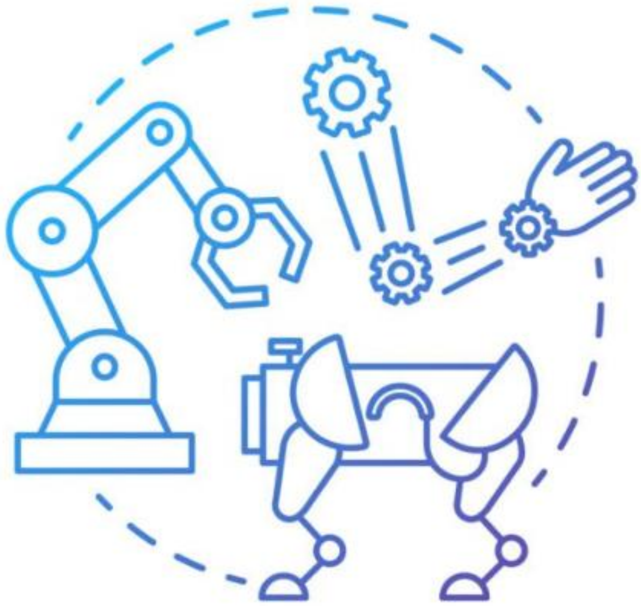
Міністерство освіти і науки України
Харківський національний університет радіоелектроніки
кафедра комп’ютерно-інтегрованих технологій, автоматизації, робототехніки та безпекової інженерії (КІТАРБІ)



МАТЕРІАЛИ

ІІІ Всеукраїнської конференції
«Комп’ютерно-інтегрованих технологій, автоматизації та робототехніки»
(Computer-integrated technologies, automation and robotics)
CITAR`26
14-15 травня 2026

[електронне видання]

Харків 2026

УДК: 005:004.896:62-65:338.3

Комп'ютерно-інтегрованих технологій, автоматизації та робототехніки 2026: матеріали III -ої Всеукраїнськой конференції, Харків, 14-15 травня 2026.: тези доповідей / [редкол. І.Ш. Невлюдов (відповідальний редактор)].-Харків: [електронний друк], 2026. – 97 с.

У збірник включені тези доповідей, які присвячені сучасним автоматизованим технологіям Industry 4.0 та їх впровадження; інформаційні управляючі системи технологічного призначення; математичні методи в системах автоматизації; розробка та програмування в робототехніці; штучний інтелект та машинне навчання в автоматизації; інтеграція технологій у виробництві та промисловості; сенсорні технології та взаємодія людини з роботами в Industry 5.0; ефективність використання роботизованих систем у виробництві; етика та правові аспекти в робототехніці; Інтернет речей та Інтегровані системи в комп'ютерно-інтегрованих технологіях, автоматизації та робототехніки; технологічні виклики та інновації у світі робототехніки.

Computer-integrated technologies, automation and robotics 2026: Proceedings of III st All-Ukrainian Conference, Kharkiv, May 14-15, 2026: Thesises of Reports / [Ed. I.Sh. Nevlyudov (chief editor).] .-Kharkiv .: [electronic version], 2026. - 97 p.

The collection includes abstracts devoted to modern automated technologies of Industry 4.0 and their implementation; information control systems for technological purposes; mathematical methods in automation systems; development and programming in robotics; artificial intelligence and machine learning in automation; integration of technologies in production and industry; sensor technologies and human interaction with robots in Industry 5.0; efficiency of using robotic systems in production; ethics and legal aspects in robotics; Internet of Things.

ЗМІСТ

METHODS FOR IMPROVING THE ENERGY EFFICIENCY OF SMALL LANGUAGE MODELS FOR AUTONOMOUS ROBOTICS

A. S. Andreiev, S. V. Sotnik
Kharkiv National University of Radio Electronics
Ukraine, 61166, Kharkiv, Nauky av., 14
E-mail: anton.andreiev@nure.ua

Abstract: This work investigates methods for increasing the energy efficiency of small language models in autonomous mobile robotics applications. A comparative analysis of the relationship between the computational complexity of models, energy consumption, and autonomous operation time of mobile robots is conducted. Quantization, pruning, and knowledge distillation methods are considered as key approaches to reducing the computational load. Recommendations are given for choosing the optimal model compression strategy depending on the requirements of a specific robotics task and hardware platform.

Keywords: energy efficiency, small language models, mobile robotics, neural network compression methods, battery life.

АНАЛІЗ МОЖЛИВИХ КІБЕРЗАГРОЗ МЕРЕЖЕВОЇ БЕЗПЕКИ

А. С. Андреєв, С. В. Сотник
Харківський національний університет радіоелектроніки
Україна, 61166, Харків, пр. Науки 14
E-mail: anton.andreiev@nure.ua

***Анотація*:** Ця робота досліджує методи підвищення енергоефективності малих мовних моделей у застосуваннях автономної мобільної робототехніки. Проведено порівняльний аналіз взаємозв'язку між обчислювальною складністю моделей, енергоспоживанням та часом автономної роботи мобільних роботів. Розглядаються методи квантування, обрізання та дистиляції знань як ключові підходи до зменшення обчислювального навантаження. Надані рекомендації щодо вибору оптимальної стратегії стиснення моделей залежно від вимог конкретного завдання робототехніки та апаратної платформи.

***Ключові слова*:** енергоефективність, малі мовні моделі, мобільна робототехніка, методи стиснення нейромереж, час автономної роботи.

RELEVANCE OF THE WORK. The rapid development of robotics and its integration into various areas of human activity have been reflected in previous studies [1-10]. Automation of technological processes is an integral part of modern industrial production and the development of cyber-physical systems [11-16]. The integration of large language models (LLMs) into mobile robotic systems opens up new possibilities for natural language control, mission planning, and human-robot interaction. However, the direct use of LLM on onboard computing platforms with limited energy resources is impossible due to excessive memory and computing power requirements. For example, a model with 7 billion parameters in FP32 format requires about 28 GB of RAM, which is critically incompatible with typical embedded platforms (NVIDIA Jetson, Raspberry Pi, etc.).

Small language models (SLMs), containing between 100 million and 3 billion parameters, are a promising compromise between functionality and resource intensity. However, even for SLMs, the issue of minimizing energy consumption remains relevant, as it directly determines the robot's battery life. Research into the relationship between the computational complexity of a model (number of parameters, number of floating point operations – FLOP), energy consumption, and battery life is a relevant scientific and applied problem.

The purpose of this work is to analyze and compare methods for improving the energy efficiency of SLM for autonomous mobile robots. To achieve this goal, the following tasks have been defined: review of the main methods of neural network compression; analysis of their impact on computational complexity and energy consumption; comparative assessment of the relationship between model characteristics and the autonomous operation time of mobile platforms.

MATERIALS AND RESEARCH RESULTS. The computational complexity of language models is determined by the number of parameters and multiply-accumulate (MAC) operations. For transformer architectures, the computational complexity of a forward pass scales as $O(n^2 \cdot d)$, where n is the context length and d is the dimension of the hidden layer. This quadratic dependence makes reducing the model size critical for deployment on onboard platforms.

Quantization is a method of reducing the bit depth of weights and model activations. The transition from FP32 to INT8 reduces memory usage by a factor of 4 and increases performance on hardware accelerators that support integer arithmetic. Quantization to INT4 allows for 8x compression, but may be accompanied by accuracy degradation for complex planning tasks. GPTQ and AWQ methods demonstrate acceptable quality even with 4-bit quantization due to layer-level error compensation.

Pruning (thinning) involves removing insignificant weights or entire neurons/attention heads. Structural pruning, which removes entire channels or layers, allows for real acceleration on standard hardware without special sparse matrix multiplication libraries. Removing 30 – 50 % of attention heads in transformer models is usually accompanied by a loss of quality of no more than 5 % while reducing energy consumption by 20 – 35 %.

Knowledge distillation is an approach in which a small «student» model learns to reproduce the behavior of a large «teacher» model. This allows a compact model to approach the quality of much larger counterparts. For example, the Phi-2 model (2,7B parameters), trained using knowledge distillation, demonstrates results comparable to LLaMA-2 13B on a number of benchmarks, with almost five times less energy consumption.

Table 1 presents a comparative analysis of the relationship between model compression characteristics, their computational complexity, and practical impact on the autonomous operation time of a mobile robot using a typical 100 Wh onboard battery and an NVIDIA Jetson Orin NX 16GB computing platform.

Table 1 – Comparison of SLM energy efficiency improvement methods and their impact on the autonomy of mobile robots

Method / characteristic	Quantization (INT8)	Quantization (INT4)	Structural pruning (40%)	Distillation of knowledge
Memory compression	×4	×8	×1,5 – 2	×4 – 10
Reduction in FLOP count	×2 – 4	×4 – 8	×1,5 – 2,5	×4 – 12
Average energy consumption (W)	8 – 12	5 – 8	10 – 15	4 – 10
Battery life (hours)	8 – 12	12 – 20	6 – 10	10 – 25
Quality degradatio (relative)	1 – 3 %	3 – 8 %	3 – 7 %	<3 %
The need for retraining	Minimal	Partial	Significant	Significant
Hardware support	Wide	Limited	Wide	Wide

Comparison of SLM energy efficiency methods and their impact on mobile robot autonomy in Fig. 1. Figure 1 shows that power consumption (W) is a stronger predictor of battery life than memory compression ratio alone: structural pruning achieves only ×1,5 – 2 memory reduction yet draws up to 15 W, resulting in the shortest runtime among all methods. Conversely, knowledge distillation combines the widest compression range (×4 – 10) with the lowest power floor (4 W), pushing the upper autonomy bound to 25 h.

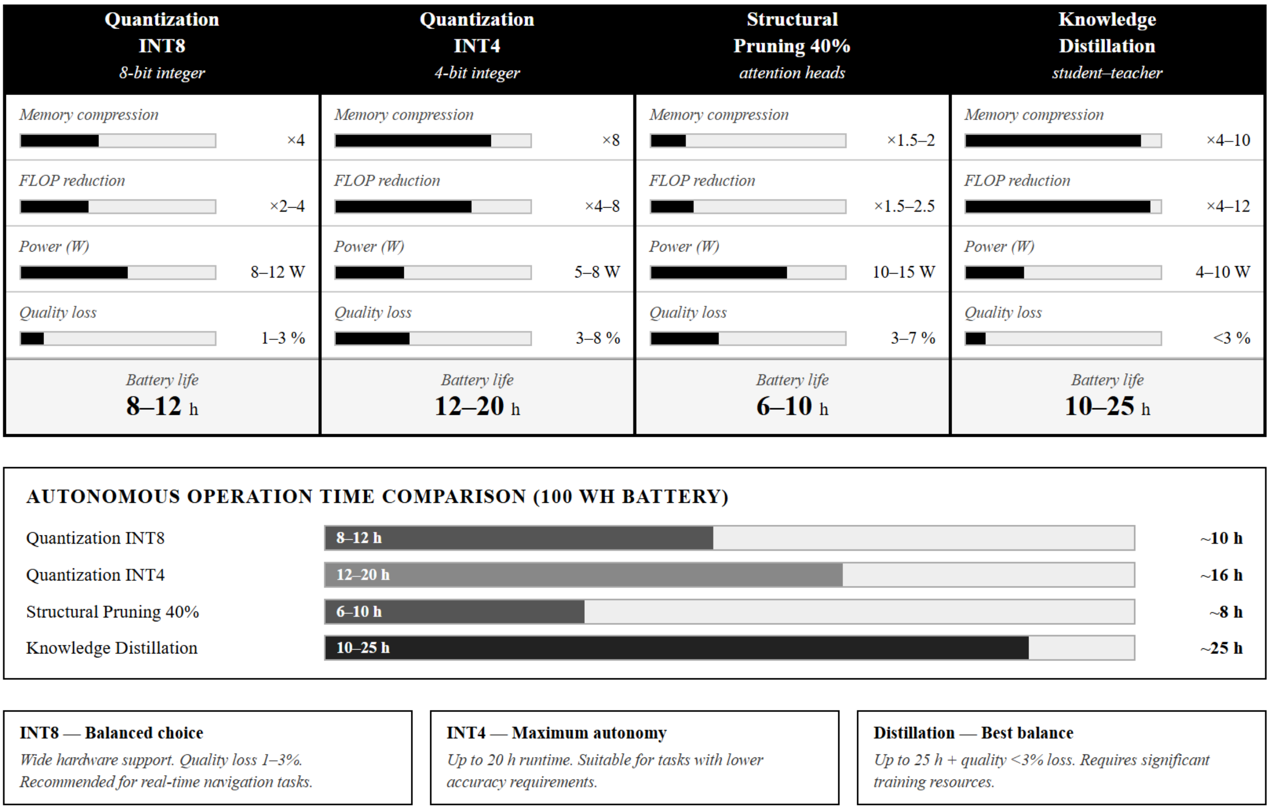


Figure 1. Comparison of SLM energy efficiency methods and their impact on mobile robot autonomy

The gap between INT4 quantization and distillation – both achieving comparable memory footprints – is therefore explained primarily by differences in residual computational overhead rather than model size.

The choice of compression strategy should be guided by the specific requirements of the robotic task and the available hardware platform. For systems deployed on platforms with broad INT8 accelerator support – such as NVIDIA Jetson Orin – where real-time responsiveness is critical, INT8 quantization represents the most practical entry point due to its minimal retraining overhead and negligible accuracy penalty. When the mission profile prioritizes extended autonomous operation over precision, such as long-range patrol or environmental monitoring, INT4 quantization is preferable provided the target hardware supports low-bit inference. Structural pruning is most appropriate when memory bandwidth rather than compute is the primary bottleneck, or when the deployment environment constrains model storage but not inference latency. Knowledge distillation is recommended for production deployments where both energy efficiency and task quality are non-negotiable, and where sufficient computational infrastructure exists offline for the training phase; this strategy is particularly well-suited for multi-mission platforms that must operate across diverse task complexities without model switching.

CONCLUSIONS. The work analyzes the main methods of improving the energy efficiency of small language models for autonomous robotics: quantization, structural pruning, and knowledge distillation. A clear correlation has been established between the computational complexity of the model (number of parameters, FLOP) and the autonomous operation time of mobile robots: a 4 – 8-fold reduction in the number of operations leads to a 1,5 – 2,5-fold increase in autonomy time, depending on the hardware platform. Quantization to INT8 is the most balanced approach with broad hardware support and minimal quality degradation (1 – 3 %), making it suitable for most practical real-time navigation and planning tasks. Quantization to INT4 is appropriate for tasks with lower accuracy requirements, where maximum autonomy is a priority. Knowledge distillation provides the best balance between quality and energy efficiency, but requires significant computational resources during the training phase. A promising direction for further research is the development of adaptive output strategies that dynamically regulate the accuracy of calculations depending on the current battery charge level and the complexity of the input request, which will allow optimal use of the mobile robot's energy resources in real-world missions.

LIST OF REFERENCES

1. Sotnik, S. V. (2025). Support systems for robotics: principles, algorithms and development prospects. Journal of natural sciences and technologies. – 4(2). – pp. 419-430

2. Taran, A., & et al. (2026). Impact of 5G/6G Networks on the Development of IOT, Robotics, and Autonomous Systems. Low Latest and Mass Connection of Devices. All-Ukrainian Conference “Intelligent Civil Safety Technologies and Robotic Systems for Emergency and Rescue Operations” (ICSTRO-2026) February 12-13, 2026. – pp. 114-118

3. Nevliudov, I. S., & et al. (2025). Design of the structure and motion control system of a stationary robot manipulator for construction work. Нові технології в будівництві. – №47. – pp. 37-45

4. Yakimenko, A., & et al. (2026). Robotics in Logistics – From Autonomous Trucks to Amazon's Picking Robots. All-Ukrainian Conference “Intelligent Civil Safety Technologies and Robotic Systems for Emergency and Rescue Operations” (ICSTRO-2026) February 12-13, 2026. – pp. 86-90

5. Sukhomlinova, D. A. & et al. (2025). Analysis of main tasks and challenges in drone remote control. Information Technologies and Automation – 2025 / Proceedings of the XVIII International Scientific and Practical Conference. Odessa, October 30-31, 2025. – Odessa, ONUT Publishing House. – pp. 1059-1061

6. Sukhomlinova, D. & et al. (2025). Optimization of drone trajectory algorithms. Manufacturing & Mechatronic Systems 2025: Proceedings of IX st International Conference, Kharkiv, October 25-26, 2025: Thesises of Reports. – pp. 46-49

7. Andreiev A. S., & et al. (2024). Analysis of robotics platforms for educational and research purposes. Комп'ютерні ігри та мультимедіа як інноваційний підхід до комунікації - 2024 / Матеріали IV Всеукраїнської науково-технічної конференції молодих вчених, аспірантів і студентів, Одеса, 26-27 вересня 2024 р. – pp. 25-27

8. Lykho, T. A., & et al. (2024). Pattern recognition and computer vision technologies in decision support systems of robotic systems. Proceedings of the XVII International scientific and practical conference «Information technologies and automation – 2024». – pp. 645-648

9. Bielik, M. S. & et al. (2025). Analysis of the structures of mobile platforms for promoter robots. The 5th International scientific and practical conference “Global trends in science and education” (June 2-4, 2025) SPC “Sci-conf.com.ua”, Kyiv, Ukraine. – pp. 319-325

10. Yechevskyi, A. D., & et al. (2025). Research of orientation methods of autonomous mobile robots in industrial conditions. «Computer-integrated technologies, automation and robotics» CITAR-2025. – pp. 115-119

11. Fesenko, A. O., & et al. (2025). Basic provisions for developing an automated system for monitoring microclimate parameters in industrial premises. Information Technologies and Automation – 2025 / Proceedings of the XVIII International Scientific and Practical Conference. Odessa, October 30-31, 2025. – Odessa, ONUT Publishing House. – pp. 194-197

12. Sotnik, S. V. (2025). Study of Shrinkage of Thermoplastics in Injection Molding. Scientific Bulletin of Valahia University - Materials and Mechanics. – Vol. 21. – Iss. 24. – pp. 13-17

13. Vasylchenko, Y., & et al. (2025). Development of Security and Fire Alarm Integrated Automation System at Enterprise. WSEAS Transactions on Systems. – № 24. – pp. 642-664

14. Сотник, С. (2025). Розробка автоматизованої інформаційно-пошукової системи вибору маніпулятора промислових роботів. Електромеханічні і енергозберігаючі системи. – № 1 (68). – С. 52-58

15. Sotnik, S. (2025). A Modular System for Gear Calculations: A Comprehensive Computational Approach. DESIGN, CONSTRUCTION, MAINTENANCE. – vol. 5. – pp. 273-289

16. Taran, A., & et al. (2026). AI as a Developer Tool: Github Copilot and Other Artificial Intelligence Assistants. All-Ukrainian Conference “Intelligent Civil Safety Technologies and Robotic Systems for Emergency and Rescue Operations” (ICSTRO-2026) February 12-13, 2026. – pp. 67-71