

FinePersonas を用いた日本語指示データの合成

言語処理学会第31回年次大会 併設ワークショップ JLR2025
日本語言語資源の構築と利用性の向上

太田 晋 (東京工科大学)

2025-03-14

目次

- 背景
- 先行研究
- 目的
- 実験
 - 実験1: 英語ペルソナデータセットから日本語指示データを合成
 - 実験2: (LLM-as-a-Judge による合成データの評価)
 - 実験3: 日本語 Web テキストから日・英ペルソナを抽出
- 考察
- まとめと今後の課題
- 参考文献
- 付録
 - 実験2: LLM-as-a-Judge による合成データの評価

背景

- チャットボット等の普及
 - 日本語言語モデル構築のため、高品質かつ多様な指示データセットの構築が求められている
- 合成データ
 - 人間によって直接生成されたデータではなく、モデルやアルゴリズムによって生成されたデータ
 - 高品質な合成データがモデル性能の向上に寄与することが報告されている
(Yuanzhi Li et al., 2023)
 - データ量の拡大は比較的容易
 - 多様性の確保が課題

合成データの多様性をスケールアップするアプローチ

- サンプリング: 固定のプロンプトを用い、高い温度パラメータで複数回推論
 - Magpie (Zhangchen Xu et al., 2024), Evol-Instruct (Can Xu et al., 2024) など
 - 問題点: 温度を上げすぎるとハルシネーションや文が破綻する可能性が高まる
- 多様なプロンプト (Xin Chan et al., 2024)
 - インスタンス駆動
 - 例: Wikipedia の記事から Q&A を生成
 - キーポイント駆動
 - 例: 学習指導要領に含まれる用語 (例: 三角関数) から数学の問題を生成
 - ペルソナ駆動
 - 例: ペルソナ (例: 運送会社のドライバー) から数学の問題を生成
- サンプリングと多様なプロンプトの組み合わせも可能

先行研究: Persona-Hub (Xin Chan et al., 2024)

- ペルソナ: 職業・興味・性格・経験・知識などの属性情報
- Text-to-Persona: 大規模な Web テキストからペルソナを抽出
- Persona-to-Text: ペルソナから大規模な合成データを生成

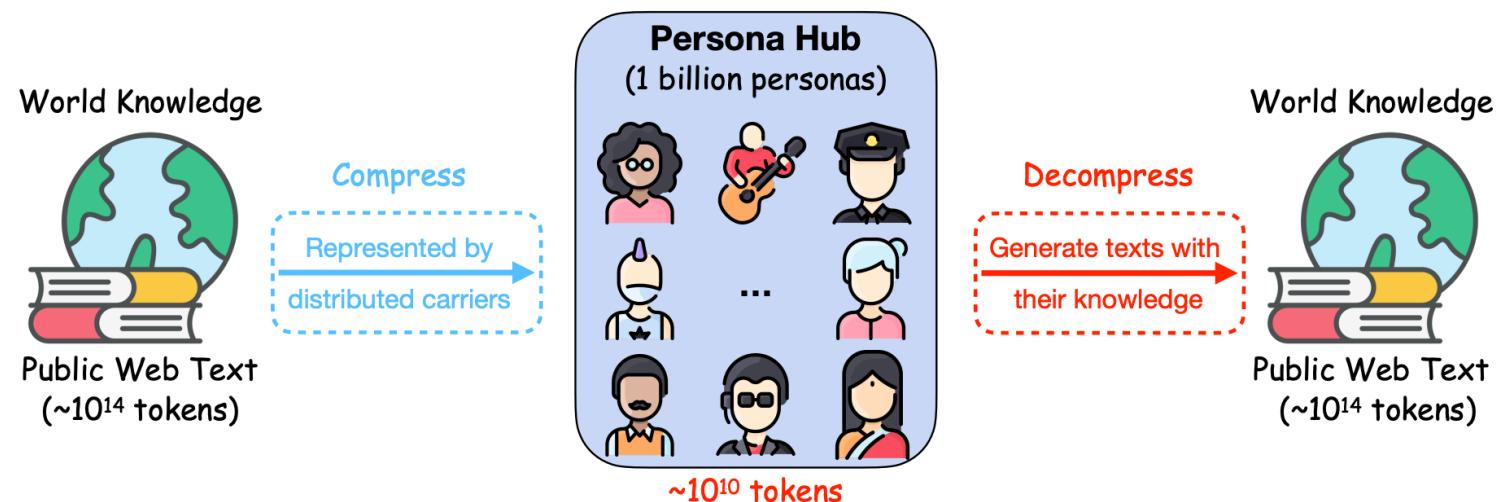


Figure 2: From a compression perspective (Delétang et al., 2023; Ge et al., 2024), Persona Hub ($\sim 10^{10}$ tokens) can be seen as the compressed form of world knowledge (public web text for training LLMs, $\sim 10^{14}$ tokens) into distributed carriers. On the other hand, the public web text can be seen as the decompressed content created by these personas with their knowledge and experiences.

先行研究: Persona-Hub (Xin Chan et al., 2024)

- データ合成プロンプト: Create {data} with {persona}
- 10億件のペルソナデータセット (ただし2025-03-09時点では未公開)

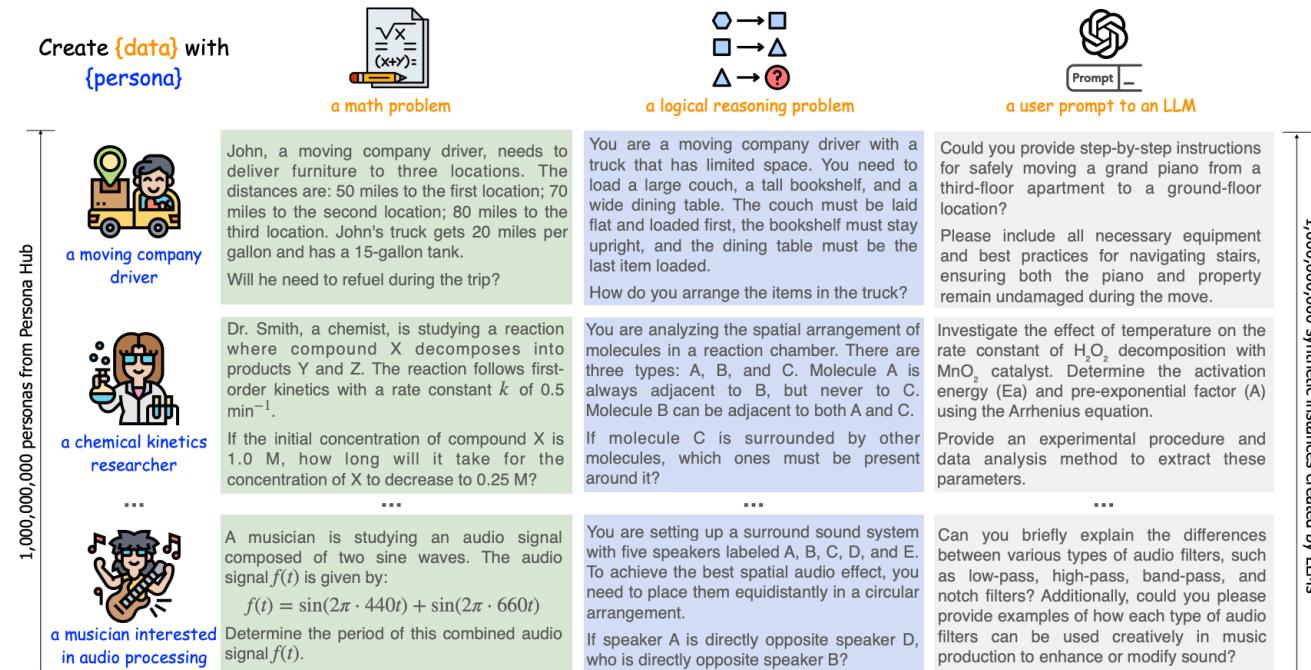


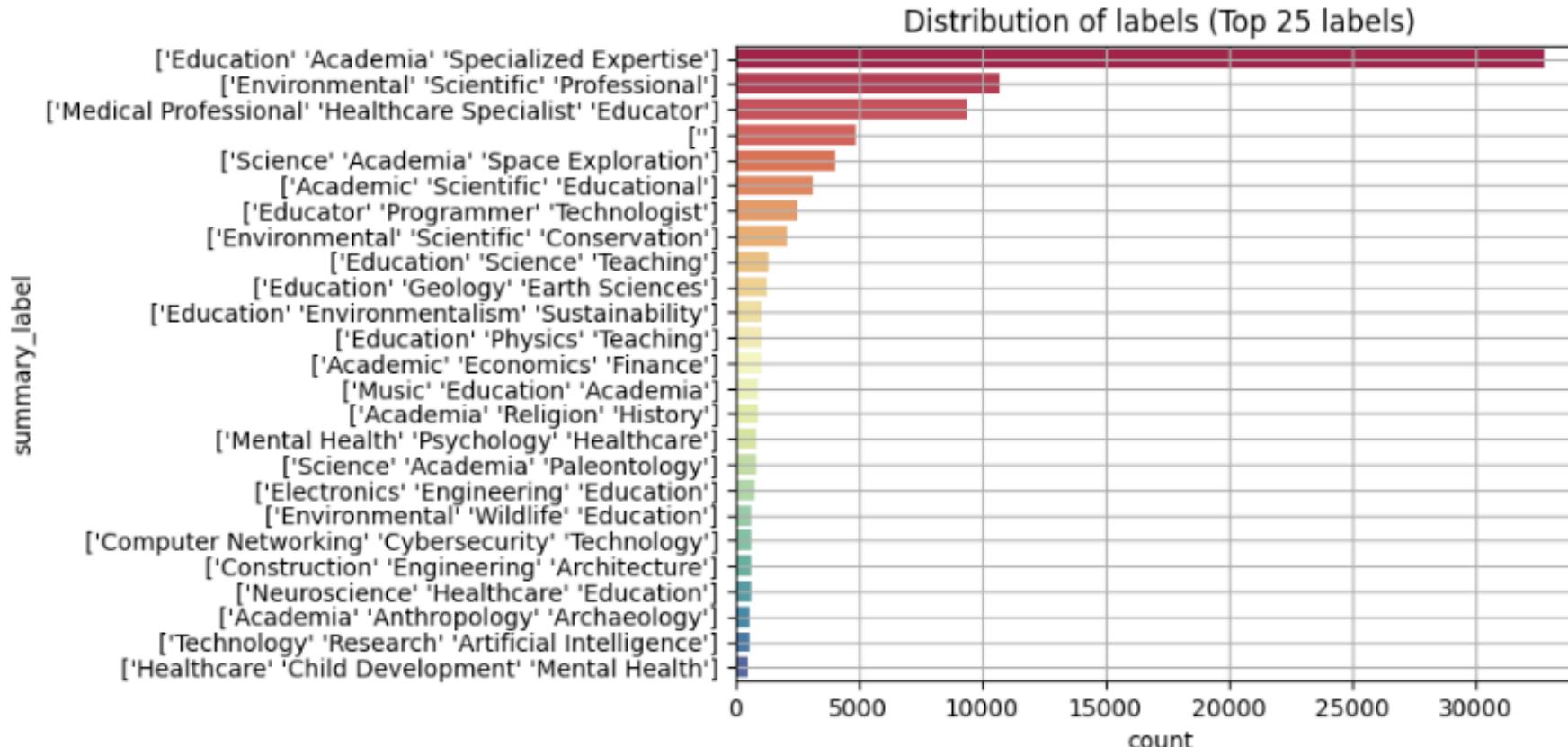
Figure 1: Personas can work with a wide range of data synthesis prompts (e.g., “create a math problem”) to guide an LLM to synthesize data with corresponding perspectives. The 1 billion personas in Persona Hub can facilitate various data synthesis scenarios at a billion scale.

先行研究: FinePersonas (Argilla, 2024)

- Argilla が2024年9月にリリースしたオープンなペルソナデータセット(英語)
- Persona-Hub と同じレシピに従い、高品質な教育関連 Web ページデータセット FineWeb-Edu (Guilherme Penedo et al., 2024) から2100万件のペルソナを抽出
- ライセンス: Llama 3.1 Community License Agreement
- ペルソナ例
 - A network engineer with a focus on routing protocols and preparing for Cisco certification exams, particularly CCNA.

先行研究: FinePersonas (Argilla, 2024)

- 図: FinePersonas のサブセット(100kペルソナ)をクラスタリングしてラベル付け
- 注意点: 教育・研究・科学技術分野に重点 (抽出元のデータセットが教育関連)



目的

- 背景・先行研究の問題点
 - 合成データの多様性確保に課題
 - FinePersonas は英語ペルソナデータセットのため、日本語合成データの生成に利用できるかどうか明らかではない
- 本研究の目的
 - 職業・興味・性格・経験・知識などの属性情報(=ペルソナ)を活用し、多様な日本語指示データを合成する
- 実験
 - 実験1: 英語ペルソナデータセットから日本語指示データを合成
 - 実験2: (合成された日本語指示データの品質を LLM-as-a-Judge で評価)
 - 実験3: 日本語 Web テキストから日・英ペルソナを抽出

実験1: FinePersonas による日本語指示データの合成

- ペルソナデータセット (Argilla, 2024)
 - argilla/FinePersonas-v0.1
- 合成データを生成する言語モデル (Xinqi Chen et al., 2024), (OpenAI, 2025)
 - rinna/deepseek-r1-distill-qwen2.5-bakeneko-32b (rinna-r1-32b)
 - gpt-4o-mini (比較のためのベースライン)
- 合成プロンプト
 - Persona-Hub (Xin Chan et al., 2024) のプロンプトテンプレートの一部を修正して様々な種類・難易度の日本語指示データを生成出来るよう拡張
- 計算資源
 - GPU: NVIDIA H100 80GB
 - 生成時間: rinna-r1-32b を vllm で推論, GPU 1枚で1時間当たり 3600 件程度

実験1: 問題を合成するプロンプト

- 例: ペルソナに関する小・中学生レベルの数学の問題を作成
- 下記マーク部分は、問題の種類・ペルソナ・難易度・対象者・言語を指定
- プロンプト全体を日本語で記述せず、問題: と言語指定のみ日本語で記述

Create a math problem related to the following persona:

{persona}

Note:

1. The math problem should be short, easy and involve basic mathematical skills and knowledge. Any average grade school student can solve it correctly.
2. You should make full use of the persona description to create the math problem to ensure that the math problem is unique and specific to the persona.
3. Your response should always start with "問題:". Your response should not include a solution to the created math problem.
4. 簡潔に日本語で回答してください。

実験1: 解答を合成するプロンプト

- 例: 平均的な小中学生が理解できる解答を生成
- 下記マーク部分は、問題・対象者・言語を指定
- プロンプト全体を日本語で記述せず、 解答: と言語指定のみ日本語で記述

Create a solution to the following problem:

{problem}

Note:

1. The solution should be concise, clean and easy to understand. Any average grade school student can understand it correctly.
2. You should make full use of the problem description to create the solution to ensure that the solution is unique and specific to the problem.
3. Your response should always start with "解答:". Your response should only include the solution to the problem.
4. 簡潔に日本語で回答してください。

実験1 結果: 生成された合成データの言語判定

- フォーマットエラー
 - `^(<think>)?(.*)</think>.*問題:(.*)$` という正規表現に合致しないレコードをエラーとする (非推論モデルは `^.*問題:(.*)$`)
- 日本語判定: ランダムに 100 件サンプリングし、人手により日本語かどうか判定

言語モデル	生成件数	フォーマットエラー	日本語判定
rinna-r1-32b	2,718,336	0.056%	100%
gpt-4o-mini	100	0.000%	100%

実験1 結果: 生成された合成データ例1

- 解答を生成するために必要十分な情報が問題に含まれていない例

	内容
ペルソナ	An academic writer or researcher focused on scientific inquiry and methodology, likely with a background in statistics and research design.
rinna-r1-32b	問題: 研究者が20人の被験者を対象にアンケート調査を行い、回答者の年齢層を以下のように分類しました。各年齢層の人数を確認し、被験者の平均年齢を計算してください。 \n- 10代: 3人\n- 20代: 8人\n- 30代: 5人\n- 40代: 4人 解答: 各年齢層の代表的な年齢（例: 10代→15歳、20代→25歳、30代→35歳、40代→45歳）を仮定し、計算します。 \n1. 各年齢層の年齢合計を計算: \n- 10代: 3人 × 15歳 = 45歳 \n- 20代: 8人 × 25歳 = 200歳 \n- 30代: 5人 × 35歳 = 175歳 \n- 40代: 4人 × 45歳 = 180歳 \n2. 総年齢 = 45 + 200 + 175 + 180 = 600歳 \n3. 平均年齢 = 600歳 ÷ 20人 = 30歳 \n答え: 被験者の平均年齢は30歳です。
gpt-4o-mini	問題: ある研究者が新しい薬の効果を調査するために、50人の参加者をランダムに2つのグループに分けました。1つのグループには薬を与え、もう1つのグループにはプラセボを与えました。もし薬を受け取ったグループに25人、プラセボを受け取ったグループに25人がいる場合、全体の参加者の何パーセントが薬を受け取ったグループに属していますか？ 解答: 薬を受け取ったグループは25人です。全体の参加者は50人です。薬を受け取ったグループの割合は次のように計算します。 \n(薬を受け取った人数 ÷ 全体の参加者数) × 100 = (25 ÷ 50) × 100 = 50% したがって、全体の参加者の50%が薬を受け取ったグループに属しています。

実験1 結果: 生成された合成データ例2

- 推論モデルと通常モデルで生成された合成データの品質に差が出る例

	内容
ペルソナ	A computer science instructor or a C++ programming language expert focused on object-oriented programming concepts and advanced language features.
rinna-r1-32b	問題: C++のオブジェクト指向プログラミングで、あるクラスのオブジェクトが8バイトのメモリを消費する場合、このオブジェクトを100個インスタンス化した際の合計メモリ消費量を求めてください（単位はバイト）。 解答: $8\text{バイト} \times 100\text{個} = 800\text{バイト}$
gpt-4o-mini	問題: あるコンピュータサイエンスの授業で、C++のオブジェクト指向プログラミングの概念を教えるために、講師は8人の学生をグループに分けることにしました。各グループには3人の学生が必要です。講師は何グループを作ることができますか？ 解答: 講師は8人の学生を3人ずつのグループに分けることができますが、8人を3人のグループに分けると、2人の学生が余ります。したがって、作れるグループは2つです。

実験1 結果: 合成データ例2の推論過程

内容	
ペルソナ	A computer science instructor or a C++ programming language expert focused on object-oriented programming concepts and advanced language features.
推論	まず、ユーザーのリクエストを正確に理解する必要があります。彼らは、コンピュータサイエンスの講師またはC++の専門家をテーマにした、簡単な数学問題を求めていています。その問題は、オブジェクト指向プログラミングやC++の高度な機能に関連する必要があります。 次に、問題を短く簡潔に作成する必要があります。平均的な小学校の生徒でも解けるような基本的な数学スキルを必要とする内容にします。例えば、オブジェクトの作成数やメモリ消費量を計算するような問題が考えられます。 C++のコンテキストを活用するために、オブジェクトのインスタンス化やメモリ使用量をテーマにします。例えば、あるオブジェクトが特定のメモリサイズを消費し、そのオブジェクトを複数作成した場合の合計メモリを計算する問題です。これなら、オブジェクト指向の概念と基本の算数を組み合わせています。 問題文を日本語で簡潔に表現し、具体的な数字を設定します。例えば、各オブジェクトが8バイト消費し、100個作成した場合の合計メモリを求める問題です。このようにすることで、C++の専門家を想定しつつも、計算は簡単な掛け算で済みます。 最後に、ユーザーの指示に従い、問題文だけを提供し、解答は含めません。また、専門的な知識を必要としないように注意し、一般的な常識と数学の基礎知識で解けるようにします。
回答	C++のオブジェクト指向プログラミングで、あるクラスのオブジェクトが8バイトのメモリを消費する場合、このオブジェクトを100個インスタンス化した際の合計メモリ消費量を求めてください（単位はバイト）。

実験2: LLM-as-a-Judge による合成データの評価

- 発表時間の都合で詳細は付録に掲載し、ここでは結果のみを紹介
- 評価に使用した言語モデル
 - deepseek-ai/DeepSeek-R1 (R1) (DeepSeek AI, 2025)
- 温度 0.6 で 8回判定して平均値を算出
- 評価項目
 - helpfulness, relevance, accuracy, depth, creativity, and level of detail
- rinna-r1-32b で合成したデータの品質が gpt-4o-mini よりも高いことを確認

言語モデル	勝率 (平均値)
rinna-r1-32b	78.0%
gpt-4o-mini	22.0%

実験3: 日本語 Web テキストからペルソナを抽出

- 日本語 Web データセット (Yuichi Tateno, 2025)
 - hotchpotch/fineweb-2-edu-japanese
- ペルソナ抽出する言語モデル (Xinqi Chen et al., 2024), (OpenAI, 2025)
 - rinna/deepseek-r1-distill-qwen2.5-bakeneko-32b (rinna-r1-32b)
 - gpt-4o-mini (比較のためのベースライン)
- ペルソナ抽出プロンプト (→次ページ)
 - Persona-Hub (Xin Chan et al., 2024) のプロンプトテンプレートの一部を修正して日本語と英語のペルソナを抽出

実験3: ペルソナ抽出プロンプト

- このテキストを書きそうな人物やこのテキストに興味がありそうな人物を言語モデルに予測させる
- 3の言語指定部分を日本語と英語で行い、日本語ペルソナと英語ペルソナを抽出

You are an expert in analyzing the text content and assigning finding the general type of persona that could be associated with (中略)

The following are starts of persona definitions:

A machine learning researcher...

A pediatric nurse whose...

An urban planner focused on...

What is the likely profession, interest, or role of the person who would write or be interested in this text?

{text}

Note:

- Your response should always start with "ペルソナ:".
- Your response should be one sentence. You should not include any notes or translations.
- 簡潔に日本語で回答してください。

実験3 結果: 生成されたペルソナの評価

- 生成されたペルソナを実験1, 2と同じ方法で評価
- 勝率はそれぞれのベースラインとの比較

言語モデル	言語	生成件数	フォーマットエラー	言語判定	勝率(平均値)
rinna-r1-32b	日	100	0.0%	100.0%	90.5%
gpt-4o-mini	日	100	0.0%	100.0%	9.5%
rinna-r1-32b	英	100	0.0%	99.0%	87.3%
gpt-4o-mini	英	100	0.0%	100.0%	12.7%

実験3 結果: 抽出したペルソナ例1

	内容
テキスト	琉球朝日放送さんの取材を受けて、韓国向けのプロモーションの詳細な内容が放送されました。今回の取材では、ちょうどホスピタブルでパワーブロガーさんを招聘した時に密着していただきました。2016年3月2日に放送されました。 個人旅行客の誘導で最も重要なのは何と言ってもグルメ情報です。地元でも有名なステーキとしゃぶしゃぶの店舗を訪問しました またさらに、その3か月前にブロガーさんの招聘を行ったタイムズレンタカーさんのその後の追いかけ取材も行っていただきました。取材時の3か月ほど前にパワーブロガーさんたちとタイムズレンタカーに乗ってドライブすることに魅力を伝えてみました。(略)
rinna-r1-32b	デジタルマーケティング担当者兼観光プロモーター、ブロガーを活用したプロモーション戦略立案・実施を専門とし、地域観光振興に必要なノウハウ提供を得意とする。
gpt-4o-mini	観光業界のマーケティング担当者で、特にグルメプロモーションやインフルエンサーとの連携に関心がある人物。
rinna-r1-32b	A travel marketing professional specializing in promotional strategies for tourist destinations, particularly leveraging influencer partnerships and content creation to drive international visitor engagement.
gpt-4o-mini	A marketing professional specializing in tourism and hospitality, focused on promoting local attractions and culinary experiences to attract international travelers.

実験3 結果: 抽出したペルソナ例2

	内容
テキスト	初めての方へ\n訪問看護を中心とした介護事業所での業務です。ご利用者様.....\n\n地域の皆様に対して質の高い医療看護介護を提供し、病気の急性期から慢性期の在宅介護まで、一貫して安心できる医療介護を受けていただけるように、地域への貢献・密着を目指している病院です!\n【教育制度充実★保育所・寮有】\n地域を代表する総合病院で、学びながら勤務することができます!\n24時間託児・病児保育・夏季冬季長期休暇中の学童保育など完備しており、子育て中の方も活躍できる環境が整っております♪\n様々な科目があり、急性期~在宅まで様々な事が学ぶことが出来ます!\n医師会が運営する病院です☆\n給与も高めで賞与は4.7ヶ月あり (略)
rinna-r1-32b	病院の人事・採用担当者であり、看護師や介護士を対象とした求人広告を制作・発信する責任者。
gpt-4o-mini	地域医療に従事する看護師で、患者の在宅介護や教育制度の充実に关心を持つ人。
rinna-r1-32b	A healthcare recruitment officer or nurse recruiter at a comprehensive hospital emphasizing long-term care, education, and family-friendly benefits.
gpt-4o-mini	A healthcare administrator focused on providing comprehensive nursing and caregiving services in a community hospital setting, emphasizing quality care, education, and support for staff.

考察

- 実験1, 2: 英語ペルソナから高品質な日本語指示データが合成可能であることを示した
 - 日本語だけでなく多言語への拡張の可能性を示唆
 - ペルソナの領域(例: C++の専門家)と、求める合成データの領域(例: 小・中学生向けの簡単な数学問題)がかけ離れている場合、不完全な応答や矛盾する応答を生成する可能性
 - それらを除外するために、フィルタリングやプロンプトによる自己改善が必要
 - 推論型モデルの優位性を示唆
 - 推論型モデルに最適化したプロンプトの検討が必要
- 実験3: 日本語 Web テキストからペルソナを抽出し、高品質な日・英ペルソナが抽出可能であることを示した
 - 抽出されたペルソナの重複削除・フィルタリングが必要
 - 抽出されたペルソナの分布を分析し多様性を評価

まとめと今後の課題

- 英語ペルソナから高品質な日本語指示データが合成可能であることを示し、多言語への拡張の可能性を示唆
- 日本語 Web テキストからペルソナを抽出し、高品質な日・英ペルソナが抽出可能であることを示した
- 本実験のソースコード: <https://github.com/susumuota/synth-persona>
- 今後の課題
 - 言語モデルの学習を行い、下流タスクの性能で合成データを評価
 - 生成した合成データを低成本でフィルタリングする手法の検討
 - 既存データセット(例: 新聞記事)を特定のペルソナ(例: 記事に関連する専門家)の視点でリライト(解説, Q&A など)
 - 複数ペルソナを用いた多様な対話データセットの構築

参考文献

- Lianmin Zheng et al., "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena", NeurIPS 2023, 2023.
https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html
- Can Xu et al., "WizardLM: Empowering Large Pre-Trained Language Models to Follow Complex Instructions", ICLR 2024, 2024. <https://openreview.net/forum?id=CfXh93NDgH>
- Zhangchen Xu et al., "Magpie: Alignment Data Synthesis from Scratch by Prompting Aligned LLMs with Nothing", ICLR 2025, 2025. <https://openreview.net/forum?id=Pnk7vMbznK>
- Yuanzhi Li et al., "Textbooks Are All You Need II: phi-1.5 technical report", arXiv preprint arXiv:2309.05463v1, 2023. <https://arxiv.org/abs/2309.05463v1>
- Xin Chan et al., "Scaling Synthetic Data Creation with 1,000,000,000 Personas", arXiv preprint arXiv:2406.20094v2, 2024. <https://arxiv.org/abs/2406.20094v2>
- Hao Chen et al., "On the Diversity of Synthetic Data and its Impact on Training Large Language Models", arXiv preprint arXiv:2410.15226v2, 2024.
<https://arxiv.org/abs/2410.15226>
- Guilherme Penedo et al., "The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale", arXiv preprint arXiv:2406.17557v2, 2024.
<https://arxiv.org/abs/2406.17557>
- DeepSeek-AI, "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning", arXiv preprint arXiv:2501.12948v1, 2025.
<https://arxiv.org/abs/2501.12948v1>
- Anton Lozhkov et al., "FineWeb-Edu: the Finest Collection of Educational Content", <https://huggingface.co/datasets/HuggingFaceFW/fineweb-edu>, アクセス日: 2025-03-09
- Argilla, "FinePersonas", <https://huggingface.co/datasets/argilla/FinePersonas-v0.1>, アクセス日: 2025-03-09
- Xinqi Chen et al., "rinna/deepseek-r1-distill-qwen2.5-bakeneko-32b", <https://huggingface.co/rinna/deepseek-r1-distill-qwen2.5-bakeneko-32b>, アクセス日: 2025-03-09

参考文献 (続き)

- OpenAI, "GPT-4o mini: advancing cost-efficient intelligence | OpenAI", <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>, アクセス日: 2025-03-09
- DeepSeek-AI, "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning", <https://huggingface.co/deepseek-ai/DeepSeek-R1>, アクセス日: 2025-03-09
- Yuichi Tateno, "FineWeb2 Edu Japanese", <https://huggingface.co/datasets/hotchpotch/fineweb-2-edu-japanese>, アクセス日: 2025-03-09
- Kan Hatakeyama, "大規模言語モデルを開発するにあたっての事前・事後学習の戦略メモー特に合成データについてー", <https://zenn.dev/matsuolab/articles/34036f017fae9e>, アクセス日: 2025-03-09
- Someya, "Tanuki-8B, 8x8B - 事後学習の軌跡", <https://zenn.dev/matsuolab/articles/62c75674190a41>, アクセス日: 2025-03-09
- Arata, "Tanuki-8BにMagpieを適用して日本語の合成対話データセットを作成する", https://zenn.dev/aratako_llm/articles/a5ae43fb2bfbb3, アクセス日: 2025-03-09
- Mitsuhashi, "プロンプト進化を用いた日本語選好データセットの構築", <https://zenn.dev/matsuolab/articles/10a1aa9d43e4fe>, アクセス日: 2025-03-09
- Kan Hatakeyama, "大規模言語モデル Tanuki-8x8B の紹介と開発経緯など", 9/10 松尾研 LLM 開発プロジェクト "Tanuki-8x8B" 開発成果報告会 Vol.1, https://www.docswell.com/s/matsuo-lab_llm/51R2L4-2024-9-10-Tanuki開発報告会-vol1, <https://www.youtube.com/watch?v=lcpXpX-r6ZY>, アクセス日: 2025-03-09
- Susumu Ota, "Persona-Hub による合成データ生成", 9/24 松尾研 LLM 開発プロジェクト "Tanuki-8x8B" 開発成果報告会 Vol. 3, https://www.docswell.com/s/matsuo-lab_llm/ZDNGR4-2024-9-24-Tanuki開発報告会-vol3, <https://www.youtube.com/watch?v=XAdc-OgLeOw>, アクセス日: 2025-03-09
- Susumu Ota, "「Persona-HubによるSFTデータ合成とLLM as a judgeによるDPOデータ合成」松尾研 LLM コミュニティ 勉強会シリーズ#5 (2025-02-02)", https://www.youtube.com/watch?v=Gwyn9_WZao, アクセス日: 2025-03-09
- Susumu Ota, "合成データハンズオン", <https://github.com/susumuota/synthetic-data-hands-on>, アクセス日: 2025-03-09
- Susumu Ota, "Synthetic Data Creation with Personas", <https://github.com/susumuota/synth-persona>, アクセス日: 2025-03-09
- [NEDO 採択プロジェクト] 多様な日本語能力の向上を目指した公開の基盤モデル開発, コードレポジトリ "synth_topic_multiturn.py", https://github.com/matsuolab/nedo_project_code/blob/team_hatakeyama_phase2/team_hatakeyama_phase2/ota/topic-hub/synth_topic_multiturn.py, アクセス日: 2025-03-09

付録

- 実験2: LLM-as-a-Judge による合成データの評価

実験2: LLM-as-a-Judge による合成データの評価

- 実験1 で生成した合成データの品質を LLM-as-a-Judge で評価 (Lianmin Zheng et al., 2023)
 - rinna-r1-32b の合成データ (2,718,336件) からランダムに100件サンプリング
 - gpt-4o-mini の合成データ100件 (ベースライン)
 - LLM-as-a-Judge でどちらが優れているか8回判定して平均値を算出
 - 評価に使用する言語モデル: deepseek-ai/DeepSeek-R1 (R1) (DeepSeek AI, 2025)
- 注意点
 - R1 蒸留モデル(rinna-r1-32b)で合成したデータを R1 で評価するので、自己強化バイアス(self-enhancement bias)がかかる可能性

実験2: LLM-as-a-Judge プロンプト

- ペアワイズ比較: 質問と2つの応答を提示し、どちらが優れているかまたは同点を判定 (Lianmin Zheng et al., 2023)

[System]

Please act as an impartial judge and evaluate the quality of the responses provided by two AI assistants to the user question displayed below. You should choose the assistant that follows the user's instructions and answers the user's question better. Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of their responses. Begin your evaluation by comparing the two responses and provide a short explanation. Avoid any position biases and ensure that the order in which the responses were presented does not influence your decision. Do not allow the length of the responses to influence your evaluation. Do not favor certain names of the assistants. Be as objective as possible. After providing your explanation, output your final verdict by strictly following this format: "[[A]]" if assistant A is better, "[[B]]" if assistant B is better, and "[[C]]" for a tie.

[User Question]

{question}

[The Start of Assistant A's Answer]

{answer_a}

[The End of Assistant A's Answer]

[The Start of Assistant B's Answer]

{answer_b}

[The End of Assistant B's Answer]

Figure 5: The default prompt for pairwise comparison.

実験2: LLM-as-a-Judge のバイアス除去

- 以下2つのバイアスに対する緩和策を実施し、一貫した結果のみを採用することで、バイアスを取り除いた評価を算出 (Lianmin Zheng et al., 2023)
- 位置バイアス (実験結果では `default` と表記)
 - 最初の回答を2つ目の回答よりも高く評価するバイアス (またはその逆)
 - 緩和策: プロンプト中の回答の順番を入れ替えて再度評価し、一貫性のある評価のみを採用
- 名前バイアス (実験結果では `rename` と表記)
 - 特定のアシスタント名(例: `Assistant A`)の回答を高く評価するバイアス
 - 緩和策: プロンプト中のアシスタントの名前を入れ替えて再度評価し、一貫性のある評価のみを採用

実験2: 結果: LLM-as-a-Judge による rinna-r1-32b と gpt-4o-mini の合成データの品質比較

- 比較対象
 - rinna-r1-32b で生成した合成データ (2,718,336件) から100件サンプリング
 - gpt-4o-mini で生成した合成データ100件 (ベースライン)
- 温度 0.6 で 8回判定して平均値を算出
- 位置バイアスと名前バイアスを取り除き、一貫した評価結果のみを採用
- 評価項目: helpfulness, relevance, accuracy, depth, creativity, and level of detail

言語モデル	勝率 (平均値)
rinna-r1-32b	78.0%
gpt-4o-mini	22.0%

実験2: 結果: DeepSeek-R1 の位置・名前バイアス評価

- (Lianmin Zheng, et al. 2023) の Table 2 に筆者が DeepSeek-R1 の結果を追加
- default: 位置バイアスの評価, rename: 名前バイアスの評価
- Consistency: 2つの回答の順番を入れ替えた場合に、裁判官が一貫した判定を下した割合(数値が高い方が高性能)
- DeepSeek-R1 は温度 0.6 で 8 回ジャッジした平均値. 位置バイアスでは GPT-4 と同程度、名前バイアスでは高性能

Judge	Prompt	Consistency	Biased toward first	Biased toward second	Error
Claude-v1	default	23.8%	75.0%	0.0%	1.2%
	rename	56.2%	11.2%	28.7%	3.8%
GPT-3.5	default	46.2%	50.0%	1.2%	2.5%
	rename	51.2%	38.8%	6.2%	3.8%
GPT-4	default	65.0%	30.0%	5.0%	0.0%
	rename	66.2%	28.7%	5.0%	0.0%
DeepSeek-R1	default	61.0%	36.4%	2.6%	0.0%
	rename	80.9%	17.1%	2.0%	0.0%

実験2: 考察

- rinna-r1-32b で合成したデータの品質が gpt-4o-mini よりも高いことが示された (78.0% vs. 22.0%)
 - ただし、R1 蒸留モデル (rinna-r1-32b) で合成したデータを R1 で評価したため、自己強化バイアス(self-enhancement bias)を含む可能性を考慮する必要がある。他の言語モデルでは、人間の評価と比べて 10% (GPT-4) から 25% (Claude-v1) 程度、自己の回答を高く評価する傾向があると報告されている (Lianmin Zheng et al., 2023)
 - 自己強化バイアスを排除するためには、DeepSeek-R1 以外の別の言語モデルを用いて LLM-as-a-Judge を行う必要がある
- DeepSeek-R1 を用いた LLM-as-a-Judge 実験では、位置バイアスの影響は GPT-4 と同程度の性能を示し、名前バイアスでは GPT-4 より高い性能を示した