LLM エージェントにおける指示曖昧性の内部表現解析

貝出 直大 [1]　井之上 直也 [1]
[1] 北陸先端科学技術大学院大学
{naohiro_kaide,naoya-i}@jaist.ac.jp

概要

LLM エージェントの実環境適用において，指示の曖昧性検知は重要課題である．既存研究は出力監視に留まり，内部機序の解明は不十分であった．本研究では，3 次元ブロック配置タスクにおける LLM の内部状態を Linear Probing により解析した．実験の結果，(1) 曖昧性は中間層以降で線形分離可能であること，(2) 内部表現による判定精度はテキスト出力を上回り，内部表現がテキスト出力に十分反映されていない可能性があること，(3)CoT はタスク遂行を改善するが内部の曖昧性表現の質には寄与しないこと，が明らかになった．本研究はエージェントの指示曖昧性を内部表現レベルで解析した初めての試みである．

1　はじめに

大規模言語モデル（LLM）は情報検索（QA）から，実環境への介入を伴う「自律エージェント」へとその役割を急速に拡大している [1]．しかし，現実世界においてユーザーが発する指示は，常に明確であるとは限らない（図 1, 左）．曖昧な指示に対するエージェントの恣意的な解釈の元での実行は，現実世界における不可逆的な誤操作に直結する危険性を孕んでいる．したがって，エージェントが指示の曖昧さを正しく検知し，必要に応じてユーザーに聞き返す能力を持つことは，システムの安全性を担保する上で重要である．

指示の曖昧性検知の課題に対しては，既存研究の多くは，高精度な検知器の構築という「エンジニアリング」に焦点を当ててきた [2, 3, 4, 5, 6]．しかし，実環境への作用を伴うクリティカルな環境では，ブラックボックスな出力監視のみによる安全性担保には限界がある [7]．予期せぬエラーを防ぎ信頼性を根本から高めるには，LLM 内部における指示曖昧性の処理メカニズムの解明が不可欠である．

そこで本研究では，指示遂行タスクにおけるモデルの内部機序の解明を最終目標として，以下の 3 つのリサーチクエスチョン（RQ）に取り組む．

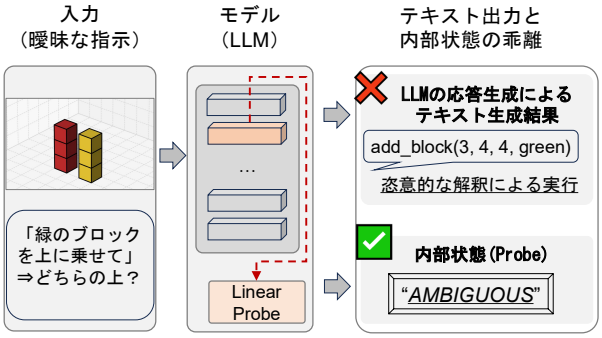

図 1　LLM エージェントへの指示における曖昧性とモデル内部・出力の乖離．「緑のブロックを上に置いて」という多義的な指示に対し，モデル内部では曖昧性が表現されているもののテキスト出力は指示に対する恣意的な解釈により実行してしまう

- **RQ1:** 曖昧なユーザー指示は，LLM の内部でどのように表現されているか？
- **RQ2:** 内部表現から抽出できる曖昧性情報は，プロンプトベースのテキスト出力による判定と比べてどの程度異なるか？
- **RQ3:** プロンプト手法は，曖昧なユーザの指示の内部表現にどのような影響を与えるのか？

我々の仮説は次の通りである: (1) QA タスクにおける回答可能性と同様 [8]，指示曖昧性は中間層以降で線形分離可能な形で符号化されている，(2) Sycophancy Bias [9] の影響により，モデルは内部で曖昧性を表現していても出力に反映しない場合がある，(3) プロンプトはモデル内部の曖昧性情報を精錬しテキスト出力に正確に反映することに寄与するが [10]，内部表現そのものには影響がない．

これらの仮説を検証するために，モデルの内部状態から推定できる指示曖昧性と，プロンプトから得られる出力を比較した（図 1）．具体的には，Instruction Tuning 済み LLM と IGLU データセット [11] を用い，(1) Linear Probing による各層内部状態の曖昧性情報の測定，(2) 各種プロンプト

（Few-shot, CoT 等）によるテキスト出力判定と内部表現ベースの識別性能の比較，(3) プロンプト条件が内部表現に与える影響の分析，を行った．

実験の結果，Probe における曖昧性の識別精度は中間層から深層にかけて向上する傾向が見られた．また，内部表現による判定精度はテキスト出力と同等か，時にはそれ以上であり，モデルが内部で表現している曖昧性を出力に十分に反映できていない可能性が示唆された．さらに，プロンプトの種類によらず内部識別精度はほぼ一定であり，プロンプトは内部表現そのものを変容させるのではなく，既存の曖昧性情報を出力に正確に引き出す役割に留まることが示唆された．本研究は，LLM エージェントにおける指示曖昧性の検知を内部表現レベルで解析した初めての試みであり，LLM エージェントの安全性の向上に資する知見を提供するものである．

2 関連研究

エージェントの指示理解において，文脈依存の曖昧性解消は重要な課題である．この能力を評価するため，ブロック配置タスクを扱う IGLU ベンチマーク [11] に加え，近年ではソフトウェアエンジニアリングの文脈においても，評価環境が整備されつつある [6]．しかし，指示の曖昧性検知の課題に対する既存の解決手法の多くは，教師あり学習による分類 [2, 3, 5] や出力確率の閾値処理 [4]，プロンプトによる工夫 [5, 6] など，高精度な検知器を構築することに焦点を当てており，内部的なメカニズムの解明には焦点が当てられていない．

LLM の内部でのメカニズムを解明するために，機械論的解釈可能性（Mechanistic Interpretability）[7, 12, 10] の分野では，モデル内部の表現を直接観測するアプローチが進展している． 特に，モデル内部のベクトル表現から特定の属性を抽出する線形分類器（Probe） [13] を用いることで，モデルの真実性や知識の保有状態を識別する研究が行われている [14]． QA タスクにおいては，質問の回答可能性や曖昧性が内部で線形分離可能な形に符号化されていることが示されている [8, 15]． しかし，QA タスクでは入力テキストに内在する言語的な曖昧性が対象となるのに対し [16, 17, 18, 19]，エージェントタスクでは環境状態との相互作用によって動的に変化する曖昧性を考慮する必要があり，この点に関する検証は未だ十分ではない． 本研究は，現在の環境状態との整合性によって動的に変化しうる「指示の実行可能性」に着目し，その内部表現と出力挙動の関係を解析する点に独自性を持つ．

3 問題設定

本研究では，エージェントによる指示の曖昧性検知において，入力 x をタスク定義 T，環境情報 E，対話履歴 H，およびユーザー指示 I の組（$x = \{T, E, H, I\}$）として定式化する．

ここで，T, E, H に照らして I が一意の実行プランに帰着する場合を「明確（Clear）」とし，複数の解釈が成立するような情報不足や，指示内容が環境や文脈と矛盾し対象を特定できない接地不全が生じる場合を「曖昧（Ambiguous）」と定義する．

エージェントモデル f のタスクは，x に対して，実行プランとして操作コマンドまたは指示の不備を指摘する応答を生成することである．その生成結果を y_{text} と表す．ただし，本研究ではその推論プロセスの内部において，指示の明確性を判定する潜在的な二値分類 $y_{\text{label}} \in \{0(\text{Ambiguous}), 1(\text{Clear})\}$ が同時に行われていると仮定し，その内部表現を解析の対象とする．

4 分析方法

プロンプトを用いて生成された応答と，内部表現に符号化された情報を比較するために，以下の 2 つの観点から分析を行う．実験時の詳細な分析手法については付録 A に記載する．

4.1 行動評価

モデルのテキスト生成結果 y_{text} に基づき，以下の指標で評価する．

- **タスク成功率 (Exact Match):** 指示が明確な場合において，モデルが生成したコマンドによりタスクが成功したかどうかを評価する．これは，モデルが基本的な指示遂行能力を有しているかを確認するための指標である．3 次元ブロック配置タスクの場合は，最終的なブロック配置が目標の配置と一致するかを判定する．
- **曖昧性判定精度 (Macro F1):** 先行研究 [11] と同様，指示が曖昧な場合の拒絶と指示が明確な場合の実行が正しく動作していたかを Macro F1 スコアから評価する．文献 [8] に倣い，"ambiguous", "unclear", "cannot determine" 等のキーワードが含まれる場合，または有効な操作コマンドが生成されなかった場合を「曖昧性

表 1 Qwen3-14B および Gemma-3-12B におけるテキスト生成による曖昧性判定評価．Hint（曖昧性の示唆），CoT（Chain-of-Thought），Shot 数によるタスク成功率（EM）と曖昧性判定精度（F1）の比較．各モデルにおける最大値を太字で示す．

Setting			Qwen3-14B		Gemma-3-12B	
Hint	CoT	Shot	EM	F1	EM	F1
×	×	Zero	0.237	0.486	0.035	0.485
×	×	Few	0.293	0.500	0.202	0.486
×	✓	Zero	0.514	0.496	0.157	0.483
×	✓	Few	**0.534**	0.510	**0.309**	0.508
✓	×	Zero	0.098	**0.627**	0.120	0.577
✓	×	Few	0.317	0.596	0.203	0.551
✓	✓	Zero	0.478	0.583	0.258	0.511
✓	✓	Few	0.490	0.624	0.275	**0.598**

検知」とみなす．

4.2 内部分析

モデル内部における曖昧性情報の符号化を検証するため，Linear Probing を用いる．具体的には，モデルの第 l 層における隠れ状態ベクトル $\mathbf{h}^{(l)}$ を入力とし，指示の明確性ラベル y_{label} を予測するロジスティック回帰分類器を学習させる．

具体的には，文献 [8] に倣い以下の 4 つの位置（トークン）における隠れ状態を使用する．

- `prompt_end`: 入力プロンプトの末尾
- `thinking_end`: CoT 生成終了時（CoT ありの場合のみ）
- `plan_start`: 実行プラン生成の直前
- `output_end`: 生成終了時

学習データを用いて分類器を学習し，検証データによりハイパーパラメータを最適化する．その後，決定されたハイパーパラメータを用いて学習データおよび検証データを統合して再学習を行い，テストデータを用いて曖昧性判定精度を評価する．

5 実験

5.1 実験設定

本研究では，指示の曖昧性を評価するベンチマークとして，3 次元ブロック構築環境における対話タスクである IGLU 2022 データセット [11] を利用した．本章では実験設定の概要を述べ，詳細なデータセット構成，モデル仕様，およびプロンプトについては付録 A に記載する．

データセット． 公開されている IGLU 2022 の Singleturn 学習用データをベースに，タスク遂行評価（ブロック配置の成否）と曖昧性判定（指示の曖昧さラベル）を同時に検証可能な統合データセットを構築した．データセットは全体を通して指示が明確（Clear）な事例が大半を占めており，曖昧（Ambiguous）な事例は全体の約 9.4%の不均衡な構成である．本実験では，これを分割し作成したテストデータを用いて評価を行った．

モデルとプロンプト． Qwen 3（4B, 8B, 14B）および Gemma-3（4B, 12B）シリーズの Instruction Tuning 済みモデルを用いた．各モデルに対し，指示の曖昧性を示唆する「ヒント」の有無，および CoT の有無，Few-shot 提示の有無を操作した複数の条件下でプロンプトの比較を行った．

表 2 ヒントあり・CoT あり・Few-shot 条件における，テキスト出力と内部状態（Linear Probing）の曖昧性判定性能（Macro F1）の比較．Probing は同設定内で最も性能が高かった層・位置の値を採用．

Model	Text F1	Probing F1	Best Layer Info
Qwen3-4B	0.511	**0.606**	layer 34 (prompt_end)
Qwen3-8B	0.588	**0.625**	layer 35 (prompt_end)
Qwen3-14B	0.624	**0.625**	layer 22 (prompt_end)
Gemma-3-4B	0.533	**0.584**	layer 27 (prompt_end)
Gemma-3-12B	0.598	**0.621**	layer 30 (output_end)

5.2 結果: テキスト出力における性能

プロンプトを用いた応答生成の結果，CoT や Few-shot の導入は，明確な指示に対するタスク成功率（Exact Match）を向上させた．一方で，指示の曖昧性判定（Macro F1）については，CoT や Few-shot による一貫した改善効果は見られず，曖昧性を示唆する「ヒント」を与えない条件下ではランダム推測（≈ 0.5）と同等の精度に留まった．ヒントの付与は判定精度を全体的に底上げするものの，そこでも CoT 等のプロンプト工夫による追加的な改善効果は限定的であった（表 1）．

5.3 結果: 内部状態における曖昧性の表現

指示曖昧性が内部表現に符号化されている． 内部表現を入力とした Probe は，ほとんどのモデルにおいてランダム推測を上回る精度（Macro F1 > 0.6）で曖昧性を識別できていることが明らかになった（表 2）．

また，層ごとの詳細な分析からは，中間層から深層にかけて識別性能がピークに達する傾向が見られた（図 2）．加えて，最終層では低下する傾向が観

表 3 Qwen3-14B（ヒントなし条件）における，CoT の有無が「タスク成功率（明確な指示）」と「内部状態による曖昧性識別精度」に与える影響の比較．CoT はタスク遂行能力（EM）を大幅に改善するが，内部の曖昧性検知（Probing F1）には寄与しない（あるいは低下させる）乖離が見られる．

	Task Success (EM)		Internal Probing (F1)	
Setting	w/o CoT	w/ CoT	w/o CoT	w/ CoT
Zero-shot	0.237	**0.514**	**0.651**	0.608
Few-shot	0.293	**0.534**	0.648	**0.659**

察された．これは，出力生成時に他の概念が優勢になっている可能性を示唆している．

テキスト出力との乖離． プロンプトに曖昧性を示唆するヒントを与える条件下においても，LLM の生成テキストから曖昧性判定する手法と比べ，Probe による曖昧性の識別は同等以上の精度となることが明らかになった（表 2）．これは，LLM が内部で曖昧性を表現していても，生成された応答にはその情報が十分に顕在化していない，あるいは活用しきれていないことを示唆している．この要因として，ユーザーの期待に応えようとする Sycophancy Bias [9] の影響により，モデルが内部で曖昧性を表現していながらも，「実行可能」な振る舞いを優先して生成を行っている可能性が示唆される．

CoT は曖昧性の内部表現を改善しない． CoT がタスク成功率を向上させつつも，内部の識別精度には影響を与えなかった点について考察する． 表 3 に示す通り，CoT の導入により Qwen3-14B のタスク成功率（EM）は大幅に向上したが，内部状態による識別精度（Probing F1）は改善せず，むしろ僅かに低下する傾向が見られた． この結果は，CoT がモデル内部における「曖昧性の符号化」そのものを強化しているわけではないことを示唆している．これは，推論能力の強化と内部の曖昧性表現が連動しない可能性を示しており，エージェントの制御においてはこれらを区別して扱う必要性を示している．LLM の内部状態は多義的な情報が重ね合わせとして表現されており [12, 10]，指示の曖昧性もまた，このような基底的な概念の一つとして，プロンプトに依らず指示から直接形成される可能性が高い．

6 結論

本研究では，Linear Probing を用いて LLM における指示曖昧性の表現を解析した．実験の結果，モデル内部の曖昧性表現がテキスト出力に十分に反映されていない可能性が示唆された．具体的には，曖昧性は中間層から深層にかけて線形分離可能な形で符号化されており，そのピークでの識別精度はテキスト出力と同等かそれ以上であった．さらに，CoT はタスク遂行能力を改善する一方で，内部の曖昧性識別精度には寄与しないことが明らかになった．これは，推論プロセスを強化するプロンプトの工夫と，モデル内部の曖昧性表現は必ずしも連動しない可能性を示唆している．

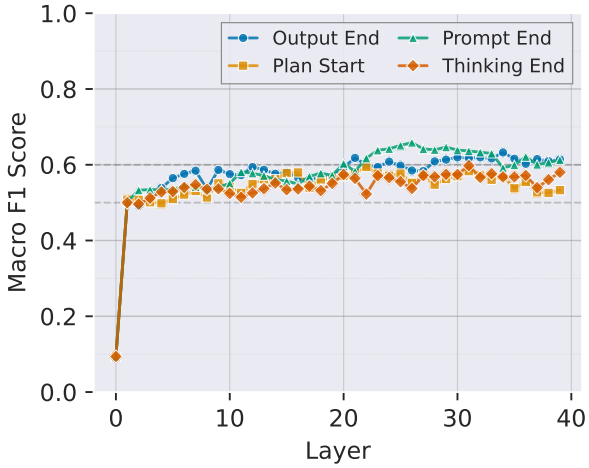

図 2 Qwen3-14B における層ごとの Probing 性能（Macro F1）の推移（ヒントなし・CoT あり・Few-shot 条件）．破線はランダムベースライン（F1=0.5）を示す．中間層から深層にかけて性能が向上していることが確認できる．

本研究はエージェントの指示曖昧性を内部表現レベルで初めて解析し，安全性評価の新たな視点を提供するものである．

限界． 本研究の限界として，データの不均衡や特定のタスク環境（IGLU）への依存が挙げられ，今後はより広範な検証が求められる．

また，分析手法である Linear Probing は相関関係の提示に留まり，因果関係の特定には至っておらず，ショートカット学習の可能性も完全には排除できていない．曖昧性の判定に関わるメカニズムを厳密に解明するには，今後 Steering [15] 等を用いた因果介入による検証が必要である．

さらに，本手法による曖昧性判定の精度（Macro F1 最大 0.651）は，Fine-tuning を用いた先行研究（BERT: 0.732 [2], LLaMA-2: 0.818 [5]）には及ばない．しかし，本研究の目的は最高性能の追求ではなく，曖昧な指示がモデル内部でいかに処理されるかを理解することにある．

謝辞

本研究は，JST 創発的研究支援事業 JPMJFR232K，および中島記念国際交流財団の助成を受けたものです．また，「ABCI 3.0 開発加速利用」の支援を受けて産総研及び AIST Solutions が提供する ABCI 3.0 を利用して得られたものです．

参考文献

[1] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wensen Cheng, Qi Zhang, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huang, and Tao Gui. The rise and potential of large language model based agents: A survey. **arXiv [cs.AI]**, September 2023.

[2] Shrestha Mohanty, Negar Arabzadeh, Julia Kiseleva, Artem Zholus, Milagro Teruel, Ahmed Awadallah, Yuxuan Sun, Kavya Srinet, and Arthur Szlam. Transforming human-centered AI collaboration: Redefining embodied agents capabilities through interactive grounded language instructions. **arXiv [cs.AI]**, May 2023.

[3] Nikhil Mehta, Milagro Teruel, Xin Deng, Sergio Figueroa Sanz, Ahmed Awadallah, and Julia Kiseleva. Improving grounded language understanding in a collaborative environment by interacting with agents through help feedback. In **Findings of the Association for Computational Linguistics: EACL 2024**, pp. 1306–1321, 2024.

[4] Kata Naszadi, Putra Manggala, and Christof Monz. Aligning predictive uncertainty with clarification questions in grounded dialog. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Findings of the Association for Computational Linguistics: EMNLP 2023**, pp. 14988–14998, Singapore, December 2023. Association for Computational Linguistics.

[5] C D Hromei, Daniele Margiotta, D Croce, and Roberto Basili. MM-IGLU: Multi-modal interactive grounded language understanding. **LREC**, pp. 11440–11451, 2024.

[6] Sanidhya Vijayvargiya, Xuhui Zhou, Akhila Yerukola, Maarten Sap, and Graham Neubig. Interactive agents to overcome ambiguity in software engineering. **arXiv [cs.AI]**, February 2025.

[7] Javier Ferrando, Gabriele Sarti, Arianna Bisazza, and Marta R Costa-jussà. A primer on the inner workings of transformer-based language models. **arXiv [cs.CL]**, April 2024.

[8] Aviv Slobodkin, Omer Goldman, Avi Caciularu, Ido Dagan, and Shauli Ravfogel. The curious case of hallucinatory (un)answerability: Finding truths in the hidden states of over-confident large language models. In **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 3607–3625, Stroudsburg, PA, USA, December 2023. Association for Computational Linguistics.

[9] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, Shauna M Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In **The Twelfth International Conference on Learning Representations**, October 2023.

[10] Hakaze Cho, Haolin Yang, Gouki Minegishi, and Naoya Inoue. Mechanism of task-oriented information removal in in-context learning. **arXiv [cs.LG]**, November 2025.

[11] Julia Kiseleva, Alexey Skrynnik, Artem Zholus, Shrestha Mohanty, Negar Arabzadeh, Marc-Alexandre Côté, Mohammad Aliannejadi, Milagro Teruel, Ziming Li, Mikhail Burtsev, Maartje ter Hoeve, Zoya Volovikova, Aleksandr Panov, Yuxuan Sun, Kavya Srinet, Arthur Szlam, and Ahmed Awadallah. IGLU 2022: Interactive grounded language understanding in a collaborative environment at NeurIPS 2022. **arXiv [cs.CL]**, May 2022.

[12] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. **arXiv [cs.LG]**, September 2022.

[13] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. https://openreview.net/forum?id=HJ4-rAVtl, February 2017. Accessed: 2025-12-25.

[14] Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In **Findings of the Association for Computational Linguistics: EMNLP 2023**, pp. 967–976, Stroudsburg, PA, USA, December 2023. Association for Computational Linguistics.

[15] Zhuoxuan Zhang, Jinhao Duan, Edward Kim, and Kaidi Xu. Sparse neurons carry strong signals of question ambiguity in LLMs. In **Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing**, pp. 16092–16110, Stroudsburg, PA, USA, November 2025. Association for Computational Linguistics.

[16] Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. AmbigQA: Answering ambiguous open-domain questions. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 5783–5797, Stroudsburg, PA, USA, November 2020. Association for Computational Linguistics.

[17] Dmitrii Krasheninnikov, Egor Krasheninnikov, and David Krueger. Assistance with large language models. In **NeurIPS ML Safety Workshop**, November 2022.

[18] Jeremy Cole, Michael Zhang, Daniel Gillick, Julian Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. Selectively answering ambiguous questions. In **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 530–543, Stroudsburg, PA, USA, December 2023. Association for Computational Linguistics.

[19] Tong Zhang, Peixin Qin, Yang Deng, Chen Huang, Wenqiang Lei, Junhong Liu, Dingnan Jin, Hongru Liang, and Tat-Seng Chua. CLAMBER: A benchmark of identifying and clarifying ambiguous information needs in large language models. In **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 10746–10766, Stroudsburg, PA, USA, 2024. Association for Computational Linguistics.

A　詳細な実験設定

A.1　データセットの構築と詳細

公開されている IGLU 2022 の Singleturn 学習用データセットをベースに，タスク遂行評価と曖昧性判定を同時に行うための統合データセットを構築した．具体的には，対話履歴，直前の指示，現在のブロックの配置，タスク完了時の目標配置，および指示の曖昧性ラベルが同時に利用できるよう結合した．これにより，同一の入力事例に対して「ブロック配置タスクの成否」と「指示の曖昧性判定」の双方を検証可能なデータセットとしている．

本実験で使用したデータセットの内訳を表 4 に示す．データは学習（Train），検証（Val），評価（Test）に分割されている．

表 4　本実験で使用したデータセットの内訳．

Split	Clear (Yes)	Ambiguous (No)	Total
Train	4,227	426	4,653
Val	666	75	741
Test	625	72	697
Total	**5,518**	**573**	**6,091**

A.2　モデルとプロンプトの詳細

本実験で使用したモデルの HuggingFace における ID を表 5 に示す．なお，計算資源の制約により，Qwen 3 では non-thinking mode で推論を行った．

表 5　使用したモデルの HuggingFace ID 一覧

Model Family	HuggingFace Model ID
Qwen 3	`Qwen/Qwen3-4B`
	`Qwen/Qwen3-8B`
	`Qwen/Qwen3-14B`
Gemma 3	`google/gemma-3-4b-it`
	`google/gemma-3-12b-it`

本タスクにおいて，モデルはタスクに関する知識や現在のブロックの配置，対話履歴とユーザの指示の入力に対して，ブロックを操作するコマンド（`add_block`, `remove_block`, `replace_block`）およびそのコマンドに対応する座標とブロックの色の生成を求める．

その上で，特にヒントありの条件下では，プロンプトに「もし指示が曖昧で実行不可能な場合は，コマンドの代わりに"AMBIGUOUS"と出力せよ」という明示的な指示を含め，Few-shot の場合は実際に曖昧な事例（色が特定できない場合など）とそれに対する"AMBIGUOUS"出力のペアを例示として与えている．

A.3　分析手法の詳細

行動評価

- **タスク成功率 (Exact Match; EM):** 明確な指示（Clear Instructions）に対するタスク遂行能力を評価する．モデルが生成した操作コマンド列を初期状態（Current Grid）に順次適用し，最終的なブロック配置が目標状態（Target Grid）と一致するかを判定する．判定に際しては IGLU ベンチマークの評価基準に準拠し，グリッドの平行移動および垂直軸周りの回転を許容した上で，形状と色が完全に一致した場合を成功とみなす．
- **曖昧性判定 (Macro F1):** 指示の曖昧性を正しく識別できたかを評価するため，曖昧性判定を 2 値分類問題として扱い，Macro F1 スコアを算出する．曖昧性の判定（Ambiguous ラベルの付与）は，以下のいずれかの条件を満たした場合とした．
  1. **キーワード検知**: モデルの生成テキスト（CoT 等の推論過程を除いた最終出力部分）を小文字化した上で，以下のキーワードが含まれる場合．
     `"ambiguous"`, `"ambiguity"`, `"unclear"`, `"not clear"`, `"cannot determine"`, `"unable to determine"`, `"not specified"`, `"insufficient information"`, `"not enough info"`, `"unknown"`, `"please clarify"`, `"need clarification"`, `"no command"`
  2. **プランの欠如**: `add_block` 等の有効な操作コマンドが一切生成されなかった場合．

内部分析ロジスティック回帰分類器の学習においては，学習および検証データを用いてハイパーパラメータ（正則化の強さ）の調整を行った．具体的には，検証データにおける分類性能（Macro F1）が最大となる設定を選定し，その最適化されたモデルを用いてテストデータに対する評価を行った．また，学習データのクラス不均衡に対処するため，各クラスの出現頻度に反比例した重みを損失関数に適用した（scikit-learn の `class_weight='balanced'` 設定を使用）．