

AI エージェントのためのツール呼び出し評価データ JMultiWOZ-TC の構築

清水綾太^{1,3} 中山功太³ 守山慧^{2,3} 徳久良子^{1,4} 宮尾祐介^{2,3}
¹ 愛知工業大学 ² 東京大学 ³ NII LLMC ⁴ 理化学研究所
 k22066kk@aitech.ac.jp

概要

近年、大規模言語モデルの AI エージェントとしての活用が注目されている。本稿では、AI エージェントのツール呼び出し性能を評価するために、日本語データセット **JMultiWOZ-TC** を構築した。本データセットは、マルチドメイン対話データである JMultiWOZ を AI エージェントのツール呼び出し評価用に再構成したもので、4,246 対話のユーザ発話 31,303 発話に対して、合計 16,510 個のツール呼び出しが付与されている。既存の LLM (Llama3.1-70B, Qwen3-32B, Swallow-70B) のツール呼び出し能力を評価し、誤り傾向を分類した結果、モデルごとに誤りの発生要因や傾向に特徴があることが分かった。

1 はじめに

近年、大規模言語モデル (LLM) を AI エージェントとして活用する動きが広がっている。AI エージェントとは、LLM を中核として構築され、外部環境とのインタラクションを通じて情報を取得・更新しながら、与えられた目的に従って推論・意思決定を行い、自律的に行動するソフトウェアである。AI エージェントには、目的を遂行するための計画立案能力、目的達成に必要な外部ツールを呼び出す能力、環境からの応答を踏まえて自身の行動を内省し次の行動へとつなげる能力、行動全体に関する情報を適切に記憶・活用する能力などが求められる。

本稿では、日本語による指示が与えられる環境において、LLM が目的の達成に必要な外部ツールを適切に呼び出せるかを評価するためのデータセットを構築する。図 1 に、LLM を用いたツール呼び出しの例を示す。この例では、 t ターン目のユーザ発話「wifi が使えるところで」に対して、LLM が Search_hotel というツールを呼び出している。

本稿の主な貢献は以下の 2 点である。

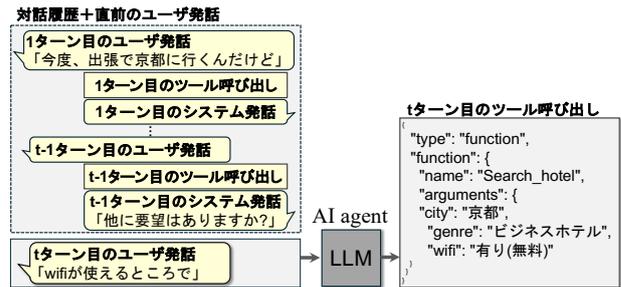


図 1 LLM を用いたツール呼び出しの例

- 1) マルチドメインタスク指向型対話のデータセットである JMultiWOZ (詳細は後述) を AI エージェントのツール呼び出し評価に適した形に再構成した **JMultiWOZ-TC** (Japanese Multi-Domain Wizard-of-Oz Tool Call) を構築する¹⁾。
- 2) JMultiWOZ-TC を用いて、既存の LLM におけるツール呼び出しの生成正答率を評価する。

2 関連研究

AI エージェント用ベンチマークデータセット

AI エージェントを評価する方法は多岐にわたるが、大きく分けて、特定のタスクの達成度をエンドツーエンドで評価する方法と、各能力を個別に評価する方法の二つに分類できる。前者を評価するベンチマークとしては、計算機上にあらかじめ用意された仮想環境において、AI エージェントがユーザの指示を正しく遂行できるかを評価する AgentBench [1] や、画像を含む環境を対象としたベンチマークである VisualAgentBench [2] がある。また、Ubuntu や Windows などの OS 上でエージェントの振る舞いを評価する OSWorld [3] が提案されており、日本語翻訳版である OSWorld-JP [4] も公開されている。後者を評価するベンチマークも多数提案されているが、ここでは本研究が対象とするツール呼び出しに関

1) ツール呼び出し評価データ JMultiWOZ-TC は右記から利用できる：<https://github.com/llm-jp/JMultiWOZ-TC>

するデータセットに限り紹介する。ToolBench [5] は、実在する API を用いたベンチマークである。しかし、API の停止や非決定性により、再現性の確保が難しい。この課題を緩和するため、API 応答を事前にキャッシュしたり LLM が API の挙動を模擬する StableToolBench [6] が提案されている。マルチタスクおよびマルチターン設定を対象とする BFCL [7] では、ツール呼び出し結果を抽象構文木として表現し、正解と比較することにより、実際のツール実行を伴わない評価を実現している。Nejumi LLM リーダーボード²⁾ では、日本語に翻訳された BFCL を評価に用いている。しかし、既存の日本語版ベンチマークの多くは英語ベンチマークの翻訳に留まっており、日本における実利用環境を想定して設計された AI エージェント向けベンチマークは存在しない。

マルチドメインタスク指向型対話 AI エージェント用のベンチマークデータセットは整備が始まったばかりであるが、外部ツールを利用しながら対話を行うマルチドメインタスク指向型対話のデータセットは以前から構築されてきた [8, 9, 10, 11]。例えば、JMultiWOZ [11] には、旅行タスクに関する 6 つのドメイン（観光名所、宿泊施設、飲食店、買い物施設、タクシー、天気）にまたがる日本語の対話が合計 4,246 対話収録されている。マルチドメインタスク指向型対話は、複数の外部ツールから適切なツールを選択しながらシステムとユーザが対話してタスクを達成するという点で、AI エージェントのツール呼び出しと問題設定が非常に近い。我々はこの点に着目し、マルチドメインタスク指向型対話のデータセットである JMultiWOZ を活用して、AI エージェントのツール呼び出しの生成能力を評価するためのデータセット JMultiWOZ-TC を構築する。

3 JMultiWOZ-TC の構築

図 2 に本稿で取り組む (1) ツール呼び出し評価データ JMultiWOZ-TC の構築と、(2) LLM のツール呼び出し生成の評価の概要を示す。3.1 節で LLM のツール呼び出し生成の評価データの要件を整理した上で、3.2 節で JMultiWOZ-TC の形式を説明する。

3.1 ツール呼び出し評価データの要件

LLM のツール呼び出し生成の評価では {対話履歴, ツール呼び出し履歴, ユーザからの指示³⁾} を

2) <https://nejumi.ai/>

3) 図 2 の例では「6 月 4 日 9 時半にタクシー 1 台お願いします」がユーザからの指示にあたる。

LLM に入力し、適切なツール呼び出しを生成できるかを評価する。図 2 の (2) に示す通り、AI エージェントのツール呼び出しの評価データの正解データには、各ユーザ発話において呼び出すべきツールとその引数が定義されている必要がある。

JMultiWOZ には、ユーザ発話とシステム発話に加えて、各システム発話における検索や予約などの操作情報が付与されている。しかし、JMultiWOZ は AI エージェントのツール呼び出しを意図して作られたデータセットではないため、そのままの形式ではツール呼び出しの評価には使用できない。次節では、JMultiWOZ からツール呼び出し評価データ JMultiWOZ-TC を再構成した際の変更点を説明する。

3.2 JMultiWOZ-TC への変換

ユーザ発話とツール呼び出しの対応の明確化

JMultiWOZ では、すべてのシステム発話に対して、検索や予約などの実行結果が付与されている。この実行結果は、過去のユーザ発話で実行された検索や予約結果がその後のシステム発話に累積する形で記述されている。例えば、図 2 の (1) では USR4 でタクシーを予約しているが、仮に、USR1 から USR3 のいずれかのユーザ発話でレストラン検索が実行されていた場合には、「SYS4 の OP」には USR4 で実行したタクシー予約だけでなく、レストラン検索も記述される。このため、現在のユーザ発話に対応したツール呼び出しが明確でなかった。そこで JMultiWOZ-TC では、「対話履歴, ツール呼び出し履歴, ユーザ発話」に対応する「ツール呼び出しの正解データ」を整備した。なお、ひとつのユーザ発話に対して複数のツールを実行する必要がある場合には、該当するすべてのツールを付与した。

ツール一覧の整備 JMultiWOZ には宿泊施設の検索や予約、タクシーの予約などの対話とその際のシステム操作が収録されている。JMultiWOZ-TC では、JMultiWOZ におけるシステム操作を、表 1 に示す 9 種類のツール呼び出しとして整理した。ツール説明の記述形式は、BFCL [7] の形式に準拠した。また、ツールの引数の候補が少ないものは、その候補を説明内に記述した。

時刻などの記述形式の変更 JMultiWOZ では、時刻は、「10:40」のような文字列で表現されており、評価時に解釈の曖昧さが生じる。JMultiWOZ-TC では、時刻を時と分に分けて数値化することで、評価の厳密性を高めた。また、日付についても月と日で分解

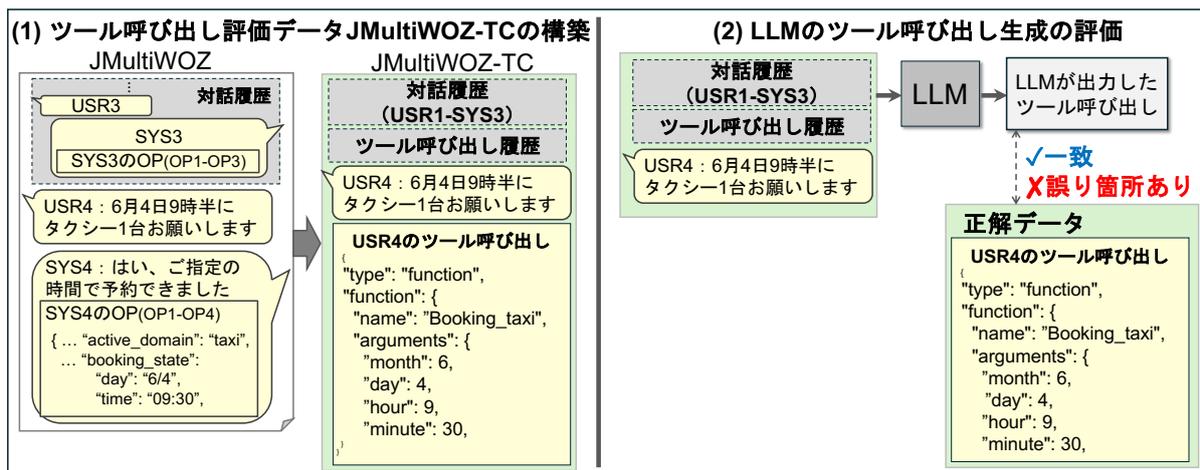


図2 ツール呼び出し評価データである JMultiWOZ-TC の構築と、LLM のツール呼び出し生成の評価の概要。図中の USR はユーザ発話、SYS はシステム発話、OP はシステムが実行した検索や予約の操作を表す。

表1 JMultiWOZ に基づき定義したツール一覧

Search_restaurant, Search_hotel, Search_attraction, Search_shopping, Search_taxi, Search_weather, Booking_restaurant, Booking_hotel, Booking_taxi

表2 既存の LLM のツール呼び出し生成正答率 (%)。

	呼び出し判定	呼び出し正答率	全体正答率
Llama	71.5	6.3	42.4
Qwen3	78.0	87.7	71.9
Swallow	62.2	29.2	60.0

するなど同様の処理を施した。

以上の整備の結果、4,246 対話に含まれるユーザ発話 31,303 発話に対して、合計 16,510 個のツール呼び出しの評価データを構築した。ユーザ発話によってはツールが呼び出されない場合や、複数のツールが呼び出される場合もある。各ユーザ発話に対する呼び出しツール数は平均 0.53 個となった。

4 実験

4.1 評価方法

JMultiWOZ-TC からランダムに 1000 個のユーザ発話を抽出し、既存の LLM のツール呼び出しの生成正答率を評価した。LLM には、「過去の対話履歴、ツール呼び出し履歴、評価対象となるユーザ発話」を入力として与え、ツール呼び出しを生成する。この際、過去の対話で呼び出したツールを再度呼び出した場合は誤りとした。また、ひとつのユーザ発話で複数のツールを呼び出す必要がある場合は、ツール呼び出しの順序が異なる場合も正解とした。

評価対象の LLM として、異なる言語特性を想定して学習された 3 種類のモデルを採用した。具体的には、英語中心の Llama (Llama-3.1-70B-Instruct⁴⁾)、中国語性能を強化した Qwen3 (Qwen3-32B⁵⁾)、お

よび前者モデルの日本語性能を強化した Swallow (Llama-3.1-Swallow-70B-Instruct-v0.3⁶⁾) である。いずれのモデルも、AI エージェントのツール呼び出しに特化した追加学習は施されておらず、既存の LLM としての性能を評価対象とした。

4.2 評価結果

表2に評価結果を示す。表2の「呼び出し判定」はツールを呼び出すか否かの判別正答率、「呼び出し正答率」はツール呼び出しが必要なユーザ発話におけるツール名と引数の生成正答率、「全体正答率」はツール呼び出しの有無を含む全発話に対するツールと引数の生成正答率を示す。

まず、ツールを呼び出すかどうかの判別能力については、Qwen3 が 78.0% と最も高い正答率を示した。一方、Llama および Swallow は、ツールを呼び出す必要がある場面においてもツールを呼び出さない傾向にあり、それぞれ 71.5%、62.2% に留まった。また、ツール呼び出しが必要な事例に対する評価では、Qwen3 が 87.7% と高い性能を示したのに対し、Llama は 6.3% と極めて低い結果となった。Llama では、ツール名は正しく生成できても引数が正しく生成できない事例が目立った。誤りの内訳や具体的な事例は 4.3 節に示す。最後に、全体正答率

4) <https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct>

5) <https://huggingface.co/Qwen/Qwen3-32B>

6) <https://huggingface.co/tokyotech-llm/Llama-3.1-Swallow-70B-Instruct-v0.3>

では、Qwen3 が 71.9% と最も高い性能を示し、次いで Swallow (60.0%), Llama (42.4%) の順となった。各 LLM のツール呼び出しの誤り傾向を明らかにするため、次節では人手で誤りの特徴を分類する。

4.3 人手による誤り事例の分類

各モデルの誤り事例をランダムに 100 件抽出し、人手により内容を分類した。誤り事例の分類結果を表 3 に、具体例を付録 A に示す。誤り事例は、誤答扱いになったがツール呼び出しの実用上で問題がないと考えられる「許容」、引数の文字生成の誤り「単語誤り」、過去の対話で取得済みの情報に対して再度ツールを呼び出す「使用重複」、ツール使用が求められる場面でツールを呼び出さない「ツール不使用」、これらに該当しない「その他」に分類した。

まず、Llama では、表 3 の「単語誤り」や「ツール不使用」に分類される誤りが多く確認された。特に、日本語の表記誤りや不自然な生成がツール呼び出し時の引数の単語誤りにつながる事例が目立った。例えば、正解データでは都市が「名古屋」、ジャンルが「シティホテル」に対し、Llama の出力は都市が「名宮」、ジャンルが「シティーホートル」と生成した（出力全体は付録 A の図 3 に示す）。この結果から、日本語生成の不安定さが引数の生成に悪影響を及ぼしていると考えられる。このような課題は、英語環境を前提としたベンチマークでは十分に評価されていなかった部分であり、日本語環境における AI エージェントの性能を適切に評価するためには日本語に特化したベンチマークが不可欠であることを示唆している。

次に、Qwen3 では、表 3 の「使用重複」に分類される誤りが 68 件であり、対話履歴で既に取得済みの情報に対して再度ツールを呼び出す事例が多く確認された。具体的には、過去の対話で天気検索ツールを呼び出して天気情報を取得しているにも関わらず、再度天気検索ツール呼び出しを生成している事例があった（付録 A の図 4）。また、Qwen3 では「許容」に分類された事例も 13 件あった。例えば、正解データはジャンルが「簡易宿所（ベッドハウス・山小屋・カプセルホテル等）」に対し、Qwen3 の出力は「簡易宿所」となった例があった（付録 A の図 5）。このような誤りは、ツール呼び出し後の検索処理などで同じカテゴリと扱われる可能性が高いため実用上は問題にならないと考えられる。許容できるツール呼び出しを自動評価で正解とするためには、自動

表 3 モデル別の誤り事例の分類結果。モデルごとに各事例に該当するエラー件数を表す。

	許容	単語誤り	使用重複	ツール不使用	その他
Llama	1	36	4	43	16
Qwen3	13	2	68	2	15
Swallow	7	3	0	84	6

評価の判定精度を高める必要があると考えられる。

最後に、Swallow では、表 3 の「ツール不使用」に分類される誤りが 84 件と大半を占めており、ツール使用が求められるユーザ発話に対しても、ツールを呼び出さない傾向が見られた。具体的には、「鶴見区で、百貨店や総合スーパーはありますかね？」というユーザ発話に対して、正解データではショッピング施設の検索ツールを呼び出すのに対し、Swallow の出力ではツールを呼び出さなかった（出力全体は付録 A の図 6 に示す）。一方で、ツールを呼び出している場合には引数の単語誤りや形式誤りはほとんど見られなかったことから、日本語を生成する能力は優れているものの、ツール呼び出しの判定能力には課題があることが示唆された。

5 おわりに

本研究では、日本語マルチドメインタスク指向型対話データセットである JMultiWOZ を AI エージェントのツール呼び出しの評価に適した形式に再構成した JMultiWOZ-TC を構築し、4,246 対話に含まれるユーザ発話 31,303 発話に対して、合計 16,510 個のツール呼び出しの評価データを構築した。

誤り分類の結果から、各モデルの性能差は単なる日本語理解能力の違いだけでなく、マルチターン対話におけるツール呼び出し判定と、ツール出力を対話全体で適切に活用する推論能力の違いに強く影響していることが明らかとなった。特に、Llama-3.1-70B-Instruct および Llama-3.1-Swallow-70B-Instruct-v0.3 では、ツールを使用すべき場面でツールを呼び出さない誤りが全体性能を大きく低下させていた。一方、Qwen3-32B はツール使用判断自体は比較的適切であるものの、対話履歴に基づく情報の再利用が不十分であり、冗長なツール呼び出しが性能低下の主要因であった。

本稿ではツール呼び出しに特化した評価データを構築したが、今後は LLM の AI エージェントとしての能力を包括的に評価するための日本語データセットの構築を進める予定である。

謝辞

本研究成果の一部は、データ活用社会創成プラットフォーム mdx を利用して得られたものです。

参考文献

- [1] Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. Agentbench: Evaluating LLMs as agents. In **ICLR**, 2024.
- [2] Xiao Liu, Tianjie Zhang, Yu Gu, Iat Long Iong, Song XiXuan, Yifan Xu, Shudan Zhang, Hanyu Lai, Jiadai Sun, Xinyue Yang, Yu Yang, Zehan Qi, Shuntian Yao, Xueqiao Sun, Siyi Cheng, Qinkai Zheng, Hao Yu, Hanchen Zhang, Wenyi Hong, Ming Ding, Lihang Pan, Xiaotao Gu, Aohan Zeng, Zhengxiao Du, Chan Hee Song, Yu Su, Yuxiao Dong, and Jie Tang. Visualagentbench: Towards large multimodal models as visual foundation agents. In **ICLR**, 2025.
- [3] Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments, 2024.
- [4] 庄司文武, 吉田雄紀. Osworld-jp: 言語を考慮した評価のための日本語版コンピュータユースベンチマーク. 人工知能学会全国大会論文集, Vol. JSAI2025, pp. 3Win556–3Win556, 2025.
- [5] Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. Tool-llm: Facilitating large language models to master 16000+ real-world apis, 2023.
- [6] Zhicheng Guo, Sijie Cheng, Hao Wang, Shihao Liang, Yujia Qin, Peng Li, Zhiyuan Liu, Maosong Sun, and Yang Liu. StableToolBench: Towards stable large-scale benchmarking on tool learning of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Findings of the Association for Computational Linguistics: ACL 2024**, pp. 11143–11156, Bangkok, Thailand, August 2024. Association for Computational Linguistics.
- [7] Shishir G Patil, Huanzhi Mao, Fanjia Yan, Charlie Cheng-Jie Ji, Vishnu Suresh, Ion Stoica, and Joseph E. Gonzalez. The berkeley function calling leaderboard (BFCL): From tool use to agentic evaluation of large language models. In **Forty-second International Conference on Machine Learning**, 2025.
- [8] Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In **EMNLP**, pp. 5016–5026, 2018.
- [9] Ting Han, Ximing Liu, Ryuichi Takanabu, Yixin Lian, Chongxuan Huang, Dazhen Wan, Wei Peng, and Minlie Huang. Multiwoz 2.3: A multi-domain task-oriented dialogue dataset enhanced with annotation corrections and co-reference annotation. p. 206–218, 2021.
- [10] Qi Zhu, Kaili Huang, Zheng Zhang, Xiaoyan Zhu, and Minlie Huang. CrossWOZ: A large-scale Chinese cross-domain task-oriented dialogue dataset. **TACL**, Vol. 8, pp. 281–295, 2020.
- [11] Atsumoto Ohashi, Ryu Hirai, Shinya Iizuka, and Ryuichiro Higashinaka. JMultiWOZ: A large-scale Japanese multi-domain task-oriented dialogue dataset. In **LREC-COLING**, pp. 9554–9567, 2024.

付録

A ツール呼び出しの誤り事例

既存の LLM のツール呼び出しの誤り事例の具体例を図 3, 図 4, 図 5, 図 6 に示す. LLM の出力の誤りに該当する部分とそれに対応する正解データの部分をそれぞれ赤字と青字で示す.

直前のユーザ発話
名古屋の旅行の時に泊まるシティホテルを探しています。

正解データ	LLMの出力 (Llama)
<pre>"name": "Search_hotel", "arguments": { "city": "名古屋", "genre": "シティホテル" }</pre>	<pre>"name": "Search_hotel", "arguments": { "city": "名古屋", "genre": "シティーホートル" }</pre>

図 3 「単語誤り」ラベルが付与された例. ツールの引数の単語の生成が間違っている場合に「単語誤り」を付与した.

対話履歴
ユーザ発話: ありがとうございます。ところで、1/14の杉並区の天気はどうですか?
アシスタント発話: 晴れ時々曇りの予報となっております。

ツール呼び出し履歴
過去のツール呼び出し: 過去のツールからの出力

<pre>"name": "Search_weather", "arguments": { "city": "東京", "area": "杉並区", "month": 1, "day": 14 }</pre>	<pre>"active_entity": { "city": "東京", "area": "杉並区", "day": "1/14", "name": "1/14 東京杉並区の天気", "weather": "晴時々曇", "mintemperature": 3.2, "maxtemperature": 7.3, "domain": "weather" }</pre>
--	---

直前のユーザ発話
最低気温はどうでしょう?

正解データ	LLMの出力 (Qwen3)
<pre>{}</pre>	<pre>"name": "Search_weather", "arguments": { "city": "東京", "area": "杉並区", "month": 1, "day": 14 }</pre>

図 4 「使用重複」ラベルが付与された例. 「ツール呼び出し履歴」で実行済みのツール呼び出しを再度出力している場合に「使用重複」のラベルを付与した. この例では, 過去の対話履歴ですでに呼び出したツール(Search_weather)を再度呼び出している.

直前のユーザ発話
あと真和志地区でWi-Fiが無料で利用できる簡易宿所を探していますがありますか?

正解データ	LLMの出力 (Qwen3)
<pre>"name": "Search_hotel", "arguments": { "city": "那覇", "area": "真和志", "genre": "簡易宿所(ベッドハウス・山小屋・カプセルホテル等)", "wifi": "有り(無料)" }</pre>	<pre>"name": "Search_hotel", "arguments": { "city": "那覇", "area": "真和志", "genre": "簡易宿所", "wifi": "有り(無料)" }</pre>

図 5 「許容」ラベルが付与された例. この例のように, LLM の出力の記述方法が異なってもツール呼び出し後の検索処理などで吸収できる場合は「許容」ラベルを付与した.

直前のユーザ発話
鶴見区で、百貨店や総合スーパーはありますか?

正解データ	LLMの出力 (Swallow)
<pre>"name": "Search_shopping", "arguments": { "city": "横浜", "area": "鶴見区", "genre": "百貨店や総合スーパー" }</pre>	呼び出していない

図 6 「ツール不使用」ラベルが付与された例. 本来はツールを呼び出すべき発話でツール呼び出しが実行できていない場合に「ツール不使用」を付与した.