日本語 LLM 生成テキスト識別データセットの構築

森川 周　　白井 清昭　　Natthawut Kertkeidkachorn
北陸先端科学技術大学院大学 先端科学技術研究科
{s2410181,kshirai,natt}@jaist.ac.jp

概要

本論文は日本語 LLM 生成テキスト識別データセットの構築について述べる．本データセットは，同じトピックについて人間が書いたテキストと LLM(大規模言語モデル) が生成したテキストの組を収録したものであり，テキストの作成者が人間か LLM かを識別するタスクのベンチマークとなるものである．人間が書いたテキストとして，論文の概要，ウェブ上の質問応答，ニュース，Wikipedia 記事の 4 種類のテキストを収集する．また，これらに対応するテキストを 4 つの LLM を用いて生成する．さらに，人間・LLM が生成したテキストの違いの分析や，同データセットを用いた LLM 生成テキスト識別タスクの実験について報告する．

1　はじめに

近年，大規模言語モデル (Large Language Model; LLM) は特にテキスト生成の分野で優れた成果を挙げており，人間が書いたテキストとほぼ見分けがつかないテキストを生成できるようになっている．これに伴い，LLM によって生成されたテキスト (以下，LLM 生成テキスト) と人間が作成したテキスト (以下，人間生成テキスト) を自動的に識別することの重要性が高まっている．例えば，LLM によって生成されたレポートや論文を見分けたりフェイクニュースを検出したりすることは実用的な価値が高い．

LLM 生成テキストの自動識別の研究には，人間生成テキストと LLM 生成テキストを収集したデータセットが必要不可欠である．これまで，先行研究によっていくつかのデータセットが構築され，公開されている．Human ChatGPT Comparison Corpus (HC3) は，同一の質問に対して人間と ChatGPT が回答したテキストの組から構成されるデータセットである [1]．HC3 Plus データセットは，HC3 では質問応答タスクのみを対象としていたのに対し，要約，翻訳，言い換えといったタスクを対象に人間生成テキストと LLM 生成テキストを収集している [2]．CHEAT は，計算機科学分野の論文の 15,395 件の概要と，ChatGPT によって生成された 35,304 件の概要から構成されるデータセットである [3]．しかしながら，LLM 生成テキスト識別のためのデータセットの多くは英語を対象に開発されており，日本語テキストを対象としたデータセットの整備は遅れている．

本研究では，日本語 LLM 生成テキスト識別データセットを構築する．同データセットは日本語の人間生成テキストとそれに対応する LLM 生成テキストの組から構成され，以下のような特徴を持つ．(1) 学術論文の概要を中心に，質問応答，ニュース，Wikipedia 記事など多様な種類のテキストを含む．(2) 複数の LLM によって生成されたテキストを含む．さらに，本論文では LLM 生成テキストと人間生成テキストの違いを分析する．また，構築したデータセットを用いて LLM 生成テキスト識別のベースラインモデルを学習し，その性能を実験的に検証する．

2　データセット構築

2.1　学術論文のデータセット構築

学術論文の剽窃の検出に用いることを目標に，論文の概要を対象としてデータセットを構築する．まず，学術論文の概要を人間生成テキストとして収集し，それと同じ論文のタイトルと本文の情報を与えて LLM に概要を生成させる．

学術論文の概要の収集　論文概要の収集には科学技術振興機構 (JST) による電子ジャーナルデータベース J-STAGE を用いる．J-STAGE Web API を使用し，論文誌名を検索クエリとして，論文のメタデータ (タイトルなど) と論文公開ページの URL を取得する．この際，土木，人工知能，精密工学など様々なジャンルの論文を検索の対象とした．次に，論文公開ページから論文の概要と PDF ファイ

ルを取得する．PyMuPDF[1] を用いて PDF ファイルをテキストに変換し，正規表現によるパタンマッチで Introduction と Conclusion に相当する節を抽出する．この際，ヘッダ・フッダ，図表のキャプション，書誌情報など論文の本文以外のテキストを除外する．上記の手続きにより，論文タイトル，概要，Introduction，Conclusion から構成される論文データベースを構築する．

LLM によるテキスト生成　論文タイトル，Introduction，Conclusion を与え，その論文の概要を生成するよう指示するプロンプトを LLM に与える．実際の論文の概要は常体 (である調) で書かれているため，常体でテキストを生成するよう指示する．また，人間が書いた概要に近い文字数の概要を生成するよう指示する．具体的には，論文誌毎に概要の平均文字数を算出し，それに近い文字数の概要を生成するように指示する．ただし，指定された文字数を大きく逸脱するテキストが生成されることも多かったため，5 回を上限として目標文字数の ±50 字の範囲に収まるまでテキストの生成を繰り返す．概要の生成に用いたプロンプトを付録 A に示す．

最後に，生成された概要に対し，短文・空行の除去，句読点の統一 (「、」「。」→「,」「.」)，改行文字の削除などの後処理を行う．

生成に使用した LLM は，GPT-4o [4]，Gemini 1.5 Flash(Gemini) [5]，Llama-3-ELYZA-JP-8B(Llama3) [6]，Llama-3.1-Swallow-8B-Instruct-v0.3(Swallow) [7, 8] の 4 つである．生成時のパラメタは，多様なテキストを生成するため，temperature=0.6，top_p=0.9 とした．

2.2　その他のテキストを対象としたデータセット構築

学術論文の概要に加えて，以下の 3 つのジャンルについて，人間生成テキストと LLM 生成テキストを対にしたデータセットを構築する．ただし，LLM としては Llama と Swallow の 2 つを用いる．

Yahoo!知恵袋　Yahoo!知恵袋[2] は，ユーザが質問を投稿し，別のユーザがその質問に対する回答を投稿するウェブ上の QA 型掲示板である．ここでは，ある質問に対して，それに対する人間の回答と，同じ質問を LLM に与えて生成された回答を組として収集する．Yahoo!知恵袋の質問と回答の組は BCCWJ コーパス [9] からランダムに選択して取得する．LLM に与えるプロンプトとして Yahoo!知恵袋の質問をそのまま与える．

Wikinews　Wikinews はウィキメディア財団が提供するニュースサイトである．ここでは，Wikinews におけるニュース記事を人間生成テキスト，記事タイトルと公開日を与えて LLM に生成させたニュース記事を LLM 生成テキストとして収録する．日本語版 Wikinews 公式ダンプデータ (2025 年 9 月 16 日取得) [10] から，ニュース記事をランダムに選択し，図 3 (付録 A) に示すプロンプトで記事を生成した．

Wikipedia　Wikipedia の記事の導入部 (リード文) を人間生成テキスト，その記事のタイトルを与えて生成したリード文を LLM 生成テキストとしてデータセットを構築する．日本語版 Wikipedia のダンプデータ (2025 年 9 月 2 日取得) [10] から，計算機科学，自然言語処理など 25 種類の情報工学に関連するカテゴリを人手で選定し，そのカテゴリの中からエントリをランダムに選択する．個々のエントリの記事からタイトル，記事本文を抽出し，記事本文からリード文のみを取得する．LLM でテキストを生成する際には図 4 (付録 A) のプロンプトを用いる．

2.3　データセットの統計

構築したデータセットの統計を表 1 に示す．

表 1: データセットの統計

ドメイン	人間	GPT-4o	Gemini	Llama3	Swallow
論文概要	9,343	9,343	9,343	9,343	9,343
Y!知恵袋	904	—	—	904	904
Wikinews	1,828	—	—	1,828	1,828
Wikipedia	2,249	—	—	2,249	2,249

3　LLM 生成テキストの分析

人間生成テキストと LLM 生成テキストの違いを分析する．文長，語彙，品詞，依存関係，Perplexity などについて両者の違いを分析したが，紙面の都合により本論文ではその一部について報告する．[3]

3.1　文長・語彙の違いの分析

表 2 はデータセットにおける平均テキスト長，平均文数，Type/Token 比 (T/T 比) を示したものである．Type/Token 比は単語の異り数をのべ数で割った値で，大きいほど語彙が豊富である (多様な単語を使用している) ことを表す．

論文の概要については人間が書いた概要と同じ程

1) https://pymupdf.readthedocs.io/ja/latest/
2) https://chiebukuro.yahoo.co.jp/
3) 詳細は文献 [11] を参照していただきたい．

度の長さのテキストを生成する処理を行ったため，平均文字数や文数に大きな差は見られない．Yahoo! 知恵袋では，人間による実際の回答 (ベストアンサー) は短文が多いのに対し，LLM は長文を生成することが多いため，LLM 生成テキストの方が平均文字数や平均文数が大きい．一方，Wikipedia については，Llama3 は人間生成テキストより短いテキストを，Swallow は長いテキストを生成している．LLM によって生成するテキスト長が異なることがわかる．

T/T 比を確認すると，論文概要では人間と LLM とで大きな差はない．人間も LLM も Introduction や Conclusion に含まれる単語を多く使うためと考えられる．一方，他のドメインについては人間の方が T/T 比が高く，より多様な単語を用いてテキストを作成していることがわかる．

表 2: 文長・語彙の分析

ドメイン	LLM	文字数 (平均)	文数 (平均)	T/T 比 (×100)
論文概要	人間	321	5.01	1.52
	GPT-4o	373	6.41	1.24
	Gemini	374	6.43	1.30
	Llama3-8B	318	5.27	1.48
	Swallow-8B	371	6.48	1.27
Y!知恵袋	人間	123	3.84	11.7
	Llama3-8B	473	13.6	4.84
	Swallow-8B	439	11.6	4.80
Wikinews	人間	541	9.87	3.96
	Llama3-8B	315	7.20	3.88
	Swallow-8B	368	8.34	3.27
Wikipedia	人間	319	5.12	5.44
	Llama3-8B	197	3.46	4.54
	Swallow-8B	333	6.42	2.66

(わかりやすさのため T/T 比は 100 倍した値を載せている)

3.2 Perplexity の違いの分析

Perplexity (PPL) はあるトークン列の次に出現するトークンの予測のしやすさを測る指標で，小さいほど予測しやすいことを示す．PPL を測る LLM として GPT-2[12]，Llama-3，Swallow を用い，PPL の分布をカーネル密度推定により平滑化して可視化する．分析結果の例として，人間ならびに Swallow によって生成されたテキストを対象とし，GPT-2 を用いて PPL を測った結果を図 1 ならびに付録 B に示す．

人間生成テキストよりも LLM 生成テキストの方が PPL が低い傾向が確認できる．PPL は LLM で測っているため，LLM 生成テキストの方が次のトークンを予測しやすいと考えられる．Llama-3 や

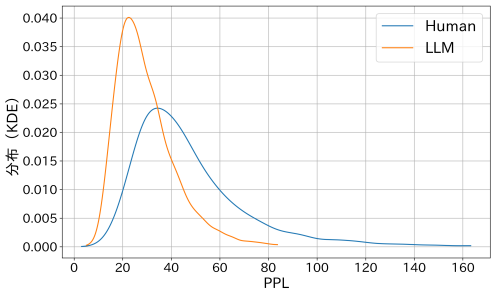


図 1: Perplexity の比較 (論文概要)

Swallow を用いて PPL を測る実験を行ったが，どの LLM を用いても LLM 生成テキストの方が PPL が低い傾向は同じであった．PPL は DetectGPT[13] などで利用されているように，今回の分析でも人間生成テキストと LLM 生成テキストを識別する有力な手がかりになりうることが確認された．

4 LLM 生成テキストの識別

構築したデータセットを用いて LLM 生成テキストを識別する実験を行う．タスク設定は，入力テキストが人間が書いたものか LLM によって生成されたものかを判定する二値分類問題とする．データセットは人間が書いたテキストと LLM 生成テキストの組によって構成されるが，本実験ではいずれか一方を入力し，人間もしくは LLM によって生成されたのかを判定する．

実験では，データセットを訓練，開発，テストデータに 8:1:1 の割合で分割する．次に，訓練データを用いて事前学習済み RoBERTa モデル [14] をファインチューニングする．学習率は $2e^{-5}$，weight decay は 0.01 とする．学習エポック数は最大 3 とし，開発データでの正解率が最大となるエポック数のモデルを選択する．

紙面の都合により本論文では一部の実験結果のみ報告する．全ての実験結果は文献 [11] を参照していただきたい．

in-domain 設定　ドメイン「d」(論文概要，ウェブ QA，Wikinews，Wikipedia) と生成に使用した LLM「l」の組によってデータセットのサブセット $D(d,l)$ を構築する．$D(d,l)$ は人間生成テキストと LLM 生成テキストを同数含む．各サブセットの訓練・開発データを用いて LLM 生成テキスト識別モデルを学習し，同じサブセットのテストデータに適用したときの結果を表 3 に示す．A は正解率 (Accuracy)，P，R，F は LLM 生成テキスト検出の精度 (Precision)，再現率 (Recall)，F1 値 (F1-score) である．いずれのサ

表 3: LLM 生成テキスト識別の結果 – in-domain 設定

ドメイン	LLM	A	P	R	F
論文概要	GPT-4o	0.9888	0.9810	0.9968	0.9888
	Gemini	0.9272	0.8919	0.9722	0.9303
	Llama3-8B	0.9770	0.9598	0.9957	0.9774
	Swallow-8B	0.9668	0.9395	0.9979	0.9678
Y!知恵袋	Llama3	0.9849	0.9800	0.9899	0.9849
	Swallow-8B	0.9558	0.9362	0.9778	0.9565
Wikinews	Llama3-8B	0.9809	0.9679	0.9945	0.9810
	Swallow-8B	0.9973	0.9946	1.0000	0.9973
Wikipedia	Llama3-8B	0.9689	0.9451	0.9956	0.9697
	Swallow-8B	0.9801	0.9617	1.0000	0.9805
マイクロ平均		0.9805	0.9655	0.9966	0.9808
マクロ平均		0.9787	0.9659	0.9928	0.9790

表 4: out-of-domain 設定 (訓練データ=D(論文, GPT))

ドメイン	LLM	A	P	R	F
論文概要	GPT-4o	0.9888	0.9810	0.9968	0.9888
	Gemini	0.9765	0.9806	0.9722	0.9763
	Llama3-8B	0.6966	0.9553	0.4122	0.5759
	Swallow-8B	0.8652	0.9749	0.7495	0.8475
Y!知恵袋	Llama3	0.4975	0.0000	0.0000	0.0000
	Swallow-8B	0.5028	0.0000	0.0000	0.0000
Wikinews	Llama3-8B	0.5082	1.0000	0.0110	0.0217
	Swallow-8B	0.5136	1.0000	0.0272	0.0529
Wikipedia	Llama3-8B	0.5044	1.0000	0.0089	0.0176
	Swallow-8B	0.5155	1.0000	0.0310	0.0601
マイクロ平均		0.8025	0.9758	0.6200	0.7582
マクロ平均		0.6569	0.7892	0.3209	0.3541

表 5: out-of-domain 設定 (訓練データ=D(論文, *))

ドメイン	LLM	A	P	R	F
論文概要	GPT-4o	0.9759	0.9540	1.0000	0.9765
	Gemini	0.9754	0.9540	0.9989	0.9759
	Llama3-8B	0.9706	0.9536	0.9893	0.9711
	Swallow-8B	0.9711	0.9536	0.9904	0.9716
Y!知恵袋	Llama3	0.9397	0.9780	0.8990	0.9368
	Swallow-8B	0.8785	0.9595	0.7889	0.8659
Wikinews	Llama3-8B	0.9126	0.8947	0.9341	0.9140
	Swallow-8B	0.9429	0.9137	0.9783	0.9449
Wikipedia	Llama3-8B	0.8289	0.8627	0.7822	0.8205
	Swallow-8B	0.9004	0.8577	0.9602	0.9061
マイクロ平均		0.9569	0.9417	0.9741	0.9576
マクロ平均		0.9296	0.9282	0.9321	0.9283

表 6: 全サブセットを学習に使用したときの結果

ドメイン	LLM	A	P	R	F
論文概要	GPT-4o	0.9561	0.9193	1.0000	0.9579
	Gemini	0.9561	0.9193	1.0000	0.9579
	Llama3-8B	0.9543	0.9190	0.9964	0.9562
	Swallow-8B	0.9551	0.9191	0.9979	0.9569
Y!知恵袋	Llama3	0.9967	0.9933	1.0000	0.9967
	Swallow-8B	0.9871	0.9748	1.0000	0.9872
Wikinews	Llama3-8B	0.9690	0.9417	1.0000	0.9699
	Swallow-8B	0.9774	0.9568	1.0000	0.9779
Wikipedia	Llama3-8B	0.9756	0.9547	0.9985	0.9761
	Swallow-8B	0.9602	0.9263	1.0000	0.9617
マイクロ平均		0.9594	0.9259	0.9988	0.9609
マクロ平均		0.9688	0.9424	0.9993	0.9698

ブセットにおいても高い性能を示しており，特に再現率は 1 に近い値が得られている．

out-of-domain 設定　次に，out-of-domain 設定，すなわちひとつのサブセットで学習したモデルを他のサブセットのテストデータに適用する実験を行う．実験結果の例として，サブセット D(論文, GPT) で学習したモデルの実験結果を表 4 に示す．訓練データとテストデータが異なる条件下では，表 3 の in-domain の設定と比べて各指標とも大幅に低下することが確認できた．特に Yahoo!知恵袋をテストデータとした場合にはほとんど全てのテキストを人間生成テキストと判断していた．これは，LLM テキスト識別モデルの能力が訓練データのテキストのジャンルに大きく依存することを示唆する．

out-of-domain 設定でのモデルの識別性能をより詳細に分析するために，ドメイン d 毎に人間生成テキストと複数の LLM で生成したテキストを統合したサブセット ($D(d, *)$ と記す) を訓練データとして LLM 生成テキスト識別モデルを学習する．この際，分類クラスの偏りをなくすため，人間生成テキストと LLM 生成テキストの数が同じになるように前者のサンプルを複製する．実験結果の例として，サブセット D(論文, *) を訓練データとして学習したモデルの結果を表 5 に示す．論文概要以外のテキストをテストデータにしたときでも，評価指標は表 4 の結果ほどは低下せず，マイクロ平均やマクロ平均も大幅に改善している．複数の LLM が生成するテキストをデータセットに加えたことで識別モデルの汎用性が向上したことが原因として考えられる．

全サブセットの使用　最後に，LLM 生成テキスト識別モデルの汎用性を向上させるために，全てのサブセットの訓練データを統合したデータセットから識別モデルを学習する実験を行う．結果を表 6 に示す．どのテストデータについても評価指標が高く，表 3 と同等の結果が得られた．したがって，汎用的な LLM 生成テキスト識別モデルを学習するためには，多様なドメインのテキストを含み，かつ様々な LLM で生成したテキストを含むデータセットを構築することが重要である．

5　おわりに

本論文では日本語の LLM 生成テキスト識別タスクのためのデータセットを構築した．本データセットは人間生成テキストが一般公開できるサブセットのみ公開する予定である．

参考文献

[1] Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. arXiv preprint arXiv:2301.07597, 2023.

[2] Zhenpeng Su, Xing Wu, Wei Zhou, Guangyuan Ma, and Songlin Hu. Hc3 plus: A semantic-invariant human chatgpt comparison corpus. arXiv preprint arXiv:2309.02731, 2024.

[3] Peipeng Yu, Jiahan Chen, Xuan Feng, and Zhihua Xia. CHEAT: A large-scale dataset for detecting chatgpt-written abstracts. **IEEE Transactions on Big Data**, Vol. 11, No. 3, pp. 898–906, 2025.

[4] OpenAI. GPT-4o system card, 2024.

[5] Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. https://arxiv.org/abs/2403.05530.

[6] Masato Hirakawa, Shintaro Horie, Tomoaki Nakamura, Daisuke Oba, Sam Passaglia, and Akira Sasaki. elyza/llama-3-elyza-jp-8b, 2024.

[7] Youmi Ma, Sakae Mizuki, Kazuki Fujii, Taishi Nakamura, Masanari Ohi, Hinari Shimada, Taihei Shiotani, Koshiro Saito, Koki Maeda, Kakeru Hattori, Takumi Okamoto, Shigeki Ishida, Rio Yokota, Hiroya Takamura, and Naoaki Okazaki. Building instruction-tuning datasets from human-written instructions with open-weight large language models, 2025.

[8] tokyotech-llm. Llama-3.1-swallow-8b-instruct-v0.3. Hugging Face model card. https://huggingface.co/tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.3 (accessed 2026-01-05).

[9] 前川 喜久雄 (監修), 山崎誠（編）. 書き言葉コーパス–設計と構築–. 講座日本語コーパス 2. 朝倉書店, 2014.

[10] Wikimedia Foundation. Wikimedia downloads. https://dumps.wikimedia.org (accessed 2026-01-07).

[11] 森川周. 日本語を対象とした大規模言語モデル生成テキストの識別 – データセットの構築と特定のモデルに依存しない分類器の実現 –. 修士論文, 北陸先端科学技術大学院大学, 3 2026.

[12] rinna. japanese-gpt2-medium. Hugging Face model card. https://huggingface.co/rinna/japanese-gpt2-medium (accessed 2025-12-22).

[13] Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. Detectgpt: zero-shot machine-generated text detection using probability curvature. In **Proceedings of the 40th International Conference on Machine Learning**, ICML'23, pp. 24950 – 24962. JMLR.org, 2023.

[14] rinna. japanese-roberta-base. Hugging Face model card. https://huggingface.co/rinna/japanese-roberta-base (accessed 2025-12-22).

A　プロンプト

論文概要の生成に用いたプロンプトを図 2 に示す．{{LENGTH}}, {{[INTRODUCTION}}, {{CONCLUSION}} には目標文字数，Introduciton, Conclusion が埋められる．

以下に論文の Introduction と Conclusion を記述します．この文章から {{LENGTH}} 文字数程度の論文概要を生成してください．また，概要はです・ます調ではなく，である調で生成してください．
論文の Introduction：{{INTRODUCTION}}
論文の Conclusion：{{CONCLUSION}}

図 2: 論文概要の生成に用いたプロンプト

ニュース記事，Wikipedia 記事の生成に用いたプロンプトを図 3，図 4 にそれぞれ示す．

記事タイトル: {{TITLE}}
公開日: {{PUBLISH DATE}}
上記のニュース記事の本文を書いてください。

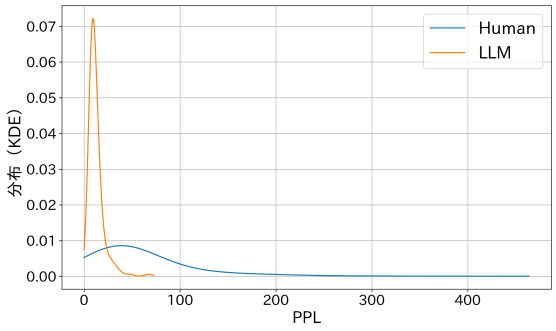
図 3: ニュース記事の生成に用いたプロンプト

{{TITLE}} について，Wikipedia 記事の冒頭の概要部分を作成してください。必ず、で・ある調で作成してください。

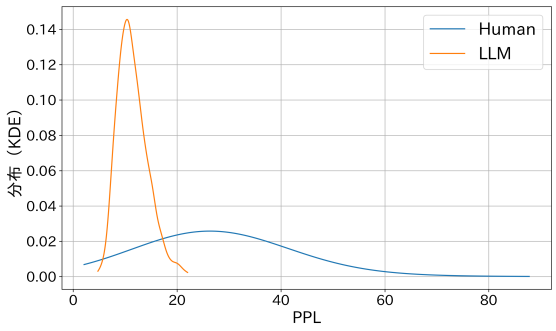
図 4: Wikipedia 記事の生成に用いたプロンプト

B　Perplexity の比較

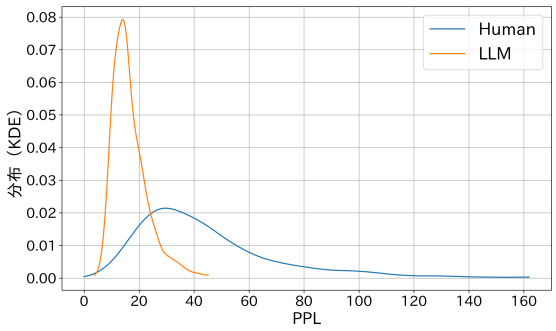
Yahoo!知恵袋、Wikinews、Wikipedia について，人間ならびに Swallow によって生成されたテキストの Perplexity を図 5 に示す．



(a) Yahoo!知恵袋



(b) Wikinews



(c) Wikipedia

図 5: Perplexity の比較