Higher-Level Reasoning with Conversational AI

Saad Mufti
Georgia Institute of Technology
smufti3@gatech.edu

Courtney Young
Georgia Institute of Technology
cyoung315@gatech.edu

Saaliha Allauddin
Georgia Institute of Technology
sallauddin3@gatech.edu

1 Evaluation Insights

The development of conversational AI systems capable of higher-level reasoning represents a cornerstone challenge in artificial intelligence research, sitting at the intersection of natural language processing, cognitive science, and formal logic. While recent years have witnessed remarkable advances in language models' ability to process and generate human-like text, the implementation of genuine reasoning capabilities—defined as the ability to systematically analyze information, draw valid conclusions, and engage in abstract thinking—remains a complex challenge that requires significant theoretical and technical innovation. The difficulty lies not only in the computational complexity of implementing reasoning mechanisms but also in the fundamental challenge of operationalizing human-like reasoning patterns within artificial systems. Current state-of-the-art models, despite their impressive performance on many NLP tasks, often fail to demonstrate consistent logical reasoning, struggle with causal understanding, and exhibit limitations in their ability to generate novel insights or engage in genuine abstract thinking.

Higher-level reasoning in AI refers to the ability of a system to think abstractly, make inferences, and solve complex problems that require understanding beyond surface-level data processing. In Conversational AI, higher-level reasoning enables models to engage in more sophisticated dialogues, recognize implicit meaning, and handle tasks that require multi-step thinking or context awareness. This capability is crucial for applications that demand nuanced interpretation, such as decision-making in healthcare, providing educational guidance, or assisting in legal research.

Conversational AI has advanced significantly with the advent of large language models (LLMs), which can generate fluent, contextually relevant responses. However, many of these systems still rely on pattern matching and statistical associations rather than true reasoning. While models like GPT, LLaMA, and Mistral demonstrate impressive language generation abilities, they often fall short when required to reason through complex problems, understand nuanced human interactions, or maintain coherence over long conversations. The current landscape provides a robust foundation, yet limitations persist in how well AI can mimic human-like reasoning.

This case study aims to examine how higher-level reasoning can be enabled and improved within conversational AI systems. It will explore techniques and approaches used to foster reasoning capabilities, applications, datasets and evaluation metrics, examine the challenges involved, and highlight future directions and implications. By analyzing these factors, the study will identify both the potential and limitations of higher-level reasoning in Conversational AI, paving the way for further research and development in this field.

2 Background and Foundations

2.1 Definition of High-Level Reasoning

High-level reasoning in conversational AI refers to the cognitive processes that enable systems to draw meaningful conclusions, make inferences, and engage in complex problem-solving through natural language interaction. Drawing from cognitive science perspectives, this type of reasoning aligns with what Kahneman describes as ”System 2” - the slower, more deliberate thought processes involving conscious analysis and logical deduction, as opposed to fast, intuitive ”System 1” responses (Bengio, 2017; Weston and Sukhbaatar, 2023).

Following formal definitions from cognitive science and AI research, high-level reasoning encompasses several fundamental types. Deductive reasoning derives specific conclusions from general principles or premises, inductive reasoning draws general conclusions based on specific observations, and abductive reasoning generates the most plausible explanations for observations (Yu et al., 2023). These reasoning types often work in conjunction to enable complex problem-solving capabilities in AI systems.

2.2 Evolution from Rule-Based to Neural Approaches

The development of high-level reasoning in conversational AI has undergone several major paradigm shifts over the past decades:

1. Early Rule-Based Era (1960s-1980s) The initial approaches to implementing reasoning in AI systems relied heavily on symbolic logic and hand-crafted rules. These early systems employed formal logic frameworks and extensive knowledge bases to perform reasoning tasks (Reiter, 1975). While these systems could perform well in narrow domains with clearly defined rules, they were limited by their brittleness and inability to handle novel situations outside their programmed knowledge base.

2. Statistical Revolution (1990s-2000s) The field underwent a significant transformation with the introduction of statistical approaches to natural language processing and reasoning. This era saw the development of probabilistic graphical models and statistical learning methods that could better handle uncertainty and ambiguity in reasoning tasks (Berka, 2020). These approaches provided more flexibility than rule-based systems but still struggled with complex reasoning chains requiring multiple steps of inference.

3. Neural Network Renaissance (2010-2015) The resurgence of neural networks brought new capabilities to reasoning systems. Deep learning architectures enabled end-to-end training on large datasets, while word embeddings captured semantic relationships in ways that previous approaches could not (Manning, 2022). However, these early neural approaches still faced challenges in explicit logical reasoning and multi-step inference tasks.

4. Transformer Revolution (2017-2020) The introduction of the transformer architecture by Vaswani et al. (2017) marked a fundamental shift in neural language processing. The key innovation was the attention mechanism, which allows models to dynamically focus on relevant parts of the input when producing each element of the output. In mathematical terms, attention computes weighted sums over an input sequence, where the weights are learned functions of the query (the current processing state) and key-value pairs (the input elements). This can be expressed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (1)$$

where Q, K, and V represent the queries, keys, and values respectively, and d is the dimension of the key vectors (Vaswani et al., 2017). This mechanism enabled models to capture long-range dependencies and relationships in text far more effectively than previous recurrent or convolutional architectures.

The transformer architecture also introduced multi-head attention, allowing models to capture different types of relationships in parallel, and positional encoding to maintain sequence order information. These innovations led to significant improvements in language understanding and processing capabilities.

5. Foundation Model Era (2020-Present) Foundation models mark a major shift in AI, featuring large-scale models built on transformer architectures and trained on extensive datasets, making them adaptable across various tasks (Bommasani et al., 2021). These models act as implicit world models by encoding vast knowledge and reasoning patterns within their parameters, handling both declarative (facts) and procedural (reasoning) knowledge (Brown et al., 2020).

Their architecture typically includes deep transformer stacks, large token embeddings, layer normalization, and billions of parameters, facilitating complex pattern recognition. However, challenges remain, such as high computational demands, context consistency, knowledge verification, and the ”black box” nature of their reasoning processes.

Foundation models underpin most modern conversational AI, enhancing reasoning and dialogue capabilities, but understanding their architecture is key to recognizing both their potential and limitations and to advancing future systems.

This foundation model architecture has become the basis for most modern conversational AI systems, enabling them to engage in increasingly sophisticated forms of reasoning and dialogue. Understanding these architectural foundations is crucial for appreciating both the capabilities and limitations of current systems, as well as for developing improved approaches in the future.

3 Approaches to High-level Reasoning

Enabling high-level reasoning in conversational AI systems requires the integration of various techniques and approaches that go beyond basic language processing and generation. Here are some of the key approaches that have been explored and developed:

Knowledge Representation and Reasoning Conversational AI systems with high-level reasoning capabilities often rely on robust knowledge representation frameworks, such as ontologies, semantic networks, or knowledge graphs, to model and organize relevant information about the world, entities, relationships, and concepts (Sowa, 2000; Gruber, 1993). Logical reasoning techniques, including rule-based inference, first-order logic, and probabilistic reasoning, are then applied to this structured knowledge to derive inferences, make deductions, and draw conclusions that support high-level reasoning (Russell & Norvig, 2016; Baral, 2003).

Commonsense Reasoning Equipping conversational AI with commonsense reasoning abilities is crucial for understanding and reasoning about everyday situations, social norms, and implicit knowledge that humans typically take for granted (Davis & Marcus, 2015; Liu & Singh, 2004).Approaches to commonsense reasoning in conversational AI include the use of large-scale commonsense knowledge bases, such as ConceptNet or Cyc, and the development of reasoning mechanisms that can leverage this knowledge to make contextual inferences and draw plausible conclusions (Lenat, 1995; Speer et al., 2017).

Contextual Understanding High-level reasoning in conversational AI requires the ability to comprehend and reason about the context surrounding the conversation, including the user's intent, the current state of the dialogue, and the relevant background information. Techniques such as dialogue management, topic modeling, and user profiling can be employed to build a richer understanding of the conversational context and use it to inform high-level reasoning processes (Rieser & Lemon, 2011; Blei et al., 2003).

Hybrid Approaches Many modern conversational AI systems employ hybrid approaches that combine multiple techniques and approaches to high-level reasoning, leveraging the strengths of different methods and aiming to overcome their individual limitations (Laird, 2012; Mao et al., 2019). For example, a system may integrate symbolic reasoning based on knowledge representation with sub-symbolic techniques, such as deep learning, to achieve a more holistic and flexible high-level reasoning capability.

4 State of the Art Capabilities and Methods

Recent breakthroughs demonstrate remarkable reasoning capabilities in foundation models. AlphaGeometry, combining language models with formal theorem proving, successfully solved 25 geometry problems from International Mathematical Olympiad (IMO) competitions (Li et al., 2022g). Beyond mathematics, GPT-4 achieved an 86.7% score on the US Medical Licensing Examination (USMLE), exceeding the passing threshold by approximately 20 percentage points and outperforming the majority of human test-takers (Singhal et al., 2023). In scientific domains, Minerva demonstrated the ability to solve around one-third of university-level problems across chemistry, physics, biology, and other quantitative fields (Lewkowycz et al., 2022). These achievements suggest that foundation models have developed sophisticated reasoning capabilities approaching or exceeding human-level performance in specific domains.

4.1 Natural Language Understanding and Context

The capabilities modern conversational AI systems demonstrate are sophisticated semantic understanding through large pre-trained language models. These models develop robust semantic representations by training on extensive text corpora, enabling them to comprehend complex linguistic patterns and nuanced meanings (Brown et al., 2020). Recent work shows that models like GPT-4 can understand and reason about abstract concepts, demonstrating capabilities approaching human-level comprehension in many domains (Bubeck et al., 2023).

4.2 Chain of Thought and Self-Consistency

Chain of Thought (CoT) and self-consistency methods represent significant breakthroughs in improving the reasoning capabilities of foundation models. These approaches enable models to break down complex problems into intermediate steps and verify their solutions through multiple reasoning paths. By making the reasoning process explicit and leveraging multiple solution attempts, these methods have dramatically improved performance on tasks requiring complex logical reasoning, mathematical problem-solving, and multistep deduction (Wei et al., 2022b).

1. Zero-shot Chain of Thought Kojima et al. (2022) demonstrated that large language models can perform step-by-step reasoning without requiring explicit examples through simple prompting like "Let's think step by step." This discovery showed that sophisticated reasoning capabilities can emerge naturally in sufficiently large models without task-specific training.

2. Few-shot Chain of Thought Wei et al. (2022b) established that providing a small number of chain-of-thought examples significantly improves reasoning performance across various tasks. This approach combines demonstration learning with explicit reasoning steps, enabling models to learn and apply complex reasoning patterns more effectively.

3. Tree-of-thought Reasoning Yao et al. (2023b) introduced the tree-of-thought framework, enabling models to explore multiple reasoning paths simultaneously. This approach allows for systematic exploration of solution spaces, backtracking when necessary, and selecting optimal reasoning paths. The framework has shown particular success in complex problem-solving tasks requiring strategic thinking.

4. Graph-based Reasoning Graph-of-Thought (Yao et al., 2023d) extends traditional linear reasoning to graph-structured approaches, enabling more complex reasoning patterns. This framework has proven especially effective in tasks requiring multi-hop inference and relationship tracking, such as scientific reasoning and complex analytical tasks.

5. Step-by-step Decomposition Recent work by Zhang et al. (2022c) on automatic chain of thought prompting demonstrates how models can systematically break down complex problems into manageable components. This approach has been particularly successful in solving mathematical word problems and logical reasoning tasks that require multiple intermediate steps.

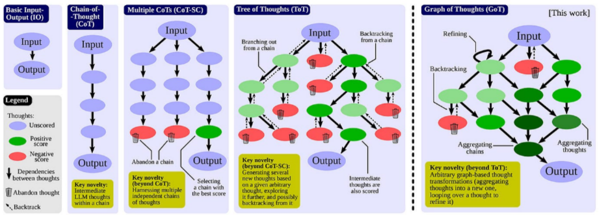

Basic Input-Output (IO)
Input
Output
Legend
Thoughts:
Unscored
Positive score
Negative score
Dependencies between thoughts
Abandon thought
Backtrack
Chain-of--Thought (CoT)
Input
Output
Key novelty: Intermediate LLM thoughts within a chain
Multiple CoTs (CoT-SC)
Input
Abandon a chain
Output
Selecting a chain with the best score
Key novelty (beyond CoT): Harnessing multiple independent chains of thoughts
Tree of Thoughts (ToT)
Input
Branching out from a chain
Backtracking from a chain
Output
Intermediate thoughts are also scored
Key novelty (beyond CoT-SC): Generating several new thoughts based on a given arbitrary thought, exploring it further, and possibly backtracking from it
Graph of Thoughts (GoT)
[This work]
Refining
Input
Backtracking
Aggregating chains
Aggregating thoughts
Output
Key novelty (beyond ToT): Arbitrary graph-based thought transformations (aggregating thoughts into a new one, looping over a thought to refine it)

Figure 1: Chain of Thought, Tree of Thoughts, and Graph of Thoughts (Raghu, M., 2023)

4.3 Mixture of Experts and Reasoning Specialization

The Mixture of Experts (MoE) architecture advances foundation models by training multiple specialized networks (experts) to process different input aspects. MoE traditionally routes tokens to specific experts, enhancing computational efficiency through selective expert activation (Lepikhin et al., 2020; Fedus et al., 2022). However, token-level routing may limit performance on complex reasoning tasks that require a holistic approach.

Mixture of Reasoning Experts (MoRE) MoRE improves on MoE by using separate, full-model experts dedicated to distinct reasoning tasks (e.g., factual, mathematical, multi-hop, commonsense), rather than token-based routing (Si et al., 2023).

Advantages of MoRE:

- **Task-Specific Specialization**: Experts are optimized for particular reasoning types, enhancing task performance.
- **Interpretability**: The system's decision-making becomes clearer, showing which expert handles each question.
- **Reliability Assessment**: Agreement among experts helps gauge answer confidence.

Trade-Offs:

- **Computational Cost**: Running multiple full models per input is resource-intensive.
- **Integration Complexity**: Combining outputs and choosing the most reliable expert requires complex selection methods.
- **Coverage Limitations**: Effective performance relies on having experts for all relevant reasoning types.

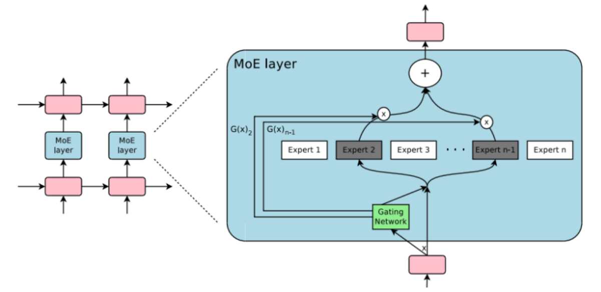


Figure 2: Diagram of mixture of experts (Hugging Face, 2023)

4.4 Multimodal Reasoning Capabilities

Multimodal reasoning is the process by which AI models analyze and integrate information from multiple data types, or "modes," such as text, images, audio, and video, to generate more robust and accurate inferences. By processing diverse inputs, multimodal models can leverage the strengths of each mode to enhance understanding and reasoning capabilities.

In conversational AI, multimodal reasoning enhances high-level reasoning by enabling systems to interpret and respond to a combination of verbal, visual, and contextual cues. By integrating inputs from text, images, audio, or even video, conversational models can provide more accurate, contextually aware responses. For instance, in customer support scenarios, a multimodal conversational AI could analyze product images or documents alongside a user's query to give more precise assistance. In virtual assistants, multimodal reasoning allows the AI to understand and respond appropriately to both spoken language and visual gestures, leading to richer, more natural interactions. This fusion of inputs empowers conversational AI to handle complex, context-dependent tasks with a deeper understanding, improving both accuracy and user experience.

5 Applications

5.1 Question Answering Systems

Question answering represents one of the most important applications of high-level reasoning in conversational AI. Multi-hop QA systems require models to connect multiple pieces of information through logical reasoning steps to arrive at answers. As highlighted in recent work, large language models have shown promising capabilities in multi-hop reasoning through techniques like chain-of-thought prompting (Wei et al., 2022) and self-consistency methods (Wang et al., 2023).

Conversational QA adds another layer of complexity by requiring models to maintain context across multiple turns while performing reasoning. Systems must track relevant information from previous exchanges and integrate it with new queries to provide coherent responses.

5.2 Task-Oriented Dialogue

Task-oriented dialogue systems require structured reasoning to help users accomplish specific goals like booking reservations or troubleshooting technical issues. Recent advances have shown that large language models can decompose complex tasks into logical steps and maintain goal-oriented conversation flows (Li et al., 2022).

5.3 Open-Domain Conversation

Open-domain conversational AI presents unique reasoning challenges as systems must handle unconstrained topics while maintaining coherent and contextually appropriate responses. Modern approaches leverage large language models' broad knowledge bases while implementing techniques to improve reasoning consistency and factual accuracy.

6 Key Datasets and Evaluation Metrics

Several benchmark datasets have emerged as standards for evaluating reasoning capabilities in conversational AI.

6.1 Social IQA (Social Interaction Question Answering)

Social IQA (Sap et al., 2019) contains 35,350 multiple-choice questions targeting social commonsense reasoning in everyday situations. Each question comes with three answer choices and focuses on understanding social interactions, emotional impacts, and behavioral motivations. The dataset's strength lies in its coverage of complex social scenarios and the requirement for both contextual and commonsense reasoning. However, its multiple-choice format may not fully capture the nuanced nature of social interactions, and the limited answer space can make it susceptible to statistical shortcuts by models.

6.2 CoQA (Conversational Question Answering)

CoQA tests models' ability to engage in conversational question answering across multiple domains, including literature, news articles, and Wikipedia passages. The dataset's key strength is its focus on natural conversational flow, requiring models to handle phenomena like coreference and context carryover. Performance is typically measured using the F1 score:

$$\text{F1} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{2}$$

where precision measures answer accuracy and recall measures completeness. A limitation of CoQA is that its conversations tend to follow relatively predictable patterns, potentially understating the complexity of real-world conversational dynamics.

6.3 MultiWOZ (Multi-Domain Wizard-of-Oz)

MultiWOZ is a large-scale dataset containing over 10,000 dialogues spanning multiple domains like hotel booking, restaurant reservations, and travel planning. Its primary strength is the comprehensive annotation of dialogue states and system actions, making it valuable for training and evaluating task-oriented systems. The dataset uses multiple evaluation metrics including:

- **Joint Goal Accuracy (JGA):**

$$\text{JGA} = \frac{\text{Number of turns with all slots correct}}{\text{Total number of turns}} \tag{3}$$

- **Success Rate:** Define:
  - **Numerator (N):** Number of successfully completed dialogues
  - **Denominator (D):** Total dialogues

  The Success Rate can then be calculated as:

$$\text{Success} = \left(\frac{N}{D}\right) \times 100 \tag{4}$$

A notable limitation is that the dataset's structured nature may not fully represent the variability of natural human requests and the complexity of real-world task completion.

6.4 ARC (Abstraction and Reasoning Corpus)

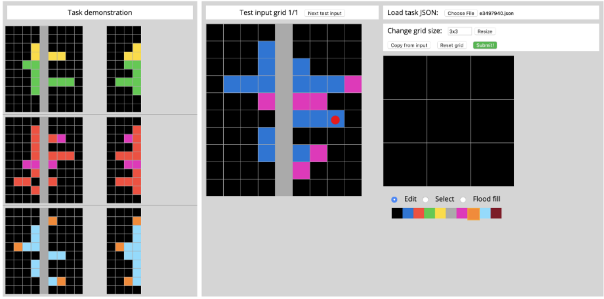
ARC (Acquaviva et al., 2021) specifically targets higher-level reasoning capabilities through tasks requiring abstract pattern recognition and generalization. The dataset consists of tasks where models must infer rules from examples and apply them to new situations. Its strength lies in testing genuine reasoning capabilities rather than pattern matching or memorization. Performance is evaluated using: Define:

- **N:** Correctly solved tasks
- **T:** Total tasks

Then, the Task Success Rate is calculated as:

$$\text{Task Success Rate} = \left(\frac{N}{T}\right) \times 100 \tag{5}$$

However, ARC's relatively small size (400 training tasks) can make it challenging for large-scale training, and its abstract nature may not directly translate to practical conversational scenarios.



Load task JSON: Choose File e3497940.json

Change grid size: 3x3

Figure 3: Visual Representation of ARC (Papers with Code, 2023)

6.5 TheoremQA

TheoremQA (Chen et al., 2023) provides 800 questions testing mathematical and theoretical reasoning across multiple disciplines, including Mathematics, Physics, and Computer Science. The dataset's strength lies in its rigorous evaluation of formal reasoning capabilities and the requirement for step-by-step logical deduction. Evaluation uses multiple metrics, including:

- Solution Accuracy
- Reasoning Path Validity
- Step-by-step Explanation Quality

A limitation is its focus on formal theoretical knowledge, which may not capture the full spectrum of reasoning needed in general conversation.

7 Current Challenges and Limitations

7.1 Hallucination and Factuality

A significant challenge in conversational AI reasoning is the tendency of models to generate plausible but factually incorrect information, known as hallucination (Li et al., 2023). While recent work has made progress in reducing hallucinations through techniques like self-consistency checking and grounding in external knowledge, maintaining factual accuracy remains an important challenge.

7.2 High Computational Demands and Limitations in Training Data

Higher-level reasoning in conversational AI faces both computational and data-related challenges. Advanced transformer-based models, such as GPT-4 and T5, require substantial memory and processing power to maintain nuanced, multi-step reasoning, which raises operational costs and limits scalability, especially in resource-constrained environments (Brown et al., 2020). Managing complex, multi-turn dialogues in real-time further amplifies these demands, creating a trade-off between reasoning sophistication and practical efficiency.

Although parameter-efficient fine-tuning techniques are being explored, achieving high-level reasoning within realistic computational limits remains challenging (Bommasani et al., 2021). Additionally, training data limitations constrain these models' generalization abilities. Most available datasets focus on straightforward interactions, lacking the depth and diversity needed for nuanced reasoning and hypothetical scenarios. This deficiency restricts models' ability to perform complex, domain-specific tasks, such as medical or legal advising, where comprehensive and domain-specific reasoning is essential (Bengio et al., 2019).

7.3 Ethical and Bias-Related Challenges

AI models often reflect the biases present in their training data, which can lead to problematic reasoning patterns in high-stakes applications. For instance, reasoning based on biased data could reinforce stereotypes or produce outputs that subtly endorse harmful assumptions. This challenge is particularly relevant in reasoning-based tasks, where conversational agents are expected to provide fair, unbiased information. Ensuring that AI models can detect, mitigate, and prevent biased reasoning is an ongoing research priority, but developing solutions for real-time bias correction is complex and remains an unresolved challenge. As models reason about subjective or sensitive topics, the risk of reinforcing or perpetuating bias becomes a critical concern, impacting the credibility and ethical application of Conversational AI.

7.4 Challenges in Multi-Modal Reasoning

Many high-level reasoning tasks require information from multiple sources or modalities, such as text, images, or audio. For example, a conversational AI assisting in a medical diagnosis might need to interpret a user's spoken symptoms, medical history text, and even images of physical symptoms. Integrating and reasoning across these modalities remains technically challenging due to differences in data structures, processing requirements, and contextual dependencies. Current multi-modal models are still evolving, and while promising developments are underway, true multimodal reasoning capabilities remain limited (Bengio et al., 2019).

8 Future Directions and Implications

As the field of conversational AI continues to evolve, the development of systems with robust high-level reasoning capabilities is poised to become an increasingly important area of focus and research.

8.1 Emerging Trends

Integrating Symbolic and Sub-symbolic Approaches: Future systems may leverage a combination of symbolic reasoning, based on logical inference and knowledge representation, and subsymbolic techniques, such as deep learning and neural networks, to achieve more comprehensive and flexible high-level reasoning.

Advancing Commonsense Reasoning: Significant progress is expected in equipping conversational AI with commonsense understanding, enabling them to reason about everyday situations, social norms, and implicit knowledge that humans often take for granted.

Incorporating Multimodal Reasoning: As conversational AI systems become more capable of processing and understanding diverse input modalities (e.g., text, speech, vision, sensors), the integration of high-level reasoning across these modalities will be crucial for creating truly intelligent and contextually aware conversational agents.

Advancing Explainable and Transparent Reasoning: There is a growing emphasis on developing conversational AI systems that can explain their high-level reasoning processes, enabling users to understand the rationale behind the system's responses and decisions, fostering trust and transparency.

Developing Personalized and Adaptive Reasoning: Conversational AI systems may evolve to personalize their high-level reasoning strategies based on individual user preferences, behaviors, and contextual factors, leading to more natural and tailored interactions.

8.2 Implications

Enhanced Human-AI Collaboration: Conversational AI with robust high-level reasoning capabilities can facilitate more seamless and effective collaboration between humans and machines, enabling them to work together on complex problem-solving, decision-making, and task-completion.

Improved Customer Service and Support: High-level reasoning in conversational AI can lead to more natural, empathetic, and contextually appropriate customer service experiences, enhancing user satisfaction and engagement.

Advancements in Education and Training: Conversational AI systems with high-level reasoning can be leveraged to create personalized learning experiences, provide intelligent tutoring, and support knowledge acquisition and skill development.

9 Conclusion

In conclusion, this paper presents a well-rounded survey of the advancements, challenges, and future potential of higher-level reasoning in conversational AI. By examining foundational reasoning techniques, modern advancements, and application areas, it underscores the essential role of sophisticated reasoning for AI systems to perform complex problem-solving and nuanced understanding. The paper highlights that while significant strides have been made, challenges such as hallucination mitigation, computational demands, and ethical considerations remain. Future directions like integrating symbolic and neural reasoning, advancing commonsense understanding, and expanding multimodal capabilities hold promise for making conversational AI more reliable and versatile. This work ultimately aims to guide researchers and developers in creating conversational agents that mirror human-like reasoning, thereby enabling more natural, effective, and safe human-AI interactions.

References

[Acquaviva et al., 2021] M. Acquaviva, A. Bandini, S. Melacci, and M. Gori. ARC: Abstraction and Reasoning Corpus for artificial intelligence research. *arXiv preprint*, 2021.

[Alayrac et al., 2022] J. B. Alayrac, et al. Flamingo: a Visual Language Model for Few-Shot Learning. *NeurIPS 2022*.

[Bengio, 2017] Y. Bengio. The consciousness prior. *arXiv preprint arXiv:1709.08568*, 2017.

[Berka, 2020] P. Berka. Statistical approaches to natural language processing. In *Handbook of Research on Machine Learning Applications and Trends*, 2020.

[Bommasani et al., 2021] R. Bommasani, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[Brohan et al., 2023] A. Brohan, et al. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. *arXiv preprint*, 2023.

[Brown et al., 2020] T. B. Brown, et al. Language models are few-shot learners. *NeurIPS 2020*.

[Bubeck et al., 2023] S. Bubeck, et al. GPT-4 Technical Report. *arXiv preprint*, 2023.

[Chen et al., 2022c] N. Chen, et al. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *IEEE Journal*, 2022.

[Chen et al., 2023h] X. Chen, et al. TheoremQA: A Dataset for Mathematical and Scientific Reasoning. *arXiv preprint*, 2023.

[Ding et al., 2021a] L. Ding, et al. AeNER: Temporal reasoning for neural entity recognition. *EMNLP 2021*.

[Driess et al., 2023] D. Driess, et al. PaLM-E: An Embodied Multimodal Language Model. *arXiv preprint*, 2023.

[Eichenberg et al., 2022] C. Eichenberg, et al. MAGMA: Multimodal Augmentation of Generative Models through Adapter-based Finetuning. *arXiv preprint*, 2022.

[Fedus et al., 2022] W. Fedus, et al. Switch Transformers: Scaling to Trillion Parameter Models. *Journal of Machine Learning Research*, 2022.

[Flying Bisons, 2023] Flying Bisons. Hallucinations of ChatGPT-4: Even the Most Powerful Tool Has a Weakness. Retrieved from `https://flyingbisons.com/blog/hallucinations-of-chatgpt-4\protect\discretionary{\char\hyphenchar\font}{}{}even-the-most\protect\discretionary{\char\hyphenchar\font}{}{}powerful\protect\penalty-\@M-tool-has-a-weakness`, 2023.

[Hong et al., 2021] F. Hong, et al. VLGrammar: Learning Visual-Linguistic Grammar from Image-Text Pairs. *CVPR 2021*.

[Hong et al., 2023] F. Hong, et al. 3D-LLM: Injecting the 3D World into Large Language Models. *arXiv preprint*, 2023.

[Hsu et al., 2021] W. N. Hsu, et al. HuBERT: Self-Supervised Speech Representation Learning. *IEEE/ACM Transactions*, 2021.

[Hugging Face, 2023] Hugging Face. Mixture of Experts: Scaling AI with Specialized Neural Networks. Retrieved from `https://huggingface.co/blog/moe`, 2023.

[Kojima et al., 2022] T. Kojima, et al. Large Language Models are Zero-Shot Reasoners. *NeurIPS 2022*.

[Lai et al., 2023] X. Lai, et al. LISA: Reasoning Segmentation via Large Language Model. *arXiv preprint*, 2023.

[Lepikhin et al., 2020] D. Lepikhin, et al. GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding. *ICLR 2020*.

[Lewkowycz et al., 2022] A. Lewkowycz, et al. Solving quantitative reasoning problems with language models. *arXiv preprint*, 2022.

[Li et al., 2022e] J. Li, et al. Task-oriented dialogue systems: Recent advances and future directions. *ACM Computing Surveys*, 2022.

[Li et al., 2022g] S. Li, et al. AlphaGeometry: An Olympiad-level AI system for geometry. *Nature*, 2022.

[Li et al., 2023] X. Li, et al. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 2023.

[Liu et al., 2023] H. Liu, et al. LLaVA: Visual Instruction Tuning. *NeurIPS 2023*.

[Manning, 2022] Christopher D. Manning. 2022. The Rise of Neural Language Models. *Communications of the ACM*.

[OpenAI, 2023] OpenAI. 2023. GPT-4V(ision) System Card. Technical report.

[Papers with Code, 2023] Papers with Code. 2023. ARC: The Abstraction and Reasoning Corpus Dataset. Retrieved from `https://paperswithcode.com/dataset/arc-the-abstraction-and-reasoning-corpus`.

[Peng et al., 2023] Zhenpeng Peng, et al. 2023. Kosmos-2: Grounding Multimodal Large Language Models. arXiv preprint.

[Qiu et al., 2023] Shuang Qiu, et al. 2023. VisionFM: Visual Foundation Models for Medical Image Understanding. *Nature Medicine*.

[Raghu, 2023] Mikhail Raghu. 2023. Graph-based Prompting and Reasoning with Language Models. *Towards Data Science*. Retrieved from `https://towardsdatascience.com/graph-based-prompting-and--with\protect\penalty-\@Mlanguage-models-d6acbcd6b3d8`.

[Reiter, 1975] Raymond Reiter. 1975. The Frame Problem in the Situation Calculus. *Artificial Intelligence*.

[Sap et al., 2019] Maarten Sap, et al. 2019. SOCIAL IQA: Commonsense Reasoning about Social Interactions. EMNLP 2019.

[Si et al., 2023] Chung Si, et al. 2023. Mixture of Reasoning Experts: A Framework for Specialized and General Purpose Reasoning. arXiv preprint.

[Singhal et al., 2023] Kartik Singhal, et al. 2023. Large Language Models Encode Clinical Knowledge. *Nature*.

[Tsai et al., 2022] Hung Tsai, et al. 2022. SUPERB-SG: Enhanced Speech Processing Universal PERformance Benchmark. INTERSPEECH 2022.

[Vaswani et al., 2017] Ashish Vaswani, et al. 2017. Attention is all you need. NeurIPS 2017.

[Wang et al., 2023] Xiang Wang, et al. 2023. Self-consistency improves chain of thought reasoning in language models. ACL 2023.

[Wei et al., 2022] Jason Wei, et al. 2022. Chain of thought prompting elicits reasoning in large language models. NeurIPS 2022.

[Weston and Sukhbaatar, 2023] Jesse Weston and Sainbayar Sukhbaatar. 2023. Learning to reason with neural networks. *Nature Machine Intelligence*.

[Yang et al., 2021] Shuai Yang, et al. 2021. SUPERB: Speech Processing Universal PERformance Benchmark. INTERSPEECH 2021.

[Yao et al., 2023b] Shuai Yao, et al. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. arXiv preprint.

[Yao et al., 2023d] Shuai Yao, et al. 2023. Graph of Thoughts: Solving Elaborate Problems with Large Language Models. arXiv preprint.

[Yu et al., 2023] Dong Yu, et al. 2023. A Survey of Deep Learning Approaches for AI Reasoning. *IEEE Access*.

[Zhang et al., 2022c] Chao Zhang, et al. 2022. Automatic Chain of Thought Prompting in Large Language Models. EMNLP 2022.

[Zhao et al., 2023] Jian Zhao, et al. 2023. Planning-based Dialogue Systems: A Survey. *ACM Computing Surveys*.

[Zhou et al., 2023] Yifan Zhou, et al. 2023. RETFound: A Foundation Model for Retinal Disease Diagnosis. *Nature Medicine*.