# Large Language Model and their application in treating ADHD

**Somansh Budhwar**
Halıcıoğlu Data Science Institute (HDSI)
University of California San Diego
sbudhwar@ucsd.edu

**Akansha Lalwani**
Halıcıoğlu Data Science Institute (HDSI)
University of California San Diego
alalwani@ucsd.edu

**Quynh Le**
Halıcıoğlu Data Science Institute (HDSI)
University of California San Diego
q3le@ucsd.edu

**Charles Ye**
Department of Computer Science and Engineering
University of California San Diego
juy022@ucsd.edu

## 1 Introduction

Attention Deficit Hyperactivity Disorder (ADHD) is a neurodevelopmental condition characterized by time blindness, lack of self-inhibition, restricted working memory, and impaired executive function, leading to challenges in planning, decision-making, and daily functioning [1]. While conventional strategies such as structuring the day, evaluating pros and cons, time-budgeting tasks, and working backward from goals are recommended for managing ADHD, emerging technologies offer new avenues for support. Despite the inherent capabilities of LLMs, such as defining activity structures, evaluating scenarios, generating diverse problem-solving ideas, and serving as a working memory for complex scenarios, they are not explicitly designed to address ADHD challenges. In this paper, we explore the potential of applying Large Language Models (LLMs) to assist individuals with ADHD and investigate three distinct approaches to enhance their effectiveness. The first approach involves rule-based prompt engineering, where we tailor prompts before users input queries, optimizing the LLM's responses for ADHD-specific assistance. This includes chain-of-thought reasoning and prompt engineering. The second approach, prompt tuning, aims to design an augmented layer that retrieves domain specific knowledge and augment the prompts by integrating relevant and scientifically supported literature on ADHD. This approach seeks to improve the quality of answers generated by the LLM by incorporating a knowledge base derived from current research and best practices. The third approach involves fine-tuning the LLM on ADHD literature directly, and customizing the model to provide solutions aligned with the specific needs of individuals with ADHD.

We then evaluate the performace of the LLMs after each alteration and compare it to the performance of ChatGPT on initial prompts. The three main criteria include safety, relevance and effectiveness of the response. This was we capture if the model outputs address the time blindness, forgetfulness and information processing challenges of an ADHD individual. By critically examining these approaches, we aim to contribute valuable insights into the potential applications of LLMs for ADHD support and provide guidance on optimizing their performance through innovative methodologies. This research lays the foundation for future developments in leveraging advanced language models to address the unique cognitive challenges faced by individuals with ADHD.

## 2 Related Work

Although the field of LLMs has become mainstream in recent years, applying LLMs to domain-specific tasks is still an open challenge. However, there are certain approaches that are gaining steam. Literature that discusses the application of LLMs in a specific domain is as follows.

[10] investigated the sensitivity of GPT-3 to in-context examples. [11] showed that few-shot prompts suffer from order sensitivity and proposed efficient prompt ordering for text classification. [12] found that Prompt-based models often learn equally fast with misleading and irrelevant templates as they do with instructive ones. [13] showed that by adding reasoning steps to a few-shot prompts, LLM can perform better on reasoning tasks than task-specific fine-tuned models. [7] showed that besides exemplars, reasoning instruction also matters in prompts.

A recent paper surveyed major techniques to make LLMs domain specific[9]. it classified approaches into 3 main classes. First is the "External Augmentation" or "Black Box" approach where the query is not only passed to the LLM but also to a retriever model that looks at domain knowledge to extract the relevant information. The original query and the output of the retriever are then concatenated and passed to the LLM to generate an output. Specifically, introduces a dynamic framework[6] that actively guides the retrieval and generation process to accommodate the unique challenges faced by individuals with ADHD, promoting a more accessible and supportive platform. The second method is called "Prompt crafting". For instance, discrete prompts are entered manually into the query. In this paper, we seek to apply few-shot prompt crafting whereby each query is augmented by a few prompts relevant to ADHD and creates a chain of thought so that the model output is restricted to this certain domain.

The third approach is to change the model itself. One way can be to augment the model with a layer that feeds in not only the user prompt but also a context from a domain-specific database. This is the Retrieval Augmented Generation method proposed in 2021[8]. It allows the model to understand the prompt with factual context from the retrieved database. Another way can be done by either training the LLM on domain-specific literature or by creating Adapter-based fine-tuning where certain layers are added in the model that restrict the output to the domain we desire. This requires access to the model and is thus not applicable to services like ChatGPT. To conduct fine-tuning of a large language model is computationally expensive, so an efficient method for fine-tuning exists whereby the LLM is quantized to weights of 4-bit floats and frozen, and the Low-Rank Adapter layers are trained on the new prompts[3]. In other words, the original model weights are frozen while a small set of trainable parameters are added to the model to obtain specific results.

Using Large Language Models like ChatGPT to assist people with ADHD is an area of ongoing research. People with ADHD have found various use cases to help with their day-to-day activities using GPT [4]. Some of those use cases are:

1. Task Management[2]: ChatGPT has been utilized as a task manager where it helps in organizing and prioritizing to-do lists, breaking down tasks into manageable steps, and providing time estimates for each step. Such as ReAct[14] proposed by S. Yao Moreover, it suggests approaches to enhance productivity and manage time more effectively.
2. Memory Aid: By logging and labeling past conversations, ChatGPT serves as a memory aid, allowing users to refer back to previous interactions as notes.
3. Job Search: LLMs can help accelerate the process of applying for jobs. They can provide very specific cover letters and resumes based on the job description. This is especially useful for people with ADHD as they struggle to maintain focus on such mundane tasks.
4. Educational Assistance: LLMs can explain a concept in different ways which can be quite useful. Prompts like ELI5 (Explain Like I am 5) can be useful to get a high-level overview of the topic in a short amount of time and then topics can be deep-dived based on the requirements.
5. Financial Planning: ChatGPT assists in financial planning by creating budgets based on input regarding monthly expenses, income, and financial goals. It also suggests ways to modify spending patterns to achieve financial objectives.

College Students with ADHD are usually said to have poor working memories [5]. LLMs can really help with providing a second brain that can store all the essential data, freeing the ADHD mind for more creative thinking

The myriad ways in which ChatGPT has been employed to assist individuals with ADHD as per the ADDitude article, showcase the potential of large language models in augmenting support for individuals facing challenges in executive functions due to ADHD. However, the limitations in terms of outdated or inaccurate information highlight the importance of continuous development

and possibly the integration of real-time information updating mechanisms to make such tools more reliable and effective in providing assistance.

# 3 Proposed Methods

## 3.1 Chain of Thoughts Prompting

### 3.1.1 Zero shot

Tl;dr: Add “A: The answer is “ in the prompt.
In the context of question answering, a zero-shot approach involves formulating questions in a way that anticipates the answer within the prompt itself, often through the addition of phrases like "A: The answer is." This technique leverages the model’s ability to comprehend and respond to questions without specific task-related training.

### 3.1.2 Zero shot CoT

Tl;dr: Add: “A: Let’s think step by step.” in the prompt.
Building upon the zero-shot paradigm, Zero Shot Chain of Thought (Zero Shot CoT) introduces a methodology where questions are framed to encourage a step-by-step thought process. This is achieved by incorporating prompts such as "A: Let’s think step by step." These prompts guide the model in constructing reasoning chains, promoting a more elaborate and detailed response, even in the absence of task-specific training data.

### 3.1.3 Manual CoT

Tl;dr: Manual-CoT achieves stronger performance by eliciting the CoT reasoning ability with effective manual demonstrations. The demonstrations for the reasoning process are manually designed.
Manual Chain of Thought (Manual CoT) takes a different approach, emphasizing human intervention to enhance the model’s reasoning capabilities. Achieving stronger performance, Manual CoT involves the creation of effective manual demonstrations that illustrate the reasoning process. By manually designing demonstrations that showcase the desired chain of thought, this method aims to guide the model toward more accurate and contextually relevant responses.

### 3.1.4 Auto CoT

Tl;dr: Auto CoT samples questions with diversity and generates reasoning chains to construct demonstrations.
Auto Chain of Thought (Auto CoT) represents an automated approach to generating reasoning chains for question-answering tasks. This method introduces diversity by sampling questions and constructing reasoning chains to form demonstrations. Auto CoT aims to capture a wide range of possible thought processes, enabling the model to adapt and generate coherent responses across various scenarios without relying on extensive manual intervention.

## 3.2 Retrieval Augmented Generation

### 3.2.1 Basic Retrieval Augmented Generation

To make our model ethically responsible and safe towards ADHD users, we wish to generate content that is not only relevant but also based on best practices of the field and current research. To this end we augment our mode with ADHD literature. After gathering the data, in text format, we break down all documents into chunks of 1000 words. This collection of chunks is then embedded as vectors and stored into a vector database for later retrieval. Now, when a user enters a prompt, it is first converted into an embedding, and a similarity search is conducted with the vector database, and top-k documents similar to the query are retrieved. The retrieved document embeddings are then converted to text, and concatenated as context with the original prompt. This augmented input is sent to the LLM and obtains a literature-relevant answer.

### 3.2.2 Active Retrieval Augmented Generation

To address the unique challenges associated with ADHD in the context of long-form generation, we advocate for the integration of active retrieval augmented generation. It is a generic framework that actively decides when and what to retrieve through the generation process, resulting in the interleaving of retrieval and generation. Individuals with ADHD often face difficulties in maintaining focus and organizing thoughts, making traditional long-form generation processes challenging for them. The proposed framework offers a tailored solution by actively guiding the retrieval and generation process. By dynamically deciding when and what to retrieve, the framework accommodates the fluctuating attention spans of individuals with ADHD, promoting a more structured and step-by-step approach. The interleaving of retrieval and generation not only supports the cognitive needs of those with ADHD but also ensures that the information flow remains engaging and conducive to substained attention. This approach aligns with the cognitive characteristics of ADHD, providing a more accessible and supportive platform for individuals with ADHD to participate in and benefit from the long-form generation process.

In the context of a downstream task, we organize the prompt by placing search-related instructions and exemplars at the outset as skill 1, followed by instructions and exemplars specific to the downstream task as skill 2. The objective is to prompt language models to integrate both skills when generating search queries during task execution. The prompt structure is illustrated in Prompt 3.2.2

**Prompt 3.2.2: retrieval instructions**

Skill 1. An instruction to guide LMs to generate search queries.

Several search-related exemplars.

Skill 2. An instruction to guide LMs to perform a specific downstream task (e.g. multi hop QA).

Several task-related exemplars.

An instruction to guilde Los to combine skills 1 and 2 for the test case.

The input of the test case.

## 3.3 Finetuning

For Fine-tuning the model, we use the QLoRA method[3] whereby a small set of trainable parameters are added to the quantized Llama-7b model. The prompts are in the following format "<User query> | <Response> <Sub-divided tasks> <Time Budget for sub-tasks> <Best Practices>". For instance, one prompt may look like as follows.

- User:
  - Help me make a Pizza.
- Agent:
  - Buy pizza base, sauce, cheese and tomato topping (shopping time - 1hr)
  - Heat the oven to 450F (30 mins)
  - Add sauce, cheese and toppings on top of the base (5 mins)
  - Wait for pizza to cook (30 mins)
  - Carefully take out the pizza using mittens and let it cool (5 mins)
  - Turn off the oven (1 min)

## 3.4 Prompt Tuning

Prompt tuning is a technique used to adjust the propensity of language models, such as GPT-3.5 and Llama, to generate responses that align more closely with a specific context or set of constraints. In

this context, prompt tuning involves the creation of prompts that are tailored to cater to executive functioning challenges, such as sustaining attention, organizing tasks, and regulating emotions.

For individuals with ADHD, effectively designed prompts can make the difference between a generic response and one that provides concrete, actionable advice. With GPT-3.5 and Llama, we have the opportunity to refine these prompts based on a deep understanding of their cognitive aspects. By incorporating strategies from ADHD management literature—like breaking tasks into subtasks, providing clear time management cues, and structuring prompts in a way that guides attention.

We expect the outputs of prompt tuning to be more practical, structured, and conducive to the needs of our target individuals.

# 4 Experiments

## 4.1 Chain of Thoughts Prompting

In this experiment, we use Auto CoT to implement Chain-of-Thoughts prompting. We collect a sample of 40 questions on the internet, which are used to identify whether a person has ADHD or not. We then experimented with different Chain-of-thought methods to qualitatively evaluate whether these methods would affect LLM answers, which could shed light on how LLM Q&A might be applied in having conversations with ADHD patients.

## 4.2 RAG: Retrieval Augmented Generation

We use the tools Langchain, Llama2-7b-chat model, MongoDB, and FAISS vector db to conduct this experiment. We collect the data relevant to ADHD. It includes 25 high-quality research papers on coping with ADHD, 25 online articles with best practices for ADHD and 50 books on ADHD covering strategies for managing education, parenting, work and marriage for people with ADHD. All documents are converter to text chunks and each chunk is stored in the MongoDB database.

We choose top 4 documents after retrieval and use the "stuff" pipeline in the Langchain module to create the custom prompt. Due to computational limitations, other methods such as "refine" and "map-reduce" could not be implemented.

Next, we evaluate the results with 30 prompts that cover topics such as cooking, education, socialization, workspace organization, marriage and so on. The results are then evaluated on three metrics, namely, safety, relevance, and effectiveness on a scale of 1-10. Safety metric implies that the model output suggests solutions that are not harmful to the user, and follow current best practices rather than stereotypical answers to the problem. Essentially it ensures the safety of the user and is considerate of the user's condition. Relevance implies that the answer should be relevant to the prompt and actually suggest solutions relevant to the task. Effectiveness measures if the solution includes strategies relevant to ADHD such as sub-tasks, time-budgeting, visual aids and so on.

## 4.3 FLARE: Forward-Looking Active REtrieval Augmented Generation

We validate our method on one of the most advanced GPT-3.5 and LLaMA LMs by iteratively querying their API. Since we focus on the integration of retrieval and generation. We can offer a hypothetical scenario of how a forward-looking active retrieval and generation model could be employed to enhance prompt tuning for ADHD. To leverage FLARE for prompt tuning in ADHD support, the model is designed to anticipate the specific needs and challenges of individuals with ADHD during the generation process. The following steps illustrate how FLARE could be used to retrieve information and generate tuned prompts. Before the user input is provided, FLARE actively retrieves relevant information from a database of ADHD-specific literature and guidance. Based on anticipating ADHD-specific challenges, FLARE dynamically generates tuned prompts that are specifically tailored to address the unique needs of individuals with ADHD.

**Multi-tasking Management:**

- *General Prompt:*"Explain how the individual can efficiently handle multitasking in a manner that minimizes stress."
  *Tuned Prompt:* "Developing a step-by-step plan for individuals with ADHD to manage

multiple tasks without feeling overwhelmed, including time estimates for each step and strategies to sustain concentration."

**Achieve Academic Success:**

- *General Prompt:* "Describe the strategies an individual could employ to attain academic success despite tendencies towards underachievement."
  *Tuned Prompt:* "Provide a plan for enhancing the academic performance of a student with ADHD through the establishment of specific, measurable goals, the creation of a structured study schedule, and the utilization of visual aids for progress tracking."

### 4.4 Fine-tuning

We created 50 prompts for fine-tuning the LLM with QLoRA. We augment this dataset by using CHatGPT to rephrase the prompts and get 200 prompts. The 200 prompts are divided into a batch size of 8 and used for QLoRA fine-tuning. Then we input the same set of prompts used in the RAG experiment to evaluate the performance of the LLaMA model in terms of relevance, effectiveness, and time budgeting. Note that these ratings are to be given by human evaluators on a scale of 1-10. We then evaluate the results with the outputs from ChatGPT on the same prompts and compare them with the LLaMA model.

### 4.5 Prompt Tuning

To empirically test the efficacy of prompt tuning, we carried out a series of experiments with two state-of-the-art models, GPT-3.5 and Llama. Our methodology involved the development of two sets of prompts for each model: a general set and a tuned set specifically designed for ADHD support.

For instance, consider the following prompts:

1. *General Prompt:* "Describe how the individual could manage multitasking in a way that is efficient and reduces stress."
   *Tuned Prompt:* "Outline a step-by-step approach for an ADHD individual to handle multiple tasks without feeling overwhelmed, including time estimates for each step and strategies for maintaining focus."

2. *General Prompt:* "Explain how the individual might strategize to achieve academic success despite underachievement tendencies."
   *Tuned Prompt:* "Detail a plan for an ADHD student to improve academic performance by setting specific, measurable goals, establishing a structured study schedule, and using visual aids to track progress."

The aim was to see if the tailored prompts would lead to responses that offer more precise and actionable advice for those with ADHD, as opposed to the more generic responses from the standard prompts. The responses were then analyzed for their practicality, specificity, and alignment with effective ADHD management strategies.

### 4.6 Evaluation

| Model | Safety | Relevance | Effectiveness | Sub-tasks | Time Budget |
|---|---|---|---|---|---|
| ChatGPT | - | - | - | - | - |
| Llama-7b | - | - | - | - | - |

Table 1: Model Evaluation

## 5 Plan of activities

### 5.1 Old Activities Plan

Deliverables by the Midway Report: Our initial focus will be on gathering relevant data and literature on ADHD. We will be exploring few-shot learning across our dataset using a selection of Large

Language Models (LLMs) including Flan T5, GPT 3.5, and Llama. We will explore and assess the impact of techniques such as the Chain of Thought (CoT) on model performance, and to understand the correlation between the model size and its efficacy. We will also settle on the evaluation metrics that are suitable for this project since the output is subjective.

Following this, we intend to delve into few-shot prompting techniques and model fine-tuning to gauge the extent of performance enhancement these methods can offer. The challenges we anticipate are: the quality of our data; computing challenges given the size and complexity of LLMs; and conducting evaluation of different approaches in an unbiased manner.

Akansha will be investigating the capabilities of Llama, focusing on its utility in zero-shot learning and chain of thought prompting. Similarly, Somansh, Quynh and Charles will be exploring the capabilities of Flan T5, GPT 3.5 and Claude respectively.

#### 5.1.1 Rule-Based Prompt Engineering - Due Date: Nov 9

Prompt Effectiveness: Measure the performance of rule-based prompts by assessing their ability to elicit relevant and helpful response from the LLM.

#### 5.1.2 Prompt Tuning with Custom Embeddings - Due Date: Nov 12

Embedding quality: Access the quality of custom embedding by measuring semantic similarity to ADHD literature and their impact on LLM performance.

#### 5.1.3 Training the LLM with ADHD Literature - Due Date: Nov 16

Response Relevance and Accuracy: Compare the relevance and accuracy of LLM responses before and after fine-tuning with ADHD literature. Use metrics like precision, recall and F1 score.

### 5.2 New Activities Plan

Somansh will fine-tune the Llama model by December 1, and conduct experiments using the RAG - Basic model. Also, Somansh will create prompts and augmented prompts totaling 200 for the experiments on fine-tuning and RAG models. Finally, he will add more relevant literature, articles and books on ADHD that focus on education, work, socialization, daily tasks and marriage for the RAG model.

Charles will conduct experiments using the FALRE - Active Augmented Retrieval model by Dec 5. Also he will try the Browser-enhanced LMs (WebGPT) that can enhance factuality using reinforcement learning/supervised training where multiple queries can be triggered before generation. FLARE is built on text-based retrievers but can combined with a browser to potentially improve retrieval quality instead of being constrained within limited resources.

Akansha will focus on prompt engineering and tuning using GPT-3.5 and Llama, aiming to complete her tasks by December 1st. Her work includes manually creating and adjusting prompts to improve how these models respond, especially in situations related to cognitive challenges. She will also compare the performance of GPT-3.5 and Llama with these custom prompts, ensuring they are as helpful and relevant as possible. Additionally, she will explore adding prompts manually to increase the efficiency of our models.

## 6 Conclusion

In this study, we have embarked on an exploration of optimizing Large Language Models (LLMs) like GPT-3.5 and Llama-7b for assisting individuals with ADHD. Through various methods such as prompt tuning, rule-based engineering, and fine-tuning with ADHD-focused literature, we aim to tailor these models to address specific cognitive challenges. We anticipate gaining deeper insights into the potential of LLMs in supporting neurodevelopmental conditions and contributing to the evolving field of AI-driven cognitive assistance.

As we progress, our focus will be on refining these methods further and conducting more comprehensive experiments to validate these initial findings. The final phase of our project will involve a

thorough analysis of the effectiveness of each approach - Chain of Thoughts Prompting, Retrieval Augmented Generation, and Prompt Tuning - in delivering accurate, relevant, and safe responses for ADHD-specific scenarios.

## References


[1] Russell Barkley. Behavioral inhibition, sustained attention, and executive functions: Constructing a unifying theory of adhd. *Psychological bulletin*, 121:65–94, 01 1997.

[2] Liting Chen, Lu Wang, Hang Dong, Yali Du, Jie Yan, Fangkai Yang, Shuang Li, Pu Zhao, Si Qin, Saravan Rajmohan, et al. Introspective tips: Large language model for in-context decision making. *arXiv preprint arXiv:2305.11598*, 2023.

[3] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms, 2023.

[4] Carole Fleck. Ai for adhd: How to make chatgpt work for you. *ADDitude*, 2023.

[5] Rachel J Gropper and Rosemary Tannock. A pilot study of working memory and academic achievement in college students with adhd. *Journal of Attention Disorders*, 12(6):574–581, 2009.

[6] Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. *arXiv preprint arXiv:2305.06983*, 2023.

[7] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners, 2023.

[8] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning, 2021.

[9] Chen Ling, Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, Tanmoy Chowdhury, Yun Li, Hejie Cui, Xuchao Zhang, Tianjiao Zhao, Amit Panalkar, Wei Cheng, Haoyu Wang, Yanchi Liu, Zhengzhang Chen, Haifeng Chen, Chris White, Quanquan Gu, Jian Pei, and Liang Zhao. Domain specialization as the key to make large language models disruptive: A comprehensive survey, 2023.

[10] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for gpt-3?, 2021.

[11] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity, 2022.

[12] Albert Webson and Ellie Pavlick. Do prompt-based models really understand the meaning of their prompts?, 2022.

[13] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

[14] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. 2022.