Beyond Agreeability: The Critical Thinking Gap Between Large Language Models and Human Cognition

Wen Gong (Digital-Duck)

Abstract: Large Language Models (LLMs) have demonstrated impressive capabilities in various natural language processing tasks, generating human-like text, translating languages, and even writing different kinds of creative content. However, a crucial

necessary for true understanding and general intelligence. This deficiency is illustrated through the "Emperor's Court" analogy, where a lack of critical feedback leads to flawed reasoning and a "chamber echo effect." We focus on the specific challenge of commonsense reasoning as a key example of this System 2 gap and discuss the implications for AI development and the importance of maintaining human agency. We further propose that developing a "model of self," involving meta-cognition, is a necessary step towards more balanced and robust AI systems.

Introduction:

The rapid advancement of Artificial Intelligence (AI), particularly in the realm of Large Language Models (LLMs), has sparked widespread excitement and speculation about the future of technology and the potential arrival of Artificial General Intelligence (AGI). LLMs have demonstrated remarkable abilities in generating human-like text, translating languages, writing various creative content, and even producing precise working codes and software applications. These impressive feats have led to a surge of interest and investment in AI, with many predicting a transformative impact on various industries and aspects of human life. However, amidst this enthusiasm, it is crucial to maintain a critical perspective and carefully examine the fundamental limitations of current AI systems. This paper argues that while LLMs excel at certain cognitive tasks, primarily those involving pattern recognition and prediction—which align with what psychologists call System 1 thinking (fast, intuitive, and associative)—they fundamentally lack crucial aspects of human cognition, most notably the capacity for independent critical thinking, abstract reasoning, and a robust "model of self"—capabilities associated with System 2 thinking (slow, deliberate, and analytical). This deficiency has significant implications for AI development, the responsible deployment of AI systems, and the broader societal impact of this rapidly evolving technology. To illustrate this critical gap, we draw a parallel with historical power structures, specifically the "Emperor's Court" analogy, demonstrating how a lack of critical feedback and independent thought can lead to flawed reasoning and a "chamber echo effect," mirroring the limitations observed in current LLMs. We will then delve into a technical analysis of this deficiency, focusing on the specific challenge of commonsense reasoning as a key example of the gap between current AI and human-level general intelligence. Finally, we will propose research directions that could lead to more balanced, robust and ethical AI systems capable of more sophisticated forms of cognition.

The Emperor's Court Analogy (System 1 in Action):

Throughout history, rulers have often surrounded themselves with advisors and courtiers. While wise rulers sought out advisors who provided honest counsel and critical feedback, others found themselves surrounded by "yes-men"—individuals who primarily agreed with the ruler's opinions, reinforcing existing beliefs and suppressing dissent. This dynamic, often referred to as a "chamber echo effect," can have disastrous consequences, leading to flawed decisions, a lack of innovation, and ultimately, harm to the ruler and their realm. A compelling example is the reign of Zhu Yuanzhang, the founding emperor of the Ming Dynasty in China (1368-1644). Zhu Yuanzhang, after rising from humble beginnings to become emperor, was known for his suspicious nature, autocratic rule, and tendency to distrust his advisors. He established a highly centralized government and often relied on advisors who catered to his whims and avoided contradicting him, fearing his wrath. This created an environment where critical feedback was suppressed, and dissenting voices were silenced. The lack of independent evaluation and critical thinking within the court created a "chamber echo" where the emperor's own thoughts, biases, and sometimes even paranoia, were constantly amplified and reflected back to him, without any external correction. This lack of critical input led to numerous misjudgments, purges of officials, and policies that ultimately hampered the long-term stability and prosperity of the Ming Dynasty. Note: The selection of Zhu Yuanzhang as one historic data point was inspired by a 50 episode TV series titled "传奇皇帝朱元璋 Legendary emperor Zhu Yuanzhang". Similar examples exist in history, and even in current organizations and affairs.

This historical example provides a powerful analogy for understanding the limitations of current LLMs. Like an emperor surrounded by "yes-men," LLMs are trained on vast amounts of data and are designed to generate outputs that align with the statistical patterns they have learned from that data. They are essentially trained to "agree" with the data, reinforcing existing biases and lacking the ability to independently evaluate the validity, truthfulness, or even the coherence of the information they process. This is a clear manifestation of System 1 thinking: fast, automatic responses based on learned associations, without the deliberate, reflective analysis characteristic of System 2. Just as the Emperor's court lacked the critical thinking necessary to challenge the Emperor's views, LLMs lack the internal mechanisms to critically evaluate the information they are given as input or the outputs they generate. This "chamber echo effect" in LLMs can manifest in various ways, such as the amplification of biases present in the training data, the generation of factually incorrect or nonsensical outputs that are nonetheless grammatically correct and contextually plausible within the limited scope of the training data, and the inability to adapt to novel situations that require a deeper understanding of context, meaning, and real-world knowledge.

Technical Analysis: The System 1/2 Divide in AI

System 1: Pattern Recognition Without Understanding (Current AI)

Current AI systems, particularly those based on deep learning architectures like Large Language Models (LLMs), excel at identifying statistical patterns within vast datasets. They operate through a process of pattern matching, predicting the most probable next word or sequence of words based on the statistical relationships they have learned. This allows them to perform remarkably well on tasks like text generation, translation, and summarization. However, this proficiency is achieved through a form of "surface-level understanding." While LLMs can manipulate symbols effectively, generating grammatically correct and contextually relevant text, they lack a true comprehension of the underlying meaning or the real-world implications of their outputs. This is analogous to System 1 thinking in humans: fast, automatic, intuitive, and associative. It's a system that excels at rapid responses and pattern recognition but lacks the depth of analysis and critical evaluation. Furthermore, these models are susceptible to biases present in their training data and often struggle to generalize to unseen situations or contexts.

The Challenge of Commonsense Reasoning (A System 2 Deficiency)

Commonsense reasoning requires understanding and applying everyday knowledge about the world. It involves abilities like understanding physical laws, social norms, and causal relationships. Current AI struggles significantly with this, indicating a deficiency in System 2 thinking. Examples like the Visual Commonsense Reasoning (VCR) dataset (Zellers et al., 2018), SWAG (Zellers et al., 2018), and research on physical and social commonsense (Bisk et al., 2020; Sap et al., 2019) demonstrate this. These failures highlight the lack of key System 2 components: causal reasoning, counterfactual thinking, and mental models of the world. For example, understanding that a "suit of armor cannot conduct an orchestra" (Bisk et al., 2020) requires knowledge of physical properties, human capabilities, and social contexts – knowledge that current AI lacks. This lack of grounding in the physical world is a significant obstacle for current AI systems.

The Missing Ingredient: Meta-cognition and the "Model of Self" (Essential for System 2)

Meta-cognition, "thinking about thinking," is crucial for critical thinking and a "model of self." It enables reflection on cognitive processes, bias identification, and reasoning evaluation. Current AI lacks this, operating in a "black box" fashion. This prevents them from understanding why they make mistakes, correcting biases, planning learning, or assessing output certainty. This lack directly impairs commonsense reasoning. Without reflecting on their understanding, they cannot identify inconsistencies or flawed inferences.

The "Aha!" Moment: Insight and the Role of Parallel Processing (A System 2 Phenomenon)

A distinctive characteristic of human cognition is the experience of sudden insight, often referred to as an "aha!" moment. These moments of clarity often arise after a period of focused thought, incubation, or even seemingly unrelated experiences. They represent a shift from System 1's associative thinking to System 2's deliberate and analytical processing. The human brain, being a massively parallel neural network, can process information simultaneously across multiple interconnected regions. This parallel processing allows for the integration of diverse information and the emergence of novel connections and insights. These insights often feel spontaneous or unexpected, even though they are the result of underlying cognitive processes.

A classic example of this is the story of Isaac Newton and the falling apple. While observing an apple falling from a tree, Newton is said to have had a sudden insight into the nature of gravity, realizing that the same force that pulled the apple to the ground also kept the moon in orbit around the Earth. This "aha!" moment, triggered by a seemingly simple observation, led to the formulation of his groundbreaking theory of universal gravitation. This exemplifies how the human brain can connect seemingly disparate pieces of information to generate novel and profound insights.

Current AI systems, while employing parallel processing within their neural network architectures, lack the global integration and flexible connectivity of the human brain. They operate primarily in a feed-forward manner, processing information sequentially through layers of neurons. This architecture makes it difficult for them to generate the kind of sudden insights that characterize human creativity and problem-solving. The lack of a "model of self" and meta-cognition further hinders their ability to engage in the kind of reflective processing that often precedes "aha!" moments.

Beyond Scaling: The Need for New Approaches

The "scaling hypothesis"—that simply increasing model size and data will lead to AGI—is challenged by these limitations. While scaling improves performance on some tasks, it does not address the fundamental lack of System 2 capabilities. New approaches are needed, including:

- Integrating symbolic reasoning and knowledge graphs (Ahn et al., 2021).
- Developing methods for causal inference.
- Exploring architectures that support meta-cognition.
- Embodied AI and grounded cognition.

Implications:

The limitations of current AI have significant real-world implications:

- Misleading Marketing and Hype: The marketing of AI often overstates its capabilities, creating unrealistic expectations and obscuring limitations. This can lead to misinformed decisions and a lack of preparedness for the actual capabilities and limitations of deployed systems.
- Ethical Concerns: Deploying AI systems lacking critical thinking in high-stakes domains (healthcare, law, finance) raises serious ethical concerns. Flawed reasoning, bias amplification, and a lack of accountability can have severe consequences.
- Impact on Human Work and Society: Overreliance on AI could diminish human critical thinking skills and agency. If we become accustomed to accepting AI outputs without question, we risk losing our ability to make informed judgments.
- The Importance of Human-in-the-Loop Systems: Human oversight is crucial, especially in situations requiring critical thinking or commonsense judgment. Human-in-the-loop systems, where humans provide feedback and intervention, are essential for responsible AI deployment.

Recommendations:

To address these limitations and move towards more robust and beneficial AI, we recommend:

- Focus on System 2 Capabilities: Research should prioritize developing AI systems with System 2 functions like causal reasoning, meta-cognition, and abstract thinking.
- Hybrid AI Architectures: Combining deep learning with symbolic AI, knowledge graphs, and other techniques offers a promising path towards more balanced systems.
- Embodied AI and Grounded Cognition: Embodiment can provide the grounded experience necessary for developing commonsense understanding.

- Developing Better Evaluation Metrics: New metrics are needed to assess System 2 capabilities, going beyond simple accuracy to measure reasoning, causal understanding, and meta-cognition.

- Promoting Public Understanding of AI Limitations: Public education is crucial to prevent unrealistic expectations and ensure responsible AI development and deployment.

Conclusion:

This paper has argued that while Large Language Models (LLMs) have demonstrated impressive capabilities in various natural language processing tasks, they fundamentally lack the capacity for independent critical thinking, a key characteristic of human cognition. By applying the System 1/2 framework, we have shown that current LLMs operate primarily through System 1 thinking—fast, intuitive, and associative—excelling at pattern recognition and prediction. However, they struggle with tasks that require System 2 thinking—slow, deliberate, analytical reasoning, and meta-cognition. The "Emperor's Court" analogy effectively illustrated how a lack of critical feedback and independent thought, mirroring the limitations of current LLMs, can lead to flawed reasoning and a "chamber echo effect." We focused on the specific challenge of commonsense reasoning as a prime example of this System 2 deficiency, highlighting the lack of causal reasoning, counterfactual thinking, and robust mental models of the world in current AI systems. Furthermore, we argued that meta-cognition, the ability to "think about thinking" and develop a "model of self," is a crucial missing ingredient for achieving true general intelligence. Simply scaling up current models is unlikely to bridge this gap; instead, new approaches are needed that explicitly target the development of System 2 capabilities. These include integrating symbolic reasoning and knowledge graphs, developing methods for causal inference, exploring new architectures that support meta-cognition, and pursuing research in embodied AI and grounded cognition. Ultimately, fostering public understanding of AI's current limitations and promoting responsible AI development are essential to ensure that this powerful technology is used for the benefit of humanity.

Acknowledgments:

The author gratefully acknowledges the significant assistance provided by Gemini (Google AI) (version 2.0 Flash Experimental) in conducting literature reviews, synthesizing information, generating text, and refining the arguments presented in this paper. Gemini's ability to access and process vast amounts of information was invaluable in exploring the complex relationship between current AI capabilities and the pursuit of Artificial General Intelligence.

The author also acknowledges Albert W. Gong for suggesting System 1/2 framework by psychologist Daniel Kahneman in the context of this paper.

Appendix:

Data-Copilot: A Generative AI System for Understanding Human Cognition

Data-Copilot is a generative AI system designed to facilitate self-service data analytics through natural language interaction. It acts as an "interpreter" between analysts and data, allowing users to "talk to data" and derive insights more efficiently.

Architecture:

Data-Copilot employs a Retrieval Augmented Generation (RAG) architecture, combining the power of Large Language Models (LLMs) with external knowledge sources. The system consists of the following key components:

- Analyst (User): Users interact with the system by asking questions in natural language.
- Prompt: The user's question serves as the initial prompt.
- Knowledge Base: A vector store (default: ChromaDB) containing database schema information (DDL) and documentation is searched for relevant context based on the prompt.
- LLM: The user's prompt and the retrieved context from the Knowledge Base are provided as input to the LLM.
- Code Generation: The LLM generates SQL queries or Python code to interact with the database.
- Database Interaction: The generated code is executed against the specified database (default: SQLite, with support for other SQL databases).
- Data Processing and Visualization: The results from the database are processed and visualized using libraries like Plotly.
- Insight Generation: The visualized data and any generated text explanations are presented to the user as insights.
- Feedback Loop: User feedback is collected to improve the system's performance over time.
- AI-generated Follow up questions: The system can suggest follow-up questions to the analyst. (Include the architecture diagram here)

Configuration:

Data-Copilot offers a flexible configuration interface (see Figure 1) that allows users to customize the system according to their needs. Users can specify:

- Data Base: The type and location of the database to be queried.
- Knowledge Base: The vector store to be used for RAG.
- GenAI Model: The specific LLM to be used for code generation and natural language processing. The system supports various models from different providers, including OpenAI, Anthropic, Google, Alibaba, and open-source models through Ollama. (Include the configuration page screenshot here – "Figure 1")

Database Interaction:

Data-Copilot provides a dedicated interface for interacting directly with the configured database (see Figure 2). This interface includes:

- SQL Editor: A text area for writing and executing SQL queries. This allows users to perform complex data manipulations and retrievals beyond what can be easily expressed in natural language.
- Schema Display: The schema of the selected table is displayed, providing users with a clear understanding of the data structure.
- Entity-Relationship Diagram (ERD): A visual representation of the database schema, showing the relationships between different tables (see Figure 3). This facilitates data exploration and helps users understand how different data elements are connected. (Include the "Query Database" page screenshot here – "Figure 2")

(Include the ERD screenshot here – "Figure 3")

Knowledge Base Management:

Data-Copilot's Knowledge Base is a crucial component for grounding the LLM's responses and enabling accurate data retrieval. The "Train Knowledge-base" page (see Figure 4) provides an interface for managing and populating this Knowledge Base. Key functionalities include:

- Knowledge Base Content Display: The current content of the Knowledge Base is displayed in a table, showing the associated question, the corresponding SQL query or documentation text, and a unique ID.
- Knowledge Management: Users can remove individual entries or clear the entire Knowledge Base. Entries can also be removed using their vector IDs, indicating the use of vector embeddings for efficient retrieval.
- Knowledge Ingestion: Users can add information to the Knowledge Base in several ways:

- Schema Addition: Database schema information (DDL) can be added to provide the LLM with a structural understanding of the data.
- Documentation Addition: Business documentation and other relevant text can be added to provide context and background information.
- Question/SQL Pair Addition: Users can manually add question-SQL pairs to train the LLM to generate accurate queries for specific types of questions. (Include the "Train Knowledge-base" page screenshot here – "Figure 4")

Natural Language Interaction and Retrieval Augmented Generation (RAG):

Data-Copilot's core functionality revolves around natural language interaction and Retrieval Augmented Generation (RAG). The "Ask RAG" page (see Figure 5) provides the interface for this interaction.

The process unfolds as follows:

- The user enters a natural language prompt.
- Data-Copilot uses the prompt to query its Knowledge Base, retrieving relevant context, such as database schema information and documentation.
- The LLM receives the user's prompt and the retrieved context as input.
- The LLM generates a corresponding SQL query to retrieve the requested data from the database.
- The generated SQL query is executed, and the results are returned as a DataFrame.
- The LLM may also generate Python code to further process the data and create visualizations.
- The results are displayed as a table and a chart.
- The LLM generates a natural language response, providing context and explanation based on the retrieved information from the Knowledge Base. (Include the "Ask RAG" page screenshot here – "Figure 5")

Movie Explorer Use Case:

Data-Copilot's capabilities extend beyond basic data retrieval. The "Movie Explorer" use case (see Figure 8 and 9) demonstrates its ability to handle more complex natural language queries and provide richer, contextualized responses. This use case also demonstrates the crucial difference between querying the LLM directly and using Retrieval Augmented Generation (RAG).

RAG Mode (Figure 8, Left Side):

When RAG is enabled, Data-Copilot grounds the LLM's responses in the connected database. This allows users to explore and discover information based on structured data. For example, a user can ask: "Show me top 10 drama movies from the 1990s with over 100k votes (Mainstream hits)." Data-Copilot translates this into a SQL query that filters movies by genre, release year, and number of votes, returning the top 10 results. The system also generates a chart visualizing the results. This mode is ideal for exploring data and discovering patterns or trends.

LLM Direct Mode (Figure 9, Right Side):

When RAG is disabled, the user's question is sent directly to the LLM without any grounding in the database. This allows the LLM to leverage its pre-trained knowledge to provide general information. For example, a user can ask for more information about a specific movie, such as "tell me more about movie 'The Shawshank Redemption', where can I watch it online." In this case, Data-Copilot relies on the LLM's internal knowledge to provide a detailed summary, including plot details, key themes, and characters. The system also attempts to provide information about streaming availability, demonstrating its ability to integrate external knowledge sources. This mode is useful for obtaining general information or summaries about specific entities.

Key Difference:

The key difference between these two modes is that RAG uses the database as a source of truth, ensuring that the information provided is accurate and grounded in real data. The direct LLM mode, on the other hand, relies on the LLM's internal knowledge, which may be incomplete, inaccurate, or outdated. This highlights the importance of RAG for mitigating the limitations of LLMs, such as hallucinations and lack of access to up-to-date information.

(Include the "Ask RAG" page screenshots showcasing the Movie Explorer queries, clearly distinguishing between RAG mode and direct LLM mode – "Figure 8 and 9")

Note Taking:

Data-Copilot also provides a built-in note-taking feature (see Figure 6) that allows users to store and organize relevant information within the application. This feature includes:

- Note List: A table displays all saved notes, including their title, a brief excerpt, associated URLs, tags, activity status, and update timestamps.
- Note Editor: A form for creating and editing notes, with fields for the title, URL, note content, and tags.

- Note Management: Users can save, delete, and download their notes as a CSV file. This feature allows users to maintain context and keep relevant information readily available during their data analysis workflow.

(Include the "Notes" page screenshot here – "Figure 6")

Open-Source Technologies and Acknowledgements:

Data-Copilot is built upon and leverages several open-source technologies and resources (see Figure 7), which we gratefully acknowledge:

- Vanna.ai: Provides functionality for chatting with SQL data, serving as a foundation for the user interaction aspects of Data-Copilot.
- Retrieval Augmented Generation (RAG): The core architecture of Data-Copilot is based on RAG principles, enabling the system to combine the strengths of LLMs with structured data for accurate and insightful responses.
- Hugging Face: We leverage Hugging Face transformers and libraries to access a wide range of pre-trained language models (LLMs) and tools for fine-tuning and integration within Data-Copilot. This allows for flexibility in choosing the most appropriate LLM for different user needs.
- Chroma: Data-Copilot utilizes Chroma, a vector store, to efficiently manage and store the Knowledge Base. This enables fast retrieval of relevant context for the LLM, improving the overall performance of the system. (Include the "Open Source Technologies" page screenshot here – "Figure 7")

Future Work and AI Embodiment:

Data-Copilot is a continuously evolving platform, and future development efforts will focus on several key areas:

- Enhanced User Interaction: Expanding natural language capabilities to support even more complex queries and data exploration workflows.
- Task Management Integration: Integrating task management functionalities to assist users in organizing and completing data analysis tasks within the application.
- User Interaction Data Capture: Capturing user interaction data (queries, responses, insights) to provide personalized user experiences and improve the system's performance over time. This data can be anonymized and aggregated to learn about user behavior patterns and preferences.
- Explainable AI and Trustworthiness: Implementing explainable AI techniques to provide users with transparency into how Data-Copilot arrives at its conclusions. This will build trust and confidence in the system's capabilities.

Here's where the concept of AI embodiment comes into play. By capturing user interaction data and utilizing explainable AI, Data-Copilot can develop a deeper understanding of how users approach data analysis. This information can be used to personalize the user experience, recommend relevant insights, and potentially suggest follow-up actions. Over time, Data-Copilot can evolve into a more "embodied" AI, becoming an integral part of the user's data analysis workflow and offering a level of collaboration and assistance that transcends simple data retrieval.

Conclusion:

Data-Copilot represents a novel approach to data analysis, empowering users with natural language interaction and leveraging the power of LLMs. Its design fosters self-service data exploration while mitigating the limitations of LLMs through Retrieval Augmented Generation (RAG). Additionally, the planned future work on user interaction data capture and explainable AI holds promise for building an AI embodiment that can offer a more personalized and collaborative data analysis experience. We believe that Data-Copilot has the potential to significantly enhance the way users interact with and derive insights from data.

Source code: https://github.com/digital-duck/data-copilot

Demo videos:

- Data Copilot demo 2024-12-10
- [Movie Explorer demo](to be made)

References:

- Ahn, D., Shin, S., & Choi, H. (2021). Improving Commonsense Reasoning in Language Models via Knowledge Graph Transformations. arXiv preprint arXiv:2106.04419.
- Ba, J., & Caruana, R. (2014). Do Deep Nets Really Need to be Deep?. Advances in Neural Information Processing Systems, 27.
- Bisk, Y., Zellers, R., Lin, X., & Choi, Y. (2020). Can a Suit of Armor Conduct an Orchestra? Common Sense Reasoning for Physical Interactions. arXiv preprint arXiv:2005.00693.
- Hendrycks, D., Burns, C., Basart, S., Zemel, R., Misra, S., Parikh, J., ... & Jurafsky, D. (2020). Measuring massive multitask language understanding. arXiv preprint arXiv:2009.03300.

- Sap, M., Rashkin, H., Chen, D., Le Bras, R., & Choi, Y. (2019). Social IQA: Commonsense Reasoning about Social Interactions. arXiv preprint arXiv:1904.09728.
- Zellers, R., Bisk, Y., Schwartz, R., & Choi, Y. (2018). SWAG: A Large-Scale Adversarial Dataset for Commonsense Reasoning. arXiv preprint arXiv:1808.05326.
- Zellers, R., Yatskar, M., Yih, S. W., & Choi, Y. (2018). From Recognition to Cognition: Visual Commonsense Reasoning. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 6720-6729.
- Daniel Kahneman. 2011 book: Thinking, Fast and Slow. https://www.wikiwand.com/en/articles/Thinking,_Fast_and_Slow
- Legendary emperor Zhu Yuanzhang 传奇皇帝朱元璋