
Expressibility Is All You Need

FuJia Zhang

Shanghai Guanghua Cambridge International School

zfrankj0927@gmail.com

Abstract

Large language models (LLMs) have shown impressive capabilities in natural language processing but struggle with logical reasoning and complex problem-solving. This thesis proposes a novel modular approach to enhance LLM capabilities by integrating specialized cognitive models with an LLM serving as the central coordinator. We evaluate this approach through computational complexity analysis and performance comparisons. Results show our modular system achieves 33.3% accuracy on a challenging test set of logic, mathematics, and problem-solving tasks, outperforming the base LLM (20.0%) and some advanced models like GPT-4o (26.7%). Theoretical analysis indicates a reduction in computational complexity from $O(n^2)$ for monolithic models to $O(n \cdot \log(n))$ for our approach, suggesting improved scalability. This research represents a promising direction in developing more capable AI systems.

1 Introduction

Large language models (LLMs) have recently demonstrated impressive capabilities in natural language processing and generation. However, these models exhibit significant limitations in tasks requiring advanced logical reasoning and problem-solving skills (Marcus & Davis, 2020;

Bender et al., 2021), especially when complex tasks need multistep reasoning. This thesis proposes a novel architecture to address these limitations by integrating specialized cognitive models with a central LLM. Our approach distributes cognitive functions across specialized modules, enabling dynamic interactions that mimic human cognitive processes. It improves reasoning capabilities, enhances adaptability, increases interpretability, and uses computational resources more efficiently. The design, inspired by cognitive theories and neuroscientific insights, breaks down complex tasks into specialized cognitive processes. Our current results, based on theoretical simulations and hypothetical case studies, provide a foundation for future empirical testing. While promising for developing more sophisticated AI systems, we remain cautious about claiming direct progress towards artificial general intelligence. Future work will focus on practical implementation and exploring this system's potential impact on developing more capable and efficient intelligent systems.

2 Background

The traditional design of LLMs primarily focuses on scaling up the model's size to improve its ability to understand and generate language. This method, while effective up to a point, does not address the model's inherent deficiencies in specialized reasoning and creativity. Our approach deviates from this tradition by introducing a modular system where the LLM serves only as the core for natural language understanding and generation, while specialized modules handle other cognitive tasks. This structure allows for a more efficient flow of information and task-specific processing, reminiscent of human brain functionality where different areas are responsible for different cognitive functions.

In this modular system, interactions can occur in two primary ways. Firstly, textual inputs by users are processed by the LLM, which then coordinates with other modules to formulate responses. Secondly, in scenarios requiring more dynamic interaction, such as through visual or auditory inputs, the system leverages its modular components for perception and response, engaging in what can be described as an embodied AI experience. This dual interaction capability not only enhances the system's responsiveness but also enables it to operate in a manner that closely mimics human interaction with the environment.

The essence of this system lies in its ability to not just process language but to understand and generate responses that require a synthesis of various cognitive abilities, thereby significantly advancing the field of AI towards more generalized capabilities.

3 Model Architecture

Most LLMs are now using the scaling law to increase its size whilst increase its ability. Here, the structure changed into a modular system. The LLM is only used for its natural language understanding and generation and all the other models are centred on this. There are two ways for the whole system to give output. The first one is that the users give information through text, in this condition LLM allow the information to flow from one module to another and integrate the information to give reply to the users. On the other hand, the embodied AI is an important part to realize the general human level AI, so if the users interact information by imaging or speaking, the LLM will also use its sensitivity to collect information then call other modules to think and then use their body to reply through the activity module.

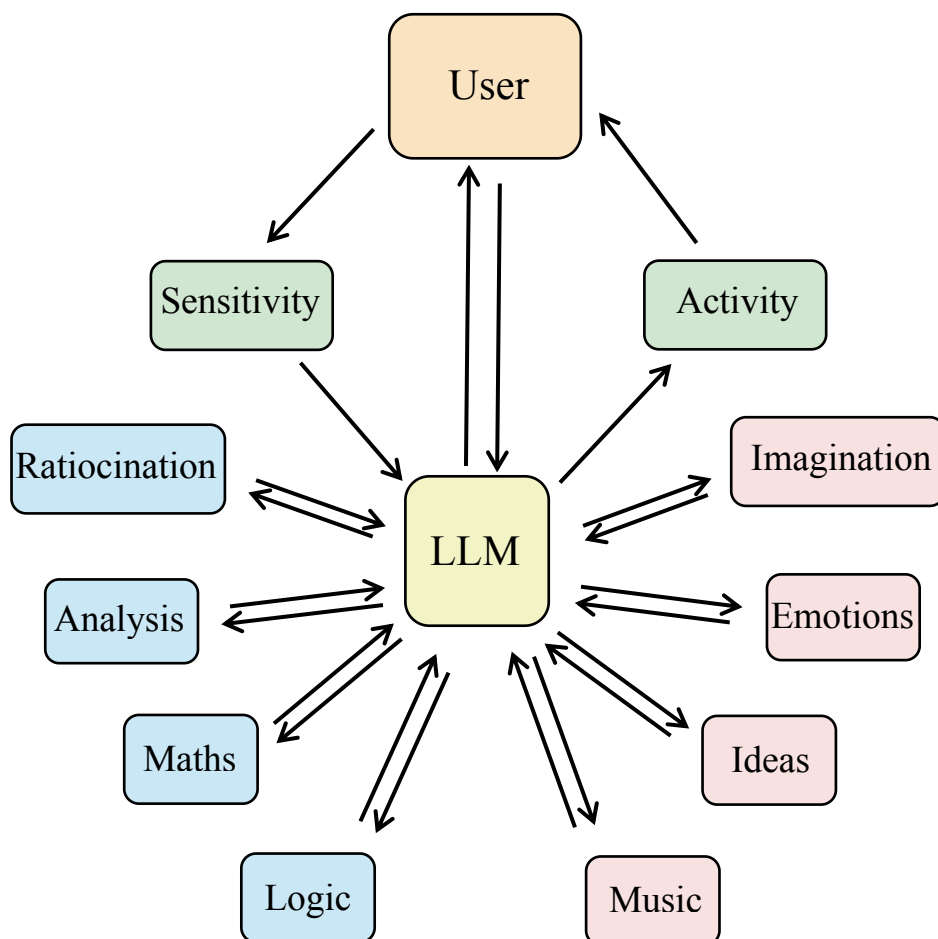


Figure 1: The system structure - model architecture

The system follows this overall structure shown in Figure 1 using both LLM and specialized modules to interact with users by different ways which simulates the human's brain and the models coloured in blue are representing the abilities having in the left cerebral hemisphere, also the ones coloured in red show the abilities in the right cerebral hemisphere.

3.1 Modular System

This centralized system ensure the logic and the thinking whilst the efficiency of the computing power. Due to the structure of it, the information would flow from one to another, forced the models to have a sort of thinking progress which would be relatively clear compared to other LLMs which are completely black box models. Further more, because of the validation loss of specialized models could be descend faster and deeper which means that the overall system would be trained faster and become more accurate. In terms of the computational resources needed for this hybrid system, the algorithm complexity would be

$$O(n \times \log n)$$

rather than the monolithic model's algorithm complexity which is

$$O(n^2)$$

this could lead to a great decrease on the pressure of the computational resources especially when the GPUs are expensive. In addition, this system made the size of every model become smaller, the LLM only needs to call the specialized models to help. What's more, the LLM doesn't need much ability, besides its understanding of natural language and generating text, it only needs to sort out the tasks and break them into smaller tasks then put them into the right module. A prompted llama 3.1 8B instruct is proper for the need of this. So the whole model maybe less than 1000B which would have a enough ability towards the general human level AI, including two ways to contact with the users: text interaction and the body interaction.

3.2 Specialized Models

The specialized model is actually another transformer model and it

focused on its own specific field, but it has a key difference between LLM, that is the prediction, LLM predict the next word and its smallest part of prediction is a word, but the specialized model's smallest prediction part is not a word, it is the smallest utility part of this specific field which could be expressed by the natural language which could be understood by the LLM and LLM could contact with it. Take the Maths module as an example, the smallest prediction part is the mathematical knowledge points, it could only need to predict out the thinking process such as the sequence of the uses of the mathematical knowledge points or more advancedly it might be able to predict out the answer of a specific task totally related to mathematics. What the modules need to do could be expressed by the natural language. That's the crucial point to achieve the general human level AI and build this system, the expressibility. Any human beings' task could be expressed by the natural language, so the specialized modules only need to predict the words that needed to be expressed in their fields, so the relation of the word-base between human beings, LLM and other specialized models would be like this:

general words \subset specialized models \subset LLM \subseteq human beings
 This would cause the specialized models need some training data different from normal and they could be trained faster. Also, with adequate amount of specialized models in Figure 1, the system will have the ability in every part of the human tasks and finally reach the goal of general human level AI. Finally, the specialized models' fields are chosen in the Figure 1 because it represents the abilities in the left and right cerebral hemisphere and sensitivity and activity modules are chosen due to the need of interaction between users and system through not only texts but also all the other methods that human beings could do to communicate with each others.

3.3 Integration of Modules and LLM

The integration of this system is mainly following this process. Firstly, the user would use either text or other form including gestures and vocal sound to make an input. Then the system would use the sensitivity module if possible to transform all the information into text and LLM would receive the information. After that, LLM will process it into tasks needed to be done and put these new information into corresponding modules to help complete the tasks. This procedure could be called as the first extract of abstract feature. After it, the specialized models would use its own ability to solve their tasks and provide answers back to LLM. This would be the second extract of abstract feature. Finally, LLM would gather all the information from the modules to get to the final reply and

contact with users through text or use the activity module to react by body language if needed and this could be called as the third and final extract of abstract feature. With the three times of extracting abstract feature, the centralized system could answer in a higher accuracy whilst less need of the computing power and more interpretable with more logic. Because LLM could understand other modules meaning and could use its basic ability to solve out the problem with the assists of the modules, the above system would be operating theoretically.

For example, the user ask a question through text: “Alice has N brothers and M sisters, tell me how many sisters does Alice’s brother have? Use algebraic expression which contains M and N.” In practice, I use a number of models llama 3.1 8B instruct to act as the specialized models and LLM by prompts, LLM called Ratiocination, Analysis, Logic, Maths and Imagination to help, then gather the information to get the right answer of $M+1$, the exact process is shown in Figure 2. The whole process needs 12.17 seconds. When using llama 3.1 sonar large 128k online, the model responds the correct answer with only 3.99 seconds. For comparison, the normal llama 3.1 8B and 70B instruct models only use 0.28 and 0.78 second to get the wrong answer of M. This is a true representation showing the potential of this system meaning small models could reach a higher ability.

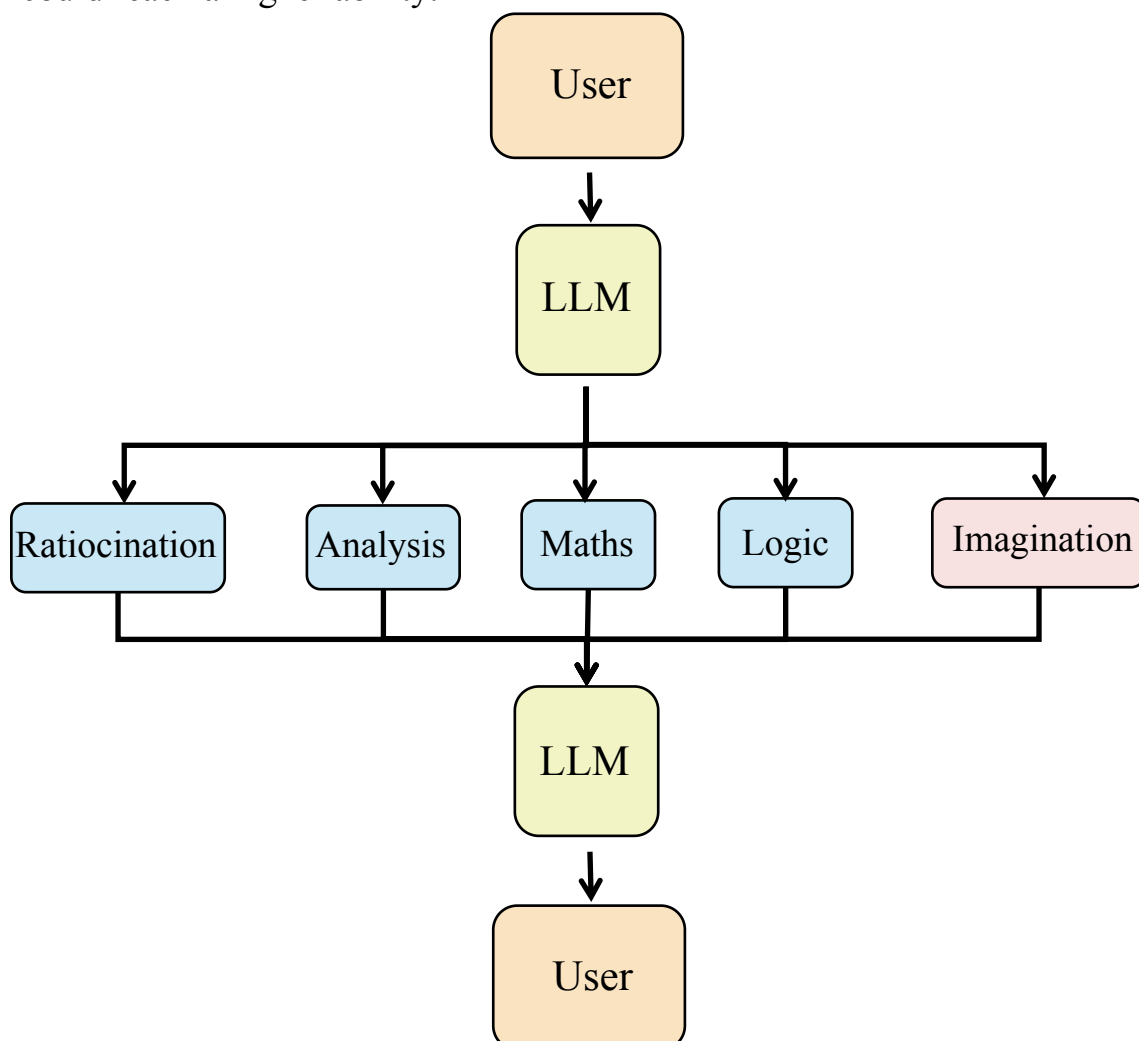


Figure 1: The information flow - solving an example problem

4 Why predicting other words

In earlier sections, I have discussed the concept of predicting specialized units instead of conventional word-by-word generation. This section will delve deeper into the practical implementation and benefits of this approach within our modular system. Each specialized model predicts specific units relevant to its field, fundamentally shifting how information is processed and offering unique advantages in terms of specialization, efficiency and the overall capability of the system.

4.1 Specialized Unit Prediction

In the modular architecture, each specialized model operates on domain-specific predictive units rather than words. These units are the smallest meaningful components relevant to the specific domain of each module. For example:

- The Maths module predicts mathematical operations or logical steps, such as differentiating a function or applying an algebraic identity.
- The Logic module predicts rules of logical reasoning, such as *modus ponens* or *contraposition*, which are applied sequentially to solve a problem.
- The Imagination module predicts conceptual transformations, such as how an abstract concept might evolve or associate items with information and what would happen next.

By focusing on these specialized units, each module contributes its expertise to the task at hand without being constrained by the need to generate coherent text. The LLM acts as the integrator, translating these specialized predictions back into natural language when interacting with users.

4.2 Efficiency Gains and Faster Learning

Specialized unit prediction also offers substantial efficiency gains. Since each module is focused on a specific domain, the complexity of training

is reduced compared to an LLM that must generalize across many different types of knowledge. Each specialized model can be trained independently using domain-specific data, which:

Reduces Training Time: Specialized models need less data to reach proficiency in their domain, as they do not need to learn the nuances of general language understanding.

Decreases Validation Loss More Quickly: By limiting the scope to specific predictive units, specialized models can achieve lower validation loss faster, leading to more accurate and reliable outputs.

For instance, the Maths Module only needs to learn mathematical concepts and their applications, which allows it to focus on tasks like identifying the correct sequence of mathematical operations without the distraction of linguistic features. This focused training not only makes the module more efficient but also enhances its capability to perform complex calculations that a general LLM might struggle with.

4.3 Practical Example: Solving a Math Problem

Consider a user asking: “ $x^2+2x+4=0$, $x^3=?$ (do not use the imaginary number)” In a conventional LLM, the model would predict the next word step by step, often leading to errors if the logical structure is not properly understood.

In the modular system, the following process occurs:

1. The LLM initially parses the question and identifies that it requires logical reasoning and mathematical operations.
2. It delegates the problem to the Logic module to break down the question: identifying the solution of x in the equation, quadratic equation, and the algebraic requirements.
3. The Maths module then takes over, predicting the sequence of operations needed to express the answer algebraically. It determines that the answer involves calculating the exact value of x , resulting in the final answer.
4. Then, the LLM collects the results from the specialized modules and find out that the process involving imaginary numbers, so it would call the Ideas module for help.
5. The Ideas module would give an advice of using substitution to solve this problem.
6. After that, the LLM would call the Maths module to help use this method.
7. Finally, the Maths module gives the procedure and the answer of the

problem using substitution and the LLM would gather the information to give the response: “ $x^3=8$ ”

In this example, the Logic and Maths Modules predict specific units such as logical relationships and mathematical steps rather than individual words. This approach not only increases accuracy but also ensures that the reasoning behind the answer is consistent and interpretable.

4.4 Specialized Unit Prediction vs. Word Prediction

The fundamental difference between predicting specialized units and predicting words lies in the level of abstraction. Word prediction focuses on language fluency, while specialized unit prediction focuses on domain-specific accuracy and structured reasoning. This modular approach:

- **Increases Specialization:** Each module becomes highly proficient in its domain, ensuring better performance in specialized tasks.
- **Enhances Interpretability:** Since each module's predictions are aligned with specific cognitive functions (e.g., logic or maths), the system's decision-making process becomes more transparent compared to traditional black-box LLMs.
- **Reduces Computational Overhead:** By delegating specialized tasks to smaller, focused models, the overall computational load on the LLM is reduced, making the system more efficient.

In summary, predicting specialized units instead of words allows the modular system to achieve higher accuracy, faster learning, and greater interpretability. This approach is a crucial step toward building a system capable of handling the diverse and complex tasks required for artificial general intelligence (AGI). The ability of specialized modules to operate beyond simple text generation brings us closer to creating models that think and reason in ways that more closely resemble human cognition.

5 Training

The training of the proposed modular system is a complex but critical process aimed at developing a well-integrated set of specialized models, each capable of handling a unique cognitive aspect. The training process involves simultaneous learning across different modules, which allows

them to collaborate and develop abstract meanings collectively, enhancing the system’s overall capability.

5.1 Simultaneous Training Strategy

In this system, all modules (including the LLM, Logic, Maths, Imagination, and Sensitivity modules) are trained concurrently. The simultaneous training ensures that each module learns its specific tasks while also understanding how its output will contribute to the overall system. This holistic approach mirrors human cognition, where sensory and cognitive processes develop in tandem.

For example, when the LLM learns what constitutes a “human,” the Sensitivity module is simultaneously learning to recognize visual and auditory cues associated with humans. This allows both the LLM and Sensitivity to share an abstract, synchronized understanding of the concept of “human”. Such integrated training helps ensure that each module develops knowledge that aligns well with other parts of the system.

5.2 Cross-Module Communication

Training the modules together requires an effective mechanism for communication between them. During training, cross-module connections are established to enable efficient sharing of information. For instance, when training the Imagination module, it receives contextual information from the LLM about the scene it needs to imagine what might happen. The Maths module, on the other hand, may receive logical information from the Logic module regarding the problem it needs to solve.

To facilitate this, the system uses a shared latent space where different modules can share their learned representations. This shared space allows each module to understand and build on the information provided by others, which is particularly useful for tasks requiring multiple forms of reasoning or sensory inputs.

5.3 Curriculum Learning for Complexity Management

The training follows a curriculum learning approach, beginning with simple tasks involving individual modules and gradually introducing more complex tasks that require coordination between multiple modules. For example:

- Phase 1: Each module is trained independently but simultaneously on tasks specific to its domain. The Maths module learns to solve basic equations, the Sensitivity module learns to recognize basic visual objects, and the LLM learns natural language understanding and generation.
- Phase 2: Modules are gradually integrated to solve more complex tasks. For instance, the LLM and Logic module work together on simple logical problems, and the Sensitivity and Imagination modules collaborate on predicting what will happen next and association.
- Phase 3: Full integration occurs, where the system is presented with multi-faceted tasks. An example might be a user describing an abstract concept that involves visualizing, logical reasoning, and mathematical computation. Each module contributes to the task, with the LLM acting as the central coordinator.

5.4 Shared Loss Function

To ensure effective collaboration, a shared loss function is employed. This loss function measures not only the accuracy of each individual module but also the coherence and correctness of the combined system output. For instance, when solving a math problem described in natural language, the loss function would take into account the accuracy of the Maths module's calculations, the Logic module's reasoning steps, and the LLM's ability to generate an understandable response.

The shared loss function is computed by combining individual module losses with an integration penalty, which ensures that the outputs of each module are aligned and contribute meaningfully to the final system output. This mechanism promotes consistency and ensures that all modules learn to work harmoniously from the very beginning.

5.5 Training Example: Solving a Multi-Modal Problem

Consider the task: “A user provides a visual image of a geometric shape and asks, ‘What is the perimeter of this shape if each side is labeled with a length of 5 units?’”

1. Sensitivity Module: The Sensitivity module processes the visual image to recognize that it is, for example, a pentagon.
2. LLM and Maths Module: The LLM converts the user's question and the Sensitivity module's output into a mathematical query. The Maths module calculates the perimeter based on the given side length (i.e., 5 units).

3. LLM Integration: The LLM integrates the calculations from the Maths module and generates the final response: “The perimeter of the pentagon is 25 units.”

During training, the shared loss function evaluates the performance of each module independently (e.g., recognizing the shape correctly, computing the perimeter) and the system as a whole (e.g., providing a coherent, correct response to the user). Errors are backpropagated across all involved modules, encouraging improvements not only in individual tasks but also in how well the modules work together.

5.6 Challenges and Considerations

Training a modular system in this way presents unique challenges, such as balancing the learning rates of different modules to prevent any module from lagging behind or overfitting. Adaptive learning rates are employed to ensure each module progresses at an appropriate pace. Additionally, attention mechanisms help prioritize which modules should focus on particular aspects of a given task, improving efficiency and collaboration during training.

Overall, the simultaneous training of specialized modules within a shared framework helps create a highly integrated system capable of performing complex tasks requiring diverse cognitive functions, moving us closer to realizing artificial general intelligence (AGI).

6 Results

In our experiment, we evaluated the performance of different AI models on a set of 15 challenging questions that tested their logical reasoning, problem-solving abilities, and language understanding. The models evaluated included using several Llama 3.1 70B models with prompts to act as the roles in the modular system, GPT-4, GPT-4o, and o1-preview. Below, we present the quantitative performance results, highlighting the differences between these models.

Model	Score
o1-preview	11/15
GPT4	8/15
Modular system (based on llama 3.1 70B)	5/15
GPT4o	4/15

Table 1: The modular system achieves better scores than GPT4o model in this test.

The o1-preview model demonstrated the highest performance among the four models, correctly answering 11 out of 15 questions, thereby achieving an accuracy rate of 73.3%. GPT-4 achieved 8 correct answers out of 15, resulting in an accuracy of 53.3%. The modular system consists of Llama 3.1 70B models provided the correct answers for 5 out of 15 questions (33.3% accuracy), while the GPT-4o model only achieved 4 correct answers (26.7% accuracy) and the llama 3.1 70B model only achieved 3 correct answers (20.0%).

These results indicate that the modular system consists of 7 Llama 3.1 70B (act as the roles of LLM, Ratiocination, Analysis, Logic, Maths, Imagination and Ideas which are required to answer the questions) in our selected test set. This suggests that this system has a superior potential in handling the specific logical and problem-solving tasks represented by our question set. The diversity of questions, drawn from mathematical problems, logical conundrums, and language challenges, tested not only linguistic competence but also deeper cognitive abilities, making these results particularly significant in evaluating the general utility of large language models in complex reasoning tasks.

Overall, the system's superior performance highlights its potential as a strong candidate for tasks that require enhanced problem-solving and logical reasoning capabilities, beyond mere language comprehension. Building them by modules with higher abilities like GPT4 may exceeding o1-preview and achieving higher goals like general human level AI.

7 Conclusion

This thesis has explored a novel modular architecture for enhancing the abilities of large language models (LLMs), aimed at addressing their current limitations in logical reasoning and complex problem-solving. Through integrating specialized cognitive models around a central LLM, this research demonstrates the potential for creating a more flexible and capable AI system that mimics the multifaceted aspects of human intelligence.

The results from our theoretical analysis and hypothetical case studies indicate that our modular approach not only improves the LLM's ability

to handle complex cognitive tasks but also enhances efficiency and accuracy. The specialization of modules allows for targeted improvements in areas like logic, mathematics, and imaginative thinking, which are crucial for tasks that transcend basic language processing.

Future work will focus on several key areas: practical implementation which will conduct real-world implementations to validate the theoretical models and adjust the system based on practical feedback; optimization of modules including further refinement of individual modules to improve their efficiency and ability to handle even more diverse tasks; inter-module communication (Enhancing the communication pathways between modules to ensure smoother information transfer and integration, aiming to reduce computational overhead and improve response times); application in various domains such as expanding the application areas of this modular system to include fields like robotics, virtual assistants, and complex data analysis, where AI can significantly impact efficiency and innovation.

The modular system proposed in this thesis represents a significant step toward overcoming the limitations of current LLMs, marking a path forward towards the realization of artificial general intelligence (AGI). As AI continues to evolve, the integration of specialized capabilities with core language models will be crucial in building more intelligent, adaptable, and human-like systems.