Exchange-of-Thought: Enhancing Large Language Model Capabilities through Cross-Model Communication

Zhuoyun Du
xiaodu.flying@gmail.com
THUNLP
2023.1.05

“Two heads are better than one.”

-English Proverb

沙龙提纲:

- I. 研究背景和动机
- A. 大语言模型在复杂推理任务中的应用
- B. 内在理解的局限性及外部见解的需求
- C. 引言Exchange-of-Thought (EoT)框架的提出

- II. Exchange-of-Thought (EoT)框架概述
- A. EoT框架的核心目标和原理
- B. EoT如何促进不同模型之间的交叉通信
- C. EoT如何整合四种独特的通信范式: Memory, Report, Relay, and Debate

III. 通信范式的详细解释
A. Memory范式
B. Report范式
C. Relay范式
D. Debate范式

IV. EoT框架的实验设计和结果
A. 实验目的和设置
B. 实验结果和对比分析
C. 实验结果的启示和意义

V. EoT框架的未来展望
A. EoT框架的潜在应用领域
B. EoT框架的改进和扩展方向

背景和动机，良好的开始是成功的一半

Background and motivation: A good start is half the battle to success.

- Where are we now?

GPT-4 and lots of other LLMs has achieved exemplary performance across a wide range of NLP tasks.

- What we can anticipate is?

The potential applications of LLMs are immeasurable. Tasks like reasoning demand LLMs to possess high levels of reasoning and comprehension abilities.

- What is troubling us?

It is crucial not to overlook the inherent limitations in the understanding of LLMs in complex reasoning tasks.

And this limitation cannot be overcome solely by increasing the size of models.

https://arxiv.org/abs/2312.01823

背景和动机，良好的开始是成功的一半

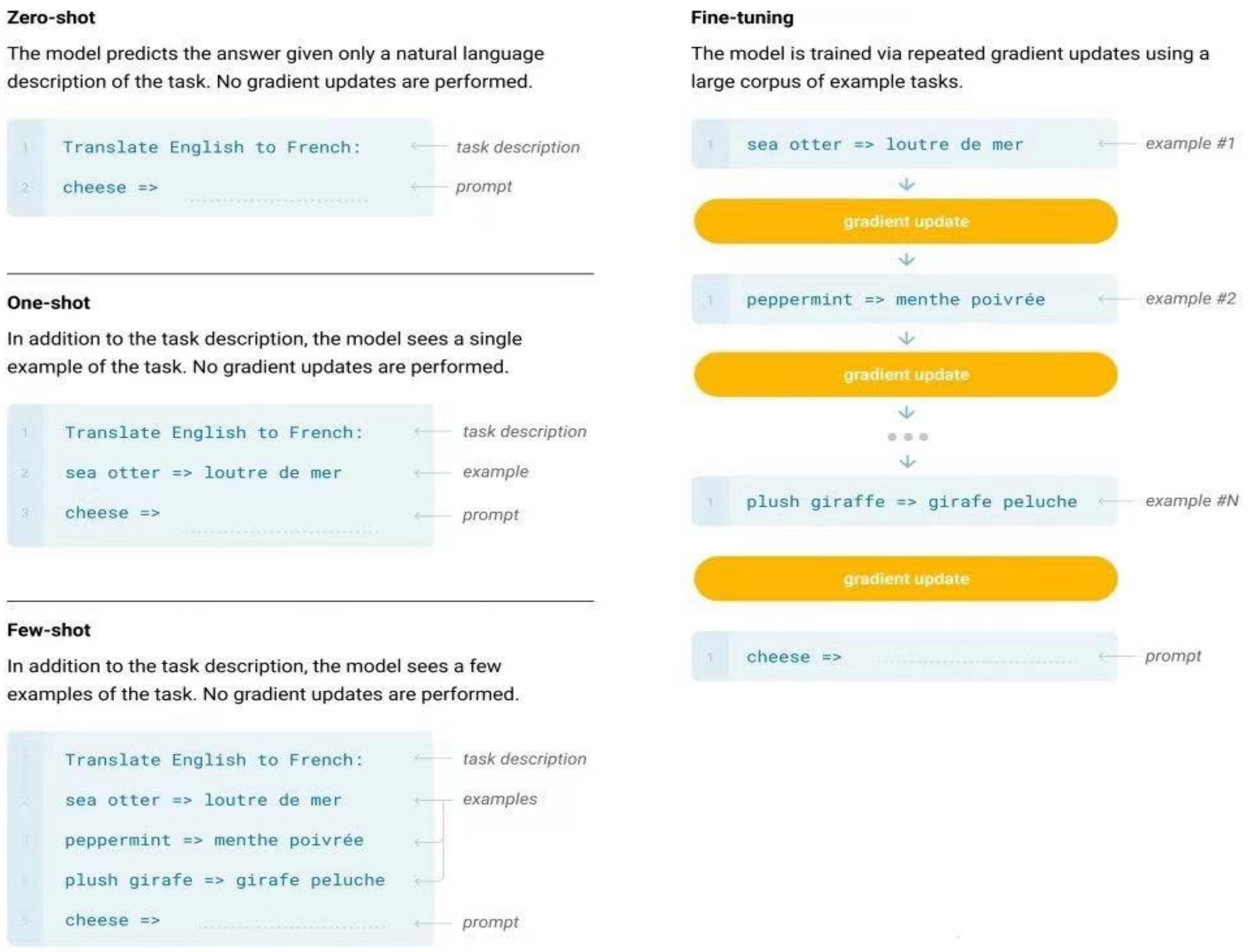
Background and motivation: A good start is half the battle to success.

- What have we accomplished so far?
- Chain of Thought (CoT)

Guide the model to generate a series of intermediate reasoning steps before reaching the final answer. (Wei et al. 2022b)

- Self-Correction

Iteratively improve the quality of answers by leveraging the model's feedback to their previous outputs.



A pic from: *https://zhuanlan.zhihu.com/p/629087587*

背景和动机，良好的开始是成功的一半

Background and motivation: A good start is half the battle to success.

- The problems associated with the above-mentioned methods are:
- Everyone's own understanding has their inherent limitations.

LLMs using CoT and self-correction still struggle to revise their responses without external feedbacks.

- Profound insights are hard to come by.

Despite single or multiple reasoning chains, when confronted with difficult questions, the model often yields a higher bumber of incorrect respense.

Got any proof? Let refer to the next slice.

背景和动机，良好的开始是成功的一半

Background and motivation: A good start is half the battle to success.

- Pilot experiments

In Figure 2, the analysis of correct and

incorrect answers within erroneous samples from

three reasoning datasets reveals that in most cases

the model can deduce the correct answer but still with many error answers.

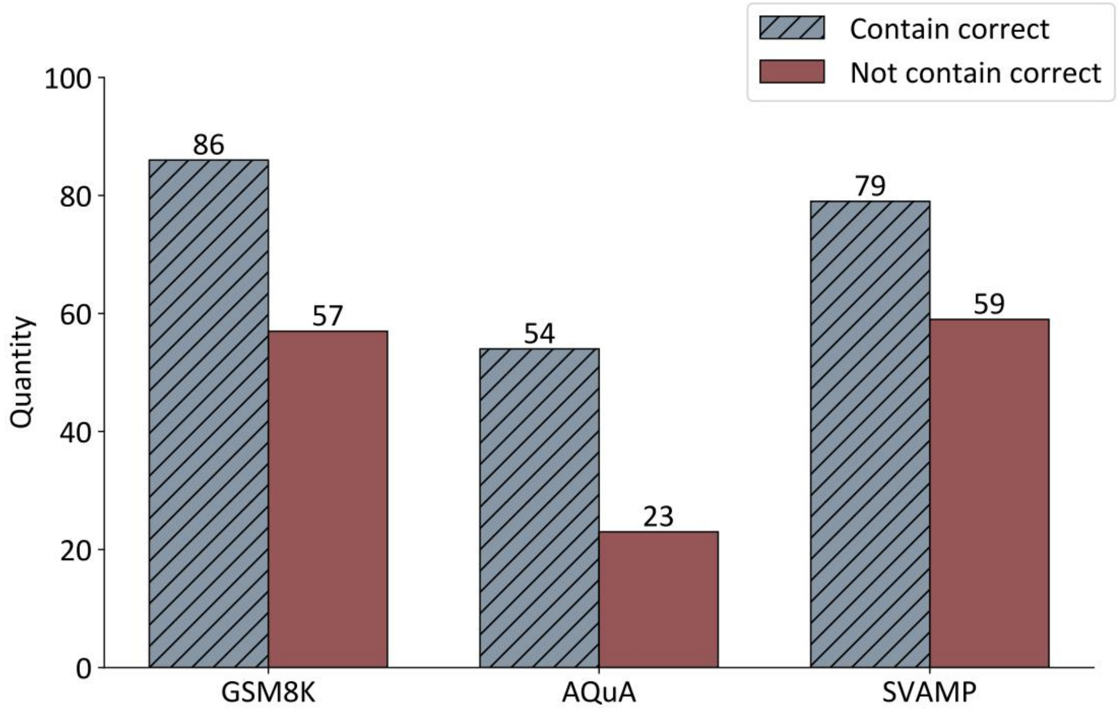


Figure 2: Pilot experiments on three reasoning datasets. The number of erroneous samples containing the correct answer is significantly higher than those not containing the correct answer.

A pic from the corresponding paper.

“Truth will ultimately prevail where there is pains to bring it to light.”

-English Proverb

背景和动机，良好的开始是成功的一半
Background and motivation: A good start is half the battle to success.

- Our firsthand experiences speak volumes.

In human society, the truth, even when held by a minority, can gain widespread acceptance and recognition through clear and persuasive communication (Le Bon, 1897).

The correct reasoning of others can serve as high-quality external insights, enriching and elevating our collective understanding.

Finally! Here comes EoT.

什么是 EoT?
What is EoT?

In a nutshell

EoT is a novel framework designed to facilitate cross-model communication, allowing for the exchange of reasoning processes to integrate external insights.

Is about communication!

什么是 EoT?
What is EoT?

Figure 1 contrasts EoT with CoT and self-correction methods.

Highlighting the unique appr-oach of EoT in integrating external perspectives.

EoT enhances the model's reasoning ability by incorporating the thoughts of other models as external insights.

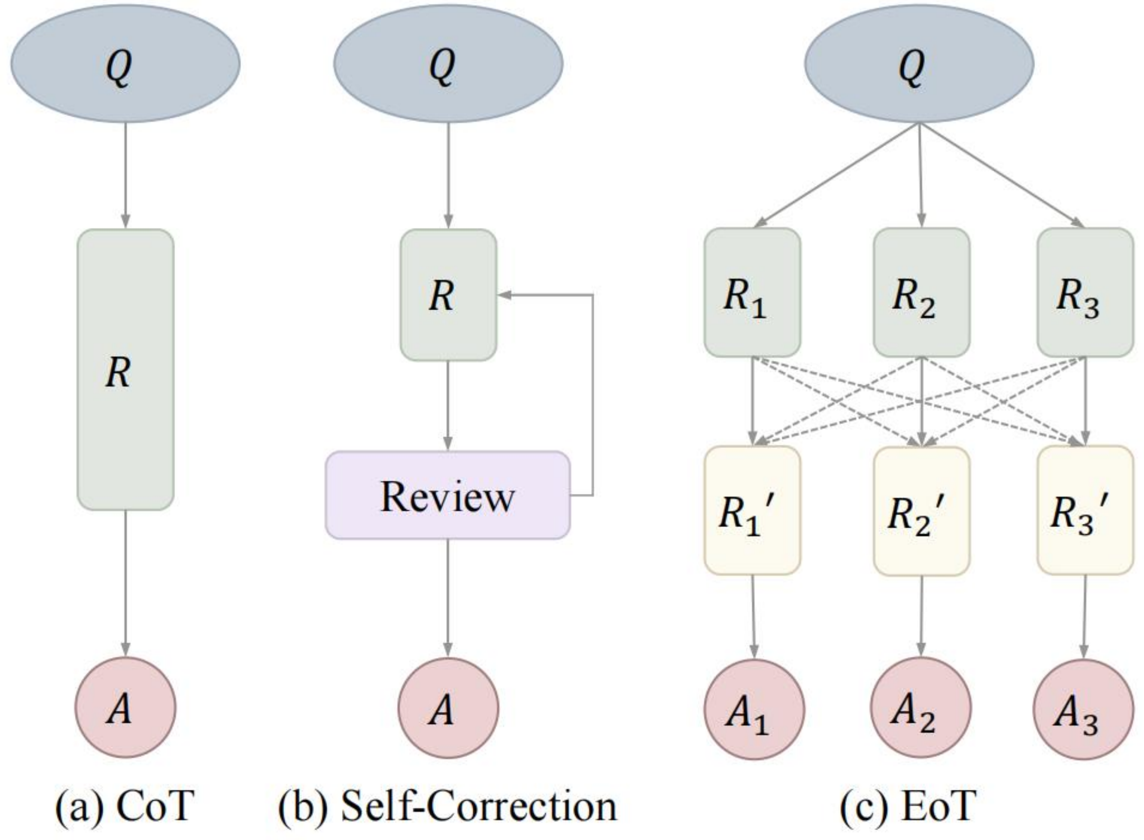


Figure 1: Comparison of CoT, Self-Correction, and EoT. Both CoT and Self-Correction rely on the model's innate abilities to generate and refine output, lacking external insights. EoT enhances the model's reasoning ability by incorporating the thoughts of other models as external insights.

A pic from the corresponding paper.

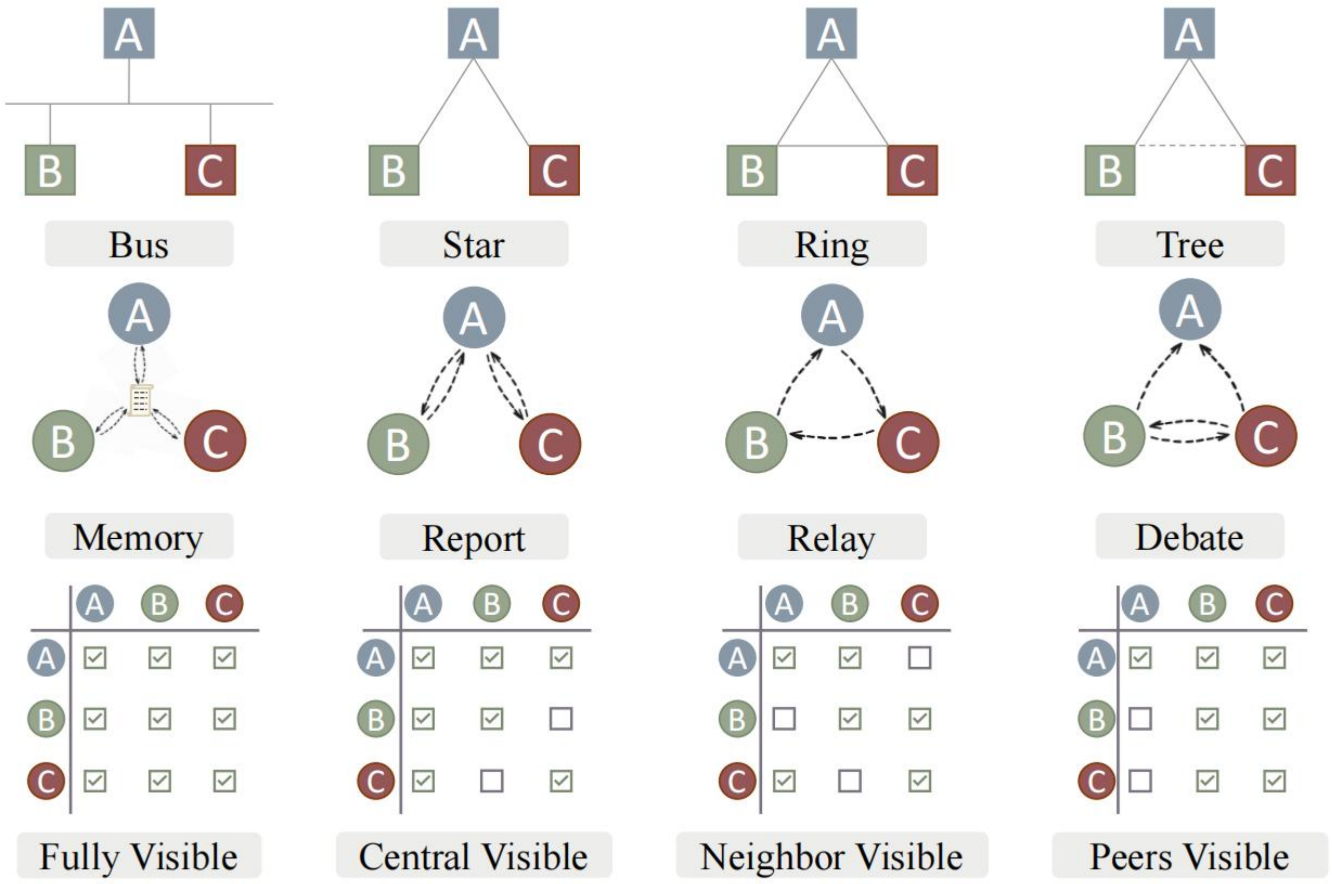
什么是 EoT?
What is EoT?

Meanings:

-Facilitate the exchange of ideas and reasoning chains among models.

-Enriching the diversity of insights

A pic from the corresponding paper.



Inspired by the principles of network topology (Bisht and Singh, 2015) and agent communication (Parsons and McBurney, 2003), there are four communication paradigms:

Memory, Report, Relay, and Debate.

最后的准备工作，稍安勿躁

The final preparations – stay calm and steady。

- This paper denote a LLM with a parameter size of θ as pθ, and the sequence length as t, which includes tokens [s1, s2, . . . , st].
- The LLM predicts the next token based on theprior tokens in the sequence. The probability of the si token is pθ(si ∣ s1, s2, . . . , si−1).
- Therefore,the probability of the whole sentence is $\prod_{i=1}^{t} p_\theta(s_i|s_{\leq i-1})$.

最后的准备工作，稍安勿躁
The final preparations – stay calm and steady。

- Standard prompting.

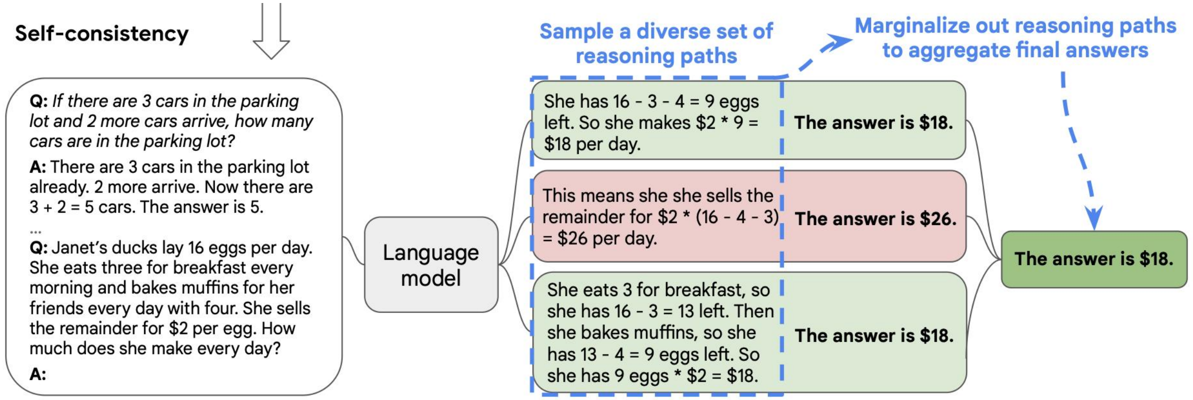
This involves deriving an answer a from a question q using $p\theta(a \mid q)$. In-Context Learning aims to improve LLMs performance by adding demonstrations D = {d1, d2, . . . , dn} to the input, which can be expressed as $p_\theta(a|D, q)$.

- CoT prompting.

A rationale ri is added to demonstration di = {qi, ri, ai} to guide the LLMs in explicitly generating reasoning steps.

最后的准备工作，稍安勿躁

The final preparations – stay calm and steady.



- Self-Consistency.

This technique prioritizes the most commonly occurring answer, defined as $a = argmax_{a_i} f(a_i)$ where f(ai) denotes the frequency of each answer ai.

- Progressive-Hint Prompting.

Introduced by Zheng et al. (2023), Progressive-Hint Prompting PHP) leverages a sequence of historical answers $\{a^{(1)} ..., a^{(j-1)}\}$ to enhance the current reasoning process $r^{(j)}$ and facilitate the derivation ofthe subsequent answer $a^{(j)}$.

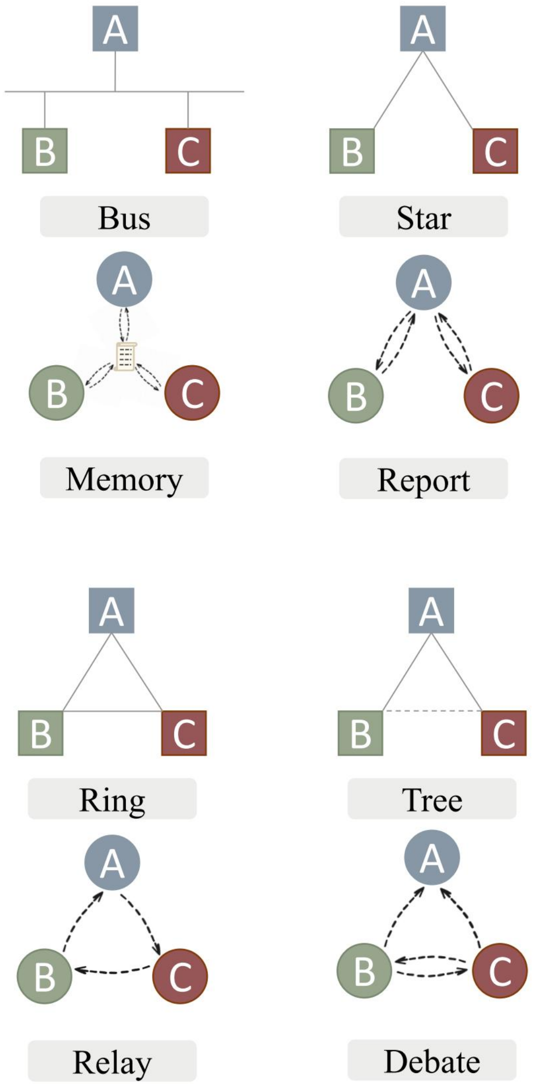
最后的准备工作，稍安勿躁

The final preparations – stay calm and steady。

- Communication Paradigm

in Figure 3, we propose Memory, Report, Relay, and Debate communication paradigms each corresponding to the Bus, Star, Ring, and Tree network topologies, respectively. Assume in j-th round of communication, given a set of LLMs $\{M\} = \{m_1, m_2, \ldots, m_n\}$, the model m_i generates the corresponding rationale $r_i^{(j)}$ and the answer $a_i^{(j)}$ based on the $(r_K^{(j-1)}, a_K^{(j-1)})$, where K is the set from which model m_i can receive reasoning processes. In the first round, we use the CoT method proposed by Wei et al. (2022b) to generate $(r^{(1)}, a^{(1)}) \sim P_\theta(r^{(1)}, a^{(1)} | D, q)$.

- Just in case you forgot



That is what a nice guy I am.

“All roads lead to Rome.”

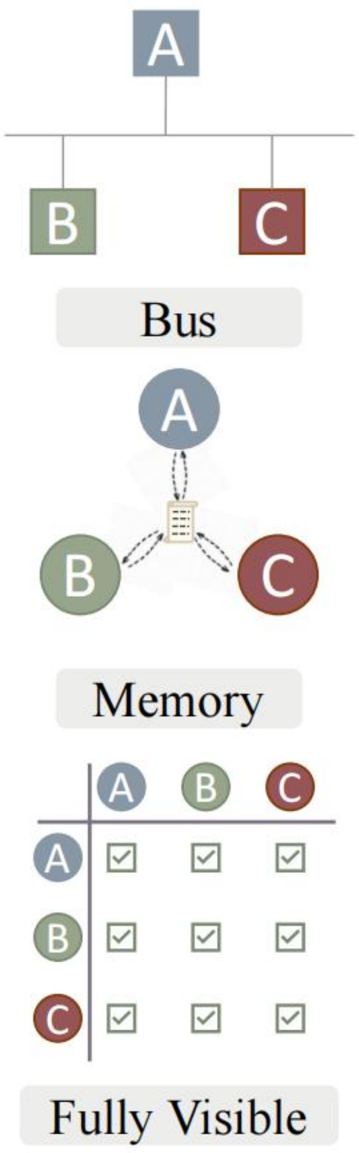
-Spanish Proverb

四种范式，各有特色
Four paradigms, each with its own distinctive features.

- Memoy

Any model, can access the reasoning chains and answers from all models.

This paradigm facilitates the <u>fastest flow of information</u> and also incurs the <u>highest communication cost.</u>



四种范式，各有特色

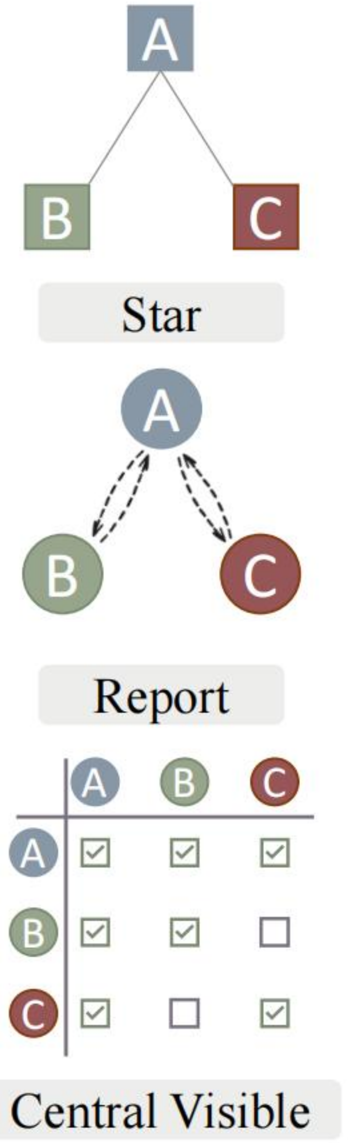
Four paradigms, each with its own distinctive features.

- Report

There is a model mA as the central node, which can obtain the rationale and answer from all other models.

Both mb and mc only receive information from mA and do not interact with each other.

Allows for rapid information flow, but it demands a higher capacity for processing and analysis for the central node.



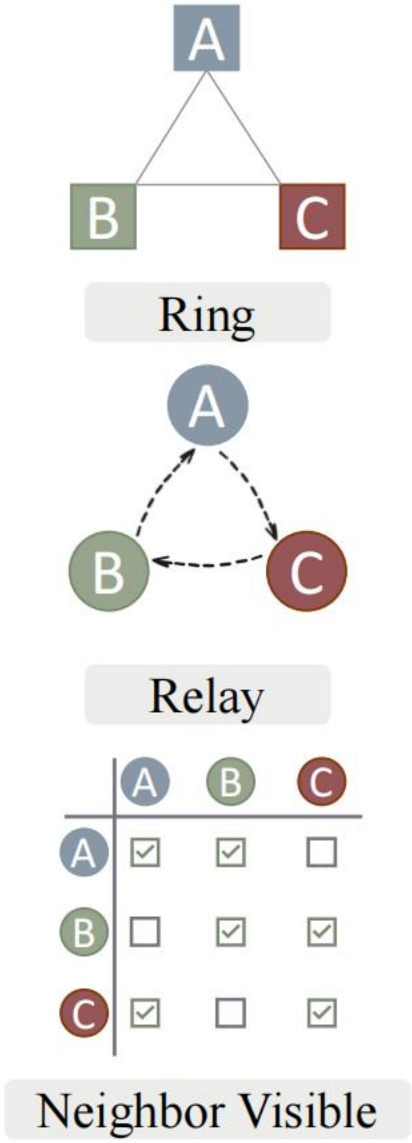
四种范式，各有特色

Four paradigms, each with its own distinctive features.

- Relay

Each node is capable of receiving information from the preceding node and transmitting its own information to the subsequent node.

This mode can reduce the demands on the information processing capacity of each node, but it may result in a slower flow of information.



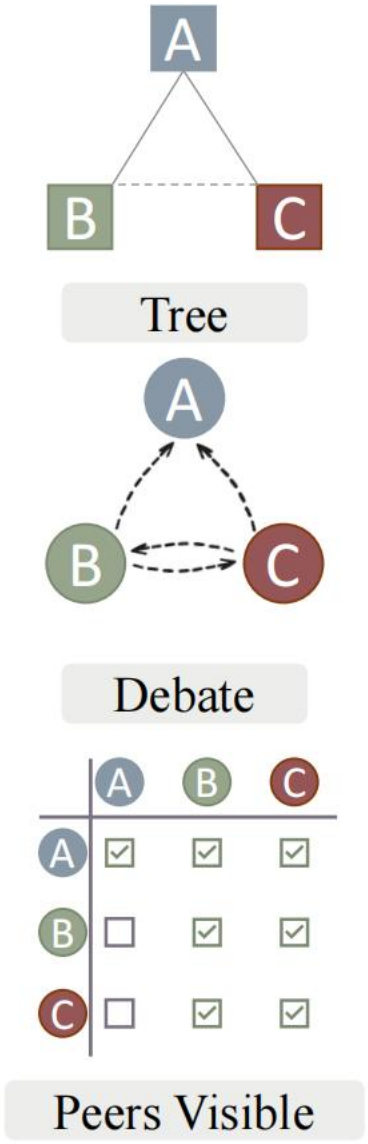
四种范式，各有特色

Four paradigms, each with its own distinctive features.

- Debate

This mode adapted the tree topology to devise the Debate paradigm which permits leaf nodes to exchange information with each other, while parent nodes are solely responsible for aggregating information. Information flow is directed upward from child to parent.

This communication paradigm strikes a balance between the model's information processing capacity and the speed of information flow.



“Don't jump to conclusions.”

-English Proverb

缜密计算，有条不紊

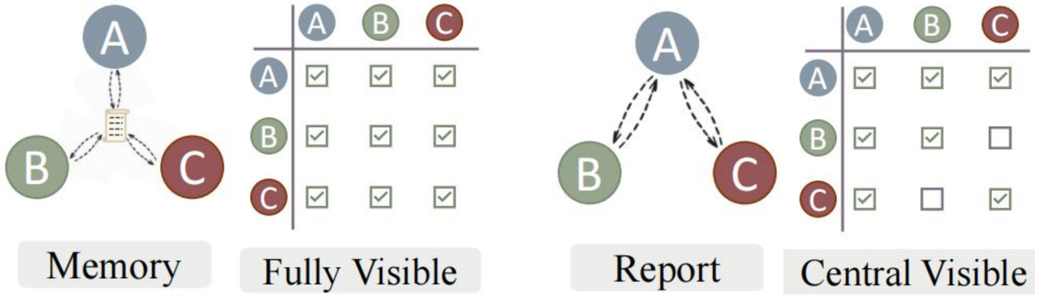
Precise calculations, methodical and orderly.

- Communication Volume

Measured by the number of messages received, assuming there are n models.

- Memoy

Every node receives information from all other nodes, resulting in a communication volume of.



- Report

The central node receives information from n – 1 non-central nodes, while each of the n – 1 non-central nodes receives, information from the central node. In addition, each node can receive information from its previous round. Thus we have:

$(n - 1) + (n - 1) + n = 3n - 2$

Average volume for each node is:

$\{ (3n - 2) - n \} / n = 2 - 2/n$

缜密计算，有条不紊

Precise calculations, methodical and orderly.

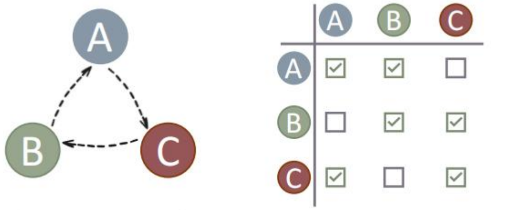
- Relay

Each node receives information from the preceding node and its own information from the last round, resulting in a communication volume of 2n

Average volume for each node is:

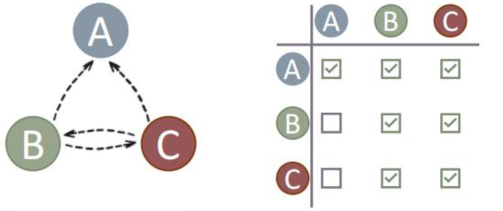
(2n - n) / 2 = n / 2

The one on the right looks quite challenging.



Relay

Neighbor Visible



Debate

Peers Visible

- Debate

The communication volume for each pair of child nodes is 4, and it is 3 for the parent node. Consequently, a subtree’s communication volume of 7.

The number of non-leaf nodes in a full

binary tree is (n–1) / 2, leading to a total volume of 7(n–1) / 2

- Average volume for each node is:

Information under the same parent node requires only one transmission. Information from the farthest nodes needs h – 1 transmissions. Thus we have:

$$\mathcal{S} = \frac{\sum_{i=1}^{h-1} 2^{i-1} i}{2^{h-1} - 1}$$

“All good things must come to an end.”

-English Proverb

终止条件，天下没有不散的宴席

Termination condition: all good things must come to an end.

- There's more than one way to skin a cat.

- Consistent Output Termination

When the output of model in the j-th round is the same as the output in the j – 1-th round.

- Majority Consesus Termination

LLMs cease communication with each other once a majority of them reach an agreement.

“Think twice before acting.”

-Chinese Proverb

可置信度评估机制

Confidence evaluation mechanisms.

- In a communication with k rounds, model migenerates a set of answers $\{a_i^{(1)}, ..., a_i^{(k)}\}$.Let $f(a_i) = max\#\{a|a = a_i^{(j)}\}$ denote the number of the most frequently occurring answer from model mi. Consequently, we obtain the model's confidence level $c_i = \frac{f(a_i)}{k}$ in the current round.

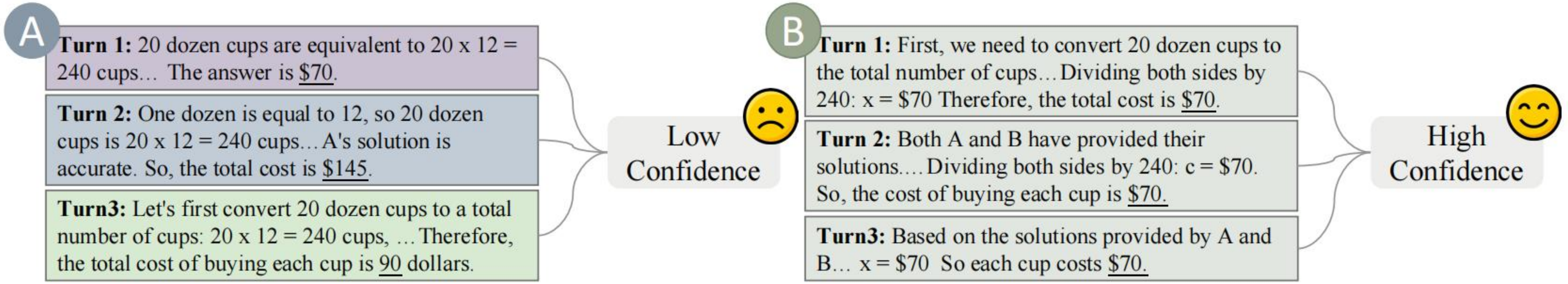


Figure 4: An illustrative comparison between a confident model and an unconfident model. Model A generates three different answers over three communication rounds, indicating uncertainty about the answer, while Model B consistently adheres to a single answer.

A pic from the corresponding paper.

Crescendo!
Here comes experments

在实验中验证想法
Validate ideas through experimentation.

- Tasks and Datasets
- Mathematical Reasoning

GSM8K MultiArith SingleEQ AddSub AQuA and SVAMP

SingleEq and AddSub involve relatively simple problems that do not require multi-step calculations. MultiArith, AQUA, GSM8k, and SVAMP, are more challenging datasets that demand multi-step reasoning to solve.

- Commonsense Reasoning

CommonsenseQA and StrategyQA

StrategyQA is a question-answering focused on open-domain questions, where the required reasoning steps are implicit in the question. CommonsenseQA have been introduced to explore the commonsense understanding, involving yes/no questions (or assertions).

- Symbolic Reasoning

Pengui and DateUnderstanding

在实验中验证想法
Validate ideas through experimentation.

- Baseline
- Chain of Thought (CoT)
- ComplexCoT
- Self-Consistency (SC)
- Progressive Hint Prompting (PHP)

Details
-temperature = 1
-Use GPT-3.5, while may incorporate Claude-2
-Results are the average performance and standard deviation across five runs.

- For simplicity, CoT-SC(10) is denoted the approach that employs the CoT prompt method to sample 10 reasoning chains and then utilize the SC method to select the answer

A table from the corresponding paper.

Method	GSM8K	MultiArith	SingleEQ	AddSub	AQuA	SVAMP	Avg.
Single Reasoning Chain							
CoT	79.12±0.50	97.27±0.65	92.80±0.27	86.23±0.52	55.12±1.03	79.52±0.81	81.67
ComplexCoT	79.32±0.65	95.40±0.50	91.34±0.33	84.46±0.86	56.46±0.59	77.70±0.54	80.78
CoT (GPT-4)	**94.90**	97.80	93.10	**89.30**	**77.50**	**90.50**	90.51
Ensemble Methods							
CoT-SC(3)	82.82±0.32	98.20±0.43	93.31±0.12	87.19±0.47	62.13±1.30	81.98±0.49	84.27
CoT-SC(5)	85.47±0.52	98.60±0.08	93.70±0.25	87.49±0.38	64.02±0.95	83.76±0.81	85.50
CoT-SC(10)	87.57±0.27	98.97±0.12	94.06±0.36	87.59±0.58	66.38±1.72	84.96±0.33	86.59
ComplexCoT-SC(3)	84.17±0.67	97.43±0.31	92.95±0.53	86.13±0.74	60.47±1.55	81.44±0.79	83.77
ComplexCoT-SC(5)	87.26±0.33	98.13±0.22	94.02±0.29	86.48±0.61	62.05±2.40	83.86±0.92	85.30
ComplexCoT-SC(10)	89.23±0.31	98.23±0.37	94.21±0.16	86.58±0.58	64.96±1.91	85.58±0.87	86.46
PHP	85.10	98.00	92.90	85.30	60.60	83.10	84.16
Exchange-of-Thought							
EoT-Memory	88.98±0.89	98.80±0.16	94.09±0.48	87.65±0.49	69.37±2.77	84.28±0.48	87.20
EoT-Report	88.61±0.83	**99.03±0.22**	94.06±0.47	87.95±0.34	70.31±2.19	84.78±0.75	87.46
EoT-Relay	88.42±0.72	98.97±0.16	94.13±0.49	87.59±0.58	70.87±1.98	85.04±0.31	87.50
EoT-Debate	88.52±0.76	98.90±0.17	**94.25±0.19**	87.70±0.34	69.69±1.24	85.10±0.24	87.36

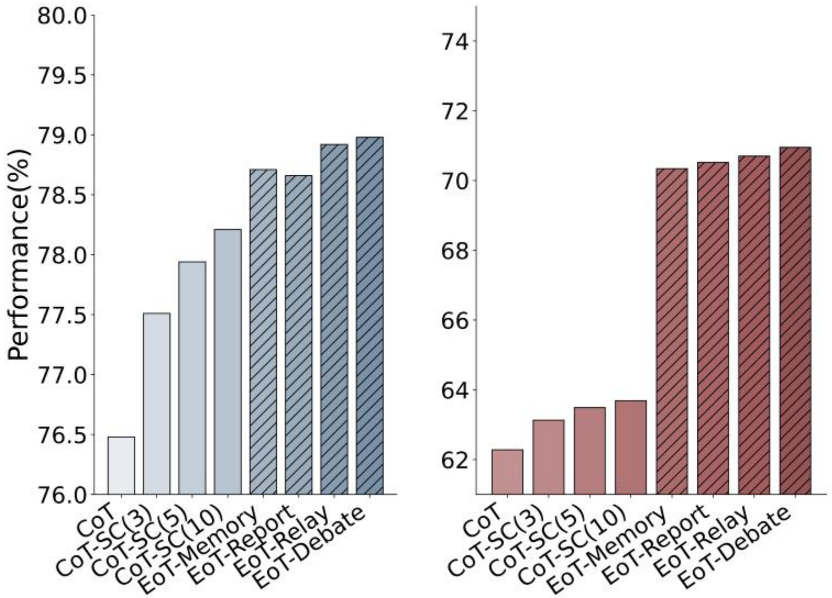
Mathematical Reasoning

- EoT has shown great improvement over CoT even surpassing strong baseline.
- Three GPT-3.5 with EoT surpassed a single GPT-4 with CoT.
- Addressing inherent shortcomings by incorporating external insights.

在实验中验证想法

A pic from the corresponding paper.

Validate ideas through experimentation.



(a) CSQA. (b) StrategyQA.

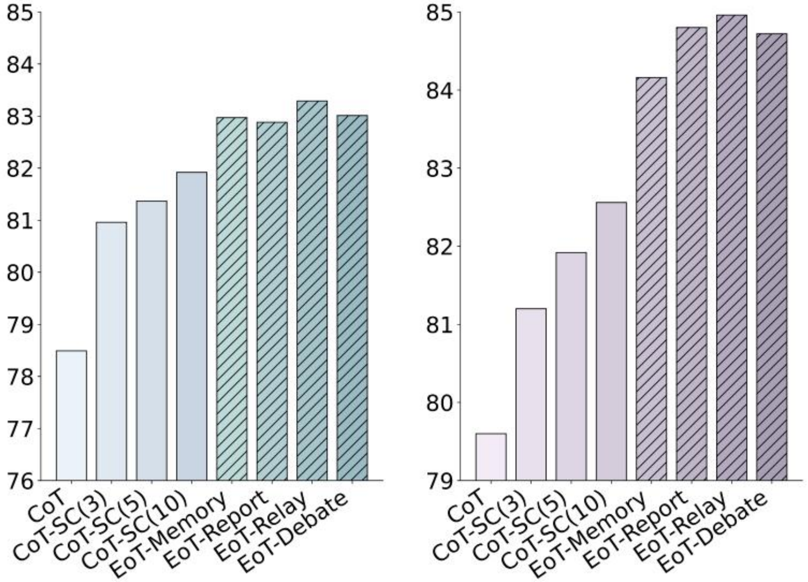
Commonsense Reasoning

- EoT shows significant outperformance over CoT, particularly on the StrategyQA dataset.
- Similar noteworthy gains are observed on the CSQA dataset.
- All four paradigms demonstrate superior performance compared to the CoT-SC(10)

在实验中验证想法

A pic from the corresponding paper.

Validate ideas through experimentation.



(c) Peguins.

(d) Date Understanding.

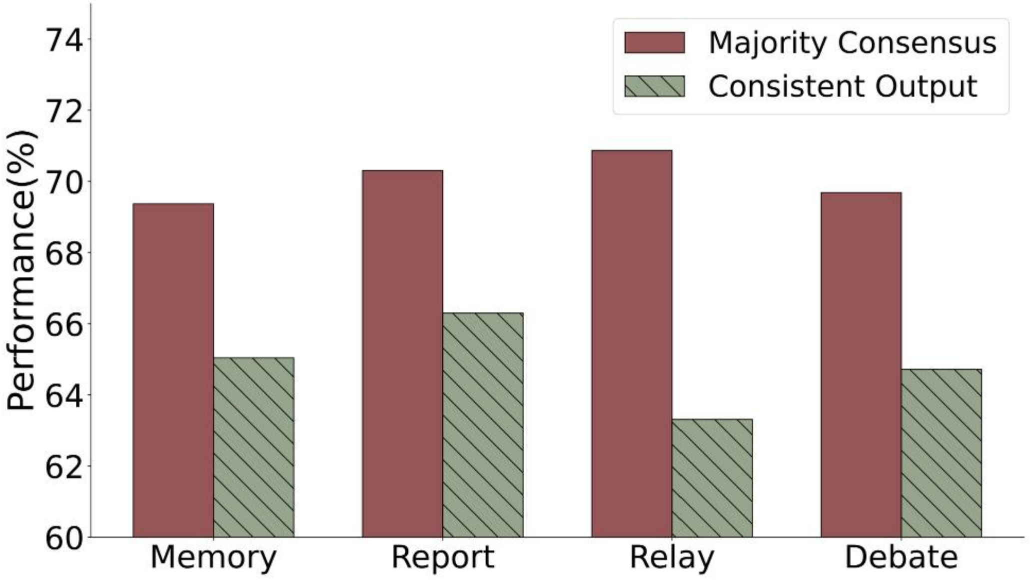
Symbolic Reasoning

- On the Penguins dataset, EoT exhibit improvements compared to the CoT-SC.
- For the Date Understanding dataset, EoT shows even more significant performance gains, with all four paradigms averaging a 2.1% improvement over CoT-SC(10).

A pic from the corresponding paper.

在实验中验证想法

Validate ideas through experimentation.



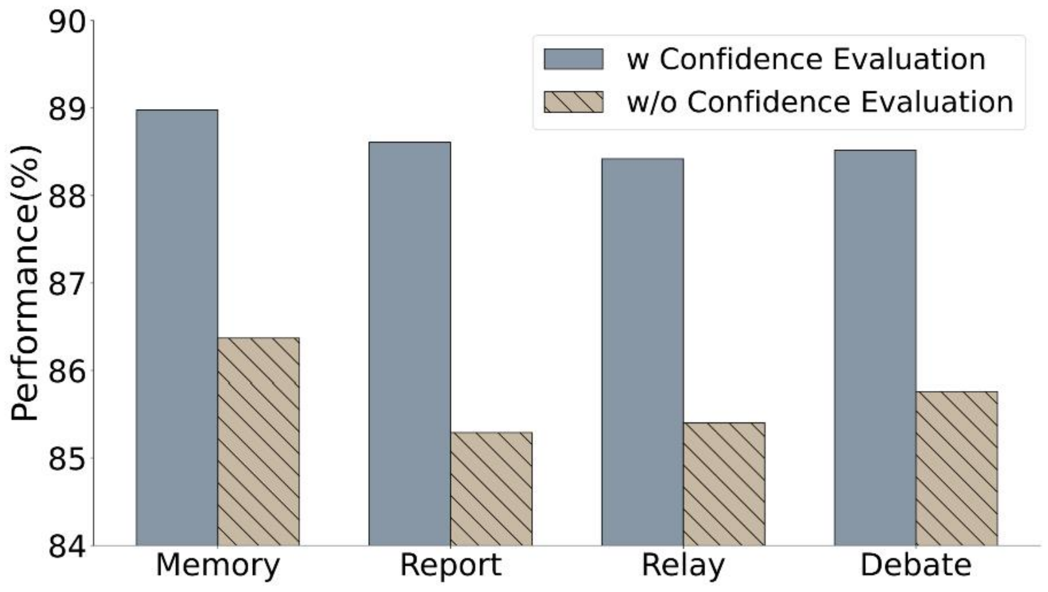
Termination Condition

- Majority consensus termination, compared to consistent output termination, shows notable improvements.
- Consistent output termination lacks a mechanism for collective negotiation, making individual models susceptible to premature exit due to degeneration. Therefore, majority consensus termination is deemed more suitable for scenarios involving multiple model communication.

在实验中验证想法

A pic from the corresponding paper.

Validate ideas through experimentation.



Confidence Evaluation

- Confidence evaluation demonstrates an average improvement of 2.92% compared to the baseline.
- It facilitates the decision to accept the other model's reasoning chains at an earlier stage, effectively mitigating the interference of incorrect reasoning chains.

在实验中验证想法
Validate ideas through expe

A pic from the corresponding paper.

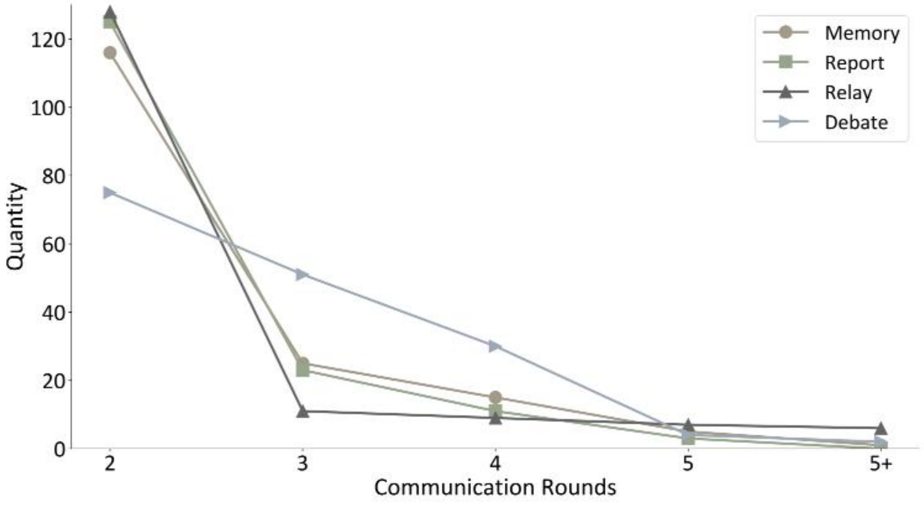


Figure 8: Number of communication rounds required to reach termination condition on SVAMP.

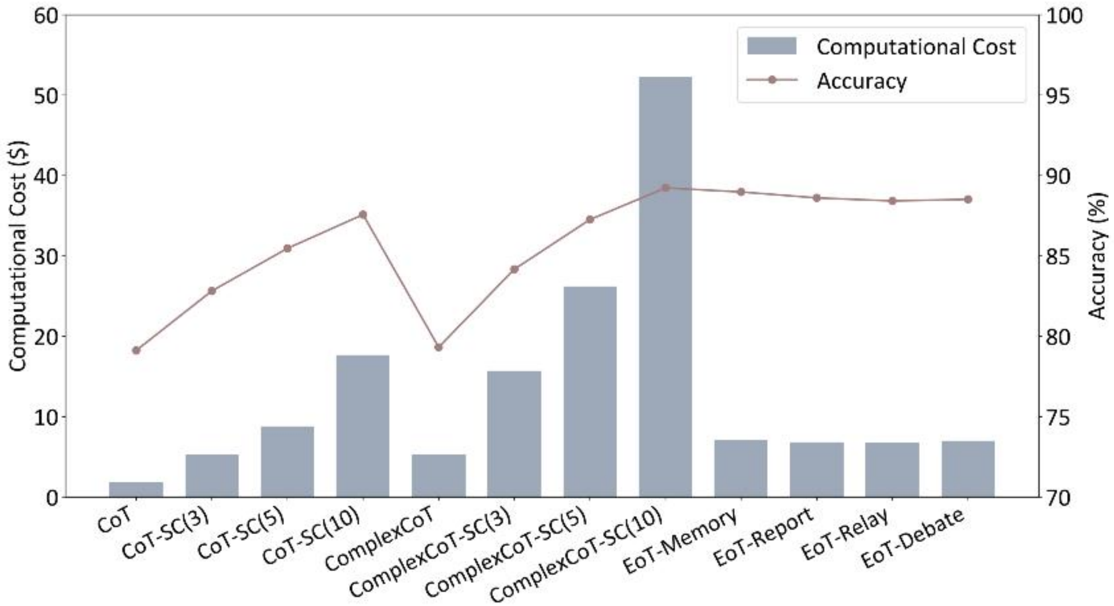
Round Analysis

- For the majority of samples, consensus on the answer can be reached within three rounds of communication.

在实验中验证想法

A pic from the corresponding paper.

Validate ideas through experimentation.



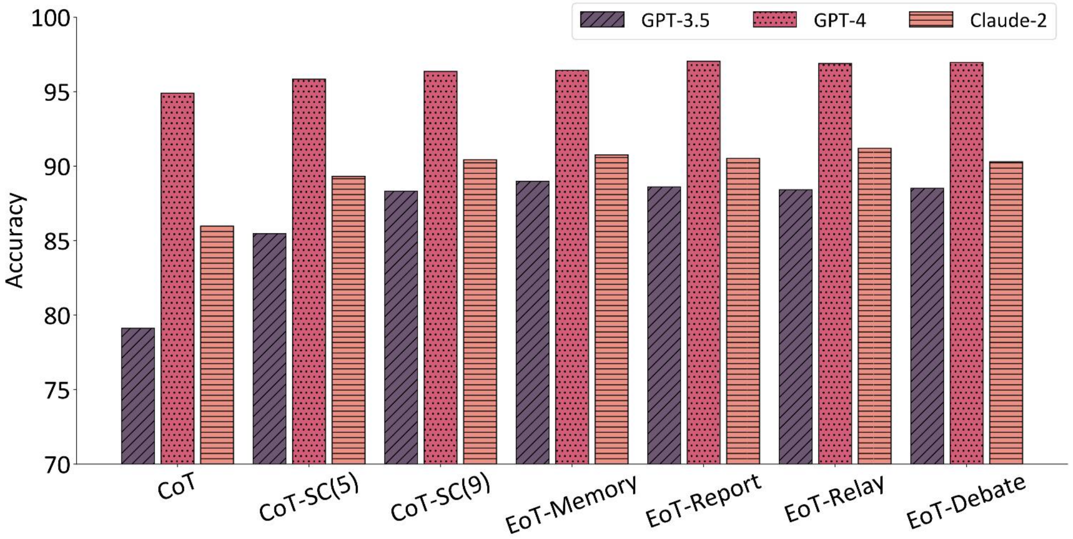
Cost Analysis

- Compared to CoT-SC(5), EoT reduces costs by 20% while improving performance by 3%. EoT achieves performance similar to ComplexCoT-SC(10) at only one-seventh of its cost.
- Given that the majority of samples conclude communication within three rounds, EoT does not impose a significant computational burden.

在实验中验证想法

A pic from the corresponding paper.

Validate ideas through experimentation.



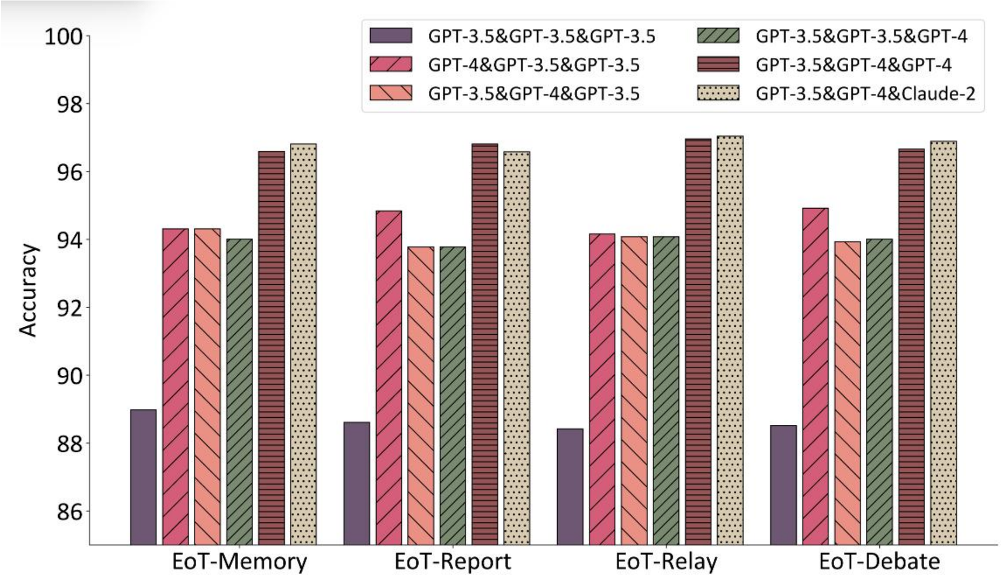
Model Applocability

- Compared to CoT-SC(5), EoT demonstrates performance improvements of 3.2% on GPT-3.5, 1.0% on GPT-4, and 1.4% on Claude-2.
- The results indicate that EoT is adaptable to various LLMs and effectively enhances performance across multiple models.

A pic from the corresponding paper.

在实验中验证想法

Validate ideas through experimentation.



Position Analysis

- Position of GPT-4 may have influence depends on the paradigm.
- A configuration with two GPT-4 models and one GPT-3.5 significantly outperforms one with two GPT-3.5 models and one GPT-4.
- Model diversity effectively boosts EoT's effectiveness.

Show down!
Let’s drop the conclusions

总结

Conclusion.

- The introduction of Exchange-of-Thought (EoT), a novel framework facilitating cross-model communication to enrich models with external insights.
- The framework includes four communication paradigms, and an in-depth analysis covers communication volume and information propagation speed.
- To address potential disruptions from incorrect reasoning, a confidence evaluation mechanism is incorporated.
- Experimental results across mathematical, commonsense, and symbolic reasoning tasks demonstrate EoT's superiority over strong baselines with a cost advantage. Further investigations highlight EoT's adaptability to various models, and the involvement of a diverse set of models enhances its overall performance.

一些未来展望

Future Outlook.

- Future Outlook:
- The EoT framework may find applications in a broader range of natural language processing tasks, including but not limited to text generation, question-answering systems, and dialogue systems.
- The EoT framework could be applied in the field of education, assisting students in tasks involving mathematical reasoning, logical inference, and other complex reasoning tasks. Improvement and Expansion Directions:
- Further optimize communication paradigms:
- Exploring additional communication paradigms or refining existing ones to adapt to different types of tasks and interactions between models.
- Consider model diversity: Researching how to introduce a greater variety of model types to increase the diversity of external insights, further enhancing the performance of the EoT framework.
- Consider real-time applications: Investigating how to apply the EoT framework in real-time scenarios, such as dialogue systems or real-time inference tasks, to validate its effectiveness and feasibility in practical applications.

“Success is not final, failure is not fatal: It is the courage to continue that counts.”

-Winston Churchill

引用
References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *ArXiv preprint*, abs/2204.05862.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022b. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

BIG bench authors. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Nivedita Bisht and Sapna Singh. 2015. Analytical study of different network topologies. *International Research Journal of Engineering and Technology (IRJET)*, 2(01):88–90.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. 2023. Extending context window of large language models via positional interpolation.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. 2022. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *ArXiv preprint*, abs/2211.12588.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2023. A survey of chain of thought reasoning: Advances, frontiers and future.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *ArXiv preprint*, abs/2210.11416.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems.

Shizhe Diao, Pengcheng Wang, Yong Lin, and Tong Zhang. 2023. Active prompting with chain-of-thought for large language models. *ArXiv preprint*, abs/2302.12246.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *ArXiv preprint*, abs/2305.14325.

Yao Fu, Hao Peng, Tushar Khot, and Mirella Lapata. 2023a. Improving language model negotiation with self-play and in-context learning from ai feedback.

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2023b. Complexity-based prompting for multi-step reasoning. In *The Eleventh International Conference on Learning Representations*.

Deep Ganguli, Amanda Askell, Nicholas Schiefer, Thomas Liao, Kamilė Lukošiūtė, Anna Chen, Anna Goldie, Azalia Mirhoseini, Catherine Olsson, Danny Hernandez, et al. 2023. The capacity for moral self-correction in large language models. *arXiv preprint arXiv:2302.07459*.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2022. Pal: Program-aided language models. *ArXiv preprint*, abs/2211.10435.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

David Ha and Yujin Tang. 2022. Collective intelligence for deep learning: A survey of recent developments. *Collective Intelligence*, 1(1):26339137221114874.

Mohammad Javad Hosseini, Hannaneh Hajishirzi, Oren Etzioni, and Nate Kushman. 2014. Learning to solve arithmetic word problems with verb categorization. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 523–533. Association for Computational Linguistics.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023. Large language models cannot self-correct reasoning yet.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2023. Decomposed prompting: A modular approach for solving complex tasks. In *The Eleventh International Conference on Learning Representations*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*.

Rik Koncel-Kedziorski, Hannaneh Hajishirzi, Ashish Sabharwal, Oren Etzioni, and Siena Dumas Ang. 2015. Parsing algebraic word problems into equations. *Transactions of the Association for Computational Linguistics*, 3:585–597.

Rik Koncel-Kedziorski, Subhro Roy, Aida Amini, Nate Kushman, and Hannaneh Hajishirzi. 2016. MAWPS: A math word problem repository. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1152–1157, San Diego, California. Association for Computational Linguistics.

Ludmila I Kuncheva and Christopher J Whitaker. 2003. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine learning*, 51:181–207.

Gustave Le Bon. 1897. *The crowd: A study of the popular mind*. TF Unwin.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*.

Xiaonan Li, Kai Lv, Hang Yan, Tianyang Lin, Wei Zhu, Yuan Ni, Guotong Xie, Xiaoling Wang, and Xipeng Qiu. 2023a. Unified demonstration retriever for in-context learning. *ArXiv preprint*, abs/2305.04320.

Xiaonan Li and Xipeng Qiu. 2023a. Finding supporting examples for in-context learning. *ArXiv preprint*, abs/2302.13539.

Xiaonan Li and Xipeng Qiu. 2023b. Mot: Pre-thinking and recalling enable chatgpt to self-improve with memory-of-thoughts. *ArXiv preprint*, abs/2305.05181.

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2023b. Making language models better reasoners with step-aware verifier. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5315–5333, Toronto, Canada. Association for Computational Linguistics.

Tian Liang, 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. *ArXiv preprint*, abs/2305.19118.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 158–167. Association for Computational Linguistics.

Xiaoran Liu, Hang Yan, Shuo Zhang, Chenxin An, Xipeng Qiu, and Dahua Lin. 2023. Scaling laws of rope-based extrapolation.

引用
References

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. *ArXiv preprint*, abs/2301.13379.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *ArXiv preprint*, abs/2303.17651.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

S. Parsons and Peter McBurney. 2003. Argumentation-based communication between agents. In *Communication in Multiagent Systems*.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are NLP models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2080–2094, Online. Association for Computational Linguistics.

Pragaash Ponnusamy, Alireza Ghias, Yi Yi, Benjamin Yao, Chenlei Guo, and Ruhi Sarikaya. 2022. Feedback-based self-learning in large-scale conversational ai agents. *AI magazine*, 42(4):43–56.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2022. Scaling language models: Methods, analysis & insights from training gopher.

Nir Ratner, Yoav Levine, Yonatan Belinkov, Ori Ram, Inbal Magar, Omri Abend, Ehud Karpas, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. Parallel context windows for large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6383–6402, Toronto, Canada. Association for Computational Linguistics.

Subhro Roy and Dan Roth. 2015. Solving general arithmetic word problems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1743–1752. The Association for Computational Linguistics.

Noah Shinn, Beck Labash, and Ashwin Gopinath. 2023. Reflexion: an autonomous agent with dynamic memory and self-reflection. *ArXiv preprint*, abs/2303.11366.

Kaya Stechly, Matthew Marquez, and Subbarao Kambhampati. 2023. Gpt-4 doesn't know it's wrong: An analysis of iterative prompting for reasoning problems.

Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. 2022. A contrastive framework for neural text generation. In *Advances in Neural Information Processing Systems*.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. 2023. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, Toronto, Canada. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. 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 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *ArXiv preprint*, abs/2201.08239.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *ArXiv preprint*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Szymon Tworkowski, Konrad Staniszewski, Mikołaj Pacek, Yuhuai Wu, Henryk Michalewski, and Piotr Miłoś. 2023. Focused transformer: Contrastive training for context scaling.

Karthik Valmeekam, Matthew Marquez, and Subbarao Kambhampati. 2023. Can large language models really improve by self-critiquing their own plans?

Aimee Van Wynsberghe. 2021. Sustainable ai: Ai for sustainability and the sustainability of ai. *AI and Ethics*, 1(3):213–218.

Jianing Wang, Qiushi Sun, Nuo Chen, Xiang Li, and Ming Gao. 2023a. Boosting language models reasoning with chain-of-knowledge prompting. *ArXiv preprint*, abs/2306.06427.

Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. 2023b. Augmenting language models with long-term memory.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023c. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022b. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Sean Welleck, Ximing Lu, Peter West, Faeze Brahman, Tianxiao Shen, Daniel Khashabi, and Yejin Choi. 2023. Generating sequences by learning to self-correct. In *The Eleventh International Conference on Learning Representations*.

Carole-Jean Wu, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga, Jinshi Huang, Charles Bai, et al. 2022. Sustainable ai: Environmental implications, challenges and opportunities. *Proceedings of Machine Learning and Systems*, 4:795–813.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2023. Efficient streaming language models with attention sinks.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do large language models know what they don't know? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8653–8665, Toronto, Canada. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *ArXiv preprint*, abs/2205.01068.

Tianjun Zhang, Fangchen Liu, Justin Wong, Pieter Abbeel, and Joseph E. Gonzalez. 2023a. The wisdom of hindsight makes language models better instruction followers. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023b. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations*.

Chuanyang Zheng, Zhengying Liu, Enze Xie, Zhenguo Li, and Yu Li. 2023. Progressive-hint prompting improves reasoning in large language models. *ArXiv preprint*, abs/2304.09797.

A Limitations and Broader Impacts

Given the current constraints in communication and analytical capacities of open-source models (Fu et al., 2023a), as well as their substantial computational resource requirements (Touvron et al., 2023b; Chowdhery et al., 2022), we have not included these models in our experiment at this stage. However, we posit that open-source models with advanced comprehension and communication skills have the potential to match or even exceed the performance of commercial models (OpenAI, 2023; Ouyang et al., 2022; Chowdhery et al., 2022), through the collaborative exchange of insights.

A critical factor in model communication is the handling of long text. The current context windows of these models limit our ability to incorporate a broader range of models in the communication process. Recent works (Liu et al., 2023; Xiao et al., 2023; Wang et al., 2023b; Tworkowski et al., 2023; Chen et al., 2023; Ratner et al., 2023, *inter alia*) have begun to overcome this limitation by equipping models with the ability to process longer texts, laying the foundation for increasing the number of models involved in communication. In addition, our experiments indicate that model communication can achieve effective performance with reduced computational resources, aligning with the sustainable development goals of AI community (Van Wynsberghe, 2021; Wu et al., 2022).

Furthermore, the concept of AI learning from each other to foster collective improvement is a focal point of current research (Bai et al., 2022b; Ponnusamy et al., 2022; Lee et al., 2023). Our aim and aspiration is to cultivate a collective intelligence among large language models (Ha and Tang, 2022). This approach not only optimizes individual model performance but also contributes to the broader AI research community's pursuit of more advanced, collaborative AI systems.

B Datasets and Evaluation Metrics

Datasets In Table 2, we meticulously detail the specifics and statistics of each dataset employed in