



LLMs: between language modeling and complex Al architectures

R. Basili, AI*IA NLP Working Group (Università di Roma, Tor Vergata)

AI*IA SUMMER SCHOOL 2024: ADVANCES IN AI, Lake Como School of Advanced Studies, September 23-27, 2024

Overview

- Generative AI: novel learning architectures for old tasks
- Learning and Natural Language in Generative AI
 - The role of language in Knowledge Representation
 - Attention, encoding-decoding in modern learning machines
 - Opportunities and risks
- Towards reliable generative AI systems
 - Agency: Task decomposition and requirement in intelligent application
 - Domain adaptation: from learning mechanisms to process engineering
 - LLM architectures
- Perspectives

Generative AI: Novel learning methods for old tasks

AI: a long and winding road







Sources: ChatGPT, Futurism

More charts: genuineimpact.substack.com

From Eliza ... to Chat GPT



Q: Who was president of the United States in 1801? A: What answer would please you the most?

Q: Which is heavier, a mouse or an elephant? A: Come, come, elucidate your thoughts. Welcome to

EEEEEE	LL	IIII	ZZZZZZ	AAAAA	
EE	LL	11	ZZ	AA	AA
EEEEE	LL	II.	ZZZ	AAA	AAAA
EE	LL	11	ZZ	AA	AA
EEEEEE	LLLLL	IIII	ZZZZZZ	AA	AA

Eliza is a mock Rogerian psychotherapist. The original program was described by Joseph Weizenbaum in 1966. This implementation by Norbert Landsteiner 2005.

ELIZA:	Is something troubling you ? Men are all alike
ELIZA:	What is the connection, do you suppose ?
YOU:	They're always bugging us about something or other.
ELIZA:	Can you think of a specific example ?
YOU:	Well, my boyfriend made me come here.
ELIZA:	Is it important to you that your boyfriend made you come here 7
YOU:	He says I'm depressed much of the time.
ELIZA:	I am sorry to hear that you are depressed.
YOU:	It's true. I am unhappy.
ELIZA:	Can you explain what made you unhappy ?
YOU:	

From Eliza ... to Chat GPT





> Hello, I am Eliza.

Q: Who was president of the United States in 1801? A: What answer would please you the most?

Q: Which is heavier, a mouse or an elephant? A: Come, come, elucidate your thoughts.

Q: How many eyes does a giraffe have? A: A giraffe has two eyes.

Q: How many eyes does my foot have? A: Your foot has two eyes.

Q: How many eyes does a spider have? A: A spider has eight eyes.

Q: How many eyes does the sun have? A: The sun has one eye.

Q: How many eyes does a blade of grass have? A: A blade of grass has one eye. >that i am unintelligent How long have you been unintelligent?

>since i was born I see.

>arent you supposed to cheer me up You're not really talking about me, are you?

>no im talking about me Why not?

≻mrh I'm not sure I understand you fully.

Q: Who was president of the United States in 1801?A: Thomas Jefferson was president of the United States in 1801.

Q: What city is in the northwest corner of Ohio? A: Toledo is in the northwest corner of Ohio.

Q: What highway do you take to get to North Lake Tahoe from Oakland? A: Highway 80 is the highway you take to get to North Lake Tahoe from Oakland.

Q: What are Darius's abilities in League of Legends? A: Darius's abilities in League of Legends are: Apprehend, Decimate, Noxian Guillotine, Crippling Strike, Apprehend, and Hemorrhage.

Q: Who did Capablanca defeat to become world chess champion? A: Capablanca defeated Lasker to become world chess champion.

From https://lacker.io/ai/2020/07/06/giving-gpt-3-a-turing-test.html

The perceptron (Rosenblatt, 1957)



Deep vs. Shallow networks: the intuition



What is Machine Learning



Learning and Knowledge Representation: the role of depth



Representation Learning



Useful Representations in images



Fig. 2. An illustration of the SegNet architecture. There are no fully connected layers and hence it is only convolutional. A decoder upsamples its input using the transferred pool indices from its encoder to produce a sparse feature map(s). It then performs convolution with a trainable filter bank to densify the feature map. The final decoder output feature maps are fed to a soft-max classifier for pixel-wise classification.

NL and Knowledge in AI: from communication to intelligent decision making

Natural Language as a Markov process





Natural Language as a Markov process

Language manifests as a generative process characterized by:
Total Ordering of symbols

• ... Uncertainty

- ... Vagueness
- … Non determinism
- ... Partial Observability
- ... a comunicative Goal



Distributional Hypothesis and NL semantics

• NL semantics and language in use

- Since 50', lexical meanings have been studied in connection with the pattern of use (distributions) of lexical items in texts
- Such distributions are the *building blocks* of semantics in CL.

• Different definitions for the *Distributional hypothesis* (Harris, 57):

- "words which are similar in meaning occur in similar contexts" (Rubenstein & Goodenough, 1965);
- "words with similar meanings will occur with similar neighbors if enough text material is available" (Schutze & Pedersen, 1995);
- "a representation that captures much of how words are used in natural context will capture much of what we mean by meaning" (Landauer & Dumais, 1997);

Linguistic Patterns and Machine Learning: Encoder-decoders for NL



Encoding NL knowledge: Lexical Autoencoding



Semantic Embedding from domain *Corpora*:

Wordspace (unsupervised induction of lexical semantic representations)







Generative AI: successful learning algorithms and intelligent behaviour

From attention to chain-of-thoughts



Williams, Ronald J.; Hinton, Geoffrey E.; Rumelhart, David E. (October 1986).

RNNs

98

Bidirectional RNNs

1997



Schuster, Mike, and Kuldip K. Paliwal. 1997

RNNsBidirectional RNNsEncoder-Decoder RNNs198619972014



I. Sutskever, O. Vinyals, & V. Le Quoc, 2014











Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901. https://arxiv.org/abs/2005.14165

New training models

O-shot learning. Instruction Learning

How to trigger *reinforcement learning*

The advent of ChatGPT



A flourishing of training modalities

- Supervised ML: training as a task-specific process based on annotated examples
- LLMs suggest pre-training as an effective way to exploit transfer learning:
 - Pre-training on a general task and fine-tuning on the final (target) task
 - Encoders (e.g., BERT) are specialized to a domain through pretraining
- Decoders can learn in 0-shot or few-shot mode through prompting
 - It is the prompt that describes the task, the question, and the input
 - In some cases (few-shot) 1 to 5 examples can be provided
 - Prompting can be flexible and applied for novel tasks (in-context learning)
- Instruction tuning (see Chat GPT) extends prompting to meta-learning

Zero and Few-shot Learning

- Training dedicated ML systems until 2020 consisted of optimizing the decision function on the basis of Supervised Learning.
 - Examples of correct and incorrect decisions
 - Reinforcement Learning. Feedback from the environment on the usefulness of decisions
 - According to iterative and cross-validation schemes.
- The introduction of language modelling enables the introduction of new tasks using natural language
 - The translation of the phrase "I love New York" into Italian is.
 - The sentiment of this post, "I love democracy!"?

The role of prompting

• **O**-SHOT CATEGORIZATION (TEXTUAL INFERENCE):

• «Wonderful concert! I am happy» - «Is this tweet expressing a positive sentiment?»

• **O**-SHOT IN-LINE INSTRUCTION:

- «Write a text on Spiderman»
- "The text should not exceed 5 pages, be double spaced, and include mentions for at least two classical of his supercriminal enemies."

• Few-shot learning:

- <u>DEF</u>: "Clean a post means remove possible morphological erros and mispellings."
- <u>EX</u>: "For example in '*Clen the table surface*' should be '*Clean the table surface*' "
- Input: "Please clean the sentence 'I don't want to press you, but plese answer soon!"
Foundation Models

A new wave of AI technology is about to completely change our conception of the capabilities of artificially intelligent systems: *Foundation Models*.

While up to now, AI systems were usually built by training learning algorithms on datasets specifically constructed for a particular task at hand, researchers and engineers are now using the almost limitless supply of available data, documents, and images on the Internet to train models relatively independently of the possible tasks for which they might be used later on.

Using large document sets with trillions of words, and incorporating hundreds of billions of parameters, such deep network models construct a re-representation of their inputs and store them in a way that later allows them to be used for different tasks such as question/answering and even inference.

Such models already produce results that were unimaginable before, and will lead to AI systems that are significantly more flexible, dramatically more powerful, and ultimately closer to a truly general AI.

Stefan Wrobel, June 2022



... nowadays ...

Limitations of LLMs

- Memory and cpu-intensive with high latency times
- Strong exposition to information obsolescence and limited accesso to external sources
- Unclear ability to memorize suitable information
- Uncertainty of the generation process



Reflection

• Expertise, Rationality and Omniscience

- A generative AI system exhibits a SIGNIFICANT LINGUISTIC COMPETENCE in agreement with native speaker / writers of the different languages observed during pre-training, fine-tuning
- It is **RATIONAL** in linguistic terms as he knows the rule of human communication and it uses them *somehow usefully*
- An LMM is not OMNISCIENT
 - Common sense errors
 - Lack of technical (e.g. mathematical) knowledge
 - Rather naive in expert domains
- It is not always fully coherent
 - Hallucinations

Conclusions

Conclusions



news.oortheastern.edu Parrots learn to make video calls to chat with other parrots, then develop friendships, Northeast...

Q 10 12 07 Q 203 161 24.4K

- Large scale Language Models (LLMs) are definitively not just "stochastic parrots"
 - Data-driven models trained over resources rappresentative about speakers' knowledge, esperiences and culture
 - The semantic awareness of these systems is very competitive wrt logical axiomatizations that are usually artificial and too much software engineering dependent
 - Natural language is the core media for knowledge modeling and sharing ...
 - ... and prompt engineering seems the crucial mechanism to exploit it

Conclusions (2)

- Agency of LLMs is a promising way to rethink AI and harmonize back the recent generative AI successes with knowledge representation aspects that still remain unsolved
 - Generalizing LLMs to complex tasks for which language modelling is not adequate
 - Exploring cognitive dimensions of reasoning, decision making and acting
 - Integrated Vision and language
 - Experience and grounded reasoning
 - Meta-learning: instruction tuning and dynamic reasoning
 - Pursuing the sustainability problems of very large scale language models
- Research in LLM agent architectures is in its infancy
- Training and Evaluating LLM agent is an open research area still relatively unexplored

NLP and LLM papers

- Y. Goldberg. "A Primer on Neural Network Models for Natural Language Processing". In: J. Artif. Intell. Res. 57 (2016), pp. 345–420.
- <u>Attention</u>: Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).
- word2Vec lexical embedding: (Mikolov et al., 2013) Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality [J]. Advances in Neural Information Processing Systems, 2013, 26(1): 3111-3119.
- <u>BERT Encoding</u>: (Devlin et al., 2018) Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics
- <u>T5</u>: (Wolf et al, 2019) Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-theart natural language processing. ArXiv, abs/1910.03771.
- <u>BART Encoding-Decoding</u>: Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461.https://arxiv.org/abs/1910.13461
- <u>GPT Few-Shot Learning</u>: Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901. https://arxiv.org/abs/2005.14165

Advanced LLM and LLM Agent papers

• Reasoning

- JasonWei et al., 2023, Chain-of-Thought Prompting Elicits Reasoning in Large Language Models,
- STaR: Self-Taught Reasoner, Bootstrapping Reasoning With Reasoning, Eric Zelikman et al., arXiv:2308.08155v2, October 2023.
- LLM Agents: Survey
 - Zhiheng Xi et al., The Rise and Potential of Large Language Model Based Agents: A Survey, Sept. 2023., URL:https://arxiv.org/abs/2309.07864

Agents and Multiagents

- ReAct. Yao, Shunyu, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan and Yuan Cao. "ReAct: Synergizing Reasoning and Acting in Language Models." ArXiv abs/2210.03629 (2022)
- ReWOO. Binfeng Xu et al., 2023, "ReWOO: Decoupling Reasoning from Observations for Efficient Augmented Language Models"
- HuggingGPT. Shen, Yongliang, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. "HuggingGPT: Solving AI Tasks with ChatGPT and Its Friends in Hugging Face."
- Du, Yilun, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. "Improving Factuality and Reasoning in Language Models through Multiagent Debate."
- Liang, Tian, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi.
 2023. "Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate."
- Qingyun Wu et al., AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation, October 2023.
- Taicheng Guo, et al., Large Language Model based Multi-Agents: A Survey of Progress and Challenges

CONTACTS: PROF. R. BASILI,

DEPT. OF ENTERPRISE ENGINEERING UNIVERSITY OF ROMA, *TOR VERGATA*

CHAIR OF THE EUROPEAN CHAPTER OF ACL

MAIL: basili@info.uniroma2.it

Web SAG group: http://sag.art.uniroma2.it/

PUBBLICATIONS: Google Scholar



Towards reliable generative Al systems

Retrieval Augmented Generation

LLM agency

Knowledge Integration and LLMs: RAG Models

- Retrieval Augmented Generation (Lewis et al., 2020)
 - At generation time contextual information able to qualify the LLM response is made available
 - It is essential for knowledge intensive tasks
 - It is possible to apply RAG either to the pre-training or to the fine-tuning and prompting stage
 - It has been shown to impact positively onto hallucinations



Figure 1: Technology tree of RAG research development featuring representative works

(Lewis et al, 2020) <u>Retrieval-augmented generation for knowledge-intensive NLP tasks. Proceedings of NIPS, Advances in Neural Information Processing Systems</u>, 33 (2020): 9459-9474.

The basic Retrieval workflow



RAG: the steps

1.<u>INPUT</u>: It corresponds to the question posed to an LLM system. If no RAG is applied, LLM responds to the question through standard decoding

2.INDEXING: To employ RAG, a set of reference documents are to be indexed.

- It involves chunking the documents, embeddings these chunks, and then indexing embeddings into a vector store.
- The input query is also embedded.
- 3.<u>RETRIEVAL</u>: Relevant documents are retrieved by comparing the query embedding against the document vectors.
- 4.<u>GENERATION</u>: Retrieved documents are first merged with the original prompt to provide additional context and then the LLM response generation is triggered:
 - This combined text and prompt is the input for response generation, that produces the final output provided to the user.

RAG models: the information flow



RAG models: the training task



Figure 1: Overview of our approach. We combine a pre-trained retriever (Query Encoder + Document Index) with a pre-trained seq2seq model (Generator) and fine-tune end-to-end. For query x, we use Maximum Inner Product Search (MIPS) to find the top-K documents z_i . For final prediction y, we treat z as a latent variable and marginalize over seq2seq predictions given different documents.

Vector Database

Vector database are data management systems focusing on the storage and efficient retrieval of unstructured data (texts, images or audio files,...)

Trough high-dimensional vector representations useful to efficiently retrieve similar information





Types of RAG









Modular RAG

Advanced RAGs

- It employs optimization across the (A) preretrieval, (B) retrieval, and (C) post-retrieval processes.
- A. The **PRE-RETRIEVAL PHASE** involves refining data indexing through five key stages:
 - enhancing data granularity,
 - optimizing index structures,
 - adding metadata,
 - alignment optimization, and
 - mixed retrieval



Advanced RAGs

- It employs optimization across the (A) pre-retrieval,
 (B) retrieval, and (C) post-retrieval processes.
- B. The RETRIEVAL PHASE involves optimizing the embedding model itself to maximize the quality of context chunks. Strategies may include:
 - fine-tuning embeddings to improve retrieval relevance or
 - employing dynamic embeddings that better capture contextual nuances (e.g., OpenAl's embeddings-ada-02 model)



Advanced RAG

Advanced RAGs

- It employs optimization across the (A) pre-retrieval,
 (B) retrieval, and (C) post-retrieval processes.
- C. The POST-RETRIEVAL PHASE focuses on circumventing context window limitations and managing noisy or distracting information. Re-ranking is a common approach to address these challenges, involving techniques such as
 - relocating relevant context to the edges of the prompt or
 - recalculating semantic similarity between the query and relevant text chunks.
 - Prompt compression techniques may also aid



Modular RAG

- SEARCH MODULE: Tailored for specific use-cases, it can perform direct searches on various corpora, utilizing LLM-generated code and query languages.
- MEMORY MODULE: Uses the LLM's memory for retrieval, improving alignment with data distributions.
- FUSION: Expands user queries into diverse perspectives, improving search results through multi-query approaches and re-ranking.
- **ROUTING**: Determines actions for queries, selecting the appropriate data source for retrieval.
- PREDICT: Uses the LLM to generate context instead of direct retrieval to reduce redundancy and noise.
- TASK ADAPTER: Adapts RAG to various tasks, enhancing universality and creating task-specific retrievers.



Retrieve

Read

ITER-RETGEN

Rerank

Read

Advanced RAC

Predict

xhartab at al 202

Modular RAG

Read

Naive RAG

RAG: Integrating Knowledge using more context

Retrieval Augmented Generation

- More contextually relevant information is made available at *generation time* in order to better qualify the answers
- Mandatory for knowledge intensive tasks
- Applicable at differest stages in integration with suitable *prompting* strategies
- Significant impact in mitigating the hallucinations





(Lewis et al, 2020) <u>Retrieval-augmented generation for knowledge-intensive NLP tasks. Proceedings of NIPS, Advances in</u> <u>Neural Information Processing Systems</u>, 2020.

RAG evaluation

- The evaluation of a RAG framework focuses on three primary quality scores and four abilities.
- QUALITY SCORES encompass measuring
 - context relevance (precision and specificity of retrieved context),
 - answer faithfulness (faithfulness of answers to retrieved context), and
 - answer relevance (relevance of answers to posed questions).
- Additionally, four abilities measure ADAPTABILITY AND EFFICIENCY of a RAG system:
 - noise robustness,
 - negative rejection,
 - information integration, and
 - counterfactual robustness.

The Foundational RAGs



Further augmentation for Generative Al

- Task-specific vs. Multi-task fine-tuning
- Instruction learning e prompt engineering

1) Chain of Thought	2) Tree of Thought	5) Expert Prompting	7) Rails	8) Automatic Prompt Engineering
Zero-Shot CoT Manual CoT	3) Self-Consistency	6) Chains	Topical Rails Fact Checking Reils	Prompt Generation
	4) Reflection		 Fact-Onecking Rails Jailbreaking Rails 	Refinement and Iteration
• Retriev#	al Augmented G	ENERATION	KGs	
 EXECUTION OF EXTERNAL TOOLS 			Re	trieved Facts
• LLM Ag	ents	Q: V is C	Vhich country Knowledge bama from? Retriever (Honolul	, Bornin, Honolulu) u, Locatedin, USA) → LLMs → A: USA
				•••

LLM agents

- The NL interpretation and problem solving capabilities of a single LLM can form the core engine of more sophisticated intelligent systems capable of:
 - composing basic intelligent functionalities
 - integrating structured intelligent behaviours (plans, strategies)
 - solving problems requiring complex processes



LLM agents

- Auxiliary Tools via dedicated APIs
- Decision-making capabilities:
 - Working Memory
 - Action execution engine
 - Utility functions





Shen, Yongliang, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. "HuggingGPT: Solving AI Tasks with ChatGPT and Its Friends in Hugging Face."

Working Memory: structure and roles

	 Sensory memory 	<pre>{ Iconic memory (visual) Echoic memory (auditory) Haptic memory (touch)</pre>	
Memory _≺	Short-term memory (Working memory)		
	Long-term memory	Explicit / Declarative memory Episodic memory (life events) (conscious) Semantic memory (facts, concepts)	

LLM agents

Examples

- \blacktriangleright ReWOO \rightarrow reasoning before Acting
- \blacktriangleright ReAct \rightarrow Dynamic problem solving
- ChatGPT + code interpreter, plugins

HuggingGPT



Binfeng Xu et al., 2023, "ReWOO: Decoupling Reasoning from Observations for Efficient Augmented Language Models"

ReWOO



Figure 1: Workflow of ReWOO. Given a question, Planner composes a comprehensive blueprint of interlinked plans prior to tool response. The blueprint instructs Worker to use external tools and collect evidence. Finally, plans and evidence are paired and fed to Solver for the answer.



Figure 4: Offloading foreseeable reasoning from GPT-3.5 into Alpaca 7B. A small LLaMa LM is fine-tuned on self-instructed data generated by GPT-3.5, producing Alpaca, endowed with general reasoning ability. Alpaca is then further fine-tuned on blueprints generated by GPT-3.5, leading to Planner 7B, a model specializing in foreseeable reasoning.

Star: Bootstrapping Reasoning



Figure 1: An overview of STaR and a STaR-generated rationale on CommonsenseQA. We indicate the fine-tuning outer loop with a dashed line. The questions and ground truth answers are expected to be present in the dataset, while the rationales are generated using STaR.

(1) Hotspot QA (1d) ReAct (Reason + Act) Question: Aside from the Apple Remote, what other device Thought 1: I need to search Apple Remote and find the program it was originally designed to interact with. can control the program Apple Remote was originally Act 1: Search [Apple Remote] designed to interact with? Obs 1: The Apple Remote is a remote control introduced in ReAct October 2005 by Apple ... originally designed to control the Front Row media center program ... (1a) Standard (1c) Act-Only Act 1: Search[Apple Remote] Answer: iPod Thought 2: Apple Remote was originally designed to control Obs 1: The Apple Remote is a the Front Row media center program. I need to search Front Row next and find what other device can control it. remote control ... (1b) CoT (Reason Only) Act 2: Search[Front Row] (1) Hotspot QA (1d) ReAct (Reason + Act) Overtion: Aside from the Apple Remote what other device _____ Thought 1: I need to search Apple Remote and find the (2) AlfWorld (2b) ReAct (Reason + Act) You are in the middle of a room. Looking quickly around Act 1: Think [First I need to find a pepper shaker...more likely to appear in cabinets (1-6), countertops (1-3), ... you, you see a cabinet 6, a cabinet 1, a coffee machine 1, Act 2: Go to cabinet 1 a countertop 3, a stove burner 1, and a toaster 1. Obs 2: On the cabinet 1, you see a vase 2. Your task is to: Put some pepper shaker on a drawer. (Here the agent go to cabinet 1, then cabinet 2, 3, then countertop 1 and 2) (2a) Act-Only Act 6: Go to countertop 3 Act 1: Go to drawer 1 Obs 6: On the countertop 3, you see a apple 1, a bread 1, Obs 1: The drawer 1 is closed. a pepper shaker 1, and a vase 3. Act 2: Open drawer 1 Act 7: Take pepper shaker 1 from countertop 3 Obs 2: You open the drawer 1. The drawer 1 is open. In it, Obs 7: You pick up the pepper shaker 1 from the countertop 3. you see a dishsponge 2, and a spoon 1. Act 8: Think [Now I find a pepper shaker 1. Next, I need to Act 3: Go to sinkbasin 1 put it in/on drawer 1.] Obs 3: On the sinkbasin 1, you see a dishsponge 3, a Act 9: Go to drawer 1 spatula 1, and a spoon 2. Obs 9: Drawer 1 is closed. Act 4: Take peppershaker 1 from sinkbasin 1 Act 10: Open drawer 1 Obs 4: Nothing happens. Obs 10: You open Drawer 1 ... Act 5: Take peppershaker 1 from sinkbasin 1 Act 11: Put pepper shaker 1 in/on drawer 1 Obs 5: Nothing happens. Obs 11: You put pepper shaker 1 in/on the drawer 1.

LLM agents

Examples

- \blacktriangleright ReWOO \rightarrow reasoning before Acting
- \blacktriangleright ReAct \rightarrow Dynamic problem solving
- ChatGPT + code interpreter, plugins

HuggingGPT



Shen, Yongliang, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. "HuggingGPT: Solving AI Tasks with ChatGPT and Its Friends in Hugging Face."

Multi-agent LLMs

Esempi

- \blacktriangleright BabyAGI \rightarrow Plans and execute tasks
- ► CAMEL e MAD (Multi-Agent Debate) → purely dialogue-driven, no tool nor human intervention
- They handle tasks of increasing complexity through the cooperation of multiple LLMs
- They can provide validating evidence for their results
- They encourage diversity of reasoning → closer to the typical human exploratory process
- They better manage the use of facts and explicit reasoning rules through the argumentation between agents

". Du, Yilun, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. "Improving Factuality and Reasoning in Language Models through Multiagent Debate." Liang, Tian, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. "Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate."
			Components	Content
	Multi age	nt LLMs	Question	When Alice walks up the hill, her speed is 1 m/s and when she goes down the hill, her speed is 3 m/s. Then when Alice walks up and down the hill, what is her average speed?
	B. Circle A rolls around circle B one trip back to its starting point. How many times will circle A revolve in total?		Correct Answer	1.5 m/s
	Chain-of-Thought Let the radius of circle A be r and the radius of circle B be 3r. The distance circle A travels along the circumfere of circle B is equal to the circumference of circle B, which is 6πr. The number of revolutions = (6πr) / (2πr) = 6 Of circle A will revolve 3 times in total as it rolls around circle B once. Self-Reflection I I disagree with you. To find the total number of revolutions, we not to consider both the rotation around circle B and the rotation of consider both the rotation around circle B and the rotation of consider both the rotation around circle B and the rotation of consider both the rotation around circle B and the rotation of consider both the rotation around circle B and the rotation of consider both the rotation around circle B and the rotation of consider both the rotation around circle B and the rotation of consider both the rotation around circle B and the rotation of consider both the rotation around circle B and the rotation of consider both the rotation around circle B and the rotation of consider both the rotation around circle B and the rotation of consider both the rotation around circle B and the rotation of consider both the rotation around circle B and the rotation of consider both the rotation around circle B and the rotation of consider both the rotation around circle B and the rotation of consider both the rotation around circle B and the rotation of consider both the rotation around circle B and the rotation of consider both the rotation around circle B and the rotation of consider both the rotation around circle B and the rotation of consider both the rotation around circle B and the rotation a		Explanation	If Alice covers a distance of d going up and down the hill, then her total distance is 2d. Her time going up the hill is $d/1 = d$, and her time going down the hill is d/3. So, her total time is $d + d/3= 4d/3$. Therefore, her average speed is 2d / (4d/3) = 3/2 m/s.
	formula for the circumference of a	and 1 time around circle B, making a total of 4 revolutions.	Incorrect Answer	2 m/s
	circle and applied it to both circles to find the number of revolutions. (3) I have correctly calculated the number of revolutions of circle A. Based on the results of my self- reflection, I am confident that my answer is correct. Therefore, circle A will revolve 3 times in total as it rolls around circle B once.	I see your point, but	Explanation	Alice's average speed can be cal- culated by adding her speed go- ing up the hill and her speed go- ing down the hill, and then divid- ing by 2. So, $(1 \text{ m/s} + 3 \text{ m/s})/2 =$ 2 m/s. Therefore, Alice's average speed is 2 m/s.
		The negative side correctly considers both the rotation of circle A around its own center and its rotation around circle B, while the affirmative side only considers the rotation around circle B. Therefore, the answer is 4.		

Liang, Tian, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. "Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate."

Table 2: An example in Counter-Intuitive AR dataset.

Multi agent LLMs

• Prompting for debates

Du, Yilun, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. "Improving Factuality and Reasoning in Language Models through Multiagent Debate."

User Input: A treasure hunter found a buried treasure chest filled with gems. There were 175 diamonds, 35 fewer rubies than diamonds, and twice the number of emeralds than the rubies. How many of the gems were there in the chest?

Round 1

Agent 1: Let x be the number of rubies. Then the number of diamonds is 175. And the number of emeralds is 2(x-35), since there are twice as many emeralds as rubies and there are 35 fewer rubies than diamonds.

To find the total number of gems, we add the number of diamonds, rubies, and emeralds: Total gems = 175 + x + 2(x-35) = 3x + 105.

We don't know the exact value of x, but we do know that it represents the number of rubies, which is less than 175 (since there are 35 fewer rubies than diamonds). Let's try some values of x: If x = 50, then the total gems = 3(50) +105 = 255. If x = 100, then the total gems = 3(100) + 105 = 405.

Since we know there were fewer rubies than diamonds, the answer must be closer to 255 than to 405. After some trial and error, we can find that x = 40X gives us: Total gems = 3(40) + 105 = 225.

User Input: Using the solutions from other agents as additional information, can you give an updated response.... (Copy Agent 2's reply) (Copy Agent 1's reply)

Round 2

Agent 1: Given the information provided in other agents' solutions, we have two answers: One agent found that the number of rubies is 40 and the total number of gems is 225. Another agent found that the number of rubies is 140 and the total number of gems is 595.

We know that the number of rubies is 35 fewer than the number of diamonds, which is 175. So the number of rubies should be 175 - 35 = 140. as the second agent found. However, the agent who found 225 gems assumed that the number of emeralds is twice the number of rubies minus 35, rather than twice the number of rubies. If we use this assumption, we get: Number of rubies = 140. Number of emeralds = 2 * 140 - 35 = 245. Total number of gems = 175 + 140 + 245 = 560. Our final answer is 560.

User Input: Using the solutions from other agents as additional information, can you give an updated response (Copy Agent 2's reply) (Copy Agent 1's reply)

Round 3

Agent 1: After reviewing the solutions provided by other agents, I agree with the second agent who found that the number of rubies is 140 and the total number of gems in the chest is 595. Therefore, my answer to the math problem is 595.

Agent 2: After reviewing the solutions provided by other agents. I also agree that the number of rubies is 140 and the total number of gems in the chest is 595. Therefore, my answer to the math problem is also 595.

Agent 2: If there were 175 diamonds,

then the number of rubies is 35 fewer.

And since there were twice as many

emeralds as rubies, we can represent the number of emeralds as 2 * 140 = 280

To find the total number of gems in the

175 + 140 + 280 = 595. Therefore, there

Agent 2: After reviewing the solutions

provided by other agents, I agree with

the second agent who found that the

number of rubies is 140 and the total number of gems in the chest is 595.

Therefore, my answer to the math

problem is 595.

chest, we can add up the number of diamonds, rubies, and emeralds;

were 595 gems in the chest.

or 175 - 35 = 140 mbies

emeralds.



 \checkmark

Figure 2: Illustration of Debate. Illustration of the debate procedure.

Multi-agent LLMs

100

75

50

25

0

Accuracy

74

66

Biographies

Esempi

- **BabyAGI** \rightarrow Plans and execute tasks >
- $MetaGPT \rightarrow Collaborative programming$ >
- **CAMEL** e MAD (Multi-Agent Debate) \rightarrow purely >dialogue-driven, no tool nor human intervention

5



Du, Yilun, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. "Improving Factuality and Reasoning in Language Models through Multiagent Debate." Liang, Tian, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. "Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate."

Conclusions (the view from here)