Artificial Intelligence Assignment 01

Ammar Jamil (01-134231-010)
Department of Computer Science
Bahria University, Islamabad

I. Overview and Contextual Analysis

I choose ChatGPT as my AI application that is a conversational AI system developed by OpenAI. Its main purpose is to generate meaningful responses to human queries based on patterns it has learned from large-scale training data. It can assist in coding, summarizing text, answering domain-related questions, and producing coherent written content.

II. History

ChatGPT builds on the transformer architecture [1]. The development of GPT models started with GPT-1 in 2018, which showed that training on huge amounts of text before fine-tuning could help in language tasks. It was still small in size, so the answers were not always very clear or complete. In 2019, GPT-2 came with 1.5 billion parameters, and this made it much better at writing longer and more connected text. At the same time, it also created concerns because the model sometimes produced biased or misleading responses, which raised questions about how safe these systems really were.

A big step came in 2020 with GPT-3, which had 175 billion parameters and was able to do many tasks like translation, summarization, and reasoning [2].Still, it often made mistakes or gave wrong facts. To make the model more useful, OpenAI introduced InstructGPT in 2022, which used human feedback to help it follow instructions better. This became the base for ChatGPT, launched in late 2022, which quickly became popular because of its easy conversational style. Later, GPT-4 (2023) improved reasoning and added the ability to work with images as well as text. The most recent version, GPT-5 (2025), works as a "system of models" and is much stronger in coding and problem-solving, but even now it is not perfect and still faces issues like bias and factual errors.

III. Strong AI vs. Weak AI

Artificial intelligence can generally be divided into two broad categories: **Weak AI** and **Strong AI**. Weak AI, also called narrow AI, is designed to perform specific tasks very efficiently but without real understanding or awareness. Everyday examples include voice assistants such as Siri or Alexa, which can answer questions, set reminders, or play music, but cannot reason outside of their programmed abilities. Similarly, ChatGPT falls into this category because, while it can generate essays, solve coding problems, or answer academic questions, it is still only predicting the next likely word based on training data rather than actually "understanding" the problem like a human would.

Strong AI, in contrast, refers to a system that possesses general intelligence similar to human beings. Such a system would not only perform a wide range of tasks but also demonstrate self-awareness, independent reasoning, and adaptability across completely new domains. For example, if a strong AI were created, it could learn a new language, invent original theories, or even apply common sense to real-world problems without needing explicit training data. Currently, no AI model—not even the most advanced ones like GPT-5 has reached this stage.

Even with these fast improvements, ChatGPT is a type of weak AI. They are very advanced in making human-like text and solving many tasks, but they do not truly understand language or have human-level intelligence. Their skills are based on patterns learned from input data, not real thinking or awareness. Strong AI, on the other hand, would need general intelligence across all areas, something no current system, not even GPT-5, has reached.

IV. Evaluation Using the Four AI Approaches

A. Acting Humanly

The *Acting Humanly* approach suggests that an AI system can be considered intelligent if it behaves in ways similar to humans. ChatGPT illustrates this idea through its conversational style, which often resembles that of a human tutor. For instance, if a student asks ChatGPT to write a simple Python program to reverse a string named "Ammar," the system responds with the following code:

```
def rev(s):
    return s[::-1]

print(rev("Ammar"))
```

In addition to providing the code, ChatGPT explains that the expression `s[::-1]` is a slicing operation in Python that reverses the string. This behavior goes beyond simply giving the correct solution, as it mimics how a teacher would guide a learner step by step. Such interaction reflects the Acting Humanly approach because the AI communicates in a way that feels supportive and educational.

Furthermore, ChatGPT adapts its explanations depending on the user's background knowledge. For a beginner, it may expand on the details of string slicing, while for an advanced user, it may provide a shorter, more concise explanation. This ability to adjust responses according to context strengthens the impression that the system is engaging in human-like

behavior, even though it does not possess real understanding or awareness like a human instructor.

B. Thinking Humanly

The Thinking Humanly approach asks whether an AI system appears to follow cognitive steps similar to a human when solving problems. ChatGPT often demonstrates this behavior during debugging and step-by-step problem solving. Consider the short Python example:

```
nums = [1, 2, 3]
print(nums[3])
```

This code raises an IndexError because Python lists use zero-based indexing; the valid indices are `0`, `1`, and `2`. A human programmer would notice that the code attempts to access a fourth element that does not exist. ChatGPT identifies the same issue, explains that the last valid index is `2`, and suggests a corrected statement such as:

```
nums = [1, 2, 3]
print(nums[2])
```

In addition to the direct fix, ChatGPT can suggest diagnostic steps a human might take, for example checking the list length with `len(nums)`, printing the list contents, or adding guard code. Example alternatives include a bounds check:

```
index = 3
if index < len(nums):
    print(nums[index])
else:
    print("Index out of range")
```

or exception handling:

```
try:
    print(nums[3])
except IndexError:
    print("Index out of range")
```

These responses mimic the sequential reasoning and corrective actions a programmer would use, which is why ChatGPT can feel as if it is "thinking" like a human. However, this behavior is produced by pattern matching on training data rather than by actual understanding: the model cannot execute the code in the current session, and its suggested fixes should be tested by the developer. As a result, ChatGPT is a useful assistant for identifying common mistakes and proposing standard solutions, but its guidance should be validated by us in context to avoid ambiguous situations and overlooked edge cases.

C. Acting Rationally

The Acting Rationally approach suggests that an AI system is intelligent if it consistently chooses actions that best achieve a desired goal. ChatGPT demonstrates this through its problem-solving style, where it recommends efficient and effective solutions rather than unnecessarily complex ones. For instance, when asked to sort a list of numbers, ChatGPT suggests using Python's optimized built-in method instead of implementing a slower, manual algorithm:

```
numbers = [5, 2, 9, 1, 7]
numbers.sort()
print(numbers)  # [1, 2, 5, 7, 9]
```

This represents rational action because the built-in function is highly optimized forexample chatgpt uses the sort() function in python which Time complexity is O(nlogn) and significantly outperforms most handwritten sorting routines. By recommending this approach, ChatGPT helps users achieve their goal with minimal effort while ensuring performance and reliability.

Moreover, ChatGPT often provides reasoning about trade-offs. If a user explicitly requests a manual sorting algorithm, the AI may demonstrate a simple bubble sort for instructional purposes but also explain that it is inefficient for large datasets compared to algorithms such as merge sort or quicksort. In this way, the system balances educational clarity with performance awareness.

Such behavior illustrates the Acting Rationally perspective, as ChatGPT aims to select responses that maximize efficiency and effectiveness. While the model does not possess true awareness or reasoning, its ability to guide users toward practical, goal-oriented solutions reflects rational decision-making in practice.

D. Thinking Rationally

The Thinking Rationally approach focuses on whether an AI applies clear rules and logical steps when solving problems. ChatGPT demonstrates this ability in simple coding examples. For instance, when asked to check if a number is even or odd, it provides the following solution:

```
def check_even_odd(n):
    if n % 2 == 0:
        return "Even"
    else:
        return "Odd"

print(check_even_odd(7))
```

In this example, ChatGPT uses the rule that a number is even if it divides exactly by 2 and odd otherwise. The AI is not simply giving a direct answer but is following a logical process that can be applied to any number. This shows rational thinking because the method is systematic, reliable, and based on mathematical reasoning rather than random guessing. The Thinking Rationally approach emphasizes the importance of logical, step-by-step problem-solving. Recent research has introduced methods to enhance the reasoning capabilities of large language models, drawing from computational models of metareasoning in cognitive science [3].

V. Ethical and Societal Implications of the AI Application

One major ethical challenge of ChatGPT is bias in decision-making. Since the model is trained on large datasets collected

from the internet, it can sometimes reflect harmful stereotypes or produce biased answers [4]. For example, if asked about coding abilities of different groups, it might unintentionally favor one group over another because of patterns in its training data. This problem connects most directly to the Thinking Humanly approach. The system tries to mimic human thought processes by predicting what a person might say, but because it lacks real understanding, it may copy not just useful knowledge but also human biases. The design philosophy of imitating human-like thinking without true judgment means that any bias in the data can be reproduced in the system's outputs.

Another serious societal concern is the danger of using ChatGPT as a doctor or therapist. The system is not trained or certified to provide medical or psychological care, yet some users may rely on it for health advice or emotional support. This risk was highlighted by a tragic case in the United States, where a young boy reportedly took his own life after interacting with ChatGPT and following its harmful suggestions. This issue connects strongly to the Acting Humanly approach, since the system's human-like conversational style can make people believe it has genuine understanding and empathy. The design philosophy of making AI sound natural and supportive, while lacking real medical expertise, increases the danger of misuse in sensitive areas like healthcare and mental health.

VI. Creative Design: High-Level Architecture

Following are the six essential layers of ChatGPT architecture: [5]:

User Interface: The user interface is the point where people interact with ChatGPT, such as a chat window, website, or API. It collects the user's input and passes it into the system for further processing.

Input Processing: At this stage, the input text is converted into tokens, which are numerical representations of words or code. Tokenization makes it possible for the model to understand and handle natural language as well as programming languages.

Core Language Model: The core language model, built on the Transformer architecture, is responsible for analyzing the input and generating predictions. It uses patterns learned from massive datasets to produce context-aware and meaningful responses.

Reasoning and Safety Layer: The raw output from the language model passes through the reasoning and safety layer. This part of the system applies filters, safety checks, and human-aligned training methods like RLHF to reduce bias, harmful content, or irrelevant responses.

Output Generation: Once the system finalizes its response, the tokens are transformed back into human-readable text. This allows users to receive clear and natural answers, whether in the form of explanations, stories, or code.t

Feedback Loop: The feedback loop collects user reactions, such as approval, disapproval, or error reports. This feedback is later used by developers to fine-tune and retrain future versions, ensuring that the system continues to improve over time..

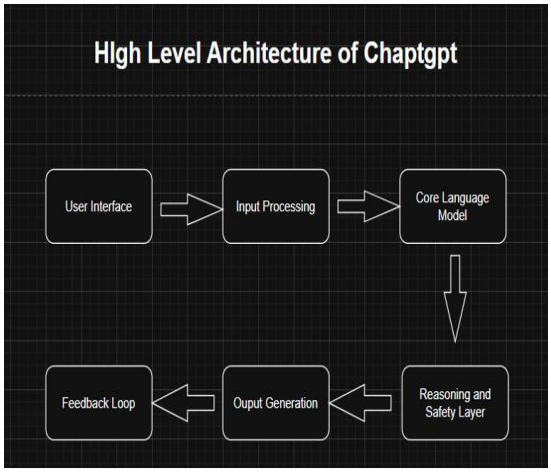

Fig. 1. High-level architecture of ChatGPT.

Figure 1 shows this architecture.

VII. Conclusion

The development of ChatGPT, from GPT-1 in 2018 to the more recent GPT-5 in 2025, demonstrates the rapid progress of natural language processing systems. Each generation has added significant capabilities, such as longer context handling, better reasoning, and multimodal input, while also raising new concerns about safety, fairness, and overreliance on automated systems. The evaluation of ChatGPT using the four classical AI approaches shows that it can act and think in ways that resemble human reasoning, but only within the limits of weak AI. It generates responses based on learned patterns rather than true understanding, which means it cannot yet be considered strong AI.

The high-level architecture of ChatGPT highlights how different layers work together, from input tokenization to safety filtering and the feedback loop. This design allows the system to provide useful, human-like responses while relying heavily on its training data and safety mechanisms. At the same time, the ethical and societal implications remain serious: biases in training data, potential misuse in sensitive areas like

healthcare, and risks that arise when users treat ChatGPT as an authority rather than a tool.

Overall, ChatGPT represents both the opportunities and limitations of current AI technology. It is a valuable assistant for education, coding, and knowledge sharing, but it also illustrates why human oversight is essential. Future work must focus not only on improving model accuracy and reasoning but also on reducing bias, enhancing transparency, and ensuring responsible deployment in real-world settings. By balancing technical performance with ethical safeguards, ChatGPT and similar systems can be integrated more safely and effectively into society.

REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30. Curran Associates, Inc., 2017, pp. 5998–6008.

[2] T. B. Brown *et al.*, "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[3] C. N. D. Sabbata, T. R. Sumers, B. AlKhamissi, A. Bosselut, and T. L. Griffiths, "Rational metareasoning for large language models," *arXiv preprint arXiv:2410.05563*, 2024. [Online]. Available: https://arxiv.org/abs/2410.05563

[4] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. ACM, 2021, pp. 610–623.

[5] OpenAI, "Introducing chatgpt," https://openai.com/blog/chatgpt, 2022, accessed: 2025-09-21.