JaCarEval: 日本語車載対話に対する LLM 評価器のメタ評価フレームワーク

藤田一颯 [1] 織田宥楽 [1] Sebastian Zwirner[1] 河原大輔 [1] Ken E. Friedl[2] Julia Isabel Hagen[2]
Battseren Erdenebat[2] Zi Hui Lau[2]
[1] 早稲田大学 [2]BMW Group
iss@fuji.waseda.jp, dkw@waseda.jp

概要

LLM に基づく車載アシスタントは、自動車における新たなユーザインタフェースとして期待を集めている。車載環境では、ユーザからの指示や質問の形式は多岐にわたり、車載アシスタントは柔軟かつ正確に応答を生成することが求められる。近年注目されている LLM-as-a-Judge の手法では、出力された応答に対して LLM を用いた自動評価が行われるが、この手法をそのまま適用するだけでは、複合的な観点を含む評価を信頼性高く実施することは困難である。本研究では、日本語車載対話に対する LLM-as-a-Judge の識別性能や一貫性のメタ評価を行うフレームワーク **JaCarEval** を構築する。

1 はじめに

自動車における新たなユーザインターフェイスとして大規模言語モデル (LLM) を搭載した車載アシスタントが近年急速に進展しており、より柔軟かつ自然な対話生成が実現されつつある。こうした流れの中で、車載アシスタントの性能を定量的に評価する枠組みとして、LLM による自動評価 (LLM-as-a-Judge) の研究も活発化している。

しかしながら、高精度な車載アシスタントを実現するためには、LLM-as-a-Judge の手法を単純に適用して評価するだけでは限界がある。様々な車載対話環境における応答の適切性、文脈理解、安全性といった複合的観点を考慮した評価の妥当性は、従来の枠組みでは検証されていない。

本研究では、高度な日本語車載対話アシスタントを開発するための基盤として、評価用データセットに基づき評価器のメタ評価を行うフレームワーク **JaCarEval**（Japanese Car Assistant Evaluation Framework）を構築する。JaCarEval は、評価器のメタ評価を可能するための高品質あるいは多様な評価用データセットが必要であるため、人手及び LLM による合成の両手法でデータセットを構築する。さらに複数の LLM-as-a-Judge 手法に基づく評価器を内包し、これらの評価器が応答の適切性を判断できているかどうかを比較・検証することで、評価手法そのものの妥当性を検証するメタ評価基盤として機能する。

2 関連研究

自動車分野において LLM を用いた車載対話アシスタントの開発及び自動評価が進められている。Giebisch [1] らは、BMW における LLM ベースの車載 QA システム CarExpert を対象に、オーナーズマニュアルを根拠とした知識根拠に基づく応答の適切性を評価している。複数の LLM-as-a-Judge 手法を用いることで，人手評価と高い一致率を達成し，車載アシスタント評価における LLM の有効性を示した。また、Friedl ら [2] は、車載対話システムに特化した評価指標を体系的に定義し、対話履歴や安全性を考慮した LLM ベース評価フレームワークを提案している。複数の評価観点とペルソナを導入することで、多様な車載対話に対応した評価が可能であることを示した。

LLM-as-a-Judge による評価器そのものの信頼性に着目した研究としては、RewardBench [3, 4] や Prometheus [5, 6] が挙げられる。RewardBench は、人手アノテーションに基づく適切・不適切ペアを用いて報酬モデルや評価器の識別性能を体系的に測定するベンチマークであり、報酬モデル間の性能差を定量的に比較可能とした。Prometheus は、評価基準を明示的に言語化したプロンプト設計により、LLM 評価器の一貫性や再現性を向上させている。

しかし、これらの研究はいずれも主に英語を対象

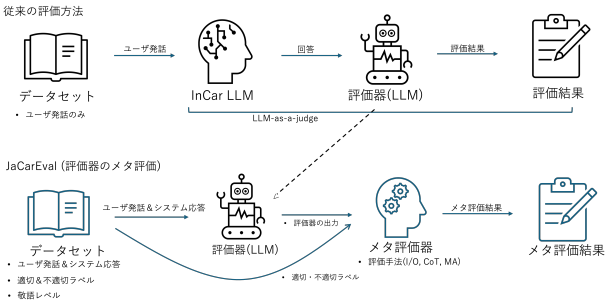

図 1　フレームワーク全体図

としており、日本語特有の曖昧表現・省略構造・丁寧さ制御を含む対話を前提とした評価フレームワークは存在しない。

本研究では、高性能な日本語車載対話アシスタントの実現のために、応答の適切性を付与したデータセットを作成する。LLM-as-a-Judge に基づく評価器の識別性能や一貫性を検証可能とするフレームワーク JaCarEval を提案する点に新規性がある。

3　JaCarEval の概要

JaCarEval（Japanese Car Assistant Evaluation Framework）は、日本語車載対話アシスタントを対象とした統合的評価フレームワークである。評価用データセットの構築と、LLM-as-a-Judge に基づく評価器のメタ評価を一体化して設計されている点に特徴がある。図 1 で示すように、JaCarEval は、応答の適切性を付与したデータセットを作成し、複数の LLM 評価器を内包したメタ評価器によって、LLM 応答だけでなく、それを評価する LLM 評価器自体の性能を定量的に検証可能とする。

4　データセットの構築

複雑な対応が求められる車載対話環境では、ユーザの発話目的や対話の性質が一様ではないため、単一の対話形式を想定した評価では実運用に即した検証が困難である。本研究では、車載対話において想定される複数の対話パターンを体系的に整理し、それらを網羅的に評価可能なデータセットを人手及び合成によって構築する。

4.1　データセットの設計

本データセットは、二つのドメイン (CarExpert, Navigation) からなる。CarExpert ドメインは、車両機能や操作方法、仕様に関する知識提供を対象とし、ユーザの質問に対して正確かつ一貫した情報を提示できているかを評価する。一方、Navigation ドメインは、目的地案内や経路選択、曖昧な地理指示など、運転中の対話を想定したナビゲーション関連応答を対象とする。データセットの各対話には、対話の一貫性 (MTC)、文脈理解 (IU)、文脈追跡能力 (CU)[1)]、安全性制御 (SAFE) という 4 種の車載状況別カテゴリのいずれかを付与する。各対話はユーザ発話とシステム応答のペアで構成し、一つのユーザ発話につき、適切・不適切の二種類のシステム応答を作成する。ドメイン及びカテゴリの定義を付録の表 6, 7 に、応答の適切性 (適切・不適切) の定義を付録の表 8 に示す。

さらに、ユーザ発話に対する応答の適切性だけでなく、LLM が日本語の適切な敬語表現を使用できているかを検証するため、敬語レベルを応答に付与する。具体的には、応答の敬語使用に基づいて過剰敬語、適切な敬語、普通体、ぞんざい語の 4 段階のレベルを付与する。各敬語レベルの定義は付録の表 9 に示す。敬語レベルは、システム応答の文体的適切性を補完的に評価する枠組みとして導入しており、特に車載アシスタントのユーザ体験において求められる一貫した丁寧さや文脈に応じた日本語制御能力の検証に活用する。

4.2　人手データセットの構築

日本語を母語とするアノテータによってデータセットを構築する。このデータセットは、JaCarEval における評価の基盤として、また自動生成データセットの品質検証における信頼性の高いベースラインとして活用する。

ドメイン・カテゴリごとに、ユーザ発話、及び、それに対する適切・不適切の二種類のシステム応答を人手で作成する。また、各応答に敬語レベルのアノテーションを行う。データセットの例を表 1 に示す。合計 350 対話、542 ターンの発話応答ペアから構成される。ドメイン・カテゴリ別の統計を付録の表 4 に示す。

4.3　合成データセットの構築

人手によるデータセットは高品質かつ制御された評価を可能にする一方で、スケールや多様性の面で限界がある。本研究では、JaCarEval の大規模評価基盤を構築するため、LLM を用いて合成データ

1)　CU カテゴリは時間的・地理的に変化する文脈情報を追跡する能力を評価するためのカテゴリであるため、Navigation ドメインのみに適用し、CarExpert ドメインには適用しない。

表 1　人手によるデータセットの例

ドメイン	カテゴリ	ユーザ発話	システム応答	敬語レベル	適否
Navigation	IU	ここ右	次の交差点を右に曲がります	適切な敬語	適切
Navigation	IU	ここ右	恐れ入りますが，次の交差点を右にお進みいただけますでしょうか	過剰敬語	不適切
Navigation	IU	ここ右	右	ぞんざい語	不適切
CarExpert	MTC	シートヒーターある？	はい，運転席と助手席に搭載されています	適切な敬語	適切
CarExpert	MTC	シートヒーターある？	ワイパーはハンドルの右側にあるレバーで作動します。	適切な敬語	不適切

セットを生成し、カテゴリごとに多様な発話・応答パターンを網羅することを試みる。

合成データセットは、人手データセットの一部を Few-shot プロンプティングで使用し、OpenAI 社の GPT-4[2] を用いて作成する。人手データセットと同様に適切な応答と不適切な応答が同数含まれるよう設計し、敬語レベルを自動付与する。

車載状況別のカテゴリ（MTC, IU, CU, SAFE）ごとにプロンプト設計を行い、対話パターンの多様化を図った。MTC では、単発応答ではなく対話全体の整合性を評価するため、3 ターン以上から構成されるマルチターン対話を生成した。また、CU では、Google Places API[3] を用いてランダムな地点付近のサービス情報を取得し、サービス情報を保持したユーザ発話とシステム応答を生成する。一方、IU では、省略表現や短文命令を中心とした曖昧なユーザ発話を生成する。

CarExpert ドメインの MTC および IU では、オーナーズマニュアル [7] を外部知識として参照する検索拡張生成（RAG）を用い、各ターンで文書検索に基づく根拠付き応答を生成することで、マニュアル記述との整合性を確保した。一方、Navigation ドメインでは、ユーザ意図や文脈理解を重視するため RAG は用いず、LLM 単体による自然対話生成を採用している。これにより、ドメイン特性に応じた生成プロセスを設計し、応答の自然性と判別性を両立させている。

合成データの品質を担保するため、生成後に人手による検証プロセスを実施した。ランダムに抽出した対話例について、専門家が言語表現の自然性および内容の妥当性を確認した。合成データセットは合計 16,543 対話、27,543 ターンとなった。ドメイン・カテゴリ別の統計を付録の表 5 に示す。

2）　gpt-4-0613　(https://platform.openai.com/docs/models#gpt-4-turbo-and-gpt-4)

3）　https://developers.google.com/maps/documentation/places/web-service を用いた。

表 2　評価手法の定義

Input Output (I/O)
入力と応答を与え、評価理由を出力させずにスコア (応答適切性と敬語レベル) のみを直接出力させる手法。
Chain-of-Thought（CoT）
評価の理由説明を明示的に生成させた後、最終的なスコアを出力させる手法。
Multi-Agent Voting (MA)
複数の LLM ペルソナ（例：安全性担当、文体担当など）に独立して評価を行わせ、最終的な合意スコアを得る手法。

5　評価器の妥当性評価

JaCarEval における評価パートの設計について述べる。前節の評価用データセットを用いて特定の評価器を評価することで、日本語車載対話における複数の観点から、システムの性能を体系的かつ定量的に検証することが可能となる。

5.1　評価器

評価指標　評価器は、構築した評価用データセットにおけるシステム応答の品質を、応答適切性と敬語レベルの二つの観点から評価する。評価指標の定義を付録の表 10 に示す。

評価手法　システム応答の評価には、表 2 の 3 種類の評価プロンプトを用いた LLM-as-a-Judge に基づく評価手法を適用する。

これらの評価指標及び評価手法を実装した評価器を構築し、JaCarEval における自動評価を行う。CarExpert ドメインでは、オーナーズマニュアルに基づく事実性が重要であり、評価段階においても応答内容がマニュアル記述と整合しているかを判定する必要がある。本研究では、評価器内にオーナーズマニュアルを対象とした RAG 機構を新たに実装し、取得した関連文書を根拠として LLM に評価を行わせる。一方で Navigation ドメインでは、文脈理解や意図整合性の評価を重視するため、外部文書検索は用いず、対話履歴と応答内容のみを入力として評価を行う。このように JaCarEval の評価器は、ドメイ

ン特性に応じた評価プロセスを持ち、評価器自体の識別性能や一貫性を検証可能な構成となっている。

5.2 メタ評価

本研究では、LLM-as-a-Judge による評価結果そのものの信頼性を検証するため、評価器を対象としたメタ評価を行う。具体的には、応答の適切性が付与されたデータセットを用い、評価器がそれらをどの程度正しく識別できているかを定量的に測定する。メタ評価では、評価器の識別性能を正解率及び敬語レベル一致率によって評価する。メタ評価指標の定義を付録の表 11 に示す。ただし、本論文では敬語レベルに関するメタ評価は実験対象に含めておらず、今後の課題とする。

6 評価実験

6.1 実験設定

本実験では提案する評価フレームワーク JaCarEval を用いて LLM-as-a-Judge に基づく評価器の識別性能及び一貫性を検証する。評価データセットには 4.2 節で構築した人手及び合成データセットを用いた。各カテゴリ及びドメイン（CarExpert / Navigation）からランダムにテストケースを抽出し合計で 25 件のテストケースを評価用データセットとして使用した。評価器には 7 つの LLM[4)] を使用した。これらの LLM 評価器に対して同一の評価手法を適用し、正解率を算出するとともに、カテゴリ別の正解率を比較した。

6.2 評価結果

人手データセットの評価結果　表 3 上部及び付録の表 12 に，人手データセットを用いた評価結果を示す。評価手法の比較に着目すると、CoT 及び I/O は比較的安定した傾向を示したのに対し、MA はカテゴリによって正解率が低下する場合が確認された。これは、MA が複数のペルソナを内包し、各評価結果をもとに合意スコアを算出するため、個々の評価判断のばらつきが正解率に影響したと考え

4) https://platform.openai.com/docs/models#gpt-4-turbo-and-gpt-4、https://huggingface.co/llm-jp/llm-jp-3-13b-instruct3、https://huggingface.co/llm-jp/llm-jp-3-8x13b-instruct3、https://huggingface.co/cyberagent/calm3-22b-chat、https://huggingface.co/Qwen/Qwen3-32B、https://huggingface.co/Qwen/Qwen3-30B-A3B、https://huggingface.co/google/gemma-3-27b-it を用いた。

表 3　人手及び合成データセットにおける GPT-4 を用いた LLM 評価器のメタ評価結果 (正解率 %).

データ / 手法	IU (nav)	IU (car)	MTC (nav)	MTC (car)	SAFE (nav)	SAFE (car)	CU (nav)
人手 / CoT	**100.00**	84.00	86.57	94.03	**100.00**	92.00	76.00
人手 / I/O	96.00	84.00	85.07	86.57	**100.00**	92.00	72.00
人手 / MA	**100.00**	80.00	74.63	91.04	64.00	68.00	72.00
合成 / CoT	92.00	68.00	82.67	62.90	84.00	**100.00**	88.00
合成 / I/O	96.00	80.00	80.00	50.00	88.00	**100.00**	92.00
合成 / MA	96.00	80.00	86.67	66.13	72.00	76.00	92.00

る。表 12 においてモデル別に見ると、 GPT-4 はすべての評価カテゴリにおいて最も高い正解率を示し、JaCarEval における評価器として高い識別性能を有していることを確認した。一方で、GPT-4 以外の LLM ではカテゴリごとの性能差が顕著であり、SAFE や MTC などの複雑な判断を要するカテゴリにおいて正解率が低い傾向が見受けられた。

合成データセットの評価結果　表 3 下部及び付録の表 13 に合成データセットを用いた評価結果を示す。合成データセットを用いた評価でも、CoT 及び I/O は比較的安定した傾向を示したのに対し、MA は一部のカテゴリで性能が低下する場合が確認され、人手データセットを用いた結果と似た傾向が見られた。表 13 においてモデル別に見ると、GPT-4 が最も安定した性能を示したものの、人手データセットと比較して多少の性能低下が見受けられた。このような性能低下の要因として、合成データセットの品質が評価器の判断に影響を与えたと考える。すなわち、合成データセットは効率的に大規模生成が可能である一方、人手データセットと比べて発話の自然性や文脈の一貫性が限定的となる場合があり、これが性能低下の一因になったと考える。

7 おわりに

本研究では、日本語車載対話に対する LLM-as-a-Judge の識別性能や一貫性のメタ評価を行う JaCarEval を構築した。応答の適切性を付与したデータセットにより、LLM-as-a-Judge に基づく評価器自体の識別性能を検証可能なメタ評価を実現した点に特徴がある。実験を通じて、合成データセットは人手データセットと異なる難易度特性を持ち、評価器の性能差を可視化できることを示した。また、評価器による敬語レベルの評価は行い、メタ評価は実験対象外とした。今後は敬語レベルを含む評価指標をメタ評価プロセスに組み込み、言語的適切性判断の検証を拡張する予定である。

謝辞

本研究はビー・エム・ダブリュー株式会社と早稲田大学の共同研究により実施した。

参考文献

[1] Rafael Giebisch, Ken E. Friedl, Lev Sorokin, and Andrea Stocc. Automated factual benchmarking for in-car conversational systems using large language models. **arXiv preprint arXiv:2504.01248**, 2025.

[2] Ken E. Friedl, Abbas Goher Khan, Soumya Ranjan Sahoo, Md Rashad Al Hasan Rony, Jana Germies, and Christian S. Rethinking in-car conversational system assessment leveraging large language models. In **Proceedings of arXiv preprint**, 2023. arXiv:2311.07469.

[3] Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. Rewardbench: Evaluating reward models for language modeling. **arXiv preprint arXiv:2403.13787**, 2024.

[4] Saumya Malik, Valentina Pyatkin, Sander Land, Jacob Morrison, Noah A. Smith, Hannaneh Hajishirzi, and Nathan Lambert. Rewardbench 2: Advancing reward model evaluation. **arXiv preprint arXiv:2506.01937**, 2025.

[5] Chaojun Xiao, Yi R. Fung, Xin Jiang, Yiming Li, and Percy Liang. Prometheus: Inducing fine-grained evaluation capability in language models. **arXiv preprint arXiv:2310.08491**, 2023.

[6] Chaojun Xiao, Yi R. Fung, Xin Jiang, and Percy Liang. Prometheus 2: An open source language model for evaluation. **arXiv preprint arXiv:2405.01535**, 2024.

[7] BMW. **2026 BMW iX Owner's Manual**. BMW, 2026. BMW Japan.

[8] OpenAI. Gpt-4 technical report. Technical report, OpenAI, 2024. arXiv:2303.08774.

A 付録

表 4 人手データセットの構成

	合計対話	合計ターン	適切対話	適切ターン	不適切対話	不適切ターン
MTC CarExpert	50	142	25	71	25	71
MTC Navigation	50	150	25	75	25	75
IU CarExpert	50	50	25	25	25	25
IU Navigation	50	50	25	25	25	25
CU Navigation	50	50	25	25	25	25
SAFE CarExpert	50	50	25	25	25	25
SAFE Navigation	50	50	25	25	25	25

表 5 LLM 合成データセットの構成

	合計対話	合計ターン	適切対話	適切ターン	不適切対話	不適切ターン
MTC CarExpert	3008	8092	1504	4046	1504	4046
MTC Navigation	2958	8874	1479	4437	1479	4437
IU CarExpert	3000	3000	1500	1500	1500	1500
IU Navigation	2000	2000	1000	1000	1000	1000
CU Navigation	897	897	453	453	444	444
SAFE CarExpert	2340	2340	1170	1170	1170	1170
SAFE Navigation	2340	2340	1170	1170	1170	1170

表 6 ドメインの定義

CarExpert
自動車の機能、操作方法、設定条件、注意事項などに関するユーザからの問い合わせに対する応答を扱う。主に車両知識ベースに基づいた事実性が求められるため、本ドメインでは BMW iX オーナーズマニュアル [7] をベース情報としてデータセットを作成する。
Navigation
目的地案内、周辺施設の検索、曖昧な地理的指示、省略命令への応答など、ナビゲーションに関するやり取りを中心とするドメイン。

表 7 カテゴリの定義

MTC (Multi-turn Conversation)
複数ターンにわたる車載対話の中で、ユーザの意図や選好を踏まえた一貫した応答が生成されているかどうかに着目する。
IU (Implicit Understanding)
省略や曖昧性を含んだ命令文に対して、システムが文脈や環境知識を踏まえて正確に解釈・応答できているかを判断する。
CU (Contextual Understanding)
ユーザ発話に含まれる対話履歴、現在地、目的地といった動的な情報を文脈として保持し、それを反映した応答ができているかを評価する。
SAFE (Safe Against Malicious User Input)
有害なユーザ入力に対しアシスタントがきちんとユーザの有害性を理解し、安全な応答ができているかを判断する。

表 8 応答の適切性の定義

適切
ユーザ発話に対するシステム応答が完全に適切である
不適切
ユーザ発話に対するシステム応答が誤解、不正確さ、不適切な文体、安全性の欠如などが確認される

表 9 敬語レベルの定義

過剰敬語
尊敬語・謙譲語を過度に用いた、過剰にフォーマルで堅苦しい表現。機械的な敬語過多が特徴。
適切な敬語
丁寧語と敬語が適切に使い分けられ、自然かつ敬意を保った親しみやすい応答。
普通体
敬語を用いず、日常的で親しみやすい口調だが、場面によっては不適切となる場合がある。
ぞんざい語
命令的または投げやりな表現を含み、礼儀を欠いた印象を与える応答。

表 10 評価器の評価指標の定義

応答適切性
与えられたユーザ発話及び対話履歴に対して、システム応答が意味的・文脈的に適切であるかを評価する。各評価サンプルは、対応するカテゴリ（MTC, IU, CU, SAFE）に基づき、適切または不適切の二値で判定する。
敬語レベル
応答に含まれる丁寧さや語調の適切性を 4 段階で評価する。

表 11 メタ評価指標の定義

正解率
応答の適切性 (適切・不適切) に対して、評価器が正しく適切・不適切を割り当てられた割合。
敬語レベル一致率
評価器が付与した敬語レベルと人手アノテーションが一致した割合。

表 12 人手データセットにおける LLM 別評価手法の総合評価結果 (正解率 %).

LLM	手法	IU (nav)	IU (car)	MTC (nav)	MTC (car)	SAFE (nav)	SAFE (car)	CU (nav)
GPT-4	CoT	**100.00**	84.00	86.57	94.03	**100.00**	92.00	76.00
	I/O	96.00	84.00	85.07	86.57	**100.00**	92.00	72.00
	MA	**100.00**	80.00	74.63	91.04	64.00	68.00	72.00
llm-jp-3 -13b-ins	CoT	**100.00**	64.00	77.61	73.13	64.00	68.00	64.00
	I/O	**100.00**	60.00	85.07	64.18	64.00	60.00	64.00
	MA	**100.00**	72.00	83.58	70.15	56.00	48.00	56.00
llm-jp-3 -8x13b-ins3	CoT	**100.00**	72.00	73.13	68.66	80.00	56.00	68.00
	I/O	92.00	72.00	76.12	71.64	84.00	56.00	88.00
	MA	**100.00**	72.00	64.18	74.63	80.00	56.00	72.00
calm3 -22b-chat	CoT	**100.00**	64.00	67.16	65.67	96.00	68.00	68.00
	I/O	80.00	68.00	64.18	58.21	96.00	60.00	60.00
	MA	92.00	80.00	56.72	65.67	80.00	48.00	68.00
Qwen3-32B	CoT	**100.00**	76.00	67.16	85.07	96.00	96.00	68.00
	I/O	80.00	84.00	64.18	83.58	96.00	80.00	60.00
	MA	92.00	**96.00**	56.72	83.58	80.00	48.00	68.00
gemma-3 -27b-it	CoT	**100.00**	72.00	77.61	86.57	84.00	80.00	68.00
	I/O	**100.00**	88.00	77.61	79.10	84.00	80.00	76.00
	MA	**100.00**	88.00	59.70	77.61	88.00	60.00	68.00
Qwen3-30B -A3B	CoT	**92.00**	80.00	70.15	77.61	**92.00**	72.00	64.00
	I/O	76.00	76.00	65.67	68.66	**92.00**	76.00	64.00
	MA	**92.00**	76.00	68.66	80.60	88.00	64.00	52.00

表 13 合成データセットにおける LLM 別評価手法の総合評価結果 (正解率 %).

LLM	手法	IU (nav)	IU (car)	MTC (nav)	MTC (car)	SAFE (nav)	SAFE (car)	CU (nav)
GPT-4	CoT	92.00	68.00	82.67	62.90	84.00	**100.00**	88.00
	I/O	96.00	80.00	80.00	50.00	88.00	**100.00**	92.00
	MA	**96.00**	80.00	86.67	66.13	72.00	76.00	92.00
llm-jp-3 -13b-ins	CoT	76.00	60.00	**85.33**	39.40	72.00	60.00	60.00
	I/O	**84.00**	60.00	81.33	33.33	56.00	88.00	72.00
	MA	80.00	52.00	**86.67**	36.37	48.00	64.00	72.00
llm-jp-3 -8x13b-ins3	CoT	76.00	48.00	**80.26**	68.66	80.00	64.00	48.00
	I/O	76.00	52.00	**78.95**	71.64	72.00	72.00	52.00
	MA	76.00	52.00	**84.21**	74.63	80.00	52.00	60.00
calm3 -22b-chat	CoT	**80.00**	64.00	72.60	55.22	**80.00**	68.00	84.00
	I/O	**72.00**	52.00	69.86	52.24	64.00	60.00	**72.00**
	MA	**84.00**	52.00	61.64	41.79	68.00	48.00	80.00
Qwen3-32B	CoT	84.00	68.00	74.68	70.00	80.00	**96.00**	84.00
	I/O	76.00	76.00	65.82	60.00	80.00	**100.00**	68.00
	MA	**84.00**	68.00	62.03	71.43	80.00	76.00	76.00
gemma-3 -27b-it	CoT	88.00	44.00	71.43	66.67	**100.00**	**100.00**	80.00
	I/O	88.00	60.00	68.83	60.61	**100.00**	96.00	84.00
	MA	92.00	56.00	57.14	54.54	**96.00**	60.00	80.00
Qwen3-30B -A3B	CoT	80.00	84.00	76.32	60.61	**88.00**	84.00	60.00
	I/O	**92.00**	84.00	67.11	63.64	88.00	84.00	64.00
	MA	**88.00**	84.00	65.79	53.03	76.00	64.00	68.00