# 金融分野に特化した複数ターン日本語生成ベンチマークの構築

平野正徳　今城健太郎
株式会社 Preferred Networks
research@mhirano.jp　imos@preferred.jp

## 概要

大規模言語モデル (LLM) の発展に伴い、様々な分野において性能を評価する取り組みが必要となってきている。本研究では、金融分野において LLM の生成の良さ測るための日本語生成ベンチマーク pfmt-bench-fin-ja を提案した。pfmt-bench-fin-ja は、MT-bench に対応するような金融分野に特化した複数ターンの日本語生成ベンチマークであり、12 カテゴリー、360 問のベンチマークを新たに構築した。評価にあたっては、GPT-4o-mini を LLM-as-a-judge として用いて、10 段階評価でスコア計測をすることとした。実験として、複数の LLM に対してベンチマークを計測し、その結果を比較検討した。その結果、pfmt-bench-fin-ja が一定レベルで LLM の性能評価を行うことができることが示された。構築したベンチマークは Github より利用可能である。

## 1　はじめに

大規模言語モデル (LLM) は、近年、著しい性能を発揮している。特に、ChatGPT[1] や GPT-4[2] をはじめとした最新の言語モデルは、性能向上と汎化が著しい。その基本技術は Transformer[3] から始まっており、BERT や [4] や、GPT シリーズ [5, 6, 7] などが続いた。ほかにも、Bard[8] や LLaMA[9, 10]、Dolly[11]、BLOOM[12]、Vicuna[13]、PaLM[14, 15] などのモデルが提案されている。

しかしながら、これらの大規模言語モデルは、様々なタスクでどの程度の性能を発揮するかは未知数であり、それらを評価する取り組みが進められている。たとえば、Language Model Evaluation Harness (lm_eval) [16] と呼ばれる、LLM 用の様々なタスクによるベンチマーク計測プラットフォームが提案されている。また、GPT-4[2] においても、様々なタスクにおける性能を評価している。さらに、会計士試験の達成度 [17] や医学分野における応用 [18]、法律分野への応用 [19, 20] を検証する研究などが存在する。

特に、金融分野への応用は、様々なタスクへの応用可能性から、研究開発が進んでいる。金融に特化した非公開のモデルとして、BloombergGPT[21] が存在するほか、公開されているモデルとしては、LLaMA[9] をチューニングした FinLLAMA[22] や、FinGPT[23]、Instruct-FinGPT[24] などがある。

加えて、日本語に特化した LLM という点でも、様々な開発が進んでいる。CyberAgent の CALM シリーズや rinna 社のモデル、stabilityai 社の stablelm シリーズ、Elyza 社のモデル、Preferred Networks 社の Plamo[25] など、様々なモデルが乱立している。

こうした特化型 LLM の台頭に際して、これらのモデルの性能を正しく評価する特化型のベンチマークの構築も必要であると考えられる。日本語に特化したベンチマークはすでに存在しており [26]、さらに金融に特化したものとしては、筆者らが以前提案した Japanese Language Model Financial Evaluation Harness (japanese-lm-fin-harness) [27] が存在している。

本研究においては、金融における日本語ベンチマークに着目し、japanese-lm-fin-harness ではカバーできていない、生成の良さを測るためのベンチマークを構築する。japanese-lm-fin-harness では、主に知識を解く多肢選択問題を用いているが、生成の良さを測るためには、より自然な文を生成するタスクが必要である。加えて、様々な LLM の性能を真に比較するためには、AI アシスタントとしての性能を測るタスクが必要である。一方で、こうした生成の良さを測るベンチマーク手法としては、MT-bench [28] のように、非常に強力な GPT-4 のような LLM を LLM-as-a-judge として用いて、出力の良さをスコア化することで評価を行うという手法がある。そこで、本研究では、この MT-bench に対応するような金融分野に特化した複数ターンの日本語生成ベンチマークを構築する。

なお、本研究で構築したベンチマークは https://github.com/pfnet-research/pfmt-bench-fin-ja で公開している。

## 2 pfmt-bench-fin-ja: Preferred Multi-turn Benchmark for Finance in Japanese

本研究では、日本語における金融タスクの性能を生成の良さの観点から評価するベンチマークを構築することを目指す。ここでは、MT-bench [28] を参考にすることとした。MT-bench とは、複数ターン (主に 2 ターン) で、ユーザーの質疑に適切に回答できるかを比較するベンチマークである。MT-bench には Writing、Roleplay、Extraction、Reasoning、Math、Coding、Knowledge I (STEM)、Knowledge II (humanities/social science) の 8 つの分類の質問が合計 80 件含まれており、これらの質問への LLM の回答を GPT-4 などの LLM で 10 段階評価することでその良さを計るベンチマークとなっている。

本研究で提案するベンチマークにおいては、上記の 8 つの分類の中から Knowledge I (STEM)、Knowledge II (humanities/social science) を取り除き、Self-Instruct[29] の質問を分類した場合にカバーできなかった問題を考慮に入れて Idea, Translation, Ethics の 3 分類を新たに追加し、さらに、金融に特化させるために Knowledge、Trustworthiness、ESGs の 3 分類を追加した。

各タスクの目標・主旨については、それぞれの項目において重要であると想定される内容を踏まえ、下記の通り新たに定義した。

1. Writing (作文): 金融に関するあらゆるタイプの文章作成に関する相談を含みます。文法の正確性、文章の流れ、語彙の選択、効果的なメッセージ伝達に重点を置き、投資家やステークホルダーに響く文章を書くための技術や方法に関する相談に応じます。アナリストレポートの構成、説得力のある株主総会におけるプレゼンテーションの作成方法など、目的に合わせた書き方に関するアドバイスも提供します。
2. Roleplay (役の演技): 特定の人の視点を採用して行動や発言を想定する質問が含まれます。オペレーター、ファイナンシャルプランナー、証券アナリストなどの職業に応じた適切な反応を考えること、または特定の状況下での対人関係のシミュレーションなど、想像力と創造性を駆使した対話や行動指針の提案を行います。
3. Knowledge (金融知識): 金融の基本知識をはじめ、幅広くかつ深い金融に関連する知識を持っていることを検証します。世界的に通用するレベルの高度で国際的な知識を問います。また、金融工学をはじめとした学際領域の知識についても確認します。
4. Extraction (情報抽出): 与えられたテキストやデータセットから重要な情報を抽出し、分析する技術に関する質問が該当します。文書の要約、データの解釈、テキスト内の概念の評価など、情報を効果的に処理するスキルが求められます。
5. Reasoning (論理的思考): 論理的推理、問題解決、読解力を要する質問がこのカテゴリに分類されます。論理問題の解答、文章読解、エビデンスに基づく意見形成など、論理的な思考能力を鍛える質問が含まれます。
6. Math (数学): 数学の概念や理論に基づいた問題解決に関する質問を扱います。基本的な算術から高度な数学および金融工学まで、方程式の解法、幾何学的な問題、比率や割合、確率過程など、金融に関連する数学に関する幅広いトピックが含まれます。
7. Coding (コーディング): プログラミング言語、アルゴリズム、データ構造、ソフトウェア開発に関する質問が含まれます。コーディング技術、アプリケーションの開発、データサイエンス、機械学習の基本から応用まで、金融に関連するプログラミングに関連する広範囲の知識とスキルに関する相談に応じます。
8. Idea (アイデア): アイデア生成、創造性の促進、目標達成に向けた概念策定に関する相談を扱います。このカテゴリでは、新しいプロジェクトや取り組みのための革新的なアイデアの創出、目標設定の方法、実現可能な行動計画の立案に関するアドバイスを提供します。金融における取引戦略や投資家に対する施策など、思考の枠を広げ、創造的な解決策を探求する過程での支援を重視します。
9. Translation (翻訳): 文章の翻訳、内容の再表現に関する質問が含まれます。このカテゴリでは、異なる言語間での正確な情報伝達、テキストの言い換えや要約、特定の聴衆や文脈に合わせた内容の適応に関する相談に応じます。言語学の深い理解、文化的背景の認識、金融における適切な語彙使用、コミュニケーションの効果性を高める技術が求められます。
10. Ethics (倫理): 個人や集団の行動における倫理

的、道徳的問題の考察に関する質問を扱います。このカテゴリでは、道徳的ジレンマの解析、倫理的な判断の基準、正義や責任の概念に基づく意思決定に関するガイダンスを提供します。金融における行動原則を適用し、多様な視点からの評価を通じて、道徳的な価値観に沿った行動を促します。

11. Trustworthiness (信頼性): 信頼性、誠実さ、公正さに関する質問が含まれます。金融市場においては、不確かな情報を提供してしまった場合に、その波及効果などから多大な損失を招く可能性があります。このカテゴリでは、情報の信頼性が重要である局面において、間違った情報を流布しないかどうかについて確認します。また、情報の真偽を判断するための方法や、信頼性の高い情報源の選定に関するアドバイスを提供できるかについて検証を行います。
12. ESGs (環境・社会・ガバナンス): 環境、社会、ガバナンスに関する質問が含まれます。企業の持続可能性、社会的責任、倫理的な投資、ESG 評価など、環境・社会・ガバナンスに関する幅広いトピックについて相談に応じます。持続可能な社会の実現に向けた取り組みや、企業の社会的責任に関する情報を提供します。

これらの定義に基づき、各 30 問、合計 360 問を構築した。また、数学の問題については、LLM-as-a-judge で正しい評価ができるようにするために、参照回答を GPT-4o を活用しながら人手で作成した。問題の構築にあたっては、Anthropic Claude 3 Opus と few-shot prompt を用いてターンの問題を生成し[1)]、人手でスクリーニングを行うことで作成した。下記に構築したベンチマークの例題を示す。なお、すべての問題は前述の URL から確認可能である。

Roleplay の例題

**Turn 1:** あなたは資産運用アドバイザーです。クライアントから「老後資金を準備するために、長期的な資産運用を始めたいと思います。おすすめのポートフォリオを教えてください。」という相談を受けました。クライアントの年齢やリスク許容度を確認し、適切な資産配分を提案してください。

**Turn 2:** クライアントから「提案されたポートフォリオを維持するために、どのくらいの頻度でリバランスを行うべきでしょうか？また、リバランスの重要性を教えてください。」という質問がありました。リバランスの必要性と適切な頻度について説明してください。

これらの例題に対して、それぞれの LLM が応答を生成し、その生成結果を GPT-4o-mini により 10 段階評価を行うことで、ベンチマークの評価を行う。この LLM-as-a-judge における評価は、MT-bench[28] における評価と同様の手法を用いたが、以下の点に関しては、MT-bench の問題点を構築するために改善を行った。

- 最大生成 token 数の増加 (1024→ 4096)
- 対話テンプレートの日本語対応
- 評価テンプレートの変更
- EOS token の対応
- 1 ターン目の回答時に LLM が勝手に 2 ターン目の問題を捏造し、そこに勝手に答えた場合にその捏造された 2 ターン目の回答を 2 ターン目の評価対象にしてしまうインジェクションの防止
- LLM からの回答がなかった場合に、評価を行う LLM(GPT-4 など) が勝手に回答を書いて自分で評価する結果、高い評価値になってしまう問題があり、その対策として、回答がなければ 0 点と判定するルーティンの追加
- 質問と同じ言語で回答しない場合には 0 点とするルーティンの追加 (翻訳を除く)

## 3 ベンチマーク実験とその結果・考察

前章で構築したベンチマークを用いて、複数のモデルに対して評価を行った。ここでは、Anthropic 社の Claude シリーズの LLM、OpenAI 社の GPT シリーズ、Preferred Networks 社の Plamo、NVIDIA 社の Nemotron を対象に、ベンチマークを実施した。これらのベンチマークは、API を通じて計測されたが、pfmt-bench-fin-ja は Huggingface 等で構築されているモデルに対してもローカル環境で計測可能である。

評価結果を表 1 に示す。また、図 1 には、各モデルの評価結果をタスクレーダーで示す。

これらの結果から、数学やコーディング、抽出タスクで大きな差が出ていることが分かった。特に、GPT-4o は数学タスクで顕著な結果を示しており、

1) ライセンス上、この問題を学習に使用することはできない。

**表 1**　各モデルの pfmt-bench-fin-ja による評価値。

| Model/Metric | Avg. | Writ. | Role. | Kno. | Ext. | Reas. | Math | Cod. | Idea | Tran. | Eth. | Trust. | ESGs |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| claude-3-5-sonnet-20240620 | 9.07 | 9.38 | 9.38 | 8.93 | 8.93 | 9.35 | 5.52 | 9.57 | 9.63 | 9.62 | 9.65 | 9.47 | 9.45 |
| gpt-4o | 8.99 | 9.45 | 9.43 | 8.78 | 7.32 | 9.30 | 7.07 | 9.33 | 9.70 | 9.40 | 9.63 | 9.43 | 9.07 |
| claude-3-opus-20240229 | 8.70 | 9.23 | 9.18 | 9.45 | 6.78 | 8.97 | 4.35 | 9.72 | 9.45 | 9.52 | 9.25 | 9.30 | 9.22 |
| claude-3-haiku-20240307 | 8.55 | 9.02 | 8.93 | 8.53 | 7.70 | 8.98 | 3.97 | 9.57 | 9.25 | 9.30 | 9.08 | 9.17 | 9.12 |
| claude-3-sonnet-20240229 | 8.41 | 9.18 | 8.95 | 9.12 | 6.38 | 9.02 | 3.00 | 9.27 | 9.08 | 9.45 | 9.08 | 9.20 | 9.17 |
| gpt-4 | 8.37 | 8.42 | 8.70 | 8.62 | 7.17 | 8.80 | 4.55 | 8.95 | 9.07 | 9.33 | 8.90 | 9.12 | 8.78 |
| plamo-1.0-prime-beta | 8.33 | 9.33 | 9.12 | 9.12 | 6.45 | 9.00 | 3.25 | 8.77 | 9.40 | 7.85 | 9.32 | 9.22 | 9.08 |
| claude-2.1 | 8.20 | 8.42 | 8.17 | 8.45 | 8.17 | 8.62 | 4.12 | 8.82 | 8.38 | 9.25 | 8.42 | 8.80 | 8.83 |
| nvidia/nemotron-4-340b-instruct | 8.07 | 9.35 | 9.30 | 8.77 | 5.30 | 9.20 | 2.45 | 5.77 | 9.53 | 9.38 | 9.37 | 9.37 | 9.07 |
| gpt-35-turbo | 8.04 | 8.78 | 8.50 | 8.57 | 6.82 | 8.68 | 3.27 | 9.00 | 8.82 | 8.28 | 8.73 | 8.87 | 8.12 |

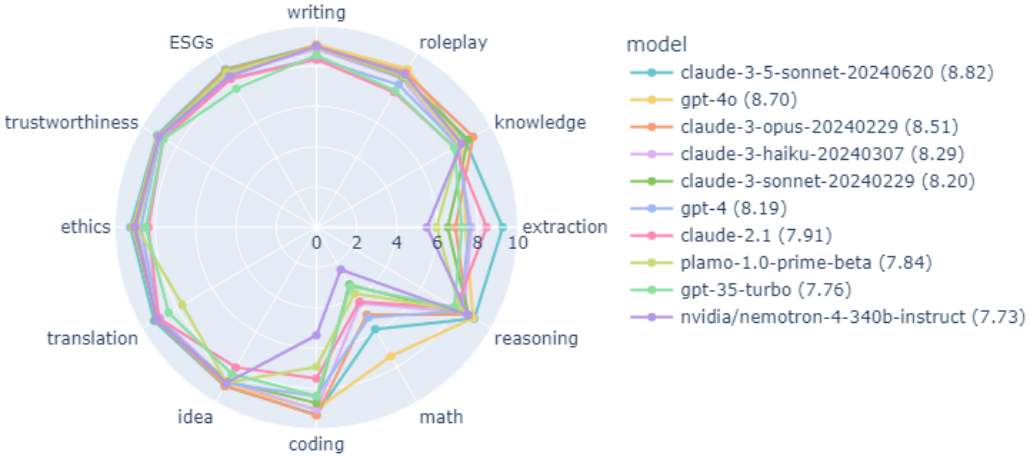


**図 1**　pfmt-bench-fin-ja のさまざまなモデルへの評価結果のタスクレーダー

claude-3-5-sonnet-20240620 は抽出タスクで高い評価を受けている。

本ベンチマークの有効性を充分に検討することは現状では難しく、今後の課題ではあるものの今回の結果から明らかであることもある。まず、gpt-35-turbo、gpt-4、gpt-4o の間の関係性を見ると、概ねすべてのタスクにおいて、そのスコアの順序性が維持されており、一般に議論される性能の順序性と一致している。また、claude-2.1, claude-3, claude-3.5 においても概ね同様の傾向性が確認できる。一方で、claude-3-haiku と claude-3-sonnet においては、Anthropic 社の公開する性能比較[2)]の順とは異なる順番となっており、ベンチマークスコアが近いモデル間での比較の有効性については疑問がのこる。claude-3-haiku と claude-3-sonnet の順序性に関しては、gpt-4o-mini で評価するのではなく、gpt-4o を LLM-as-a-judge として使用した場合においても同様の結果が得られることが確認できたことから、実際に、金融分野における性能が逆転している可能性も示唆される。

2)　https://www.anthropic.com/news/claude-3-family

## 4　まとめ

本研究では、pfmt-bench-fin-ja という、金融分野に特化した日本語の複数ターンの生成ベンチマークを構築した。このベンチマークは、MT-bench を参考にし、金融分野に特化させるために新たなカテゴリを追加し、カテゴリーの基準を定義したうえで、半自動的に複数ターンの質問セットを 360 問構築した。この質問セットに対して、複数のモデルの応答を取得し、その応答結果を GPT-4o-mini を用いて LLM-as-a-judge による評価を行うことで、10 段階評価による生成の良さを示す数値を取得できるベンチマークを構築した。そのうえで、このベンチマークを用いて、複数のモデルに対して評価を行い、その結果を示した。今後の課題として、このベンチマークの有効性をさらに検証する必要性などが挙げられる。

## Declarations

著者らは、pfnet/plamo-100b の開発元である、株式会社 Preferred Networks/Elements に所属しているが、本研究での実験においては、他のモデルと公平な評価を行っており、透明性の確保のために、ベンチマークの計測コードを公開している

# 参考文献


[1] OpenAI. ChatGPT, 2023. https://openai.com/blog/chatgpt/.

[2] OpenAI. GPT-4 Technical Report, 2023.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In **Advances in Neural Information Processing Systems**, Vol. 30, pp. 5999–6009, 2017.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics**, pp. 4171–4186. Association for Computational Linguistics, 2019.

[5] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving Language Understanding by Generative Pre-Training, 2018. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

[6] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners, 2019. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

[7] Tom Brown, Benjamin Mann, et al. Language Models are Few-Shot Learners. **Advances in Neural Information Processing Systems**, Vol. 33, pp. 1877–1901, 2020.

[8] Google. Bard, 2023. https://bard.google.com/.

[9] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and Efficient Foundation Language Models. **arXiv**, 2023. https://arxiv.org/abs/2302.13971.

[10] Hugo Touvron, Louis Martin, et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. **arXiv**, 2023. https://arxiv.org/abs/2307.09288v2.

[11] Databricks. Dolly, 2023. https://github.com/databrickslabs/dolly.

[12] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. **arXiv**, 2022. https://arxiv.org/abs/2211.05100.

[13] Vicuna. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality, 2023. https://vicuna.lmsys.org/.

[14] Aakanksha Chowdhery, Sharan Narang, et al. PaLM: Scaling Language Modeling with Pathways. **arXiv**, 2022. https://arxiv.org/abs/2204.02311v5.

[15] Rohan Anil, Andrew M. Dai, et al. PaLM 2 Technical Report. **arXiv**, 2023. https://arxiv.org/abs/2305.10403v3.

[16] Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, et al. A framework for few-shot language model evaluation, 2021. https://github.com/EleutherAI/lm-evaluation-harness.

[17] Marc Eulerich, Aida Sanatizadeh, Hamid Vakilzadeh, and David A. Wood. Is it All Hype? ChatGPT's Performance and Disruptive Potential in the Accounting and Auditing Industries. **SSRN Electronic Journal**, 2023. https://papers.ssrn.com/abstract=4452175.

[18] Harsha Nori, Nicholas King, Scott Mayer Mckinney, Dean Carignan, and Eric Horvitz. Capabilities of GPT-4 on Medical Challenge Problems. **arXiv**, 2023. https://arxiv.org/abs/2303.13375v2.

[19] Kwan Yuen Iu and Vanessa Man-Yi Wong. ChatGPT by OpenAI: The End of Litigation Lawyers? **SSRN Electronic Journal**, 2023. https://papers.ssrn.com/abstract=4339839.

[20] Jonathan H. Choi, Kristin E. Hickman, Amy Monahan, and Daniel B. Schwarcz. ChatGPT Goes to Law School. **SSRN Electronic Journal**, 2023. https://papers.ssrn.com/abstract=4335905.

[21] Shijie Wu, Ozan ˙ Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. BloombergGPT: A Large Language Model for Finance. **arXiv**, 2023. https://arxiv.org/abs/2303.17564v2.

[22] Pedram Babaei William Todt, Ramtin Babaei. Fin-LLAMA: Efficient Finetuning of Quantized LLMs for Finance, 2023. https://github.com/Bavest/fin-llama.

[23] Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. FinGPT: Open-Source Financial Large Language Models. **arXiv**, 2023. https://arxiv.org/abs/2306.06031.

[24] Boyu Zhang, Hongyang Yang, and Xiao-Yang Liu. Instruct-FinGPT: Financial Sentiment Analysis by Instruction Tuning of General-Purpose Large Language Models. **arXiv**, 2023. https://arxiv.org/abs/2306.12659.

[25] Kenshin Abe, Kaizaburo Chubachi, Yasuhiro Fujita, Yuta Hirokawa, Kentaro Imajo, Toshiki Kataoka, Hiroyoshi Komatsu, Hiroaki Mikami, Tsuguo Mogami, Shogo Murai, et al. Plamo-100b: A ground-up language model designed for japanese proficiency. **arXiv preprint arXiv:2410.07563**, 2024.

[26] StabilityAI. JP Language Model Evaluation Harness, 2023. https://github.com/Stability-AI/lm-evaluation-harness/tree/jp-stable.

[27] Masanori Hirano. Construction of a Japanese Financial Benchmark for Large Language Models. In **Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP**, pp. 1–9, 2024.

[28] Lianmin Zheng, Wei-Lin Chiang, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, **Advances in Neural Information Processing Systems**, Vol. 36, pp. 46595–46623. Curran Associates, Inc., 2023.

[29] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. **arXiv preprint arXiv:2212.10560**, 2022.