Improving the inference performance of LLM with code

***Fei Ding** [1,*]

Abstract

Large Language Models (LLMs) have shown exceptional generative abilities in various natural language and generation tasks. Large language models (LLMs) have demonstrated remarkable performance on a variety of natural language tasks based on just a few examples of natural language instructions, reducing the need for extensive feature engineering. However, LLM is relatively weaker in reasoning and problem-solving abilities. We propose a new construction that solves the problem of insufficient logical mathematics and logical ability.

Keywords: Large Language Model

Contents

∗. *Corresponding author email: dingfei@email.ncu.edu.cn

1 Introduction

With the remarkable progress made by large language models such as GPT-4, ChatGPT, Google Gemini, Llama-2 (Touvron et al., 2023), and Mistral (Jiang et al., 2023) in NLP research, machines are now capable of performing a wide range of language tasks that were previously believed to be exclusive to humans (OpenAI, 2023; Brown et al., 2020; Zhao et al., 2023). Performs well on language tasks such as Hellaswag (Zellers et al., 2019), Winogrande (Sakaguchi et al., 2021), PIQA (Bisk et al., 2020) and ARC-Easy . However, Logical reasoning is a critical component of intelligence and is essential for many practical applications, including question-answering systems (Khashabi, 2019) and conversational agents (Beygi et al., 2022).

Several datasets have been proposed such as (Clark et al., 2020; Tian et al., 2021; Joshi et al., 2020; Saeed et al., 2021), LogiQA (Liu et al., 2021) , ReClor (Yu et al., 2020) , FOLIO (Han et al., 2022) and ProntoQA (Saparov and He, 2023) that demonstrate the relatively weak ability of these LLMs to reason logically over natural language text.

Large language models also perform poorly in mathematics and code such as GSM8K (Cobbe et al., 2021) with maj@8 , MATH (Hendrycks et al., 2021) with maj@4 , Humaneval (Chen et al., 2021) and MBPP (Austin et al., 2021) .

We delved into D. Kahneman's theory of thinking fast and slow (Kahneman, 2011), and we propose a new simple AGI architecture (called DingFei model, for Artificial General Intelligence) where divide AGI into emotional brain, rational brain, and bottom brain. The emotional brain reacts through intuition, and the modified LLM can serve as an emotional brain. The rational brain is responsible for logical reasoning. The bottom brain is the underlying code that supports the operation of the emotional brain and the rational brain, as well as rules written in code. The collaboration of three brains solved the problem mentioned earlier. We also introduced the concept of skills and scratch paper, ensuring 100% accuracy. With the addition of autonomous learning, we achieved Artificial General Intelligence 1.0.

1.1 objective

We propose the DingFei model to achieve Artificial General Intelligence 1.0. AGI needs to possess many capabilities that would naturally be included in a notion of human intelligence. Examples of these capabilities are generalizability, adaptability, robustness, explainability, causal analysis, abstraction, common sense reasoning, ethical reasoning (Rossi and Mattei, 2019), as well as a complex and seamless integration of learning and reasoning supported by both implicit and explicit knowledge (Littman et al., 2021). We delve into the mechanisms that enable humans to possess these capabilities, which helps us understand how to imbue AI systems with these competencies. (Rossi and Loreggia, 2019; Booch et al., 2021).

2 The emotional brain

The emotional brain is primarily driven by intuition over careful consideration, providing quick responses to straightforward questions. Intuition is often generated after reading a sufficient amount of data. They are tightly linked to the availability of huge datasets and computational power (Marcus, 2020). However, these answers can occasionally be incorrect due to unconscious biases or their reliance on heuristics and other shortcuts (Gigerenzer and Brighton, 2009), and typically lack explanations.

3 The rational brain

Intuition often fabricates false facts (i.e. *hallucination*). But logical reasoning requires precise answers, and this is when the rational brain needs to be used.

3.1 Training data

Textbooks and a large number of related exercises from kindergarten, elementary school, middle school, Senior high school, and University.

3.2 Training

How to train the rational brain? For example, mathematics. Let the LLM read textbooks, and then summarize skills for each knowledge point . Then Use the skills learned to answer practice questions. If they are incorrect, correct and improve the skill. A skill is a set of knowledge, rules, and operational flowcharts. Then the skill is converted into executable code and stored in the rational brain, ready to be used when needed. Connect relevant skills with each other.

4 Scratch paper

Psychological research reveals a fascinating insight: even children and adults can have their problem-solving prowess significantly dampened by irrelevant information (Hoyer et al., 1979; Pasolunghi et al., 1999; Marzocchi et al., 2002) . Similarly,We conducted a large number of experiments that demonstrated the performance of large language models can be affected by irrelevant context, leading to incorrect results. To solve this problem, we proposed the concept of draft paper, ensuring a success rate of 100% when dealing with complex issues. I will show the details later.

5 The bottom brain

In the bottom brain, human beings can set the behavioral norms for AI.The bottom brain runs directly on the hardware and provides support for both the rational brain and the emotional brain. I will introduce more details about the rational brain and the bottom brain in the future.

6 Effect

6.1 Prediction

For example, a math problem.

CHATGPT:

chatgpt at Figure 1.Everything was correct until it came to multiplication at the end . Although ChatGPT was trained with a large amount of mathematical computation data to improve the accuracy of mathematical computations, it obviously cannot write out the correct answer to a three-digit multiplication problem based on intuition alone.

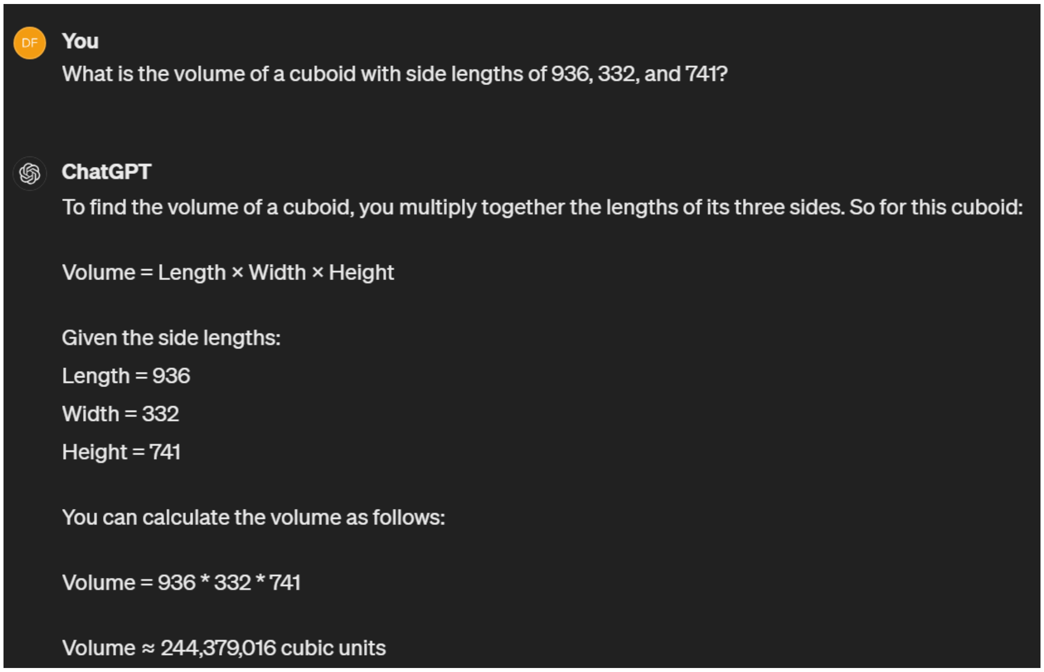

Figure 1: chatgpt

Mistral:

Mistral at the top of Figure 2. Similarly, Mistral also did a lot of optimization in mathematics, but there was an error in the calculation of multiplication.

The DingFei mode:

The DingFei model at the bottom of Figure 2. We used cross-computation between the rational and emotional brains to avoid giving an answer directly through intuition alone, thereby ensuring the precision of the results . Here are the details:

At line 1 , the rational brain was tasked with thinking, using the skill "formula for the volume of a rectangular prism," to arrive at line 1.

At line 2 , the rational brain was tasked with thinking, using the skill "Associative Property of Multiplication", to arrive at line 3.

At line 3 , uses the scratch paper to calculate 936×332 and obtains the result in line 4.

At line 4 , the rational brain was tasked with thinking, using the skill "Multiplication of large numbers", to arrive at line 5.

At line 5 , uses the scratch paper to calculate each multiplication separately and obtains the result in line 6.

At line 6 , the rational brain was tasked with thinking, using the skill "Adding Multiple Numbers", to arrive at line 7.

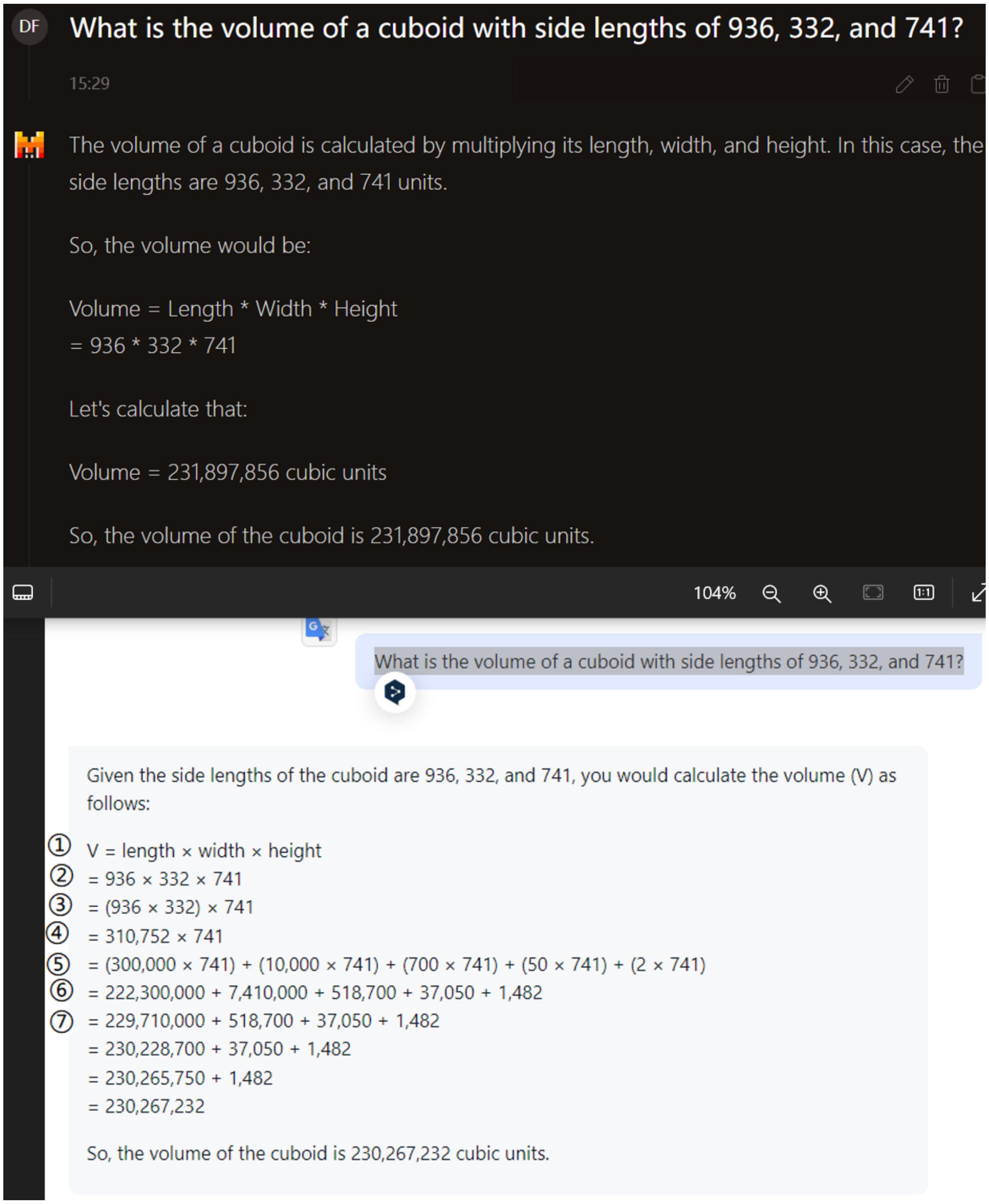

Figure 2: Top: mistral, Bottom: DingFei model

7 Conclusion

The DingFei model can achieve 100% accuracy in complex problems.Due to insufficient computational power , we have only trained part of the mathematics textbooks, and the accuracy rate of doing exercises can reach 100%. In the future, we will continue to train mathematics and other discipline . We expect the accuracy rate to remain at 100%.

References

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

Sajjad Beygi, Maryam Fazel-Zarandi, Alessandra Cervone, Prakash Krishnan, and Siddhartha Jonnalagadda. Logical reasoning for task oriented dialogue systems. In *Proceedings of the Fifth Workshop on e-Commerce and NLP (ECNLP 5)*, pages 68–79, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.ecnlp-1.10. URL `https://aclanthology.org/2022.ecnlp-1.10`.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, 2020.

Grady Booch, Francesco Fabiano, Lior Horesh, Kiran Kate, Jonathan Lenchner, Nick Linck, Andreas Loreggia, Keerthiram Murgesan, Nicholas Mattei, Francesca Rossi, and Biplav Srivastava. Thinking fast and slow in AI. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 15042–15046, 2021.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf`.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. Transformers as soft reasoners over language. *arXiv preprint arXiv:2002.05867*, 2020.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *ArXiv*, abs/2110.14168, 2021.

Gerd Gigerenzer and Henry Brighton. Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science*, 1(1):107–143, 2009.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Luke Benson, Lucy Sun, Ekaterina Zubova, Yujie Qiao, Matthew Burtell, et al. Folio: Natural language reasoning with first-order logic. *arXiv preprint arXiv:2209.00840*, 2022.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

William J Hoyer, George W Rebok, and Susan Marx Sved. Effects of varying irrelevant information on adult age differences in problem solving. *Journal of gerontology*, 34(4):553–560, 1979.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Pratik Joshi, Somak Aditya, Aalok Sathe, and Monojit Choudhury. TaxiNLI: Taking a ride up the NLU hill. In Raquel Fernández and Tal Linzen, editors, *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 41–55, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.conll-1.4. URL `https://aclanthology.org/2020.conll-1.4`.

Daniel Kahneman. *Thinking, Fast and Slow.* Macmillan, 2011.

Daniel Khashabi. *Reasoning-Driven Question-Answering for Natural Language Understanding.* University of Pennsylvania, 2019.

Michael L. Littman, Ifeoma Ajunwa, Guy Berger, Craig Boutilier, Morgan Currie, Finale Doshi-Velez, Gillian Hadfield, Michael C. Horowitz, Charles Isbell, Hiroaki Kitano, Karen Levy, Terah Lyons, Melanie Mitchell, Julie Shah, Steven Sloman, Shannon Vallor, and Toby Walsh. Gathering Strength, Gathering Storms: The One Hundred Year Study on Artificial Intelligence (AI100) 2021 Study Panel Report. *Stanford University*, 2021.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. Logiqa: a challenge dataset for machine reading comprehension with logical reasoning. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3622–3628, 2021.

Gary Marcus. The next decade in AI: Four steps towards robust artificial intelligence. *arXiv preprint arXiv:2002.06177*, 2020.

Gian Marco Marzocchi, Daniela Lucangeli, Tiziana De Meo, Federica Fini, and Cesare Cornoldi. The disturbing effect of irrelevant information on arithmetic problem solving in inattentive children. *Developmental neuropsychology*, 21(1):73–92, 2002.

OpenAI. Gpt-4 technical report, 2023.

Maria Chiara Pasolunghi, Cesare Cornoldi, and Stephanie De Liberto. Working memory and intrusions of irrelevant information in a group of specific poor problem solvers. *Memory & Cognition*, 27:779–790, 1999.

F. Rossi and N. Mattei. Building ethically bounded AI. In *33rd*, 2019.

Francesca Rossi and Andrea Loreggia. Preferences and ethical priorities: thinking fast and slow in AI. In *Proceedings of the 18th international conference on autonomous agents and multiagent systems*, pages 3–4, 2019.

Mohammed Saeed, Naser Ahmadi, Preslav Nakov, and Paolo Papotti. RuleBERT: Teaching soft rules to pre-trained language models. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1460–1476, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.110. URL `https://aclanthology.org/2021.emnlp-main.110`.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

Abulhair Saparov and He He. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=qFVVBzXxR2V`.

Jidong Tian, Yitian Li, Wenqing Chen, Liqiang Xiao, Hao He, and Yaohui Jin. Diagnosing the first-order logical reasoning ability through LogicNLI. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3738–3747, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.303. URL `https://aclanthology.org/2021.emnlp-main.303`.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. Reclor: A reading comprehension dataset requiring logical reasoning. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=HJgJtT4tvB`.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.