Expressivity of Transformer architectures: recent developments, existing limitations and realistic consequences

u2110450

March 12, 2025

'You'll Never Stop at One! I'll Take You all on!'
Optimus Prime (Transformers, 2009) describing the exponential-time problems.

Word Count: 1097/1100

1 Introduction

Recent developments in the field of machine learning techniques have pushed the practical capabilities of modern 'AI' models to previously unknown levels. Available Transformer architecture-based LLMs, such as ChatGPT, have become omnipresent and hailed as candidates to achieve 'Artificial General Intelligence' (OpenAI, 2023). Evolving from generating a coherent text, they are now expected to 'reason' i.e. use computation. In reality, these models can technically be unable to perform such tasks. In this essay I analyse two recent papers '*Tighter Bounds on the Expressivity of Transformer Encoders*' (Chiang, Cholak et Pillay, 2023) and '*The Expressive Power of Transformers with Chain of Thought*' (Merril et Sabharwa, 2024) that define their limitations through theoretical analysis, and interpret their potential implications[1]. I first introduce the theoretical framework used in the essay and present papers' results which demonstrate formal limitations of the existing Transformer architectures. I then link these results to existing trends in 'AI' and question the claims about future AGI systems.

2 Frameworks

Recently several empirical experiments (Bhattamishra et al., 2022, Delétang et al., 2023, Dziri et al., 2023) have questioned the realistic capabilities of *Transformers*: a neural network model architecture that replaces *recurrence* technique with *attention* (Vaswani et al., 2017), suggesting that lack of the memory component substantially narrows down their problem-solving capabilities. Traditionally, complexity of RNN architectures, such as LSTMs, have been evaluated through the Chomsky Language Hierarchies (Chomsky, 1956), where establishing a lower and upper bounds for a given 'algorithm' assigns it to a corresponding complexity class, and determines which problems it is capable of solving (Immerman 1981). However due to fundamental differences in their architecture (Hao et al. 2022) this method is not a viable option for simple Transformers. There were two methods introduced to evaluate complexity of Transformers: first order logic and circuit complexity, used by the following papers.

3 Papers

Chiang et al. (2023) aim to characterise Transformers through a formal system and theoretically identify their feasible power and limitations. By introducing and defining a designated logic FOC[+;MOD], the paper relates fixed and arbitrary-precision[2] transformer encoders with a first order logic with counting quantifiers[3] (Chiang et al., 2023), allowing them to establish their upper and lower bounds. For fixed-precision transformer encoders they find the upper bound to be at FOC[+;MOD] i.e. they can only recognize languages within this class.[4,5]

This yields two results:

- In comparison to previous research these findings tighten the known bound of complexity for Transformers, as FOC[+;MOD] is contained within the TC^0 circuit complexity class, previously suggested by Bhattamishra et al. (2020).
- Without the properties of recurrence and 'memory', they cannot solve *sequential problems* such as graph connectivity (*NL* class) (Fig 1) or linear equalities (*P* class) (Fig 2), which reiterates the limitations raised by the empirical evidence (Zhang et al., 2023, Bhagavatula et al., 2024).

[1] The aim of this essay is not to examine the technical proofs of the analysed works, as it would go beyond the scope of this work, but to apply their findings to future AI developments.

[2] A fixed-precision encoders are such that they 'round' their computations due to limitations in precision(Merrill et Sabharwal, 2022).

[3] More complexity classes can be found at Complexity Zoo (Aaronson, 2024).

[4] A system recognises a set of languages when it is able to identify whether a certain string belongs to a formal language. In reality, papers use *Transduction Tasks* (Delétang et al., 2022)

[5] For arbitrary-precision transformers, FOC[+;MOD] is their lower bound.

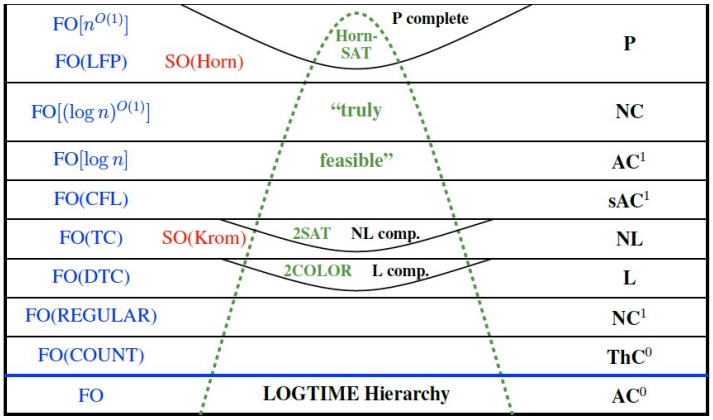


Figure 1: Descriptive Complexity Classes (Immerman 1999)

The lack of memory has been partially answered by introducing the *Chain of Thought* (CoT) technique (Wei et al., 2022). The intuition behind the Chain of Thought is that it allows introducing recurrence, and therefore partially sequential problem solving, to the Transformer architecture (Wei et al. 2022). By introducing a dedicated *layer norm hash*, the model can precisely retrieve information from previous states (*exact matching*). In the context of a Turing Machine with a sequence of states on its 'tape', imagine the CoT Transformer accepting a *state vector* at (t) only if it exactly matches the value from the previous 'tape-cell' $(t\text{-}1)$. Such sequential mechanisms can improve the 'reasoning' (or *computing*) capabilities of the models (Wei et al., 2022). The second paper 'The Expressive Power of Transformers with Chain of Thought' by Merrill et Sabharwal (2024) expands the previous findings of Chiang et al. (2023), by analysing a CoT Transformer-decoder through a circuit complexity approach - they relate transformers to a given circuit complexity class, similarly to Furst et al. (1984). They argue this technique extends the computational power of the Transformer models. The paper introduces two CoT methods with: logarithmic and linear number of decoding steps, and then compares the results with a classical model.

As suggested by Chiang et al. (2023) a standard Transformer model without recurrence CoT(1), can solve only some problems within the TC^0 class. However, introducing the intermediate steps of CoT substantially extends the formal complexity of transformers to:

- For Logarithmic Steps CoT($\log n$), the upper bound is expanded to L class. This expands the upper bound of transformers only to a limited extent.
- For Linear Steps CoT(n), transformer decoders can recognise languages within NL class, meaning they are able to solve some graph connectivity problems (Fig 3) (Merrill et Sabharwal, 2024).

This results with:

1. Intuitively, CoT enhances ability to solve sequential problems, however to an extent,
2. for example, if O(n) = O(n^c) for some constant c, i.e. we introduce polynomial-steps for CoT[6], the upper bound of Transformers would further increase to P, which would substantially increase their capabilities, for example solving linear problems '*Find x s.t. Ax = B*',
3. Such advanced transformers can finally be linked with Chomsky Hierarchies (Merrill et Sabharwal, 2024), where CoT(n) models are able to recognise all regular languages (Reg), and are upper bounded within the context-sensitive languages (CSL) (Kuroda, 1964).

[6]This is unlikely due to feasibility constraints of running polynomial-time intermediate steps.

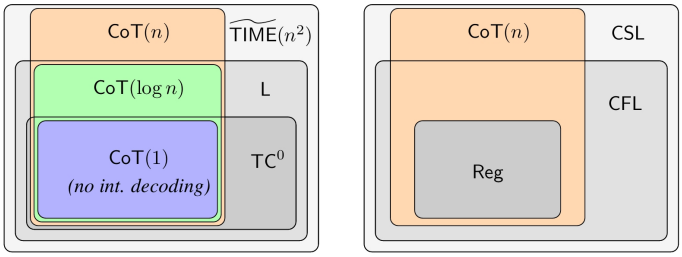


Figure 2: Complexity and Expressivity classes of various Transformers with and without CoT (Merrill et Sabharwal, 2024), where the right panel classifies them within the Chomsky Language Hierarchy.

4 What's the fuzz?

4.1 Attention is not all you need

Transformers were first introduced to enable faster training due to parallelisation (Vasvani et al. 2017), which enabled the current 'AI boom', as these architectures were easy to train and scale. Chiang et al. (2023) and Merrill et Sabharwal (2024) demonstrate their realistic limitations and suggest there exists a tradeoff between parallelisation and complexity capabilities of a model. We can train models faster and with less resources, but their computing-abilities will suffer. CoT is a technique to mitigate such limitations, hence it is being introduced in newest iterations of LLMs (Fig 3), but at a practical cost.

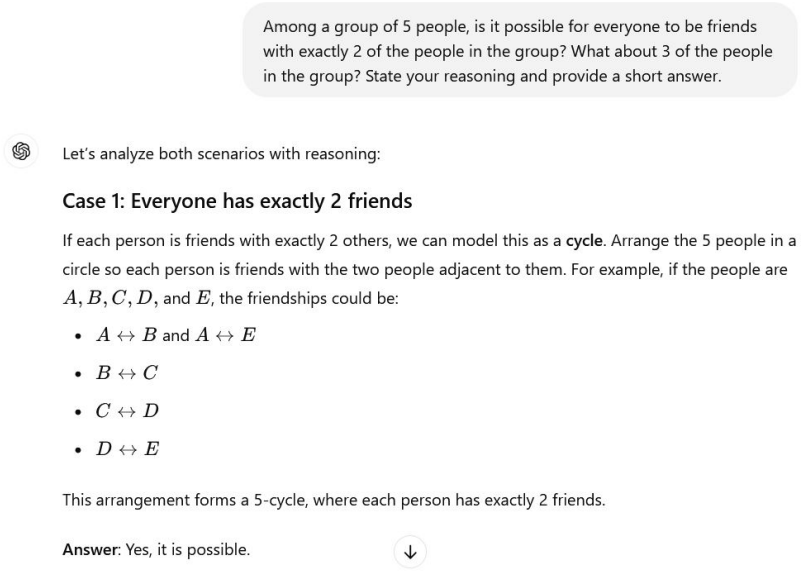


Figure 3: Newest ChatGPT iteration employs CoT and hence provides answers to graph connectivity questions.

4.2 '*50% chance of AGI by 2029*' (Shane Legg, 2023) and scalability 'laws'

Despite optimistic outlooks these results suggest existing state-of-the-art 'AI' architectures are not universal: solving computing problems, achieving high NLP scores and being easy to train and deploy, cannot be achieved by one. Therefore, expectations for future 'AI' may need to be revised. Unless a new architecture is introduced, the marginal gains from increased number of parameters or CoT steps may not be enough to fulfil our broad futuristic expectations for LLMs. Furthermore, these findings undermine the vision of achieving an AGI, and pose a broader question: 'What do we expect from, and can achieve, with 'AI architectures'? Should we strive for universal programs that will be able to

operate as 'all-rounders', or should we design programs specialising in one field - just as writers and mathematicians specialise in their dedicated subjects. These questions, particularly the feasibility of achieving an AGI, will be raised in further course essays.

5 Bibliography

1. Aaronson, S. (2011). 'Why philosophers should care about computational complexity'. MIT Press.

2. Bahdanau, D. (2014). 'Neural machine translation by jointly learning to align and translate'.

3. Bhattamishra et al. (2020). 'On the Ability and Limitations of Transformers to Recognize Formal Language'. *Empirical Methods in Natural Language Processing 2020*.

4. Chiang D., Cholak P., Pillay A. (2023). Tighter Bounds on the Expressivity of Transformer Encoders. Proceedings of Machine Learning Research, PMLR.

5. Chomsky N. (1956). 'Three models for the description of language. IRE Transactions on Information Theory'. 2 (3): 113–124. doi:10.1109/TIT.1956.1056813. S2CID19519474. Archived (PDF) from the original on 2016-03-07

6. Dean, W. and Naibo, A. (2024). 'Artificial intelligence and inherent mathematical difficulty'. arXiv preprint arXiv:2408.03345.

7. Dean, W. (2019). 'Computational complexity theory and the philosophy of mathematics.' *Philosophia Mathematica* 27, 3 (2019), 381–439.

8. Delétang, G., Ruoss, A., Grau-Moya, J., Genewein, T., Wenliang, L.K., Catt, E., Cundy, C., Hutter, M., Legg, S., Veness, J. and Ortega, P.A. (2022). 'Neural networks and the chomsky hierarchy'. arXiv preprint arXiv:2207.02098.

9. Dziri, N., Lu, X., Sclar, M., Li, X.L., Jiang, L., Lin, B.Y., Welleck, S., West, P., Bhagavatula, C., Le Bras, R. and Hwang, J.(2024). **??**'Faith and fate: Limits of transformers on compositionality'. *Advances in Neural Information Processing Systems*, 36.

10. Furst M., Saxe J., Sipser M. (1984). 'Parity, circuits, and the polynomial-time hierarchy.', *Mathematical Systems Theory*, 17(1):13–27. https://doi.org/10.1007/BF01744431.

11. Hao, Y., Angluin, D. and Frank, R. (2022). 'Formal language recognition by hard attention transformers: Perspectives from circuit complexity.' *Transactions of the Association for Computational Linguistics*, 10, pp.800-810.

12. Harel D. (1986) *Algorithmics: the spirit of computing.* Addison-Wesley Longman Publishing Co., Inc., USA.

13. Immerman N. (1982) 'Upper and lower bounds for first order expressibility.' *Journal of Computer and System Sciences*, Volume 25, Issue 1, 1982, Pages 76-98, ISSN 0022-0000, https://doi.org/10.1016/0022-0000(82)90011-3.

14. Immerman, N. (1999) *Descriptive Complexity.* Springer Verlag.

15. Kaplan J., McCandlish S., Henighan T., Brown T.B., Chess B., Child R., Gray S., Radford A., Wu J., Amodei D. 2020 'Scaling laws for neural language models.' CoRR.

16. Legg Sh. (2023). '2028 AGI, New Architectures, Aligning Superhuman Models.' *Dwarkesh Podcast.*

17. Merrill, W. and Sabharwal, A. (2022). 'The parallelism tradeoff: Limitations of log-precision transformers.' *Transactions of the Association for Computational Linguistics*, 11, pp.531-545.

18. Merrill, W. and Sabharwal, A. (2024). The expresssive power of transformers with chain of thought, ICLR.

19. Merrill, W. (2024). 'The Parallelism Tradeoff: Understanding Transformer Expressivity Through Circuit Complexity.' *The Simons Institute.*

20. Minsky, M., Papert, S. (1969). *Perceptrons: An Introduction to Computational Geometry.* Cambridge, MA, USA: MIT Press.

21. OpenAi, 2023. (2023) 'Planning for AGI and beyond.'

22. Pérez, J., Marinković, J. and Barceló, P., 2019. 'On the turing completeness of modern neural network architectures.' arXiv preprint arXiv:1901.03429.

23. Russell, S. J., Norvig, P., and Davis, E. (2010) *Artificial intelligence: a modern approach.* 3rd ed. Upper Saddle River, NJ, Prentice Hall.

24. Udandarao, V., Prabhu, A., Ghosh, A., Sharma, Y., Torr, P.H., Bibi, A., Albanie, S. and Bethge, M. (2024) No "zero-shot" without exponential data: Pretraining concept frequency determines multimodal model performance. arXiv preprint arXiv:2404.04125.

25. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V. and Zhou, D., 2022. 'Chain-of-thought prompting elicits reasoning in large language models.' *Advances in neural information processing systems,* 35, pp.24824-24837.

26. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit J., Jones L., Gomez, Aidan N; Kaiser, Ł., Polosukhin I. (2017) 'Attention is All you Need. Advances in Neural Information Processing Systems.' 30. Curran Associates, Inc.

27. Zhang, R., Frei, S. and Bartlett, P.L. (2023). 'Trained transformers learn linear models in-context.' arXiv preprint arXiv:2306.09927.