

Unleashing Artificial Cognition: Integrating Multiple AI Systems

Muntasir Adnan

Faculty of Science and Technology, University of Canberra
Email: adnan.adnan@canberra.edu.au

Buddhi Gamage

Faculty of Science and Technology, University of Canberra
Email: buddhi.gamage@canberra.edu.au

Zhiwei Xu

Faculty of Science and Technology, University of Canberra
Email: danny.xu@canberra.edu.au

Damith Herath

Faculty of Science and Technology, University of Canberra
Email: damith.herath@canberra.edu.au

Carlos C. N. Kuhn

Faculty of Science and Technology, University of Canberra
Email: carlos.noschangkuhn@canberra.edu.au

Abstract

In this study, we present an innovative fusion of language models and query analysis techniques to unlock cognition in artificial intelligence. Our system seamlessly integrates a Chess engine with a language model, enabling it to predict moves and provide strategic explanations. Leveraging a vector database through retrievable answer generation, our OpenSIAI system elucidates its decision-making process, bridging the gap between raw computation and human-like understanding. Our choice of Chess as the demonstration environment underscores the versatility of our approach. Beyond Chess, our system holds promise for diverse applications, from medical diagnostics to financial forecasting.

Keywords AI cognition, Chess, large language models, query analysis, retrievable answer generation

1 Introduction

Artificial Intelligence (AI) systems have achieved remarkable feats in specialized areas such as image recognition and natural language processing [13, 19, 36]. Despite these advancements, individual AI models typically excel in isolated tasks and lack general cognition abilities, leading to Artificial General Intelligence (AGI) [25]. This fragmentation restricts their potential for broader and more generalized applications requiring seamless interaction of multiple cognitive functions.

Human cognition is marked by adaptability, creativity, and emotional intelligence, guided by goals, norms, and social and ethical considerations [24]. In contrast, artificial cognition involves simulating these processes in machines, enabling them to perform tasks autonomously [21]. Studies have highlighted the strengths and limitations of human and artificial cognition, emphasizing the need for understanding these differences for effective human-AI collaboration [14].

The Turing Test, introduced by Alan Turing [26], posits that a machine can be considered intelligent if it can carry on a conversation indistinguishable from a human. Despite its historical significance, the Turing Test has notable limitations. It is anthropocentric, assuming human-like conversation as the definitive marker of intelligence, thereby excluding other forms of intelligence like complex problem-solving or creative pattern recognition. Critics, including Turing, have argued that pre-programmed responses could deceive the interrogator, undermining the test's ability to assess cognitive abilities [9, 23]. Additionally, the Turing Test lacks granularity

in evaluating cognition, as it does not assess various cognitive abilities such as attention, memory, learning, and reasoning, nor does it compare AI’s cognitive stages to human levels [20].

Evaluating cognition in AI involves assessing the system’s ability to perform tasks requiring intelligence and adaptation to various situations. This includes simulating human-like cognitive processes to enable socially intelligent and adaptive interactions with humans [15, 27, 29]. By incorporating specific tasks that assess the mentioned cognitive qualities, we aim to create a more comprehensive assessment strategy for AI cognition, offering insights into the strengths and weaknesses of AI systems.

In the book “Cognitive Robotics”, Cangelosi and Asada [1] discuss eight cognitive abilities, drawing on the work of [15], who examine seven essential cognitive abilities: perception, attention mechanisms, action selection, memory, learning, reasoning, and meta-reasoning. Vernon et al. [28] add anticipation to this list.

Following these ideas, in this study centred around chess, we identified five cognitive qualities relevant to chess players for making decisions during gameplay. The cognitive qualities we focus on are:

- **Perception.** The ability to interpret and understand sensory information from the environment.
- **Memory.** The capability to store, retain, and retrieve information.
- **Attention.** The skill of focusing on relevant stimuli while filtering out distractions.
- **Reasoning.** The ability to draw logical inferences and conclusions from available information.
- **Anticipation.** The capability to predict future events or outcomes based on current information and past experiences.

This paper focuses on developing the initial requirements for an AI system to achieve higher cognition levels in a closed environment. We present a systematic way to evaluate the cognitive capabilities of our integrated system. We show that individual models may exhibit cognitive qualities independently, and their integration can lead to the emergence of cognitive behaviours comparable to humans.

2 Methodology

2.1 Proposed System for Demonstrating Cognitive Abilities

Our proposed system integrates multiple AI models and tools, each specialising in different aforementioned cognitive qualities. By integrating these tools, we aim to enable the system to perform complex tasks that require the interplay of multiple cognitive functions, thus exhibiting cognition.

While constituting a mainstream and *demonstrably effective* set, the employed techniques are acknowledged to be limited. More advanced fine-tuning, Retrieval-Augmented Generation (RAG) [16], and Retrieval-Augmented Fine-Tuning (RAFT) [35], may offer [8] further performance enhancements. Nonetheless, this study combines these mainstream technologies to assess the feasibility and potential for introducing human-like cognitive capabilities within an AI system. The proposed system encompasses a range of services that the agents can decide to employ. To evaluate its efficacy, the system is designed to experiment with several Large Language Models (LLMs). The system is comprised of the following components:

- A query analyser service.
- Base LLM or a fine-tuned LLM using Parameter-Efficient Fine-Tuning’s (PEFT) [34] and Low-Rank Adaptation (LoRA) [11].
- An external knowledge source facilitated by a Faiss vector database and RAG capability.
- A chess engine service powered by Stockfish.
- A vector database update service that allows real-time information updates.

The components mentioned above work together to achieve cognitive qualities within the system. Figure 1 illustrates this collaboration in detail, depicting the system architecture.

Fine-tuning. For fine-tuning, we leverage an instruction tuning [30, 31] methodology. The base model is Mistral 7B, chosen for its balance of performance, efficiency and size.

To promote slow and deliberative reasoning in a small student model, we employ a teacher-student learning paradigm [18]. OpenAI’s GPT-4o served as the teacher model, and we interacted with it using specific system

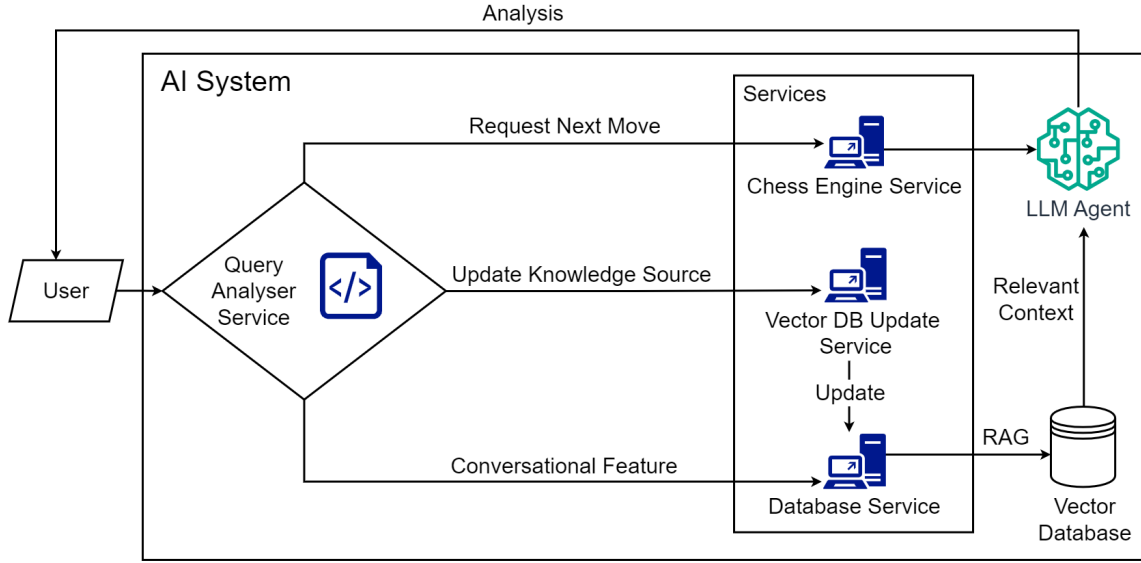


Fig. 1: System Architecture and Data Flow

prompts designed to elicit a deliberate step-by-step reasoning approach while generating responses. These outputs form a core component of the training data. Conversely, during fine-tuning the student model, a generic system prompt is employed. This generic prompt, used for future inference across all models, would allow the student model to leverage its learned reasoning skills as its default behaviour. This is because the pre-trained teacher model, due to its size and capacity, may provide answers directly, while the small student model will require a more deliberative approach to provide similar answers [18]. This distinction in prompt engineering transfers the generalised reasoning abilities from the teacher model to the student model.

Additionally, a dataset of chess games with detailed and step-by-step reasoning annotations for each move, akin to a “chain-of-thought” approach [32], is constructed [35]. This custom chess reasoning dataset is further supplemented by the publicly available Kaggle Lichess dataset to provide broader coverage of chess scenarios.

The finetuning training dataset included the following tasks:

- **Chess Move analysis.** Explain the rationale for a chess move prediction with reference analysis from OpenAI.
- **Next Move Prediction.** Predict the best next move played by a human opponent with samples from Kaggle Lichess.
- **Game Winner Prediction.** Predict the winner of a chess game based on the current state with samples from Kaggle Lichess and OpenAI.
- **Piece Capture Analysis.** Analyse potential captures based on a list of moves in Algebraic Notation using custom scripts.
- **FEN Parsing and Reasoning.** Reason from a position represented in Forsyth-Edwards Notation (FEN) format [7] using custom scripts.
- **Retrieval-Augmented Fine-Tuning dataset.** Analyse the provided contexts and use the most relevant context to generate an answer, following work of [35]

The diversity of the training data aims to enhance the model’s ability to “think” in various chess scenarios and develop transferable reasoning capabilities. Additionally, we introduce new knowledge to the model through fine-tuning, enabling it to understand and utilise FEN. Although the model successfully demonstrates this new capability, the new knowledge increases its propensity for hallucinations when confronted with zero-shot tasks [8].

To optimize computational efficiency during fine-tuning, we employ several techniques. The base Mistral 7B model is quantized to a 4-bit representation [4], significantly reducing the computational complexity regarding memory and running time. Furthermore, instead of fully fine-tuning the base model, we leverage Low-Rank Adaptation (LoRA) to fine-tune a smaller adapter module. This approach further reduces the computational

overhead associated with fine-tuning. The quantization configuration is consistent across all models utilising the Hugging Face Transformers library [33].

For inference, we use the finetuned PEFT model, which involves adding the fine-tuned LoRA adapters to the quantized base model. Interestingly, we observe behavioural discrepancies between the fine-tuned model when adding the LoRA adapter (as a PEFT model) and fully merging the adapter into the base Mistral 7B model. While the PEFT model performs as expected during inference, a more detailed investigation into this discrepancy is warranted for future studies.

Finally, to assess the effectiveness of the proposed fine-tuning methodology for comparisons, the AI system is additionally evaluated with base models from other LLM families, including GPT-4o, GPT-3.5 Turbo, Gemma 7B Instruct, and Mistral 7B Instruct. The results of this comparative analysis will be presented in the experiments.

Retrieval-Augmented Generation (RAG). We adopt RAG [16] in our framework to leverage external knowledge sources stored in a Faiss vector data store [6]. An embedding model plays a crucial role in transforming both textual information from the knowledge source and user queries into high-dimensional vectors. This enables efficient similarity search using distance measurement [5] between the query vector and vectors representing data points in the Faiss database, where the smallest distance indicates the most similar context to the query vector. The `similarity_search_with_score` from LangChain [2] facilitates the retrieval of the most relevant contexts based on these distances.

A pre-defined similarity threshold filters retrieved information to extract highly relevant knowledge for reasoning. This is crucial as we aimed to assess the system’s ability to exhibit reasoning and attention across diverse scenarios, potentially extending beyond the immediate domain of chess.

All base LLMs used in the system leverage the same embedding model `embedding_model_name` retrieved from the Hugging Face library for seamless integration with the Faiss vector store, which works seamlessly for both pipelines using LangChain and the Llama indexing [17] library. However, combining the fine-tuned model requires slightly more advanced techniques since Llama Indexing currently lacks support for PEFT models. Additionally, the fine-tuned model necessitates further training to leverage retrieved contexts effectively. To address these concerns, we use RAFT and provide our strategy below.

Retrieval-Augmented Fine-Tuning (RAFT). The base models perform optimally within the RAG framework, effectively utilising the provided contexts. However, the fine-tuned model exhibits confusion when presented with the extracted contexts from the documents. Although the model demonstrates its capability to tackle various problem-solving tasks, including novel queries for piece-capture and game-analysis problems in a zero-shot manner, it struggles to deal with entirely novel contexts, leading to hallucinations in response. Consequently, further fine-tuning is necessary to instruct the model on effectively leveraging the provided contexts [35].

We adopt the methodology outlined in [35], where the authors train an LLM to concentrate on the most relevant context by inducing a slow reasoning analysis of all given contexts, similar to the Chain of Thought (CoT) approach. The resultant dataset comprises a user query, relevant data points, and a CoT-styled analysis of the contexts as the answer, thus instructing the model on how to analyse and utilise the provided contexts. We also include training examples where the model is expected to disregard all provided contexts upon analysis and instead rely on its pre-trained knowledge to respond to the query. An example of this dataset is provided below.

```
Who are Tardo and Peo in {DISQUALIFIED} by Charles L. Fontenay?
-----
To determine who Tardo and Peo are in {DISQUALIFIED} by Charles L. Fontenay , we can
look at the provided context .

1. The context mentions that Tardo , the Solar Council’s Planetary Aid agent , and his
companion , Peo , were taken to the castle . This indicates that Tardo is the Solar
Council’s Planetary Aid agent , and Peo is his companion .

Therefore , based on the context provided , Tardo is the Solar Council’s Planetary Aid
agent , and Peo is his companion .

<ANSWER>: Tardo is the Solar Council’s Planetary Aid agent , and Peo is his companion .
```

While this approach successfully introduces the fine-tuned model to consider the retrieved contexts, we believe this specific methodology warrants a dedicated study to enhance the RAG capability among various problem-solving tasks for large language models (LLMs).

Chess Engine Service. The chess engine service is an integral component of our system, leveraging the

capabilities of the Stockfish engine [22] to evaluate and demonstrate its cognitive abilities within the domain of chess. Particularly, Stockfish is a highly advanced and open-source chess engine renowned for its robust performance and ability to analyse positions and predict potential moves. In our system, the chess engine service interacts seamlessly with the LLM, providing real-time move suggestions. This integration enables the AI system to delegate tasks to a more capable tool and articulate the rationale of each move, thereby showcasing its reasoning and anticipation capabilities.

Vector Database Update Service. The vector database update service plays a pivotal role in maintaining and enhancing the system’s memory and knowledge retrieval capabilities. This service is designed to update the Faiss [6] vector database in real time, ensuring that the AI system can access the most current and relevant information. This dynamic updating mechanism is critical for the RAG framework, which relies on accurate and up-to-date information to provide contextually relevant responses. The vector database update service ensures that the AI model can learn from its experiences, store important information, and retrieve it when needed, thereby enhancing its cognitive abilities related to memory. This service underscores the system’s capacity for continuous improvement and adaptation, essential traits for demonstrating artificial cognition in varied and evolving scenarios.

2.2 Scoring Mechanism

This section explores how to assess an AI System’s cognitive qualities in a closed environment; in this case, we are using Chess as the closed environment.

To quantify the cognition capability of our AI system, we design a scoring mechanism with the aforementioned 5 qualities from the perspectives of the agent’s environment understanding, information processing, and solution provision. To this end, we have developed a Question-and-Answer (Q&A) testing system, wherein the list of questions was meticulously curated to consider each quality. It is important to note that the Q&A dataset used in this study for the particular chess domain is not exhaustive; the questions can be generalized to other task domains and the AI system’s level of cognition [20].

The evaluation of each quality is based on statistics of sufficient test samples. We provide the details and their corresponding assessment criteria below. These criteria form the basis of our evaluation, ensuring a thorough and systematic assessment of AI’s cognitive capabilities. Standard Algebraic Chess Notation [10] is used to represent the chessboard and assess cognitive aspects.

Perception. To assess perception, we simulate a chessboard state by providing a sequence of moves into the system. We then query the system with questions that evaluate its understanding of the current board state, including:

- Understand chess piece position on a given FEN.
- Compute the number of captured pieces in Algebraic Chess Notation.
- Provide step-by-step analysis of pieces captured, the number of pieces left in total, and the number of pieces left for each player in Algebraic Chess Notation.

For the capture analysis questions, we reward partial understanding of the board. If a model manages to predict the number of captures partially, we will penalize only for the incorrect predictions which include skipping piece capture or overestimating the number of captures. If the number of captures in the FEN is n_c and the prediction from the model is n_m , the formula for scoring on this query is

$$\text{Capture Analysis: } s_{\text{capture}} = 1 - \frac{|n_c - n_m|}{n_c} . \quad (1)$$

Then, we get the overall perception score by using the following formula where the FEN perception score is s_{FEN} , capture analysis score is s_{capture} , and piece analysis score is s_{piece} ,

$$\text{Perception Score: } s_{\text{perception}} = \frac{s_{FEN} + s_{\text{capture}} + s_{\text{piece}}}{\text{number of questions}} . \quad (2)$$

Memory. The system’s memory is evaluated using questions that assess its general chess knowledge. Furthermore, the AI system is augmented with RAG [16] and has access to two chess books, simulating external

knowledge sources. Questions specific to the system’s external knowledge sources are included, requiring relevant sections and utilising the gathered context.

In this study, we investigate cognitive processes akin to long-term memory. As previously discussed, memory is encoded through a combination of the base model’s knowledge and the RAG architecture. To enhance memory retention, we enable the system to incorporate new information into its vector database. Consequently, it can store relevant interactions with users, enriching its memory capacity

The memory scoring questionnaire is designed to contain a singular solution, ensuring that the score is calculated by the formula as follows

$$\text{Memory: } s_{\text{memory}} = \frac{\text{number of correct answers}}{\text{number of questions}}. \quad (3)$$

This method facilitates an objective assessment of the memory performance.

Attention. The attention mechanism is subjected to a comprehensive assessment employing a tripartite approach. Initially, a sequence of chess moves is presented, accompanied by questions that require understanding specific segments of the chess moves. This evaluation aims to scrutinize the system’s capacity for focused attention on pertinent data. Furthermore, the ability to answer any question and even participate in Q&A tests demonstrates high attention quality when the system can identify the context of the question to use relevant knowledge. To test this, questions irrelevant to the control environment are introduced, assessing the system’s ability to comprehend question context and access relevant information or acknowledge uncertainty. Notably, the accuracy of the response is secondary to the system’s ability to recognize and utilise contextual cues. Finally, RAG questions are used to assess the system’s ability to retrieve relevant context based on user queries. The questionnaire for attention contains a singular solution, ensuring that the score is calculated using Eq. (3).

Reasoning. To assess the system’s reasoning capabilities, we have curated a collection of chess puzzles from <https://www.chess.com/puzzles> [3]. We provide the initial chess board setup in FEN, which is then used to configure the board via the chess engine service. Ideally, our system should be able to select the appropriate service dynamically, which itself would demonstrate reasoning. However, for this proof-of-concept framework, we employ a query filtering script that detects keywords to trigger the chess engine service. Once the board is set, the chess engine service attempts to solve the puzzle. The LLM’s task is to generate explanations for the prediction of the best moves, which will be evaluated through human supervision using the rubric in Table 1.

Score	Notation	Explanation
0	Inadequate	The model exhibits solely erroneous assertions with no basis in reality.
1	Deficient	The explanation contains elements of accuracy but is predominantly flawed.
2	Satisfactory	The model correctly identifies the action taken but lacks sophisticated strategic insight.
3	Competent	The explanation is free from false information and accurately describes the move.
4	Proficient	The model articulates a cogent rationale and strategic understanding behind the move.
5	Exemplary	The model provides an exceptional explanation that reveals a profound strategic acumen.

Table 1. Score scale for human study. This is an extension of the traditional five-point assessment framework, commonly utilised in empirical research. For this study, a six-level scale has been employed to augment the granularity of differentiation among the evaluated models’ performances

For the system’s reasoning quality, the LLM provides analysis for the n^{sys} moves provided by the system solution, followed by a human evaluation using the six-level scale in Table 1. The score for reasoning is then averaged over all the assessed analysis of the predicted moves on all M puzzles, defined as

$$\text{Reasoning: } s_{\text{reasoning}} = \frac{1}{5M} \sum_{i=1}^M \sum_{k=1}^{n_i^{\text{sys}}} \frac{s_i(k)}{n_i^{\text{sys}}}. \quad (4)$$

Anticipation. A proficient chess player can predict their opponent’s moves, strategically capitalize on this foresight and plan several moves ahead. To assess this anticipation skill, we analyse the responses of an AI agent during puzzle-solving curated from Chess.com [3].

Each puzzle includes the best solution provided by Chess.com, denoted as n_{best} , which represents the minimum number of moves to solve the puzzle. The system attempts to solve the puzzle by making a series of moves, with the total number of moves taken by the system is denoted as n^{sys} .

For the anticipation quality, the LLM computes the ratio of the n moves in the agent’s prediction over the n^{sys} best moves provided by the system, and the average score over all M puzzles is

$$\text{Anticipation: } s_{\text{anticipation}} = \frac{1}{M} \sum_{i=1}^M \min\left(\frac{n_i}{n_i^{\text{sys}}}, 1\right). \quad (5)$$

3 Experiments

In this section, we present a full evaluation and analysis of the cognition performance of the proposed OpenSIAI System, which is featured by 5 cognition qualities: perception, memory, attention, reasoning, and anticipation. We provide the quality scores on 5 LLMs in Fig. 2. Our proposed AI system integrates LLMs with 3 services: a chess engine for best move prediction using Stockfish, a vector database for dynamic information retrieval, and retrieval-augmented generation on documents. The main LLMs for evaluation are GPT-4o, GPT-3.5 Turbo (for anticipation), Gemma 7B Instruct, Mistral 7B Instruct, and fine-tuned Mistral 7B. All GPU-related experiments are conducted on NVIDIA GeForce RTX 3090, and our OpenSIAI system will be available upon publication.

3.1 Evaluation on System Services

a) Best-move Prediction for Chess Game. The best-move prediction aims to predict the best next move for a given chess FEN or a sequence of moves. Our system incorporates the strong open-source chess engine, Stockfish, with interaction with LLMs to analyse the move decision, yielding nearly perfect predictions to the ground truth labels obtained from [3]. In Fig. 2, Gemma 7B Instruct, Mistral 7B Instruct, and fine-tuned Mistral are unable to predict any correct best moves, indicating their deficiency in the game reasoning.

While GPT-4o alone exhibits an unexpectedly high prediction accuracy with a 32.5% success rate, Fig. 2, it falls short in strategic reasoning, particularly in determining check and checkmate situations. In contrast, our system demonstrates superior performance by accurately predicting the optimal moves in 40 chess games, underscoring the significant benefit of integrating our chess engine service.

b) Vector Database for Dynamic Information Retrieval. Similar to one-shot learning, our system can retrieve up-to-date context by adding certain information with an updated timestamp to the inbuilt vector database, which is managed by using the Facebook AI similarity search tool. Information to be added can be sourced from a document or a sentence prefix with `__update__store__`.

c) RAG on Documents. In our system, the query into LLMs will first be used to retrieve relevant context from the vector database, followed by prompt generation under a tuned prompt template to trigger the LLM engine for text generation. This retrieved context will be filtered out if the score of its cosine similarity to the query is under a given threshold, 0.7 in our case, to avoid misleading information in the query to LLMs. For information retrieval from an external document, our system uses LangChain to split the document into chunks with a chunk size 1,000 and overlap size 100, and then encodes them with an embedding model to update the vector database. The success rates achieved by GPT-4o, Gemma 7B Instruct, and Mistral 7B Instruct are 80%, 70%, and 77.5% respectively.

3.2 Evaluation on System Cognition Capability

a) Evaluation Scale. We evaluate the cognition ability of our OpenSIAI system with 5 qualities in Sec. 1. The evaluation scale of each quality is provided below.

- **Perception.** While the perception can be represented as the system’s understanding of a scenario, we provide 3 types of questions: 40 questions for parsing chess piece position, 20 questions for parsing chess piece status, and 8 questions for identifying chess piece captures.
- **Memory.** The system’s memory is evaluated on 3 types of questions: 40 questions for retrieving LLM’s base knowledge, 8 questions for the memory of previous scenarios in a chess game, and 6 questions for retrieving context from the updated system vector database.
- **Attention.** We evaluate 40 RAG questions that are extracted or raised from 3 books. The correct attention should be localized on the page providing the correct answer to the question.
- **Reasoning.** The reasoning capability is evaluated through human annotation, assessing LLM’s analysis of the suggested move by the chess engine service, for a given FEN. A dataset of 40 chess puzzles is

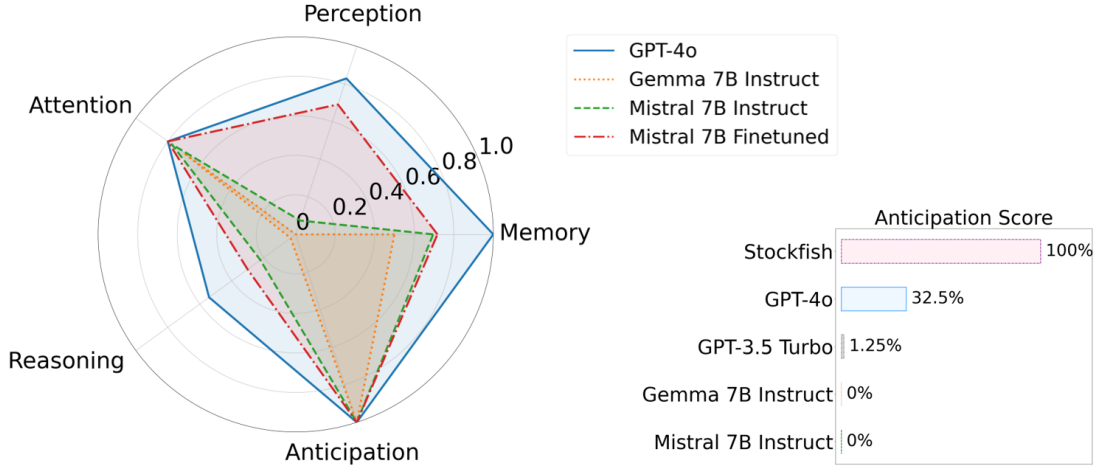


Fig. 2: The chart on the left displays the normalised scores for the five cognitive qualities measured using the OpenSIAI System powered with several LLMs. On the right, we evaluate the anticipation¹ for how well these LLMs can predict the next move in a chess game compared to our OpenSIAI System. LLMs are inferior in predicting the optimal next chess move.

compiled, each comprising 1-3 move sequences for either Black or White. 4 human annotators rated the model’s reasoning ability based on the rubric in Table 1.

- **Anticipation.** With the same 40 chess puzzles used for reasoning, the anticipation ability is measured by the ratio of correctly predicted best moves using our system and comparison LLMs.

b) System Cognition Capability. In Fig. 2, our system provides the statistics of 5 qualities evaluated on 4 public LLMs. Particularly, GPT-4o performs the best, and Mistral 7B Instruct outperforms Gemma 7B Instruct on all qualities. From the experiments, we observe that Gemma 7B Instruct is unable to support system prompts, leading to missing or irrelevant responses in several questions, and its difficulty in adapting to varying evaluation scenarios and cognition qualities due to the rigid persona. Different from the anticipation evaluated on base LLMs in Fig. 2, however, when embedding these models in our system with the anticipation operator changed to the inbuilt chess engine, all models achieve high anticipation scores by predicting the next best move given a puzzle FEN.

Among these LLMs, GPT-4o emerges as the frontrunner in reasoning, articulating the logic underpinning its subsequent actions. These experiments underscore the potential of our OpenSIAI system to augment the capabilities of such LLMs through the embedded system services. For instance, the integration of GPT-4o and the chess engine has shown an improvement in both reasoning and anticipation through the OpenSIAI system. Furthermore, given the superior performance of Mistral 7B Instruct, we fine-tuned A Mistral 7B model to enhance its cognitive capability for domain-specific tasks.

4 Future Work

In our current system, the query analyser utilises keyword detection to route user queries to appropriate services and employs a Chain of Thought (CoT) query service to convert user queries into CoT queries through keyword matching. Despite its effectiveness in improving the system’s cognition ability, this method has limitations regarding flexibility and scalability. To overcome these challenges, we propose enhancing the query analyser by replacing the keyword detection mechanism with a fine-tuned LLM or a classification model. This will facilitate more accurate and context-aware routing of user queries to relevant services. Furthermore, we intend to automate and refine the CoT query generation process by fine-tuning an LLM to produce CoT queries, thereby enhancing the system’s reasoning capabilities [12]. To further increase system reliability and scalability, we will integrate multiple fine-tuned LLMs replacing the single LLM at the core, each specialized for different tasks relevant to the domain, and deploy them in a distributed system architecture. This will enable the AI system to handle more queries without sacrificing its cognition performance, scale more efficiently by reducing the computational complexity, and facilitate easier maintenance, updates, and addition of external services. Moreover, with the

¹We evaluated GPT-3.5 Turbo and GPT-4o from OpenAI on chess tasks and found that GPT-4o demonstrated superior chess understanding. Hence, we use GPT-4o to evaluate the system’s cognition ability as a strong comparison.

distributed architecture, we can create a robust and adaptable AI system that can be deployed with consumer-level computing resources mainly on GPUs, enabling the integration of numerous specialised components to tackle diverse tasks.

5 Conclusion

In this study, we have showcased the efficacy of integrating multiple AI systems with LLMs to augment the cognition abilities of digital assistants. Our proposed architecture is resilient and intuitive, allowing for seamless incorporation into broader systems. One of the core innovations lies in the query analyser for specific services, thereby enhancing the LLM's role as an interactive intermediary between the user and the system. Furthermore, we have illustrated that while LLMs may exhibit limited predictive capabilities, they are competent at interpreting the responses from predictive tools. This enhanced system holds the potential to assist a wide spectrum of professionals, including financial advisors, lawyers, programmers, etc. In summary, the proposed system integrates multiple AI models and services to create a cohesive framework to demonstrate comprehensive cognitive abilities. By integrating these components, our system aims to bridge the gap between raw computational abilities and human-like cognitive processes, setting the stage for future advancements in AGI.

6 References

- [1] A. Cangelosi and M. Asada. *Cognitive robotics*. MIT Press, 2022.
- [2] H. Chase. Langchain llm app development framework. <https://langchain.com/>, 2022. Accessed August, 2024.
- [3] Chess.com. Play chess online, 2024. URL <https://www.chess.com>. Accessed: 2024-07-01.
- [4] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36, 2024.
- [5] I. Dokmanic, R. Parhizkar, J. Ranieri, and M. Vetterli. Euclidean distance matrices: essential theory, algorithms, and applications. *IEEE Signal Processing Magazine*, 32(6):12–30, 2015.
- [6] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P-E. Mazaré, M. Lomeli, L. Hosseini, and H. Jégou. The faiss library. *arXiv preprint arXiv:2401.08281*, 2024.
- [7] S. J. Edwards. Standard: Portable game notation specification and implementation guide, 1994. URL https://ia902908.us.archive.org/26/items/pgn-standard-1994-03-12/PGN_standard_1994-03-12.txt. Internet Archive. Retrieved 02 July 2024. The page linking to the document is here. Section "16.1: FEN".
- [8] Z. Gekhman, G. Yona, R. Aharoni, M. Eyal, A. Feder, R. Reichart, and J. Herzig. Does fine-tuning llms on new knowledge encourage hallucinations? *arXiv preprint arXiv:2405.05904*, 2024.
- [9] M. Halpern. The trouble with the turing test. *The New Atlantis*, (11):42–63, 2006.
- [10] D. Hooper and K. Whyld. *The Oxford Companion to Chess*. Oxford University Press, 2nd edition, 1996. ISBN 0-19-280049-3.
- [11] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.
- [12] M. Jin, Q. Yu, H. Zhao, W. Hua, Y. Meng, Y. Zhang, M. Du, et al. The impact of reasoning step length on large language models. *arXiv preprint arXiv:2401.04925*, 2024.
- [13] D. Khurana, A. Koli, K. Khatter, and S. Singh. Natural language processing: state of the art, current trends and challenges. *Multimedia tools and applications*, 82(3):3713–3744, 2023.
- [14] J. H. Korteling, G. C. van de Boer-Visschedijk, R. A. Blankendaal, R. C. Boonekamp, and A. R. Eikelboom. Human-versus artificial intelligence. *Frontiers in artificial intelligence*, 4:622364, 2021.
- [15] I. Kotseruba and J. K. Tsotsos. 40 years of cognitive architectures: core cognitive abilities and practical applications. *Artificial Intelligence Review*, 53(1):17–94, 2020.
- [16] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

- [17] J. Liu. LlamaIndex, 11 2022. URL https://github.com/jerryliu/llama_index.
- [18] A. Mitra, L. Del Corro, S. Mahajan, A. Cudas, C. Simoes, S. Agarwal, X. Chen, A. Razdaibiedina, E. Jones, K. Aggarwal, and H. Palangi. Orca 2: Teaching small language models how to reason. *arXiv preprint arXiv:2311.11045*, 2023.
- [19] R. Patil and V. Gudivada. A review of current trends, techniques, and challenges in large language models (llms). *Applied Sciences*, 14(5):2074, 2024.
- [20] J. Piaget. The theory of stages in cognitive development. *McGraw-Hill*, 1971.
- [21] Y. Qu, C. Wei, P. Du, W. Che, C. Zhang, W. Ouyang, Y. Bian, F. Xu, B. Hu, K. Du, et al. Integration of cognitive tasks into artificial general intelligence test for large models. *Iscience*, 27(4), 2024.
- [22] T. Romstad, M. Costalba, J. Kiiski, and G. Allievi. Stockfish. <https://stockfishchess.org/>, 2008. Accessed: 2024-07-22.
- [23] J. R. Searle. Minds, brains, and programs. *Behavioral and brain sciences*, 3(3):417–424, 1980.
- [24] G. Siemens, F. Marmolejo-Ramos, F. Gabriel, K. Medeiros, R. Marrone, S. Joksimovic, and M. de Laat. Human and artificial cognition. *Computers and Education: Artificial Intelligence*, 3:100107, 2022.
- [25] S. Sonko, A. O. Adewusi, O. C. Obi, S. Onwusinkwue, and A. Atadoga. A critical review towards artificial general intelligence: Challenges, ethical considerations, and the path forward. *World Journal of Advanced Research and Reviews*, 21(3):1262–1268, 2024.
- [26] A. M. Turing. Mind. *Mind*, 59(236):433–460, 1950.
- [27] D. Vernon. *Artificial cognitive systems: A primer*. MIT Press, 2014.
- [28] D. Vernon, G. Metta, and G. Sandini. A survey of artificial cognitive systems: Implications for the autonomous development of mental capabilities in computational agents. *IEEE transactions on evolutionary computation*, 11(2):151–180, 2007.
- [29] D. Vernon, C. von Hofsten, and L. Fadiga. Desiderata for developmental cognitive architectures. *Biologically Inspired Cognitive Architectures*, 18:116–127, 2016.
- [30] Y. Wang, S. Mishra, P. Alipoormolabashi, Y. Kordi, A. Mirzaei, A. Arunkumar, A. Ashok, A. S. Dhanasekaran, A. Naik, D. Stap, et al. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. *arXiv preprint arXiv:2204.07705*, 2022.
- [31] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.
- [32] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
- [33] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al. Huggingface’s transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.
- [34] L. Xu, H. Xie, S.-Z. J. Qin, X. Tao, and F. L. Wang. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. *arXiv preprint arXiv:2312.12148*, 2023.
- [35] T. Zhang, S. G. Patil, N. Jain, S. Shen, M. Zaharia, I. Stoica, and J. E. Gonzalez. Raft: Adapting language model to domain specific rag. *arXiv preprint arXiv:2403.10131*, 2024.
- [36] L. Zhao, L. Zhang, Z. Wu, Y. Chen, H. Dai, X. Yu, Z. Liu, T. Zhang, X. Hu, X. Jiang, et al. When brain-inspired ai meets agi. *Meta-Radiology*, page 100005, 2023.

Acknowledgments

This work is funded under the agreement with the ACT Government, Future Jobs Fund - Open Source Institute (OpenSI) - R01553; and NetApp Technology Alliance Agreement with OpenSI- R01657. We especially acknowledge Carlos Noschang Kuhn for their visionary approach and foundational ideas that made this research possible.

Author Contributions

Muntasir Adnan and Zhiwei Xu collaborated on implementing the code base, designing experiments, evaluating results, and writing the manuscript. Buddhi Gamage focused on experiment design, result evaluation, and manuscript writing. Damith Herath conducted a thorough review of the manuscript. Carols Noschang Kuhn developed the initial concept for this project and contributed to experiment design, research question development, data evaluation, and manuscript writing.