# Can Large Language Models Derive High-Level Cognition from Low-Level and Fragmented Foundational Information?

**Yang Liu**[1], **Xiaoping Wang**[1*], **Kai Lu**[2*]

[1]College of Computer Science and Electronic Engineering, Hunan University, China
[2]College of Computer, National University of Defense Technology, China
liuyang0542@hnu.edu.cn, xpwang@hnu.edu.cn, kailu@nudt.edu.cn

## Abstract

As one of the key technologies leading to Artificial General Intelligence (AGI), Large Language Models (LLMs) have achieved remarkable accomplishments. Exploring the capabilities of LLMs is crucial for scientific research, and many studies propose new challenges from various aspects to explore the boundaries of capabilities in LLMs. This paper attempts to push the challenges of information understanding, synthesizing and reasoning to the extreme, in order to explore the boundaries of more advanced dimensional cognitive capabilities in LLMs. It is defined as the task of High-Level Cognition (HLC), which involves obtaining high-level conclusions from low-level and fragmented foundational information. To evaluate HLC, we construct a dataset based on soccer matches. Experiments and analysis on this dataset show that current state-of-the-art LLMs lack the ability to effectively solve the task of HLC, because their performance is equivalent to random-level. However, by fine-tuning Llama3-8B-Instruct, there are improvements of 14.4%, 48.1%, and 19.4% over random-level in three types of evaluation tasks. This indicates that LLMs have great potential to solve the task of HLC.

**Model & Dataset Details** — https://github.com/nlpmy/hlc

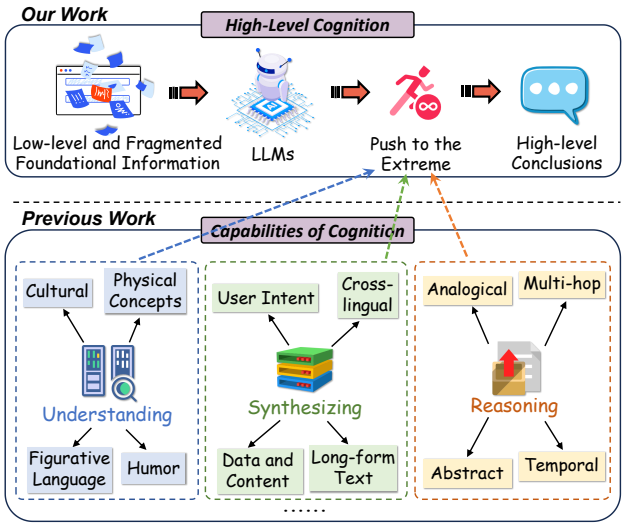


Figure 1: An illustration of our work compared to the previous work. To explore the HLC in LLMs, an attempt that pushing the challenges of information understanding, synthesizing and reasoning to the extreme in this paper.

## 1 Introduction

Large Language Models (LLMs) have demonstrated significant progress in a wide range of Natural Language Processing (NLP) tasks, including intelligent dialogue (Yi et al. 2024; Joko et al. 2024), commonsense reasoning (Krause and Stolzenburg 2024; Zhao, Lee, and Hsu 2024), text comprehension (Chen and Leitch 2024; Säuberli and Clematide 2024), and code generation (Wang and Chen 2023; Jiang et al. 2024b). These demonstrate their strong capabilities in handling complex language tasks. However, the capabilities and potential of LLMs remain an open field. Exploring the boundaries of capabilities in LLMs is crucial for advancing their intelligence. There are many new challenges about the capabilities of LLMs are proposed, which involve understanding (Zhu et al. 2024; Hessel et al. 2022), synthesizing (Hu et al. 2024; Salvador et al. 2024), and reasoning (Gendron et al. 2023; Stechly, Marquez, and Kambhampati 2023). They guide researchers to consider how to improve LLMs, thereby achieving more human-like cognitive intelligence. Nevertheless, the boundaries of more advanced dimensional cognitive capabilities in LLMs remain unresolved, because the standards and methods for measuring in complex cognitive tasks are still not clear enough. To address this, the challenges of information understanding, synthesizing, and reasoning are pushed to the extreme in this paper. It is the task of High-Level Cognition (HLC).

HLC is obtaining high-level conclusions from low-level and fragmented foundational information. In a soccer match, the low-level foundational information includes players' performance and ball events, and the high-level conclusions refer to the tactical characteristics of teams, which cannot be directly extracted from foundational information. To achieve this, the relationship between all low-level foundational information (e.g., passing, shooting) is pushed to the boundaries of understanding. Then an implicit global structure is

*Corresponding author.

formed by pushing all fragmented foundational information to the boundaries of synthesizing. Finally, high-level conclusions (e.g., playing styles, strengths and weaknesses) are drawn by pushing the structure to the boundaries of reasoning. This process is similar to the professional knowledge and experience in human. Therefore, the HLC task is an integrated task of understanding, synthesizing, and reasoning, and it requires the formation of systematic thinking according to a specific cognitive framework.

Exploring the HLC in LLMs is a new challenge. Different from traditional NLP tasks, which often involve simple extraction (Singh 2018), summarization (Awasthi et al. 2021), or text rewriting (Xu et al. 2019), a more complex cognitive framework is required in the HLC task. It cannot be addressed by traditional knowledge enhancement methods (e.g., retrieval-augmented generation (Lewis et al. 2020), knowledge graphs (Hogan et al. 2021), database queries (Chandra 1988)). Moreover, different from tasks involving mathematical or programming skills, the results from HLC task reflect the inherent analytical and reasoning abilities in LLMs, because there is no pre-existing chain of thought (Wei et al. 2022) framework to rely on. There are some obstacles in the research process, and a major is the lack of specialized datasets, because traditional datasets do not meet the complexity required for generating high-level conclusions. Additionally, there are no clear, objective benchmarks for evaluating HLC, and no robust metrics or standard frameworks to accurately assess it.

To address these issues, a dataset for evaluating HLC is constructed in this paper, and it is based on soccer matches. For providing objective evaluation results, the outcomes of HLC are categorized into three types: (a). Multiple-Choice Question (MCQ), which requires LLMs to choose multiple from the existing statements about playing styles; (b). Single-Choice Question (SCQ), which requires LLMs to choose one from the existing statements about playing styles; (c). True/False Question (TFQ), which requires LLMs to determine whether a sentence describing the playing styles is correct. Experimental results show that current state-of-the-art LLMs cannot derive HLC from low-level and fragmented foundational information, because their performance is equivalent to random-level. However, it is found that the HLC can be endowed by fine-tuning. The illustration of our work is shown in Figure 1.

The main contributions of this paper are as follows:

- The HLC task is first proposed, and it is currently one of the most challenging tasks for LLMs. The HLC task lies outside the current LLMs capability boundaries, because the performance of the state-of-the-art LLMs is equivalent to random-level.
- A dataset named *MatchIntel* is constructed to evaluate the performance of LLMs on HLC tasks. The dataset contains both low-level and high-level information about soccer matches. It provides an experimental benchmark for exploring HLC in the future.
- Preliminary testing for the HLC tasks is conducted, and two conclusions are drawn: 1). current state-of-the-art LLMs lack the ability to effectively solve HLC tasks. 2). Fine-tuning can improve the performance of LLMs on HLC tasks to some extent, but it does not fully solve such tasks. These highlight the need for further academic investigation.

The rest of the paper is organized as follows: related work is introduced in Section 2; the proposed scheme is described in Section 3; experiments and analysis are provided in Section 4; limitations are presented in Section 5; Finally, some conclusions are drawn in Section 6.

## 2 Related Work

The cognition of LLMs has been studied across various topics, focusing on the boundaries of capabilities in information understanding, synthesizing, and reasoning. Research findings in these are diverse, with some studies highlighting the limitations of LLMs in a specific capability, while others propose methods to enhance the specific capability in LLMs.

### 2.1 Understanding

A contextual understanding benchmark is introduced to evaluate LLMs' linguistic comprehension within context (Zhu et al. 2024), including both document-based and dialogue-based scenarios. Experimental results indicate that LLMs face challenges in nuanced contextual understanding, especially in in-context learning settings. The capability of understanding humor in LLMs is evaluated (Hessel et al. 2022). There are three tasks derived from the New Yorker Cartoon Caption Contest, and both multi-modal and language-only models are investigated. The study found that current LLMs are still unable to recognize, understand, and evaluate humor as effectively as humans.

### 2.2 Synthesizing

A benchmark SportsMetrics is developed to evaluate the capability of information fusion in LLMs (Hu et al. 2024). This benchmark presents challenges such as adapting to new match rules, interpreting lengthy descriptions, managing scrambled narratives, and analyzing key statistics in match summaries. It highlights the great potential of LLMs in information fusion, though limitations remain. The capability of LLMs for the Semantic Overlap Summarization (SOS) task is studied (Salvador et al. 2024). It involves summarizing overlapping information from multiple narratives. Experimental results show that LLMs struggle with the SOS task, indicating room for improvement in this area.

### 2.3 Reasoning

To evaluate abstract reasoning of LLMs, a framework is constructed using both text and visual datasets (Gendron et al. 2023). The experiments demonstrated that LLMs lack the capability for abstract reasoning, and existing techniques for improving NLP tasks cannot enhance the capability. The effectiveness of iterative prompting strategy in LLMs is explored (Stechly, Marquez, and Kambhampati 2023). Its purpose is to improve the accuracy about reasoning problems. In a self-critique iterative framework for graphic color problems, it is found that iterative prompting sometimes led to worse performance compared to generating a single answer.

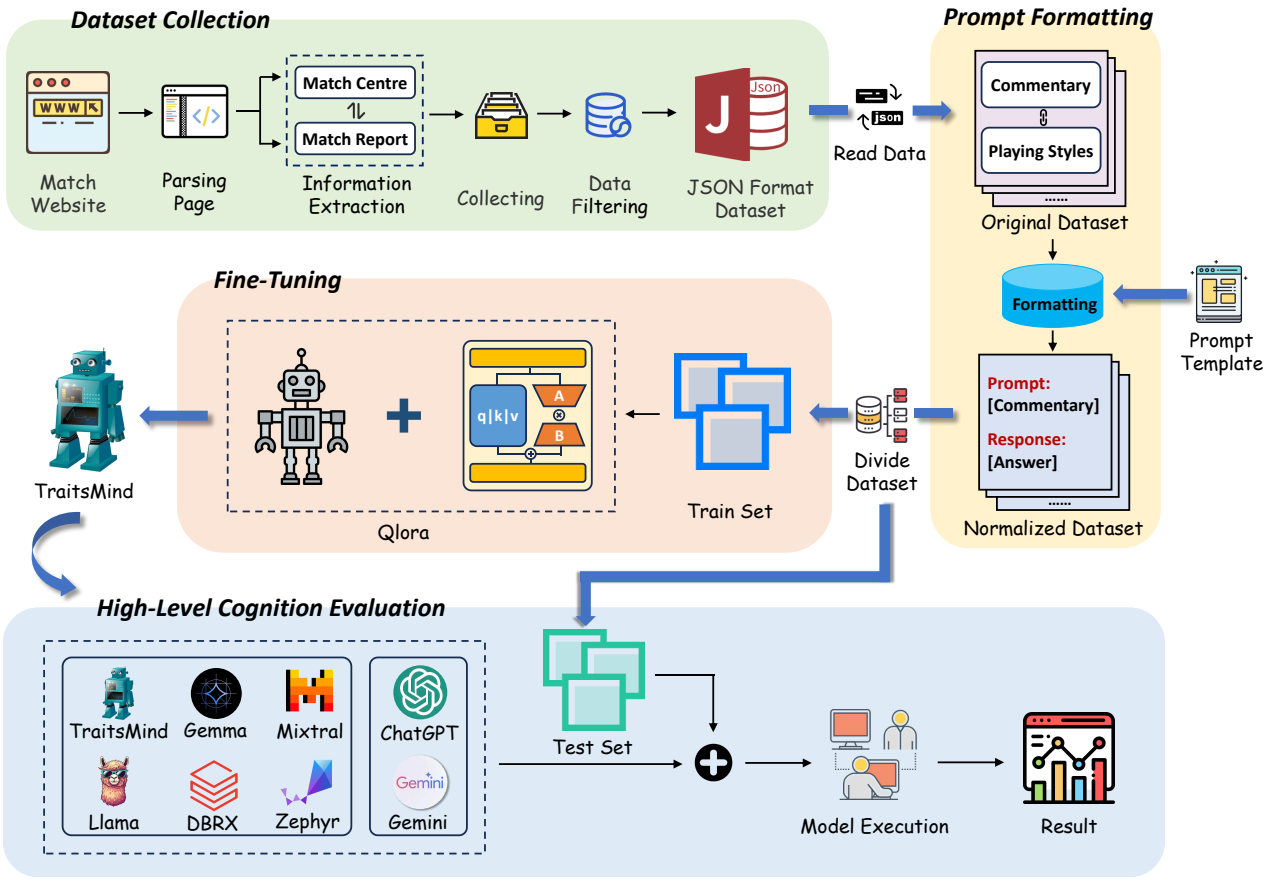


Figure 2: An overview of the proposed scheme. A dataset containing commentary texts and playing styles is collected from a professional soccer data website. Then it is normalized using the prompt template. After that, the dataset is divided into a train set and a test set. The train set is used to fine-tune LLMs, and the test set is used to evaluate the high-level cognition of LLMs.

These studies acknowledge the potential and recognize the boundaries of LLMs across different cognitive topics. However, they ignore the more advanced dimensional cognition in complex cognitive tasks, and it is the integration of multiple cognitive topics. To address this gap, the HLC task is performed in this paper, based on data from soccer matches.

## 3 The Proposed Scheme

The flowchart of the proposed scheme is shown in Figure 2. Firstly, a substantial corpus of data about soccer matches is collected from a sports reporting website, and then stored as a JSON dataset. Subsequently, the prompt template is designed, and it formats the information by reading soccer match commentary texts and statements about playing styles from the dataset, thereby optimizing the structure of the dataset. After that, the normalized dataset is partitioned into two sets: a train set and a test set. Finally, the train set is used to fine-tune LLMs, and the test set is used to evaluate the performance of HLC in LLMs.

The dataset collection, prompt formatting, fine-tuning LLMs, and high-level cognition evaluation are described in details below.

### 3.1 Dataset Collection

To meet the standards of HLC, a dataset containing 52,500 samples is collected from a professional soccer data website[1] and named *MatchIntel*. Each sample records the situation of a match, which includes soccer match commentary texts and playing styles of two teams. The soccer match commentary texts are described with details of players' performance or ball events at each moment, and they are preserved as primitive observational records, without complex processing and analysis. For example, Wigan vs. Morecambe in the 2021/2022 season of England League One, the commentary text from the 40th to 44th minute is intercepted as

*"40 | Morecambe | Cole Stockton has shot blocked (Standing, Out of box, Right footed, Open play)*
*43 | Wigan | Stephen Humphrys has attempt saved (Standing, High to the right, Out of box, Left footed, Open play)*
*43 | Wigan | Stephen Humphrys wins a corner (To the right)*
*43 | Morecambe | Trevor Carson makes a save (Diving, Parried safe)*

[1]https://www.whoscored.com


(a) (b) (c)

Figure 3: The prompt templates for three evaluation tasks: (a). Multiple-Choice Question (MCQ); (b). Single-Choice Question (SCQ); (c). True/False Question (TFQ). The red part represents that it needs to be filled in.

*44 | Wigan | GOAL! Jason Kerr scores , Assisted by Tom Naylor (Standing, Low to the left, 6-yard box, Big Chance, Right footed, From corner)"*.

The actions of players are described, but they are not presented in a clear and comprehensible description. Instead, some incoherent soccer terms are listed, which represent the low-level and fragmented foundational information.

The statements about playing styles are categorized into a finite field and consist of 12 types. They cannot be directly extracted from the commentary texts. For example, *"Had a large quantity of possession in their opponent's half"* is one of them. It means that a team constantly passes, controls, and seeks opportunities to attack in the opponent's half, while the opponent's team may be forced to defend and have little chance to get the ball. To obtain this accurately, understanding, synthesizing, and reasoning must be applied to all soccer terms of the commentary texts, and all of these capabilities need to be pushed to the boundaries. Therefore, the statements about playing styles are high-level conclusions, and the dataset can be used to evaluate HLC in LLMs.

## 3.2 Prompt Formatting

For providing objective evaluation results, some HLC evaluation tasks need to be created. When a soccer commentary text is inputted into LLMs, the appropriate statements about playing styles are expected to generate by LLMs. If there is no guidance, these statements will fail to be generated due to the randomness of LLMs' output. Therefore, to ensure the output is focused on the finite field, prompt templates need to be designed. Specifically, the model is defined as a soccer match analyst at the beginning, and a soccer commentary text is provided. Then some statements about playing styles are presented, and a decision needs to be made based on these statements. The prompt templates are designed for the following three tasks:

**Multiple-Choice Question (MCQ)** The playing styles of each match are part of all the statements. Therefore, predicting the ground truth is essentially a multiple-choice question. To simplify the evaluation process, the prompt template is designed to include only four statements, and at least two correct statements are provided. However, existing research indicates that LLMs are sensitive to the numbering and positioning of options when handling choice questions (Zheng et al. 2023; Li and Gao 2024; Pezeshkpour and Hruschka 2023). Thus, in the process of designing, the position of presented statements is not fixed, and LLMs are required to output multiple statements, without any changes. The MCQ prompt template is shown in Figure 3(a).

**Single-Choice Question (SCQ)** The SCQ is a type of question that contains multiple options, and only one is correct while the others are not. After analyzing the dataset, each team has more than three incorrect statements and at least one correct statement. Therefore, when a team is selected, one correct statement and three incorrect statements can be chosen. Then these four statements are randomly arranged to construct a single-choice question. Finally, LLMs are required to choose one from them, and output the statement without any changes. The SCQ prompt template is shown in Figure 3(b).

**True/False Question (TFQ)** LLMs are required to make a judgment on the sentence constructed by the prompt about a team's playing styles. A correct *"include"* sentence contains all correct statements, while a correct *"not include"* sentence contains all incorrect statements. Additionally, an

incorrect *"include"* sentence contains all correct statements and one incorrect statement, while an incorrect *"not include"* sentence contains all incorrect statements and one correct statement. All these statements are randomly listed in the prompt template. After that, LLMs are required to judge the sentence. If it is considered as correct, *"True"* should be output; otherwise, *"False"* should be output. Figure 3(c) shows the prompt template for TFQ.

The dataset is normalized by the three prompt templates and then used to evaluate HLC of LLMs.

### 3.3 Fine-Tuning

The train set is used to fine-tune a LLM for analyzing soccer matches. The Llama3-8b-Instruct (Meta 2024) is chosen as the foundational model, because of its superior performance among the open-source models, especially in processing complex language tasks. Due to the constraints of computational resources, the Parameter-Efficient Fine-Tuning (PEFT) technique (Han et al. 2024) is used. Specifically, the method of Qlora (Dettmers et al. 2024) is adopted, and it is used for fine-tuning by 4-bit precision, significantly reducing the computational burden without compromising the performance. Moreover, a set of compact, learnable Low-rank adapters (Lora) weights (Hu et al. 2021) is introduced, which can be adjusted through backpropagation. Finally, the learning capability of model is enhanced, and effective adaptation to task requirements is allowed. The fine-tuned model is named *TraitsMind*, specifically designed to analyze the playing styles of teams in soccer matches.

### 3.4 HLC Evaluation

The test set is used to evaluate HLC of LLMs. To obtain comprehensive results, popular open-source and closed-source LLMs, along with the *TraitsMind*, are employed for testing. The data is inputted into the model, and then the output is presented in a fixed format, because the output format is defined in prompt templates. Finally, the statements are extracted from the output, and a match is made with the ground truth. The higher the matching degree, the better the performance of the model.

## 4 Experiment

### 4.1 Experimental Setup

**Datasets** The experiment is performed on the dataset *MatchIntel*, with a total of 52,500 samples. 49,500 samples are allocated for the train set, and the remaining 3,000 for the test set. Additionally, the test set is divided into three parts, with each evaluation task containing 1,000 samples.

**Comparison Models** Some open-source LLMs are selected for testing, including Zephyr (Tunstall et al. 2023), Gemma (Team et al. 2024a), Llama3 (Meta 2024), Mixtral (Jiang et al. 2024a), and DBRX (Team et al. 2024b). Advanced performance in text understanding, reasoning, and generation has been demonstrated by these models. To enhance the credibility of the experiment, two closed-source LLMs, ChatGPT (Achiam et al. 2023) and Gemini (Team et al. 2023), are employed for interactive testing through their APIs.

**Evaluation Metrics** Accuracy (Makridakis 1993) is used to evaluate performance. The output is considered accurate when it matches the ground truth exactly.

**Parameter Details** LLMs are fine-tuned using Qlora with 4-bit quantization. A low-rank adaptive approach (rank 64, alpha 16, dropout 0.1) is applied to the query and key-value components. Fine-tuning is done with a single iteration on the train set using the *paged_adamw_32bit* optimizer (Kingma and Ba 2014), initiated with a learning rate of 1e-4 and weight decay rate of 0.01. The maximum input length is capped at 3200 tokens, and the half-precision (FP16) is used for training efficiency. During inference, a conservative and stable output is maintained to improve the accuracy. Therefore, temperature is set to 0 for ChatGPT and Gemini, while 0.01 for other LLMs. The top_p is set to 1.0 for all LLMs.

### 4.2 Experimental Results

The threshold of HLC needs to be determined in three evaluation tasks, and it is the baseline of performance. If the performance of LLMs significantly surpasses the threshold, HLC is considered to be exhibited. Otherwise, LLMs lack the ability to effectively solve HLC tasks. The random-level can be used as the threshold, because it is the expected performance that LLMs can achieve without any cognitive abilities. Pure speculation and actual cognition can be clearly distinguished by it. To get the values of thresholds, combinatorial mathematics and probability theory (Spitzer 1956) are applied, because these tasks have a clear probability about correct answers. Finally, the expected accuracy plus 3 standard deviations (99.7% confidence interval) as the values for thresholds, which can be calculated as 0.118 for MCQ, 0.291 for SCQ, and 0.547 for TFQ. The experimental results of LLMs in three evaluation tasks are shown in Table 1.

From the results, it can be concluded that the LLMs cannot derive HLC from low-level and fragmented foundational information.

- The performance of LLMs is equivalent to the random-level, because they do not significantly surpass the thresholds in three evaluation tasks. Therefore, it indicates that LLMs lack the ability to effectively solve HLC tasks.
- If there is HLC in LLMs, it should become more prominent as model parameters are increased. However, in MCQ and TFQ evaluation tasks, better performance is exhibited by Llama3-Instruct with 8B, rather than the model with 70B. This indicates that HLC is not present in LLMs.
- The *TraitsMind* outperforms all LLMs in three evaluation tasks. In particular, a significantly improved performance compared to its foundational model Llama3-8b-Instruct. Its performance surpasses the thresholds by 14.4%, 48.1%, and 19.4% in three evaluation tasks, respectively. This indicates that fine-tuning can endow the LLMs with HLC, but the performance improvement is only to some extent and cannot fully solve such tasks.

| Model | Params | Evaluation Task | | |
|---|---|---|---|---|
| | | MCQ | SCQ | TFQ |
| *Baseline* | | | | |
| Random | - | 0.118† | 0.291† | 0.547† |
| *Closed-source models* | | | | |
| GPT-3.5-Turbo (Achiam et al. 2023) | - | 0.106 | 0.286 | 0.508 |
| Gemini-Pro (Team et al. 2023) | - | 0.116 | 0.288 | 0.509 |
| *Open-source models* | | | | |
| Zephyr-Beta (Tunstall et al. 2023) | 7B | 0.094 | 0.193 | 0.438 |
| Gemma-1.1-Instruct (Team et al. 2024a) | 7B | 0.113 | 0.219 | 0.484 |
| Llama3-Instruct (Meta 2024) | 8B | 0.111 | 0.267 | 0.519 |
| Mixtral-Instructv0.1 (Jiang et al. 2024a) | 46.7B | 0.108 | 0.260 | 0.534 |
| Llama3-Instruct (Meta 2024) | 70B | 0.084 | 0.288 | 0.434 |
| DBRX-Instruct (Team et al. 2024b) | 132B | 0.060 | 0.250 | 0.530 |
| *The proposed model* | | | | |
| TraitsMind | 8B | **0.135** | **0.431** | **0.653** |
| *(Relative Impr. over Random)* | | *(+14.4%)* | *(+48.1%)* | *(+19.4%)* |

Table 1: Accuracy of LLMs in three HLC evaluation tasks. Two closed-source models, six open-source models and the proposed model are adopted for performance comparison. The data marked with † represent the theoretical values obtained by calculation.

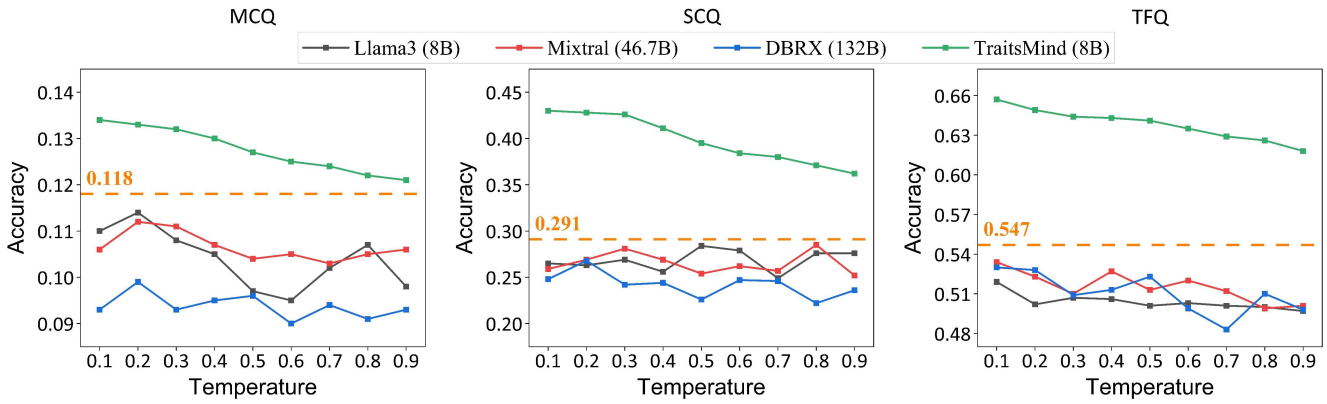


Figure 4: The relation between accuracy and the temperature in three evaluation tasks.

### 4.3 Ablation Studies

To further explore HLC, detailed ablation studies are conducted on temperature parameters, foundational models and quantized values.

**Sensitivity to the Temperature** *Is the high-level cognition endowed by fine-tuning stable?* In the inference stage of LLMs, lower temperature settings lead to more stable and conservative outputs, thereby improving credibility (Savelka et al. 2023). If HLC is stable in LLMs, changes in temperature will significantly affect its accuracy. Therefore, the performance of LLMs at different temperatures is explored, and the experimental results are shown in Figure 4. It can be observed that, in three evaluation tasks, the accuracy of *TraitsMind* shows a significant decrease as the temperature increases, while the accuracy of other LLMs does not exhibit a noticeable trend. It indicates that the HLC can be endowed by fine-tuning, and once endowed, it is stable in LLMs.

**Sensitivity to the Foundational Model** *What is the source of high-level cognition?* Three versions of the Llama series model, with their corresponding dialogue versions, are chosen as the foundational models for fine-tuning. The results are listed in Table 2. It can be seen that new-generation LLMs as the foundational models provide better performance than old-generation LLMs, because new-generation LLMs are pre-trained in larger and higher quality datasets, and the optimization of training modes is further advanced. Also, the architecture is more comprehensive, and the ability to understand contexts is significantly improved, which lead to the richer implicit knowledge is enabled by these new-generation LLMs. Therefore, the implicit knowledge

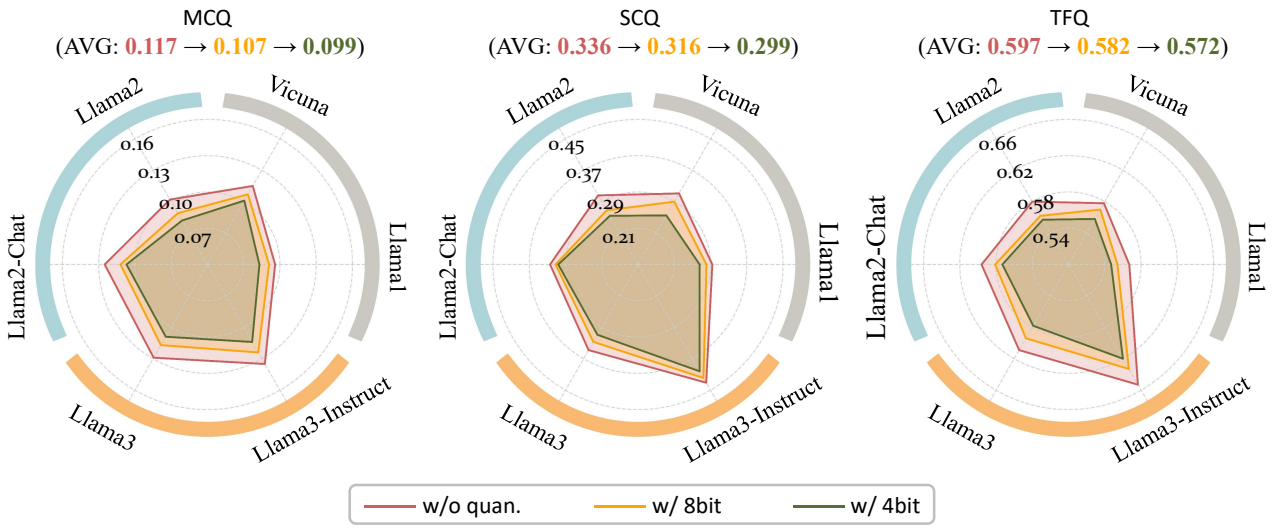


Figure 5: The accuracy comparison on different quantization. The series for Llama and their corresponding dialogue models are adopted, and they are quantized at 8-bit and 4-bit.

| Model (Params) | MCQ | SCQ | TFQ |
|---|---|---|---|
| Llama1 (7B) | 0.096 | 0.294 | 0.567 |
| Vicuna-v1.5 (7B) | 0.115 | 0.311 | 0.578 |
| Llama2 (7B) | 0.102 | 0.306 | 0.580 |
| Llama2-Chat (7B) | 0.125 | 0.324 | 0.596 |
| Llama3 (8B) | 0.129 | 0.348 | 0.609 |
| Llama3-Instruct (8B) | 0.135 | 0.431 | 0.653 |

Table 2: Effects of fine-tuning different foundational models. The accuracy of three series for Llama are evaluated.

is recognized as the source of HLC. Moreover, by fine-tuning the dialogue version LLMs, better performance is shown, which indicates that supervised fine-tuning (Dong et al. 2023) and reinforcement learning (Kaelbling, Littman, and Moore 1996) can boost the HLC in LLMs.

**Sensitivity to the Quantization** *How does the source influence high-level cognition?* After fine-tuning, the LLMs are quantized at 8-bit and 4-bit. Experiments are conducted on three evaluation tasks, and the results are shown in Figure 5. As the quantization bit-width decreases, the accuracy gradually declines, because lower quantization bit-width can lead to a loss of the language processing ability in LLMs. The language processing ability is a manifestation of the implicit knowledge in LLMs, because it is a direct reflection of the extensive and subtle information, which LLMs have internalized during the training, and these are not explicitly programmed, but emerges from the patterns and associations learned from large amounts of data. In higher-precision representations, the LLMs retain and utilize its implicit knowledge better, thereby exhibiting stronger HLC. Therefore, the fidelity of implicit knowledge significantly influences HLC.

## 5 Limitations

There are two limitations. First, the LLMs are fine-tuned only under the initially set hyperparameters. It endows the HLC in a certain extent, but the highest level is not achieved. Therefore, further improve the accuracy of LLMs in these evaluation tasks by optimizing the hyperparameters can be believed. Second, the LLMs are extremely sensitive to prompts. Even if the same meaning is conveyed, different expression styles may change the accuracy of experiments. In this paper, the design of prompts is a preliminary attempt, but further improve the performance by optimizing the prompts is credible. Future works can explore the sensitivity of LLMs to prompts and hyperparameters.

## 6 Conclusion

In this paper, we propose the HLC task, which is currently one of the most challenging tasks for LLMs. To evaluate the performance of LLMs on such task, a specialized dataset *MatchIntel* and three evaluation tasks are created. Experimental results show that the performance of current state-of-the-art LLMs is equivalent to the random-level, indicating that they lack the ability to effectively solve HLC tasks. Additionally, the HLC can be endowed by fine-tuning, but the performance improvement is only to some extent and cannot fully solve such tasks. Therefore, the HLC of LLMs needs further academic investigation in the future. To the best of our knowledge, this paper is the first study to explore the boundaries of more advanced dimensional cognitive capabilities in LLMs. We hope that their capability of processing complex cognitive tasks can be enhanced from this study, enabling better comprehension and simulation of human thinking. However, HLC involves more subtle processes of language processing, which are not entirely transparent or predictable. Future work will focus on improving the interpretability, controllability, and generality.

## References


Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Awasthi, I.; Gupta, K.; Bhogal, P. S.; Anand, S. S.; and Soni, P. K. 2021. Natural language processing (NLP) based text summarization-a survey. In *2021 6th International Conference on Inventive Computation Technologies (ICICT)*, 1310–1317. IEEE.

Chandra, A. K. 1988. Theory of database queries. In *Proceedings of the seventh ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, 1–9.

Chen, C.; and Leitch, A. 2024. LLMs as Academic Reading Companions: Extending HCI Through Synthetic Personae. *arXiv preprint arXiv:2403.19506*.

Dettmers, T.; Pagnoni, A.; Holtzman, A.; and Zettlemoyer, L. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Dong, G.; Yuan, H.; Lu, K.; Li, C.; Xue, M.; Liu, D.; Wang, W.; Yuan, Z.; Zhou, C.; and Zhou, J. 2023. How abilities in large language models are affected by supervised fine-tuning data composition. *arXiv preprint arXiv:2310.05492*.

Gendron, G.; Bao, Q.; Witbrock, M.; and Dobbie, G. 2023. Large language models are not strong abstract reasoners. *arXiv preprint arXiv:2305.19555*.

Han, Z.; Gao, C.; Liu, J.; Zhang, S. Q.; et al. 2024. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608*.

Hessel, J.; Marasović, A.; Hwang, J. D.; Lee, L.; Da, J.; Zellers, R.; Mankoff, R.; and Choi, Y. 2022. Do androids laugh at electric sheep? humor" understanding" benchmarks from the new yorker caption contest. *arXiv preprint arXiv:2209.06293*.

Hogan, A.; Blomqvist, E.; Cochez, M.; d'Amato, C.; Melo, G. D.; Gutierrez, C.; Kirrane, S.; Gayo, J. E. L.; Navigli, R.; Neumaier, S.; et al. 2021. Knowledge graphs. *ACM Computing Surveys (Csur)*, 54(4): 1–37.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Hu, Y.; Song, K.; Cho, S.; Wang, X.; Foroosh, H.; Yu, D.; and Liu, F. 2024. SportsMetrics: Blending Text and Numerical Data to Understand Information Fusion in LLMs. *arXiv preprint arXiv:2402.10979*.

Jiang, A. Q.; Sablayrolles, A.; Roux, A.; Mensch, A.; Savary, B.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Hanna, E. B.; Bressand, F.; et al. 2024a. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Jiang, J.; Wang, F.; Shen, J.; Kim, S.; and Kim, S. 2024b. A Survey on Large Language Models for Code Generation. *arXiv preprint arXiv:2406.00515*.

Joko, H.; Chatterjee, S.; Ramsay, A.; de Vries, A. P.; Dalton, J.; and Hasibi, F. 2024. Doing Personal LAPS: LLM-Augmented Dialogue Construction for Personalized Multi-Session Conversational Search. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 796–806.

Kaelbling, L. P.; Littman, M. L.; and Moore, A. W. 1996. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4: 237–285.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Krause, S.; and Stolzenburg, F. 2024. From Data to Commonsense Reasoning: The Use of Large Language Models for Explainable AI. *arXiv preprint arXiv:2407.03778*.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33: 9459–9474.

Li, R.; and Gao, Y. 2024. Anchored Answers: Unravelling Positional Bias in GPT-2's Multiple-Choice Questions. *arXiv preprint arXiv:2405.03205*.

Makridakis, S. 1993. Accuracy measures: theoretical and practical concerns. *International journal of forecasting*, 9(4): 527–529.

Meta, A. 2024. Introducing meta llama 3: The most capable openly available llm to date, 2024. *URL https://ai. meta. com/blog/meta-llama-3/. Accessed on April*, 26.

Pezeshkpour, P.; and Hruschka, E. 2023. Large language models sensitivity to the order of options in multiple-choice questions. *arXiv preprint arXiv:2308.11483*.

Salvador, J.; Bansal, N.; Akter, M.; Sarkar, S.; Das, A.; and Karmaker, S. K. 2024. Benchmarking LLMs on the Semantic Overlap Summarization Task. *arXiv preprint arXiv:2402.17008*.

Säuberli, A.; and Clematide, S. 2024. Automatic generation and evaluation of reading comprehension test items with large language models. *arXiv preprint arXiv:2404.07720*.

Savelka, J.; Agarwal, A.; An, M.; Bogart, C.; and Sakr, M. 2023. Thrilled by your progress! Large language models (GPT-4) no longer struggle to pass assessments in higher education programming courses. In *Proceedings of the 2023 ACM Conference on International Computing Education Research-Volume 1*, 78–92.

Singh, S. 2018. Natural language processing for information extraction. *arXiv preprint arXiv:1807.02383*.

Spitzer, F. 1956. A combinatorial lemma and its application to probability theory. *Transactions of the American Mathematical Society*, 82(2): 323–339.

Stechly, K.; Marquez, M.; and Kambhampati, S. 2023. Gpt-4 doesn't know it's wrong: An analysis of iterative prompting for reasoning problems. *arXiv preprint arXiv:2310.12397*.

Team, G.; Anil, R.; Borgeaud, S.; Wu, Y.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Team, G.; Mesnard, T.; Hardin, C.; Dadashi, R.; Bhupatiraju, S.; Pathak, S.; Sifre, L.; Rivière, M.; Kale, M. S.; Love,

J.; et al. 2024a. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Team, M. R.; et al. 2024b. Introducing dbrx: A new state-of-the-art open llm, 2024. *URL https://www. databricks. com/blog/introducing-dbrx-new-state-art-open-llm. Accessed on April*, 26.

Tunstall, L.; Beeching, E.; Lambert, N.; Rajani, N.; Rasul, K.; Belkada, Y.; Huang, S.; von Werra, L.; Fourrier, C.; Habib, N.; Sarrazin, N.; Sanseviero, O.; Rush, A. M.; and Wolf, T. 2023. Zephyr: Direct Distillation of LM Alignment. arXiv:2310.16944.

Wang, J.; and Chen, Y. 2023. A Review on Code Generation with LLMs: Application and Evaluation. In *2023 IEEE International Conference on Medical Artificial Intelligence (MedAI)*, 284–289. IEEE.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Xu, Q.; Qu, L.; Xu, C.; and Cui, R. 2019. Privacy-aware text rewriting. In *Proceedings of the 12th International Conference on Natural Language Generation*, 247–257.

Yi, Z.; Ouyang, J.; Liu, Y.; Liao, T.; Xu, Z.; and Shen, Y. 2024. A Survey on Recent Advances in LLM-Based Multi-turn Dialogue Systems. *arXiv preprint arXiv:2402.18013*.

Zhao, Z.; Lee, W. S.; and Hsu, D. 2024. Large language models as commonsense knowledge for large-scale task planning. *Advances in Neural Information Processing Systems*, 36.

Zheng, C.; Zhou, H.; Meng, F.; Zhou, J.; and Huang, M. 2023. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*.

Zhu, Y.; Moniz, J. R. A.; Bhargava, S.; Lu, J.; Piraviperumal, D.; Li, S.; Zhang, Y.; Yu, H.; and Tseng, B.-H. 2024. Can Large Language Models Understand Context? *arXiv preprint arXiv:2402.00858*.