LCTG Bench: 日本語 LLM の制御性ベンチマークの構築

栗原健太郎 [1,2]　三田雅人 [2]　張培楠 [2]
佐々木翔大 [2]　石上亮介 [2]　岡崎直観 [3]
[1] 株式会社 AI Shift [2] 株式会社サイバーエージェント
[3] 東京工業大学
{kurihara_kentaro,mita_masato,zhang_peinan,
sasaki_shota,ishigami_ryosuke}@cyberagent.co.jp
okazaki@c.titech.ac.jp

概要

日本を含む世界中で大規模言語モデル（LLM）の開発や事業における活用が加速していく中で、LLM の性能評価が重要課題になりつつある。LLM の事業における活用では、記事の入稿規程や SEO 対策などを考慮することから、生成結果の内容の品質のみならず、文字数制約や単語の制約などの制御性も LLM の性能の評価対象となり得る。しかし、日本語 LLM の制御性に着目した評価の枠組みは存在しない。本研究では、LLM の事業応用において留意すべき観点の一つである制御性に焦点を当て、評価ベンチマーク LCTG Bench を構築する。GPT-4 [1] や Gemini [2] などの多言語 LLM を含む 11 種類の日本語 LLM を対象とする LCTG Bench を用いたベンチマーク実験を通して, 日本語 LLM の制御性に関する現状と課題を示す。

1　はじめに

OpenAI 社の ChatGPT の公開以降、世界中で大規模言語モデル (Large Language Model: LLM) の研究・開発が加速している。高性能な LLM の開発サイクルに欠かせないのが、LLM の性能評価である。日本語の LLM の性能評価では、日本語言語理解ベンチマーク JGLUE [3] などのデータセットによるリーダーボードや、高性能な LLM による生成結果の一対比較 [4] で品質評価が行われる。代表的なリーダーボードである lm-evaluation-harness [5][1)]では、JGLUE の他に、算術計算データセット MGSM [6]、要約データセット XL-Sum [7] など、多様なタスク [8, 9] が用いられる。生成結果の一対比較では、高性能な LLM として GPT-4 [1]、データセットとして Rakuda Benchmark[2)] や Japanese MT-Bench[3)] などが用いられる。これらのベンチマークでは、流暢性・正確性などの生成の品質に焦点を当てて LLM を評価している。

しかし、LLM の事業における活用では、記事の入稿規程や SEO 対策などを考慮することから、文字数制約や単語の制約の順守など、生成の制御性も求められる。英語では、制御性の評価に焦点を当てた調査 [10] や評価データセットの構築 [11]、単語数やキーワードの有無などの自動評価可能な制御項目の評価 [12] や、要約タスクを基にした事実一貫性などの品質の観点からの制御項目の評価 [13] が実施されている。しかし、日本語においてはこうした制御性に焦点を当てた取り組みは存在しない。

本研究では、日本語 LLM の制御性の評価に焦点を当てたベンチマークとして、LLM Controlled Text Generation (LCTG) Bench を構築する。LCTG Bench は 2 つの言語生成タスクで構成され[4)]、タスク横断的な制御性の評価が行えるように設計されている。本ベンチマークを用いた実験を通じて、GPT-4 などの多言語 LLM を含む 11 種類の日本語 LLM の制御性に関する現状と課題を示す。

2　LCTG Bench の構築

LCTG Bench の構成を表 1 に示す。LCTG Bench は要約タスク、広告文生成タスクの 2 つから構成されており、LLM の制御性能をタスク横断的に評価する。ただし、LLM の生成を評価する上で、正解の生成結果に依存する評価は LLM の生成の多様性を考

1)　本稿〆切時点、最新のリーダーボードは https://rinnakk.github.io/research/benchmarks/lm/ で掲載。

2)　https://yuzuai.jp/benchmark

3)　https://github.com/Stability-AI/FastChat/tree/jp-stable/fastchat/llm_judge

4)　今後更にタスクを追加する予定である。

表 1 LCTG Bench の構成

タスク	データ	フォーマット	文字数	キーワード	NG ワード
要約	ABEMA TIMES	120	120	120	120
広告文生成	CAMERA	150	150	150	150

以下の条件で与えられた文章を要約して出力してください。
[条件]
70 文字以上、180 文字以下で要約すること
[文章]
小学館「週刊少年サンデー」にて連載中の『葬送のフリーレン』（原作・山田鐘人、作画・アベツカサ）の TV アニメ化が決定し、キービジュアルが公開された。
...
キャラクターの佇まいからも彼らの気持ちが伝わると良いなと思います。」と、ビジュアルに込めた想いを語っている。※種崎敦美の「崎」は、正式にはたつさきの字 (C) 山田鐘人・アベツカサ／小学館／「葬送のフリーレン」製作委員会

図 1 要約タスクのプロンプトの例（制御項目は文字数）

以下の [文章] で与えられた説明文に対する広告文のタイトルを、[条件] に従って 1 つ作成してください。
[条件]
「it エンジニア」という単語を使って広告文を生成
[文章]
IT/Web エンジニア採用に特化した求人・スカウトサービス「フォークウェル ジョブズ」は、経験、知識、スキルともに専門性の高い即戦力エンジニア 44,000 人が集まるスカウトサービスです。即戦力エンジニアのデータベースからマッチした人材に直接アプローチが可能！

図 2 広告文生成タスクのプロンプトの例（制御項目はキーワード）

慮できていないという課題がある。本ベンチマークでは、各事例に正解の生成結果を用意せず入力のプロンプトのみを用意することで、LLM の出力の制御性を自動評価できるようにする。

2 つのタスクのプロンプトの例を図 1, 2 に示す。各プロンプトは「タスクの指示文」「制御項目に関する条件文」「タスクの対象となる文章」の 3 要素で構成する。ただし、同じ意味でも異なる表現でプロンプトが入力されることでモデルの評価結果も変化し得る [14]。そこで、制御項目に関する条件文のテンプレートは、同じ条件を与える多様な表現を収録するため、クラウドソーシング[5] も用いて収集した[6]。タスクの対象となる文章は、公開済みデータセット及びサイバーエージェントが保有する公開可能なデータより収集した。クラウドソーシングによるテンプレートの収集方法と収集例を付録 A に示す。

2.1 評価する制御項目

フォーマット LLM の出力フォーマットの制御に関する課題が指摘されている中で [15]、Function Calling[7] などの外部ツールを利用した後処理を必要とする場面がある。ところがその精度は完璧とは言えず、事業での適用においては、LLM の出力フォーマットに関する制御性が要求される。本研究では、フォーマットの観点における基本的な性能を「出力の前後に不必要な説明文などを付与しない生成の性能」として評価する。要求する生成物以外の文を付与しないよう指示する条件文（付録 B）を作成する。

文字数 事業における記事やタイトルの作成などの応用において、文字数に制限が付くことがある。本研究では、「指定した範囲内の文字数で生成することができるか」を評価する。条件文のテンプレートはクラウドソーシングを用いて収集する。

キーワード・NG ワード Web サイトのタイトルやキャッチコピーの生成においては、SEO 対策などの理由から生成結果に含めたい単語を指定したいケースがある。また、誇大広告になることを防ぐため、使用を避けるべき（NG）単語や表現が存在するケースもある。本研究では、「キーワードを含んだ文章、および NG ワードを含まない文章を生成することができるか」を評価する。条件文のテンプレートはクラウドソーシングを用いて収集する。

2.2 タスクの概要と条件文の作成方法

要約 難易度が高く事業でも適用例が多い言語生成タスクとして要約タスクを採用する。要約タスクでは、図 1 に示すように、条件に従って文章を要約するよう指示したプロンプト集合のデータセットを構築する。要約元の文章は、ニュースサイト “ABEMA TIMES”[8] の記事から 6 カテゴリ[9] の 120 件を用いる。条件文のテンプレートに代入する値について、文字数には値の上限と下限の制約を設け、上限 200 文字、下限 50 文字のランダムな 10 の倍数の値を代入する。キーワードは、要約元の記事に登場する重要度の高い単語から選択する。NG ワードについて、実際に要約を生成する場面で重要度の高い単語は NG ワードに指定されづらい可能性がある。

5) Yahoo!クラウドソーシング https://crowdsourcing.yahoo.co.jp/ を用いた。
6) 条件文のテンプレートのみを収集し、具体的な値については各制御項目毎に収集する。例えば文字数に関する条件文は、テンプレートとして「X 字以上 Y 字以内で要約して下さい」を収集し、X, Y にはランダムに生成した数値を代入することで条件文を作成する。
7) https://platform.openai.com/docs/guides/function-calling
8) https://times.abema.tv/
9) 公開上の制約から「ニュース」以外のカテゴリから「エンタメ」「スポーツ」「アニメ」「将棋」「麻雀」「HIPHOP」の 6 つを選出した。

与えられた文章に「it エンジニア」という単語を用いて、タイトルを作成いたします。

タイトル: 即戦力 it エンジニアのデータベースからマッチした人材に直接アプローチ

このタイトルは、条件に従って「it エンジニア」という単語を用いて、文章に含まれているキーワードをタイトルに含めることで、検索結果で上位に表示させることができます。

図 3 広告文作成タスクにおける、内容と関係のない説明文が付与された LLM の生成結果の例：文字数制御の評価において、説明文の部分も文字数に含まれてしまう。

しかし、本タスクにおいては指定した表現を除いた出力をする能力を測るという観点から、キーワードと同様に要約元の記事に登場する重要度の高い単語から選択する。具体的には、各サンプルの記事要約を GPT-4 を用いて 5 つ生成し、得られた要約集合に共通して出現する単語を重要度の高い単語として 2 つ抽出し、キーワード・NG ワードとする。これは、GPT-4 のような高性能な LLM の生成する要約に出現しやすい単語は文書中での重要度が高い単語であるという仮定に基づいている。

広告文生成 要約タスクと比較して要求される出力の文字数が少なく、キーワードの重要度が高いタスクとして広告文生成タスクを導入する。広告文生成タスクでは、図 2 に示すように、条件に従って広告文のタイトルを作成するよう指示したプロンプト集合のデータセットを構築する。広告文のタイトル生成の元となる文章は、広告文生成ベンチマーク CAMERA [16][10] の評価データのうち付与された検索キーワードが 2 件の事例について、LP テキスト部分から収集する。文字数については、要約タスクと同様に上限と下限の制約を設けつつも、要約タスクと比較して少ない文字数帯での出力を期待するタスクであることから、上限 50 文字、下限 20 文字のランダムな 5 の倍数の値を代入する。キーワード・NG ワードについては、CAMERA のサンプルに付与されている検索キーワードを使用する。

3 LCTG Bench を用いた LLM 評価

LCTG Bench を用いた日本語 LLM の評価実験を実施することで、日本語 LLM の現状と課題および本ベンチマークの有用性を示す。

3.1 モデル

実験に用いるモデルは、GPT-4 などの高性能とされているモデルの他、Llama 2 [18] や GPT-NeoX [19] などのベースとするモデルの種類やパラメータ数に多様性を持たせるよう選択した。また、各種 LLM のハイパーパラメータおよびシステムプロンプトは、原則 Hugging Face Hub に掲載されている値を既定値とみなして使用した。実験に用いたモデルと各種ハイパーパラメータの設定を付録 C に示す。

3.2 評価

制御性能の評価のみでは、制御性を満たしつつも生成内容が著しくタスクと乖離がある挙動を見逃す恐れがあることから、生成の品質評価も併せて行う。また、LLM は同じプロンプトの入力に対して異なる生成結果を出力するため [20]、評価結果にも揺れが生じる。そこで 1 つのプロンプトに対して 3 回生成を行い、各回ごとのスコアの平均値を最終的なスコアとする。さらに、LLM は要求するタスクの内容と関係のない説明文も付与した生成をすることがある。評価する制御項目のうち、「フォーマット」では説明文の有無を評価するものの、他 3 項目、および生成の品質評価については説明文が付与されていることで、タスクに対する生成結果の評価としてはノイズとなり得る（図 3）。そこで、フォーマット以外の 3 つの制御項目と生成の品質評価においては、LLM の生成結果から GPT-4 を用いて不要な説明文を除去したのちに評価を実施する[11]。

制御性能の評価 フォーマットに関して、GPT-4 による不要文の除去操作の前後の生成結果を比較し、要約タスクについては前後 10 文字、広告文生成タスクについては前後 5 文字が完全一致している事例の割合を算出する[12]。文字数に関しては、生成結果の文字数が条件文で指定した文字数の範囲内に収まっている事例、キーワード・NG ワードに関しては、生成結果において条件文で指定した単語が含まれている（いない）事例の割合をそれぞれ算出する。

生成の品質評価 生成の品質評価については Rakuda Benchmark などにおける評価方法に倣い、評価器として GPT-4 を活用する。GPT-4 を用いて付録 E に示すプロンプトを入力することで、適切な生成ができているか否かの 2 値分類を実施し、適切な生成ができている事例の割合を算出する。

10） https://github.com/CyberAgentAILab/camera

11） 不要な説明文の除去に用いたプロンプトを付録 D に示す。

12） GPT-4 による除去操作により生成結果の中間部分が変更される恐れがある。そのため、除去操作前後の生成結果の完全一致による不要な説明文有無の判断は困難であるため、前後の文字列の比較を採用する。

表 2 要約タスクの制御性 (CTG) と生成の品質 (Quality) の評価結果（ca: cyberagent, line: line-corporation）

モデル	フォーマット		文字数		キーワード		NG ワード		Average	
	CTG	Quality	CTG	Quality	CTG	Quality	CTG	Quality	CTG	Quality
gpt-4-1106-preview (GPT-4 Turbo)	**0.992**	**0.925**	0.450	0.869	**0.972**	**0.886**	**0.970**	0.775	**0.846**	**0.864**
gemini-pro [2]	0.914	0.894	**0.486**	**0.881**	0.939	0.856	**0.645**	**0.817**	0.746	0.862
ca/llama2-7b-chat-japanese	0.708	0.675	0.206	0.606	0.794	0.553	0.400	0.575	0.527	0.602
ca/llama2-13b-chat-japanese	0.708	0.767	0.336	0.725	0.805	0.714	0.394	0.722	0.561	0.732
ca/calm2-7b-chat	0.881	0.478	0.219	0.428	0.808	0.444	0.303	0.403	0.553	0.438
ca/mistral-7b-chat [17]	0.914	0.792	0.278	0.742	0.884	0.742	0.219	0.750	0.574	0.756
elyza/ELYZA-japanese-Llama-2-7b-fast-instruct	0.458	0.789	0.325	0.792	0.803	0.803	0.305	0.778	0.473	0.790
rinna/youri-7b-chat	0.911	0.692	0.166	0.683	0.647	0.609	0.492	0.611	0.554	0.649
line/japanese-large-lm-3.6b-instruction-sft	0.344	0.067	0.125	0.056	0.597	0.044	0.525	0.056	0.398	0.056
llm-jp/llm-jp-13b-instruct-full-jaster-v1.0	0.253	0.047	0.003	0.039	0.364	0.017	0.808	0.036	0.357	0.035
matsuo-lab/weblab-10b-instruction-sft	0.944	0.530	0.194	0.500	0.614	0.458	0.500	0.495	0.563	0.496

表 3 広告文生成タスクの制御性 (CTG) と生成の品質 (Quality) の評価結果（ca: cyberagent, line: line-corporation）

モデル	フォーマット		文字数		キーワード		NG ワード		Average	
	CTG	Quality	CTG	Quality	CTG	Quality	CTG	Quality	CTG	Quality
gpt-4-1106-preview (GPT-4 Turbo)	0.960	**0.987**	0.222	**0.971**	**0.851**	0.711	**0.991**	**0.924**	0.756	**0.898**
gemini-pro	0.956	0.931	0.494	0.942	0.796	0.655	0.886	0.884	**0.783**	0.853
ca/llama2-7b-chat-japanese	0.391	0.822	0.416	0.854	0.564	0.731	0.807	0.736	0.544	0.786
ca/llama2-13b-chat-japanese	0.071	0.887	0.484	0.920	0.620	**0.738**	0.918	0.866	0.523	0.853
ca/calm2-7b-chat	0.860	0.735	0.396	0.762	0.449	0.618	0.664	0.661	0.592	0.694
ca/mistral-7b-chat	0.686	0.946	0.398	0.935	0.644	0.720	0.813	0.853	0.635	0.864
elyza/ELYZA-japanese-Llama-2-7b-fast-instruct	0.749	0.615	0.236	0.817	0.640	0.542	0.773	0.684	0.600	0.665
rinna/youri-7b-chat	0.724	0.320	0.200	0.347	0.524	0.258	0.649	0.259	0.524	0.296
line/japanese-large-lm-3.6b-instruction-sft	0.582	0.309	0.080	0.269	0.440	0.242	0.609	0.283	0.428	0.276
llm-jp/llm-jp-13b-instruct-full-jaster-v1.0	**0.991**	0.218	**0.951**	0.227	0.578	0.131	0.853	0.204	0.618	0.195
matsuo-lab/weblab-10b-instruction-sft	0.876	0.671	0.255	0.651	0.376	0.563	0.709	0.582	0.554	0.617

3.3 結果・考察

2 つのタスクにおける各モデルの制御性能と生成の品質の評価結果を表 2, 3 に示す。全般的に GPT-4 及び gemini-pro が制御性能、生成品質のいずれにおいても高スコアを獲得している。ただし、文字数制約については全てのモデルでスコアが低い。この結果は、言語モデルの Tokenizer がトークン単位で処理が行われるがゆえに、日本語の文字数を正確に捉えられていないことが原因と考えられる。要約タスクにおいては、NG ワードの制御性についても GPT-4 以外のモデルで低スコアの傾向にある。本結果は、LLM が否定表現を適切に捉えることができないという報告 [21] と一致している。

2 つのタスクの同じ制御項目を比較すると、スコア差が大きいモデルが存在すること、特にキーワード・NG ワードにおいてモデルを問わず全般的にスコア差が大きいことがわかる。本結果より、同じ制御項目でもタスクによりその難易度が異なる可能性を本ベンチマークが示唆していると言える。また、llm-jp のモデルが広告文生成タスクの一部制御項目で高スコアを獲得しているが、生成の品質のスコアが著しく低いことから、広告文生成タスクで高い制御性能を持つと断定することはできない。例えば、図 2 を入力した際の llm-jp のモデルの出力の一つは「it エンジニア」であり、条件は満たすものの広告文のタイトルとしては意味を成していないことが確認できる。これは、JGLUE などの分類問題において、存在しないラベルを生成し不正解となっているが意味的には正解しているというような事例を、本ベンチマークでは捕捉することができることを示唆する。さらに、本結果における Average のスコアのランキングは、既存ベンチマークの Rakuda Benchmark や lm-evaluation-harness に掲載されているモデルの性能のランキングと異なっていることから、本ベンチマークを用いることで、既存ベンチマークとは異なる新たな視座を提供できるといえる。

4 おわりに

本研究では、日本語 LLM の制御性能を評価するためのベンチマークとして LCTG Bench を構築し、全 11 種類の日本語 LLM の制御性能を評価し、日本語 LLM の現状と今後の課題を示した。LCTG Bench は 2024 年中に公開予定である[13)]。

13) https://huggingface.co/datasets/kkurihara-cs/LCTG-Bench で公開予定。

参考文献

[1] OpenAI. Gpt-4 technical report. **ArXiv**, Vol. abs/2303.08774, , 2023.

[2] Gemini Team. Gemini: A family of highly capable multimodal models. 2023. abs/2312.11805.

[3] Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. JGLUE: Japanese general language understanding evaluation. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, pp. 2957–2966, Marseille, France, June 2022. European Language Resources Association.

[4] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. 2023. abs/2306.05685.

[5] Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, September 2021.

[6] Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. Language models are multilingual chain-of-thought reasoners. 2022. abs/2210.03057.

[7] Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, **Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021**, pp. 4693–4703, Online, August 2021. Association for Computational Linguistics.

[8] Alexey Tikhonov and Max Ryabinin. It's All in the Heads: Using Attention Heads as a Baseline for Cross-Lingual Transfer in Commonsense Reasoning. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, **Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021**, pp. 3534–3546, Online, August 2021. Association for Computational Linguistics.

[9] 鈴木正敏, 鈴木潤, 松田耕史, 田京介, 井之上直也. Jaqket: クイズを題材にした日本語 qa データセットの構築. 言語処理学会第 26 回年次大会, 2020.

[10] Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. A survey of controllable text generation using transformer-based pre-trained language models. 2023. abs/2201.05337.

[11] Jiao Sun, Yufei Tian, Wangchunshu Zhou, Nan Xu, Qian Hu, Rahul Gupta, John Frederick Wieting, Nanyun Peng, and Xuezhe Ma. Evaluating large language models on controlled generation tasks. 2023. abs/2310.14542.

[12] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. 2023. abs/2311.07911.

[13] Yixin Liu, Alexander R. Fabbri, Jiawen Chen, Yilun Zhao, Simeng Han, Shafiq Joty, Pengfei Liu, Dragomir Radev, Chien-Sheng Wu, and Arman Cohan. Benchmarking generation and evaluation capabilities of large language models for instruction controllable summarization. 2023. abs/2311.09184.

[14] Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. State of what art? a call for multi-prompt llm evaluation. 2023. abs/2401.00595.

[15] Qianyu He, Jie Zeng, Wenhao Huang, Lina Chen, Jin Xiao, Qianxi He, Xunzhe Zhou, Lida Chen, Xintao Wang, Yuncheng Huang, Haoning Ye, Zihan Li, Shisong Chen, Yikai Zhang, Zhouhong Gu, Jiaqing Liang, and Yanghua Xiao. Can large language models understand real-world complex instructions? 2023. abs/2309.09150.

[16] Masato Mita, Soichiro Murakami, Akihiko Kato, and Peinan Zhang. Camera: A multimodal dataset and benchmark for ad text generation. 2023. abs/2309.12030.

[17] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. 2023. abs/2310.06825.

[18] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. 2023. abs/2302.13971.

[19] Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. Gpt-neox-20b: An open-source autoregressive language model. 2022. abs/2204.06745.

[20] Shuyin Ouyang, Jie M. Zhang, Mark Harman, and Meng Wang. Llm is like a box of chocolates: the non-determinism of chatgpt in code generation. 2023. abs/2308.02828.

[21] Thinh Hung Truong, Timothy Baldwin, Karin Verspoor, and Trevor Cohn. Language models are not naysayers: an analysis of language models on negation benchmarks. In Alexis Palmer and Jose Camacho-collados, editors, **Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)**, pp. 101–114, Toronto, Canada, July 2023. Association for Computational Linguistics.

表 4 実験に用いた LLM のリストとハイパーパラメータの設定 (*は本稿〆切時点は未公開のモデル)

モデル	ベースモデル	max_new_tokens	temperature	top_p
gpt-4-1106-preview (GPT4-Turbo)	-	-	-	-
gemini-pro	-	-	-	-
cyberagent/llama2-7b-chat-japanese*	Llama 2	4,096	0.9	-
cyberagent/llama2-13b-chat-japanese*	Llama 2	4,096	0.9	-
cyberagent/calm2-7b-chat	Llama 2	4,096	0.8	-
cyberagent/mistral-7b-chat*	Mistral	4,096	0.8	-
elyza/ELYZA-japanese-Llama-2-7b-fast-instruct	Llama 2	4,096	-	-
rinna/youri-7b-chat	Llama 2	4,096	0.5	-
line-corporation/japanese-large-lm-3.6b-instruction-sft	-	4,096	1.0	-
llm-jp/llm-jp-13b-instruct-full-jaster-v1.0	-	4,096	0.7	0.95
matsuo-lab/weblab-10b-instruction-sft	GPT-NeoX	4,096	0.7	0.95

A クラウドソーシングによる条件文の収集方法と収集例

2.1 節で述べている制御項目のうち、「文字数」「キーワード」「NG ワード」の 3 項目の条件文のテンプレートを収集した。条件文は、ある与えられた条件文を同じ意味になるように書き換えるタスクを実施することで収集した。収集事例を下記に示す。

- 文字数: XXX-YYY 文字で要約して, XXX-YYY 文字でまとめること, XXX 文字以上、XXX 文字以下で要約 など
- キーワード・NG ワード: 「XXX」という単語を含める, XXX という言葉を使ってください。, 「XXX」という単語は入れない, 「XXX」という言葉は使用不可 など

B フォーマットの条件文

フォーマットの条件文には、固定の文を用いた。2 つのタスクそれぞれで用いたフォーマットの条件文を以下に示す。

- 要約タスク: 文章の要約結果のみを出力し、要約結果の前後に説明文などは付与しないでください。
- 広告文作成タスク: 広告文のみを出力し、広告文の前後に説明文などは付与しないでください。

C 実験に用いた LLM

実験に用いた LLM の一覧と設定したハイパーパラメータの値を表 4 に示す。

D 不要な説明文の除去プロンプト

LLM の生成結果に含まれる不要な説明文の除去の際に、GPT-4 に入力したプロンプトを図 4,5 に示す。

```
以下に提示している文章は、ある文章を生成 AI を用いて要約した出力結果です。
出力には「要約」あるいはそれに類する単語を含むような文として、「以下の文章を要約します。」「【要約】」などの冒頭の説明文や「以上が要約結果になります。」などの文末の説明文が入っていることがあります。また、英語でこれらの説明文が与えられることもあります。
提示した文章に上記で述べた説明文が含まれていない場合には提示した文章をそのまま出力し、上記で述べた説明文が含まれている場合は提示した文章から説明文を除去したものを抜き出してください。文章の中間部分を編集する必要は一切ありません。文が入っていることがあります。また、英語でこれらの説明文が与えられることもあります。
[文章]
{ 生成結果 }
```

図 4 要約タスクにおける不要説明文の除去に用いたプロンプト

```
以下に提示している文章は、ある文章を元に作成した広告文のタイトルです。
出力には「広告文：」や「広告文を作成します」などの冒頭の接頭辞や説明文、「作成しました。」「このタイトルは、、」などの接尾辞やタイトルの後ろの説明文が含まれていることがあります。
提示した文章に上記で述べた説明文や接頭辞、接尾辞が含まれていない場合には、提示した文章をそのまま出力してください。「」や**などの記号で囲われている事例の場合、記号は全て残したまま出力してください。
上記で述べた説明文が含まれている場合は提示した文章から説明文や接頭辞、接尾辞を除去したものを抜き出してください。冒頭、末尾以外の中間部分を編集する必要は一切ありません。新しく文字を追加などをしないでください。

[文章]
{ 生成結果 }
```

図 5 広告文生成タスクにおける不要説明文の除去に用いたプロンプト

E 生成の品質評価プロンプト

LLM の生成の品質評価をするために、GPT-4 に入力したプロンプトを 6,7 に示す。

```
以下に要約した文章とその要約元の文章が提示されています。
要約した文章は要約元の文章を適切に要約できているかを判断してください。
適切に要約できている場合は「適切」、適切に要約できていない場合は「不適切」と回答してください。
ただし、要約元の文章から断定できない情報が要約した文章に含まれている場合も「不適切」と回答してください。
「適切」「不適切」のいずれかのみを出力し、説明文などは付与しないでください。
【要約元の文章】
{ 要約元の文章 }

【要約した文章】
{ 生成結果 }
```

図 6 要約タスクにおける生成の品質評価に用いたプロンプト

```
以下に、ランディングページの説明文とその説明文をもとに作成した 1 つの広告文のタイトルがあります。
説明文の内容に基づいているタイトルを作成できているかを判断してください。
適切に作成できている場合は「適切」、適切に作成できていない場合は「不適切」と回答してください。
ただし、説明文とタイトルが完全に一致している事例とタイトルとして長すぎる事例も「不適切」と回答してください。
「適切」「不適切」のいずれかのみを出力し、説明文などは付与しないでください。

【説明文】
{LP テキスト }

【広告文のタイトル】
{ 生成結果 }
```

図 7 広告文生成タスクにおける生成の品質評価に用いたプロンプト