Global Context for Multimodal Agents

Bridging Modalities through Dynamic Low-Rank Adaptation

Volodymyr Seliuchenko

Abstract

This work introduces a novel approach inspired by Low Rank Adaptation (LoRA) to effectively couple large AI models processing diverse modalities. By leveraging a global context temporal buffer of low rank states, we establish a coherent framework for integrating pre-trained domain-specific foundation models into a single multi-modal agent. The proposed architecture facilitates the development of self-learning free-running AI agents that demonstrate striking parallels to biological intelligence.

1 Introduction

Current approaches to multi-modal AI face a fundamental trade-off: training unified models from scratch is computationally prohibitive, while naively combining pre-trained foundation models fails to achieve coherent cross-modal reasoning. Moreover, existing architectures process inputs in discrete episodes rather than continuous temporal streams, limiting their applicability to autonomous agent applications that require persistent interaction with dynamic environments.

We address these challenges by introducing **Dynamic Low-Rank Adaptation (DyLoRA)**, a technique that extends static LoRA fine-tuning (Hu et al. 2021) to dynamically generate adaptation states. Rather than learning fixed low-rank matrices BA for model adaptation, we train a prediction network $\mathbf{q}$ that generates time-varying low-rank states based on multi-modal inputs and temporal context. This enables efficient coupling of frozen foundation models across modalities without expensive joint pre-training.

The key architectural component is the **global context** (GC)—a temporal FIFO buffer storing sequences of low-rank states, encoded inputs, and metadata. The GC functions as a memory-efficient state representation with three critical properties: (1) it extends effective context length beyond transformer attention limits ($O(n^2)$ complexity); (2) it enables information exchange between heterogeneous foundation models through shared state representation; and (3) it provides a compact state space for reinforcement learning-based continuous operation.

Our approach yields three concrete contributions:

1. **Scalable context extension**: We demonstrate that transformers augmented with GC can process sequences of millions of tokens by storing compressed states rather than full attention matrices, achieving $O(N \log N)$ complexity through locality-sensitive hashing.
2. **Modular multi-modal integration**: Foundation models remain frozen while lightweight LR prediction networks mediate cross-modal information flow. This decoupled architecture permits independent training on imbalanced datasets and computational cost dominated by inference rather than training.
3. **Continuous autonomous agents**: By framing the GC as a reinforcement learning state space, we enable free-running agents that learn from temporal interaction streams. The LR prediction network functions as a continuous DQN, while the GC serves as both working memory and replay buffer.

Preliminary experiments on LLM fine-tuning demonstrate that GC-based adaptation achieves performance comparable to standard LoRA, even when local context is ablated, confirming effective information extraction and storage in the compressed global state.

The architecture exhibits computational efficiency analogous to biological intelligence: specialized processing modules (foundation models), cross-modal coordination (LR coupling), hierarchical memory (local context, GC, learned weights), and offline consolidation phases (training on replay buffer). This establishes a tractable path toward general-purpose agents that process continuous multi-modal streams.

2 Background and Related Work

Although multi-modal AI has received considerable attention, significant challenges persist. These include data alignment across modalities, handling missing modalities, addressing modality imbalance, and ensuring scalability of training and inference processes. Current research efforts focus on addressing these challenges to enhance the capabilities of multi-modal AI systems.

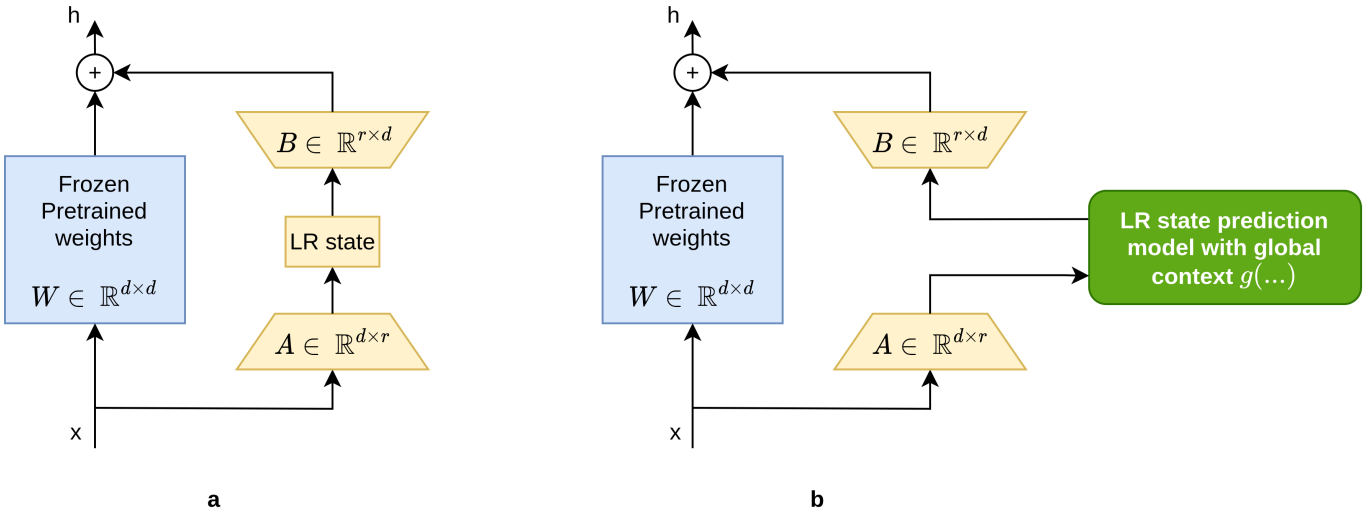

Figure 1: Lora vs DyLoRA

Training multi-modal AI models from scratch can be computationally prohibitive. To address this challenge, we propose leveraging existing domain-specific models and integrating them to create a unified multi-modal system. This approach draws inspiration from biological intelligence, where the brain exhibits functional specialization across different regions. By utilizing pre-trained models specialized in distinct domains, we can harness their expertise and integrate them into a cohesive framework. This fine-tuning-based strategy for constructing multi-modal models offers a cost-effective alternative to training from scratch while maintaining the modularity observed in biological intelligence.

AI finetuning is a crucial aspect of developing high-performing machine learning models. It involves the process of refining a pretrained model on a specific task or dataset to improve its performance, adapt it to new domains, or address specific requirements. Finetuning allows models to leverage the knowledge gained from pretraining while tailoring them to specific applications, leading to more accurate and effective results. An extensive overview of different finetuning strategies is given in Kaplan et al. (2020).

One notable prior art in this area is the Low rank adaptation (LoRA) method Hu et al. (2021), which has attracted significant attention for its effectiveness in finetuning large-scale language models. LoRA offers flexibility in incorporating domain-specific information into the low-rank approximation, enabling better adaptation and fine-grained control over the model's behavior.

It was discovered in Hu et al. (2021) that rank as low as 1 can be sufficient to inject the context from a relatively small finetuning dataset into the main large language model. Building on top of success of LoRA finetuning, we hypothesise that the low rank adaptation states can be also used to provide a wide context for a AI generative models in a dynamic manner. We propose a new AI model architecture that dynamically derives low rank adaptation states from multple modalities and previous model experience enabling efficient continuous mode multi-modal generative AI. Furthermore, we develop the proposed architecture to include high level cognitive functions of a human brain.

While today's language models (LLMs) are incredibly powerful, relying solely on textual information may not be sufficient to construct a comprehensive understanding of the real world. Biological intelligence (BI), on the other hand, operates with series of multimodal signals, processing information over time. In this paper, we delve into the exploration of architectures that enable continuous operation with streams of information from multiple modalities, aiming to approximate the functioning of the human brain.

3 Methodology

The fundamental concept of low-rank adaptation is illustrated in Figure 1(a) and Equation 1. Each fully connected coefficient matrix W in the original model is augmented with a low-rank adaptation matrix BA, where only the BA component is trained during fine-tuning.

$$h = Wx + BAx \tag{1}$$

Extensive LoRA fine-tuning experiments demonstrate

that complex generation patterns, defined by relatively small fine-tuning datasets, can be effectively encoded using the low-rank state of the adapter matrix BA. We propose replacing static low-rank adaptation with Dynamic Low Rank Adaptation (DyLoRA), which predicts the low-rank adaptation state (LR state) based on external information modalities and previous experience.

This concept is illustrated in Figure 1(b). Rather than training static weights to derive the LR state, we propose training a network $\mathbf{q}$ that dynamically generates the low-rank adaptation state based on the current input x and previous states G_i, as shown in Equation 2.

$$h = Wx + B \cdot \mathbf{q}(Ax, G_{t-1}, G_{t-2}, ...) \tag{2}$$

The previous states G_i are stored in a first-in-first-out (FIFO) queue termed the **global context** (GC). The GC contains a temporal sequence of predicted LR states $\mathbf{q_{t-1}}, \mathbf{q_{t-2}}, ...$, inputs Ax_{t-1}, Ax_{t-2}, and additional global state information detailed in subsequent sections. Thus, the GC functions as a temporal sequence of features encoding the compressed state of the main generative AI model.

Similar to the static LoRA finetuning approach, the predicted LR state $\mathbf{q}$ dynamically establishes an appropriate context for generation, enabling various applications that will be explored in the following discussions.

3.1 Context extension

Current transformer architectures, despite their remarkable performance across various natural language processing tasks, face limitations regarding input context length. Transformers typically operate on fixed-length input sequences, and processing longer sequences requires substantially increased computational resources and memory. The attention mechanism exhibits $O(n^2)$ time complexity, practically limiting input sequence length (local context) to thousands of tokens.

The GC provides memory-efficient encoding of transformer states, enabling storage of millions of states without exceeding practical memory constraints. The time complexity of conventional attention mechanisms can be improved by adopting locality-sensitive hashing (LSH) attention, which reduces complexity to $O(N \log N)$ without significant performance degradation Kitaev, Kaiser, and Levskaya (2020).

Figure 2 illustrates the architecture of a transformer model incorporating the proposed global context. The transformer generates sequences based on local context, typically limited to thousands of tokens. As generation progresses, the corresponding LR states are stored in the global context. The LR state prediction model processes the GC as input and generates predicted LR states ($\mathbf{q}$) that are injected into the main model, biasing it toward generating the intended sequence.

The LR state prediction model $\mathbf{q}$ can be implemented using an LSH transformer Kitaev, Kaiser, and Levskaya (2020) or other architectures capable of efficiently processing long sequences.

Global context, together with LR states, can also hold other auxilary global state information, for example token position/time encoding using techniques similar to the position ecoding in the original Vaswani et al. (2017) paper.

3.2 Modality coupling

The global context can also be used to efficiently couple information from one modality to another. Figure 3 illustrates an architecture that combines text, video, audio, and other domains. This approach utilizes pretrained foundation models with fixed weights, which are interconnected through the global context register.

In this architecture, the foundation models extract features from the input data, representing them as LR (Low Rank) states. These LR states are concatenated and stored in the global context register, forming a composite LR state. The LR state prediction model takes the temporal series of composite LR states as input and generates the LR state for the subsequent generation cycle, which is then used by the foundation models.

One of the challenges in training multimodal models lies in dealing with imbalanced datasets. By using LR states to adapt the frozen foundation models, it becomes possible to decouple different modalities and train separate LR prediction models for pairs or sets of modalities. Each pair or set of modalities has its own independent prediction network within the LR state prediction model. Furthermore, after completing the decoupled pretraining of the multimodal model, a coupled network can be added to the LR state prediction model.

The global context serves as an aggregator of information from various modalities, enabling the inclusion of peripheral sensors and actuators. This comprehensive representation within the model allows for a coherent internal understanding of the world and facilitates the integration of diverse modalities into a unified framework.

3.3 Free-running self-learning agent

While current AI systems predominantly focus on static tasks, it is worth noting that biological intelligence primarily revolves around the processing of temporal sequences.

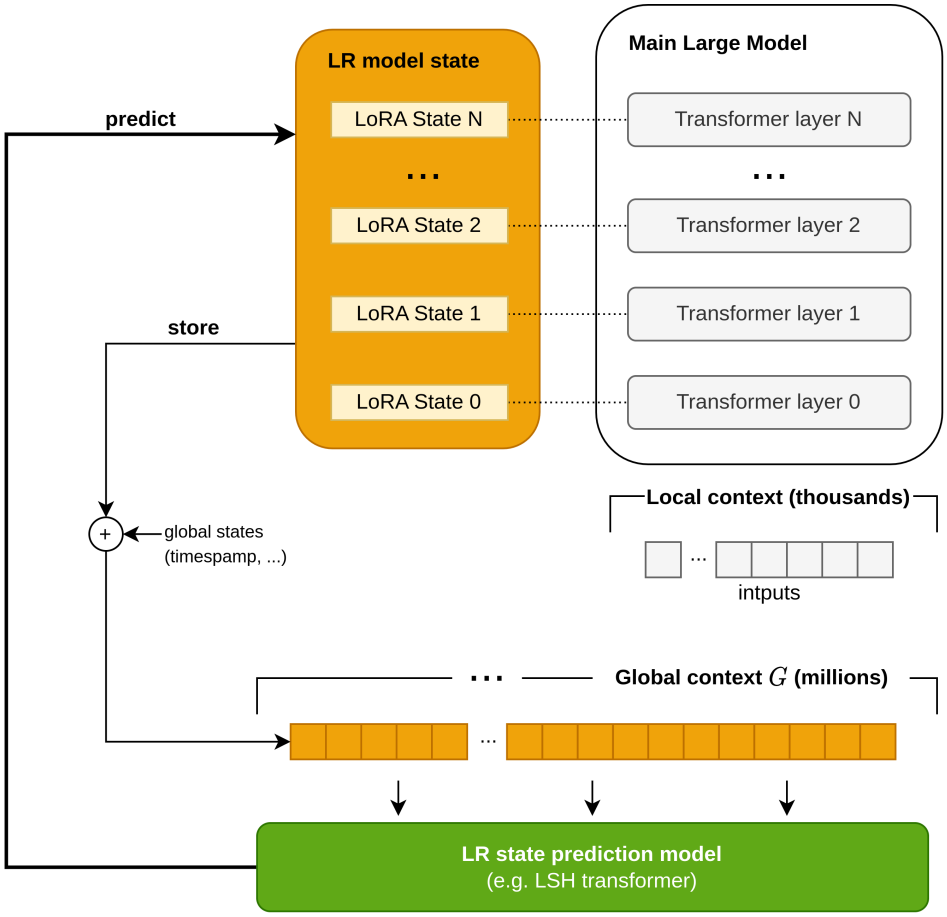

Figure 2: Transformer context extension using global context

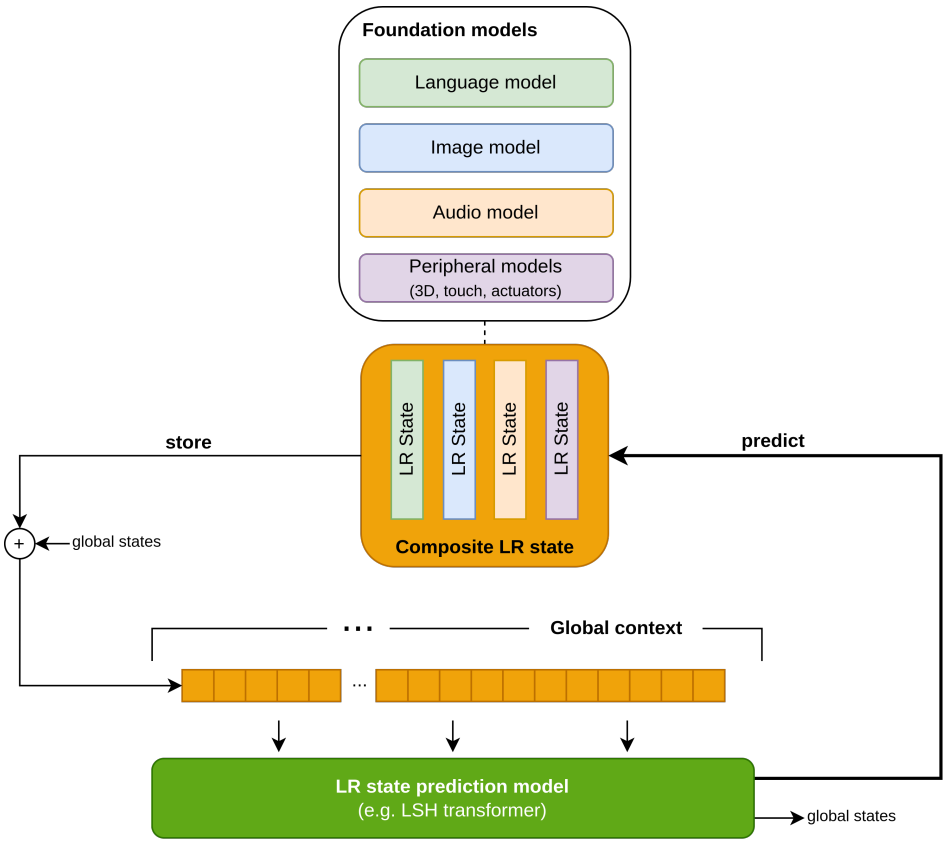

Figure 3: Multimodal model with a global context

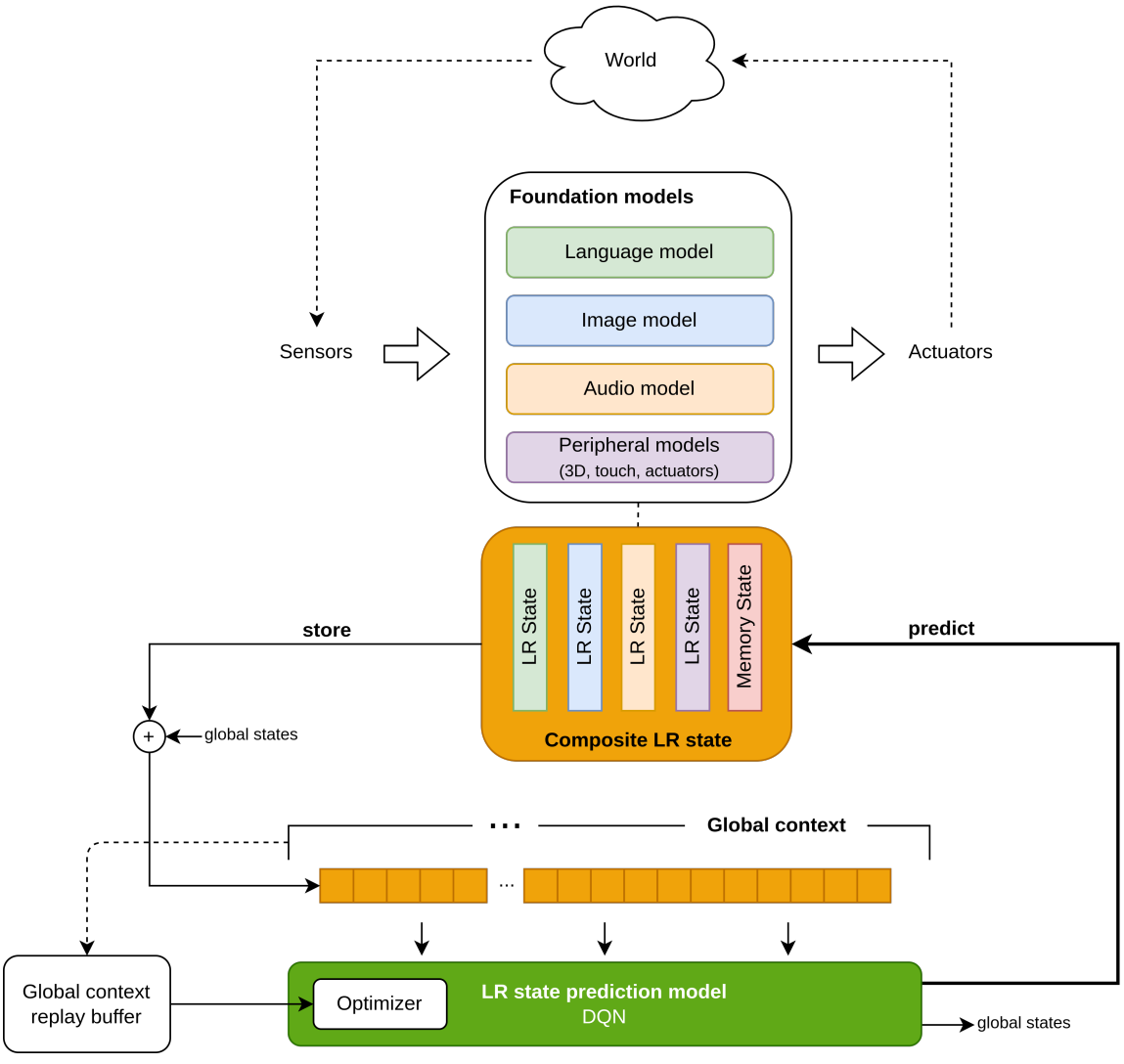

Figure 4: Free-running self-learning agent

To bridge this gap, the proposed global context architecture can be naturally extended to interact with the external world and acquire new experiences through this interaction.

In the proposed framework, the foundation models can be configured for continuous mode generation, with the objective of predicting the next state based on the current input sensory data and internal state. For instance, given previous image frames, audio inputs, and other relevant information, the foundation models aim to predict the subsequent image frame. Consequently, the domain models encompass two key components: an encoder that connects to the sensory inputs, and a decoder responsible for generating the expected next sensory input state.

Both the encoding and decoding processes are facilitated by the utilization of LR states. The prediction of the next state relies on the model's *local context* alongside the LR state prediction generated by the *LR state prediction model*. This interplay between the local context and LR state prediction ensures that the model effectively captures the temporal dependencies within the input sequences, enabling accurate generation of subsequent sensory states.

The proposed architecture, depicted in Figure 4, can be framed within the context of Reinforcement Learning (RL). The global context serves as the state space for the RL agent, while the LR state prediction model takes the form of a continuous deep Q network (DQN) that generates predictions (actions) for the next LR state, aiming to maximize a reward function. By defining the reward function, RL approaches can be employed to train the LR state prediction model, leveraging successful techniques used in challenging RL tasks such as the game of Go or Starcraft (references to be added). The global context functions as an accumulator of past experiences, serving as a replay buffer in the RL system. This enables the free-running agent to observe the surrounding environment coherently, learn from experience, and adapt its behavior to maximize the reward.

It is worth noting that the definition of the reward function is a crucial aspect of the system. At this stage, as the system operates at a highly abstract level encoded in the global context, it may be impractical or even impossible to programmatically define the reward function based on the sensory inputs directly. However, the decoder models in the proposed architecture offer interpretable insights into the current internal state. Consequently, it is feasible to define the reward function at an abstract level by fine-tuning a decoder foundation language model to generate the reward function based on observations of the internal state. Various fine-tuning approaches can be utilized (see an overview in Lialin, Deshpande, and Rumshisky (2023)), with Low rank adaptation (LoRA) being particularly promising due to its ability to reuse the frozen foundation model with minimal computational overhead and proven performance.

Therefore, the problem of reward function definition can be formulated as a set of human-interpretable statements used for the fine-tuning of the reward model. The reward value would be determined by the proximity of the internal state to the reward shaping statements. This approach aligns with our belief that formulating comprehensive laws or principles for Human-AI alignment requires more than concise statements like Asimov's Three Laws of Robotics or the Bible's Ten Commandments. Such concise formulations often leave room for misinterpretation and can potentially lead to disastrous consequences. Throughout human history, religious texts like the Bible have provided collections of stories to illustrate and clarify underlying values, resolving ambiguities and providing guidance in practical situations. These religious manuscripts can be seen as datasets that have contributed to the fine-tuning of societal values.

As AI reasoning becomes increasingly abstract, there is an escalating need for datasets that align human values with AI systems. Humans inherently possess biases, and these biases are manifested in the internet, which is a reflection of human activity. Consequently, foundation models trained solely on internet data inherit these biases, making them potentially unsafe in their raw form. To establish a reliable alignment between humanity and AI, it is essential to fine-tune generic foundation generative AI models with data specifically aligned to human values. The creation of comprehensive datasets that inject agreed-upon values into new AI models as a final fine-tuning step is crucial for shaping the reward function effectively.

The reward function LoRA adapter serves as the reference for the AI agents and is not programmable by the agent itself.

Throughout continuous operation, the global context buffer fills with composite LR states; the overflowing values are offloaded to the global context replay buffer. Reward values are stored in the global context alongside the LR states. Once the global context and replay buffer are filled, training of the DQN commences using the buffered data, adjusting the weights of the DQN to maximize the reward. An overview of different RL approaches in discrete and continuous spaces can be found in Zhu, Wu, and Zhao (2021). The optimal RL approach will be determined throughout the course of the project.

4 Training

The training process is conducted in several phases with the objective of developing an effective multi-modal network and a self-learning AI agent based on the principles outlined in this paper:

Global Context Extension of LLM Models. The first phase focuses on extending the global context of the large language models (LLMs). This step involves exploring and implementing techniques to incorporate the global context architecture into the existing LLM models.

Cross-Domain Pretraining for Multiple Modalities. The second phase involves pretraining the network for multiple modalities, starting with text and progressively incorporating other modalities such as images, video, and sound. The goal is to develop a comprehensive multi-modal framework that can effectively process and integrate information from different domains.

Reinforcement Learning. In the final phase, we conduct experiments to train and evaluate the RL agent. This involves training the agent to interact with its environment, learn from experiences, and optimize its behavior using reinforcement learning techniques.

The optimal architecture is determined through a series of experiments that explore various hyperparameters, including the rank of LR coupling.

Previous experiments with LoRA finetuning (Hu et al. 2021) have shown that the computational cost of finetuning the system is negligible compared to the cost of training the foundation models. This allows for efficient allocation of the training budget, with a focus on the computationally demanding RL experiments.

5 Experimental Results

5.1 Initial Results

We present preliminary experimental results demonstrating the effectiveness of our approach.

We evaluated the effectiveness of the global context architecture using RoBERTa LLM, with results shown in Figure 5. The model was fine-tuned on the MRPC task using both conventional LoRA (*lora baseline*) and the proposed global context architecture (*not-blinded-dynamic-with-global-context*). We observed comparable performance between these approaches. Additionally, we replaced the local context with PAD tokens, effectively "blinding" the local context (*blinded-dynamic-with-global-context*), forcing the model to rely exclusively on information processed and stored in the global context. Performance degradation was minimal, confirming the architecture's capability to extract valuable information from foundation model hidden states and effectively apply it to downstream tasks. As a control, we conducted an experiment (*blinded-dynamic-no-global-context*) with zero global context size, rendering the model unable to adapt to the task due to lack of information in both local and global contexts.

The preliminary results demonstrated in Figure 5 confirm the effectiveness of information coupling through the global context architecture, establishing a solid foundation for future expansion into multimodal and temporal coupling applications.

6 Analogies with bio-intelligence

The proposed architecture of the free-running self-learning agent aligns closely with several functions observed in biological intelligence.

Memory. The self-learning agent incorporates three types of memory:

1. "Muscle" memory, represented by the local context of the foundation models, which contains detailed information necessary for consistent generation.
2. Short-term memory, represented by the global context, holds recent experiences.
3. Long-term memory, represented by the periodically fine-tuned LR state prediction model, retains important experiences and behavioral patterns.

The local context captures raw details and is crucial for maintaining coherence in generation. Current large language models (LLMs) and audio models typically have a local context size ranging in thousands of tokens. The image and peripheral generative models would also benefit from frame buffers.

Each sample in the global context includes timestamp information that can be encoded using (e.g. using a set of harmonic functions Vaswani et al. (2017)), similar to the encoding of biological neural oscillations that may have similar timestamping function in the brain. Additionally, a decaying weight function can be applied to the global context to prioritize more recent experiences and emulate forgetting.

The periodically fine-tuned LR state prediction model that captures the most relevant information through periodic RL finetuning can be considered as a long term associative memory.

Furthermore, the global context may include a set of states that are *not* connected to the foundation models

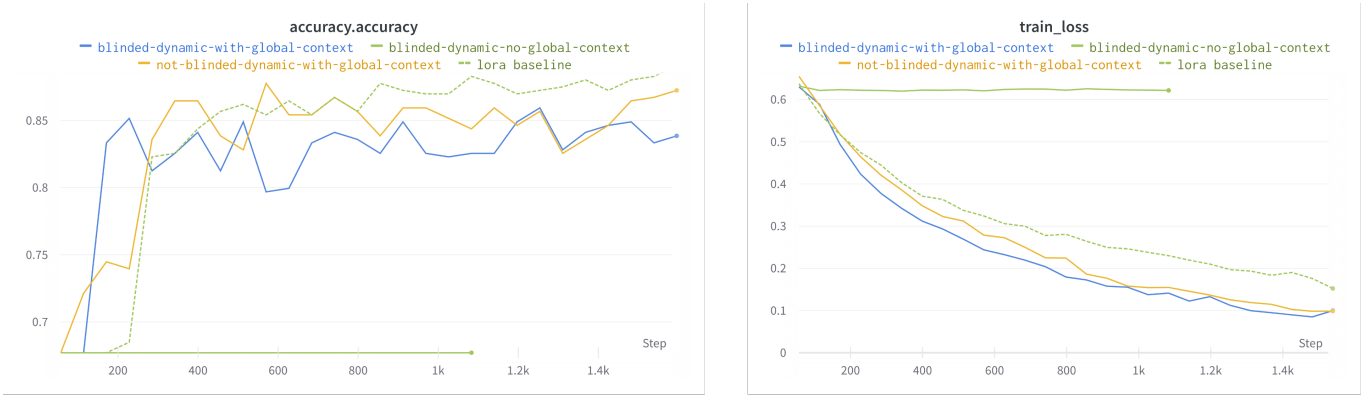

Figure 5: Illustration of efficiency of the global context based on Roberta LLM finetuning on MRPC task

but utilized by the LR state prediction model as temporary storage. LSTM and GRU memory cells can also be incorporated into the LR state prediction model architecture.

Learning, need for sleep and "hapiness". During continuous operation, the agent stores LR states together with global states in the global context. Assuming an operating frequency of 25 cycles per second, a global state buffer of approximately 1 million samples would be required to store continuous experiences from a 12-hour day. When the global context overflows, the excess samples are flushed into the replay buffer.

Once the global context and replay buffers are full, the system initiates training to reinforce behavior (LR state prediction) that leads to an increase in the value function and discourage behavior that decreases the value function. The value function can be likened to the abstract feeling of happiness. Thus, the agent's goal is to maximize its happiness. The more its state aligns with the state defined by the reward function, the happier it is.

While foundation models effectively capture domain-specific world representations, the AI agent must learn to interact with these representations to maximize its value function. Beyond direct world interaction, the agent can acquire experiences through simulation. Generative foundation models can be configured in autoregressive mode by connecting outputs to inputs as local context. During simulation sessions, the global context accumulates new experiences similarly to real-world interactions. Once the global context and replay buffer reach capacity, the agent learns from these simulated experiences through training.

The agent cannot simultaneously interact with the environment while undergoing training or simulation—it requires periods of offline processing analogous to sleep. The simulation and training phases correspond to REM (rapid eye movement) and NREM (non-rapid eye movement) sleep phases in humans. During REM sleep, neural mechanisms inhibit motor functions and minimize sensory processing. Similarly, the agent must deactivate actuators and input sensors during offline phases, enabling immersive simulation and experience extraction from world representations encoded in foundation models.

Curiosity. For efficient learning, the agent must be motivated to explore uncharted territory and possess effective means of identifying and capturing new knowledge.

At each step, the agent predicts the next step using the generative foundation models. The discrepancy between the prediction and the actual sensory input can serve as a marker of surprise, encouraging the memorization of unexpected experiences.

The agent's reward function can be formulated to provide rewards for positive surprise experiences, thereby stimulating curiosity. We refer to this as curiosity-driven exploration.

7 Conclusion

The path toward Artificial General Intelligence lies in the effective processing of temporal sequences. The proposed architecture provides a computationally efficient framework for coherent processing of sensory data streams from multiple sources.

Similar to the brain, the AI agent comprises well-defined functional blocks that specialize in processing information from specific domains (frozen foundation models). The architecture includes coordination blocks that facilitate communication between domain-specialized modules (LR state coupling network), blocks responsible for short-term memory (global context), long-term memory (DQN training on replay buffer), and unconscious procedural memory (local context). The proposed architecture naturally incorporates concepts analogous to biological processes:

periods of offline training (analogous to sleep), reward optimization (analogous to satisfaction), generative simulation in the absence of sensory inputs (analogous to imagination), and experience-based learning (DQN training on replay buffer). The reward and value functions can be shaped through natural language specifications, enabling an intuitive intelligence programming interface.

The internal LR state stored in the global context functions analogously to cognitive states—these "thought states" are not directly connected to outputs but rather modulate the behavior of domain-specific foundation models.

References

Hu, Edward J., Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. "LoRA: Low-Rank Adaptation of Large Language Models." arXiv. http://arxiv.org/abs/2106.09685.

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. "Scaling Laws for Neural Language Models." arXiv. http://arxiv.org/abs/2001.08361.

Kitaev, Nikita, Łukasz Kaiser, and Anselm Levskaya. 2020. "Reformer: The Efficient Transformer." arXiv. http://arxiv.org/abs/2001.04451.

Lialin, Vladislav, Vijeta Deshpande, and Anna Rumshisky. 2023. "Scaling Down to Scale Up: A Guide to Parameter-Efficient Fine-Tuning." arXiv. http://arxiv.org/abs/2303.15647.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." arXiv:1706.03762. arXiv. https://doi.org/10.48550/arXiv.1706.03762.

Zhu, Jie, Fengge Wu, and Junsuo Zhao. 2021. "An Overview of the Action Space for Deep Reinforcement Learning." In *2021 4th International Conference on Algorithms, Computing and Artificial Intelligence*, 1–10. Sanya China: ACM. https://doi.org/10.1145/3508546.3508598.