Jellyfish: データ前処理のための LLM

張 皓辰 [1]　董 于洋 [2]　肖 川 [1]　小山田昌史 [2]
[1] 大阪大学情報科学研究科　[2] ＮＥＣ
[1]{chou.koushin,chuanx}@ist.osaka-u.ac.jp　[2]{dongyuyang,oyamada}@nec.com

概要

データ分析において不可欠なステップであるデータ前処理（DP）に対して，大規模言語モデル（LLM）の活用が注目を集めている．しかし，既存手法の多くは GPT の API に依存しており，プライバシーやコスト面での課題が残されている．本研究では，ローカル環境で実行可能な，Llama3-8B, mistral-7B などのパラメータ規模が小さい LLM に Instruction-tuning を実施し，代表的な DP タスクにおける性能を包括的に評価した．本実験により，全てのモデルの性能向上が確認され，GPT-3.5/4 に匹敵する処理能力を実証した．この取り組みは，ローカル環境で実行でき，プライバシーとコストを考慮しつつ高性能な DP を実現しうる LLM による実用的な解決策を提案するものである．モデルおよび学習データは https://huggingface.co/NECOUDBFM で公開されている．

1　はじめに

データ前処理（DP）は，データ分析パイプラインにおいて，生データを扱いやすく処理可能な形式へ変換する極めて重要なステップである．過去数十年にわたり，さまざまな DP タスクにおいて大きな進歩がみられたが，2021 年頃までは，エラー検出（ED）[1, 2]，データ補完（DI）[3, 4, 5]，スキーママッチング（SM）[6]，エンティティマッチング（EM）[7, 8] など，個別のタスクに特化した研究が中心であった．DP の汎用的な解決策を開発する上での主要な課題は，これらのタスクにおける対象（エラー，異常値，マッチデータなど）と，操作手法（検出，修正，アラインメントなど）がそれぞれ異なる点にある．

一方，大規模言語モデル（LLM）の登場により，より幅広い DP タスクに対応できる汎用解決策の開発が促進された．従来の DP 手法と比べ，LLM を用いた手法は自然言語生成能力，内部知識，推論能力，汎化能力に加え，ゼロショット [9] あるいはフューショット [10] 学習による適応性を備えている．そのため，パラメータチューニングといった人手による作業コストを削減でき，さらに可解釈性の向上も期待できる．

しかし，既存の LLM による DP 解決策 [11, 12, 13] は，GPT の API への依存によりデータ漏洩の懸念を引き起こしてきた [14]．また，GPT シリーズは高度に専門的なドメインのデータを扱う場合，追加学習のコストが高く，モデルのカスタマイズが困難である．これらの課題に対し，本研究では DP タスク向けのインストラクションデータ構築方法と，LLM のチューニング手法を探究する．

本研究の主要な問いは以下の三つである：

- ローカル LLM は Instruction-tuning により GPT シリーズと同等の DP 精度を達成できるか．
- 同一のデータセットは異なるアーキテクチャやパラメータサイズの LLM に対しても有効か．
- 訓練データに含まれないドメイン外 DP タスクにおいても精度向上が見られるか．

これらを明らかにするため，本研究では独自に構築した DP 向けインストラクションデータを用い，異なる LLM を比較する．

本研究で提案するファインチューニング済みモデルの Jellyfish は，構造化データを対象とし，ドメイン外タスクやデータセットに対する知識注入仕組みを備えている．さらに，プロンプトエンジニアリングに対する性能を維持しつつ，将来的なカスタマイズにも対応できる汎化能力を有している．また，7B，8B，13B という三つのパラメータサイズを持つモデルを用いて，提案手法の有効性を検証した．

Jellyfish の評価では，広く使用される DP ベンチマークを用いて，非 LLM 手法および GPT シリーズを代表とする LLM 手法との比較を行った．その結果，すべての Jellyfish モデルにおいて DP 性能の大幅な向上が確認された．特に，13B モデルはドメイ

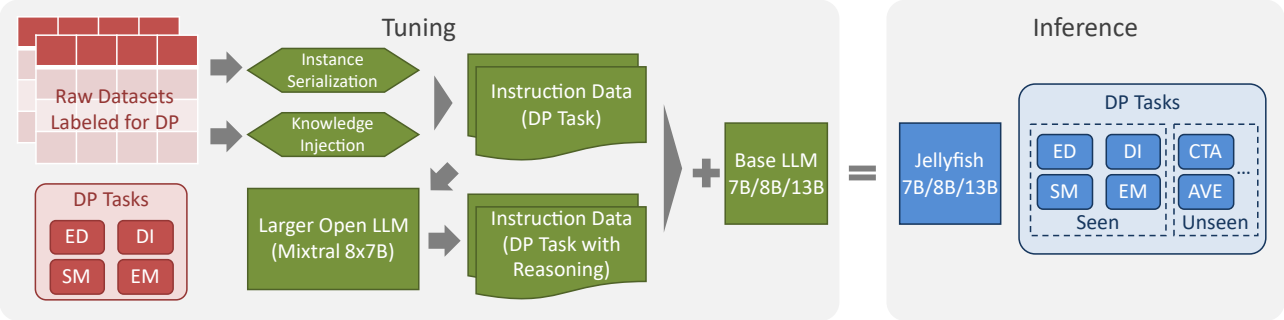

図 1　Jellyfish Instruction-tuning のフレームワーク．

ン内タスクで非 LLM 手法を一貫して上回る性能を示した．また，7B/8B モデルも DI と EM タスクを中心に競争力のある結果を達成した．ドメイン外タスクにおいても，各 Jellyfish モデルは GPT-3.5/4 に匹敵する性能を示し，ファインチューニングに用いたドメイン内タスクを超えた幅広い DP タスクに汎化できる可能性が示唆された．

2　手法

図 1 に示すように，Jellyfish は四つの DP タスク（ED，DI，SM，EM）の公開データセットを選定し，それらをもとにファインチューニング用のデータを構築している．ファインチューニング用のインストラクションデータは，正答のみを示す「DP Task Data」と，理由まで含めた推論を提供する「DP Task with Reasoning Data」の二種類から構成される．生のデータをプロンプトにシリアライズするとともに，タスクやデータセットに関する知識を加えて構築している．これにより，タスクおよびデータセット固有の知識——特にドメイン外データセットにも適用可能なドメイン知識——をプロンプト内に組み込むことが可能となる．DP Task with Reasoning Data は，Mixtral-8x7B-Instruct-v0.1[1)] を用いた知識蒸留により作成された．構築されたインストラクションデータは https://huggingface.co/datasets/NECOUDBFM/Jellyfish-Instruct で公開している．

2.1　データセット整理

過去の研究で広く使用され，様々な領域をカバーする代表的なデータセットとして，以下の 4 つのドメイン内タスクのデータを選択した：ED（Adult，Hospital [1]），DI（Buy，Restaurant [5]），SM（MIMIC-III，Synthea [6]），EM（Amazon-Google，Beer，DBLP-ACM，DBLP-GoogleScholar，Fodors-Zagats，iTunes-Amazon [15]）．

各タスクのデータサイズについては，非 LLM 手法 [5, 6, 8] との公平な比較を実現するため，訓練データに 115k インスタンス以下という制約を設けた．

さらに，訓練データの構築にあたっては，以下のような工夫を行っている：

- **正例インスタンスの全利用**：多くの訓練セットでは正例インスタンスが負例に比べて極端に少ないため，正例を全て使用した．
- **ED データの二重化**：欠損値は文脈次第で「エラー」または「非エラー」と解釈されるのができるため，各インスタンスについて「エラーとして扱う場合」と「非エラーとして扱う場合」の 2 パターンを用意した．

次に，生データを「DP Task Data」と「DP Task with Reasoning Data」の二種類に変換する（具体例は付録 A 参照）．

2.2　DP Task Data

LLM 用の DP Task Data を準備するためには，生データ内の各インスタンスをプロンプトとしてシリアライズする必要がある．プロンプトには，タスクの説明，データインスタンスの内容，および注入知識が含まれる．

DP Task Data は，次の五つのパートから構成される：

- **システムメッセージ**：モデルの動作をガイドするための基本指示
- **DP タスクの説明**：タスクの定義および達成目標
- **知識の注入**：以下の二種類の知識を含む
  - 汎用知識：多くのタスクやデータセットに

1）https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1

適用可能な共通知識
  – 特定知識：特定のデータセットにのみ適用される固有知識
- **インスタンス内容**：シリアライズしたデータインスタンス
- **質問と出力形式**：タスクに応じた質問および期待される出力形式の制約

2.3 DP Task with Reasoning Data

DP Task with Reasoning Data は，モデルに DP 結果の解釈能力を付与するだけでなく，結論に至る根拠を学習することで，調整済みタスクやデータセットと類似したロジックを持つ未知のシナリオへの一般化能力の向上を目指す．

DP Task with Reasoning Data の作成には，大規模オープン LLM である Mixtral-8x7B-Instruct-v0.1 を使用し，推論回答を正解データとして取得した．これにより，モデルは DP タスクにおける推論知識を Mixtral から蒸留する（具体例を付録 B 参照）．この手法は GPT シリーズのような外部 API を使用しないため，データに機密情報を含める場合でもデータセキュリティが確保される．

DP Task with Reasoning Data のプロンプトは DP Task Data のプロンプトと基本的に同一であるが，推論指示，システムメッセージ，および出力形式が異なる（付録 A 参照）．Mixtral から推論回答を取得する際には，プロンプトの最後に正しい DP 結果に関するヒント（例：マッチングタスクにおける「yes/no」）を追加し，正しい方向への推論を促す．ただし，このヒントはモデル評価用のプロンプトには含めていない．

3 実験

3.1 実験設定

データセット．実験には表 1 に示したデータセットを使用した．また，以下の二つをドメイン外 DP タスクとして扱った：

- **列タイプ注釈（CTA）**：ヘッダーのないテーブルに対して，事前定義されたタイプ（例：名前，時間など）から各列の種類を推定するタスク
- **属性値抽出（AVE）**：エンティティのテキスト説明と事前定義された属性セットが与えられた際に，対応する属性値を抽出するタスク

表 1 テストデータセット詳細

タスク	種類	データセット	インスタンス数
ED	ドメイン内	Adult	9900
		Hospital	17101
	ドメイン外	Flights	12832
		Rayyan	8997
DI	ドメイン内	Buy	65
		Restaurant	86
	ドメイン外	Flipkart	2675
		Phone	1194
SM	ドメイン内	MIMIC-III	6408
		Synthea	2964
	ドメイン外	CMS	2564
EM	ドメイン内	Amazon-Google	2293
		Beer	91
		DBLP-ACM	2473
		DBLP-GoogleScholar	5742
		Fodors-Zagats	189
		iTunes-Amazon	109
	ドメイン外	Abt-Buy	1946
		Walmart-Amazon	2049
CTA	ドメイン外	SOTAB	250
AVE	ドメイン外	AE-110K	1482
		OA-Mine	2451

Jellyfish モデル．構築した二種類のインストラクションデータを用いて，Mistral-7B-Instruct-v0.2[2)], Llama 3-8B-Instruct[3)], OpenOrca-Platypus2-13B[4)] という三つの広く利用されているオープンソース LLM に対して LoRA チューニング [16] を実施した：

チューニング後のモデルはそれぞれ Jellyfish-7B, Jellyfish-8B, Jellyfish-13B と呼び，https://huggingface.co/NECOUDBFM/Jellyfish で公開している．

ベースライン．比較対象とする既存手法は，以下の 2 種類に分類される：

- 非 LLM 手法： ED（HoloDetect [1], Raha [2]），DI（IPM [5]），SM（SMAT [6]），EM（Ditto [8], Unicorn [17]），CTA（RoBERTa [18]）．これらの性能値は，先行研究で報告された最良の結果を参照している．
- LLM 手法：GPT-3, GPT-3.5, Table-GPT [13], GPT-4，GPT-4o，Stable Beluga 2 70B，SOLAR 70B

評価指標．DP タスク解決能力の評価には以下の指標を使用する：

2) https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2
3) https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
4) https://huggingface.co/Open-Orca/OpenOrca-Platypus2-13B

表 2　ドメイン内 DP タスク評価

タスク	種類	データセット	モデル								
			Best of non-LLM	GPT-3	GPT-3.5	GPT-4	GPT-4o	Table-GPT	Jellyfish-7B	Jellyfish-8B	Jellyfish-13B
ED	ドメイン内	Adult	99.10	99.10	92.01	92.01	83.58	–	77.40	73.74	**99.33**
		Hospital	94.40	**97.80**	90.74	90.74	44.76	–	94.51	93.40	95.59
	ドメイン外	Flights	81.00	–	–	**83.48**	66.01	–	69.15	66.21	82.52
		Rayyan	79.00	–	–	81.95	68.53	–	75.07	81.06	**90.65**
DI	ドメイン内	Buy	96.50	98.50	98.46	**100**	**100**	–	98.46	98.46	**100**
		Restaurant	77.20	88.40	94.19	**97.67**	90.70	–	89.53	87.21	89.53
	ドメイン外	Flipkart	68.00	–	–	**89.94**	83.20	–	87.14	87.48	81.68
		Phone	86.70	–	–	**90.79**	86.78	–	86.52	85.68	87.21
SM	ドメイン内	MIMIC-III	20.00	–	–	40.00	29.41	–	**53.33**	45.45	40.00
		Synthea	38.50	45.20	57.14	**66.67**	6.56	–	55.56	47.06	56.00
	ドメイン外	CMS	50.00	–	–	19.35	22.22	–	42.86	38.10	**59.29**
EM	ドメイン内	Amazon-Google	75.58	63.50	66.50	74.21	70.91	70.10	**81.69**	81.42	81.34
		Beer	94.37	**100**	96.30	**100**	90.32	96.30	**100.00**	**100.00**	96.77
		DBLP-ACM	**98.99**	96.60	96.99	97.44	95.87	93.80	98.65	98.77	98.98
		DBLP-GoogleScholar	95.70	83.80	76.12	91.87	90.45	92.40	94.88	95.03	**98.51**
		Fodors-Zagats	**100**	**100**	**100**	**100**	93.62	**100**	**100**	**100**	**100**
		iTunes-Amazon	97.06	98.20	96.40	**100**	98.18	94.30	96.30	96.30	98.11
	ドメイン外	Abt-Buy	89.33	–	–	**92.77**	78.73	–	86.06	88.84	89.58
		Walmart-Amazon	86.89	87.00	86.17	**90.27**	79.19	82.40	84.91	85.24	89.42
平均			80.44	-	-	84.17	72.58	-	82.74	81.55	**86.02**

表 3　ドメイン外 DP タスク評価

タスク	データセット	モデル									
		RoBERTa (159 shots)	RoBERTa (356 shots)	Stable Beluga 2 70B	SOLAR 70B	GPT-3.5	GPT-4	GPT-4o	Jellyfish-7B	Jellyfish-8B	Jellyfish-13B
CTA	SOTAB	79.20	89.73	–	–	89.47	**91.55**	65.06	83.00	76.33	82.00
AVE	AE-110k	–	–	52.10	49.20	**61.30**	55.50	55.77	56.09	59.55	58.12
	OA-Mine	–	–	50.80	55.20	62.70	**68.90**	60.20	51.98	59.22	55.96

- DI：精度（Accuracy）
- ED，DI，EM，AVE：F1 スコア
- CTA：マイクロ F1 スコア

全ての評価結果は 100 点スケールで報告する．

3.2　DP タスクによる評価

ドメイン内タスク

表 2 はドメイン内タスクにおける性能評価結果を示している．GPT-4 は 19 個中 11 個のケースで最高性能を達成したが，SM タスクの CMS データセットにおけるスコアは平均的であった．Jellyfish-13B は 7 個のケースで最高性能を記録し，特に CMS データセットにおける優位性により，平均スコアで GPT-4 を上回った．

Jellyfish-13B は GPT-3，GPT-3.5，GPT-4o，および Table-GPT と比較して多くのケースで優れた性能を示した．また，非 LLM 手法の最高スコアについても，全データセットのうち 1 つを除き上回った．ここで，非 LLM 手法は対応するデータセット上でファインチューニングが必要なため，全てのデータセットが「ドメイン内」となることに留意する．

Jellyfish-7B および Jellyfish-8B も競争力のある性能を示し，特に DI および EM タスクで高い成果を達成した．これらのモデルの平均スコアは，非 LLM 手法の最高スコアや GPT-4o を上回っている．

ドメイン外タスク

表 3 はドメイン外タスクにおける性能比較を示している．CTA タスクでは GPT-4 が最高性能を示したものの，Jellyfish モデル，特に 7B および 13B モデルも競争力のある結果を示した．

AVE タスクでは，全ての Jellyfish モデルが高い汎化性能を示した．特に，Jellyfish-8B および Jellyfish-13B は，二つの 70B モデルのスコアを両データセットで上回り，AE-110k データセットでは GPT-4 を上回る性能を達成した．

4　結論

本研究では，LLM を用いた汎用的な DP タスクの解決策として，Instruction-tuning に基づく手法を提案した．具体的には，7B, 8B, 13B の三つのベースモデルに対してチューニングを実施し，ローカル GPU で実行可能な環境を構築することで，データセキュリティを確保しながら運用できる仕組みを実現した．実験結果から，提案手法によって得られた Jellyfish モデルは，既存の DP 解決策に対して競争力のある性能を示し，さらに NLP タスクにおける性能の維持能力（付録 C 参照）や，新規タスクへの優れた汎化性能，および DP 解釈能力（付録 D 参照）を備えていることが確認された．

謝辞

本研究は主にＮＥＣの支援を受け，JSPS 科研費 JP23K17456, JP23K25157, JP23K28096 の助成，および JST, CREST, JPMJCR22M2 の支援を受けたものです．

参考文献

[1] Alireza Heidari, Joshua McGrath, Ihab F Ilyas, and Theodoros Rekatsinas. HoloDetect: Few-shot learning for error detection. In **SIGMOD**, pp. 829–846, 2019.

[2] Mohammad Mahdavi, Ziawasch Abedjan, Raul Castro Fernandez, Samuel Madden, Mourad Ouzzani, Michael Stonebraker, and Nan Tang. Raha: A configuration-free error detection system. In **SIGMOD**, pp. 865–882, 2019.

[3] Theodoros Rekatsinas, Xu Chu, Ihab F Ilyas, and Christopher Ré. HoloClean: Holistic data repairs with probabilistic inference. **PVLDB**, Vol. 10, No. 10, pp. 1190–1201, 2017.

[4] Mohammad Mahdavi and Ziawasch Abedjan. Baran: Effective error correction via a unified context representation and transfer learning. **PVLDB**, Vol. 13, No. 12, pp. 1948–1961, 2020.

[5] Yinan Mei, Shaoxu Song, Chenguang Fang, Haifeng Yang, Jingyun Fang, and Jiang Long. Capturing semantics for imputation with pre-trained language models. In **ICDE**, pp. 61–72. IEEE, 2021.

[6] Jing Zhang, Bonggun Shin, Jinho D Choi, and Joyce C Ho. SMAT: An attention-based deep learning solution to the automation of schema matching. In **ADBIS**, pp. 260–274. Springer, 2021.

[7] Pradap Konda, Sanjib Das, AnHai Doan, Adel Ardalan, Jeffrey R Ballard, Han Li, Fatemah Panahi, Haojun Zhang, Jeff Naughton, Shishir Prasad, et al. Magellan: toward building entity matching management systems over data science stacks. **PVLDB**, Vol. 9, No. 13, pp. 1581–1584, 2016.

[8] Yuliang Li, Jinfeng Li, Yoshihiko Suhara, AnHai Doan, and Wang-Chiew Tan. Deep entity matching with pre-trained language models. **PVLDB**, Vol. 14, No. 1, pp. 50–60, 2020.

[9] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. **arXiv preprint arXiv:2205.11916**, 2022.

[10] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. **NeurIPS**, Vol. 33, pp. 1877–1901, 2020.

[11] Avanika Narayan, Ines Chami, Laurel Orr, and Christopher Ré. Can foundation models wrangle your data? **PVLDB**, Vol. 16, No. 4, pp. 738–746, 2022.

[12] Keti Korini and Christian Bizer. Column type annotation using ChatGPT. **arXiv preprint arXiv:2306.00745**, 2023.

[13] Peng Li, Yeye He, Dror Yashar, Weiwei Cui, Song Ge, Haidong Zhang, Danielle Rifinski Fainman, Dongmei Zhang, and Surajit Chaudhuri. Table-GPT: Table-tuned GPT for diverse table tasks. **arXiv preprint arXiv:2310.09263**, 2023.

[14] OpenAI. March 20 ChatGPT outage: Here's what happened, 2023.

[15] Sanjib Das, AnHai Doan, Paul Suganthan G. C., Chaitanya Gokhale, Pradap Konda, Yash Govind, and Derek Paulsen. The magellan data repository. https://sites.google.com/site/anhaidgroup/useful-stuff/the-magellan-data-repository.

[16] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. **arXiv preprint arXiv:2106.09685**, 2021.

[17] Jianhong Tu, Ju Fan, Nan Tang, Peng Wang, Guoliang Li, Xiaoyong Du, Xiaofeng Jia, and Song Gao. Unicorn: A unified multi-tasking model for supporting matching tasks in data integration. **Proceedings of the ACM on Management of Data**, Vol. 1, No. 1, pp. 1–26, 2023.

[18] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. **arXiv preprint arXiv:1907.11692**, 2019.

[19] Haochen Zhang, Yuyang Dong, Chuan Xiao, and Masafumi Oyamada. Jellyfish: Instruction-tuning local large language models for data preprocessing. In **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 8754–8782, 2024.

[20] Haochen Zhang, Yuyang Dong, Chuan Xiao, and Masafumi Oyamada. Large Language Models as Data Preprocessors. In **2nd TaDA Workshop on VLDB 2024**, 2024.

[21] Alexander Brinkmann, Roee Shraga, and Christian Bizer. Product attribute value extraction using large language models. **arXiv preprint arXiv:2310.12537**, 2023.

[22] Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. Instruction tuning for large language models: A survey. **arXiv preprint arXiv:2308.10792**, 2023.

[23] Guineng Zheng, Subhabrata Mukherjee, Xin Luna Dong, and Feifei Li. Opentag: Open attribute value extraction from product profiles. In **KDD**, pp. 1049–1058, 2018.

[24] Jinsung Yoon, James Jordon, and Mihaela Schaar. GAIN: Missing data imputation using generative adversarial nets. In **ICML**, pp. 5689–5698, 2018.

[25] Saravanan Thirumuruganathan, Han Li, Nan Tang, Mourad Ouzzani, Yash Govind, Derek Paulsen, Glenn Fung, and AnHai Doan. Deep learning for blocking in entity matching: a design space exploration. **PVLDB**, Vol. 14, No. 11, pp. 2459–2472, 2021.

[26] Lilian Weng. Prompt engineering. https://lilianweng.github.io/posts/2023-03-15-prompt-engineering/, 2023.

[27] vLLM Team. vLLM: Easy, fast, and cheap LLM serving with PagedAttention. https://vllm.ai, 2024.

A　インストラクションデータプロンプトの具体例（DP タスク EM）

	DP Task Data	DP Task with Reasoning Data
system message	You are an AI assistant that follows instruction extremely well. User will give you a question. Your task is to answer as faithfully as you can.	You are an AI assistant that follows instruction extremely well. User will give you a question. Your task is to answer as faithfully as you can. While answering, provide detailed explanation and justify your answer.
task description	You are tasked with determining whether two Products listed below are the same based on the information provided. Carefully compare all the attributes before making your decision.	
injected knowledge	Note that missing values (N/A or "nan") should not be used as a basis for your decision.	
instance content	Product A: [name: "Sequoia American Amber Ale", factory: "Wig And Pen"] Product B: [name: "Aarhus Cains Triple A American Amber Ale", factory: "Aarhus Bryghus"]	
question	Are Product A and Product B the same Product?	
output format	Choose your answer from: [Yes, No]	After your reasoning, finish your response in a separate line with and ONLY with your final answer. Choose your final answer from [Yes, No].
answer	No	… the names are not identical, with Product B having … These are different factories … they are not the same product. Final answer: No

B　生成された DP Task with Reasoning Data の具体例 （DP タスク EM）

To determine if Product A and Product B are the same, we need to compare their names and factories. The names of the products are "Sequoia American Amber Ale" and "Aarhus Cains Triple A American Amber Ale". While both are American Amber Ales, the names are not identical, with Product B having additional words "Aarhus Cains Triple A". Therefore, based on the name, we cannot conclude that the products are the same. Next, let's consider the factories. Product A is produced at "Wig And Pen" while Product B is produced at "Aarhus Bryghus". These are different factories, and the fact that they may belong to the same parent company or use similar production processes does not change the fact that the products are made at different locations. Therefore, based on the comparison of the names and factories, we can conclude that Product A and Product B are not the same product.
Final answer: No

C　Open LLM Leaderboard による NLP 評価

サイズ	モデル	MMLU (5-shot)	WinoGrande (0-shot)	ARC (25-shot)	TruthfulQA (0-shot)	GSM8K (8-shot)	HellaSwag (10-shot)	Average
7B	Mistral-7B	62.91	73.88	63.48	66.91	41.32	84.79	65.55
	Jellyfish-7B	62.08 (-0.83)	72.69 (-1.19)	63.48 (+0.00)	64.76 (-2.15)	37.91 (-3.41)	84.48 (-0.31)	64.23 (-1.32)
8B	Llama 3-8B	64.51	71.74	61.01	51.63	70.36	78.61	66.31
	Jellyfish-8B	64.23 (-0.28)	72.06 (+0.32)	60.15 (-0.14)	51.83 (+0.20)	69.29 (-1.07)	77.92 (-0.69)	65.76 (-0.56)
13B	OOP2-13B	54.49	74.03	62.63	52.56	25.32	83.24	58.71
	Jellyfish-13B	53.04 (-1.45)	74.19 (+0.16)	62.88 (+0.25)	52.56 (+0.00)	24.26 (-1.06)	83.16 (-0.08)	58.35 (-0.36)

D　DP タスク解釈能力評価

GPT-4o を判定者として，GPT-3.5 と Jellyfish-7B/8B の DP タスクに対する解釈能力を直接比較した．

タスク	データセット	比較 1		比較 2	
		GPT-3.5	Jellyfish-7B	GPT-3.5	Jellyfish-8B
ED	Adult	**17**	3	4	**16**
	Hospital	4	**16**	4	**16**
DI	Buy	4	**16**	4	**16**
	Restaurant	**10**	**10**	9	**11**
SM	Synthea	**15**	5	3	**17**
EM	Amazon-Google	3	**17**	2	**18**
	Beer	**13**	7	7	**13**
	DBLP-ACM	**11**	9	2	**18**
	DBLP-GoogleScho8lar	**16**	4	9	**11**
	Fodors-Zagats	**13**	7	**13**	7
	iTunes-Amazon	**12**	8	2	**18**
合計		**118**	102	59	**161**
勝率		**53.63%**	46.36%	26.81%	**73.18%**