Dynamic Inference Thought in Large Language Models

鈴木 拓真　川本 樹　三山 航
目黒 拓己　鈴木 中穂美　高木 友博
明治大学大学院
stakuma9912@gmail.com　tatsuki.00.0306@gmail.com　chinpanzi914@gmail.com
takumimeguro1321@gmail.com　nsuzuki1900@gmail.com　takagit@gmail.com

概要

本論文では、LLM が必要に応じて事前知識を参照しながら動的に推論を進める新しいフレームワークを提案する。私たちのアプローチでは、LLM による推論とその推論を改善するプロセス、そして現在の思考が十分かを LLM により判断して最終的な回答を推論する。高度な推論を複数の単純な推論に分割すること、多段推論を動的に行うことにより、推論精度の向上を達成する。ANLI データセットを用いて、このフレームワークの有効性を検証しており、先行研究と比較して、最大で 32%の精度向上を達成した。また、定性的評価も行い、本アプローチが従来手法よりも論理的に推論を行えていることを確認した。

1　はじめに

近年、ChatGPT などの大規模言語モデル (LLM) は、多くの自然言語処理や関連分野において素晴らしい性能を示しており、特に推論タスクにおいて、推論能力向上のための様々なプロンプトエンジニアリングに関する研究がなされている [1, 2, 3]。

依然として、LLM は人間には問題のない一般的な計画/推論タスクで受け入れ可能なパフォーマンスを持っておらず、それは現在最も優れた LLM である GPT-4 においても同様である [4, 5]。

これは、推論を多段で行う機構を持っておらず、このような推論プロセスを一回の呼び出しのみで実行することが求められる通常のアプローチでは、不十分なパフォーマンスを引き起こす。そして、その課題解決のために、様々な推論精度向上のための研究が行われている。

例えば、推論過程を例示する Chain of Thought(CoT)[6] や、外部メモリを活用して、LLM が事前に推論した大量の問題と推論結果を利用する Memory of Thought(MoT)[7] などがある。しかし、これらのアプローチによる推論のサポートは暗黙的であり、LLM に求められる処理の難しさは単一呼び出しのアプローチと同じため、問題は本質的に解決されていない。

そこで、私たちはこの課題に対処するために、新しいフレームワークを導入し、LLM の推論過程を明示的かつ多段にし、各推論を単純化することによる推論精度の向上を達成する。私たちのフレームワークの考え方は主に 2 つの既存研究にインスパイアされている。複雑な 1 つのタスクを複数の単純なタスクに分割して解いていくことによって精度を向上する Least-to-Most Prompting(L2M)[8] と与えられたコンテキストから文を選択する選択モジュールと、選択された文から 1-hop 推論を行う推論モジュールを使用し、二つのモジュールを交互に繰り返すことで multi-hop 推論を行う Selection-Inference(SI)[9] である。

SI では、選択と推論の反復回数をハイパーパラメータとして事前に設定する必要があるが、適切な解を出すのに必要な推論ステップ数は事前に決まることはなく、ステップ数の過不足は推論精度低下の原因となる。そのためステップ数は推論過程で動的に決定する必要があり、これは SI の本質的な欠点であるといえる。

私たちは、SI における反復のような思考の改善をいつ止めるべきかについてを、LLM に判断させ、思考の改善に関しても、必要に応じて LLM が持つ事前知識を利用し、補完することにより上記の課題に対処した。

これにより、例えば、「X はイギリスで生まれた」という仮説に対して、コンテキスト中に「X はロンドンで生まれた」というような記述しかない場合においても、「X はロンドンで生まれた」という事実と、「ロンドンはイギリスにある」という事実を組

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

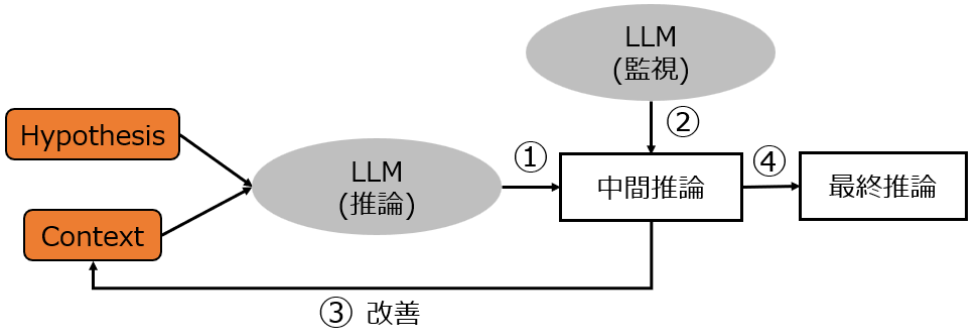

図 1　提案フレームワークの概要

み合わせて「X はイギリスで生まれた」と推論することができる。

2　関連研究

Chain of Thoughts（CoT）[6] は、モデルが問題を解決する際に、中間ステップや推論の過程を明示的に生成し、最終的な答えに至るまでの思考の流れをトレース可能にするアプローチである。これにより、精度向上とモデルの判断基準や推論プロセスの透明化を達成している。一方で、100 億パラメータ以上のモデルサイズやメモリ機能が要求されており、本質的な解決には至っていない。

Selection Inference（SI）[9] は、LLM を複数回呼び出すシンボリックアプローチである。しかし、反復回数は固定されており、事前に設定する必要があるため、タスクには必要な推論ホップの固定回数を持つ必要がある。これに対して、Faithful Reasoning[10] は SI がステップ数において反復をいつ停止するかを判断できない問題を解決し、推論が十分であるかどうかを判断する LLM「Halter」を用意した。Halter は自身で停止反復を決定することができるが、各タスクごとに追加で Language Model をファインチューニングする必要があり、Halter では様々なタスクに対して動的に対応することができない。

Tree of Thoughts（ToT）[11] は、タスク解決までの中間プロセスをツリー構造で表現し、同時並行で複数の思考を追求するアプローチである。CoT のような線形推論経路では、各サブ問題には複数の適切なオプションがある場合に思考を制限してしまう問題に対処しており、各サブ問題に対して可能なオプションを考慮して、外部のツリーサーチメカニズム（例：BFS、DFS）によって決定木を探索できる。また、LLM の評価能力を効率性向上のためのノード刈り込みに使用している。

Graph of Thoughts（GoT）[12] は、LLM によって生成された情報を任意のグラフとしてモデル化するフレームワークである。このフレームワークにより、思考の改善、思考の分解、思考の集約により柔軟に組み合わせて答えに至ることが可能になり、より人間の思考プロセスに近づけている。GoT では、思考の単位をノードとしてグラフ内に表し、これらの思考間の依存関係や接続をエッジとして描く。この構造は、人間の思考プロセスを線形のチェーンだけでなく、人間が考えるより複雑で非線形な方法を反映するネットワークとしてもモデル化が可能である。

一方で、ToT、GoT では、どのタイミングで推論を止めるかについては動的に判断することを行なっていない。推論の手順を事前に生成することで対応しているが、プランの生成には GPT-4 のような高度な推論スキルが必要であるため、CoT と同様に本質的な解決には至っていない。

3　モデル

このセクションでは、私たちが提案する、論理的で正確な推論のためのフレームワークについて記述する。提案手法の主な考え方に関しては図 1 に示しており、問題から結論までのプロセスは以下の 4 ステップに分けられる。

1. 中間推論ステップ
   与えれた入力に対する推論（ここでは中間推論とよぶ）を行う。
2. 改善の必要性判断ステップ
   問題に回答するために十分な結論が得られているかどうかを LLM が判断する。
3. 改善ステップ
   (2) で改善が必要だと判断された場合に、問題に付属するコンテキストもしくは、事前知識を用いて中間結論を改善する。

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

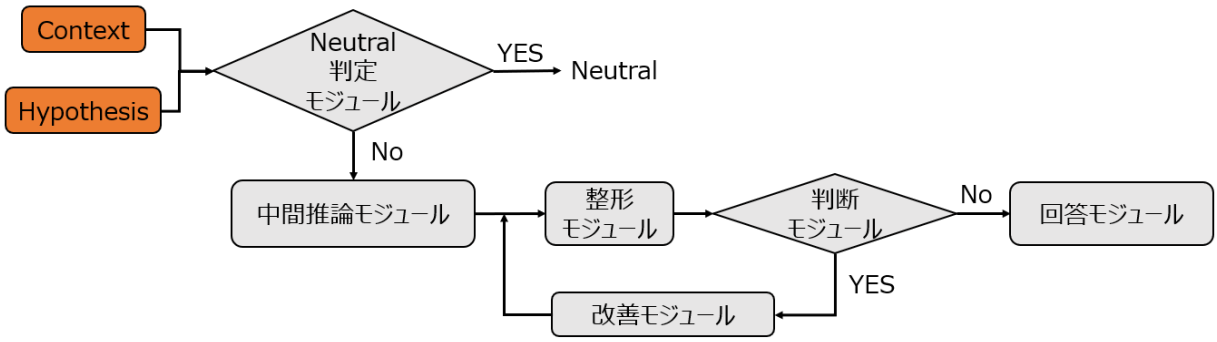

図 2　実験上における実装モデル

4. 最終推論ステップ
   (2) で改善が必要なしと判断された場合に、それまでの中間結論を最終的な結論とする。

4　Experiment

4.1　実験設定

提案するフレームワークの有効性を確認するために、ANLI データセット [13] を用いて実験を行った。各データでは、コンテキスト c と仮説 h、そして答え a が与えられており、コンテキスト c には仮説 h に関する記述がある場合 (Entailment)、関係する記述がない場合 (Neutral)、対立する記述がある場合 (Contradiction) の 3 つのパターンが存在し、モデルは各データがどのパターンに該当するかをに関して推論を行う。コンテキスト中には、回答をサポートする文章だけでなく回答には不必要な情報も含まれる。また、コンテキスト c と仮説 h の整合性判断のためには、コンテキスト中に記述された情報だけでは不十分で、一般的知識を用いて補完する必要がある場合も存在する。

私たちは、このデータセットから A1 の 1000 個のデータを使用し、評価実験を行った。モデルは、LLaMA2-13b-chat[14] を利用している。各モジュールに関しては、3 ショットプロンプトを使用しており、例題に関しては訓練データからサンプルしている。

4.2　実験における提案モデル

NLI タスクを用いた実験時には、精度向上のため細かい処理をいくつか追加しており、詳細を図 2 に示す。各モジュールに関しては以下の通りである。

1. Neutral 判定モジュール
   仮説に関する文が文脈中に存在するかを推論。Neutral の場合、多段推論の必要がないため、最初の段階で判断する。
2. 中間推論モジュール
   文脈を参照し、仮説に対して推論を行う。
3. 判断モジュール
   現在の中間推論が、仮説とコンテキストとの整合性を推論するために十分な情報を持っているかを判断する。
4. 改善モジュール
   中間推論が不十分であると判断された場合に、コンテキスト中もしくは、LLM が持つ事前知識を利用して中間推論の改善に必要な情報を生成する。
5. 整形モジュール
   中間推論に、改善モジュールで生成した情報を組み込み、新たな中間推論として整形する。
6. 回答モジュール
   中間推論を用いて、最終的な仮説とコンテキストの内容との整合性判断の推論を行う。

4.3　比較手法

- IO：モデルには仮説とコンテキストが入力され、モデルが回答のみを生成する。
- CoT：仮説とコンテキストがモデルに入力され、モデルが推論過程と回答を生成する。
- SI：モデルはコンテキスト中から 2 つの文を選択し、選択した 2 つの文から 1 つの推論を行う。この選択と推論を 1 ステップとして扱われる。SI ではステップ数はハイパーパラメータであり、今回は 3 ステップとして設定している。

比較するすべての手法において、3-shot で推論を行っている。

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

#Problem to solve
Hypothesis:
The BSF is funded by the country directly north of Mexico.

Context:
The Building Strong Families Program (BSF) is part of the Healthy Marriage Initiative funded by the U.S. Department of Health and Human Services, Administration for Children and Families, "to learn whether well-designed interventions can help couples fulfill their aspirations for a healthy relationship, marriage, and a strong family."

Correct Answer: Entailment

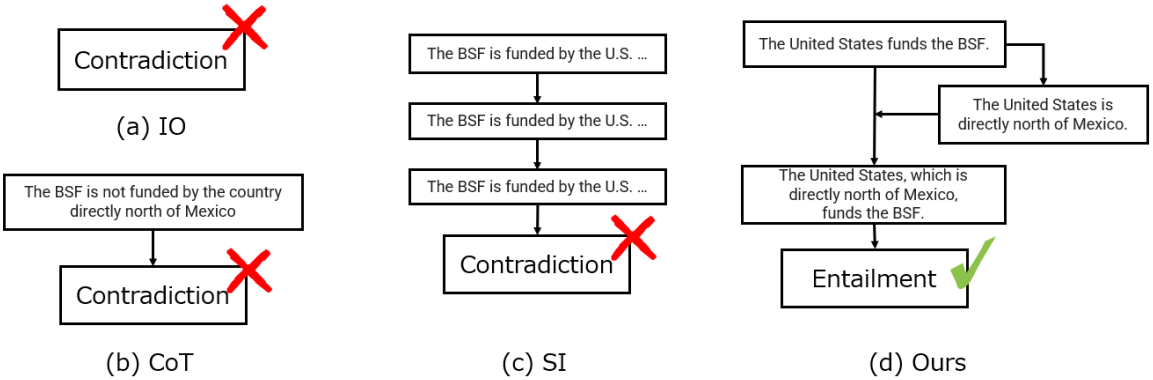

図 3　各アプローチによる出力の違い

4.4　定性評価

各方法の出力を検証することによって、私たちのフレームワークが動的に推論を進めることができるかを確認する。例題と、それに対する各手法の出力の違いは図 3 に示す。IO、CoT では、誤った回答をしており、その原因も不明、または曖昧な状態である。SI では、回答は誤ったままであるが、推論過程は明確である。最初に BFS に資金提供しているのはアメリカであることをコンテキストから推論している。しかし、コンテキストに含まれない「アメリカがメキシコの上側にある国である」という一般常識的な知識を利用できず、推論が発展せず、結果として最後の推論時に誤った回答をしてしまっている。一方で、提案手法では、思考の改善段階で、仮説の表現に従って「アメリカがメキシコの上側にある国」という必要な情報を付加することにより推論を改善を行えることが確認された。

4.5　定量評価

生成された回答が真実の回答と一致するかにより精度を測定した。定量的評価の結果を表 1 に示す。私たちが提案したフレームワークが最良の精度を達成し、本フレームワークの有効性を示している。CoT の精度は IO よりも精度が上がっており、これは LLaMA2-13b-chat の推論精度が高いことを示しており、簡単なタスクであれば十分な推論能力を持っていることがわかる。

表 1　実験結果

Method	ANLI(A1)
IO	0.39
CoT	0.416
SI	0.361
Ours	0.478

5　結論と展望

本論文では、LLM が必要に応じてコンテキスト中もしくは、事前知識を参照しながら動的に推論を進める新たなフレームワークを提案した。このフレームワークでは、推論プロセスを複数のモジュールに分割することによる各推論の簡易化や、動的な思考の改善により LLM がより精度の高い推論を行うことをサポートしている。ANLI データセットを用いた精度による定量評価や、実際の推論過程を確認する定性評価を通して本フレームワークの有効性を示した。

一方で、提案するフレームワークでは、現状単一の推論過程のために、思考が大幅に制限されている。関連研究にある ToT[12]、GoT[13] のように複数の思考を展開することは、推論精度の観点から重要な要素である。そのため、将来的に本フレームワークの複数思考への拡張を考えている。

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

参考文献

[1] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity, 2023.

[2] Vedant Gaur and Nikunj Saunshi. Reasoning in large language models through symbolic math word problems, 2023.

[3] Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey, 2023.

[4] Karthik Valmeekam, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. Large language models still can't plan (a benchmark for llms on planning and reasoning about change), 2023.

[5] Konstantine Arkoudas. Gpt-4 can't reason, 2023.

[6] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

[7] Xiaonan Li and Xipeng Qiu. Mot: Memory-of-thought enables chatgpt to self-improve, 2023.

[8] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. Least-to-most prompting enables complex reasoning in large language models, 2023.

[9] Antonia Creswell, Murray Shanahan, and Irina Higgins. Selection-inference: Exploiting large language models for interpretable logical reasoning, 2022.

[10] Antonia Creswell and Murray Shanahan. Faithful reasoning using large language models, 2022.

[11] Jieyi Long. Large language model guided tree-of-thought, 2023.

[12] Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Michal Podstawski, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. Graph of thoughts: Solving elaborate problems with large language models, 2023.

[13] Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial nli: A new benchmark for natural language understanding, 2020.

[14] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).