System 2 Reasoning Capabilities Are Nigh

Scott C. Lowe
Vector Institute
Toronto, Canada
scott.lowe@vectorinstitute.ai

Abstract

In recent years, machine learning models have made strides towards human-like reasoning capabilities from several directions. In this work, we review the current state of the literature and describe the remaining steps to achieve a neural model which can perform System 2 reasoning analogous to a human. We argue that if current models are insufficient to be classed as performing reasoning, there remains very little additional progress needed to attain that goal.

1 Introduction

The dual process theory of thought processes is long standing within psychology (Wason & Evans, 1974; Evans, 2008; Stanovich & West, 2000) and was popularized more broadly by Kahneman (2012). In this framework, human thinking capabilities are conceptualized as two distinct modes of thought. System 1 is fast, automatic, instinctive, and more emotional; System 2 is slower, effortful, deliberate, and more logical. System 1 is, in essence, unconscious thought, and System 2 is conscious thought; though there is not yet consensus on whether "ah-ha" moments which come following an incubation period are triggered by unconscious work of System 1 or 2 (Christensen, 2005; Gilhooly, 2016). Additionally, due to its instinctive and reactive nature, System 1 is more prone to bias than System 2, though System 2 is not without bias (Tversky & Kahneman, 1974).

In comparison to these two cognitive systems of thought, feed-forward neural networks are sometimes described as being analogous to System 1. Their outputs are immediate and automatic, yielded immediately without what might call "deliberation". Like with System 1, the computational system producing the output does not and can not provide an explicit explanation for *why* it produced a certain response, making interpretability challenging, even when attempting to induce it to provide an a posteriori justification for its response (Jung et al., 2022). Such systems are effectively performing pattern matching for their current stimulus against the body of data imbibed during training.

In comparison, symbolic rule-based algorithms (classical "artificial intelligence"), whether they are manually or programatically created, can provide an explanation for their reasoning. However their performance is limited because the space of the real-world is too large to be handled with a narrow set of rules that are coded in the stimulus domain (Yihe et al., 2019).

In this work, we review the existing literature in the space of reasoning from the perspective of cognitive psychology and machine learning, and we speculate on what form a neural network would need to take for it to be able to perform reasoning in the style of System 2. We argue the majority of hurdles needed to achieve this task have already been cleared, and there are a small number of pieces of the puzzle remaining. Thus complex agents, trained through deep learning, that can reason logically about the real-world will be available in the near-term, if they are not here already.

The First Workshop on System-2 Reasoning at Scale, NeurIPS 2024.

2 Background

2.1 Modalities of human thought

Historic texts indicate that ancient philosophers such as Plato believed that thinking was synonymous with an inner monologue. However, whilst an internal monologue (inner speech) is common, it is not ubiquitous and most people overestimate how often their thoughts are expressed verbally (Hurlburt et al., 2013). There is a wide variety of inner experiences across humans (Hurlburt & Heavy, 2006), and most people are surprised when they first discover that other people's internal experiences differ greatly from their own.

The modalities of human inner thought are (Hurlburt & Schwitzgebel, 2011; Hurlburt et al., 2013):

- Inner speaking/inner monologue — thoughts expressed verbally, e.g. talking to yourself, hearing your/a voice while recalling.
- Inner seeing/visual imagery — thoughts expressed visually, e.g. picturing a memory or imagining a hypothetical scene.
- Feelings — a conscious experience of emotional processes, e.g. sadness when grieving.
- Unsymbolized thinking — thoughts expressed without words or images, e.g. drinking a glass of water, without internal discussion or commentary.
- Sensory awareness — attending to a sensory aspect of the environment for an unimportant reason, e.g. hearing someone talk but seeing the light reflecting off their glasses.

Most people experience inner speech and inner imagery some of the time but not all of the time, with the majority of their thought processes unsymbolized (Hurlburt et al., 2013). However there are outliers in each direction, with some people having no inner speech (anauralia), constant inner speech, no mind's eye (aphantasia), or extremely vivid mental imagery as detailed as sensory stimuli (hyperphantasia). Day-to-day observations of people across society demonstrate, and academic studies confirm, that people are able to complete tasks irrespective of whether their internal thoughts are represented through speech, imagery, or neither (Keogh et al., 2021; Hinwar & Lambert, 2021); though the lack of inner sight does impair the ability to recall visual characteristics (Monzel et al., 2022; Bainbridge et al., 2021). Additionally, note that those possessing an inner monologue who speak multiple languages can have their inner monologue flip between languages depending on recent context. These observations lead us to hypothesize that **conscious thoughts** (i.e. System 2 thinking) **are fundamentally abstract in nature**, but can be **projected to language and visual modalities** internally.

2.2 What is System 2 reasoning?

As a thought exercise, consider the task of solving this illustrative example from the Cognitive Reflection Test (Frederick, 2005):

> A bat and a ball together cost $1.10. The bat costs $1 more than the ball. How much does the ball cost?

For illustrative purposes, please briefly solve the problem yourself before reading ahead. System 1 is responsible for the automatic response which immediately comes to mind on reading the problem: ten cents. This answer is yielded in an involuntary manner, seemingly to all who hear the question for the first time. However, by engaging System 2 we can verify whether the intuitive solution is correct, and reason about it. By reflecting on the intuitive answer, one can observe that if this were the price of the ball, the total would be $1.20, hence the answer is incorrect. Since as the total price is the difference in price between the two objects ($1) plus *twice* the price of the ball, the answer is in fact 5 cents.

If we analyze it, it appears that the instinctive response stems from pattern matching—the problem looks at first glance like other problems comparing the quantity of two items, which we have solved in the past using the "subtraction" method, hence we instinctively try to apply it here. Analogous mistakes can be seen in large language model (LLM) outputs, where they respond to a fake logic puzzle as if it were real (Richardson, 2024).

One way to conceptualize such reasoning is as a series of hypothesis generation and verification steps. If the initial hypothesis fails the verification, we then come up with a new hypothesis conditioned on the new information generated during the verification process. This process is repeated until a revised hypothesis satisfies the verification step. Such a framework is similar to the Actor-Critic reinforcement learning algorithm (Konda & Tsitsiklis, 1999), with the actor analogous to the hypothesis generator and the critic as the verifier. This could be interpreted as alternating steps of System 1 (for hypothesis generation) and 2 (critique), however such a construction is contrary to prevalent psychological models of cognitive processes such as the dual process theory.

Alternatively, human reasoning can be conceptualized as a train of thought in a continuous stream of consciousness (Potter et al., 2014; James, 2012). This framework is comparable to the chain-of-thought LLM prompting technique (Wei et al., 2022), in which a model reaches its final output by focusing on a series of steps.

2.3 Existing neural reasoning agents

Previous work has found that by prompting LLMs with an in-context example of chain-of-thought reasoning, and asking it to think step-by-step for its own answer, models can be coerced into "thinking" step-by-step (Wei et al., 2022). Providing such a prompt changes the distribution of most likely initial next tokens to be steps towards the solution instead of an immediate answer. By having the model attend to its own outputs as it progresses, it can build on its previous steps, eventually producing a final result that is more likely to be accurate (Wei et al., 2022; Li et al., 2024). However, recent work has demonstrated the majority of the gains seen when using chain-of-thought prompting can be matched by prompting with a long series of task-independent filler tokens instead, suggesting the length of the sequence and the size of the compute graph is more important than the textual output (Pfau et al., 2024). This implies the transformer can process data through unseen computations within the hidden layers of the network, unwitnessed in the chain-of-thought tokens that it outputs. Such findings may be analogous to System 2 reasoning in humans, which we noted in Section 2.1 are primarily non-symbolic but can be projected to consciously observed language streams (Hurlburt et al., 2013), though such a hypothesis is challenging to investigate due to the difficulties of interpreting deep transformer representations (Rai et al., 2024).

In the domain of closed-world games, tremendous gains were seen by applying deep reinforcement learning models that learn through self-play to optimize a value function (Silver et al., 2016, 2017, 2018). In this case, the value network can be fit to predict the likelihood of each player winning the game from a given position. The result of the game can be objectively determined by continuing to play the match and seeing who wins, providing a strong training signal to the value network. And since the value network is able to fit this task, it is able to steer the actor model effectively. Results are further improved by performing a Monte Carlo Markov chain tree search at inference time, guided by the model's predictions to prune the tree to a narrow range of feasible moves, to evaluate future game states and choose an optimal move. Such searches are similar to the Tree-of-thoughts approach to improve chain-of-thoughts reasoning (Long, 2023; Yao et al., 2023).

Similarly, when deploying LLMs on mathematics problems, step-level verification of chain-of-thought has been shown to be effective training technique (Cobbe et al., 2021; Uesato et al., 2022; Lightman et al., 2024).

3 Future steps and potential pitfalls

3.1 Learning to reason

Given the existing neural reasoning techniques, and their analogous relationship to human reasoning processes, we posit that networks already can learn to reason.

Multiple works have shown training LLMs to reason step-by-step is best achieved by step-level feedback (Zelikman et al., 2022; Pfau et al., 2024; Lightman et al., 2024). One issue for training a reasoning model at scale is thus that there is a lack of large-scale reasoning datasets to train on in which humans have written out their train of thoughts explicitly. However, such data can be acquired at modest scale (e.g. Yang et al., 2018), and, by explicitly labelling which steps are valid and which are not, such data can be used to train a verifier that predicts whether individual logical reasoning steps are sound. This verifier, similar to the rationalization evaluator used by Zelikman et al. (2022),

can then serve a similar role to the step-wise maths problem solver of Lightman et al. (2024). Using this, we can bootstrap more chain-of-thought data by tasking a pretrained LLM with chain-of-thought prompting to generate more reasoning data, and discarding outputs which contain steps which do not pass the verifier, similar to that used for maths problems by Lightman et al. (2024).

Note that the verifier is an essential part of this pipeline, and it must be accurate in order for the iterative self-distillation to be effective. But in any scenario where verification is easier than generation, the verifier (even if learnt and imperfect) can be deployed to iteratively refine and distill the generative model (Christiano et al., 2018). An alternative bootstrap formulation would be to generate a large body of chain-of-thoughts data using chain-of-thought prompting applied on a large corpus of problems with known solutions. We then train a verifier to, given a particular point in the chain-of-thought, classify whether the model will get the right answer. This verifier model will serve a similar role to the value function in self-play RL systems (Silver et al., 2018), and we can fine-tune our model to generate its step of thoughts whilst trying to maximize the verifier's probability the problem will be solved. Since such a system bears similarity to Q-learning and STaR bootstrap reasoning (Zelikman et al., 2022), it might be aptly given the name "Q*". We note that other recent work has successfully applied reinforcement learning fine-tuning to pretrained LLMs, such as reinforcement with human feedback (RLHF) (Ziegler et al., 2020) or with harmlessness feedback (Bai et al., 2022); and these methods can be improved by modifying the method to provide direct feedback (Rafailov et al., 2023; Lee et al., 2024). The implementation we propose would be a similar reinforcement learning fine-tuning stage, but with a objective focused on reasoning accuracy.

When such a model is half-way through its reasoning at deployment time, the steps that the model has already produced are provided back to it as in input to generate the next step. However, the inaccuracy of the model creates a domain-shift between its training data and the data it sees during inference—during training teacher forcing means the model only sees accurate steps, but at deployment time the previous steps can be incorrect. From a reinforcement learning perspective, such a training configuration is "off-policy", which can be corrected for by (re)training the model on its own outputs (on-policy), allowing it to learn to back-track when its inputs include incorrect steps (Kumar et al., 2024). We note that the importance of backtracking rises *exponentially* with the number of steps in the reasoning process, since every reasoning step is an opportunity to make a mistake.

All the components for this solution for a reasoning agent framework seemingly already exist in the literature, and it is even possible such a model has already been trained recently (OpenAI, 2024).

3.2 Applicability

LLMs trained only on textual data are unlikely to master reasoning about the real-world, since their observations of it are highly indirect. When humans communicate with each other, they do so with a large body of common experiences merely from both being creatures raised and living in the real world. This means that many things that are taken for granted remain unstated as they are assumed to be known by all parties in the discourse.

In order for foundation models to be able to reason efficiently about the world, we speculate they will need a world model that is built on sensory observations, not just text descriptions. More recent foundation models have made progress in this direction (Zhang et al., 2022) by being multi-modal—processing both language and visual stimuli. However, we posit that further gains will be made when using data which captures the richness of the real world through video data and (less abundantly) embodied sensorimotor data. Video data has rich features about the world, enabling the network to construct its own intuitive physics, infer cause and effect (Bardes et al., 2024).

3.3 Scaling

Will scaling laws continue to hold for chain-of-thought reasoning, or will such models hit scaling problems?

The "bitter lesson" of machine learning has been that gains from methods that can exploit generic compute scaling (e.g. larger and more flexible models, trained increasingly large datasets), in the long-run outperform gains from human-knowledge adjustments due to Moore's law (Sutton, 2019). Thus we postulate that reasoning models will naturally also benefit from utilizing general methods rather than hand-tuned routines. This is evidenced by recent work deploying LLMs on mathematical

problems (Snell et al., 2024), which found that evaluation performance increases as the amount of inference compute increases.

However, one possible obstacle is the quadratic scaling of transformers with respect to their input sequence length due to their all-to-all attention. Inefficient chain-of-thought reasoning will create excessively verbose thought-histories, greatly increasing the amount of compute required to reach the end of a chain-of-thought. This poses a challenge to efficiently utilize compute when the model's inference steps are scaled up. There have been various attempts to modify transformers to scale better (Child et al., 2019; Choromanski et al., 2021; Dao et al., 2022; Dao, 2024). Recently there have also been orthogonal efforts towards SOTA LLMs that are built using State Space Model (SSM) architectures (Gu et al., 2022; Poli et al., 2023; Gu & Dao, 2023; Dao & Gu, 2024).

More critically, as the number of entities to reason about grows, the number of potential interactions between the entities grows exponentially. This has the potential to out-scale the computational resources available to train and deploy reasoning models. However, we note that human working memory is limited to 7 ± 2 objects or chunks across a variety of tasks, where the number and size of chunks depends on the individual's familiarity with the items being held in memory (Miller, 1956). This implies that reasoning does not require all-to-all attention over objects in the thought history, rather it only requires a constant memory space. The remaining challenges are (1) items being held in memory must be appropriately compact; (2) when only a limited number of items are retained in memory, the model must learn which memories to keep and which to drop.

With regards to compactness, this is a challenge for token-based models as typically the embedding space has the same granularity as the stimulus space. Yet recent hierarchical models from the vision literature offer insights into how a hierarchical token-based model may look, in which the embedding space is more spatially compact than the stimulus representations (Liu et al., 2021; Fan et al., 2021; Li et al., 2022; Ryali et al., 2023).

With regards to selecting memories to retain, recent work on memory-augmented transformers (Bulatov et al., 2024) and on SSMs that can select and retain memories in their state-space (Dao & Gu, 2024) each provide research directions towards this goal, though there is still work to be done. Even if memory selection remains challenging, less efficient reasoning models will be possible in the meantime.

3.4 Safety concerns

As new capabilities are introduced to AI models, it is important to monitor these frontier models for potential safety risks (Phuong et al., 2024). From an AI control perspective, ML agents which can reason and strategically plan present a much larger risk than passive models which merely predict things. Like any ML model in deployment, there is a societal risk that the model's learnt biases from its training distribution will result in its behaviour diverging from human aspirations.

But more importantly, such a model raises the existential risk from AI models. Models which can reason can use their abilities to plan and strategize, potentially over the long-term. If allowed to act autonomously to achieve a goal, they may erroneously or surreptitiously plan with subgoals that involve taking control of resources they should not have access to, etc[1]. To mitigate these concerns, it is important that training data be screened to ensure it does not contain instructions we would not wish an agent to take when deployed in the wild.

Another concern regards the scrutability of reasoning agents. Current LLMs must always project their chain-of-thought reasoning steps to English, though there are concerns that their internal computation may not be fully reflected in their outputs (Pfau et al., 2024; Lyu et al., 2023; Lanham et al., 2023). From a gain of function perspective, it may be advantageous to train models that can reason in abstract concepts that do not directly correspond to tokens in the training corpus. However, we are of the opinion that steps must always be taken to ensure that model reasoning is projected into a frame (be it language or imagery) in which it can be explicitly and as completely as possible communicated to humans.

[1]We will not go into unnecessarily explicit details, as it is plausible this paper may be included in the training data of future LLMs.

4 Conclusions

We have discussed the literature surrounding the philosophy of human inner thought and reasoning, and the current neural network approaches to reasoning models. The current networks have strong analogues to processes ascribed to human reasoning. We thus argue they already achieve reasoning, though to limited degrees due to either their limited domains or lack of explicit training.

From this, we propose a pipeline which combines several existing techniques from the machine learning literature together as a candidate for how a reasoning agent could be explicitly trained to reason. By expanding the breadth of training data to include richer, raw, temporal stimuli such as video, we anticipate the model can achieve a more capable world model to anchor its representations and better reason about the real world. Thus we conclude that neural reasoning models are either already here, or if not they will be soon.

Acknowledgments and Disclosure of Funding

Many thanks to David Emerson, Iulia Eyriay, Kevin Kasa, Kristen Menou, and Michael Zhang for insightful discussions and feedback, and to Philip from AI Explained for providing the initial inspiration (AI Explained, 2024).

Resources used in preparing this work were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring[2] the Vector Institute.

References

AI Explained. o1 - What is going on? Why o1 is a 3rd paradigm of model + 10 things you might not know, Sept 2024. URL https://www.youtube.com/watch?v=KKF7kL0pGc4.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., Kerr, J., Mueller, J., Ladish, J., Landau, J., Ndousse, K., Lukosuite, K., Lovitt, L., Sellitto, M., Elhage, N., Schiefer, N., Mercado, N., DasSarma, N., Lasenby, R., Larson, R., Ringer, S., Johnston, S., Kravec, S., Showk, S. E., Fort, S., Lanham, T., Telleen-Lawton, T., Conerly, T., Henighan, T., Hume, T., Bowman, S. R., Hatfield-Dodds, Z., Mann, B., Amodei, D., Joseph, N., McCandlish, S., Brown, T., and Kaplan, J. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022. doi:10.48550/arxiv.2212.08073.

Bainbridge, W. A., Pounder, Z., Eardley, A. F., and Baker, C. I. Quantifying aphantasia through drawing: Those without visual imagery show deficits in object but not spatial memory. *Cortex*, 135:159–172, 2021. ISSN 0010-9452. doi:10.1016/j.cortex.2020.11.014.

Bardes, A., Garrido, Q., Ponce, J., Rabbat, M., LeCun, Y., Assran, M., and Ballas, N. Revisiting feature prediction for learning visual representations from video. *arXiv:2404.08471*, 2024. doi:10.48550/arxiv.2404.08471.

Bulatov, A., Kuratov, Y., Kapushev, Y., and Burtsev, M. S. Scaling transformer to 1M tokens and beyond with RMT. *arXiv preprint arXiv:2304.11062*, 2024. doi:10.48550/arxiv.2304.11062.

Child, R., Gray, S., Radford, A., and Sutskever, I. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019. doi:10.48550/arxiv.1904.10509.

Choromanski, K. M., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., Davis, J. Q., Mohiuddin, A., Kaiser, L., Belanger, D. B., Colwell, L. J., and Weller, A. Rethinking attention with performers. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=Ua6zuk0WRH.

Christensen, B. Problematic assumptions in incubation effect studies and what to do about them. *Creative Cognition: Analogy And Incubation*, 2005.

[2] https://vectorinstitute.ai/partnerships/current-partners/

Christiano, P., Shlegeris, B., and Amodei, D. Supervising strong learners by amplifying weak experts. *arXiv preprint arXiv:1810.08575*, 2018. doi:10.48550/arxiv.1810.08575.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. doi:10.48550/arxiv.2110.14168.

Dao, T. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations (ICLR)*, 2024.

Dao, T. and Gu, A. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*, 2024. doi:10.48550/arxiv.2405.21060.

Dao, T., Fu, D. Y., Ermon, S., Rudra, A., and Ré, C. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Evans, J. S. B. T. Dual-processing accounts of reasoning, judgment, and social cognition. *Annu. Rev. Psychol.*, 59(1):255–278, 2008.

Fan, H., Xiong, B., Mangalam, K., Li, Y., Yan, Z., Malik, J., and Feichtenhofer, C. Multiscale vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6804–6815, 2021. doi:10.1109/ICCV48922.2021.00675.

Frederick, S. Cognitive reflection and decision making. *Journal of Economic Perspectives*, 19(4): 25–42, December 2005. doi:10.1257/089533005775196732.

Gilhooly, K. J. Incubation and intuition in creative problem solving. *Frontiers in Psychology*, 7, 2016. ISSN 1664-1078. doi:10.3389/fpsyg.2016.01076.

Gu, A. and Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. doi:10.48550/arxiv.2312.00752.

Gu, A., Goel, K., and Ré, C. Efficiently modeling long sequences with structured state spaces. In *The International Conference on Learning Representations (ICLR)*, 2022.

Hinwar, R. P. and Lambert, A. J. Anauralia: The silent mind and its association with aphantasia. *Frontiers in Psychology*, 12, 2021. ISSN 1664-1078. doi:10.3389/fpsyg.2021.744213.

Hurlburt, R. and Heavy, C. *Exploring Inner Experience: The Descriptive Experience Sampling Method*. Advances in consciousness research. John Benjamins Pub., 2006. ISBN 9789027252005.

Hurlburt, R. and Schwitzgebel, E. *Describing Inner Experience?: Proponent Meets Skeptic*. Life and Mind: Philosophical Issues in Biology and Psychology. MIT Press, 2011. ISBN 9780262516495.

Hurlburt, R. T., Heavey, C. L., and Kelsey, J. M. Toward a phenomenology of inner speaking. *Consciousness and Cognition*, 22(4):1477–1494, 2013. ISSN 1053-8100. doi:10.1016/j.concog.2013.10.003.

James, W. *The Principles of Psychology, Vol. 1*. Number v. 1. Dover Publications, 2012. ISBN 9780486123493.

Jung, J., Qin, L., Welleck, S., Brahman, F., Bhagavatula, C., Le Bras, R., and Choi, Y. Maieutic prompting: Logically consistent reasoning with recursive explanations. In Goldberg, Y., Kozareva, Z., and Zhang, Y. (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 1266–1279, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi:10.18653/v1/2022.emnlp-main.82.

Kahneman, D. *Thinking, fast and slow*. Penguin, London, 2012. ISBN 9780141033570 0141033576.

Keogh, R., Wicken, M., and Pearson, J. Visual working memory in aphantasia: Retained accuracy and capacity with a different strategy. *Cortex*, 143:237–253, 2021. ISSN 0010-9452. doi:10.1016/j.cortex.2021.07.012.

Konda, V. and Tsitsiklis, J. Actor-critic algorithms. In Solla, S., Leen, T., and Müller, K. (eds.), *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999. URL https://proceedings.neurips.cc/paper_files/paper/1999/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

Kumar, A., Zhuang, V., Agarwal, R., Su, Y., Co-Reyes, J. D., Singh, A., Baumli, K., Iqbal, S., Bishop, C., Roelofs, R., Zhang, L. M., McKinney, K., Shrivastava, D., Paduraru, C., Tucker, G., Precup, D., Behbahani, F., and Faust, A. Training language models to self-correct via reinforcement learning. *arXiv preprint arXiv:2409.12917*, 2024. doi:10.48550/arxiv.2409.12917.

Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., Denison, C., Hernandez, D., Li, D., Durmus, E., Hubinger, E., Kernion, J., Lukošiūtė, K., Nguyen, K., Cheng, N., Joseph, N., Schiefer, N., Rausch, O., Larson, R., McCandlish, S., Kundu, S., Kadavath, S., Yang, S., Henighan, T., Maxwell, T., Telleen-Lawton, T., Hume, T., Hatfield-Dodds, Z., Kaplan, J., Brauner, J., Bowman, S. R., and Perez, E. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023. doi:10.48550/arxiv.2307.13702.

Lee, H., Phatale, S., Mansoor, H., Mesnard, T., Ferret, J., Lu, K., Bishop, C., Hall, E., Carbune, V., Rastogi, A., and Prakash, S. RLAIF vs. RLHF: Scaling reinforcement learning from human feedback with AI feedback. *arXiv preprint arXiv:2309.00267*, 2024. doi:10.48550/arxiv.2309.00267.

Li, Y., Wu, C.-Y., Fan, H., Mangalam, K., Xiong, B., Malik, J., and Feichtenhofer, C. MViTv2: Improved multiscale vision transformers for classification and detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4794–4804, 2022. doi:10.1109/CVPR52688.2022.00476.

Li, Z., Liu, H., Zhou, D., and Ma, T. Chain of thought empowers transformers to solve inherently serial problems. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3EWTEy9MTM.

Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., and Cobbe, K. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9992–10002, 2021. doi:10.1109/ICCV48922.2021.00986.

Long, J. Large language model guided tree-of-thought. *arXiv preprint arXiv:2305.08291*, 2023. doi:10.48550/arxiv.2305.08291.

Lyu, Q., Havaldar, S., Stein, A., Zhang, L., Rao, D., Wong, E., Apidianaki, M., and Callison-Burch, C. Faithful chain-of-thought reasoning. In Park, J. C., Arase, Y., Hu, B., Lu, W., Wijaya, D., Purwarianti, A., and Krisnadhi, A. A. (eds.), *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 305–329, Nusa Dua, Bali, November 2023. Association for Computational Linguistics. doi:10.18653/v1/2023.ijcnlp-main.20.

Miller, G. A. The magical number seven plus or minus two: some limits on our capacity for processing information. *Psychol. Rev.*, 63(2):81–97, March 1956.

Monzel, M., Vetterlein, A., and Reuter, M. Memory deficits in aphantasics are not restricted to autobiographical memory – perspectives from the dual coding approach. *Journal of Neuropsychology*, 16(2):444–461, 2022. doi:10.1111/jnp.12265.

OpenAI. Learning to reason with LLMs, Sept 2024. URL https://openai.com/index/learning-to-reason-with-llms/.

Pfau, J., Merrill, W., and Bowman, S. R. Let's think dot by dot: Hidden computation in transformer language models. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=NikbrdtYvG.

Phuong, M., Aitchison, M., Catt, E., Cogan, S., Kaskasoli, A., Krakovna, V., Lindner, D., Rahtz, M., Assael, Y., Hodkinson, S., Howard, H., Lieberum, T., Kumar, R., Raad, M. A., Webson, A., Ho, L., Lin, S., Farquhar, S., Hutter, M., Deletang, G., Ruoss, A., El-Sayed, S., Brown, S., Dragan, A., Shah, R., Dafoe, A., and Shevlane, T. Evaluating frontier models for dangerous capabilities. *arXiv preprint arXiv:2403.13793*, 2024. doi:10.48550/arxiv.2403.13793.

Poli, M., Massaroli, S., Nguyen, E., Fu, D. Y., Dao, T., Baccus, S., Bengio, Y., Ermon, S., and Ré, C. Hyena hierarchy: Towards larger convolutional language models. In *International Conference on Machine Learning*, pp. 28043–28078. PMLR, 2023.

Potter, M. C., Wyble, B., Hagmann, C. E., and McCourt, E. S. Detecting meaning in rsvp at 13 ms per picture. *Attention, Perception, & Psychophysics*, 76(2):270–279, Feb 2014. ISSN 1943-393X. doi:10.3758/s13414-013-0605-z.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a85b405ed65c6477a4fe8302b5e06ce7-Paper-Conference.pdf.

Rai, D., Zhou, Y., Feng, S., Saparov, A., and Yao, Z. A practical review of mechanistic interpretability for transformer-based language models. *arXiv preprint arXiv:2407.02646*, 2024. doi:10.48550/arxiv.2407.02646.

Richardson, J. To an LLM, everything looks like a logic puzzle. *LessWrong*, May 2024. URL https://www.lesswrong.com/posts/qNXXe7EDGyveC4SCp.

Ryali, C., Hu, Y.-T., Bolya, D., Wei, C., Fan, H., Huang, P.-Y., Aggarwal, V., Chowdhury, A., Poursaeed, O., Hoffman, J., Malik, J., Li, Y., and Feichtenhofer, C. Hiera: a hierarchical vision transformer without the bells-and-whistles. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., and Hassabis, D. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, Jan 2016. ISSN 1476-4687. doi:10.1038/nature16961.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., and Hassabis, D. Mastering the game of Go without human knowledge. *Nature*, 550(7676): 354–359, Oct 2017. ISSN 1476-4687. doi:10.1038/nature24270.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., and Hassabis, D. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419): 1140–1144, 2018. doi:10.1126/science.aar6404.

Snell, C., Lee, J., Xu, K., and Kumar, A. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024. doi:10.48550/arxiv.2408.03314.

Stanovich, K. E. and West, R. F. Advancing the rationality debate. *Behavioral and Brain Sciences*, 23(5):701–717, 2000. doi:10.1017/S0140525X00623439.

Sutton, R. The bitter lesson. *Incomplete Ideas (blog)*, 13(1):38, 2019. URL http://www.incompleteideas.net/IncIdeas/BitterLesson.html.

Tversky, A. and Kahneman, D. Judgment under uncertainty: Heuristics and biases. *Science*, 185 (4157):1124–1131, 1974. doi:10.1126/science.185.4157.1124.

Uesato, J., Kushman, N., Kumar, R., Song, F., Siegel, N., Wang, L., Creswell, A., Irving, G., and Higgins, I. Solving math word problems with process- and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022. doi:10.48550/arxiv.2211.14275.

Wason, P. and Evans, J. Dual processes in reasoning? *Cognition*, 3(2):141–154, 1974. ISSN 0010-0277. doi:10.1016/0010-0277(74)90017-1.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. V., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W. W., Salakhutdinov, R., and Manning, C. D. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., and Narasimhan, K. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*, 2023. doi:10.48550/arxiv.2305.10601.

Yihe, L., Lowe, S. C., Lewis, P. A., and van Rossum, M. C. W. Program synthesis performance constrained by non-linear spatial relations in Synthetic Visual Reasoning Test. *arXiv preprint arXiv:1911.07721*, 2019. doi:10.48550/arxiv.1911.07721.

Zelikman, E., Wu, Y., Mu, J., and Goodman, N. STaR: Bootstrapping reasoning with reasoning. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 15476–15488. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/639a9a172c044fbb64175b5fad42e9a5-Paper-Conference.pdf.

Zhang, C., Van Durme, B., Li, Z., and Stengel-Eskin, E. Visual commonsense in pretrained unimodal and multimodal models. In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V. (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5321–5335, Seattle, United States, July 2022. Association for Computational Linguistics. doi:10.18653/v1/2022.naacl-main.390.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., and Irving, G. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2020. doi:10.48550/arxiv.1909.08593.