概要

推論とは未知の事柄について自らの知る情報から予想・論じることであり，人間の知的機能の基盤をなす能力である．人工知能における言語に関する研究分野である自然言語処理において，言語処理モデルの推論能力は機械学習の文脈における汎化性能に直結する重要な要素である．Transformer の登場以来，言語処理モデルの言語能力は飛躍的に向上したものの，未だ推論能力に関しては多くの課題が存在している．Shortcut reasoning は，言語処理モデルの非合理的な推論のことである．具体的には，学習データと同じ分布を持つ入力に対しては有効であるが，学習データと異なる分布を持つ入力に対して有効ではない推論のことである．この非合理的な推論により，言語処理モデルの頑健性が低下することが先行研究で明らかになっている．

昨今大きな話題となっている大規模言語モデル (LLM) は，既存の言語処理モデルを凌駕する言語能力をみせている．LLM は推論においても性能の向上が観察されている．入力に際して特定のプロンプトを与えると性能が大きく向上したという結果や，外部のアルゴリズムと組み合わせて LLM に推論させる手法なども提案され，LLM に推論を行わせることに関心が寄せられている．

しかしながら，Hallucination と呼ばれる LLM の生成する反事実的な知識が大きな課題として知られている．仮に推論プロセスを正しく構築することが可能であっても，その過程で用いられる知識に誤りがあれば最終的に出力される解答に間違いが含まれる可能性がある．先行研究では，推論に際して発生した hallucination が下流の推論プロセスに次々に影響を与え，雪だるま式にエラーが増大していくことが明らかになっている．こうした問題を受け，様々な解決策が研究されている．その一つが RAG と呼ばれる手法である．RAG は入力に対して知識ベースから取得し (Retrieval)，取得した文書を参照しながら解答を生成するパラダイムのことである．これにより，hallucination が低減され，性能が向上したという実験結果が多数の研究により示されている．しかし，RAG の retrieval 機構にもいくつかの改善の余地が存在する．その代表例が，取得した k 個の文書群の中に不必要な文書が含まれている，もしくは必要な文書が含まれていないことがある点である．不必要な文書がノイズとなり，Hallucination を増長させてしまう可能性がある．また，推論のたびに retrieval が必要である点など，解決すべき課題は未だ残っている．

LLM の推論能力自体についても，依然として問題点が見受けられる．先行研究によると，LLM はその構造上，単一ステップの推論には成功するものの，部分的な推論結果を合成しながら解く必要があるような推論問題，例えば 3 桁 $\times$3 桁のような計算問題を解くことが不得意であることが示されており，我々が観察できる LLM の優れた推論は大規模な事前学習データ内に含まれる推論プロセスを使ったパターンマッチングにより行われていることが明らかになっている．

以上のような，言語処理モデル，あるいは LLM の推論に関する諸問題を受け，我々は自己認識能力を持つ，解釈可能な推論システムの実現を目指す．自己認識

能力とは，何を知っていて，何を知らないのかを説明できることである．また，解釈可能とは，どのように質問を分解して，どの知識を参照したのかを説明することが可能なことである．このシステムは，入力された質問の分解と，知識ベースへの問い合わせ，そして得られた結果の合成によって，最終的な予測を出力する．知識ベースモジュール，推論モジュール，中央制御モジュールの 3 つのモジュールから構成され，それぞれが質問に関する知識の取得，質問の分解，そして，解答の統合・出力を行う．

この実現にあたり，言語処理モデルの推論能力に関する分析・考察を経た後，自己認識可能な知識ベースモジュールの実現に向けた研究を行った．具体的には，以下の 3 つのテーマに取り組んだ．

(i) 一般性を考慮した Shortcut reasoning の自動検出では，先行研究で提案されていた手法の課題を解決し，shortcut reasoning を自動的に検出する手法を提案した．先行研究で提案されている検出手法では，shortcut reasoning の形態を事前に定義している，内部情報を用いていない，人手による評価を必要としている等の課題が存在していた．最新の研究の提案手法によってそれらの課題は克服されたものの，依然としていくつかの制約を抱えている．感情分析タスクと自然言語推論タスクの分類問題で学習された言語処理モデルを対象とした実験を行い，人間の介入無しで shortcut reasoning を検出することに加え，先行研究で明らかになっていなかった未知の shortcut reasoning を発見することに成功した．

(ii) 論理的根拠に基づく機械読解システムに向けた研究では，機械読解タスクでの Explainer における shortcut reasoning について実験を行った．Explain-then-predict は予測とその論理的根拠を出力させるパラダイムである Rationalization において頻繁に使用されるアーキテクチャであり，explainer はその一部を構成するモジュールである．Explainer は適切な推論のために入力中の必要な情報のみを抽出する，すなわち推論に不必要なノイズを低減してくれる効果が期待できることから，ある先行研究はこのアーキテクチャは頑健性を向上できるという仮説を立て，検証したものの，頑健性の向上は限定的であった．この結果を踏まえ，我々は explainer が shortcut reasoning を行っているため，頑健性の向上が実現できなかったという仮説を立てた．経験的な実験の結果，explainer の入力に破壊的な編集を加えても精度が落ちなかったことから，モデル内の shortcut reaosning が示唆された．

(iii) 自己認識が可能な知識ベースとしての大規模言語モデルの実現に向けた議論では，LLM の知識について，予測の自信度，その正解率，その知識の事前学習データにおける頻度の 3 つの要素の関係性について分析を行った．先行研究では，既存の手法で計算された LLM の予測の自信度は，正解率を良く反映していることが明らかになっている．また，自信度と正解率，正解率と事前学習における頻度の相関関係はいくつかの研究により明らかになっているが，自信度と頻度の関係性については我々の知る限り未知である．そこで，我々は LLM の知識の自信度が，正解率ではなく，知識が事前学習に現れる頻度を反映しているという仮説を立て，

予備的な検証実験を行った．実験に用いた PopQA データセットは，Wikipedia から収集した知識と，その知識の Wikipedia ページにおける月間閲覧数を頻度として近似した人気度と呼ばれる指標が付与されている．実験の結果，GPT-3.5-turbo を用いたとき，頻度が大きい知識に関して自信度と正解率が正相関するが，頻度の低い知識に対しては自信度が上がっても正解率が上がらなかったという結果を得た．この結果は興味深いものであるものの，仮説の実証には至ることができなかった．