DOI: https://doi.org/10.36910/6775-2524-0560-2025-58-14
UDC 004.8:004.42
Polyakovska Nataliya, Master, Data Scientist
https://orcid.org/0000-0003-2855-7970
SoftServe, Lviv, Ukraine

TOWARDS A PRACTICAL FRAMEWORK FOR LLMOPS: UNDERSTANDING AND BUILDING OPERATIONAL EXCELLENCE FOR LARGE LANGUAGE MODELS

Polyakovska N. Towards a Practical Framework for LLMOps: Understanding and Building Operational Excellence for Large Language Models. Effective life cycle management of large language models (LLMs) ensures their reliability and adaptability in production environments. The study's relevance is due to the rapid growth in the use of large language models in various industries, accompanied by challenges such as high computational requirements, risks of generating false results ("hallucinations"), and algorithmic bias. It is established that traditional MLOps methods do not provide adequate quality control and scalability, which requires the development of specialized operational approaches for LLM. The study aims to form an operational framework for LLMOps that covers all stages of the life cycle of large language models, from data processing to real-time monitoring and support. The research methods are based on an interdisciplinary analysis of current practices of implementing large-scale language models, identifying problematic aspects of their use, and developing practical recommendations for effective system management. The main stages of the operational framework are identified, including data preparation, model development and deployment, and control of results at the monitoring and support stages. The most critical aspects are integrating multi-level monitoring, compliance with ethical standards, and introducing automated algorithms to reduce the frequency of false positives. The study's results confirm that the proposed operational framework increases the reliability of language models in high-load environments and ensures their adaptation to dynamic changes in the query structure. The use of distributed computing methods, resource optimization, and post-processing verification helps minimize the risks associated with the performance and accuracy of model responses. Simulation stress tests are recommended to check the system's stability during peak load periods. Conclusions. The paper emphasizes the importance of continuously auditing models' operations to ensure their transparency and compliance with regulatory requirements. Prospects for further research include optimizing retraining processes and implementing energy-efficient computing resource management methods.
Keywords: LLMOps, large language models, MLOps, operational framework, ethical artificial intelligence, generative AI life cycle, inference optimization.

Поляковська Н.О. На шляху до практичної рамки LLMOps: розуміння та формування операційної досконалості для великих мовних моделей. Ефективне управління життєвим циклом великих мовних моделей (LLM) є ключовим аспектом забезпечення їхньої надійності та адаптивності у виробничих середовищах. Актуальність дослідження зумовлена стрімким зростанням використання великих мовних моделей у різних галузях, що супроводжується викликами, такими як високі обчислювальні потреби, ризики генерації хибних результатів («галюцинацій») та алгоритмічна упередженість. Встановлено, що традиційні методи MLOps не забезпечують належного рівня контролю якості та масштабованості, що вимагає розробки спеціалізованих операційних підходів для LLM. Мета дослідження – формування операційної рамки LLMOps, яка охоплює всі етапи життєвого циклу великих мовних моделей: від обробки даних до моніторингу та підтримки в реальному часі. Методи дослідження базуються на міждисциплінарному аналізі сучасних практик упровадження великих мовних моделей, виявленні проблемних аспектів їхнього використання та розробці практичних рекомендацій для ефективного управління системами. Визначено основні етапи операційної рамки, зокрема підготовку даних, розробку та розгортання моделі, а також контроль результатів на етапах моніторингу й підтримки. Найбільш критичними аспектами є інтеграція багаторівневого моніторингу, дотримання етичних стандартів і впровадження автоматизованих алгоритмів для зменшення частоти помилкових відповідей. Результати дослідження підтверджують, що запропонована операційна рамка підвищує надійність роботи мовних моделей у високонавантажених середовищах і забезпечує їхню адаптацію до динамічних змін у структурі запитів. Використання методів розподілених обчислень, оптимізації ресурсів і перевірки результатів на етапі постобробки сприяє мінімізації ризиків, пов'язаних із продуктивністю та точністю відповідей моделі. Рекомендовано впроваджувати симуляційні стрес-тести для перевірки стійкості системи під час пікових періодів навантаження. Висновки. У роботі наголошено на важливості постійного аудиту роботи моделей для забезпечення їхньої прозорості та відповідності нормативним вимогам. Перспективи подальших досліджень стосуються оптимізації процесів повторного навчання й упровадження методів енергоефективного управління обчислювальними ресурсами.
Ключові слова: LLMOps, великі мовні моделі, MLOps, операційні рамки, етичний штучний інтелект, життєвий цикл генеративного ШІ, оптимізація інференції.

Problem statement. The problem of practical implementation and operation of large language models lies in the complexity of managing their life cycle, which includes development, deployment, monitoring, and support in real time. High computational requirements, high dependence on high-quality data, and sensitivity to changes in the operating environment characterize these models. Traditional MLOps practices do not fully meet the specific needs of working with large language models, as such systems require specialized approaches to assess the relevance of results to user requests, prevent model "hallucinations," and ensure ethical compliance. At the same time, using large language models in critical

© Polyakovska N.

areas such as medicine, education, and finance requires the development of a comprehensive operational framework that covers both technical and ethical aspects.

The scientific significance lies in the creation of an interdisciplinary approach that combines knowledge from the fields of distributed computing, machine learning, large language models and cybersecurity. The practical value of such a framework lies in the possibility of scalable and controlled application of language models to process large amounts of information while minimizing the risks associated with data privacy violations or algorithmic bias. Developing a practical LLMOps framework will ensure the effective integration of models into production environments and promote the implementation of ethical standards of artificial intelligence in real-world use cases.

Analysis of the latest research and publications. The operational management of large language models (LLM) covers key aspects such as model lifecycle optimization, scalability, quality control, and ethical challenges associated with generative algorithms.

The development of large language models began with the creation of innovative architectures and training methods. In particular, A. Vaswani et al. proposed a transformer architecture that replaced recurrent and convolutional neural networks, providing high performance and parallelism of computation. This solution significantly increased the speed of processing long text sequences and laid the foundation for modern LLMs [1].

The next significant step was the study by J. Kaplan et al., who established the dependence of model performance on the number of parameters and the amount of training data. The conclusions of this work allowed developers to create large-scale models such as GPT, which demonstrate high efficiency in generative tasks [2].

A study by T.B. Brown et al. revealed the potential of few-shot learning, which allows models to achieve high performance with only a few instruction examples. This opens up opportunities for effective adaptation of models without the need for complete retraining [3].

In turn, L. Ouyang demonstrated the benefits of a reinforcement learning approach based on human feedback, which provides more accurate and relevant answers. This approach allows models to better respond to real user requests and reduce the number of false answers [4].

The information security challenges in large-scale systems are highly relevant, especially with the rise of large language models (LLMs). A study by Yifan Yao et al. emphasized LLMs' dual role in enhancing security, including code vulnerability detection, while also posing risks like user-level attacks and information leakage. Proper security measures and further research are essential to address these vulnerabilities [5]

Regarding adapting LLM to specific tasks, E.J. Hu proposed the LoRA (Low-Rank Adaptation) method, which allows for effective fine-tuning of models with minimal computational costs. This technique greatly facilitates the integration of models into workflows without the need for complete retraining [6].

M. Howard and G. Quattrocchi emphasized the importance of infrastructure automation to support scalable systems. They explored the benefits of using the Infrastructure-as-Code approach, which simplifies the management and deployment of large computing systems [7].

F. Zeng's research was devoted to distributed model training. He proposed optimizing data exchange between cluster nodes to reduce delays and improve system performance when training large models [8].

E. Frantar made a significant contribution to inference optimization by introducing the GPTQ method for quantized model training. This method can significantly reduce the amount of memory while maintaining high-generation quality [9].

M. Chen's work was devoted to speeding up inference through the use of optimized computing units. The proposed approach reduces the query execution time in production environments [10].

Evaluation plays a critical role in the quality control of language models. Y.-T. Lin and Y.-N. Chen has developed the LLM-Eval metric, which provides a multidimensional assessment of answers' coherence and actual accuracy. This study demonstrated the need to move from the traditional BLEU and ROUGE metrics to more comprehensive metrics for generative models [11].

The study by K. Kenthapadi analyzed real-time model monitoring systems that allow for quick detection of anomalies in LLM performance. The author emphasized the importance of adaptive monitoring systems to reduce the risk of decreased performance and response quality [12].

In the context of ethical challenges, L. Weidinger et al. have investigated the social risks associated with algorithmic bias and misinformation that models can generate. The authors emphasize the importance of developing ethical standards to prevent the harmful effects of LLM [13].

© Polyakovska N.

C. Borchers considered the problem of bias and "hallucinations," proposing methods of adaptive fine-tuning and correction of hints to reduce false generations. These approaches minimize the risk of false answers and increase system user confidence [14].

T. Rebedea developed practical solutions for the security of LLM applications. He introduced the concept of "guardrails"—limiters that control the output content of models. His research shows that software guardrails significantly reduce the risks of malicious generation and increase the level of security [15].

H. Inan described the integration of security modules into systems of user interaction with models. The authors emphasized the importance of such modules to ensure transparency and minimize the risk of false answers in critical scenarios [16].

The research analysis shows that significant progress in the fields of LLM and MLOps has created the prerequisites for the formation of a separate LLMOps field that combines distributed learning, adaptive tuning, and automated monitoring approaches. The main areas of research include the development of model architectures, computational optimization, response quality assurance, and minimization of bias and "hallucinations."

Highlighting previously unsolved parts of the problem. Despite the development of MLOps, large language models (LLMs) have specific needs that remain insufficiently explored. In particular, this concerns processing large amounts of unstructured data, optimizing distributed computing, and ensuring speed without losing quality. Monitoring systems without high computational costs have not been developed to detect delays and "hallucinations" in real time.

There are also unresolved issues of eliminating algorithmic bias and integrating automatic adjustments depending on the specifics of queries. Existing solutions are often focused on technical indicators but do not consider users' needs in different industries.

The proposed operating model covers all stages of the model lifecycle, including multi-level monitoring, automatic anomaly resolution, and ethical compliance. Practical recommendations for industry scenarios increase the reliability and scalability of LLMs, facilitating their effective integration into production environments.

The study aims to create a holistic LLMOps framework for effective life cycle management of large language models that considers performance, ethics, and compliance with production requirements.

Objectives of the study:

1. Analyze current approaches to MLOps and identify specific requirements for working with large language models, focusing on data processing, distributed computing, and building an operational model that covers all life cycle stages - from data preparation to supporting the model's functioning in real time.

2. Investigate methods for monitoring models in production environments focusing on detecting delays, accurate model outputs, and ethical compliance, and develop approaches to address critical risks of LLMs, such as bias and hallucinations, by integrating quality control and performance optimization mechanisms.

3. Provide practical recommendations for implementing the LLMOps framework in real-world use cases to ensure models' reliability, scalability, and adaptability in highly loaded environments.

Summary of the primary material. In modern research and practice, MLOps is considered a comprehensive approach to automating machine learning models' development, deployment, and operation. However, this approach is insufficient for large-scale language models (LLMs), as they have unique needs and challenges. One of the main differences is the scale of data processing and the need for distributed computing. Large language models operate on vast amounts of textual data that must be thoroughly cleaned, anonymized, and structured. In addition, LLM fine-tuning and inference processes require significant computational resources, making them unsuitable for standard MLOps approaches without high-performance clusters and specialized hardware. Accordingly, data processing and infrastructure management become critical steps in working with LLM (Table 1).

Table 1 – Comparison of traditional MLOps and specific requirements for LLMOps in the context of data processing and computing resources

Aspect	Traditional MLOps	Requirements for LLMOps
Data processing	Use of structured data sets with basic cleaning and normalization methods	Working with large volumes of unstructured data that requires anonymization, synthetic additions and classification
Data storage	Conventional relational databases or standard file systems	Use vector databases for quick access and search of relevant text fragments

© Polyakovska N.

Computing resources	Standard processors (CPUs) for most stages of training and inference	The need for graphics processing units (GPUs) and cluster computing to speed up training and reduce latency
Data volumes	Limited by the number of parameters and the number of datasets	Terabytes of text data to customize models and process user requests
Scalability	Local computing or small clusters	Scalable infrastructure to support distributed learning and inferences
Tools	Use of standard libraries and platforms such as TensorFlow, MLflow	Using specialized frameworks for LLM optimization, such as Hugging Face Transformers

Source: compiled by the author based on [1; 2; 8; 9].

Table 1 illustrates the key differences between traditional MLOps and LLMOps requirements. For large language models, data processing requires working with large amounts of text, often including anonymization and synthetic augmentation. Vector databases are used for data storage and faster access to information. It is also critical to have a high-performance infrastructure to ensure efficient data processing and support for parallel computing.

The life cycle of large language models (LLMs) covers all stages - from data preparation to real-time model maintenance and updating. The peculiarity of the LLM life cycle is its iterative nature and constant dependence on environmental changes. Creating an operational model involves considering such aspects as processing large volumes of unstructured data, efficient deployment and customization of models, and regular monitoring and support to ensure that models are adapted to new requests and conditions. Integrating such approaches avoids performance degradation and improves the accuracy of model responses while minimizing the risk of bias or "hallucinations." The main stages of the LLM lifecycle can be represented in the form of an operational model that considers the key tasks at each stage (Fig. 1 or Table 2, as appropriate).

Table 2 – Main stages of the life cycle of large language models and their key tasks

Life cycle stage	**Process content**	**Key tasks**
Data processing	Data collection, cleaning and anonymization using structuring algorithms and synthetic augmentation methods	Formation and maintenance of an up-to-date dataset that meets the requirements of confidentiality and relevance
Model development	Fine-tune models or use prompt engineering to adapt to specific tasks	Optimize model parameters, evaluate performance using specialized metrics, and prepare for deployment
Deployment	Integrate the model into the production environment with a scalable infrastructure	Containerize the model, ensure uninterrupted operation, and develop scaling strategies
Monitoring	Continuous data collection on model performance, analysis of performance metrics and response quality	Detect anomalies, monitor compliance with ethical standards, evaluate latency and throughput
Support and updates	Adaptation of the model to new working conditions, integration of feedback and re-training	Re-fine-tuning or updating the parameters of the prompts, integrating new data to improve accuracy and reliability

Source: compiled by the author based on [4; 5; 6; 7; 11].

© Polyakovska N.

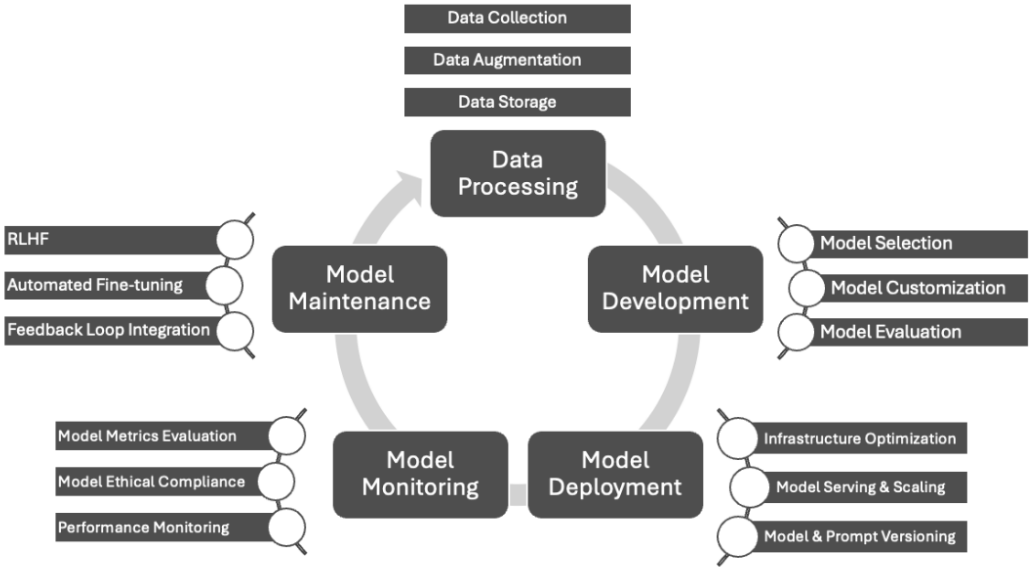

Figure 1. Main stages and components of the proposed LLMOps framework
Source: authors' own development

The proposed operational model for managing the life cycle of large language models (LLMs) consists of several interrelated stages that form a closed cycle. Such a cyclic structure ensures iterative processes and continuous improvement of the models in the production environment.

At the data processing stage, a comprehensive data collection, cleaning, and anonymization process is carried out to form a high-quality training dataset. Particular attention is paid to structuring unstructured data and applying synthetic augmentation methods to increase the volume and relevance of data without risking loss of model accuracy.

Model development involves adapting the LLM to specific tasks through fine-tuning or prompt engineering. This stage involves using optimization techniques such as LoRA or PEFT to reduce computational costs and improve efficiency. The quality of the model is evaluated using specialized metrics such as perplexity, answer consistency and query processing speed.

The model is deployed through containerization and integration into the production environment, considering the needs of scaling and ensuring stability even under high loads. At this stage, ensuring low latency and continuity of request service is important.

Monitoring the performance and quality of the model is critical for the timely detection of anomalies, including model "hallucinations" or biased results. For this purpose, modern metrics collection systems and automated systems for notifying about critical deviations are used. If degradation of results is detected, re-training or updating of hints can be initiated.

The model maintenance and updating phase involves implementing feedback and using new data to adapt the model to changing environmental conditions. This may include model re-fitting or integration of new approaches to data processing and parameter optimization. This ensures that model responses remain relevant and accurate even in dynamic use cases. Thus, the proposed LLMOps framework demonstrates not only the structured management of the model's life cycle but also scientifically substantiates the feasibility of a cyclical organization to ensure sustainability, reliability, and compliance with ethical standards. Monitoring large language models (LLMs) in production environments is key to ensuring their stability, performance, and compliance with ethical standards. Given the scale and complexity of such models, monitoring goes beyond standard metrics such as accuracy and performance. It includes an assessment of response latency, error rate, response coherence, and detection of "hallucinations." Particular attention is paid to compliance with ethical principles, as LLMs can show biased responses or generate inaccurate information, potentially creating user risks. Effective model monitoring requires the integration of automatic notification and feedback analysis systems, which allows for tracking critical deviations and initiating corrective measures in real-time (Table 3).

Table 3 – Key metrics for monitoring large language models in production environments

Monitoring metrics	Content of the indicator	Importance for model management

© Polyakovska N.

Latency of response	Time from receiving a request to receiving a response from the model	Allows you to assess whether the model meets performance requirements and how efficiently the infrastructure is used
Consistency of the answer	The level of logical integrity and meaningful sequence of answers	Helps determine whether the model generates clear and correct texts without contradictions
User feedback	User-provided ratings, reactions or qualitative feedback on whether the response was accurate, useful, or met their expectations.	Helps evaluate the model's performance from the user's perspective, measure satisfaction, and identify areas for improvement in accuracy, relevance, and usability.
Percentage of "hallucinations"	Share of responses containing false or fictitious information	Determines the need for additional training stages to increase the reliability of answers
Compliance with ethical standards	Presence of prejudice or discriminatory statements in the answers	Helps ensure that the model meets ethical standards and regulatory requirements

Source: compiled by the author based on [11; 13; 14; 15; 16].

LLM models in production environments are monitored in real time using automated metrics collection and analysis systems. For example, to assess response latency, tools monitor the speed of request processing and notify you when the acceptable limits are exceeded. High latency may indicate an overload of computing resources or the need to use additional GPU clusters. The consistency of the answer is checked both by automatic algorithms and with the involvement of a human to assess the context and content of answers in complex queries.

The rate of "hallucinations" is critical to maintaining user confidence in the system. If the model regularly generates false or fictitious answers, re-tuning or updating the prompt engineering is triggered to reduce the error rate. Ethical compliance is monitored by analyzing the content of answers for discriminatory statements or biases. To prevent violations, mechanisms for automatically blocking certain answers and regular audits by artificial intelligence experts are implemented.

Eliminating critical risks associated with the operation of large-scale language models (LLMs), such as bias and the generation of false information ("hallucinations"), is a priority to ensure their reliability and compliance with ethical standards [13]. One of the key risks is model bias, which arises from imbalances in the training data or insufficient control over the generation of responses [5]. This factor can affect both the content of the answers and their context, which poses a potential threat of incorrect information, especially in sensitive industries such as medicine, education, and law enforcement. Another significant risk is model "hallucinations" – cases where the model generates confident but false or fictitious statements that are not based on factual data [4].

To minimize such risks, it is important to implement comprehensive quality control mechanisms and optimize model performance. The selection and cleaning of training data help reduce the number of potentially biased or unreliable fragments that affect model performance. When preparing datasets, classification, and sample balancing methods are used to ensure that different points of view are represented and that certain categories of data are not dominated. In addition, one of the most effective methods is the introduction of special filters at the stage of post-processing the results, which allows the model's responses to be checked for compliance with ethical standards and the absence of potentially dangerous statements [15].

An important component is monitoring and automatic intervention systems allowing for real-time deviation detection. In particular, models can be configured to interrupt or block responses beyond the defined ethical framework or show signs of "hallucination." Using metrics, such as consistency of answers and checking for inconsistent statements, allows for maintaining high accuracy and logical integrity of the generated texts [11]. In addition, integrating models with verified knowledge databases helps narrow the generation space to factually correct data, reducing the likelihood of false statements. In practice, these approaches can be implemented as automated response evaluation modules that apply several verification algorithms to analyze the relevance of model results to current queries and their validity [12]. Such modules can operate in real time, creating multi-level protection against the impact of inaccurate data or bias risks. An additional element of control is a regular audit of model results by experts who analyze cases of false answers and adjust settings to maintain maximum transparency of the process. Implementing these methods and tools can significantly reduce critical risks and ensure the sustainability of models in various use cases.

© Polyakovska N.

Implementing the LLMOps framework in real-world scenarios should consider the need to adapt to a dynamic environment and the specifics of application tasks in various industries. One of the key aspects is to create a scalable implementation strategy with the gradual connection of additional modules and stages of verification of the system's compliance with the practical needs of users. This involves integrating the system at a pilot level to test workloads and identify potential problems before launching at full capacity. In real-world production environments, implementing the framework also involves introducing a system for regularly collecting feedback from end users to update models in response to specific requests. An important component is simulation testing, which allows the system to be tested under stressful conditions to ensure its uninterrupted operation during peak loads. This approach avoids errors in critical situations and increases the system's reliability. Thus, the implementation of the LLMOps framework becomes an effective solution for creating systems with a high level of adaptability and scalability. Integrating pilot testing, personalization, and simulation tests makes it possible to achieve stable operation of language models in real production environments while minimizing risks.

Conclusions and Prospects for Further Research. The article establishes that effective life cycle management of large language models (LLMs) requires specialized approaches for data processing, model tuning, monitoring, and adaptation to environmental changes. The main problems in this context are high computational resource requirements, the risk of "hallucinations," and algorithmic bias, which can reduce the accuracy and credibility of the results. The analysis of practical scenarios confirmed the need to implement multi-level monitoring and quality control systems that ensure timely detection of errors and support the stable operation of models.

The main recommendations are to use distributed computing technologies to scale systems, integrate mechanisms for automated response correction, and comply with ethical standards through regular audits and inspections. The proposed LLMOps operating framework increases the reliability of models, ensures their adaptability to dynamic conditions, and minimizes the risks of bias by integrating sources of verified knowledge. Prospects for further research are related to the optimization of relearning processes and the implementation of efficient energy consumption methods in scalable computing systems. It is also advisable to develop tools for assessing models' coherence and ethical compliance in different linguistic and cultural contexts, which will contribute to the formation of universal practices to maintain transparency and accountability of language models on a global scale.

References:

1. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser Ł., Polosukhin I. Attention Is All You Need. *Proceedings of the 31st International Conference on Neural Information Processing Systems, Curran Associates Inc*. 2017. P. 6000-6010. URL: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html (date of access: 02.01.2025).
2. Kaplan J., McCandlish S., Henighan T., Brown T. Scaling Laws for Neural Language Models. *ArXiv preprint.* arXiv:2001.08361. 2020. URL: https://arxiv.org/abs/2001.08361 (date of access: 02.01.2025).
3. Brown T.B., Mann B., Ryder N., Subbiah M., Kaplan J., Dhariwal P., Neelakantan A. et al. Language Models Are Few-Shot Learners. *Proceedings of the 34th International Conference on Neural Information Processing Systems, Curran Associates Inc.* 2020. P. 1877-1901. https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html (date of access: 02.01.2025).
4. Ouyang L., Wu J., Jiang X., Almeida D., Wainwright C., Mishkin P., Zhang C. et al. Training Language Models to Follow Instructions with Human Feedback. *Proceedings of the 36th International Conference on Neural Information Processing Systems, Curran Associates Inc.* 2024. P. 27730-27744. URL: https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html (date of access: 02.01.2025).
5. A survey on Large Language Model (LLM) security and privacy: The Good, The Bad, and The Ugly / Y. Yao et al. *High-Confidence Computing*. 2024. P. 100211. URL: https://doi.org/10.1016/j.hcc.2024.100211 (date of access: 16.01.2025).
6. Hu E.J., Shen D., Wallis P., Allen-Zhu Z., Li Y., Wang L., Chen W. LoRA: Low-Rank Adaptation of Large Language Models. *OpenReview.net: Website.* 2021. URL: https://openreview.net/forum?id=nZeVKeeFYf9 (date of access: 02.01.2025).
7. Howard M. Terraform - Automating Infrastructure as a Service. *ArXiv preprint.* arXiv:2205.10676. 2022. URL: https://doi.org/10.48550/arXiv.2205.10676 (date of access: 02.01.2025).
8. Zeng F., Zhao Y., Zhou X., Luo L. Distributed Training of Large Language Models. 29th International Conference on Parallel and Distributed Systems (ICPADS). *IEEE Xplore.* 2023. P. 840-847. DOI: https://doi.org/10.1109/ICPADS60453.2023.00126 (date of access: 02.01.2025).

© Polyakovska N.

9. Frantar E., Stock P., LeCun Y. GPTQ: Accurate Post-Training Quantization for Generative Pre-Trained Transformers. *ArXiv preprint*. arXiv:2210.17323. 2023. URL: https://doi.org/10.48550/arXiv.2210.17323 (date of access: 02.01.2025).
10. Zaharia M., Chen M., Ghodsi A., Jordan M. Accelerating the Machine Learning Lifecycle with MLflow. *IEEE Data Eng. Bull*. 2023. 1. P. 39-45. URL: https://people.eecs.berkeley.edu/~alig/papers/mlflow.pdf (date of access: 02.01.2025).
11. Lin Y.-T., Chen Y.-N. LLM-Eval: Unified Multi-Dimensional Automatic Evaluation for Open-Domain Conversations. *ArXiv preprint*. arXiv:2305.13711. 2023. URL: https://doi.org/10.48550/arXiv.2305.13711 (date of access: 02.01.2025).
12. Model Monitoring in Practice / K. Kenthapadi et al. *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Washington DC USA. New York, NY, USA, 2022. URL: https://doi.org/10.1145/3534678.3542617 (date of access: 15.01.2025).
13. Weidinger L., Uesato J., Biegler F., van den Driessche G., O'Brien C., Kramár J., McAleese N. Ethical and Social Risks of Harm from Language Models. *ArXiv preprint*. arXiv:2112.04359. 2021. URL: https://doi.org/10.48550/arXiv.2112.04359 (date of access: 02.01.2025).
14. Looking for a Handsome Carpenter! Debiasing GPT-3 Job Advertisements / C. Borchers et al. *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, Seattle, Washington. Stroudsburg, PA, USA, 2022. URL: https://doi.org/10.18653/v1/2022.gebnlp-1.22 (date of access: 15.01.2025).
15. Rebedea T., Berengueres J., Dinic C., Safeguard L. NeMo Guardrails: A Toolkit for Controllable and Safe LLM Applications with Programmable Rails. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations.* Association for Computational Linguistics. 2023. P. 431-445. DOI: https://doi.org/10.18653/v1/2023.emnlp-demo.40 (date of access: 02.01.2025).
16. Inan H., Olmez G., Kaya M. Llama Guard: LLM-Based Input-Output Safeguard for Human-AI Conversations. *ArXiv preprint*. arXiv:2312.06674. 2023. URL: https://doi.org/10.48550/arXiv.2312.06674 (date of access: 02.01.2025).

© Polyakovska N.