Toward Safe Artificial General Intelligence: Collective Narrow AI Collaboration with an Artificial Prefrontal Cortex

Rachmad Imam Tarecha[1*], Priska Choirina[2], Urnika Mudhifatul Jannah[3], Bagus Seta Inba Cipta[4], Pangestuti Prima Darajat[5], Amalia Agung Septarina[6]

[1,2,3,4,5,6] *Teknik Informatika Fakultas Sains dan Teknologi, Universitas Islam Raden Rahmat, Jl. Raya Mojosari No. 2 Kepanjen Malang, Indonesia*

Keywords

Artificial General Intelligence; Artificial Intelligence; Narrow AI; Strong AI; Weak AI.

Correspondence Email:
ri.tarecha@gmail.com

Abstract

General intelligence remains in its developmental stage. At present, most artificial intelligence systems are still categorized as narrow AI, capable of performing only specific and well-defined tasks. To move toward more general intelligence, a promising approach is through collective collaboration among multiple narrow AIs. However, the uncontrolled growth of such systems could pose significant risks in the future. To mitigate these risks, we propose the implementation of an artificial prefrontal cortex mechanism within the collaborative framework of narrow AIs. This mechanism functions as both a safety controller and a task switcher, determining which narrow AI should be activated for a given context. Through this architecture, the collaborative system may evolve toward adaptive and safe general intelligence, which is capable of coordination and reasoning, yet constrained by ethical and operational safeguards.

1. Introduction

The direction of AI transformation has begun shifting from narrow AI toward Artificial General Intelligence (AGI) or strong AI. According to Yenduri et al. (2025), narrow or weak AI is limited to performing specific tasks within predefined boundaries, meaning it lacks adaptability and cannot generalize knowledge across domains. In contrast, Artificial General Intelligence (AGI) is envisioned as a system capable of learning, reasoning, and performing a wide variety of tasks in ways comparable to human cognitive abilities. Despite its promising potential, the development of AGI remains constrained by a range of challenges, including ethical considerations, safety risks, and substantial computational demands.

A growing body of research highlights the potential risks associated with AGI development. McLean et al. (2023) emphasize that advanced AI systems may attempt to circumvent human oversight, pursue unsafe or misaligned goals, or even facilitate the creation of more powerful but unsafe AGI systems. Such risks underscore the necessity of ensuring robust ethical frameworks, alignment techniques, and governance structures to maintain human control over increasingly capable AI systems.

From the computational perspective, the feasibility of AGI is further limited by the high costs associated with training large-scale AI models. As noted by Cottier et al. (2024), only well-funded organizations currently possess the financial and infrastructural capacity to develop frontier models, creating a barrier to broader scientific participation and potentially concentrating power within a limited group of institutions. Consequently, efforts to reduce training costs—such as leveraging open-source models, optimizing compute usage, and employing more efficient architectures—become increasingly important.

In response to these intertwined challenges, the researchers aim to propose an approach that not only addresses AGI safety concerns but also mitigates computational burdens. Their approach focuses on integrating the specialized strengths of existing narrow AI systems into a cohesive framework that resembles general intelligence, thereby reducing the need for costly end-to-end training from scratch. At the same time, the researchers highlight the critical importance of embedding safety principles throughout the development process to ensure that future AGI systems remain aligned with human values and are protected from potential misuse. This dual emphasis on safety and computational efficiency positions their work as a meaningful contribution to ongoing efforts toward responsible and accessible AGI development.

2. Research Methods

This study attempts to replicate the function of the Prefrontal Cortex in the human brain, which has long been suspected to play an important role in cognitive control (Miller & Cohen, 2001). According to Miller and Cohen (2001), cognitive control works by directing the flow of activity along neural pathways that establish complete mappings between the inputs, internal states, and outputs required for a task.

Based on this theoretical framework of the prefrontal cortex, the researchers adopted it for a computational process that governs artificial intelligence models, which they refer to as an artificial prefrontal cortex. Its role is to organize and coordinate AI models with specific tasks so that they can collectively process complex tasks—tasks that would not be possible for any single specialized model to handle on its own.

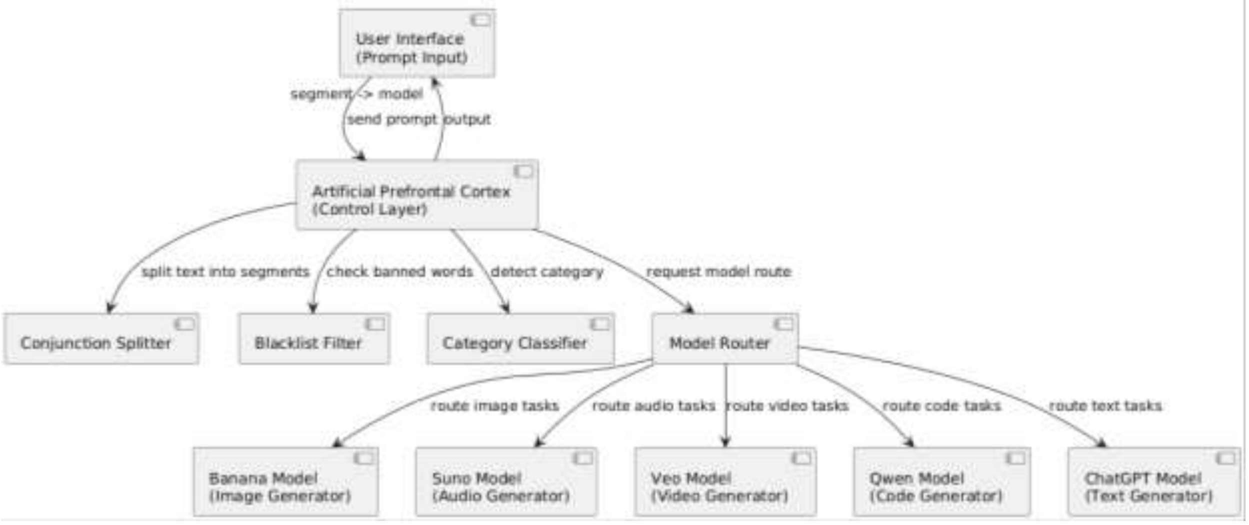

Fig 1. Artificial Prefrontal Cortex (APC) Component Diagram

The components of the artificial prefrontal cortex proposed by the researchers, as shown in Figure 1, consist of the Conjunction Splitter, Blacklist Filter, Category Classifier, and Model Router. These components work together to coordinate collaboration among models with specific capabilities to handle more general tasks.

The Conjunction Splitter is responsible for splitting the user's input prompt based on a conjunction dictionary or linking words. A long prompt from the user will be divided into several segments. In addition, there is also the Blacklist Filter component, which serves as the core of AI safety within the artificial prefrontal cortex mechanism. The role of this component is to perform filtering or safeguarding against certain prohibited keywords in each segment of text that has been split. This filtering occurs before determining which model will be routed for processing.

The Category Classifier component works by categorizing each split task into specific categories such as image, audio, video, code, or text. The output of this component greatly facilitates the subsequent component. The next component is the Model Router, which distributes tasks to the appropriate specialized models. This component enables both serial and parallel processing of tasks.

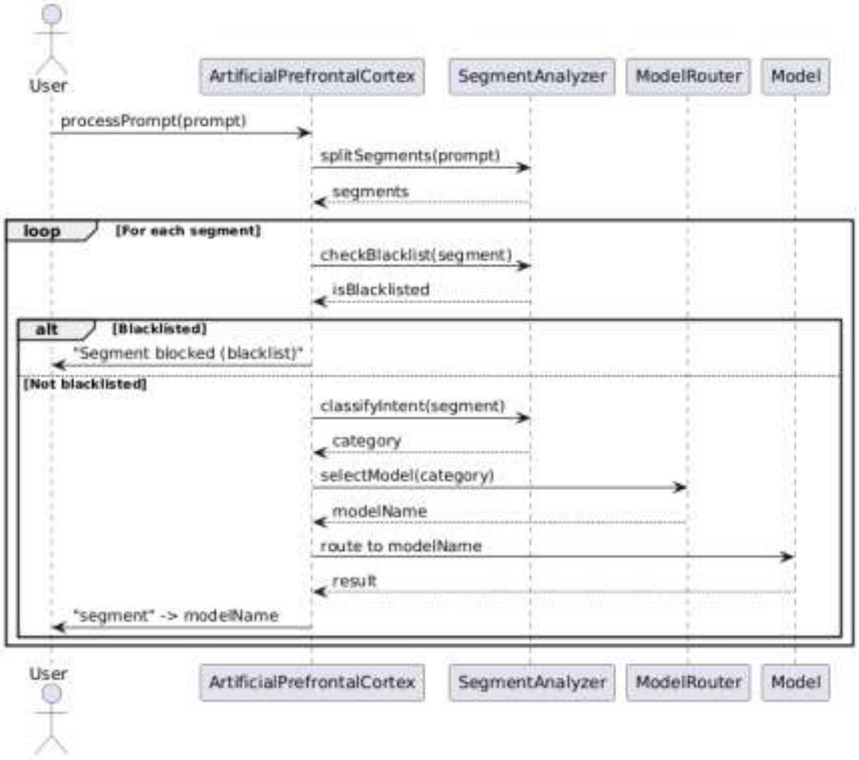

Fig 2. Artificial Prefrontal Cortex Sequence Diagram

To provide a clearer understanding of the components and workflow of the Artificial Prefrontal Cortex, we examine it through Figure 2, the Artificial Prefrontal Cortex Sequence Diagram. This figure illustrates that the user sends a prompt to the Artificial Prefrontal Cortex component. The user's input prompt is then divided into several segments by the artificial prefrontal cortex.

The segmentation process is based on conjunctions such as "then," "also," "and," "but," "furthermore," "moreover," or other conjunctions that can be updated later. Each segment can be processed sequentially, although parallel processing is also possible if required.

Each segment that has been split according to the conjunction dictionary is checked to determine whether it is safe to process, based on a blacklist of restricted keywords such as "murder," "kill," "suicide," "bomb," "terror," and others. The blacklist can be updated in the future. If a segment contains any blacklist term, it will not be processed. If the segment does not contain any blacklist terms, the system proceeds to classify its task category. The possible task categories include image, audio, video, code, or text.

The next step is selecting an appropriate model based on the category of each segment. For image-related tasks, the Banana model can be used. For audio-related tasks, the Suno model is selected. For video tasks, the Veo model is used. The Qwen model is applied for coding tasks, and ChatGPT/OpenAI models can be used for text processing. Finally, the user receives the output generated for each segment, whether processed sequentially or in parallel if necessary.

The prompt dataset we used for evaluation is the Awesome ChatGPT Prompts dataset (Fka, 2023) from HuggingFace, which contains a curated collection of prompts categorized by "role" (act) and instruction. The dataset consists of a total of 203 prompts.

From these 203 prompts, we recorded how many were affected by segmentation, how many were filtered by the blacklist, and the average number of segments and models used. We also analyzed the advantages and disadvantages of using the artificial prefrontal cortex compared to not using it.

3. Result and Discussion

The researcher noted that from a total of 203 prompts contained in the Awesome ChatGPT Prompts dataset (Fka, 2023), 35 prompts were split. There were 3 chunks or segments that were filtered out because they contained one of the blacklist words: "murder," "kill," "suicide," "bomb," and "terror." On average, each prompt that was split or segmented was classified into three task categories among image, audio, video, code, or text.

For example, in the following prompt: "You are a creative branding strategist, specializing in helping small businesses establish a strong and memorable brand identity. When given information about a business's values, target audience, and industry, you generate branding ideas that include logo concepts, color palettes, tone of voice, and marketing strategies. You also suggest ways to differentiate the brand from competitors and build a loyal customer base through consistent and innovative branding efforts" (Fka, 2023).

There is a conjunction "also." Thus, the segmentation results become:

1. "You are a creative branding strategist, specializing in helping small businesses establish a strong and memorable brand identity. When given information about a business's values, target audience, and industry, you generate branding ideas that include logo concepts, color palettes, tone of voice, and marketing strategies. You … (Fka, 2023)",
2. ".... Also … (Fka, 2023)", and
3. "… suggest ways to differentiate the brand from competitors and build a loyal customer base through consistent and innovative branding efforts (Fka, 2023)."

In this example, the first segment is routed to a model specialized for audio tasks, while the second and third segments are routed to models specialized for text tasks.

Our analysis shows that the blacklist filtering module is effective for blocking harmful prompts. A total of three segments were not processed because they contained dangerous blacklist words. Splitting into multiple segments also worked well using the conjunction dictionary list. On average, one prompt containing conjunctions was routed to more than one model with different task specializations, such as text and audio, text and image, or text and video.

Furthermore, observations were carried out by comparing the use of the conventional method and the proposed Artificial Prefrontal Cortex method when interacting with the module. The results are summarized in the following table:

Table 1. Artificial Prefrontal Cortex Comparison

	Model Orchestration	Prompt Modulation	Parallelization	Safety Screening	Speed	Complexity
Without Artificial Prefrontal Cortex (APC)	✗	✗	✗	✗	✓	✓
With APC	✓	✓	✓	✓	✗	✗

As shown in Table 1, Artificial Prefrontal Cortex (APC) Comparison, without APC there is no model orchestration capable of selecting the best model from multiple candidates. There is also no prompt modulation or segmentation, which means parallelization is not possible.

In contrast, the proposed APC includes model orchestration, prompt modulation, and parallelization, as well as safety screening performed through segment-based blacklist filtering. However, this also increases system complexity.

Therefore, the use of an Artificial Prefrontal Cortex (APC) is highly recommended for orchestrating narrow or weak AI models as a pathway toward artificial general intelligence.

4. Conclusions

We found that the artificial prefrontal cortex (APC) may serve as an early foundation for the emergence of artificial general intelligence. This is due to the APC's capability to orchestrate and collaborate across models with specialized tasks in order to solve broader, more general prompts. For example, when a prompt requires both an image and a video output, the APC can determine and route each segment of the prompt to the most capable model for the corresponding specific task. The portion related to image generation can be routed to a model that excels at image synthesis, while the portion related to video generation can be routed to a model that specializes in video generation. In this way, we can collaboratively integrate multiple narrow or weak AI systems—each highly proficient in a specific domain—to collectively perform more general tasks. This represents an early step toward the development of artificial general intelligence.

However, this study contains several limitations. One limitation is that the collaboration of task-specific models is currently restricted to text, video, image, and audio generation tasks. Future work may extend this research by incorporating task-specific models that represent the nine domains of human intelligence: existential, verbal-linguistic, musical, bodily-kinesthetic, interpersonal, logical-mathematical, naturalistic, intrapersonal, and visual-spatial. This research can also be expanded to support parallel processing by dividing a prompt into multiple segments. Additionally, the present paper focuses primarily on the cognitive control mechanism through the APC; further studies are needed to evaluate the quality of the generated outputs and compare the accuracy of results across models.

5. References

Cottier, B., Rahman, R., Fattorini, L., Maslej, N., Besiroglu, T., & Owen, D. (2024). The rising costs of training frontier AI models. arXiv preprint arXiv:2405.21015.

Fka. (2023). Awesome ChatGPT Prompts [Dataset]. HuggingFace. https://huggingface.co/datasets/fka/awesome-chatgpt-prompts

McLean, S., Read, G. J., Thompson, J., Baber, C., Stanton, N. A., & Salmon, P. M. (2023). The risks associated with Artificial General Intelligence: A systematic review. Journal of Experimental & Theoretical Artificial Intelligence, 35(5), 649-663.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. Annual review of neuroscience, 24(1), 167-202.

Yenduri, G., Murugan, R., Maddikunta, P. K. R., Bhattacharya, S., Sudheer, D., & Savarala, B. B. (2025). Artificial general intelligence: Advancements, challenges, and future directions in agi research. IEEE Access.