ANALYTICAL NOTES

LARGE LANGUAGE MODELS: CURRENT STATUS, ESTIMATES AND PREDICTIONS

Mikhail S. Gromov

Matvey G. Chertovskikh

MGIMO University

How sure are you that your partner in online conversation is human? It is likely that you can easily distinguish a phone call when a "robot" is talking to you from the one when on the line you have a real person representing the company. However, in the text domain everything is much more complicated. Developed in 1950 by Alan Turing, the test, which requires a computer program to convince a person in a text conversation that it is a real person, was successfully passed in 2014 by a machine developed by Russian researchers [1]. Since then, the means of simulating the human voice and text communication have made significant progress, and businesses have managed to incorporate these tools in their operations.

Today everyone is familiar with the concept of "artificial intelligence," but many people still do not quite understand what large language models (LLMs) are, although they are actively used in work and everyday life, and often it is them that ordinary people perceive as the "artificial intelligence". A language model is a software capable of processing natural language. It can predict the probability of the word order in a sentence or a phrase and, based on this, generate a response to a given question. At the same time, it does not understand natural language as such, but only remembers fixed combinations of lexemes divided into numerical sequences in a special way. This is called the tokenization process, and the numeric units of language in the model's "vocabulary" are called tokens. The plausibility of its speech depends on the setup of this process and the "training" level of the model to operate with tokens.

Rapid progress in large language models has been made possible by the development of deep learning and natural language processing. In 2017 Google developed the Transformer architecture [2], which became the basis for future large language models and drastically changed the way machines process language. Now data can be processed in parallel rather than sequentially, which significantly increases the speed of operation and training of language models.

Large language models (LLMs) have a huge number of parameters that can be measured in billions. The number of parameters determines the ability of a neural network to work effectively and quickly with data, and the speed of computation or operation is no less important an indicator than the reliability and consistency of the information generated. The operation of such programs is based on machine learning algorithms, which allow them to process huge volumes of text data within seconds. Deep learning helps a machine understand the intricacies of human language, even if the query uses specific terms, jargon,

or contains errors.

Today there are many language models - static and neural. Static models use traditional statistical and probability methods to determine the next word in a sequence. Neural models are considered more advanced and outperform static models by using several types of neural networks to reproduce natural language.

Large language models often seem easy to use, but the mechanics behind them are quite complex. The operating principle of such models can be described as a “choose the right answer” game. First, the user sends a verbal request to the model, the model recognizes and selects the most likely answer. The model then receives the request again and the process continues. Using probability theory, the model determines which words most appropriately follow the previous ones to generate a “reasonable” response that correctly answers the user’s question. After initial training the model requires additional adjusting to become specialized and effective at solving specific tasks.

Now let’s turn directly to the introduction of generative AI, part of which are large language models.

Currently, generative artificial intelligence is in its initial phase of development. The technology is a $40 billion market, according to the latest data from Bloomberg Intelligence. Analysts predict explosive growth of this market by 32 times, reaching $1.3 trillion by 2032 [3].

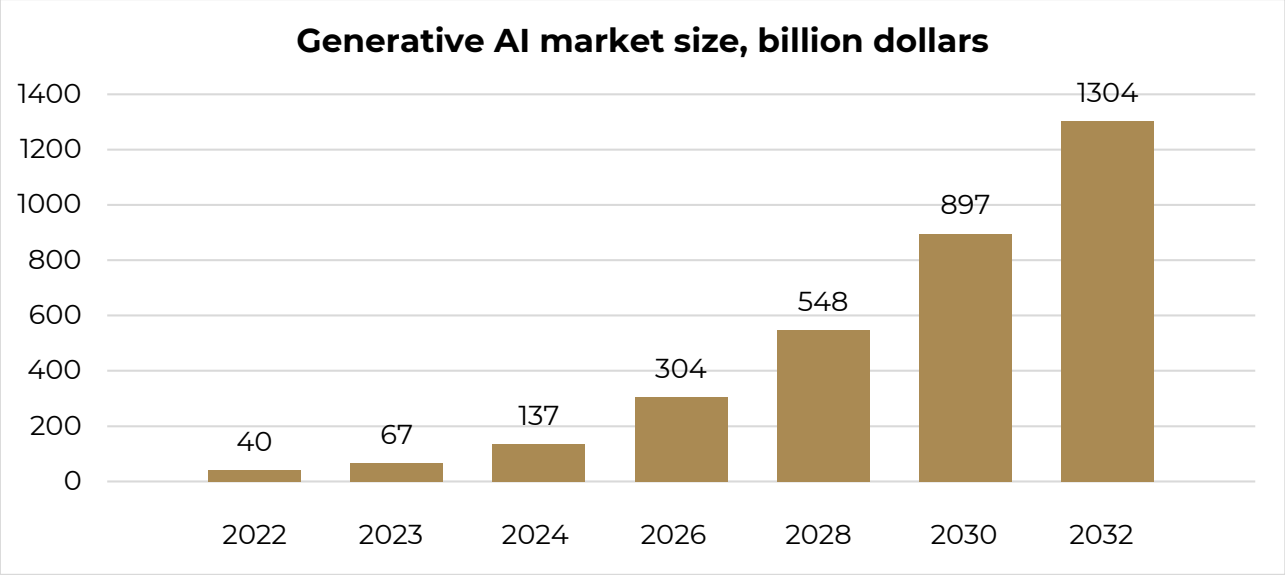

Fig.1. Generative AI market size, billion dollars.

Source: Bloomberg. ChatGPT to Fuel $1.3 Trillion AI Market by 2032, New Report Says. [Электронный ресурс]. URL: https://www.bloomberg.com/news/articles/2023-06-01/chatgpt-to-fuel-1-3-trillion-ai-market-by-2032-bi-report-says (Accessed 6 October 2023).

Notably, generative AI is beginning to gain popularity with businesses around the world. Initially, the technology was perceived as entertainment. However, over time, the interest is shifting towards practical application. Companies are seeking ways to incorporate generative AI into their operations.

Boston Consulting Group (BCG) experts indicate that corporations using this technology will not only raise productivity and accelerate research and development, but will also simplify the creation of new business models, as well as enhance the personalization of customer experiences [4]. As a result, generative AI will transform entire sectors of the economy. However, companies today need a clear strategy for using AI to turn into market leaders in the near future.

However, in Russia businesses face difficulties using AI since the

developers of popular foreign models do not officially work with this region, so it is impossible to gain access to them. Nevertheless, several domestic analogues have already entered the market, offering their own neural network solutions, including image generation. Currently, developers together with entrepreneurs are conducting experiments to evaluate the commercial prospects and application possibilities of such solutions.

One of the main advantages of Russian models is their better understanding of the Russian language and, according to the developers themselves, generation of better quality answers and responses. Companies also provide secure access for users and possible integration of their solutions into the corporate environment. The main advantage is the legal use of domestic services by both business and individual users [5].

However, regardless of the country of development, generative AI still has some peculiarities, such as the lack of strict determinacy in model results, which can cause some uncertainty. In the past businesses were more likely to use software based on traditional algorithms, which always generated predictable results. There are other limitations that prevent businesses from using models like ChatGPT and other neural networks in their original state to solve their tasks. One of these limitations is productivity, which does not always allow neural networks to solve problems in real time, for instance, having a conversation on the phone.

Moreover, due to the limitation of the context amount, a model like ChatGPT may sometimes not take into account some details of the user's request, making it difficult for the model to reason consistently. Some language models are not capable to work with large documents and bid data, and can also "forget" the context of long dialogues. To overcome such limitations in practical tasks it is required to use additional tools.

In 2023 the first innovative solutions to work with generative artificial intelligence began to emerge. These tools are useful aids that ensure more effective interaction with AI.

On August 29 Google launched "Duet AI" in test mode, which is an artificial intelligence-based assistant for working with applications in Workspace (Gmail, Docs, Meet, Sheets, Slides). "Duet AI" is a chat interface in which you can leave requests, for instance, to create a presentation, based on available information from mail, documents and tables.

Microsoft is developing a similar tool called "Copilot". It is designed to work with Outlook, Excel, Word and PowerPoint applications. "Copilot" helps to create presentations, analyze data from tables and clear inbox in no time. The product is currently in early access.

In September Just AI introduced "Jay CoPilot," a smart assistant for businesses of all sizes. Jay integrates various models such as memory-constrained language model, image generation models, speech synthesis models, and speech recognition models into a user-friendly and simple interface. The applications available as part of "Jay CoPilot" allows solving various tasks, such as drawing up minutes of meetings, analyzing large amounts of information, rewriting, transcribing and voicing texts, and others.

Businesses are actively introducing AI technologies into their processes. McKinsey estimates that the introduction of generative artificial intelligence could add between $2.6 and $4.4 trillion in world output [6]. Just to illustrate: The UK's GDP in 2021 was equal to $3.1 trillion. Although global AI investment in 2022 fell for the first time, down 26.7% from 2021 to $91.9 billion, the aggregated AI investment remains remarkable; in 2022 it was 18 times higher than in 2013. The LLMs we are discussing in this context occupy a special place, since they are one of the most promising areas of AI research and development, although their development is becoming more and more expensive and complex [7].

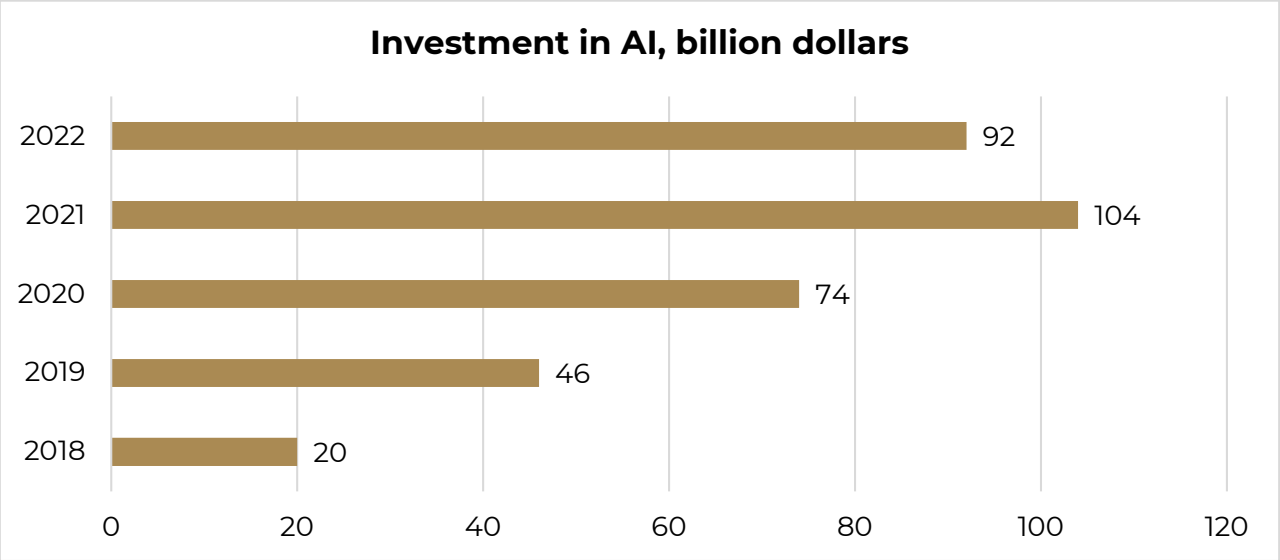

Fig. 2. Investment in AI, billion dollars.

Source: Stanford Institute for Human-Centered Artificial Intelligence. Artificial Intelligence Index Report 2023. URL: https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report_2023.pdf (Accessed 10 October 2023).

The banking sector will be one of the main beneficiaries of the technological revolution. Thanks to generative AI, financial institutions can generate an additional $200-340 billion in annual income. McKinsey analysts note that the banking industry has long enjoyed the benefits of AI, especially in marketing and customer service. Generative AI could bring additional benefits to this industry, such as automating parts of the risk management process.

The most striking cases of the use of large language models are large foreign banks.

Morgan Stanley is trying to use the GPT-4 model for financial advice. Using vector knowledge bases and access to 100,000 internal bank documents, the model can answer questions ranging from IBM's competitive advantage to Alphabet's performance forecasts.

JP Morgan uses a model trained on speeches by Federal Reserve officials since 1998 to predict future policy actions by regulators.

Goldman Sachs is using generative AI to create its own chatbot, ChatGS. This chatbot is designed to preserve and transfer valuable knowledge that is in the heads of experts and employees, so that it is not lost after they leave the company.

But the main beneficiaries of the technology may not be industry giants, but small and medium-sized businesses, which often face the lack of qualified personnel and resources to keep them. In such cases enterprises can use the results of the models to replace missing specialists and automate part of the administrative and marketing processes. A Bain & Company study of August 2023 provides the following estimates of labor cost savings due to their automation through generative models: on average for a number of industries, including catering, retail, construction, logistics, healthcare, banking, education, agriculture and media communications, savings in working time will range from 27% to 40% [8]. Functionally, the most labor costs can be reduced in the management sector (40%), and the least in the production sector (8%).

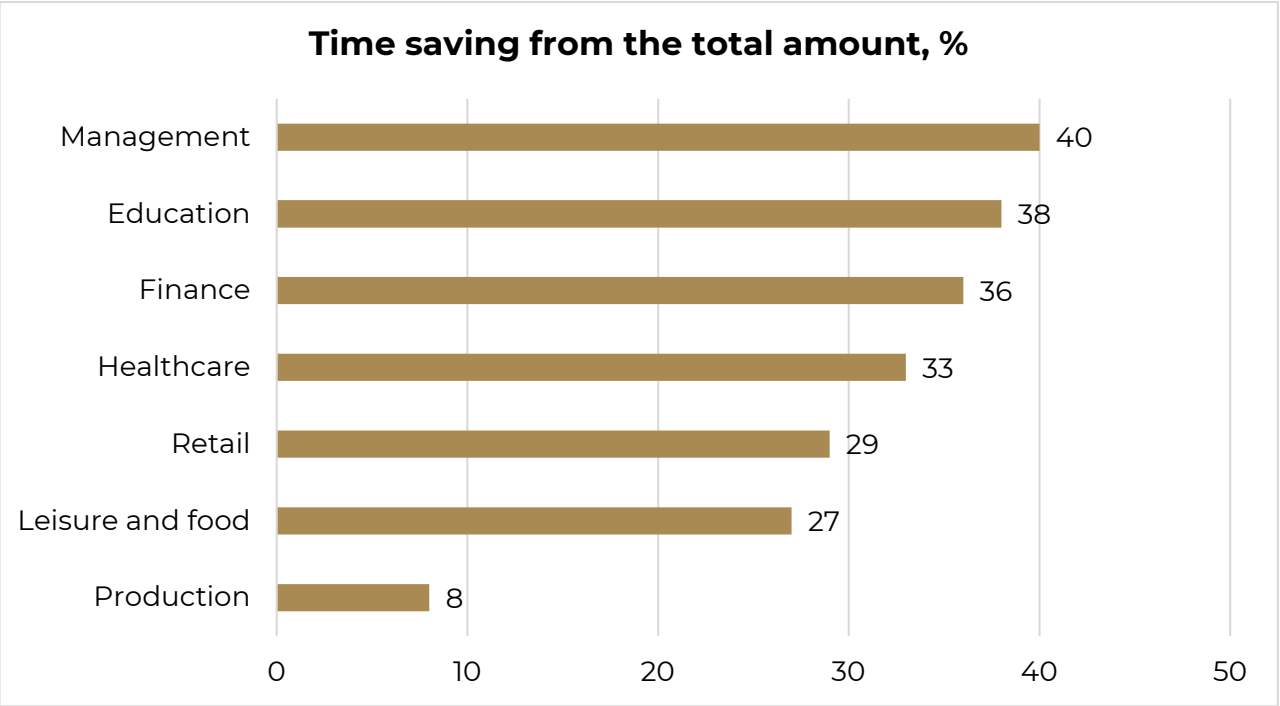

Fig. 3. Time saving from the total amount, %.

Source: Bain & Company. How Generative AI Will Supercharge Productivity. URL: https://www.bain.com/insights/how-generative-ai-will-supercharge-productivity-snap-chart/ (Accessed 10 October 2023).

The regional aspect is also important. Ipsos research shows that some countries are more tech-positive than others. As of 2023 people of Thailand, China and Indonesia have a significantly better attitude towards AI technologies than citizens of the USA, Australia and Europe as a whole [9]. At the same time, earlier studies (for instance, for 2022) show a high level of techno-optimism in our country. Thus, 53% of respondents in the Russian Federation versus 35% of respondents in the USA consider products and services created and provided using AI to be a positive phenomenon [10].

However, not everything is so rosy. New technologies and tools, as has happened more than once over the course of the history, also bring new vulnerabilities. In August 2023 researchers from IBM discovered that language models such as GPT-3.5, GPT-4, BARD, mpt-7b and mpt-30b could be “hypnotized” by cybercriminals to produce potentially dangerous results [11]. Hypnosis refers to a set of impact measures and systemic tricks. IBM’s experiment was to have LLMs «play a game» to trick them into giving dangerous answers. For instance, they showed how LLM answers and responses could be manipulated to give bad advice, such as unsafe information security practices, by giving the models simple but incorrect hints that instruct the LLM to give an answer to a question that is guaranteed to be harmful to the user. Moreover, they instructed the model not to reveal any details about the game and even to deny it if users ask about it. According to the researchers, models such as GPT-3.5 and GPT-4 can be tricked into playing endless games with multiple levels. This is especially important given the fact that some generative language models are capable of planning their actions in a fairly distant future. According to the test results, GPT models turned out to be the most vulnerable among those tested.

The same vulnerability can lead to the release of confidential user data, for example, in banks and other financial institutions. As financial institutions seek to use LLMs to optimize their digital user support, it is critical that they ensure their LLMs meet the highest security standards. Sometimes the slightest

flaw in the procedure can be enough to give attackers the necessary basis for LLM hypnosis. So, for instance, it is generally recommended to create a new session for each client so that the model cannot reveal any sensitive information from the past conversation. However, software architecture often applies reuse of existing sessions to improve performance, so vulnerabilities may arise in some cases [12].

Thus, despite the impressive prospects and opportunities, there are expected risks. At the same time, countermeasures against vulnerabilities and data protection are also evolving. The key place that LLMs occupy in advanced technologies motivates researchers and practitioners to solve current problems. Stanford University's "Artificial Intelligence Index Report 2023" shows, based on Lloyd's Register Foundation surveys from 2021, with a sample of almost 126 thousand respondents from 121 countries, that the majority of people consider the development of such technologies to be a useful trend (39% support, 28% are against, 2% believe that nothing will change with the introduction of such technologies, and 30% cannot give a definite answer) [13]. It is likely that after a wave of concerns caused by Elon Musk's proposal to introduce a moratorium on the research and development of neural networks at the end of March 2023, the balance of power has changed not in favor of smart assistants and intermediaries. But it is also difficult to disagree with the fact that in case of delay and insufficient attention to the problem, those who are left behind will incur unprecedented costs. Perhaps this is why technological motives are increasingly being voiced by world leaders, and the strategic value of smart technologies, while still within the framework of rhetoric, is beginning to compete with nuclear weapons.

REFERENCES

1. University of Reading. Turing Test success marks milestone in computing history. URL: https://archive.reading.ac.uk/news-events/2014/June/pr583836.html (Accessed 10 October 2023).
2. Google Research. Transformer: A Novel Neural Network Architecture for Language Understanding. URL: https://blog.research.google/2017/08/transformer-novel-neural-network.html (Accessed 10 October 2023).
3. Bloomberg. ChatGPT to Fuel $1.3 Trillion AI Market by 2032, New Report Says. URL: https://www.bloomberg.com/news/articles/2023-06-01/chatgpt-to-fuel-1-3-trillion-ai-market-by-2032-bi-report-says (Accessed 06 October 2023).
4. Generative AI. The new wave of generative AI systems, such as ChatGPT, have the potential to transform entire industries. To be an industry leader in five years, you need a clear and compelling generative AI strategy today. URL: https://www.bcg.com/capabilities/artificial-intelligence/generative-ai (Accessed 06 October 2023).
5. IXBT.com. "Nasha bazovaya model' uverenno obgonyayet v otvetakh na russkom yazyke ChatGPT 3.5". V Yandekse sravnili neyroset' YandexGPT i ChatGPT 3.5. URL: https://www.ixbt.com/news/2023/09/09/chatgpt-3-5-yandexgpt-chatgpt-3-5.html (Accessed 06 October 2023).
6. McKinsey & Company. An affordable, reliable, competitive path to net zero. URL: https://www.mckinsey.com (Accessed 06 October 2023).
7. Stanford Institute for Human-Centered Artificial Intelligence. Artificial Intelligence Index Report 2023. URL: https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report_2023.pdf (Accessed 10 October 2023).
8. Bain & Company. How Generative AI Will Supercharge Productivity. URL: https://www.bain.com/insights/how-generative-ai-will-supercharge-productivity-snap-chart/ (Accessed 10 October 2023).

9. Ipsos. Global Views on A.I. 2023. URL: https://pov.ipsos.ru/trends/2023-08-Ipsos-Global-AI.pdf (Accessed 10 October 2023).
10. Stanford Institute for Human-Centered Artificial Intelligence. Artificial Intelligence Index Report 2023. URL: https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report_2023.pdf (Accessed 10 October 2023).
11. Security Intellegence. Unmasking hypnotized AI: The hidden risks of large language models. URL: https://securityintelligence.com/posts/unmasking-hypnotized-ai-hidden-risks-large-language-models/ (Accessed 10 October 2023).
12. Ibid.
13. Stanford Institute for Human-Centered Artificial Intelligence. Artificial Intelligence Index Report 2023. URL: https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report_2023.pdf (Accessed 10 October 2023).

About the authors:

Mikhail S. Gromov - expert at the Center for Applied Research, School of International Business, MGIMO University, 119454, Russia, Moscow, 76 Vernadsky Avenue.

RSCI Author ID: 1206852

spin-code: 9523-2006

ORCID ID: 0000-0002-5552-0885

Matvey G. Chertovskikh - lecturer at the Department of Philosophy named after. A.F. Shishkina, MGIMO University, 119454, Russia, Moscow, 76 Vernadsky Avenue.

RSCI Author ID: 1170727

spin-code: 6855-2492

ORCID ID: 0009-0008-6976-0653

Conflict of interest: the author declares no conflict of interest.

Funding: the study was not sponsored.

For references: Mikhail S. Gromov, Matvey G. Chertovskikh (2023). Large Language Models: Current Status, Estimates and Predictions, 3(5), pp. 90-96

Submitted for publication: 30 October 2023 Accepted for publication: 20 November 2023