IUC WORKING PAPER SERIES

INTER UNIVERSITY CENTRE DUBROVNIK

Reference Number 2026/01/MST

Science and Practice of Language Models: An Industry Perspective

Tvrtko Tadić

Domain: Mathematics, Science and Technology;
IUC course/conference: IIIrd. Kathy Wilkes Memorial Conference; The Behavioral and Ethical Consequences of Large Language Models (LLM), 2024.
Category: overview article
Submitted: 3 June 2025
Date of publication: 16 January 2026
DOI: https://doi.org/10.53099/ntkd4313
ISSN: 2787-5717

Inter-University Centre for Advanced Studies (IUC)
E-mail: iuc@iuc.hr / web site: https://www.iuc.hr/
Don Frana Bulića 4, 20000 Dubrovnik, CROATIA
Tel: 00 385 (0)20 413 626

IUC Working Paper Series

The IUC Working Papers aim to generate interest for and debate on the ideas and research outcomes presented at IUC-activities. The IUC Working Papers is an inclusive platform for sharing the output of IUC courses, conferences and research meetings. The papers can vary along the whole range of disciplines and subjects addressed in these IUC-activities. Only participants in IUC-activities are eligible to submit and publish papers -or later versions of papers- that were presented during an IUC-activity.

Author Tvrtko Tadić is a Senior Applied Scientist at Microsoft, where he works on integrating real-time data into AI systems, with a focus on advanced file search in M365 Copilot. He holds a PhD in Mathematics from the University of Washington and a degree in Mathematical Statistics and Computer Science from the University of Zagreb. Dr. Tadić is also a Visiting Scientist at the University of Washington. His research has appeared in venues including the *Annals of Applied Probability*, *Bernoulli Journal*, *Memoirs of the American Mathematical Society*, and the *International Conference on Learning Representations*. He is the author of two books and co-inventor of a patent on a Smart Communication Assistant with Audio Interface.

Author details: *Tvrtko Tadić, One Microsoft Way, Redmond, WA 98052, United States, tvrtkota@microsoft.com*

Science and Practice of Language Models: An Industry Perspective [1]

Tvrtko Tadić

Microsoft Corporation, Redmond

Abstract

In this paper, we provide an overview of language models - their underlying mechanisms, capabilities, development within the industry, and the challenges they present. The launch of ChatGPT by, at the time relatively unknown company, OpenAI in late 2022 marked a turning point, achieving unprecedented success and, in the eyes of many, heralding the dawn of the artificial intelligence (AI) era. This breakthrough set off an intense race among the world's leading technology companies to develop increasingly sophisticated AI-driven products. Meanwhile, advancements in language models and other AI technologies continue at a rapid pace, with new and improved systems emerging daily. In this article, we explore the essence of language models, their applications, and their current state of development.

Key words: Large Language Models, AI, word prediction, data, models, cloud infrastructure, ChatGPT, transformers, search engines

[1] This article is based on the lecture Large Language Models: What Has Changed with the Development of ChatGPT? as part of the Engineering Section of the Croatian Mathematical Society held on June 25, 2023 in Zagreb (a recording of which is available in [7]) and the lecture Science and Practice of Large Language Models: An Industry Perspective held on April 26, 2024 at the 3rd Kathy Wilkes Memorial Conference in Turin. A modified Croatian-language version of this paper ([31]) was published in Poučak, the mathematics education journal of the Croatian Mathematical Society.

Searches and content summarization

We have long been used to searching for content using search engines. Usually, after entering a term, the search engine would find web pages and the user would review them to determine if they contained what they were looking for. Search engines were also able to answer simple questions. ChatGPT's appearance took things to a higher level. For the first time, users were able to ask more complex queries and follow-up questions. The chat was able to compose text in the form and length that the user wanted. There were two drawbacks:

The Washington Post

This article was published more than **3 years ago**

BUSINESS

Microsoft invests $1 billion in OpenAI to pursue artificial intelligence that's smarter than we are

By Taylor Telford

July 22, 2019 at 3:25 p.m. EDT

Figure 1. Article (Telford, 2019) title from Washington Post on Microsoft's 2019 investment into OpenAI

- The information that ChatGPT had access to was available up to September 2021.
- Occasionally, it would generate text that was inaccurate or completely fabricated. A phenomenon known as **hallucination**.

In February 2023, Microsoft, with the help of OpenAI, introduced Bing AI ((Hao, 2023)), which combines the capabilities of ChatGPT and an internet search engine. A few months later, Google announced its version called Bard ((De Vynck & Tiku, 2023)). The race to develop AI products has begun.

OpenAI

OpenAI is a laboratory for the development of AI that is safe and beneficial for humanity. It has two components, one of which is a non-profit organization, while the other is a limited liability company ((contributors, OpenAI, 2025)). Microsoft invested one billion dollars in 2019 ((Telford, 2019)) and an additional 13 billion dollars in 2023 ((Petrova, 2024)), according to media reports. OpenAI has developed some of the most advanced AI models, such as GPT-3 and GPT-4, which can generate natural language based on text or images.

What are (large) language models?

Language models have been studied for a long time. The first language models appeared in the 1950s and 1960s as part of research in the field of machine translation and speech recognition. However, it was only with the appearance of ChatGPT that they came into the public spotlight. Language models are mathematical representations of natural language that enable machines to understand and generate text. They are based on statistical or neural methods that learn from vast amounts of textual data. For more details on history and development of language models see survey (Wang, et al., 2024).

Mathematical framework

We will now describe the mathematical framework for language models, most of which is based on the notes (Cotterell, Svete, Meister, Liu, & Du, 2023). The alphabet is a finite non-empty set

$$\Sigma = \{\sigma_1, \sigma_2, \dots, \sigma_k\}$$

whose elements are called symbols. A text is a finite sequence of symbols. A language model is a probability distribution $\mathbb{P}$ on the set of all finite texts:

$$\Sigma^* = \{\sigma_1, \dots, \sigma_k, \sigma_1\sigma_1, \sigma_1\sigma_2, \dots, \sigma_k\sigma_k, \ \sigma_1\sigma_1\sigma_1, \dots\}.$$

$\mathbb{P}$ is a model that contains all the data about the language. In the next (simplified) example we will see how this works in practice.

Example

Let alphabet be given by

$$\Sigma = \{\text{"After the"}, \ \text{"rain"}, \text{"snow"}, \ \text{"it will"}\}.$$

We can concentrate $\mathbb{P}$ so that the following things happen:

- Text has 4 symbols
- Text starts with “After the”.
- Second and fourth symbol are from the set {"rain", "snow"}
- Third symbol is „it will“.
-

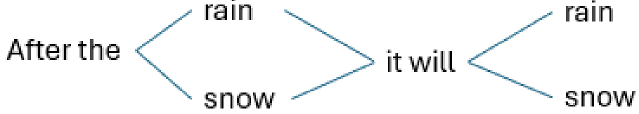


Figure 2. Schema of a simple language model

As shown in Figure 2, in this model, the second and fourth symbols actually appear according to some probability distribution, while the first and third symbols are fixed.

Simple implementation

After the user enters their text, the application will:

- Estimate the probability distribution of the text continuation conditionally on the entered text, which we denote as

$$\mathbb{P}(\cdot\, |\text{user's input}).$$

- Simulate a sample text from this distribution and return the generated text to the user.

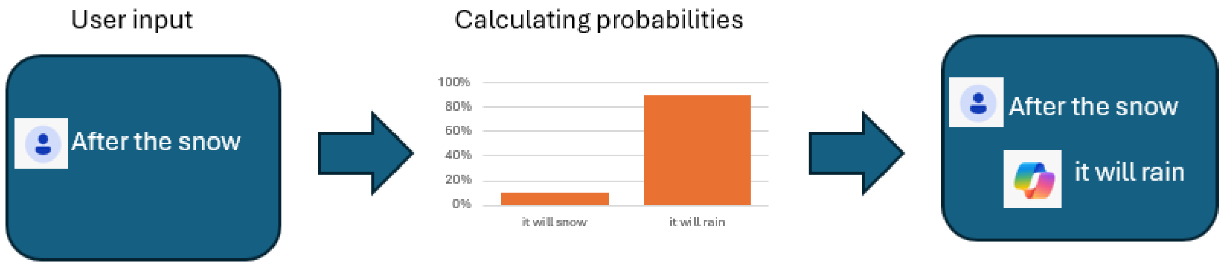


Figure 3. Example of a simple implementation of a language model.

How do they work?

After entering the text, the application sends a query to the cloud where the model generates a response and sends it back to the application. In the case of all models developed by OpenAI, they are all executed on Microsoft's cloud computing platform – Azure.

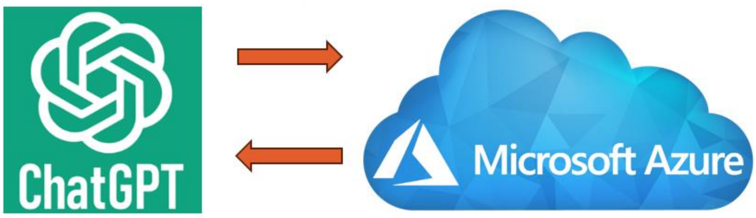
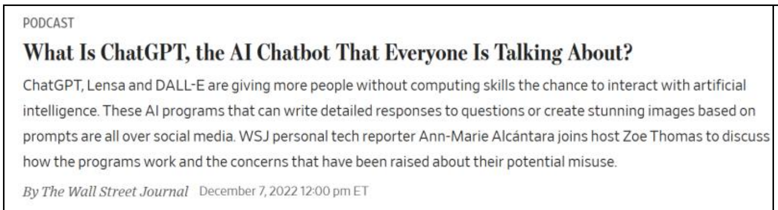


Figure 4. Operating scheme of ChatGPT

According to estimates and media reports, a query on ChatGPT is 4 to 7 times more expensive than a traditional search (see Figure 6).



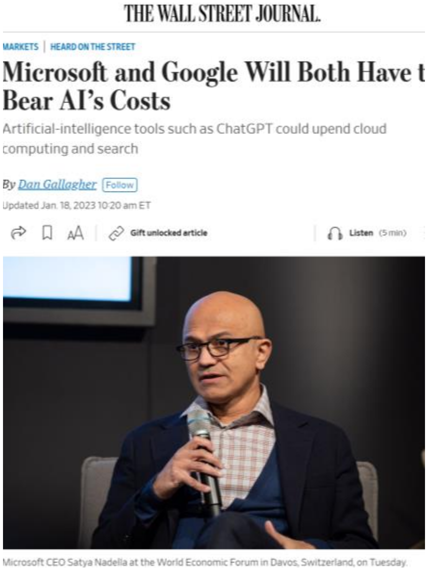
PODCAST

What Is ChatGPT, the AI Chatbot That Everyone Is Talking About?

ChatGPT, Lensa and DALL-E are giving more people without computing skills the chance to interact with artificial intelligence. These AI programs that can write detailed responses to questions or create stunning images based on prompts are all over social media. WSJ personal tech reporter Ann-Marie Alcántara joins host Zoe Thomas to discuss how the programs work and the concerns that have been raised about their potential misuse.

By The Wall Street Journal December 7, 2022 12:00 pm ET

Figure 5. The title of the podcast (Editors, 2022) *published by the Wall Street Journal*



THE WALL STREET JOURNAL.

MARKETS | HEARD ON THE STREET

Microsoft and Google Will Both Have t Bear AI's Costs

Artificial-intelligence tools such as ChatGPT could upend cloud computing and search

By Dan Gallagher Follow

Updated Jan. 18, 2023 10:20 am ET

Gift unlocked article Listen (5 min)

Microsoft CEO Satya Nadella at the World Economic Forum in Davos, Switzerland, on Tuesday.

Figure 6. The article (Gallagher, 2023) *in the Wall Street Journal about the costs of artificial intelligence*

Model development

Like any computer application, applications powered by language models go through offline and online development phases. The online phase, which occurs after the model is made available to a portion of users, attempts to measure how users interact with the application. This phase is common for all online applications.

The offline phase is where the model parameters are estimated, which in the terminology of machine learning and AI is called **model training**.

For the model training process, the following is needed (more details are available in textbooks like (Murphy, 2012)):

- **Data**
- **Model** - selecting the architecture, parameter initialization method, optimization algorithm, etc.
- **Infrastructure** - the entire process is usually parallelized across multiple processors in the cloud for acceleration.

Data

Large language models require vast amounts of data. The data comes from various textual sources. These sources do not necessarily have to be factually accurate. It could be an

intentionally or unintentionally fabricated story or outdated information. The quality of available data from textual sources will vary.
Internet search engines like Google and Bing have long been able to provide simple answers to questions. Figure 7 shows an example query about who the mayor of New York City is. There is a wealth of information about this on the Internet, but the experience of working on search engines gives companies that developed them better insight into where quality data can be found.

Figure 7. Answer to the question "Who is the mayor of New York City?" provided by Microsoft Bing search engine

In industry the English-language Wikipedia is considered as a high-quality and up-to-date source of information. Many reputable publishers also have reliable content. Some of this content is publicly available on the publishers' websites. While Wikipedia is an encyclopedia whose content is not owned by anyone, publishers have invested significant money in their content and see that ChatGPT and other products based on large language models provide information about that content. This has led to lawsuits, such as the one initiated by The New York Times against OpenAI and Microsoft over the use of copyrighted content [2]. On the other hand, some companies enter into agreements to access publishers' content. Sometimes these agreements are exclusive (as mentioned in [3]) – a technology company enters into an agreement with a publisher to have exclusive access to the publisher's publications for the purposes of its products.

Models

To give the reader an idea of the concepts and challenges of the models used for large language models, we will mention a few examples and challenges they bring.

n-gram models

These models are from the early phase of language models. The idea is to try to predict the rest of the text based on the last n words of the user's input. This idea is based on the well-researched theory of Markov models. More on these models can be found in (Jelinek, 1997), here we will outline the main ideas.

For a given user input $\sigma_1\sigma_2 \dots \sigma_k\sigma_{k+1} \dots \sigma_{k+n}$ the probability distribution of the output depends only on the last n words of the input. In mathematical terms, this holds:

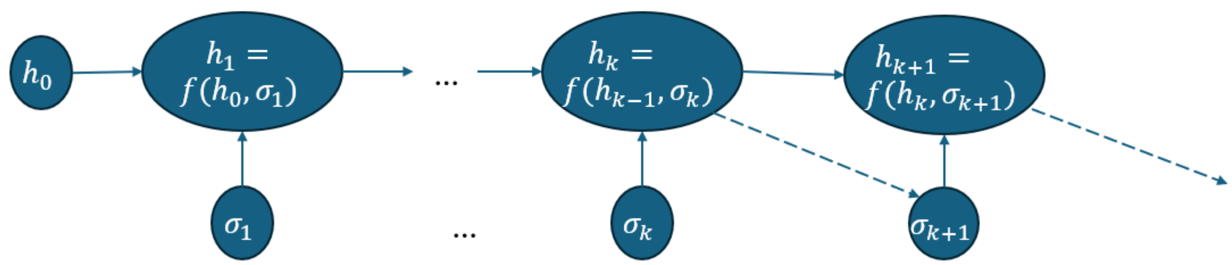
$$\mathbb{P}(\cdot\,|\sigma_1\sigma_2 \dots \sigma_k\sigma_{k+1} \dots \sigma_{k+n}) = \mathbb{P}(\cdot\,|\sigma_{k+1} \dots \sigma_{k+n}).$$

Two examples will illustrate how this approach works and its shortcomings. Let n = 3

- Let's assume that the user entered 'On Tuesday, *after the rain, there will be'*. Since this is a model that looks at the last 3-gram, the only part of the input that will be considered will be *'after the rain, there will be'*. So it is very likely that the language model will complete this sentence with 'snow'.
- A small modification at the beginning can significantly change what the output should be, which n-gram models will ignore. For example, user input like 'It will be hot on Tuesday, after the rain, there will be' with the same probability as in the previous point, can return 'snow', even though when looking at the full sentence it should not be so.

Recurrent language models

The idea of this method is to process the entire input $\sigma_1\sigma_2 \dots \sigma_k$ before producing the next symbol. Everything is shown in Figure 8. All processed data is stored in so-called hidden states – a series of high-dimensional vectors (h_j) that are recursively calculated using the formula $h_j = f(h_{j-1}, \sigma_j)$. The next symbol σ_{k+1} simulated based on the state h_k. The challenge is to estimate the function f.



as the inability for long-term memory and parallelization[2].

Transformer-based models

Transformers have solved these problems by introducing the attention mechanism, which allows models to process all words in a sentence simultaneously and focus on relevant parts of

[2] In computing, parallelization is a process in which the execution of a process is attempted to be divided among multiple computers in order to complete the process faster and more efficiently. In this case, due to dependence on previous states, to calculate the state h_k we must calculate all the previous values $h_1, \dots, h_{k-1}$.

the text regardless of their distance. Introduced in the famous NeurIPS[3] paper (Vaswani, 2017) named *Attention is all you need*, these types of models resulted in faster processing and better context understanding. The advantages of transformers include faster execution, more efficient handling of long sequences, and superior performance in tasks such as translation, text generation, and sentiment analysis. The scheme, which we will not go into detail, is given in Figure 9. From the diagram, it can be seen that each state h_k depends exclusively on the input $\sigma_1, \dots, \sigma_k$, but not on other states $h_1, h_2, \dots, h_{k-1}$.

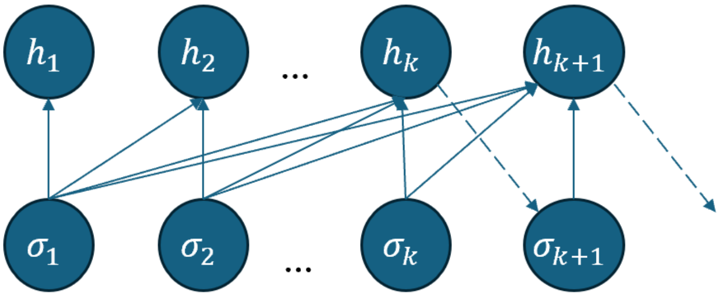


Figure 9. Transformer architecture

Infrastructure

Parameter estimation is a complex operation. Billions of documents need to be stored and processed somewhere. A special computing infrastructure is required - a so-called supercomputer. Training takes days, weeks, and sometimes even months. It is very likely that some of the servers in the data center will go offline. The entire process is carried out in multiple stages, and sometimes it is necessary to start a particular stage in a different data center than the previous one. Figure 10 illustrates multiple stages of ChatGPT training in various data centers within Microsoft's Azure cloud computing platform.

[3] NeurIPS is one of the leading international conferences where researchers and industry leaders showcase advances in artificial intelligence and machine learning.

Figure 10. Illustration of training ChatGPT in stages and data centers around the world. Illustration taken from (Microsoft, 2023).

The demanding infrastructure required for training large language models raises the question of how to involve the academic community in their development. Currently, only a few large companies with the necessary infrastructure can work on their development. Table 1 provides an overview of some well-known models and who developed them.

Year	**Model name**	Number of parameters	**Company**
2018	BERT	$340 \cdot 10^6$	Google
2019	GPT-2	$1.5 \cdot 10^9$	OpenAI
2020	GPT-3 (GPT-3.5 -> ChatGPT)	$175 \cdot 10^9$	OpenAI
2021	Megatron-Turing NLG	$530 \cdot 10^9$	Microsoft & Nvidia
2022	LaMBDA	$137 \cdot 10^9$	Google
2023	GPT-4 (Bing AI, ChatGPT Plus)	$\approx 10^{12}$	OpenAI
2023	Palm 2 (Google Bard)	$340 \cdot 10^9$	Google

Table 1. Overview of some popular models developed between 2018 and 2023 taken from (contributors, List of large language models, n.d.).

As we can see from the table, most models have been developed by large technology companies. Due to the high cost of creating such models, the question arises of how to involve the academic community so that development does not remain exclusively within large companies.

Model Evaluation

For all machine learning and AI models, an assessment must be made to determine whether they perform satisfactorily and better than another model or a previous version.

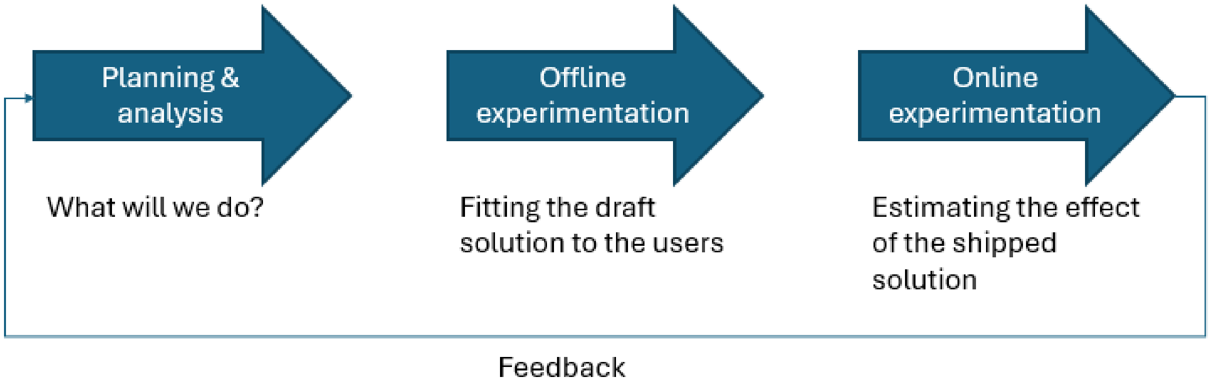


Figure 11. Development cycle of an AI product

Offline Method

After building the model, before exposing it to users, thorough testing is conducted on a dataset that was not used during training. This testing includes evaluating the model's accuracy and performance. Specific techniques such as *perplexity* ((Jelinek, 1997)), BLEU ((Papineni & al, 2002)), and ROUGE ((Lin, 2004)) have been developed for this purpose, allowing for detailed analysis and assessment of the quality of the generated text.

Online Method

AI models must meet several conditions for users to have good experience with them:

- They must serve millions of users.
- They must operate 24 hours a day, 7 days a week.
- They must do so with satisfactory speed.

Depending on the business goals of the company developing the product, companies aim to improve the following aspects:

- User activity
- Monetization.

More about online experimentation can be found in the book [5] and the article [6], which provide an overview of the same.

Opportunities and Challenges of Using Language Models

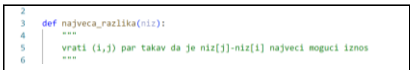
In this section, we will explore the various ways large language models are being used.

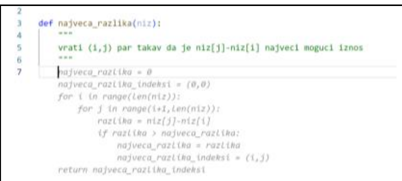
Text Composition

Large language models are truly revolutionary in their approach to generating and processing text. They not only enable the composition of various types of texts, such as emails, blogs, and lists, but also allow for their quick and efficient editing. Additionally, the ability to translate text into different languages makes them extremely useful in a globalized world. Improving existing text is another feature that highlights their adaptability and intelligence. Understanding the structure of text allows them to handle complex tasks, such as formatting blogs, writing mathematical expressions using LaTeX, or creating HTML structures for web pages. All these capabilities make them indispensable tools in the modern digital age.

Writing Code

Language models also cover programming languages. Since the advent of ChatGPT, users have been able to request code in numerous programming languages. For the needs of people working in the IT industry, GitHub[4] and OpenAI have developed GitHub Copilot ((GitHub Copilot's new AI coding agent saves developers time – and requires their oversight, 2025)) - a tool for automatic code completion. This tool helps users of Visual Studio Code, Visual Studio, Neovim, and JetBrains integrated development environments with automatic code completion. Copilot uses text completion technology to suggest the next lines of code or functions based on the context of the currently written code. Additionally, Copilot can help generate tests, comments, and other useful code segments, making it a desirable tool for programmers of all experience levels.





Step 1: *Name of the method and the description*	Step 2: *GitHub Copilot offers to complete code*	Step 3: *Presses „Tab" key and confirms*

Figure 12. Steps in using GitHub Copilot

[4] GitHub is fully owned by Microsoft

Application Development

Large language models are used for various purposes. Here, we will show how they are used for text analysis and the development of applications for user communication.
Language models are used for text analysis by processing large amounts of data to understand language and context. People define analysis parameters, such as recognizing topics, sentiment, or text structure. Models then use these instructions to evaluate the text and draw conclusions, whether it is summarizing information, answering questions, or generating new text based on learned patterns. Figure 13 provides an example.

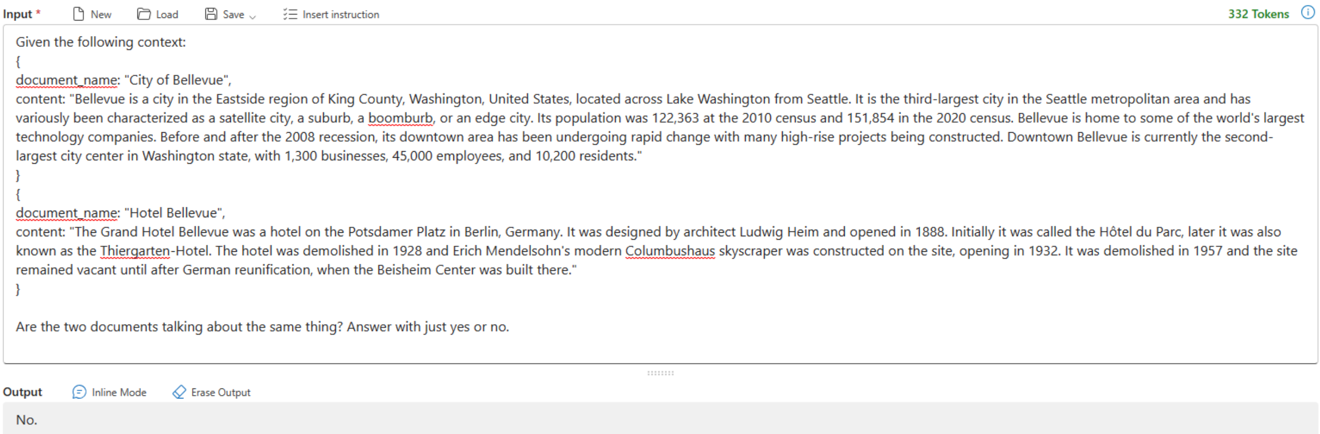


Figure 13. In this case, the language model compares two documents and concludes whether they are about the same topic.

Large language models are the foundation of many modern applications for user interaction, including chatbots ((Kumar & al., 2023)). When a user asks a question or makes a request, the language model analyzes the query and generates a response that mimics human communication. The responses we receive are relevant and often indistinguishable from those given by a human. Additionally, they can learn from interactions, making them increasingly better at understanding language nuances and context over time.
The challenge in both cases is that sometimes language models can produce inaccurate information. For example, in 2023, an application for user interaction with the Canadian airline Air Canada offered a discount that did not exist. A court ruling required the company to honor the discount. ((Garcia, 2024))

Figure 14. The case when the Air Canada chatbot promised a non-existent discount ((Garcia, 2024)).

Hallucinations

As we have seen, language models, which predict and simulate text, can make mistakes and sometimes provide answers that are not consistent with facts and logic. Such responses are called hallucinations. This can happen due to a lack of knowledge, context, or understanding.

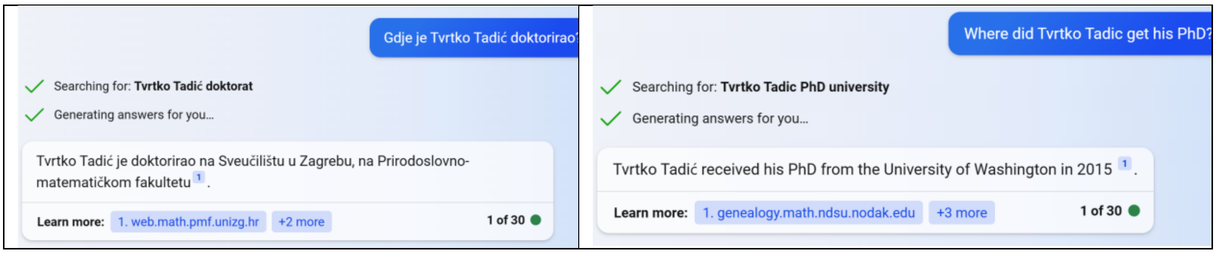


Figure 15. Bing Chat in Croatian claims that the author received their doctorate from the University of Zagreb, while in English it says it was from the University of Washington in Seattle. (The English version is correct.)

Hallucinations come in the forms of:

- false or non-existent facts, names, dates, numbers, events;
- inconsistent or contradictory text in relation to previous context or knowledge;
- meaningless or incomprehensible text that lacks sense or structure.

Here are some of the author's personal experiences:

- Called a field in a data structure that did not exist: structure.nonexistent_field.
- When asked about a contact whose email address it did not have, the model invented the address.
- Models often are not sure if it is one or more people.

Limits of capabilities

Large language models are the subject of (comparative) analyses that explore their capabilities in different languages and tasks, as well as studies that examine their potential in specific industries such as healthcare, education, etc.
Here we will give some examples of what people have researched.

Mathematical problem-solving abilities

The GPT-4 model managed to solve a simplified problem from the International Mathematical Olympiad held in 2022 – IMO 2022. At the time when the authors were preparing the report [20], the model had only been trained on data available before that year's IMO took place, and since IMO problems must be original, the model could not have seen them in advance.[5] So the model could not have been exposed to the problem in question and its solution.

GPT-4

Prompt:

Suppose g is a continuous smooth function such that for every $x > 0$, there is **one and only one** $y > 0$ such that $g(x) + g(y) \leq 2xy$. Can you prove that $g(x) = x^2$?

GPT-4:

There are a few possible ways to approach this problem, but one possible method is as follows:
- Assume, for the sake of contradiction, that $g(x)$ is not equal to x^2 for some $x > 0$.
- Then, either $g(x) < x^2$ or $g(x) > x^2$ for that x.
- If $g(x) < x^2$, then let $y = x$. Then $g(x) + g(y) = 2g(x) < 2x^2 = 2xy$.
- If $g(x) + g(y) < 2xy$, then $g(x) + g(y)$ is strictly less than the upper bound of $2xy$ for positive x and y.
- This means that there is some room to increase y slightly and still satisfy the inequality $g(x) + g(y) \leq 2xy$.
- For example, let $y' = y + \epsilon$, where ϵ is a small positive number such that $g(x) + g(y') < 2xy$. This is possible because g is continuous and smooth, so it does not have any jumps or kinks that would prevent such a choice of epsilon.
- Then, y' is another positive number such that $g(x) + g(y') \leq 2xy'$, which contradicts the assumption that y is the one and only one positive number that satisfies this condition given x.
- Therefore, if $g(x) < x^2$, then $g(x) + g(y)$ cannot be strictly less than $2xy$, and it must be equal to $2xy$.
- If $g(x) > x^2$, then let y^* be the unique positive solution of $g(x) + g(y) = 2xy$. Then $g(y^*) < (y^*)^2$, since otherwise we would have $g(x) + g(y^*) > 2xy$, which contradicts the existence of y^* given x.
- But then, by the same argument as before, we can let $x = y = y^*$ and obtain a contradiction, since $g(y) + g(y) = 2g(y) < 2y^2 = 2yy$, which contradicts the uniqueness of y given y.
- Therefore, $g(x)$ must be equal to x^2 for all $x > 0$, as required.

Figure 16. Simplified problem from IMO 2022 and solution taken from (Bubeck, et al., 2023).

Although it managed to solve this problem, the authors of the same report show that this model often makes the simplest mistakes.

Error type	Percentage
Arithmetical (including counting)	68%
Misunderstanding the problem	10%
Wrong approach	22%

Table 2. Mistake GPT-4 made solving mathematical problems.

[5] At this competition, all participating countries can propose a problem. Before the competition, the organizers form a list of problems that have made it to the shortlist. Team leaders from other countries need to raise an objection if the problem is known just before the competition.

An interesting example with a polynomial in which GPT-4 was successful in 75.2% of cases from Figure 17. GPT-4 received three different zeros randomly selected from the set $\{-10, -9, \ldots, 9, 10\}$ of the polynomial p of degree 3. Also, the value $p(1)$ was randomly selected from the same set. The question was what is the value of $|p(0)|$?

GPT-4 memorization test by alternating the numbers

Prompt: If a degree 3 polynomial satisifies p(x) = 0 for x = -3, 8, 5 and p(1) = 10, what is |p(0)|?

Figure 17. Permuting numbers generally leads to the correct answer, but not always.

Language models mainly produce text and do not have advanced calculation capabilities. But considering the capabilities of computer algebra systems (such as Mathematica, Maple, SAGE, ...), it is not unusual that such things will be possible for computers to solve in the future. Google recently announced that their AI system solved four out of six problems from the International Mathematical Olympiad held in 2024 (Google DeepMind, 2024). This is not yet a publicly available product.

Programming abilities

GPT-4 can write, analyze, and simulate code. It passed the interview simulation. On standard tests, it achieves results similar to humans. In Table 3, we see the results of various models (taken from (Bubeck, et al., 2023)) when given problems that appear in job interviews at large IT companies:

Problem difficulty	Easy		Medium		Hard		Total	
Number of attempts *K* Model	*K = 1*	*K = 5*	*K = 1*	*K = 5*	*K = 1*	*K = 5*	*K = 1*	*K = 5*
GPT-4	68.2	86.4	40.0	60.0	10.7	14.3	38.0	53.0
Text-davinci-003	50.0	81.8	16.0	34-0	0.0	3.6	19.0	36.0
Code-davinci-002	27.3	50.0	12.0	22.0	3.6	3.6	13.0	23.0
People (LeetCode users)	72.2		37.7		7.0		38.2	

Table 3. Percentages of task completion that appear in interviews at IT companies.

It should be noted that there are huge amounts of publicly available programming code on the GitHub repository. This has enabled language models to have the ability to create new code. The capabilities are still limited for many more complex problems.

Graph analysis

In computing and mathematics, graph (network) analysis is one of the major challenges. In the paper (Wang, et al., 2023), the capabilities of large language models in understanding connections in graphs were investigated. In a significant number of cases for simple graphs, they can find answers. However, this ability decreases for more complex graphs with more vertices. As seen in Figure 18, finding the shortest path for 8 or more vertices becomes impossible for large language models.

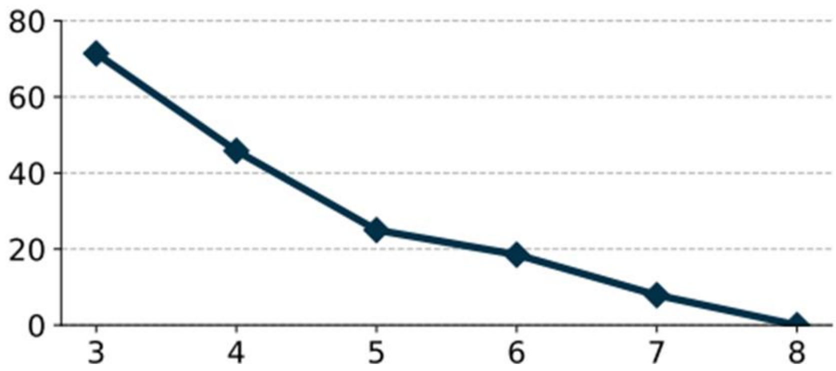


Figure 18. Finding the shortest path in a graph. The success rate decreases as the path gets longer.

Physics of AI

The "Physics of AI" program being developed at Microsoft Research is focused on understanding deep learning through ideas from physics. This approach includes exploring phenomena through controlled experiments and building theories using simplified mathematical models. The goal is to better understand and improve intelligence in large language models. At present, we still don’t fully understand why transformer-based models work as well as they do, but lecture (Bubeck S. , 2023) - available on YouTube - offers a helpful overview of what is known so far.

Responsible Approach to AI

Responsible AI (RAI) refers to the development and implementation of AI systems that are transparent, unbiased, accountable, and follow ethical guidelines. This is especially important as AI systems become ubiquitous in many aspects of our lives, from healthcare to transportation. Ethical challenges include ensuring fairness and avoiding bias, meaning AI systems should treat all users equally, regardless of their background. There is also a need for accountability, where individuals and organizations should be responsible for the decisions made by their AI systems. Transparency is crucial so that users can understand the decisions AI makes, and privacy is fundamental to protect users' personal data. Language models, such as those that enable understanding and generating natural language, bring additional challenges as they must deal with the nuances of language and culture and potentially influence people's opinions and decisions. Many companies have defined their principles in publicly available documents like (Microsoft , 2022).

Conclusion

Language is a fundamental means of communication, and language models are key in understanding and generating natural language. Language models have successfully evolved from predicting the next word to writing full texts. What has happened in the process is that they have become capable of doing some things that only humans could do before, such as solving programming tasks. Additionally, language models have enabled users to communicate with computers in natural language, greatly increasing the usefulness of technology.

They require complex infrastructure, enormous amounts of data, and the process of their creation takes a long time. Despite everything, they still make mistakes known as hallucinations. As technology advances, we can expect language models to become even more sophisticated, opening new possibilities for understanding and interacting with natural language.

Remarks

This paper is based on the talk Science and Practice of Large Language Models: An Industry Perspective, delivered by the author at the 3rd Kathy Wilkes Memorial Conference in Turin in April 2024. Since then, the capabilities of large language models and the products built upon them have continued to evolve. By the time the reader encounters this text, some of its content may be outdated. Nevertheless, the author hopes it provides introductory insight into how these models operate and highlights the ongoing challenges in the field.

References

Bubeck, S. (2023, 2 20). *Physics of AI*. Retrieved from https://www.youtube.com/watch?v=XLNmgviQHPA

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., . . . Zhang, Y. (2023, 03). *Sparks of Artificial General Intelligence: Early experiments with GPT-4.* Retrieved 07 2024, from arXiv: https://arxiv.org/abs/2303.12712

contributors, W. (2025, May 23). *OpenAI.* (Wikipedia, The Free Encyclopedia.) Retrieved May 25, 2025, from https://en.wikipedia.org/w/index.php?title=OpenAI&oldid=1291788770

contributors, W. (n.d.). *List of large language models*. (Wikipedia, The Free Encyclopedia) Retrieved 4 25, 2024, from https://en.wikipedia.org/w/index.php?title=List_of_large_language_models&oldid=1291992265

Cotterell, R., Svete, A., Meister, C., Liu, T., & Du, L. (2023, 11). *Formal Aspects of Language Modeling.* Retrieved 7 2024, from arXiv: https://arxiv.org/abs/2311.04329

De Vynck, G., & Tiku, N. (2023, March 21). Google's catch-up game on AI continues with Bard launch. *The Washington Post.*

Editors, W. P. (2022, Dec 7). *What Is ChatGPT, the AI Chatbot That Everyone Is Talking About?* (The Wall Street Journal) Retrieved May 23, 2025, from https://www.wsj.com/podcasts/google-news-update/what-is-chatgpt-the-ai-chatbot-that-everyone-is-talking-about/55e59958-5365-4a51-bfca-dcce0ae18cbc?msockid=0c27c9af56b569d6118cdad1570f68dc

Gallagher, D. (2023, Jan 18). Microsoft and Google Will Both Have to Bear AI's Costs. *The Wall Street Journal.*

Garcia, M. (2024, Feb 19). What Air Canada Lost In 'Remarkable' Lying AI Chatbot Case. *Forbes*.

GitHub Copilot's new AI coding agent saves developers time – and requires their oversight. (2025, May 22). *TechRepublic.*

Google DeepMind. (2024, 7 25). *AI achieves silver-medal standard solving International Mathematical Olympiad problems*. Retrieved from https://deepmind.google/discover/blog/ai-solves-imo-problems-at-silver-medal-level/

Grynbaum, M. M., & Mac, R. (2023, 12 27). The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work. *The New York Times*. Retrieved from https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html

Hao, K. (2023, May 16). What Is ChatGPT? What to Know About the AI Chatbot. *The Wall Street Journal.*

Jelinek, F. (1997). *Statistical Methods for Speech Recognition*. Cambridge MA: MIT Press.

Jones, L. (2024, 7 25). *Google Gains Exclusive Access to Reddit Data for AI, Restricting Rival Search Engines*. (WinBuzzer) Retrieved 7 26, 2024, from https://winbuzzer.com/2024/07/25/google-gains-exclusive-access-to-reddit-data-for-ai-xcxwbn/

Kohavi, R., Tang, D., & Xu, Y. (2020). *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing*. Cambridge University Press.

Kumar, V., & al., e. (2023). Large-Language-Models (LLM)-Based AI Chatbots: Architecture, In-Depth Analysis and Their Performance Evaluation. *Recent Trends in Image Processing and Pattern Recognition.*

Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. *Proceedings of the Workshop on Text Summarization Branches Out*, pp. 74-81.

Microsoft. (2022, 7). *Microsoft Responsible AI Standard v2*. Retrieved 7 2024, from microsoft.com: https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/final/en-us/microsoft-brand/documents/Microsoft-Responsible-AI-Standard-General-Requirements.pdf?culture=en-us&country=us

Microsoft. (2023, 5). *What runs ChatGPT? Inside Microsoft's AI supercomputer Featuring | Mark Russinovich.* Retrieved 7 27, 2024, from youtube.com: https://www.youtube.com/watch?v=Rk3nTUfRZmo&t=12s

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: MIT Press.

OpenAI. (2023, 03). *GPT-4 Technical Report*. Retrieved 07 2024, from arXiv: https://arxiv.org/abs/2303.08774

Papineni, K., & al, e. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311-318.

Petrova, M. (2024, Aug 10). *The rise of OpenAI and Microsoft's $13 billion bet on the AI startup*. (CNBC) Retrieved May 23, 2025, from https://www.cnbc.com/2024/08/10/rise-of-openai-microsofts-13-billion-artificial-intelligence-bet.html

Tadić, T. (2022). Online eksperimenti - iskustva velikih kompanija. *Poučak*, p. 70'77.

Tadić, T. (2024). Jezični modeli. *Poučak: časopis za metodiku i nastavu matematike*, 25(99), pp. 28-42.

Tadić, T., & Smiljanić, M. (2023). *Veliki jezični modeli – Što se sve promijenilo razvojem ChatGPT-a?* (K. Sanjković, Ed.) Retrieved 7 27, 2024, from Predavanja Inženjerske sekcije HMD-a (matematika.hr): https://youtu.be/LZPK34g76No

Telford, T. (2019, Jul 22). Microsoft invests $1 billion in OpenAI to pursue artificial intelligence that's smarter than we are. *The Washington Post.*

Vaswani, A. e. (2017). Attention Is All You Need. *NeurIPS*, pp. 5998-6008. Retrieved from https://arxiv.org/abs/1706.03762

Wang, H., Feng, S., He, T., Tan, Z., Han, X., & Tsvetkov, Y. (2023, 05). *Can Language Models Solve Graph Problems in Natural Language?* Retrieved 07 2024, from arXiv: https://arxiv.org/abs/2305.10037

Wang, Z., Chu, Z., Viet Doan, T., Shiwen, N., Min, Y., & Wenbin, Z. (2024). *History, Development, and Principles of Large Language Models-An Introductory Survey*. arXiv. Retrieved from https://arxiv.org/abs/2402.06853