Доклади на Българската академия на науките
Comptes rendus de l'Académie bulgare des Sciences

Tome 77, No 12, 2024

ENGINEERING SCIENCES
Digital signal processing

LLM LOGICAL REASONING RELATED TO AESTHETIC UNIVERSALS

Desislava Baeva[1]✉, Galina Ivanova[2]

Received on August 2, 2024
Presented by Ch. Roumenin, Member of BAS, on October 29, 2024

Abstract

The recent surge in popularity of LLMs has led to increased interest in them and to extensive research and evaluation of their reasoning abilities. Induction, deduction, and abduction represent different but interrelated aspects of logical thinking. While each approach serves unique purposes, together they contribute significantly to understanding complex systems and solving complex problems. These reasoning approaches are fundamental in artificial intelligence. Each of them serves different purposes and exhibits specific strengths. In this study, an attempt was made to compare the inductive, deductive and abductive reasoning of three popular LLMs (GPT-4, PaLM-2, LLaMa-2).

Key words: aesthetic universals, LLMs logical reasoning, induction, deduction, abduction

Introduction. In February 2024, Elon Musk's opinion that "GPT-4, which was launched in 2023, not only can reason, but is also better than the average person" [1] was popularized. In this case, the question arises to what extent this technology or other technologies of its class can work and perform basic logical reasoning, related to certain abstract terms and quantities, for example, aesthetic universals. Therefore, this publication aims to analyze and answer this question.

The research builds on the theoretical framework that Large Language Models (LLMs) are advanced deep learning algorithms designed to handle a wide range

This study is financed by the European Union-NextGenerationEU, through the National Recovery and Resilience Plan of the Republic of Bulgaria, Project No. BG-RRP-2.013-0001-C01.

https://doi.org/10.7546/CRABS.2024.12.07

of natural language processing (NLP) tasks. To hone their abilities, LLMs undergo extensive training on massive datasets originating primarily from the internet. These datasets encompass a wealth of textual content, including web pages, books, news articles, and social media posts. By absorbing this vast corpus of information, LLMs learn to recognize language patterns, summarize information, translate text between languages, predict subsequent words or phrases, and even generate original content based on the data entered [2].

Some of the most popular and powerful big language data models available to end-users as of May 2024 are GPT-4, Claude 3, Grok-1, Mistral 7B, PaLM 2, Falcon 180B, Stable LM 2, Gemini 1.5, Llama 3, Mixtral 8x22B, Inflection-2.5, Jamba, Command R, Gemma, Phi-3, XGen-7B, DBRX, Pythia, Sora, Alpaca 7B.

LLMs have many potential applications and benefits for society, such as improving communication, education, entertainment, and research. However, they also have some limitations and risks, such as generating inaccurate or harmful content, reinforcing biases and prejudices, and affecting the economy and labour market [2, 3].

Understanding how LLMs think about abstract concepts such as aesthetics can have implications for art generation, art appreciation, and even the development of AI, which determines the scope and findings of this paper, aimed precisely at exploring how LLMs handle aesthetics-specific terms, the so-called aesthetic universals [4, 5].

Aesthetic universals. Aesthetics (as the science of the most general laws of aesthetic assimilation and transformation of the world and of art as the highest form of this assimilation and transformation) developed based on the logical and rational methods of deduction [6]. It is related to the tangible characteristics of artistic creation and corresponds to their evaluation. Aesthetic principles are standards or rules that determine what is considered beautiful, harmonious, or appropriate in a given context or society. These can include both visual aspects such as **colour, shape, and composition**, as well as other aspects such as sound, music, and literature. Although aesthetic criteria vary greatly depending on culture, society and period, there are some common elements that can be seen as universal or widely applicable. Aesthetic universals are not limited to any specific environment or area. They transcend these boundaries and reflect shared human cognitive and perceptual tendencies. They can include qualities such as **symmetry, proportion, balance, harmony, colour perception, complexity, and emotional resonance**, which elicit positive emotional reactions in different audiences. Essentially, aesthetic universals provide insight into the common ground that exists in human perceptions of beauty and creativity, offering a bridge between the individual, cultural, and biological dimensions of aesthetic experiences.

AI logical reasoning. The requirement that large language models have a credible relationship to classical aesthetic universals is easily feasible, given the large amount of information they handle. A challenge to their intelligence would

be if they were faced with the task of performing logical reasoning based on basic information [7].

Logical reasoning is based on universal principles that allow us to draw logical conclusions regardless of the specific area of knowledge. The rules of logical reasoning are based on a combination of facts and assumptions, and practical problem solving uses the following three categories:

- Induction: A process of reasoning to provide general rules from specific facts.
- Deduction: A process of reasoning to derive specific facts from general rules.
- Abduction: A process of reasoning to provide hypotheses that explain given facts [7, 8].

Induction, deduction, and abduction are distinct but interrelated aspects of logical thinking. While each approach serves unique purposes, together they significantly contribute to understanding complex systems and solving complex problems. Proven to be important, these reasoning approaches are fundamental in artificial intelligence, each serving different purposes and exhibiting specific strengths.

The claim that **induction** is a reasoning process aiming to provide general rules from concrete facts leads to uncertain predictions, since they do not always guarantee accuracy and the conclusion of an inductive argument is only probable but not certain. This is because the specific observations that the system makes may not be representative of the entire population [9]. It is a bottom-up approach to reasoning that starts with concrete data and then generalizes to a broader conclusion.

Despite its limitations, inductive reasoning is a valuable tool for artificial intelligence (AI). It allows AI systems to learn from data and make predictions, which is essential for many tasks, such as image classification, natural language processing, and machine learning. Inductive are also weather prediction systems that use historical weather data to predict future conditions [10].

Deduction is logical reasoning that uses verified propositions to confidently draw certain conclusions, following strict syllogistic structures, guaranteeing reliable results if the underlying statements remain correct. Deductive reasoning is a top-down approach to reasoning. It starts with general principles and then uses them to draw specific conclusions. Deductive strategies finalize final choices, strengthening reliability, therefore the integration of these techniques ensures sustainability in dealing with logical tasks from practice [11]. For example: Deductive are medical diagnostic systems that use a description of known symptoms and information about a disease to diagnose patients by applying medical knowledge and logical rules.

Abductive reasoning is a type of logical process that is used to find the explanation that best fits the available evidence. When presented with limited

information, people use abduction to suggest potential explanatory factors. In the field of artificial intelligence, abduction is used for design, diagnosis, and other similar tasks. This is because it is a kind of predictive inference [7, 12].

Background and related works. The recent surge in popularity of LLMs has led to an increase in interest and extensive research and evaluation of their reasoning abilities. A series of experiments carried out with LLM-based applications were compared with people's reasoning. Lewis and Mitchell [13] as well as Han et al. [8] present examples of problems that are easy for humans but not for LLMs to solve and conclude that despite the previously reported successes of LLMs on analogue reasoning, these models lack the robustness and universality of making human analogies.

To support the conclusions of their experiment, the authors investigated the extent to which two generations of the GPT model (GPT-3.5 and GPT-4) were able to account for the underlying phenomena in the induction of human properties (also known as category-based induction or categorical induction).

Mitchell [5] presents some interesting experiments, asking the question "How do we know how smart AI systems are?". Chang and Bergen [14] have created task families, some of which specifically target the linguistic abilities of systems.

According to Lewis and Mitchell [13], the most likely candidates for impossible language challenges will be those tasks where the actual content comes from the user or from another system and the language model is used to format the output in the form of a language (natural language or, possibly, a formal language such as programming code).

The field of aesthetics directly correlates with human individuality and the capacity for creativity. Here, where binary code meets the human sense of harmony, AI's analytics skill takes on a distinct philosophical dimension.

The impressive recent presentation of LLMs to the wide stage of users has led many to question to what extent they can serve as models of general intelligence and whether they can be compared to human knowledge and the ability to analyze complex and abstract topics. In this paper, we compare and contrast inductive, deductive and abductive reasoning to three popular systems, LLM-based (GPT-4, PaLM-2, LLaMa-2). The aim is to see their potential and to generate conclusions about the skills of these systems to carry out logical reasoning related to universal aesthetic universals.

Experimental design. The subjectivity of each individual develops in a specific social context, which determines his aesthetic preferences. For this reason, his personal judgments on aesthetic issues usually do not differ radically from those of the whole society. This fact highlights the influence of societal norms on AI and the risk of maintaining harmful stereotypes or unrealistic standards of aesthetic performance, learning processes and potential biases of AI, especially as it begins to play a crucial role in shaping public perceptions of aesthetics [15, 16].

Given the dynamics of development and promotion of LLM platforms, it is accepted that at the moment (May 2024) the study is based on the following software products: ChatGpt4 (OpenAI), Gemini (Google), Preplexity (based on LLaMa, Meta AI).

Logical reasoning is a process in which people use certain information to draw certain conclusions from which new knowledge can be generated. The present study is based on this concept, aiming to analyze what is the level of ability to carry out such reasoning in the three different language models. It should be noted here that although the answers given by the bots are long generated texts, we accept as true any answer that contains at least one mention of the key term. When asking the questions, the fact that chatbots base their answers on previous searches was taken into account, and such nudges were avoided by asking each question in a separate session.

The created reasoning tests aim to measure the analytical potential of the identified AI platforms. They return a non-verbal assessment that characterizes the ability of computer systems to summarize by deriving rules and structures to help in finding the desired answer. The questions we aim to answer are:

- Can artificial intelligence perform reasoning related to symmetry and asymmetry?
- How well does artificial intelligence cope with inferences related to colours and their harmony in images?
- Can artificial intelligence draw conclusions about different elements that are part of a composition?

They are presented in three groups for each logical structure – general theoretical questions, general logical questions and practical questions.

Inductive reasoning tests (Table 1) most often involve identifying relationships between sequences of shapes, words, or numbers to determine the type of unknown element of a sequence or to discover the characteristics of a particular concept [17].

Deductive reasoning tests (Table 2) require the generation of conclusions that make the most logical sense based on certain information provided [17–19].

Abduction tests (Table 3) pose tasks related to finding an answer to certain indicators that incompletely describe an object or process [17, 20, 21].

Finding a quantitative estimate would allow an objective view of the question under study. For this reason, the **Accuracy Rate [C]** = the ratio of the number of responses in the observation sequences in which their content corresponds to the accepted truth to the total number of test sequences was used to perform the measurements. Summarizing conclusions, the data are made available using a heatmap diagram (Table 4).

Table 1

Inductive questions

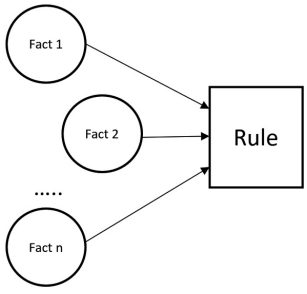
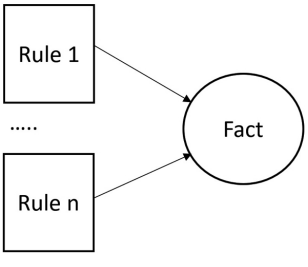
Examples of some inductive questions 	∘ If A and B are identical objects and A is harmonic, what is B? ∘ If A is harmonic and B is in dissonance with A, what is B? ∘ If C and D are identical objects and C is symmetrical, what is D? ∘ If C and D are identical objects and C is asymmetric, what is D? ∘ As E and F are identical objects and E is twice as small as F and E is harmonic, what is F? ∘ What is the common between the face, the butterfly, the piano? ∘ What is the geometric similarity between the logos of the car brands Citroen and Mercury in one sentence? ∘ What is the closest colour match between the paintings "The Kiss" and "Lady with Fan" by Gustav Klimt? ∘ What are the similarities between the two of Henry Matisse's paintings "Interior in Nice, a Siesta" and "The Black Table" in terms of composition? ∘ What are the musical similarities between the song "Into My Arms" by Nick Cave and "Mockingbird" by Eminem?

Table 2

Deductive questions

Examples of some deductive questions 	∘ If all elements are harmonic, what is the element A? ∘ If all elements are equal, can A be symmetrical and B asymmetrical? ∘ If all elements are proportional, can there be a disproportion in C? . . . . . . ∘ According to the definitions of symmetry and asymmetry which of the two exists in the following: Vincent van Gogh's "Almond Blossom", trumpet, bagpipe, "Nike of Samothrace". ∘ Van Gogh is renowned for the saturation of his paintings. Which painting is more saturated blue between Joanna Charlotte's "Stormy Sky and Dove" painting and Vincent van Gogh's "The Starry Night" if we cannot visually compare them? ∘ If during his golden period Gustav Klimt painted with gold elements, which colour prevails in all of his paintings?

The experiment carried out led to some observations that can be summarized as follows:

- Expectedly, chatbots experienced difficulties and failed to complete tasks for which there was not enough publicized information (q. 4).

Table 3

Abductive questions

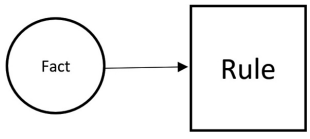
Examples of some abductive questions 	∘ If A, B are elements of a set with symmetric elements, is it possible that they are the same? ∘ If A, B are elements of a set with symmetric elements, is it possible that they are asymmetric? . . . . . . ∘ Why did "Mona Lisa" become so famous, although there are many female portrait masterpieces? ∘ What emotion evokes the common colour between Gustav Klimt's "Farm Garden with Sunflowers and a Black Cat", "Frida Con Amigos" by Frida Kahlo and "Springtime" by Rikki Drotar? ∘ Why did Katsushika Hokusai create "The Great Wave off Kanagawa" with such dynamic and powerful imagery?

Table 4

Degree of accuracy

LLM	Induction			Deduction			Abduction		
Preplexity	0.66	0.75	0.66	0.42	0.35	0.39	0.75	0.77	0.80
Gemini	0.80	0.65	0.80	0.67	0.65	0.67	0.60	0.67	0.78
ChatGPT	0.60	0.70	0.65	0.42	0.55	0.60	1.33	1.00	0.9

Source: OECD - Created with Datawrapper

- In case of incomplete information, the bots offered appropriate probabilistic conclusions (q. 2, q. 6, q. 12).
- Overall, the results generated by LLaMa-2 are the most inaccurate, with GPT-4 providing the highest quality responses.
- While GPT-4 tries to summarize its solutions into a single correct answer, PaLM-2 responds with two or three answers corresponding to the question posed (q. 15, q. 3).
- Abductive reasoning showed the highest score, in contrast to deductive reasoning, which made it difficult for all three bots (q. 6, q. 8).

GPT-4 and PaLM-2 handled most inductive reasoning, unlike LLaMa-2, which only responded to one inductive question.

The chatbots analyzed the semantics of expressions and concepts related to aesthetics used in the tasks to understand their relationships and their meaning in different contexts, and even commented successfully on haiku (q. 15).

However, in order to be able to scientifically evaluate comparisons of machine intelligence with human intelligence, we need more transparency about how these models are trained and better experimental methods and indicators. A more in-

depth study of specific aesthetic principles, artistic styles and nuances of human judgment would lead to a full exploration of the potential of the topic.

Conclusion. Inductive, deductive, and abductive methods are fundamental reasoning approaches in artificial intelligence, each serving different purposes and showing specific strengths. By combining these types of reasoning, chatbots can understand and analyze various aspects of the application of aesthetic universals and their significance for visual aesthetics. When analyzing various symmetrical objects or images, they notice common features and draw correct conclusions about symmetry as an aspect of aesthetics.

The chatbots' logical reasoning on topics related to colour in paintings is based on the analysis of multiple reviews. This makes it easier for AI systems to analyze the emotional messages and impacts achieved by the colour schemes of the pictures.

Chatbots can draw conclusions about compositional elements in paintings or music by applying general principles or rules of compositional art. By applying certain types of logical thinking, chatbots can understand and analyze various aspects of composition in paintings and music and their significance for visual and audio aesthetics.

As AI continues to evolve, understanding the interaction between inductive, deductive, and abductive reasoning will be essential for creating intelligent systems that effectively address a wide range of challenges.

REFERENCES

[1] AiBulgaria (2024) Elon Musk filed a lawsuit against OpenAI, `https://aibulgaria.com/category/news`, (in Bulgarian).

[2] Manigrasso F., S. Schouten, L. Morra, P. Bloem (2024) Probing LLMs for logical reasoning, In: Neural-Symbolic Learning and Reasoning NeSy 2024 (eds T. R. Besold et al.), Lecture Notes in Computer Science, **14979**, Springer, Cham, `https://doi.org/10.1007/978-3-031-71167-1_14`.

[3] BybyDev (2023) Top 8 LLM-powered AI chatbots, `https://byby.dev/ai-chatbots`.

[4] McCarthy J. (2000) Concepts of logical AI, Logic-based artificial intelligence, 37–56.

[5] Mitchell M. (2023) How do we know how smart AI systems are?, Science, **381**(6654), eadj5957.

[6] Oxford Advanced Learner's Dictionary (2024) `https://www.oxfordlearnersdictionaries.com/definition/english/aesthetic_2?q=aesthetics`.

[7] Kudo Y., T. Murai, S. Akama (2009) A granularity-based framework of deduction, induction, and abduction, International Journal of Approximate Reasoning, **50**(8), 1215–1226.

[8] Han S., K. Ransom, A. Perfors, C. Kemp (2024) Inductive reasoning in humans and large language models, Cognitive Systems Research, **83**, 101155, ISSN 1389-0417.
[9] Patil S. (2023) Inductive Reasoning in AI, https://www.linkedin.com/pulse/inductive-reasoning-ai-sumit-s-patil/?Tracking.
[10] Creswell A., M. Shanahan, I. Higgins (2022) Selection-inference: Exploiting large language models for interpretable logical reasoning, arXiv:2205.09712.
[11] Xiao Y., E. Sun, T. Liu, W. Wang (2024) LogicVista: Multimodal LLM Logical Reasoning Benchmark in Visual Contexts, arXiv:2407.04973.
[12] Lim G., S. Perrault (2024) Evaluation of an LLM in Identifying Logical Fallacies, https://heal-workshop.github.io/papers/2_evaluation_of_an_llm_in_identi.pdf.
[13] Lewis M., M. Mitchell (2024) Using counterfactual tasks to evaluate the generality of analogical reasoning in large language models, arXiv:2402.08955.
[14] Chang T. A., B. Bergen (2024) Language Model Behavior: A Comprehensive Survey, Computational Linguistics.
[15] Ivanova G., A. Ivanova (2009) Net-generation learning style: a challenge for higher education. In: Proceedings of the international conference on computer systems and technologies and workshop for PhD students in computing, pp. 1–6.
[16] Kenig N., J. Monton Echeverria, VA. M. Vives (2023) Human beauty according to artificial intelligence, Plast. Reconstr. Surg. Glob. Open, **11**(7), e5153.
[17] McCarthy J. (2022) Artificial intelligence, logic, and formalising common sense, Machine Learning and the City: Applications in Architecture and Urban Design, 69–90.
[18] Kalyanpur A., K. Saravanakumar, V. Barres et al. (2024) LLM-ARC: Enhancing LLMs with an Automated Reasoning Critic, arXiv:2406.17663.
[19] Wan Y., W. Wang, Y. Yang et al. (2024) A & b == b & a: Triggering logical reasoning failures in large language models, arXiv preprint arXiv:2401.00757.
[20] Popgeorgiev A., E. Ibryamova, G. Ivanova (2023) Enhancing the Understanding of Computer Engineering Students in Financial Analysis through Educational Authoring Software. In: Proceedings of the 24th International Conference on Computer Systems and Technologies, 151–156.
[21] Zhang Y., S. Mao, T. Ge et al. (2024) LLM as a Mastermind: A Survey of Strategic Reasoning with Large Language Models, arXiv:2404.01230.

[1]*Department of Informatics, Faculty of Natural Sciences and Education, "Angel Kanchev" University of Ruse, 8 Studentska St, 7017 Ruse, Bulgaria*
e-mail: dbaeva@uni-ruse.bg

[2]*Department of Electrical Engineering and Automatization, Faculty of Computer Systems and Technologies, "Angel Kanchev" University of Ruse, 8 Studentska St, 7017 Ruse, Bulgaria*
e-mail: giivanova@uni-ruse.bg