: Association for Computational Linguistics, Nov. 2021, pp. 935–940.

[10] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Gray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, R. Lowe: *Training language models to follow instructions with human feedback*, in: Advances in Neural Information Processing Systems, ed. by A. H. Oh, A. Agarwal, D. Belgrave, K. Cho, 2022.

[11] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. 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. F. Christiano, J. Leike, R. Lowe: *Training language models to follow instructions with human feedback*, in: Advances in Neural Information Processing Systems, ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh, vol. 35, Curran Associates, Inc., 2022, pp. 27730–27744.

[12] R. Rei, M. Treviso, N. M. Guerreiro, C. Zerva, A. C. Farinha, C. Maroti, J. G. C. de Souza, T. Glushkova, D. Alves, L. Coheur, A. Lavie, A. F. T. Martins: *CometKiwi: IST-Unbabel 2022 Submission for the Quality Estimation Shared Task*, in: Proceedings of the Seventh Conference on Machine Translation (WMT), Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, Dec. 2022, pp. 634–645.

[13] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov: *Proximal Policy Optimization Algorithms*, CoRR abs/1707.06347 (2017), arXiv: 1707.06347.

[14] N. Stiennon, L. Ouyang, J. Wu, D. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, P. F. Christiano: *Learning to summarize with human feedback*, in: Advances in Neural Information Processing Systems, ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin, vol. 33, Curran Associates, Inc., 2020, pp. 3008–3021.

[15] S. Tao, S. Chang, M. Miaomiao, H. Yang, X. Geng, S. Huang, M. Zhang, J. Guo, M. Wang, Y. Li: *CrossQE: HW-TSC 2022 Submission for the Quality Estimation Shared Task*, in: Proceedings of the Seventh Conference on Machine Translation (WMT), Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, Dec. 2022, pp. 646–652.

[16] D. Varga, P. Halacsy, A. Kornai, V. Nagy, L. Nemeth, V. Tron: *Parallel corpora for medium density languages*, in: Recent Advances in Natural Language Processing IV. Selected papers from RANLP-05, ed. by N. Nicolov, K. Bontcheva, G. Angelova, R. Mitkov, Amsterdam: Benjamins, 2007, pp. 247–258.

[17] J. Wang, K. Wang, B. Chen, Y. Zhao, W. Luo, Y. Zhang: *QEMind: Alibaba's Submission to the WMT21 Quality Estimation Shared Task*, in: Proceedings of the Sixth Conference on Machine Translation, Online: Association for Computational Linguistics, Nov. 2021, pp. 948–954.

[18] L. Wu, F. Tian, T. Qin, J. Lai, T.-Y. Liu: *A Study of Reinforcement Learning for Neural Machine Translation*, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 3612–3621, doi: 10.18653/v1/D18-1397.

[19] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel: *mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer*, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online: Association for Computational Linguistics, June 2021, pp. 483–498, doi: 10.18653/v1/2021.naacl-main.41, url: https://aclanthology.org/2021.naacl-main.41.

[20] Z. G. Yang, J. L. Laki, B. Siklósi: *HuQ: An English-Hungarian Corpus for Quality Estimation*, in: Proceedings of the LREC 2016 Workshop - Translation Evaluation: From Fragmented Tools and Data Sets to an Integrated Ecosystem (Portorož, Slovenia, May 24, 2016), 2016.

[21] C. Zerva, D. van Stigt, R. Rei, A. C. Farinha, P. Ramos, J. G. C. de Souza, T. Glushkova, M. Vera, F. Kepler, A. F. T. Martins: *IST-Unbabel 2021 Submission for the Quality Estimation Shared Task*, in: Proceedings of the Sixth Conference on Machine Translation, Online: Association for Computational Linguistics, Nov. 2021, pp. 961–972.

[22] D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, G. Irving: *Fine-Tuning Language Models from Human Preferences*, arXiv preprint arXiv:1909.08593 (2019).

[23] D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, G. Irving: *Fine-Tuning Language Models from Human Preferences*, 2020, arXiv: 1909.08593 [cs.CL].