Some Aspects of Artificial Intelligence Development Strategy for Mobile Technologies

Vadym Slyusar[1,*], Yuriy Kondratenko[1,2], Anatolii Shevchenko[1] and Tetiana Yeroshenko[1]

[1]*Institute of Artificial Intelligence Problems of the Ministry of Education and Science and the National Academy of Sciences of Ukraine, Kyiv, Ukraine*
[2]*Petro Mohyla Black Sea National University, Mykolaiv, Ukraine*
E-mail: swadim@ukr.net; y_kondrat2002@yahoo.com; a.i.shevchenko@ipai.net.ua; eroshenko@ipai.net.ua
[*]*Corresponding Author*

Received 31 October 2023; Accepted 07 February 2024

Abstract

The article addresses hardware-software and other key aspects of the artificial intelligence development strategy for mobile technologies. The proposed components of the strategy include a series of approaches to address issues related to the development and deployment of large language models on mobile devices, as well as suggestions for improving connectivity, memory management, and data security.

Keywords: Strategy, artificial intelligence, neural networks, large language models, Internet of Things, cloud computing, fog computing, quantum computing.

Journal of Mobile Multimedia, Vol. 20_3, 525–554.
doi: 10.13052/jmm1550-4646.2031

List of Abbreviations

AI	Artificial intelligence
LLM	Large language model
SDG	Sustainable Development Goals
NLP	Natural language processing
UNHCR	United Nations High Commissioner for Refugees
WIPO	World Intellectual Property Organization
IoT	Internet of things
FG-AI4A	The use of artificial intelligence in the agricultural sector
ATHENA	Advanced Tools for Historical Environment Assessment
FG-AI4AD	Focus Group on AI for Autonomous and Assisted Driving
WFP	World Food Programme
UNFPA EECARO	United Nations Population Fund Eastern Europe and Central Asia Regional Office
Edge	Edge devices, network edge devices, are computational devices or systems located at the periphery of a network, closer to end-users or data sources
GPT-4	Generative Pre-trained Transformer 4
GPU	Graphics Processing Unit
NPU	Neural Processing Unit
ASR	Automatic Speech Recognition
LVM	Large Vision Model
QML	Quantum Machine Learning
MIMO	Multiple-Input, Multiple-Output
MANET	Mobile Ad-hoc Networks
LIDAR	Laser Radar Probing

1 Introduction

Mobility in the modern world is one of the key characteristics that play a crucial role in the life of mankind, reflecting the transience and flexibility of societal processes, granting people the opportunity to be efficient and always at the center of events. This trend is observed in various spheres, from technology to social relations. Smartphones, tablets, wearable devices, and other mobile gadgets have become an integral part of our daily lives. People

can stay online, communicate, work, and entertain wherever they are. Mobility aids the ability to build a career, study, or relocate in search of better living conditions. Business is also adapting to the growing need for mobility. Remote work, e-commerce, and mobile apps for business reflect shifts in consumer behavior and business strategies. Thanks to electric and autonomous cars, car-sharing, electric scooters, and passenger drones, people are striving for greater mobility both within cities and beyond.

The trend of mobility has not bypassed the education sector. The spread of online courses, virtual classroom services, and educational mobile apps due to the COVID-19 pandemic allows for effective assimilation of educational programs regardless of geographical location.

Mobility is also a crucial characteristic of modern military actions. In the face of unpredictability and rapidly changing circumstances on the battlefield, the ability to quickly adapt, respond, and reallocate resources can determine the fate of combat missions. Modern military technologies, such as drones, autonomous military vehicles, and armored machinery, enhance the mobility of military operations. These technologies allow for real-time tracking of enemy maneuvers, facilitate fast movement of units, and provide the capability to carry out tasks without direct human intervention. Consequently, thanks to the ability to swiftly respond to changes on the battlefield, military strategy and tactics have become more flexible and dynamic. Relying on mobile communication systems, military commanders can execute operational management, receive data in real-time and send relevant instructions to the frontlines. Overall, mobility in modern military actions can be a decisive factor influencing the success or failure of a military operation.

However, unfortunately, this same mobility can lead to the rapid spread of military conflicts when opposing sides use mobility technologies for invasions or swift offensive operations.

In summary, by "mobile technologies" we mean a set of technologies, devices, software, and communication standards that allow users to access information, communicate, and perform various tasks using mobile devices such as smartphones, tablets, wearable electronics, etc. These technologies provide users (be it humans, robots, piloted or unmanned platforms) the ability to interact with data and other users in real-time, regardless of their physical location.

At this stage, the main trend in the development of mobile technologies lies in the application of artificial intelligence (AI) to enhance their efficiency and functionality. In this context, mobile operating systems, mobile internet access services, mobile software applications, communication standards

(such as 4G, and 5G), and other components are noteworthy as targets for AI implementation.

The field of integration between artificial intelligence and mobile technologies is rapidly evolving, so companies that implement AI-based innovations gain competitive advantages by offering users more personalized and intuitive products and services. Specifically, AI can analyze large volumes of user data to predict their needs, adapt to changes in behavior, and respond to market trends. This ensures the ability to provide more relevant content, recommendations, forecasts, and other features that enhance the user experience. High hopes are placed on AI due to the need for further automation of the increasing volume of tasks, reducing resource costs, and improving productivity.

At the same time, ensuring the sustainable development of artificial intelligence technologies for the mobility needs of society is impossible without forming an appropriate strategy.

As is known, in the field of management, strategy is an essential tool for any organization or project, helping to prevent chaotic development and ensuring coordinated efforts. Strategy defines a clear direction of actions, focusing on the main priorities and dismissing insignificant or distracting tasks. It acts as a unique roadmap for leadership, indicating how to achieve the desired future state through coordinated collaboration and mutual understanding of objectives and tasks. The constant limitation of resources and the need to allocate and direct them optimally to achieve maximum effectiveness of expected results adds particular significance to the strategy. Having a strategy helps organizations anticipate potential changes in the environment and prepares them for rapid adaptation to such changes, providing mechanisms for monitoring progress and making adjustments based on collected data. The most significant factor influencing the realization of the positive and negative potential of AI is the degree of its autonomy or self-freedom. The more freedom AI is granted, the more unpredictable the outcome will be, the more human norms and values may change, and the less it will be able to "explain" its actions to humans. Identifying potential threats and opportunities based on this allows for the development of plans for different scenarios and adaptation to future changes, ensuring readiness for innovations. In turn, a clear understanding of strategic goals and tasks can motivate achieving better results, knowing that the community's efforts are directed towards important and priority tasks. In all these aspects, the strategy acts as a means of internal and external coordination, helping society stay focused, efficient, and prepared for future challenges.

Thus, by having a clear development strategy for AI in the field of mobile technologies, states and organizations can identify priority areas of activity and focus resources on the most promising projects, ensuring confidence in the chosen course. Considering the outlined role of strategies, leading countries of the world and international organizations, including NATO, have developed and approved general strategies for the development of artificial intelligence [1, 2]. Along with this, further steps are needed for their detailing and implementation, specifying and synchronizing efforts both within individual countries and on the international stage. Evidence of this is the activity of the United Nations (UN) in the field of artificial intelligence.

2 UN Efforts in the Field of AI Development: The Context of Mobile Technologies

A comprehensive overview of the UN activities in the field of AI can be obtained by reviewing documents [3, 4]. The report [3] presents over 200 projects developed by 40 UN bodies and institutions, aimed at accelerating the achievement of the UN's Sustainable Development Goals (SDGs). Many of the projects are funded by the UN itself, while others are implemented with the support of external partners and organizations, such as the Swiss Agency for Development and Cooperation, the World Bank, and others. Moreover, some projects are part of larger programs or initiatives funded from multiple sources. Relevant projects cover a wide range of areas, from health and infrastructure to inequality and environmental issues. Document [3] provides an analysis of key trends, pointing out areas where the UN system needs to put more effort. In particular, emphasis is placed on SDG 3 (Health and Well-being), 9 (Industry, Innovation, and Infrastructure), 10 (Reducing Inequality), 13 (Responsible Consumption and Production), and 17 (Partnerships to Achieve the Goals). Over 15% of the reported projects were dedicated to AI solutions related to addressing the consequences of COVID-19 or preparedness for pandemic responses. Notably, the most common outputs of UN projects in the field of AI that can be used to address challenges hindering the achievement of SDGs are datasets and software tools, including for mobile applications.

Historically, support has been justified in the context of a series of projects [3] for the use of natural language processing (NLP), which essentially allowed the creation of a foundation for the further implementation of popular AI applications in modern mobile technologies (GPT-4, Bing, Gemini, etc.). Among such projects mentioned in [3], several

typical examples deserve special attention. In particular, the innovation service of the United Nations High Commissioner for Refugees (UNHCR) used NLP to create an efficient time-saving process, avoiding manual analysis and classification of data. This is used to assist the organization in understanding issues related to the UNHCR protection mandate, such as incitement of hatred, discrimination, xenophobia against refugees and individuals seeking shelter, and other individuals who are the concern of the UNHCR.

The Regional Innovation Centre of the United Nations Development Programme's Bangkok Regional Hub has started researching the use of NLP, machine learning (ML), and network analysis for the analysis of dense, unstructured data, while the World Intellectual Property Organization (WIPO) uses NLP in its "WIPO Translate" project, a leading software for translating specialized texts. It can be adapted and configured for other technical areas and mobile applications.

The Executive Office of the UN Secretary-General developed a radio monitoring system that can "listen" to radio stations, translate audio using machine learning models to convert speech to text and analyze content using NLP methods for display on the user interface.

These examples show that NLP is used in various directions within the UN's AI initiatives, from translation and text analysis to speech recognition and data classification, with an emphasis on user mobility.

The analogous report on the activities of the United Nations in the field of artificial intelligence for the year 2022 [4] showcases several new developments and changes compared to the 2021 edition. Notably, there was an observed increase in the number of participants and presented projects (84 new projects introduced in 2022). When analyzing the focus on achieving sustainable development goals, it should be noted that despite a strong emphasis on SDGs 3, 9, 10, and 17, there was a rise in the number of projects targeting SDG 16 (Peace, Justice, and Strong Institutions) in 2022.

Compared to the year 2021, there has been an increase in reporting on projects related to human rights, ethics, justice, agriculture, and telecommunications when "Digital Transformation" was identified as a priority thematic area. Significantly, many of the projects and initiatives discussed in the document involve the development and use of mobile applications to achieve their goals. Specifically, document [4] describes several projects that use AI in the context of mobile technologies in the fields of agriculture, transport, and telecommunications. For instance, the focus group on Artificial Intelligence and the Internet of Things (IoT) for digital agriculture (FG-AI4A) has studied the potential of new technologies, including AI and IoT, in supporting data

collection and processing, improving modeling to increase the volume of agricultural and geospatial data, and ensuring effective telecommunications for measures related to the optimization of agricultural production processes. This project involves the use of mobile technologies for data collection and communication.

The ATHENA project developed a set of artificial intelligence tools for systematizing the International Fund for Agricultural Development (IFAD) portfolio to facilitate results measurement and institutional learning, improving knowledge management by deploying AI/ML in IFAD's information and communication technologies (ICT) systems to make the project outcomes and lessons learned accessible and practical. This project may entail the use of mobile technologies for accessing and interacting with the AI toolset.

The United Nations Office on Drugs and Crime (UNODC) illegal crop monitoring program uses AI for the detection of prohibited plantings. This project also relies on the use of mobile technologies for data collection and transmission.

In the field of transport, the Focus Group on AI for Autonomous and Assisted Driving (FG-AI4AD) under the auspices of the International Telecommunication Union (ITU) has completed a project to support the standardization of services and applications based on artificial intelligence systems in autonomous and assisted driving. The primary goal is to ensure that the performance of AI on the roads will be on par with or exceed that of a competent driver. By achieving international harmony in defining minimum performance standards for AI systems, this is intended to facilitate the introduction of AI on roads and direct efforts to reduce road traffic injuries, which are the leading cause of death among children and youth aged 5–29 years. AI can play a significant role in reducing 1.3 million road deaths and 25 million injuries annually, and also contribute to safe, accessible, and sustainable transport development. The FG-AI4AD's activities concluded in September 2022 with the development of three technical reports, which were forwarded to ITU-T SG16 for further discussion:

"Automated driving safety data protocol – specification" [5];
"Automated driving safety data protocol – Ethical and legal considerations of continual monitoring" [6];
"Automated driving safety data protocol – Practical demonstrators" [7].

In the telecommunications sphere, typical projects [4] are dedicated to the use of AI in the process of creating and distributing TV and radio content, including to mobile users. AI ensures automatic extraction and localization

of content from vast archives, generating accessible services like subtitling, audio description, text-to-speech, and sign language, much faster and more accurately than operators do. AI is also being explored as an effective tool for spectrum management and radio monitoring activities, including in relation to moving radiation sources.

A rather significant group of UN projects [4] utilizes mobile applications to achieve their goals. For example, the "MEZA" project [4] involves developing an optical character recognition system based on artificial intelligence to digitize handwritten records, speeding up the collection and analysis of data on the nutrition of millions of malnourished children. The corresponding application allows the WFP and governments to promptly obtain necessary information and enhance the efficiency of combating malnutrition in remote medical clinics.

The project "AMMA App – Period & Pregnancy Tracker" [4] was dedicated to improving a mobile application named "Amma", which allows women to have informed, safe, and healthy pregnancies under the remote supervision of doctors. With the support of United Nations Population Fund (UNFPA) Eastern Europe & Central Asia Regional Office (EECARO), additional content for mobile applications was created within the project to convey more accurate information about safe pregnancy, prenatal care, analyze ultrasound images with the help of AI, and more. Every year, 10 million pregnant women download the app, and it is available in 13 languages.

This list of UN efforts can be extended and leads to the conclusion that the integration of AI and mobile technologies is a common trend. However, mainly ready-made fundamental-level solutions in the AI field are used, whereas the problem lies in developing long-term approaches to achieving maximum synergy between AI and the applied aspects of mobility. An appropriate AI development strategy in mobile technologies must take into account the specifics of mobile devices, user needs, and technical capabilities. Let's consider several key aspects that should be taken into account when forming such a strategy.

3 Key Elements of An Effective Development Strategy for AI in Mobile Technologies

3.1 Searching for New Neural Network Architectures and LLM

One of the main issues in the mentioned field is the need to optimize the architecture of AI models for mobile devices and systems, which, as we

know, have limited resources (for example, computing power, memory, and battery energy). Therefore, neural network structures, especially LLM (Large Language Model), which often require significant computational resources, should be optimally scaled for mobile devices to efficiently use resources. This will ensure high execution speed and efficiency in terms of energy consumption. Moreover, considering that internet access can be limited, AI models should operate both online and offline.

Global trends in the AI sector involve the development of hardware platforms to reduce the cost of implementing local versions of LLM and scaling them to the level of Edge devices, as well as further development of non-local LLMs. Additionally, a promising direction is the adoption of transformer-informer architectures of LLM for solving multimodal generative AI tasks, including joint processing/generation of text, images, video, and audio. A prime example of this is the integration of the GPT-4 [8] service and DALL-E3 into GPT-4V, the development and refinement of a multimodal language model for language and image processing called LLaVA (Large Language-and-Vision Assistant), which by October 2023 had already distinguished itself in tasks such as image descriptions and answering questions about their content [9]. Notably, LLaVA can be trained on a single 8-A100 GPU. Another example of the progress in the development of multimodal language models is Kosmos-2 [10] with a chatbot function similar to LLaVA's for providing insight information about images. It's easy to predict that such a service will be in demand in smartphones shortly. It's entirely possible that this approach will gradually replace traditional image segmentation technologies [11] and object detection [12, 13], although there will still be a niche for them in extremely resource-limited mobile applications. Moreover, there may be some integration of traditional neural network architectures with transformer structures.

Improvement of neural network architectures and large language models can involve a range of strategies aimed at increasing productivity, accuracy, flexibility, and interpretability. One way or another, they presuppose the development of more efficient optimization and weight initialization algorithms to accelerate learning, and the application of quantization and weight factor coarsening methods to reduce memory and computational resource requirements. Modularity and flexibility are achieved using modular architectures, which allow components to be easily replaced or added to enhance the functionality and flexibility of the system. This also enables the distribution of hyper-neural network modules in Fog and Cloud environments, among swarms of autonomous platforms, while maintaining coherent functioning

and the ability to form output results after parallel data processing. The adaptability of neural network architectures to work with new types of data or tasks also plays in their favor, for example, due to the sequential or parallel integration of specialized adapters into the architecture. Additionally, the modularity of the architecture allows it to work with large-volume LLMs by sequentially reloading its configuration into the processor in several stages. Depending on the available memory capacity of the hardware platform, for instance, loading a local version of an LLM like GPT-3 may require several hundred reloads.

With the indicated direction, there is a close connection to the development of entirely new architectures, for example, through the introduction of new types of layers or mechanisms of attention or intentions. The effectiveness of such an approach is evidenced, for example, by the application of the distributed attention mechanism in the local LLM Mistral-7B family [14], which, despite their small size, can demonstrate effectiveness comparable to some LLMs with 10 times more tuning parameters. For instance, according to data from the platform for evaluating open language models, Open LLM Leaderboard [15] on the Hugging Face portal, LLM Dolphin-2.1-mistral-7b [16] (7 billion parameters) surpassed the average indicator level of 67.06 in the ranking list, models with 70 billion parameters LLaMa-2-70b-chat-hf (66.8), Aria-70B (66.76), WizardLM 70B V1.0–GPTQ (66.47), and also 180-billion LLM Openbuddy-Falcon-180b-v13-preview0 (67.01). Similar results are demonstrated by Mistral-7B-OpenOrca [17], which, according to the MT-bench test, matched LLaMa-2-70b-chat (see Table 1) and slightly

Table 1 The comparison of the effectiveness of some LLMs

Model	*MT-bench (score)*	MMLU
GPT-4	8.99	86.40
Claude-2	8.06	78.50
GPT-3.5-turbo	7.94	70.00
Claude-1	7.90	77.00
WizardLM-70b-v1.0	7.71	63.70
Vicuna-33B	7.12	59.20
WizardLM-30B	7.01	58.70
Mistral-7B-OpenOrca	6.86	61.73
LLaMa-2-70b-chat	6.86	63.00
Vicuna-13B	6.57	55.80
Guanaco-33B	6.53	57.60
Guanaco-65B	6.41	62.10
OpenAssistant-LLaMa-30B	6.41	56.00

lagged behind this version of LLaMa-2 in the MMLU (Massive Multitask Language Understanding) indicator.

Overall, the synthesis of efficient architectures is a complex task, yet its effectiveness can significantly enhance the capabilities of mobile AI technologies when successful solutions are found. For this reason, this direction is a priority in the list of components of the AI development strategy for mobile users. It requires substantial human and hardware resources, so the optimal approach is to search for new architectures primarily by open communities of enthusiast developers and scientists with the goal of further scaling advanced LLM architectures for solving multimodal tasks of generative AI, especially for joint processing/generation of texts, audio, images/videos. In this process, it's important to develop models capable of operating in mobile applications with a large number of tokens (up to 100,000) for processing extensive texts (books, reports, groups of articles) and videos. In the context of synthesizing new architectures, deep collaboration with neurobiologists and representatives of other brain sciences will also be a key to success, aiming to leverage the results of studying natural neural organizational structures.

3.2 Improvement of the Hardware Implementation of AI

Creating a mobile artificial intelligence infrastructure based on Large Language Models is a complex task due to the vast number of parameters and computational resources required for real-time data processing and analysis. These aspects can serve as a foundation for formulating an AI development strategy in mobile technologies.

To address issues related to the creation of a mobile AI infrastructure based on LLM, a strategy is proposed that includes a series of recommendations.

Firstly, there's a need to develop lightweight models with 7–13 billion parameters that can be efficiently deployed on Edge devices, ensuring sufficient productivity and accuracy. This will reduce the computational and energy requirements since the maximum power of Edge devices should be limited to 100W, taking into account power consumption management under various scenarios. LLMs with 30 billion tuning coefficients should be considered as the complexity limit for advanced Edge devices.

The development of the mentioned hardware platforms will reduce the cost of implementing local versions of LLM and their scaling to Edge device levels and will also promote the further development of larger non-local

LLMs to be localized at data centers, companies, enterprise, and organizations. Additionally, quantization techniques should be used, relying on reduced precision data formats, such as FP16 or BF16 (BFloat16) [18], to alleviate memory requirements and improve inference speed. Notably, the BF16 format is gaining more and more supporters among machine learning hardware developers because it has an optimal precision-to-range ratio [18], making it ideal for deep learning tasks where numerical reliability and convergence are crucial.

Secondly, it is recommended to only perform inference procedures on Edge devices, whereas fine-tuning and model updates should be carried out on more powerful servers or in the cloud environment, to ensure optimal training and tuning of models before deploying them to Edge-device. Notably, such servers can be deployed in multiple tiers, incorporating intermediate Fog Computing at the level of desktop computers, laptops, user tablets, or drone control consoles, which directly interact with mobile devices, and also at cellular base stations, which provide connection to cloud services [19]. Both cloud and Fog computing are capable of compressing and processing multiple requests from many users of the AI service simultaneously. For example, compressing and aggregating tens of thousands of requests in a cloud service can significantly reduce costs for accessing such services and maintaining their 24/7 operation. An AI with functionality similar to Siri should continue to be implemented using Cloud-computing, especially when scaling this service to LLM with 100 billion parameters and more, and with fine-tuning.

Worthy of attention is the concept of integrating additional AI computing resources into traditional power banks and portable charging stations, which will serve not only as an additional power source for smartphones or other gadgets but also as a means of expanding the neural network model and an auxiliary computing resource when implementing AI algorithms in interaction with smartphones.

Ensuring data exchange in the AI Fog environment, as well as uploading updated and retrained neural network models to mobile devices, can be done not only through cable connections but also through Wi-Fi and 5G/6G cellular networks to reduce data transmission delays. Implementing advanced 5G/6G communication networks will also reduce the data transfer time between Edge devices and cloud servers. More detailed strategic directions for communication improvements will be discussed separately.

Thirdly, it is important to seek opportunities for hardware acceleration of AI algorithms on end devices, including the use of specialized accelerators for

matrix and tensor operations, parallelizing matrix multiplication, and other computational operations.

An example of this approach in the field of mobile technologies is the Qualcomm Snapdragon 8 Gen 3 [20] mobile processor, which uses the AI coprocessor NPU (Neural Processing Unit) Hexagon, which, according to the developer, provides a 98-percent increase in performance and a 40-percent increase in energy efficiency in tasks related to AI algorithm operation. The NPU Hexagon supports multimodal generative AI models with 10 billion parameters, including popular LLMs and machine vision models (LVM), as well as automatic speech recognition (ASR) based on neural network transformers. According to Qualcomm, image generation using the NPU with the Stable Diffusion neural network is done at a rate of 1 second. In the case of the LLaMa-2/Baichuan LLM language models, the NPU can process more than 20 tokens per second.

In general, it should be noted that for LLM, the token passage rate (Rate) is a very important indicator. For a person to comfortably read the responses of GPT-4, for example, it is necessary to generate text at a rate of no less than 3 words per second. Meanwhile, in Ukrainian texts, a single word often corresponds to 6–8 tokens in typical tokenization and embedding libraries. Thus, the performance indicator of the mentioned NPU, which processes 20 tokens per second, is almost borderline acceptable. At the same time, one should take into account the time for embedding the input stream and converting tokens into language that is easily comprehensible to humans.

Overall, the indicated neural processor efficiently transfers the capabilities of modern generative artificial intelligence into a chipset. This significantly speeds up labor-intensive operations that were usually offloaded to cloud computing. It's not difficult to predict that reducing the NPU's topology from the current 4 nm to 2 nm or less will eventually overcome the threshold of 20 billion LLM tuning parameters, which is quite substantial for the further democratization of mobile AI services.

Besides specialized neural processors, efficient memory management plays a significant role in mobile AI platforms, including optimizing the interface between the neural processor and memory and using high-performance High Bandwidth Memory (HBM) to accelerate data access. The aforementioned need to increase the number of tokens to 100,000 is accompanied by an increase in their length, leading to the need to store intermediate computation results. As a result, the response speed of the neural network is largely determined by the memory interface. For example, when implementing attention mechanisms, it is necessary to fully read and reload the coefficients

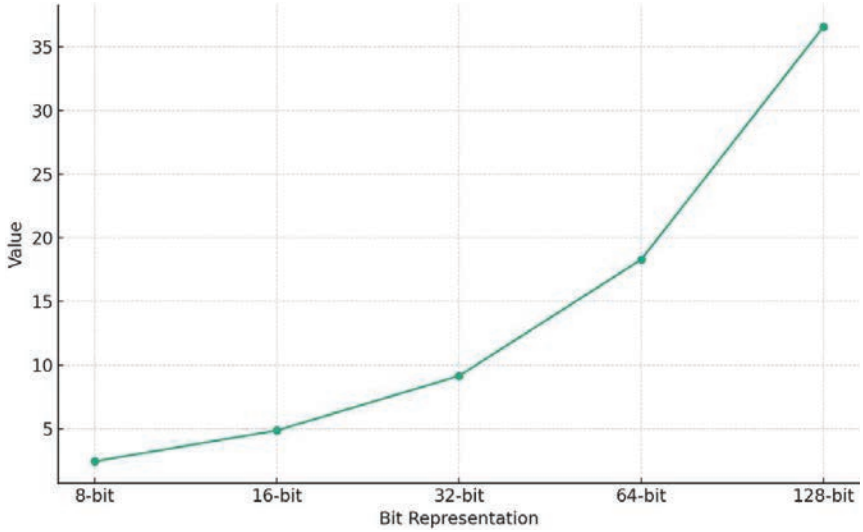


Figure 1 The dependence of normalized values of required VRAM volumes on the bit depth of weight coefficients representation for LLM LLaMa-1 7B.

of matrices Q, V, and K [21]. The problem is that the time to reload them currently exceeds the computation time. The situation is also complicated by the need to protect data by encrypting (scrambling) it during the exchange between the RAM and the AI accelerator to prevent information leakage.

Regarding memory requirements, it's recommended to specify a petabyte threshold as the desired level. The general methodology of calculating the size of Video Random Access Memory (VRAM) requirements for training LLM and inference processes is considered in [21, 22]. As an example, Figure 1 illustrates for LLM LLaMa-1 7B, based on data from [22], the impact of the bit depth of weight coefficients, plotted along the horizontal axis, on the minimum necessary VRAM sizes. The relative VRAM volumes are shown along the vertical axis, normalized to the model size (7B). Essentially, this chart demonstrates how many times the necessary VRAM resources exceed the declared volume of the LLM and reflects the nonlinear nature of the corresponding dependence on the bit depth of the model's weight coefficients. However, this requirement isn't so pressing considering the practice of reloading interim results. According to experts, with the neural network coefficient format of FP16, using 4 HBM-memory chips with a capacity of 4 Terabytes each is sufficient to cover a need of 600 Terabytes, taking into account a reasonable volume of data reloads per second.

Furthermore, the use of sequential compression techniques and efficient data representation can help reduce the required memory volume.

A similar approach of sequential loading of LLM segments can also be applied to an architecture based on the community of experts (agents) principle. As noted in [23], the use of a mixed architecture allowed the LLM Mixtral 8x7B to outperform LLaMa-2 70B and GPT3.5-turbo in a number of a variety of benchmarks. According to [23], the Mixtral 8x7B architecture combines 8 Mistral 7B models, each specializing in a specific set of tasks. The subsequent moderation mechanism ensures the adaptive selection of the most relevant response from the collective models for each user query. This approach aligns with the principle of selecting experts of specific specialization depending on the context of the task at hand. Importantly, the synergistic effect achieved by combining several specialized models lies not only in improving the quality and speed compared to larger monolithic LLMs but also in a relative reduction in the number of parameters necessary for the instance process. In particular, as stated in [23], the 47-billion Mixtral 8x7B architecture only requires 12 billion tunable weight coefficients. This significantly reduces hardware resource requirements to maintain LLM operability. Moreover, replacing a monolithic architecture of larger size with a collection of several smaller LLMs in the case of Mixtral 8x7B allows the use of the principle of sequential loading into the memory of the neuroprocessor of individual models or clusters formed from them, depending on the context of the task being solved.

Significantly, the application of a mixed architecture like Mixtral 8x7B can be scaled to other small-dimension LLMs. Specifically, this approach is suggested for use with Google's developed LLMs Gemini Nano 1, which has 1.8 billion parameters, and Gemini Nano 2 with 3.25 billion parameters [24–26]. As a result, for example, a collection of 8 or 16 LLM Gemini Nano 1 can be considered as an alternative to Mixtral 8x7B. Within this conglomerate, the bit depth of weight coefficients can vary from model to model. For instance, several Gemini Nano models may have a 4-bit weight coefficient, while others have an 8-bit one, etc. Additionally, to form a community of experts, both homogeneous sets of LLMs and their various combinations can be used. For example, 4 LLM Gemini Nano 2 can be combined with 8 LLM Gemini Nano 1, utilizing the potential of the more powerful models for processing images or videos and leaving text processing to Gemini Nano 1. Similarly, within a single architecture, such as Gemini Nano and Mistral 7B or other LLMs, can be combined. As a result, it becomes possible to maximally utilize available hardware resources, optimally tailoring the volume of

the collective LLM to the size of the available VRAM of the graphics card or the dedicated RAM segment for LLM operations.

Regarding mobile systems, the considered concept of integrating several expert LLMs within a single model also simplifies the hardware distribution of the collective LLM across multiple interacting devices. For example, in this case, part of the overall model as a cluster of several Gemini Nano 1 and Nano 2 can be located in a smartphone, while another part, for instance, Gemini Nano 1 – in smartwatches, smart headphones, or smart power banks. Their integration into a unified whole is facilitated by high-speed communication channels. Such a version of Fog Computing, if necessary, can also be implemented within a car, distributing the LLM conglomerate between the onboard computer and the smartphones or other gadgets of the passengers.

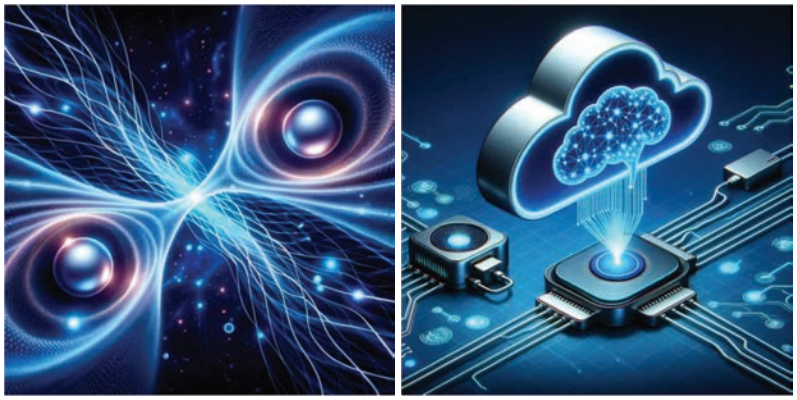
Lastly, considering security needs, a coded configuration file for hardware should be used for effective management and adjustment of hardware parameters. This approach allows for the creation of an adaptive configuration of hardware to meet the requirements of different usage scenarios, optimize power, and ensure system safety.

The described hardware aspects of the strategy for creating a mobile AI infrastructure based on LLMs should help address the outlined challenges in the short and medium term and ensure more efficient and faster natural language processing tasks on Edge devices.

3.3 Regarding the Strategy of Synergy Between Artificial Intelligence and Quantum Technologies

Quantum technologies open up broad prospects for improving and accelerating the development of artificial intelligence to meet mobility needs [19]. Their impact, especially in computational speed, can be significant: using quantum phenomena, such as superposition and entanglement (Figure 2), quantum computers can perform certain types of calculations much faster than their classical counterparts. This can reduce the training and fine-tuning time for artificial intelligence systems, thereby accelerating their development. Furthermore, quantum cloud computing based on quantum communications will contribute to scaling up large language models, and neuro-transformers for processing and generating images and video content, providing a hardware foundation for their integration and the development of so-called strong artificial intelligence.

Known as Quantum Machine Learning (QML), this scientific field that combines the principles of quantum physics and machine learning also has



a) b)

Figure 2 The development of quantum entanglement technologies in the medium-term perspective has the potential to significantly expand the capabilities of (a) quantum communications and (b) quantum computing (DALL-E3).

significant potential to enhance the effectiveness of artificial intelligence. It aims to harness the power of quantum computers to optimize machine learning algorithms. Such quantum systems can offer opportunities for effective training of models on larger datasets and more nuanced problem-solving in optimization.

Quantum technologies are also important in the context of ensuring the security of artificial intelligence systems. The application of quantum cryptography and secure quantum communications in quantum networks can contribute to the development of highly resistant-to-hacking artificial intelligence systems, ensuring their reliability and security. In this sense, quantum key distribution will provide strict security measures to protect data exchange between mobile users and cloud services.

In turn, the artificial intelligence development strategy should promote the advancement of quantum technologies themselves. At the intersection of artificial intelligence and quantum technologies, a combined strategy should be formulated that merges both fields to accelerate the progress of each. Notably, the use of machine learning can revolutionize the process of designing and fine-tuning quantum systems. AI algorithms, processing data from quantum communications, are capable of optimizing their parameters to maximize productivity and efficiency.

Artificial intelligence methods can play a key role in calibrating quantum devices and monitoring quantum systems. Detecting and correcting errors

in quantum communication systems and quantum computers requires high precision, which AI algorithms can provide.

Artificial intelligence can also contribute to quantum compilation LLM – the conversion of quantum algorithms into programs designed for execution on quantum computers. Here, machine learning becomes an indispensable tool that will significantly improve the efficiency of this process.

In totality, the interaction between quantum technologies and artificial intelligence initiates a powerful wave of progress in both fields. In a world where there is intense competition for the implementation and ownership of these technologies, such synergy can become a catalyst for significant strategic changes in the use and development of artificial intelligence. New opportunities open up for the advancement of both technologies. Thus, studying and understanding the possible impact of quantum technologies on artificial intelligence becomes a top priority for scientists, engineers, and strategists in this field.

Monitoring the mentioned synergistic relationships will allow for understanding the deep mechanisms of the influence of artificial intelligence on quantum technologies, which will open new horizons for scientific discoveries. Therefore, it is important to pay sufficient attention to the development of both fields so that they continue to coherently complement each other and contribute to societal development. This is not just a matter of scientific progress but also a strategic task, the resolution of which will affect what the future of technology will look like and their impact on our society.

3.4 Reduction of Computational Volumes Due to Special Matrix Operations That Allow Parallelizing Data Streams

A critically important path to achieve parallelization and optimization of computations in the mobile segment IT infrastructure is the introduction of new mathematical methods and the refinement of existing versions of the implementation of mathematical operations for big data processing. In this context, the application of the tensor-matrix theory of neural networks, developed, for example, in [27] based on the family of face-splitting products of matrices [28, 29], deserves attention. The corresponding mathematical apparatus allows formalizing neural network models of any complexity, avoiding the limitations of classical matrix calculations, and reducing the mathematical complexity of data processing libraries. This, among other things, will contribute to increasing the transparency of the mathematical models of neural networks and their controllability.

As an example, the penetrated product of matrices can be presented similar to [27]:

$$\mathbf{A} \circledcirc \mathbf{B} = [\mathbf{A} \circ \mathrm{B_n}] = [\mathbf{A} \circ \mathrm{B}_1\ \mathbf{A} \circ \mathrm{B}_2 \vdots \ldots \vdots \mathbf{A} \circ \mathrm{B_n} \vdots \ldots], \tag{1}$$

$$\mathbf{A} = \begin{bmatrix} \mathrm{a}_{11} & \mathrm{a}_{12} \\ \mathrm{a}_{21} & \mathrm{a}_{22} \\ \mathrm{a}_{31} & \mathrm{a}_{32} \end{bmatrix},\ \mathbf{B} = \left[\begin{array}{cc:cc:cc} \mathrm{b}_{111} & \mathrm{b}_{121} & \mathrm{b}_{112} & \mathrm{b}_{122} & \mathrm{b}_{113} & \mathrm{b}_{123} \\ \mathrm{b}_{211} & \mathrm{b}_{221} & \mathrm{b}_{212} & \mathrm{b}_{222} & \mathrm{b}_{213} & \mathrm{b}_{223} \\ \mathrm{b}_{311} & \mathrm{b}_{321} & \mathrm{b}_{312} & \mathrm{b}_{322} & \mathrm{b}_{313} & \mathrm{b}_{323} \end{array}\right], \tag{2}$$

$$\mathbf{A} \circledcirc \mathbf{B} = \left[\begin{array}{cc:cc:cc} a_{11} \cdot b_{111} & a_{12} \cdot b_{121} & a_{11} \cdot b_{112} & a_{12} \cdot b_{122} & a_{11} \cdot b_{113} & a_{12} \cdot b_{123} \\ a_{21} \cdot b_{211} & a_{22} \cdot b_{221} & a_{21} \cdot b_{212} & a_{22} \cdot b_{222} & a_{21} \cdot b_{213} & a_{22} \cdot b_{223} \\ a_{31} \cdot b_{311} & a_{32} \cdot b_{321} & a_{31} \cdot b_{312} & a_{32} \cdot b_{322} & a_{31} \cdot b_{313} & a_{32} \cdot b_{323} \end{array}\right], \tag{3}$$

where matrix **A** can be considered as the input matrix of image pixels, while each block of matrix **B** corresponds to a block of weight coefficients for several neurons in one layer of the neural network.

3.5 Neural Network Models Accounting the Specifics of Hardware Platforms

Modern neural networks are predominantly developed in high-level programming languages, such as Python. An advantage of this approach is the ability to run identical software models on various hardware platforms. In doing so, the programmed model, as a sum of knowledge laid out in the software product, remains robust since even if a particular hardware component stops functioning, the knowledge regarding the neural network architecture and its weight coefficients are preserved and can be launched on other equipment [30]. However, a drawback of this approach is the neglect of hardware-specific characteristics, resulting in excessive time and energy consumption for computations. Given this, ideally, neural network models for mobile devices should be written in Assembly language to optimize the code and speed up computations.

On the other hand, if one doesn't differentiate between software and hardware, we get what Geoffrey Hinton called "dead computations" [30]. In the context of neural networks, this means that the software model takes into account the specific analogue properties of the hardware platform and is only useful for that particular equipment. This provides advantages for the application of low-energy analogue computations. However, a compiled model cannot be scaled to other hardware resources that have certain distinctions.

One solution is to strike a balance between the abstraction level of the programming language used for neural network model development and the complexity of the programming process.

At the same time, as pointed out in [30], digital computations allow the launch of many copies of an identical model on different hardware. All these digital agents can analyze different data and share what they have learned very efficiently, averaging changes in their weight coefficients. Combined with reinforcement learning and the back propagation training method, this could potentially allow adapting neural network models to the physical peculiarities of hardware solutions without losing their scalability and relocation capabilities.

In the authors' opinion, the development of this direction deserves attention since learning by examples could, quite possibly, allow us to abandon the fundamental principle of informatics – the separation of software and hardware in the future. Moreover, the significance of the approach to training models directly on the device lies in the fact that it can not only improve the response speed of neural networks but also ensure privacy preservation.

3.6 Directions for Improving Communication Tools

A promising direction in the field of data exchange in the AI Fog environment, apart from the already mentioned implementation of quantum networks and 5G/6G communications, is the transition to data transmission protocols that do not require routing. An example of this is the Barrage Relay method developed by Trellisware Technologies, Inc. (USA) [31]. The peculiarity of the Barrage Relay concept lies in the reuse of MIMO channels by re-emitting the same packets in the same time slots by multiple nodes of the MANET (Mobile Ad-hoc Networks) network. The Barrage Relay method eliminates the need for routing protocols and allows for massive scalability of networks with very low overhead costs. Unlike traditional approaches, Barrage Relay does not require supporting infrastructure and enables each node to simultaneously transmit, receive, and relay information. The effectiveness of this approach can be enhanced by employing AI.

Another effective means of reducing the amount of traffic between mobile devices and the stationary segment of AI is augmented reality, which serves as a connecting bridge between humans and AI. Standardizing augmented reality symbols will minimize intensity and speed up human-AI interactions using mobile gadgets. In this context, AI can be involved in generating augmented reality symbols [32] and converting them back into audio or

textual messages. Importantly, integrating AI with modern IoT and Augmented/Virtual Reality (AR/VR) technologies opens up new opportunities for developing multifunctional applications.

Reducing the requirements for data transmission channels is also achieved by introducing the standardized SAPIENT protocol (Sensing for Asset Protection with Integrated Electronic Networked Technology) [33, 34]. This protocol was developed in the interests of the UK Ministry of Defence and NATO adoption [35]. The traffic advantage in the system with SAPIENT is based on using not raw data from sensors but data processed by artificial intelligence for the detection and classification of objects. They also make decisions about their operation autonomously, for example, where to look or when to zoom in on an image. This reduces the need for an operator to constantly monitor the sensor output.

Thus, new standardized communication protocols aimed at optimizing traffic through data processing by artificial intelligence point to the growing role of autonomous systems in the communication landscape.

In addition to the transport level of communications, it is also necessary to implement spectrally efficient types of signals at the physical level, for example, considering their further super-Relay resolution [36].

3.7 Key Directions for Implementing AI in the Transport Sector

Transportation with autopilots, driver assistants, and unmanned platforms form a separate niche where the features of AI integration and mobile technologies are most reflected. That's why we will pay special attention to this area, although there are undoubtedly other directions that deserve separate consideration beyond this article.

Artificial intelligence is being introduced into the automotive industry in various ways, bringing a revolution in many aspects of the design, production, and operation of vehicles, making them safer, smarter, and more efficient. These trends will continue.

Using a combination of cameras, radars, ultrasonic sensors, LIDAR, and other sensors, AI can analyze the surrounding environment, predict the actions of other road users, and make decisions about maneuvers on the road, making cars fully or partially autonomous.

Many modern cars are already equipped with intelligent driver assistance systems, such as automatic emergency braking, lane departure warning, or automatic parking. Personal voice assistants, such as Siri or Google Assistant, are integrated into cars, allowing drivers to control many car functions without manual distractions.

a) b)

Figure 3 Automotive transport is a key object of the integration strategy (a), (b) of IT and mobile technologies (DALL-E3).

AI can predict when a particular part of a car, such as brake pads or batteries, may need replacement or maintenance. Such predictive maintenance will reduce the cost of operating a car and help avoid emergencies.

With cameras and biometric sensors, some cars can recognize the driver, automatically adjusting the seat, mirrors, and multimedia system settings. Meanwhile, the AI assistant can suggest music tracks, news, or other forms of entertainment based on the preferences of the driver and passengers with service personalization. An adaptive user interface can significantly enhance the comfort of using cabin space, allowing the built-in service system to automatically adjust to user preferences.

For optimizing people's mobile capabilities, AI, with the help of a cloud service (Figure 3), can analyze real-time city traffic data and suggest optimal routes to avoid congestion.

In electric vehicles, AI is capable of optimizing the use of battery packs by predicting energy needs based on a given route, road condition scenarios, and driving style. The more the smart car "learns" from the driver's behavior and traffic conditions, the better it can adapt and offer optimal solutions. In this context, traditional neural networks for time series processing will increasingly face competitive pressure from transformer architectures, an example of which is the development of TimeGPT-1 [37].

During the car manufacturing phase, AI should continue to be used for optimizing production lines, quality control, and automating routine tasks. This will improve planning and the course of production processes. At the

same time, it should be noted that given the vast amount of personal data on mobile platforms, AI algorithms must be secure and guarantee the confidentiality of user data. Privacy and security in this field remain critical issues, especially in the context of ensuring the prevention of negative consequences from unauthorized access.

Lastly, it is particularly important to highlight the problem of ethical choice in preventing the consequences of road traffic accidents. Its resolution requires the implementation of a mechanism known as "artificial conscience" [38–40] and is closely linked to various aspects of philosophical science. The problem of forming a human perception of the environment in AI, to unify augmented reality platforms for humans and robots [41, 42], also remains unresolved. These and other aspects should be the subject of further research.

4 Conclusions

The directions of AI development for mobile technologies discussed in the article only partially cover the incredibly vast array of relevant issues and ways to address them. This is why artificial intelligence has become a subject of study for various sciences today. The divergence of interests can be so significant that researchers might not understand each other. The need for coordinating considerations at the most generalized, philosophical level increases depending on how the essence of human intellect, thinking, and consciousness is interpreted. These interpretations affect searches in more specialized fields.

In interdisciplinary areas, like the field of artificial intelligence, philosophical intuitions (as a crucial component of development strategy) play a defining role not only at the initial development stage but throughout the entire lifecycle of technologies. The discussion of AI from a philosophical perspective has numerous facets. Among them, several fundamental philosophical questions can be distinguished. Depending on the answers to these questions, different foundations will be laid for understanding the essence of human intellect, its interaction with the world, and the principles of creating general artificial intelligence. These philosophical aspects of the AI issue represent the cutting edge of scientific advancement in the context of the discussed strategy, especially if we talk about the most ambitious task of creating artificial thinking, consciousness, and intellect.

Modern AI technologies, at best, replicate probable human activities under certain conditions. An important factor influencing the realization of

AI's positive and negative potential is the degree of its autonomy or inherent freedom. A person who delegates some of their responsibilities to a computer, smartphone, or smart home system – which under certain circumstances becomes an "open book" for an AI capable of predicting and influencing the sequence of device actions – risks a moment when they cannot or do not have time to intervene in the actions of the chosen device. Thus, human civilization is entering a new phase where artificial intelligence might gain total control over society. Therefore, developing AI strategies in relevant sectors, especially in mobile technologies, will allow timely prevention of possible negative trends and maximum mitigation of potential critical consequences of unforeseen developments.

In conclusion, it should be noted that artificial intelligence opens up prospects for deeper personalization of mobile user interactions, allowing the adaptation of relevant interfaces in the form of voice assistants and chatbots. The integration of AI with augmented reality and IoT technologies creates new opportunities for enhancing user experience. However, this also raises the issue of educating users so that they can use these new possibilities most effectively. Adherence to ethical standards and potential regulation of AI development and implementation processes is essential, considering that continuous learning is necessary for AI's optimal functioning and ensuring multi-platform support. In this context, international cooperation among developers is strategically important for widespread AI implementation, capable of providing them with tools for seamless integration of the latest mobile technologies into specific applications.

References

[1] Y. Kondratenko, A. Shevchenko, Y. Zhukov, M. Klymenko, V. Slyusar, G. Kondratenko, O. Striuk, 'Analysis of the Priorities and Perspectives in Artificial Intelligence Implementation', 13th International IEEE Conference "Dependable Systems, Services and Technologies" (DESSERT'2023), Greece, Athens, October 13–15, 2023.

[2] Y. Kondratenko, G. Kondratenko, A. Shevchenko, V. Slyusar, Y. Zhukov, M. Vakulenko, 'Towards Implementing the Strategy of Artificial Intelligence Development: Ukraine Peculiarities', CEUR Workshop Proceedings, vol. 3513, 2023, pp. 106–117, https://ceur-ws.org/Vol-3513/paper09.pdf.

[3] United Nations Activities on Artificial Intelligence (AI). 2021, https://www.itu.int/dms_pub/itu-s/opb/gen/S-GEN-UNACT-2021-PDF-E.pdf.

[4] United Nations Activities on Artificial Intelligence (AI). 2022, https://www.itu.int/dms_pub/itu-s/opb/gen/S-GEN-UNACT-2022-PDF-E.pdf.

[5] Technical Report "FGAI4AD-01 – Automated driving safety data protocol – Specification", Focus Group on AI for autonomous and assisted driving (FG-AI4AD), 2022, 22 p., https://www.itu.int/dms_pub/itu-t/opb/fg/T-FG-AI4AD-2022-PDF-E.pdf.

[6] Technical Report "FGAI4AD-02 - Automated driving safety data protocol – Ethical and legal considerations of continual monitoring", Focus Group on AI for autonomous and assisted driving (FG-AI4AD), 2021, 52 p., https://www.itu.int/dms_pub/itu-t/opb/fg/T-FG-AI4AD-2021-02-PDF-E.pdf.

[7] Technical Report "FG-AI4AD-03 – Automated driving safety data protocol – Practical demonstrators", Focus Group on AI for autonomous and assisted driving (FG-AI4AD), 2022, 90 p., https://www.itu.int/dms_pub/itu-t/opb/fg/T-FG-AI4AD-2022-01-PDF-E.pdf.

[8] GPT-4. Technical Report by OpenAI, 27 March 2023, URL: https://arxiv.org/pdf/2303.08774v3.pdf.

[9] H. Liu, C. Li, Q. Wu, Y.J. Lee, 'Visual Instruction Tuning', 2023, 19 p., https://arxiv.org/abs/2304.08485.

[10] Z. Peng, W. Wang, L. Dong, Y. Hao, S. Huang, S. Ma, and F. Wei, 'Kosmos-2: Grounding Multimodal Large Language Models to the World', 2023, 20 p., https://arxiv.org/pdf/2306.14824.pdf.

[11] V. Slyusar, M. Protsenko, A. Chernukha, V. Melkin, O. Petrova, M. Kravtsov, S. Velma, N. Kosenko, O. Sydorenko, M. Sobol, 'Improving a neural network model for semantic segmentation of images of monitored objects in aerial photographs', Eastern-European Journal of Enterprise Technologies, vol. 2, no. 6 (114), 2021, pp. 86–95, doi: 10.15587/1729-4061.2021.248390.

[12] V. Slyusar, et al., 'Improvement of the object recognition model on aerophotos using deep convolutional neural network', East. Eur. J. Enterp. Technol., vol. 5, no. 2 (113), 2021, pp. 6–21.

[13] V. Slyusar, M. Protsenko, A. Chernukha, V. Melkin, O. Biloborodov, M. Samoilenko, O. Kravchenko, G. Kalinichenko, A. Rohovyi, M. Soloshchuk, 'Improvement of the model for detecting objects on aerial

photos and video in unmanned aerial systems', Eastern-European Journal of Enterprise Technologies, vol. 1, no. 9 (115), 2022, pp. 24–34, doi: 10.15587/1729-4061.2022.252876.

[14] A. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. Singh Chaplot, D. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. Renard Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, 'Mistral 7B', 2023, 9 p., https://arxiv.org/pdf/2310.06825.pdf.

[15] E. Beeching, C. Fourrier, N. Habib, S. Han, N. Lambert, N. Rajani, O. Sanseviero, L. Tunstall, T. Wolf, 'Open LLM Leaderboard', Hugging Face, 2023, https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard.

[16] E. Hartford, 'Dolphin-2.1-mistral-7b', https://huggingface.co/ehartford/dolphin-2.1-mistral-7b.

[17] Mistral-7B-OpenOrca, https://huggingface.co/Open-Orca/Mistral-7B-OpenOrca.

[18] A. Ahmadian, S. Dash, H. Chen, B. Venkitesh, S. Gou, P. Blunsom, A. Üstün, S. Hooker, 'Intriguing Properties of Quantization at Scale', 2023, 32 p., https://arxiv.org/abs/2305.19268.

[19] B. Cornet, H. Fang, H. Wang, 'Overview of Quantum Technologies, Standards, and Their Applications in Mobile Devices', GetMobile: Mobile Computing and Communications, Volume 24, Issue 4, December 2020, pp. 5–9, doi: 10.1145/3457356.3457358.

[20] A new era of possibility with on-device AI, Qualcomm, 2023, https://www.qualcomm.com/products/technology/artificial-intelligence.

[21] V. Korthikanti, J. Casper, S. Lym, L. McAfee, M. Andersch, M. Shoeybi, B. Catanzaro, 'Reducing Activation Recomputation in Large Transformer Models', 17 p., https://arxiv.org/pdf/2205.05198.pdf.

[22] S. Gunjal, 'Understanding VRAM Requirements to Train/inference with Large Language Models (LLMs)', https://medium.com/@siddheshgunjal82/understanding-vram-requirements-to-train-inference-with-large-language-models-llms-a3edd0f09d9f.

[23] P. Dwivedi, 'Mixtral 8x7 – A better and cheaper alternative to ChatGPT', https://generativeai.pub/mixtral-8x7-a-better-and-cheaper-alternative-to-chatgpt-5c251b2e714d 3/12.

[24] G. Bao, Z. Ou, Y. Zhang, 'GEMINI: Controlling The Sentence-Level Summary Style in Abstractive Text Summarization', 9 December 2023, https://arxiv.org/pdf/2304.03548.pdf.

[25] Gemini: A Family of Highly Capable Multimodal Models, Gemini Team, Google, https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf.
[26] Hafsa Farooq, 'Google's Gemini – A multimodal model pushing the boundaries of AI', 7 December 2023. https://pub.aimind.so/googles-gemini-3e4f562e727f.
[27] V.I. Slyusar, 'Key aspects of the tensor-matrix theory of analysis and processing of multichannel measuring signals in the classical and neural network approaches', The 10th International Symposium on Precision Mechanical Measurement (ISPMM'2021), 15–17 October 2021, Qingdao, China, VTC. DOI: 10.13140/RG.2.2.31722.64966/1.
[28] V.I. Slyusar, 'End-face matrix products in radar applications', Izvestiya VUZ: Radioelektronika, 41 (3), 1998, pp. 71–75.
[29] V.I. Slyusar, 'New operations of matrix products for application of radars', IEEE MTT/ED/AP West Ukraine Chapter DIPED 1997 – Direct and Inverse Problems of Electromagnetic and Acoustic Theory, art. no. 710918, 1997, pp. 73–74, doi: 10.1109/DIPED.1997.710918.
[30] G. Hinton, 'Two Paths to Intelligence', 25 May 2023, Public Lecture, University of Cambridge, https://www.youtube.com/watch?v=rGgGOccMEiY.
[31] U.S. Pat. No. 10,097,318, 'Methods and systems for reliable broadcasting using re-transmissions', October 8 2018 – Trellisware Technologies, Inc., https://patentimages.storage.googleapis.com/b3/c0/e5/360f1245cd938c/US10097318.pdf.
[32] M. Tetiana, Y. Kondratenko, I. Sidenko, G. Kondratenko, 'Computer Vision Mobile System for Education Using Augmented Reality Technology', Journal of Mobile Multimedia 17/4, 2021, pp. 555–576.
[33] SAPIENT Interface Control Document, DSTL/PUB145591, 01-Feb-2023, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1144352/SAPIENT_Interface_Control_Document_v7_FINAL__fixed2_.pdf.
[34] SAPIENT autonomous sensor system, Last updated 20 April 2023, https://www.gov.uk/guidance/sapient-autonomous-sensor-system.
[35] O. Savage, 'NATO to adopt SAPIENT as C-UAS standard'. Janes, 25 September 2023, https://www.janes.com/defence-news/news-detail/nato-to-adopt-sapient-as-c-uas-standard.
[36] V.I. Slyusar, V.G. Smolyar, 'Communication channels frequency multiplexing on the basis of superrayleigh signals resolution', Izvestiya

Vysshikh Uchebnykh Zavedenij: Radioelektronika, 46 (7), 2003, pp. 30–39.
[37] A. Garza, M. Mergenthaler-Canseco, 'TimeGPT-1', 2023, 12 p., https://arxiv.org/pdf/2310.03589.pdf.
[38] A.I. Shevchenko, 'Natural Human Intelligence – The Object of Research for Artificial Intelligence Creation', International Scientific and Technical Conference on Computer Sciences and Information Technologies, vol. 1, 2019, pp. XXVI–XXIX, 8929799, CSIT 2019, Lviv, 17–20 September 2019.
[39] A. Shevchenko, M. Klymenko, 'Developing a Model of Artificial Conscience', 15th IEEE International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT'2020, vol. 1, 23–26 Sept. 2020, Lviv-Zbarazh, 2020, pp. 51–54.
[40] B. Dresp-Langley, 'Artificial Consciousness: Misconception(s) of a Self-Fulfilling Prophecy Nobody Wants', Qeios, December 2023, https://doi.org/10.32388/DW9JBP.2.
[41] R. Duro, Y.Kondratenko (Eds.), 'Advances in Intelligent Robotics and Collaborative Automation', River Publishers, Aalborg, Denmark, 2015, doi: https://doi.org/10.13052/rp-9788793237049.
[42] Y. Kondratenko, A. Shevchenko, Y. Zhukov, G. Kondratenko, O. Striuk, 'Tendencies and Challenges of Artificial Intelligence Development and Implementation', Proceedings of the 12th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, IDAACS'2023, Vol. 1, 2023, pp. 221–226, IDAACS 2023, Dortmund, Germany, 7–9 September 2023.

Biographies

Vadym Slyusar is a Doctor of Science, Professor, Honoured Scientist and Technician of Ukraine (2008). He has received a Ph.D. in 1992, Doctor of

Sciences in 2000, Professor in 2005. His research interests include radar systems, smart antennas for wireless communications and digital beamforming, artificial intelligence, and robotics.

Yuriy Kondratenko is a Doctor of Science, Professor, Honour Inventor of Ukraine (2008), Corr. Academician of Royal Academy of Doctors (Barcelona, Spain), Head of the Department of Intelligent Information Systems at Petro Mohyla Black Sea National University (PMBSNU), Ukraine. He has received (a) a Ph.D. (1983) and Dr.Sc. (1994) in Elements and Devices of Computer and Control Systems from Odessa National Polytechnic University, (b) several international grants and scholarships for conducting research at Institute of Automation of Chongqing University, P.R.China (1988–1989), Ruhr-University Bochum, Germany (2000, 2010), Nazareth College and Cleveland State University, USA (2003), (c) Fulbright Scholarship for researching in USA (2015/2016) at the Dept. of Electrical Engineering and Computer Science in Cleveland State University. Research interests include robotics, automation, sensors and control systems, intelligent decision support systems, and fuzzy logic.

Anatolii Shevchenko graduated from the Faculty of Physics at Donetsk State University with a major in Radio Physics and Electronics. In 1985,

he defended his Ph.D. thesis, earning the academic degree of a candidate. In 1990, he completed his doctoral thesis, obtaining the academic degree of Doctor of Technical Sciences. In 1997, he was awarded the title of professor, and in 1998, he received the honorary title of Honored Scientist and Technician of Ukraine. In 2006, was elected as a Corresponding Member of the National Academy of Sciences of Ukraine in the field of Computer Systems. In 2015, he was appointed as the Director of the Institute of Artificial Intelligence of the Ministry of Education and Science of Ukraine and the National Academy of Sciences of Ukraine in Kyiv. Together with the Department of Informatics of the National Academy of Sciences of Ukraine, he initiated an international scientific journal called "Artificial Intelligence" and was appointed as its chief editor. His research interests encompass various aspects of artificial intelligence, modeling human intelligence, simulating elements of human consciousness, breakthrough technologies in the field of artificial intelligence, and multidisciplinary aspects of artificial intelligence.

Tetiana Yeroshenko obtained her Ph.D. degree in 2016 in the field of «Philosophy of Science» from the H.S. Skovoroda Institute of Philosophy (Kyiv, Ukraine). She is a research scientist at the Department of Theoretical Research in the field of artificial intelligence at the Institute of Artificial Intelligence of the Ministry of Education and Science of Ukraine and the National Academy of Sciences of Ukraine (Kyiv). Her research interests include the philosophy of science and multidisciplinary aspects of artificial intelligence.