Irena Snikhovska, Associate Professor, PhD in Linguistics
Vladyslav Korol, Bachelor student
Zhytomyr Polytechnic State University

REINFORCEMENT LEARNING FROM HUMAN FEEDBACK IN LARGE LANGUAGE MODELS

It's amazing to think what great and exciting things
people will be doing with PC's in 30 years.
(Bill Gates, circa 1989)

One of the fundamental ways how artificial intelligence, including linguistic AI known as large language models (LLMs) are produced, is the process of training whereby large amount of data is gathered and model is fitted to produce similar outputs by learning patterns in the provided data through a number of mathematician tools and techniques.

One of the key challenges in creating high-quality LLMs is ensuring it is both accurate and upholds the ethical conduct imposed on it [1, 1]. Ensuring that AI systems are unbiased [3, 23], inclusive and safe have been one of the key goals in popularising this technology, yet it has virtually nothing to do with accuracy of the generated responses. As such, the LLM could learn how to produce accurate answers, but may forgo any safety and overlook cultural differences (such as etiquette, formality levels or race titles) if it is fitted only with texts reflecting its domain of knowledge.

One more problem in ensuring artificial intelligence systems are ethical is that it takes human efforts to evaluate text for ethical norms, and the human time and efforts spent in creating AI products would make it extremely resource-consuming, expensive and unsustainable in the long run. In order to solve this problem, OpenAI has suggested the strategy of reinforcement learning from human feedback (RLHF). At its core, it relies on the idea to create a secondary AI model whose purpose is solely to filter input data (in case of LLMs, texts) based on the ethical standards, which learns and tunes itself from human feedback. This lesser AI model is called the value coach or reward model [1, 2], and coming from there, another coach is employed whose purpose is to filter texts based on their factuality, which in turn was trained by gathered data about a wide range of topics. This model is in turn referred to as the coherence coach.

As soon as the two coaches emerge and are properly tested, they can work together to bootstrap another model programmed to satisfy both coaches [1, 5; 3, 16]. As the result, the final LLM learn to be accurate generating accurate responses, but at the same time its responses follow the code of conduit laid at its foundation, and by filtering out the responses that do not satisfy one of the coaches, it iteratively develops the set of qualities and properties that the AI designers are looking for [2, 2].

The process of RLHF, although automated, plays the key role in contemporary AI development, and at its core, it stands on the live sense of morality inherent in humans. The combination of automation tools and human judgement is the efficient strategy to make AI systems more safe to interact with. For humans, RLHF and other AI trends reorient the direction of human expertise from possessing irreplaceable skills in specific domain into developing soft skills, including prompting, information extraction and summarisation that supplement domain experience, which makes it easier to produce high-quality works, learn, and otherwise boost productivity.

In conclusion, workflows like RLHF emphasise the inherent human abilities that cannot be replaced programmatically that plays a considerable role in the modern workplace. Consequently, it sets the trend in soft skills that are becoming more and more advantageous and necessary for humans in the future that are becoming the pillar of AI nowadays and in the upcoming years.

REFERENCES

1. Ziegler, M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., & Irving, G. (2019). Fine-tuning language models from human preferences. arXiv preprint arXiv:1909.08593.

2. Zhu, B., Sharma, H., Frujeri, F. V., Dong, Sh., Zhu, Ch, Jordan, M.I., Jiao, J. (2023). Fine-tuning language models with advantage-induced policy alignment. arXiv preprint arXiv:2306.02231.

3. Bai, Y. et al. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv preprint arXiv:2204.05862.