Thoughts on What is Needed for AGI and Why We Aren't There Yet

1. Static Nature of Current AI

One of the core issues with current AI is that it is static. This poses a problem when considering long-running processes—tasks that people regularly perform over extended periods. These tasks tend to be hyper-similar in nature over time. This presents an issue with static AI, which can be best understood through a Retrieval-Augmented Generation (RAG) method. As the data increases, the limited static space available for embedding hyper-similar tasks gets overwritten repeatedly. Consequently, when the AI selects data, it often retrieves irrelevant but hyper-similar information. This leads to a situation where the model has no way of discerning the exact information it needs for the precise step in a long-running task. [2][6]

2. Expanding Embedding Space and Continuous Learning

To address this issue, we need to expand the embedding space. For a language model without RAG, this involves training without forgetting—ensuring it can expand its internal representations without collapsing or losing previously learned knowledge. While difficult, this is not insurmountable. For instance, Elastic Weight Consolidation (EWC) shows that continuous learning without catastrophic forgetting is possible, but it comes with a trade-off: it saturates the internal state, preventing the model from learning further due to the protective mechanisms applied to its memory. To avoid this limitation, we need a system capable of expanding the dimensionality of the model while retaining its current state. For EWC, this might involve adding more connections between neurons, thereby expanding the dimensionality and allowing the model to store more information without reaching a saturation point. However, the challenge is that there are few effective methods to infinitely increase a model's dimensionality within current systems. Evolving architectures are limited because our current technology is not homogenous in construction; it comprises many differing components that do not work well together if you randomly start connecting them. [1][12][16]

3. Context and Computation Constraints

Neither context nor computation alone can solve this problem. In the case of RAG, the current approach is to process ever-larger amounts of data as the state of information grows. You can easily imagine this overlap of the embedding space increases. Your RAG will need to return ever more information for your next step. For a non-RAG model, this translates to an increasingly large context window, requiring exponentially more computation to generate a single reasoning thread. In long-running tasks, such as human work over extended periods, this would necessitate an exponentially increasing computational budget, even for minor decisions. This approach is computationally infeasible for the types of long-term tasks that AGI would need to handle. [3][13]

4. Reinforcement Learning (RL) Limitations

Reinforcement learning (RL) is another method, but it has limitations. While RL can work for certain tasks, it assumes a fixed problem with access to all future data within the task. RL can perform long-running tasks as long as the agent encounters data similar to what it has already seen. However, stepping out of the problem space or encountering new variations would cause the system to break down. [4][8]

5. A Path Forward: Continuous Learning with Episodic Memory

The real key to AGI lies in a continuous learning system with episodic memory. This system would process new information, updating its internal parameter memory after each step. With each new step, the previous reasoning would be committed to long-term memory, no longer requiring computation to re-perform this reasoning. This compounded reasoning could link over time, allowing the model to compute the next step using a fixed 'reasoning budget' while leveraging all previous steps as free computation. [5][14]

6. The Role of Forgetting and Dreaming

Another intriguing concept that may be key to AGI is the process of forgetting and dreaming. In biological systems, sleep and dreaming appear to cause the pruning of neural connections. While the exact reason for this is unknown, I speculate

that hallucinations or dreams may help connect information across disparate subjects in a denser format, potentially increasing the dimensionality of a fixed parameter space like the human brain. The fact that current AI models exhibit hallucinations may indicate that we are on the right path. In both biological brains and large language models, hallucination could be a necessary component of effective long-term reasoning. [7][9][10][15]

7. Preventing Factors for AGI

To summarize, the main obstacles preventing AGI are as follows:
- Lack of near-infinite dimensionality expansion methods that do not require additional parameters. We need a way to fold and twist a model's parameters to accommodate more information without significantly losing previously learned knowledge.
- Lack of episodic memory retention methods that rely on a fixed number of parameters from a base model. [11][17]

References

1. Chen, Y., et al. (2024). Improving language plasticity via pretraining with active forgetting. NeurIPS 2023. arXiv:2307.01163. https://doi.org/10.48550/arXiv.2307.01163

2. Shumailov, I., et al. (2023). The curse of recursion: Training on generated data makes models forget. arXiv:2305.17493. https://doi.org/10.48550/arXiv.2305.17493

3. Kaplan, J., et al. (2020). Scaling laws for neural language models. arXiv:2001.08361. https://doi.org/10.48550/arXiv.2001.08361

4. Mnih, V., et al. (2015). Human-level control through deep reinforcement learning. Nature, 518(7540), 529-533. https://doi.org/10.1038/nature14236

5. Ibrahim, A., et al. (2024). Simple and scalable strategies to continually pre-train large language models. arXiv:2403.08763. https://doi.org/10.48550/arXiv.2403.08763

6. Mireshghallah, F., et al. (2022). Memorization in NLP fine-tuning methods. arXiv:2205.12506. https://doi.org/10.48550/arXiv.2205.12506

7. Alemohammad, S., et al. (2023). Self-consuming generative models go MAD. arXiv:2307.01850. https://doi.org/10.48550/arXiv.2307.01850

8. Mnih, V., et al. (2013). Playing Atari with deep reinforcement learning. arXiv:1312.5602. https://doi.org/10.48550/arXiv.1312.5602

9. Gekhman, Z., et al. (2024). Does fine-tuning LLMs on new knowledge encourage hallucinations? arXiv:2405.05904. https://doi.org/10.48550/arXiv.2405.05904

10. Ha, D., and Schmidhuber, J. (2018). World models. arXiv:1803.10122. https://doi.org/10.48550/arXiv.1803.10122

11. Das, A., et al. (2024). A decoder-only foundation model for time-series forecasting. ICML 2024. Google Research. https://research.google/blog/a-decoder-only-foundation-model-for-time-series-forecasting/

12. Parisi, G. I., et al. (2018). Continual lifelong learning with neural networks: A review. arXiv:1802.07569. https://doi.org/10.48550/arXiv.1802.07569

13. Hernandez, D., et al. (2022). Scaling laws and interpretability of learning from repeated data. arXiv:2205.10487. https://doi.org/10.48550/arXiv.2205.10487

14. Madaan, D., et al. (2021). Representational continuity for unsupervised continual learning. arXiv:2110.06976. https://doi.org/10.48550/arXiv.2110.06976

15. Zhou, H., et al. (2022). Fortuitous forgetting in connectionist networks. arXiv:2202.00155. https://doi.org/10.48550/arXiv.2202.00155

16. Kirkpatrick, J., et al. (2017). Overcoming catastrophic forgetting in neural networks. PNAS, 114(13), 3521-3526. https://doi.org/10.48550/arXiv.1612.00796