# AmbiNLG: 自然言語生成のための指示テキストの曖昧性解消

丹羽 彩奈　磯 颯
Megagon Labs
{ayana,hayate}@megagon.ai

## 概要

大規模言語モデル（Large Language Models; LLMs）の台頭により、自然言語を用いた指示で多岐にわたる言語処理タスクが実行可能になった。しかし、与えられた指示が曖昧性であるためにユーザーの意図と異なるテキストが生成されることがある。特に自然言語生成（Natural Language Generation; NLG）に見られる曖昧性は広範で、人間が曖昧でない指示を書くことは難しい。そこで本研究では、曖昧性解消ベンチマークデータセット AmbiNLG の導入および自動的かつ網羅的な曖昧性解消を行う。実験より、提案手法による複数の曖昧性の明示的・網羅的な曖昧性解消の有効性を示した。

## 1　はじめに

LLMs [1, 2], 特に ChatGPT [1)] に代表されるような対話型 LLMs は、自然言語による指示を介して操作できるため、非専門家も扱える。しかし指示テキスト次第で LLMs の出力は大きく変動するため [3]、所望のテキストを生成させるには適切な指示を探索するプロンプトエンジニアリングが必須である [4]。特に自然言語は曖昧性という本質的な問題を抱えており [5, 6, 7]、ユーザーの指示が不明瞭である場合出力結果を一意に定めることができず、結果に悪影響をもたらしえる。一方で、明瞭な指示文作成のためのプロンプトエンジニアリングは容易ではないため、自動で曖昧性を解消する手法が求められる。

曖昧性を解消する先行研究では、機械翻訳における多義語 [6] や質問応答における情報不足 [8] といった特定のタスク・曖昧性のみを対象としている。しかし、先行研究のような限定的な設定における曖昧性解消では十分ではないと考えられる。まず、LLM は実社会での多様な NLG タスクで利用されているため、幅広い NLG タスクに対して解消すべき曖昧性カテゴリを洗い出し、それらを網羅的に解消する必要がある。加えて、NLG タスクの指示には曖昧になりえる要素が散在するという点も問題である。例えば「Summarize this article.」という指示だけでは、要約の長さ、焦点を当てるべき記事の内容、段落構成などの情報が与えられておらず、ユーザーが真にどのような出力を求めているかはわからない。

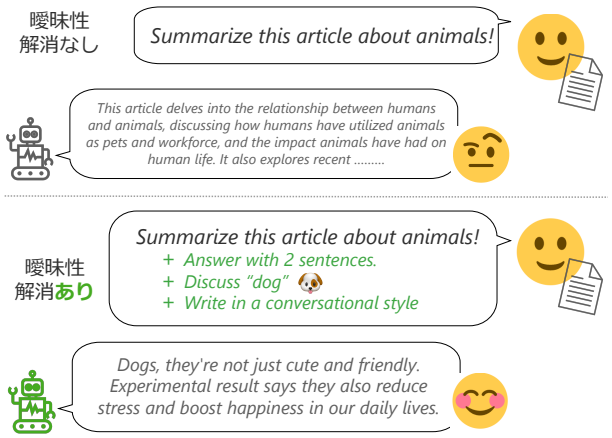



**図 1**　本研究が提案する曖昧性解消方針。指示テキストに曖昧性を解消するための補完テキストを追加する。

これらの問題に対し、本研究では (1) 様々な NLG タスクに対する指示テキストの曖昧性問題の体系化、(2) 様々な曖昧性の網羅的解消 を行う。まず既存の Super-natural instruction [9] データセットにおける曖昧性を人手で 6 つのカテゴリに体系化する。次に、データセット内の指示テキストに曖昧性カテゴリとその曖昧性を補完するための情報（補完テキスト）を自動付与することで曖昧性解消データセット AmbiNLG を構築する。そして、元の指示テキストに補完テキストを連結することで曖昧性が解消された明瞭な指示テキストを作成する（図 1）。

実験の結果、曖昧性を解消した指示によってタスク性能が最大で 9.12 ポイント向上した。さらに、特定の曖昧性のみ解消した指示と網羅的に曖昧性を解消した指示を与えた場合を比較すると、後者の方がタスク性能がさらに 8.69%以上、さらに曖昧性を解消しない元の指示と比較して 17.21%改善した。こ

1)　https://openai.com/chatgpt

れらのことから、NLG タスクの指示の曖昧性はタスクの性能を左右するため、網羅的に曖昧性を掘り下げ分類し解消する必要があることがわかる。昧性を解消することは各タスクの性能向上の面で有効であること、また人手評価により実際に解消された曖昧性の観点で出力結果が改善していることを確認した。

# 2 AmbiNLG データセット

## 2.1 指示テキストにおける曖昧性の体系化

まず、過去の NLG 研究の文献を参考に、NLG において、指示が曖昧となり得る要素を体系的に整理した [10, 11, 12, 13, 14, 15]。さらに、元の指示において特定のカテゴリに関する指定が曖昧であった場合、それを補完する追加指示を行うためのテンプレートを開発した。以下に構築した曖昧性カテゴリとそれを解消するための指示テンプレートを示す。

**CONTEXT** 背景情報や外部知識等の文脈情報が与えられていない（Additional information: _____ .）

**THEME** 焦点を当てて欲しいトピックが明確に与えられていない（Primarily discuss the following theme: _____）

**PLANNING** もたせるべき大域的な構造が明確に与えられていない（Please generate the output following the outline provided below: Outline: 1. _____ 2. ___ ...）

**LENGTH** 長さ情報（単語数や文数）が与えられていない（Answer with _____ words.）

**STYLE** もたせるべき文体やトーンが与えられていない（Write in a _____ style.）

**KEYWORD** 含めるべき単語やフレーズが与えられていない（Include _____ in your response.）

なお、具体例は Appendix の表 4 に示した。

## 2.2 AmbiNLG データセットの構築

先に定義した曖昧性カテゴリを基に、指示テキストの曖昧性とそれによる生成テキストへの影響を評価できるデータセット、AmbiNLG を構築した。具体的には、既存の指示付き NLG ベンチマークの各インスタンスに対して、各曖昧性カテゴリに対応する追加指示テンプレートを用いて、曖昧性を解消するための穴埋め作業を行い、その追加指示をした際の生成テキストを評価を可能とする。本研究では、指示付き NLG ベンチマークとして、Super-Natural Instructions (SNI) [9] の NLG データから 500 事例を選択し、質問生成や要約、事実文生成、物語生成など 29 タスク・196 種類のデータに対しアノテーションを行った。

## 2.3 アノテーション方法

AmbiNLG のアノテーションには、元の指示テキストと、入出力テキストを慎重に比較し、不足した情報を同定し、それをもとに穴埋めを行う必要があり、非常にコストがかかる。本研究では、LLM-in-the-loop により効率的かつ高品質なデータ作成を試みた [16, 17, 18]。以下に具体的な手順を示す。

**手順 1. 学習データの作成** GPT-3.5 の fine-tuning のための高品質な学習データを GPT-4 と人手作業で作成する。まず、各曖昧性カテゴリにつき事例を 100 件ずつランダムにサンプリングする。次に GPT-4 に入力テキストと出力テキスト、そして穴埋め前のテンプレートを与え、正しく出力テキストを生成するために必要な補完情報を穴埋めさせることで、曖昧性補完テキストを生成する。最後に、人手で生成結果の確認を行い、品質の高い事例のみを fine-tuning のための学習データとして選定した。なお、選定基準は Appendix A に示した。

**手順 2. GPT-3.5 によるデータ拡張** 入力文、指示テキスト、曖昧性カテゴリとテンプレートをモデルに入力し、その指示テキストの曖昧性を与えられた曖昧性カテゴリの観点で補完するようにテンプレートを穴埋めさせる。ここでは、GPT-3.5 を曖昧性カテゴリごとに別途 fine-tuning し、学習済みのモデルを全事例に対し適用した。

なお、補完テキストが出力テキストの直接的なリークにならないよう、テンプレートの穴埋め時には出力テキストを明示的に含めないことを指示に含めている。また、補完テキストと出力テキストとの単語の重複度合いを ROUGE（F 値）で評価したところ、入力テキストと出力テキストのスコアと比較して同程度、あるいは大幅に減少したことから、今回の補完テキストはリークには当たらないと考えている。詳細は Appendix B に示した。

**LLMs なしにデータ作成する曖昧性カテゴリ** 出力テキストから直接抽出できる LENGTH と KEYWORD の補完テキストは、以下のように作成した。

**LENGTH** NLTK [19] を用いて、複文であれば文数、

単文であれば単語数を抽出した。

**KEYWORD** Yake [20] を用いてキーワードの抽出し、重要度の高い Top-$n$ のキーワードを採用した[2)]。

# 3 曖昧性分析

## 3.1 曖昧性解消データセットの分析方法

AmbiNLG データセットの補完テキストを元の指示に追加することで曖昧性を解消し、その挙動を性能と曖昧性の観点から分析する。

**曖昧性解消の評価** 曖昧性が完全に解消されているか否かを判別することは難しい。そこで、補完テキストを指示文に追加したときに、追加前と比較して**より**曖昧性が補完されるか、つまり相対的な**曖昧レベル**の変動に着目することで曖昧性を評価する。具体的には、まず指示テキストを GPT-4 に与え、補完テキストを後から追加したときに追加前と比べて曖昧レベルが「増加」したか「減少」したか「不変」であるかの三値分類を行った。曖昧レベルの減少は、補完テキストが指示テキストの曖昧性補完に寄与したことを意味する。

本分析では、提案する補完テキストに加え、比較対象として別の事例・同じカテゴリの補完テキスト（RANDOM）と 指示テキストから補完テキストと可能な限り単語数の近い文を抽出したもの（OVERLAP）[3)] の 2 種類も用いる[4)]。

**タスク性能に対する曖昧性解消の効果** 補完テキストを元の指示テキストに加えることで実際のタスクを解かせた際の性能が向上するかを調べる。これは、補完テキストの内容がタスクを解くうえで有用であることを示す。指標には、本研究が対象とする多様な NLG タスクに対応するため、SNI [9] と同様に正解文との単語の重複度合いに基づく ROUGE-L F1 スコア [21] を採用する。

## 3.2 曖昧性分析結果

**曖昧性は性能に影響を与えるか** 本分析では、曖昧性補完テキスト、RANDOM、OVERLAP のテキストそれぞれを指示テキストに連結してタスクを解かせた時の性能を、判定された曖昧レベルごとに集計した結果を表 1 に示した。なお、数値は追加テキストなしの性能との差分の平均値である。わかりやすさのため、曖昧レベルが増加したものを「曖昧である」、減少したものを「明瞭である」と表記する。

| | 曖昧である | 不変 | 明瞭である |
|---|---|---|---|
| CONTEXT | -3.66 | +1.78 | **+5.86** |
| KEYWORD | -3.41 | +1.66 | **+9.12** |
| LENGTH | -6.10 | +0.72 | **+3.20** |
| STYLE | -6.18 | **-0.84** | -2.53 |
| THEME | -7.35 | -0.85 | **+3.23** |
| PLANNING | -16.10 | -0.45 | **+8.33** |

**表 1** 曖昧レベルと性能の関係。性能は追加指示テキストを与えない場合との ROUGE スコアの差分

これによると、ほぼ全ての曖昧性カテゴリで、「明瞭である」と判定されたテキストを追加することで、最大で 9.12 ポイント性能が向上する。逆に「曖昧である」と判定された追加テキストを使うとすべての曖昧性カテゴリにおいて性能は大きく低下し、最大で 16.1 ポイント悪化する。このように、指示テキストの曖昧性は確かに性能を悪影響を及ぼすこと、つまりその曖昧性を補完することが性能向上に有効であることがわかる。性能への影響の大きさは曖昧性カテゴリごとに異なり、曖昧性を解消することでほぼすべての曖昧性カテゴリで性能が大きく向上したが、STYLE では悪化した。理由としては、文体制御による単語選択に合わせて出力テキストの表層情報が変化することが挙げられる。

**カテゴリごとの網羅的な曖昧性補完は必要か** 提案手法は、NLG タスクに見られる様々な曖昧性を網羅するために、曖昧性カテゴリごとに別途補完テキストを生成させており、さらにそれらを併用できる。そこで、補完テキストを作成する際に明示的にカテゴリ名とその説明を与える場合と曖昧性カテゴリを与えない場合[5)]、また全カテゴリの補完テキストを同時に連結する場合の性能を比較することで、カテゴリを明示的に、また網羅的に扱うことの重要性について調べる。

表 2 に示したように、明示的にカテゴリ名を与えない場合は、カテゴリ名を明示的に与えた場合に比べて性能が低く、補完テキストを追加しない場合と

2) キーワード数 $n$ は、合計単語数が出力単語数の 40%以下を満たす最大数、あるいは最大で 4 とした。

3) 指示テキストが短い場合は、TEMPLATE の長さに近くなるように指示テキストを複数回繰り返す

4) この曖昧性評価の性能は、RANDOM や OVERLAP に対して全カテゴリ平均で正答率 94%を超えることを確認している（詳細は Appendix B に記載した）。つまり、GPT-4 は情報量や系列長が増えることと曖昧性が解消されることを高精度で区別できている。

5) fine-tuning なしの GPT-3.5 を用いて、可能な限り提案手法と近いプロンプトを用いて、テンプレート “Additional information: ____” を穴埋めさせた。

| | ROUGE | (最小/最大) |
|---|---|---|
| 提案手法（カテゴリ別） | +3.94 | (-2.69/+8.52) |
| カテゴリ指定なし | +0.70 | - |
| 提案手法（全カテゴリ） | **+17.21** | - |

**表 2**　異なる曖昧性解消方法による ROUGE スコア。値は曖昧性解消前との差分の平均値

| | Plain | Con. | Key. | Len. | Sty. | The. | Pln. |
|---|---|---|---|---|---|---|---|
| 人手 | - | 0.88 | 0.96 | 0.82 | 0.92 | 0.98 | 0.94 |
| G-Eval | 4.49 | 4.77 | 4.60 | 4.65 | 4.34 | 4.65 | 4.93 |

**表 3**　カテゴリごとの補完テキストに従った事例の割合（人手）と指示全体に対する IF スコア（G-Eval）

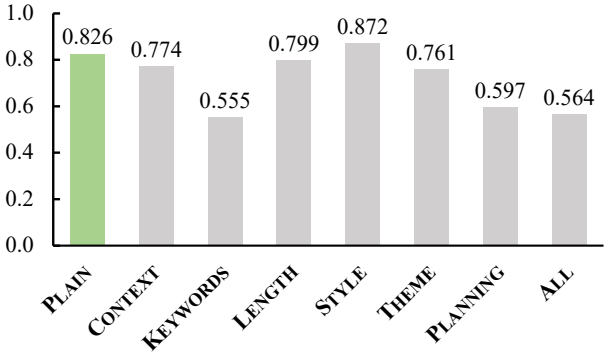
大差ない。一方で、全てのカテゴリの補完テキストを同時に与えた場合は、与えない場合に比べて17.2%もの性能向上が見られた。以上の結果から、曖昧性補完テキストを生成する際は補完させるべき観点を明確化すること、また複数のカテゴリを網羅的に扱うことが有効であることが示された。このためには、本研究のようにデータを分析し、観測される曖昧性カテゴリを体系化することも必要である。

また、生成した結果が実際に指示に従ったものかを人手評価および自動評価した。自動評価では、GPT-4 に基づく NLG のための評価フレームワーク G-Eval [22] を採用した。人手評価では、補完テキストに従っている事例の割合を調べた [6]。G-Eval は、各事例が補完テキストを追加した指示全体にどれだけ従ったかを表す Instruction Following（IF）スコアを調べた。参考値として補完テキストなしの場合のスコアも示した。各曖昧性カテゴリごとに 50 件ランダムサンプリングした結果を表 3 に示したように、人手評価では平均で 92%の事例で生成結果が曖昧性補完テキストの指示に従っていることがわかった。また自動評価でもほとんどの曖昧性カテゴリで Plain よりも指示テキスト全体に従えていることがわかった。以上より、提案手法は曖昧性補完によってタスク性能を向上させられたと考えられる。

**曖昧性は頑健性に影響を与えるか**　曖昧な指示を与えられると、LLM 自身がその曖昧性を適切に補完することを強いられる。しかしその補完方法によって出力が左右されるため、曖昧な指示は明瞭な指示と比べて性能にもばらつきが生じることで頑健性にも悪影響を与える可能性がある。そこで、曖昧

6）Keyword は指定したキーワードのいずれかが入っているか、Length は単語の場合は系列長× 0.4、文の場合は文数× 0.2 の揺れまでを許容した。

**図 2**　各事例 5 回の試行での ROUGE スコアの変動係数。




な指示と明瞭な指示をそれぞれ与えた LLM から複数の出力をサンプリングし、それらに対して性能を求め、その変動係数を比較することで、指示の曖昧性が頑健性に与える影響を調べる。temperature=1.0、top_p=0.9 を採用し、ランダムにサンプリングした 100 件の事例に対してそれぞれ 5 回ずつ推論させたときの ROUGE スコアの変動係数を図 2 に示した。参考値として、補完テキストなしの設定（Plain）での値も示した。変動係数が大きいほど、出力結果のばらつきが大きいことを指す。

結果、ほぼすべての曖昧性カテゴリにおいて、曖昧性解消により変動係数が最大で 27.1 ポイント減少し、出力性能のばらつきを抑制できることがわかった。そのため、曖昧な指示を用いた場合、モデル間の性能比較においてそれらの優位関係が入れ替わり得るため、頑健な比較のためには指示が明瞭であることを担保する必要がある。

## 4　おわりに

本研究では、自然言語生成（NLG）における指示テキストの曖昧性解消手法の提案と、曖昧性解消がタスク性能に与える影響について調査した。既存のベンチマークセットの指示に含まれる曖昧性を体系化し、その曖昧性を補完するためのテキストを曖昧性カテゴリごとのテンプレートを用いて生成した。そして、その補完テキストを元の指示に追加することで複数の曖昧性を網羅的に解消する手法を提案した。分析の結果、曖昧性解消は性能向上に重要であり、特に曖昧性の体系化および網羅性が重要であることがわかった。今後の展望としては、データの曖昧性や曖昧性解消結果の人手評価や、複数モデルを用いた曖昧性解消能力の調査を行いたい。

## 謝辞

有益な助言をしていただいた大阪大学の荒瀬由紀准教授に感謝いたします。

## 参考文献

[1] Tom Brown et al. Language models are few-shot learners. In **NeurIPS**, 2020.

[2] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In **ICLR**, 2022.

[3] Ryuto Koike, Masahiro Kaneko, and Naoaki Okazaki. How you prompt matters! even task-oriented constraints in instructions affect llm-generated text detection. **arXiv preprint arXiv:2311.08369**, 2023.

[4] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. A prompt pattern catalog to enhance prompt engineering with chatgpt. **arXiv preprint arXiv:2302.11382**, 2023.

[5] Yu Wang and Eugene Agichtein. Query ambiguity revisited: Clickthrough measures for distinguishing informational and ambiguous queries. In **HLT**, 2010.

[6] Jonathan Pilault, Xavier Garcia, Arthur Bražinskas, and Orhan Firat. Interactive-chain-prompting: Ambiguity resolution for crosslingual conditional generation with interaction. In **IJCNLP-AACL**, 2023.

[7] Zeqiu Wu, Ryu Parish, Hao Cheng, Sewon Min, Prithviraj Ammanabrolu, Mari Ostendorf, and Hannaneh Hajishirzi. InSCIt: Information-seeking conversations with mixed-initiative interactions. **TACL**, 2023.

[8] Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. AmbigQA: Answering ambiguous open-domain questions. In **EMNLP**, 2020.

[9] Yizhong Wang et al. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In **EMNLP**, 2022.

[10] Ehud Reiter and Robert Dale. Building applied natural language generation systems. **NLE**, Vol. 3, No. 1, pp. 57–87, 1997.

[11] David D. McDonald and James D. Pustejovsky. A computational theory of prose style for natural language generation. In **EACL**, 1985.

[12] Karen Kukich. Design of a knowledge-based report generator. In **ACL**, 1983.

[13] Regina Barzilay and Mirella Lapata. Modeling local coherence: An entity-based approach. In **ACL**, 2005.

[14] David Reitter, Frank Keller, and Johanna D. Moore. Computational modelling of structural priming in dialogue. In **NAACL**, pp. 121–124. Association for Computational Linguistics, 2006.

[15] Angela Fan, David Grangier, and Michael Auli. Controllable abstractive summarization. In **WNMT**, 2018.

[16] Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. Is GPT-3 a good data annotator? In **ACL**, July 2023.

[17] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. Chatgpt outperforms crowd-workers for text-annotation tasks. **PNAS**, 2023.

[18] Haopeng Zhang, Hayate Iso, Sairam Gurajada, and Nikita Bhutani. Xatu: A fine-grained instruction-based benchmark for explainable text updates. **arXiv**, 2023.

[19] Steven Bird. NLTK: The Natural Language Toolkit. In **COLING-ACL**, 2006.

[20] Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. Yake! keyword extraction from single documents using multiple local features. **Information Sciences**, Vol. 509, pp. 257–289, 2020.

[21] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In **Text Summarization Branches Out**, 2004.

[22] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In **EMNLP**, 2023.

[23] Albert Lu, Hongxin Zhang, Yanzhe Zhang, Xuezhi Wang, and Diyi Yang. Bounding the capabilities of large language models in open text generation with prompt constraints. In **Findings of the EACL**, 2023.

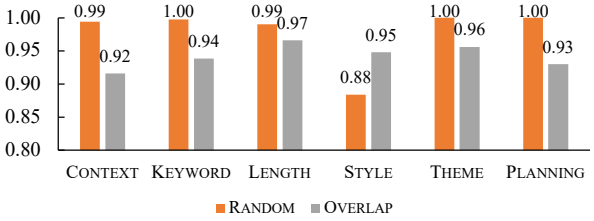
| 曖昧性 | 補完テキスト（下線部が穴埋めされた箇所を指す） |
|---|---|
| Context | Additional information: The main factors of climate change are natural phenomena and human activities. |
| Keyword | Include global warming in your response. |
| Length | Answer with around 5 sentences. |
| Style | Write in a persuasive style. |
| Theme | Primarily discuss the following theme: the impact of human activities. |
| Planning | Please generate the output following the outline provided below: Outline: 1.a brief definition, 2. causes, ... |

**表 4** 曖昧性カテゴリと “Write a summary about climate change.” という指示に対する補完テキストの例

| | Length | Context | Theme |
|---|---|---|---|
| 補完 | 1.10 | 1.08 | 23.14 |
| 入力 | 23.25 | 23.25 | 23.25 |
| | **Style** | **Keyword** | **Planning** |
| 補完 | 4.90 | 25.93 | 28.11 |
| 入力 | 23.25 | 25.36 | 31.41 |

**表 5** 補完テキストと入力テキストそれぞれの出力テキストとの ROUGE スコア

**図 3** 各曖昧性カテゴリでの曖昧性評価の正答率




# A 曖昧性解消データセット詳細

**作成プロセスの詳細** Style の穴埋めは、代表的な項目や先行研究 [23] で用いられている分類[7]への分類問題として行なった。Planning は、テキストの大域的な構造を対象としていることから、出力テキストが 2 文以上から構成される事例にのみ適用した。また、Keyword はキーワード抽出器がキーワードがあると判定した事例のみ対象とした。

**人手によるデータ選定基準** 選定基準は、(1) 内容が適切である (2) 出力テキストが何かが明示する記述を含まない (3) 出力テキストに対して明らかな間違いを含まない (4) 別の曖昧性カテゴリの記述を含まない (5) 指示テキストや入力テキストと大きく内容が重複しない ことである。

# B 分析結果詳細

**補完テキストによる出力テキストのリーク評価** 補完テキストによる出力テキストのリークについて評価するため、出力テキストとの単語の一致度合いを ROUGE で評価した結果を表 5 に示した。参考値として、入力テキストと出力テキストでのスコアも示した。これによると、すべての曖昧性カテゴリにおいて補完テキストは入力テキストでのスコアと比較して同程度あるいは大幅に低い値であった。

**GPT-4 は曖昧性を認識できるか** 3.2 節の分析で用いた追加テキスト（Random、Overlap）の曖昧性カテゴリごとの曖昧性評価の正答率をもとに、GPT-4 がもつ曖昧性の判断能力について調べる。Random が与えられた場合は、情報量が増えているが有用な情報ではなく曖昧性は解消できない。元の指示テキストと内容が競合する場合は曖昧性レベルが上昇することもある。また Overlap が与えられた場合は、情報量も増えておらず曖昧性は解消できない。そのため、曖昧レベルは同程度である。以上より、各テキストを指示テキストに追加した際の曖昧性評価の正答率 $P_R$、$P_O$ は以下のように計算される。両正答率ともに、100%に近いほど良い。

$$P_R = \frac{\textbf{減少・不変}\text{ラベルが振られた事例数}}{\text{総事例数}} \quad (1)$$

$$P_O = \frac{\textbf{不変}\text{ラベルが振られた事例数}}{\text{総事例数}} \quad (2)$$

各追加テキストの正答率を図 3 に示した。結果によると、Random、Overlap に対してほとんどのカテゴリで正答率が 90%を超えた。全カテゴリ平均ではそれぞれ 94%を超える。つまり、GPT-4 は曖昧性が補完された現象と情報量や系列長が増えた現象をある程度明確に区別できていることが示唆される。

7) ‘descriptive, expository, narrative, persuasive, directive, conversational, technical, journalistic, review, poetic, formal, informal, optimistic, assertive, dramatic, humorous, sad, passive-aggressive, worried, friendly, curious, encouraging, surprised, cooperative’