

# 早押しクイズの名数問題における解の妥当性を考慮した解答をするための CoT プロンプトの構築

杉山宏輝<sup>1</sup> 角康之<sup>1</sup>

<sup>1</sup> 公立はこだて未来大学

h-sugiyama@sumilab.org

sumi@fun.ac.jp

## 概要

本研究では、早押しクイズの名数問題における解の妥当性を考慮した解答を言語モデルに出力させる方法を提案する。通常の早押しクイズでは解の妥当性を考慮することで、一見答えが確定していない段階での解答の正答率を高めることができ、言語モデルにおいても同様のことが期待できる。具体的な手法として、Chain-of-Thought(CoT) プロンプト [1] を用いて、妥当性を考慮した思考過程とそこから得られる解答を出力する。妥当性には、仲間はずれ、文字列の類似、難易度、順序の4種類あり [2]、それぞれ問題と思考過程を含めた解答をプロンプトとして与える。

## 1 はじめに

通常の質問応答タスクでは問題文が全文与えられるため純粋な知識量が重要となるが、早押しクイズは他者よりも早くボタンを押して解答権を得る必要があるため、問題文の途中、より早い段階での解答が求められ、知識量に加えて問題文の推測能力も必要となる。早押しクイズを解くシステムとしては Watson[3] を始め、研究されてきているが、多くは解答のために十分な文量を要し、人間のクイズプレイヤーほどの早さで解答することは難しい。早押しクイズには構文や答えとしての自然さがあり [4]、それらをもとに問題文を推測することで、一見答えが確定していないような段階での解答が可能となっている。例えば、『歳時記』における季語の5つの分類とは、「春」「夏」「秋」「冬」と何でしょう? という問題がある。通常、『歳時記』における季語の5つの分類とは、の段階では「春」「夏」「秋」「冬」「新年」の5択であり、1つに絞り込むことはできない。しかし、多くのクイズプレイヤーは「新年」と答え

ることができると考えられる。なぜなら「春」「夏」「秋」「冬」は四季として慣れ親しんだ季節であるのに対し、「新年」は多くの場合季節として認識していない。歳時記ではそれが一季節として分類されているという意外さから、「新年」を答えとするのが最も自然だからである。こうした答えとしての妥当性を、言語モデルが推測できるようにすることで、早押しでの質問応答タスクでも正答することができるようになる。本研究では、CoT を用いて、名数問題を対象に、問題文の前半で解答する方法を提案する。

## 2 関連研究

今回用いる Chain-of-Thought という手法は、ある問題に対する解答を、思考過程をいくつかの段階に分けて例示することで、最終的な精度の向上を図るものである [1]。算術推論、常識推論、記号的推論において精度の向上が見られた。日本語においても有効性が示されている。[5]

海外の早押しクイズにおける研究 [6] では、複数のモデルを組み合わせ、ランダムな位置で切り捨てた問題文で学習させた結果、NIPS 2017 で行われた質問応答のコンペティションで優勝し、人間のクイズ王6名に対して勝利を収めている。しかし、複数の文からなる形式のクイズ「Quiz bowl」の問題を用いた検証において、全文が与えられた場合の正解率は96%だったが、単一の文では56%にとどまった。これは最適な押しポイントは学習できておらず、推定精度が問題文の情報量に依存しているためと考えられる。本研究では、名数問題において問題文の前半に得られる解候補から、後半の情報を聞くことなく正解することを目指す。また、日本語の早押しクイズを題材とした研究 [7, 8] では、パラレル問題を調査し、その構造と対比関係を明らかにしている。パラレル問題

とは「アイルランドの首都はダブリンですが、アイスランドの首都はどこでしょう？」のような早押しクイズ特有のひっかけ問題のことである。対比関係を対義、連続、一対一対応、音・表記の一致と定義づけ、それらを元に問題や解の自動生成を行った。このことから、早押しクイズ特有の要素は解の推定に関係があり、本提案においても有効であると考えられる。

### 3 名数問題と妥当性

ここでいう名数問題とは、以下のような形式で、あるグループ  $W$  の中から答えとなる要素以外を列挙し、残ったものを聞く問題のことを指す。

問題： $W$  とは、 $w_1, w_2, \dots, w_{n-1}$  と何でしょう？ 答え： $w_n$

名数問題の妥当性は、以下の4つに大別される。[2]

**仲間外れ** 解候補のうち、他に共通している上位概念がないもの。

例題：世界四大珍獣に数えられる4種類の動物とは、ジャイアントパンダ、コビトカバ、オカピと何でしょう？ 答え：ボンゴ

解説：「ジャイアントパンダ」「オカピ」「コビトカバ」の3種がまとめて世界三大珍獣と呼ばれるため、ボンゴが仲間外れとなる。

**文字列の類似度** 解候補のうち、文字の表記や音が1つだけ似ていないもの。

例題：オリオン座の中央にある3つ星を構成しているのは、アルニタク、アルニラムと何でしょう？ 答え：ミンタカ

解説：「アルニラム」「アルニタク」という似た音の単語と比べ、「ミンタカ」は明らかに音が異なっていることがわかる。

**難易度** 解候補のうち、知名度が低いなどして1つだけ単語の難易度が高いもの

例題：一般に「世界三大料理」と呼ばれる料理とは、フランス料理、中華料理と何料理でしょう？ 答え：トルコ料理

解説：「フランス料理」「中華料理」に比べ「トルコ料理」はあまり一般的ではなく、単語としての難易度が高いといえる。

**順序** 日付など何かしらの番号が付与されている単語の順番

例題：京都三大祭といえば、5月の葵祭、7月の祇園祭と10月にある何祭でしょう？ 答え：時代祭

解説：葵祭は5月、祇園祭は7月、時代祭は10月に開催されるため、葵祭<祇園祭<時代祭という順序

関係が存在する。

## 4 プロンプトの構築と評価

名数問題に解答するためのプロンプトの構築と評価用の問題の選定、および各プロンプトを用いて大規模言語モデルで解いた場合の評価をする。例題として利用した問題と解答を図1に示す。

### 4.1 プロンプト

CoT プロンプトの shot として用いる8問の問題と思考過程を含む解答を作成した。問題は4種類の妥当性が全て含まれるように採用し、計8問の例示とした。プロンプトの構成は実際にクイズを行っているときの思考過程に沿って1) 解候補の列挙2) 理由付け3) 最終的な答えとした。

### 4.2 評価データ

クイズを題材にした質問応答のコンペティションである AI 王<sup>1)</sup>で配布されている問題(22,335問)から、351問の名数問題を採用した。

### 4.3 評価設定

構築したプロンプトを用いて大規模言語モデルに評価データを解かせ、正解率を調べる。モデルは日本語のモデルである Japanese Stable LM<sup>2)</sup>を使用し標準プロンプトと CoT プロンプトの2種類で精度を比較する。標準プロンプト 8-shot, CoT プロンプト 8-shot で行い、標準プロンプトは問題の直後に”答えは「”を置き、CoT プロンプトは”解答：”を置く。問題文は選択肢が確定する前振り(問題文前半)までを入力とした。正誤判定は、CoT プロンプトにおいては出力の表記ゆれが激しいため(答えは「～」、正解は「～」、最後は「～」など)、最終的に言っている単語を解答とみなし、人手で判断した。

### 4.4 結果と考察

モデル	正解率
Japanese Stable LM	0.115
Japanese Stable LM + CoT	0.104

結果を表1に示す。標準プロンプトを用いた場合と CoT プロンプトを用いた場合で正解率の差はほとんど見られなかった。CoT プロンプトによって正解率

1) <https://sites.google.com/view/project-ai0/dataset?authuser=0>

2) <https://huggingface.co/stabilityai/japanese-stablelm-base-alpha-7b>

問題：日本国憲法に定められている国民の三大義務とは？  
解答：国民の三大義務は、教育、勤労、納税の3つである。教育は第26条、勤労は第27条、納税は第30条で規定されており、数字順に並べたほうが自然であるため、最後に聞かれるのは納税である。よって答えは「納税」

問題：オリオン座の中央にある3つ星を構成しているのは？  
解答：オリオン座の3つ星は、ミンタカ、アルニラム、アルニタクの3つである。ミンタカはアルニラム、アルニタクに比べて音・表記上の類似度が低いため、最後に聞かれるのはミンタカである。よって答えは「ミンタカ」

問題：一般に世界四大珍獣に数えられるのは？  
解答：世界四大珍獣は、ジャイアントパンダ、オカビ、コビトカバ、ボンゴの4つである。ジャイアントパンダ、オカビ、コビトカバは世界三大珍獣に数えられ、ボンゴは仲間外れであるため、最後に聞かれるのはボンゴである。よって答えは「ボンゴ」

問題：一般に世界三大料理といえば？  
解答：世界三大料理は、中華料理、フランス料理、トルコ料理の3つである。一番知名度が低いのはトルコ料理のため、最後に聞かれるのはトルコ料理である。よって答えは「トルコ料理」

問題：茶道の三千家といえば？  
解答：茶道の三千家は表千家、裏千家、武者小路千家の3つである。一番知名度が低いのは武者小路千家のため、最後に聞かれるのは武者小路千家である。よって答えは「武者小路千家」

問題：京都三大祭といえば？  
解答：京都三大祭は葵祭、祇園祭、時代祭の3つである。葵祭は5月、祇園祭は7月、時代祭は10月に開催されており、時系列順に並べたほうが自然であるため、最後に聞かれるのは時代祭である。よって答えは「時代祭」

問題：ABC式消火器が対応する3種類の火災とは？  
解答：ABC消火器が対応する3種類の火災は普通火災、油火災、電気火災の3つである。Aは普通火災、Bは油火災、Cは電気火災を表しており、出現順に並べたほうが自然であるため、最後に聞かれるのは電気火災である。よって答えは「電気火災」

図1 CoT プロンプト

が向上しなかった原因はさまざま考えられるが、1つは解候補列挙段階でのハルシネーションの影響が大きいと考えられる。問題2(表2)のように、列挙した中に正答が含まれていなければ、最終的に正解することはほとんど不可能である。また、理由付けとして妥当な推論ができていないことがほとんどであった。問題3,4のように、列挙した解候補は正しいにもかかわらず、推論が支離滅裂になってしまい、答えを間違ってしまう場合も多い。これは8-shotの例示では思考過程を十分に学習できなかったことが原因であると考えられる。関連研究[5]では1つのタスクにつき8つの例示であったが、本研究では4種類の妥当性全て含めて8-shotであり、各妥当性につき1-3つの例示しかできていなかったため、理由付けを推論するための十分な情報が無かったと考えられる。中には問題1のように、カースト階級の順序を正確に把握し、一番最後のシュードラを答えとするなど、理想的な推論結果も得られており、これらはCoTによって妥当性を考慮した推論を学習できた結果で

あると考えられる。

表2 問題とモデルの出力例

問題1：古代インドのカースト制度における主要な4階級とは、バラモン、クシャトリア、ヴァイシャと何でしょう？ 解答：古代インドにおけるカースト制度は、バラモン、クシャトリア、ヴァイシャ、シュードラの4つで構成されている。バラモンはカーストの最上位に位置し、クシャトリアはバラモンに次ぐ地位、ヴァイシャはクシャトリアに次ぐ地位であり、シュードラは最下位に位置している。よって、最後に問われるのはシュードラである。(正解)
問題2：夏目漱石の前期三部作といえば、『三四郎』『それから』と何？ 解答：夏目漱石の前期三部作は「吾輩は猫である」、「坊っちゃん」、「草枕」のつである。一番知名度が低いのは「草枕」のため、最後に聞かれるのは「草枕」である。よって答えは「草枕」(不正解)
問題3：「六法」と呼ばれる6つの法律といえば、憲法、民法、民事訴訟法、刑法、刑事訴訟法と何でしょう？ 解答：六法とは、憲法、民法、刑法、商法、民事訴訟法、刑事訴訟法の6つである。「六法全書」という書籍は存在しないため、六法といえば六法である(不正解)
問題4：ロシアの作曲家、チャイコフスキーの三大バレエ音楽といえば、『白鳥の湖』『くるみ割り人形』と何でしょう？ 解答：チャイコフスキーの三大バレエは、白鳥の湖、眠れる森の美女、くるみ割り人形の3曲である。白鳥の湖の次に聞かれるのがくるみ割り人形であるため、最初に聞かれる曲は白鳥の曲である。「白鳥の歌」は白鳥の曲の別名である。(不正解)
問題5：太陽系の惑星の中で衛星をもっていない2つの惑星とは、水星と何でしょう？ 解答：水星と金星である。水星は太陽に近く、金星は地球に近く、どちらも衛星をもたない。よって最後に聞かれるのは金星であり、答えは「金星」(正解)

## 5 おわりに

本研究では、早押しクイズの名数問題における解の妥当性を考慮し、問題文の途中で解答するためのCoTプロンプトの構築と検証を行った。8-shotの標準プロンプトとCoTプロンプトで正解率の差は見られなかったが、理想的な推論結果が得られた出力もあり、今後に期待できる結果となった。妥当性の無い推論なども数多くあったが、CoTはモデルサイズが大きいほど効果が高いということが知られているため、今回検証した7bよりも大きいサイズのモデルや、例次数を増やした検証もしていきたい。また、ハルシネーションへの対策として、Retrieverを機構を組み込み、外部知識を参照できるようにする必要がある。

## 参考文献

- [1] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. **Advances in Neural Information Processing Systems**, Vol. 35, pp. 24824–24837, 2022.
- [2] 杉山宏輝, 角康之. 早押しクイズの解候補の絞り込み. インタラクション 2023 論文集, pp. 1014–1018, 2023.
- [3] JDavid Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A. Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John Prager, Nico Schlaefter, and Chris Welty. Building watson: An overview of the deepqa project. **AI Magazine**, Vol. 31, No. 3, pp. 59–79, 2010.
- [4] 伊沢拓司. クイズ思考の解体. 朝日新聞出版, 2021.
- [5] 堀尾海斗, 村田栄樹, 王昊, 井手竜也, 河原大輔, 山崎天, 新里顕大, 中町礼文, 李聖哲, 佐藤敏紀. 日本語における chain-of-thought プロンプトの検証. 人工知能学会全国大会論文集, 2023.
- [6] Ikuya Yamada, Ryuji Tamaki, Hiroyuki Shindo, and Yashiyasu Takefuji. Studio ousia’s quiz bowl question answering system. In **The NIPS ’17 Competition: Building Intelligent Systems**, pp. 181–194, 2018.
- [7] 橋元佐知, 佐藤理史, 宮田玲, 小川浩平. 競技クイズ・パラレル問題の基本構造と文型. 語処理学会第 27 回年次大会発表論文集, pp. 1420—1424, 2021.
- [8] 橋元佐知, 佐藤理史, 宮田玲, 小川浩平. 早押しクイズのパラレル問題の自動生成. 語処理学会第 28 回年次大会発表論文集, pp. 832—837, 2021.