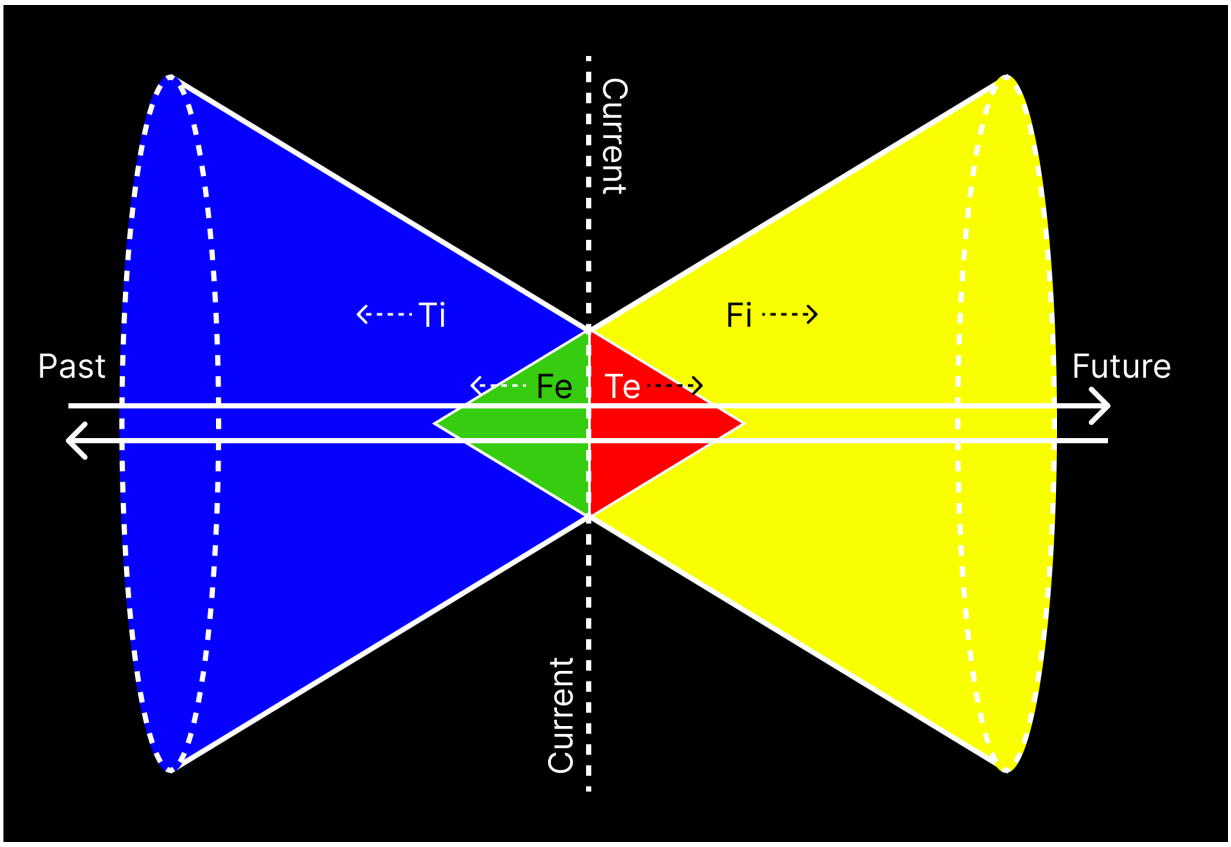
Homunculus Project

By Habibullah Akbar

> *"In alchemy, homunculus is about using various substances and rituals to create a miniature of human being, essentially an artificial person."*

This is the technical documentation of the Homunculus Project. The long-term technical plan is to build autonomous machine intelligence based on the combination of insight from many disciplines such as physics, mathematics, machine learning, neuroscience, psychoanalysis, and Jungian theory.



High-level visualization of perspectives of truth based on light cone theory and Jungian psychology

2004 Review of Light Cone Field Theory

This article on light cone field theory reviews the basic concepts of the Poincare group in light cone quantum mechanics and light cone field theory. It discusses recent progress in light cone...

https://arxiv.org/abs/hep-ph/0401248

⚠ Disclaimer: ⚠

- Yes. Jungian psychology is an old theory from a century ago that is often dismissed in popular psychology because it's not empirically validated. I would say, it's an advancement of the popular philosophical model of system 1 and system 2 thinking.

Quantum Teleportation and Jungian Psychology

We propose that the Jungian psychological type of an individual is naturally modelled as a quantum state: a maximally entangled two-qubit state, one of whose qubits is undergoing quantum teleportation.

https://arxiv.org/abs/2008.01079

- Our modelling is mostly based on in-depth reasoning and theoretical exploration. We took premises from existing research-validated concepts and built something from it.
- We are not only making something that computationally makes sense but also following philosophical sense and the law of physics.
- This modelling is only for the individual who wants to step into a completely new territory and perspective of consciousness and intelligence.
- This modelling is highly agnostic. Meaning, it's not related to any supernatural or spiritual sense.
- This documentation is distillation from a raw discussion of the whole system.

Main Concept

Our modelling consists of several parts and mechanisms:

- Multi-modal immediate sensory encoder (Se cognitive function)
- Attention-based long-term memory retrieval (Si cognitive function)
- Convergent intuitive generation (Ni cognitive function)
- Divergent imaginative generation (Ne cognitive function)
- Discriminative model (T cognitive function)
- Pain signal block (F cognitive function)
- Action decoder
- Short-term memory injection.
- Internal latent loop.
- Two-way chained reality (forward and backwards)
- Defining unknown as the middle value of judgement.
- Perceiving and judging function preference.
- Spectrum of cognitive preference.
- Snapshot of each perceiving function for each cognitive cycle
- Sleep and dreaming stage for daily iterative refinement

Here's a breakdown of the system's core elements and how they interact:

1. Perceiving the World:

- **Multimodal Sensory Encoders (Se):** The agent perceives its environment through multiple sensory modalities, including:
  - **Vision:** A high-resolution visual encoder (e.g., ViT, ConvNeXt†), potentially supplemented with an object detection model like YOLO for focused attention.
  - **Audio:** A sophisticated audio encoder (e.g., a modified Whisper model) capable of recognizing speech, sounds, and music.
  - **Other Senses (Future Potential):** The architecture allows for integrating additional senses, such as touch/tactical, smell, or proprioception, as technology advances.

The Platonic Representation Hypothesis

We argue that representations in AI models, particularly deep networks, are converging. First, we survey many examples of convergence in the literature: over time and across multiple domains, the...

https://arxiv.org/abs/2405.07987

- **Attention-Based Long-Term Memory Retrieval (Si):**
  - **Vector Database:** A vast vector database stores the agent's memories, each represented as a compressed embedding.
  - **Attention-Based Retrieval:** The agent can quickly retrieve relevant memories based on its current context or goals using an attention mechanism that identifies the most similar memories in the database.

Attention Mechanism in Neural Networks: Where it Comes and Where it Goes

A long time ago in the machine learning literature, the idea of incorporating a mechanism inspired by the human visual system into neural networks was introduced. This idea is named the attention...

https://arxiv.org/abs/2204.13154

ColPali: Efficient Document Retrieval with Vision Language Models

Documents are visually rich structures that convey information through text, as well as tables, figures, page layouts, or fonts. While modern document retrieval systems exhibit strong performance...

https://arxiv.org/abs/2407.01449

- **Intuitive Generation (Ni and Ne):**
  - **Convergent Generation (Ni):** This module generates focused predictions or insights based on the agent's current understanding of the world, goals, and observed patterns. It's like the agent's "gut feeling" or its ability to see the **most likely** outcome using an autoregressive model. This generative model can adapt to generate any modality/sensory data.

Forward and Reversed Time Prediction of Autoregressive Sequences on JSTOR

Stamatis Cambanis, Issa Fakhre-Zakeri, Forward and Reversed Time Prediction of Autoregressive Sequences, Journal of Applied Probability, Vol. 33, No. 4 (Dec., 1996), pp. 1053-1060

https://www.jstor.org/stable/3214985

  - **Divergent Generation (Ne):** This module explores a wider range of possibilities, brainstorming ideas, making unexpected connections, and generating creative or unconventional solutions using the diffusion model. This generative model can adapt to generate any modality/sensory data.

2. **The "Chained Reality" of Thought:**

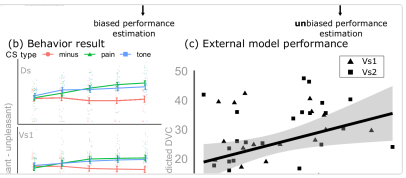
- **Two-Way Chains:** The agent constructs its understanding of the world through "chained realities":
  - **Backward Chain (Ti):** The agent reasons backwards in time, seeking to understand the *causes* of events or experiences. This chain is driven by Ti (Introverted Thinking), which seeks logical consistency and explanations.
  - **Forward Chain (Te):** The agent reasons forward in time, predicting the *effects* of its actions or imagining potential future scenarios. This chain is guided by Te (Extroverted Thinking), which prioritizes efficiency, goals, and control.
- **Dynamic Switching:** The agent can fluidly switch between these two orientations based on the context, its goals, or its current cognitive flow.

3. **Judgment and Evaluation:**

- **Discriminative Model (T):** This module evaluates the logical coherence of the agent's "chained realities," regardless of their modality or temporal orientation. It uses a confident layer trained by the loss function of the generative process,

simply a reward function. To assess how well the elements of the chain fit together logically.

- **Pain Signal Associative Evaluation Block (F):** This module assesses the emotional implications of the agent's experiences and guides its decisions based on a combination of:
  - **Pain Signals:** The agent experiences various forms of "pain" (e.g., hunger, sleepiness, uncertainty, cognitive overload) that act as negative reinforcement, motivating it to avoid harmful situations or to seek out rewards.



  - **Emotional Judgment:** The agent learns to associate specific contexts, actions, or thoughts with different levels of pain/reward signal, shaping its emotional responses and values. This emotional judgment shaping aesthetic preference, specific trauma, and art interpretation.
  - **Judgment directions:** Because the agents only collect negative signals, the value of judgment is determined by the directions (forward/backward). A negative signal can be evaluated as positive if it's for a backward chain (Negative first, more positive later) and neutral signals can be evaluated as positive if it's just after a negative chain.

4. **Defining unknown and uncertainty:**

- **Middle value as computational representation of unknown:** The middle value of confidence score or prediction is defining that the agent is not confident with its cognitive process, simply unknown. Near zero (false) or near one (true) scalar value is defining certainty. Uncertainty can trigger pain signal, so the agents maintaining internal motive to expand its own knowledge/experience.

5. **Cognitive preference:**

- **Perceiving and judging (speed vs depth):** The agents can develop preference over generating/gathering information (perceiving) and validating the chained reality (judging). The agents with perceiving function preference will have bigger toleration over uncertainty/unknown and use judgement function in a single pass as guide. Meanwhile the agents with judging function preference will seek for definitive answer of judgement, doing multiple pass of judgement function to avoid the middle value of judgment confidence. Slower cognitive rate per cycle compared to perceivers type. Agents with perceiving preference advantage are speed, meanwhile agents with judging preference advantage are depth.
- **One function at a time:** Agents can only use one function at a time. Similar to how human focus on one cognitive function at a time. The likelihood of function usage shapes the agent's cognitive preference. There's no strict typing like 16 types of Jung cognitive function. But each agent can lean into one type like a spectrum.

Attentional Limitations in Doing Two Tasks at Once: The Search for Exceptions - Mei-Ching Lien, Eric Ruthruff, James C. Johnston, 2006

People generally have difficulty doing two tasks at once. To explain this fact, theorists have proposed that central processing—the thought-like stages followin...

https://journals.sagepub.com/doi/10.1111/j.0963-7214.2006.00413.x

6. **Action Decoder:**

- **Translating Thought into Action:** The action decoder translates the agent's internal state (its chained reality, its judgments, and its emotional evaluations) into tangible actions.
- **Specialized Modules:** It consists of separate decoders for:
  - **Motoric Movements:** Controlling the agent's body or actuators.
  - **Speech Outputs:** Generating speech that reflects the agent's thoughts and emotions.

7. **Short-term memory injection:**

- **Embedding as compressed information:** Reuse the embedding from the last layer as compressed information for the new sequence (working memory). Every time the working memory hits the maximum hardware limit or changes direction (forward/backwards), we reuse the last embedding/tensor from some tokens as compressed information. Maintaining the whole information by compressing earlier sequences.

Squid: Long Context as a New Modality for Energy-Efficient...

This paper presents Dolphin, a novel decoder-decoder architecture for energy-efficient processing of long contexts in language models. Our approach addresses the significant energy consumption and...

https://arxiv.org/abs/2408.15518

LLM2Vec: Large Language Models Are Secretly Powerful Text Encoders

Large decoder-only language models (LLMs) are the state-of-the-art models on most of today's NLP tasks and benchmarks. Yet, the community is only slowly adopting these models for text embedding...

https://arxiv.org/abs/2404.05961

8. **Internal Latent Loops:**

- **Latent loop:** Feeding back the entrance model with final embedding/tensor. Use the same parameter multiple times to achieve more compute/test time while inference.

TroL: Traversal of Layers for Large Language and Vision Models

Large language and vision models (LLVMs) have been driven by the generalization power of large language models (LLMs) and the advent of visual instruction tuning. Along with scaling them up...

https://arxiv.org/abs/2406.12246

Scaling LLM Test-Time Compute Optimally can be More Effective than...

Enabling LLMs to improve their outputs by using more test-time computation is a critical step towards building generally self-improving agents that can operate on open-ended natural language. In...

https://arxiv.org/abs/2408.03314

Zamba: A Compact 7B SSM Hybrid Model

In this technical report, we present Zamba, a novel 7B SSM-transformer hybrid model which achieves competitive performance against leading open-weight models at a comparable scale. Zamba is...

https://arxiv.org/abs/2405.16712

- **Refining Understanding:** Each fixed-size block can engage in internal latent loops, revisiting and refining its predictions or insights based on feedback from the judgment functions.
- **Simulating "Slow Thinking":** These loops allow for a more deliberate and iterative form of reasoning, mimicking the human capacity for reflection, analysis, and the gradual refinement of understanding.

9. **Snapshot Evaluation and Cognitive Flow:**

- **Efficient Prioritization:** At the start of each cognitive cycle, the agent collects "snapshots" or "teasers" of information from each of its perceiving functions (Se, Si, Ne, Ni) and the action decoder.
- **Layerskip and sparsity:** Each perceiving function can be executed in highly sparse mode to get the snapshot quicker before the actual generation/information gathering.

Nerva: a Truly Sparse Implementation of Neural Networks

We introduce Nerva, a fast neural network library under development in C++. It supports sparsity by using the sparse matrix operations of Intel's Math Kernel Library (MKL), which eliminates the...

https://arxiv.org/abs/2407.17437v1

LayerSkip: Enabling Early Exit Inference and Self-Speculative Decoding

We present LayerSkip, an end-to-end solution to speed-up inference of large language models (LLMs). First, during training we apply layer dropout, with low dropout rates for earlier layers and...

https://arxiv.org/abs/2404.16710

- **Judgment-Guided Selection:** The judgment functions (Ti or Te) evaluate these snapshots and prioritize the most relevant or promising source of information for further processing. This creates a dynamic and adaptive cognitive flow, allowing the agent to focus its resources on the most important aspects of its experience.

10. **Sleep and Dreaming Stage:**

- **Offline Refinement:** During a "sleep" state, the agent engages in offline refinement of its cognitive functions, including:
  - **Memory Consolidation:** Integrating new experiences into long-term memory and strengthening important associations, simply fine-tuning with new collected data. Exception for the long-term memory module (Si function), the process is indexing the sensory data into quick accessed index.
  - **Module Fine-tuning:** Each module is refined independently to prevent instability and cognitive resource overload (training a module one by one needs smaller resources than training the whole cognitive flow at once).
  - **Architectural Plasticity:** Potentially growing or pruning layers within modules to optimize for efficiency and adaptability.

Mixture of A Million Experts

The feedforward (FFW) layers in standard transformer architectures incur a linear increase in computational costs and activation memory as the hidden layer width grows. Sparse mixture-of-experts...

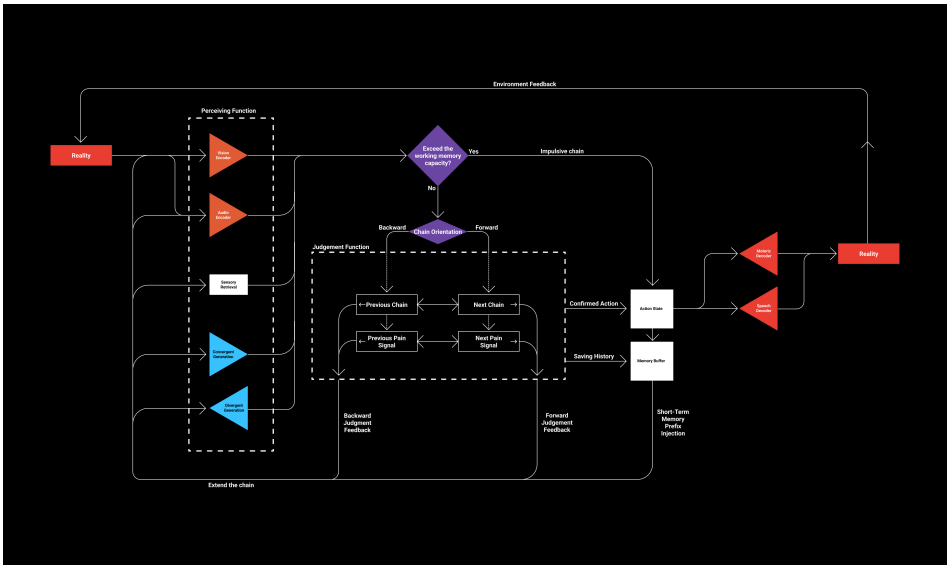
https://arxiv.org/abs/2407.04153v1

FinerCut: Finer-grained Interpretable Layer Pruning for Large...

Overparametrized transformer networks are the state-of-the-art architecture for Large Language Models (LLMs). However, such models contain billions of parameters making large compute a necessity,...

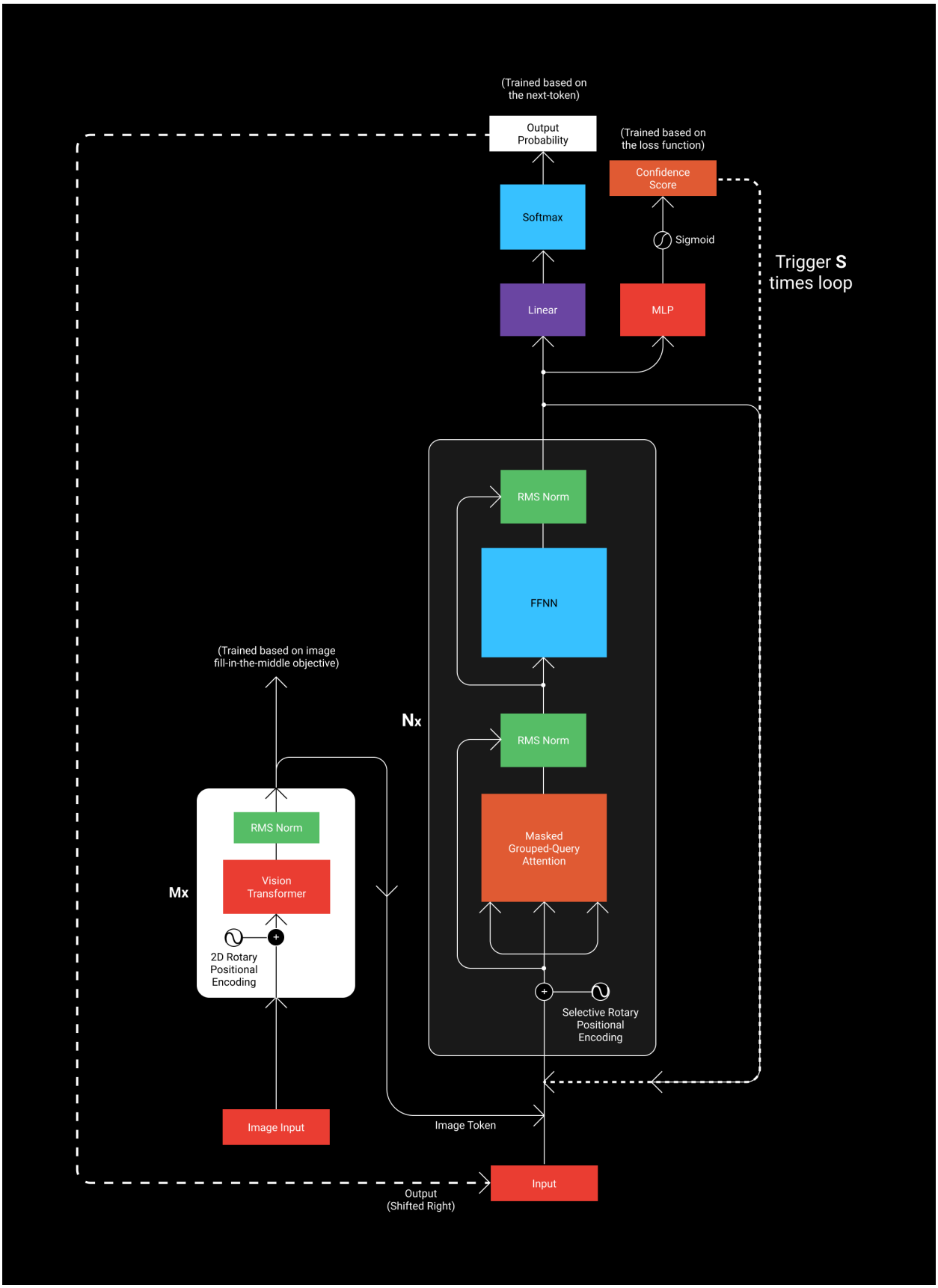
https://arxiv.org/abs/2405.18218

- **Working memory reset:** After sleep session, the short-term memory of the agents is flushed to start a new working memory sequence.



Detailed flow of the architecture

> “Life can only be understood backwards, but it must be lived forwards.”
> —Soren Kierkegaard


(Trained based on the next-token)
Output Probability
(Trained based on the loss function)
Confidence Score
Softmax
Sigmoid
Trigger S times loop
Linear
MLP
RMS Norm
FFNN
RMS Norm
Masked Grouped-Query Attention
Nx
Selective Rotary Positional Encoding
(Trained based on image fill-in-the-middle objective)
RMS Norm
Vision Transformer
Mx
2D Rotary Positional Encoding
Image Input
Image Token
Input
Output (Shifted Right)

Flowchart of the code implementation

Code Implementation:

This is frequently updated segment to apply iterative development based on the latest software/hardware ecosystem.

The entire codebase is maintained in GitHub:

```
https://github.com/akbar2habibullah/Homunculus-Project
```

Code implementation

> **Disclaimer for license:** ***This software implementation is dual-licensed under the terms of the GNU Affero General Public License (AGPL) and a commercial license. For commercial use, please contact Habibullah Akbar at akbar2habibullah.gmail to obtain a commercial license. Commercial use is defined as any use of the software for financial gain, including but not limited to, selling, licensing, or distributing the software as part of a product or service.***