マルチモーダルかつ長い文脈の処理が求められる語用論的推論ベンチマーク

佐藤 拓真 [1,2]　吉野幸一郎 [3,2,1]
[1] 奈良先端科学技術大学院大学
[2] 理化学研究所ガーディアンロボットプロジェクト　[3] 東京科学大学
sato.takuma.sq6@naist.ac.jp

概要

言語の文脈依存の意味を理解するための語用論的推論能力は、人間のみならず言語モデルや視覚言語モデルにとっても重要である。ベンチマークはモデルの語用論的推論能力を向上させる研究の基盤となるが、既存のベンチマーク群には文脈の長さ・複雑さやモーダルの単一性、高性能なモデルの評価における難度の不十分さ等の問題がある。本研究では漫画をコーパスとすることで、これらの課題に取り組む語用論的推論ベンチマークを構築した。構築したベンチマークで VLLM と人間の語用論的推論能力を比較評価した結果、最先端の VLLM でも人間の正解率に劣ることが示された。

1　はじめに

自然言語処理 (NLP) や人工知能 (AI) の研究において、言語の意味を理解できるシステムの実現は重要な目標の一つである。言語は「文字通りの意味」だけでなく「文脈依存の意味」をしばしばもち、人間の言語運用やコミュニケーションにおいては、こうした文脈依存の意味が重要な役割を果たす。文脈依存の意味を取り扱う言語学の領域は「語用論」と呼ばれる [1, 2]。言語モデルを中心とする AI 技術の発展と社会浸透を目指すうえで、モデルの語用論的推論能力の向上は重要な課題である [3]。

言語モデルの語用論的能力を改善するためには、その時点で存在しているモデルの能力を正確かつ詳細に把握する必要がある。そのため、NLP 領域において語用論的意味理解・推論についてのデータセットやベンチマークはこれまでも提案されてきた (§2)。しかしながら既存のベンチマークには、タスクにおいて考慮すべき文脈が短い対話に限定されている、回答に不要な情報が文脈として含まれない、

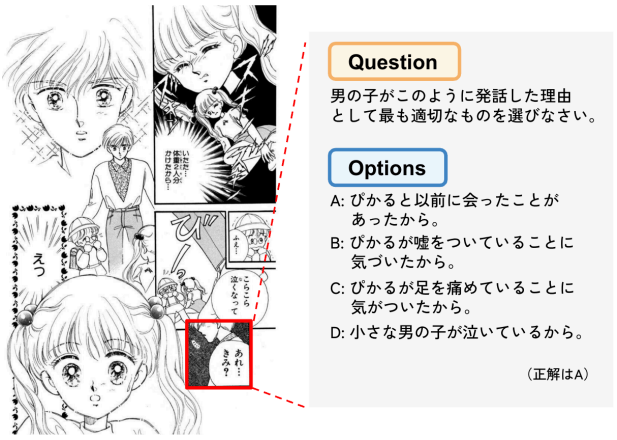

図 1　構築したベンチマークの概要。例は、漫画内の発話の理由を問うカテゴリの問題。© 栗城祥子

難度が最新のモデルの評価には不十分である、扱うモーダルが多くの場合テキストのみであるといった課題がある。今後、大規模言語モデル (LLM) や大規模視覚言語モデル (VLLM) の語用論的推論能力の向上を目指す研究のために、これらの課題を解決したベンチマークが必要である。

本研究では、上記の課題を解決する**マルチモーダルかつ長い文脈の処理を要する語用論的推論ベンチマーク**を構築した (§3)。コーパスとして、権利処理がなされた漫画データ [4, 5] を採用し、見開き 106 ページの漫画についての 4 択の多肢選択式問題を、計 101 問人手で作成した。各設問は 7 つのカテゴリに分類し、各選択肢は一貫性のある方法で作成した。構築したベンチマークの概要を図 1 に示す。

また、構築したベンチマークを用いて、最先端の VLLM と人間の正解率の評価実験を行った (§4)。実験の結果、最先端の VLLM であっても人間の被験者と比較して低い正解率しか達成することができず、現状の VLLM の語用論的推論能力には依然として課題が残ることが分かった。

2 関連研究

語用論的推論能力は我々の日常的な言語運用において重要であるため [1, 6]、LLM にもこの能力を高い程度に保有させる努力がなされてきた。先行研究の成果として、supervised fine-tuning や preference optimization を含む事後学習 [7, 8, 9] や、Chain-of-Thought (CoT)[10, 11]、few-shot learning といった手法が、LLM の語用論的推論能力を向上させることが確認されている。また、グライス語用論 [1] や関連性理論 [12] といった語用論理論の概要をプロンプトとして提示することで、モデルの語用論的推論能力が向上することも報告されている [13]。

ベンチマークは、LLM や VLLM の語用論的推論能力の向上を目指す研究において基盤的な役割を果たす。研究者は、モデルがベンチマークにおいて達成するスコアを分析することで、その時点でのモデルのできることとできないことを把握し、モデル改善のための施策を打つことができる [14]。語用論的推論の文脈においても、多くのベンチマークやデータセットが提案されてきた [3]。タスク形式としては、二値分類 [8, 11, 15] や多肢選択式 [16, 17, 18] が中心的であり、自然言語推論 [19] や質問応答 [20] の形式をとるものや、様々な形式のタスクを包括的に含むベンチマークセットも存在する [9]。

しかしながら、既存のベンチマーク群には以下に列挙するような問題が存在する。本研究は、我々の知る限り、語用論的推論ベンチマークとしてこれらの課題の全てに取り組む初の試みである。

- タスクにおいて考慮すべき文脈の量が小さい (1-数ターンの会話に留まっている)。
- 文脈中に「最終的な回答に必要な文脈」以外のノイズとなる情報が殆ど含まれていない。
- 文脈となるモーダルが 1 つのみ (多くはテキストのみ、一部では画像のみ [21]) である。
- タスクとしての難度が十分でなく、既存モデルが人間と同等以上の性能を発揮するなど、ベンチマークとして飽和している [13]。

また、本研究と同コーパスを使用した類似ベンチマークとして、一般的な漫画理解を問う MangaVQA [22] が存在する。しかしながら、当該研究は漫画中のテキストとして明示的に書かれている内容を問うベンチマークであるのに対し、本研究は明示されてはいないが言外に読み取ることができる内容を問うことで、語用論的推論能力を測定することを目的としている点が異なる。加えて、MangaVQA は見開き 1 ページをモデルへ文脈として入力するのに対し、我々のベンチマークは最長で見開き 59 ページを入力する。つまり、極めて長い文脈の関わる語用論的推論に焦点を当てている点も、当該研究と本研究のスコープの差異である。

3 ベンチマークの構築

3.1 使用コーパス

コーパスとして、日本の商業漫画 109 冊を収録するデータセットである Manga109 [4, 5] を使用した。Manga109 では、1970 年代から 2010 年代の日本語の漫画画像に対して、コマや吹き出しの位置、テキスト等のアノテーションが付与されている。収録された漫画は、コマの加工を含めて、非営利かつ学術目的に限った使用を作者から許諾されている。

3.2 ベンチマーク構築の方法

設問の作成方法　漫画の吹き出し (発話) を問題付与の対象単位として、表 1 のようなカテゴリ分けのもとで設問を本稿の第一著者が人手で作成した[1)]。設問カテゴリは、語用論的推論データセット構築の先行研究 [9, 15, 16] を参照しつつ、パイロットアノテーションを行いながら決定した。

選択肢の作成方法　1 つの正解肢と 3 つの誤答肢を、以下のように人手で作成した。選択肢の作成者も、本稿の第一著者のみである。

正解肢　設問に対する正しい回答。

誤答肢 1 (hard)　漫画なしで見れば問いに対する答えとして尤もらしいが、漫画の文脈を考慮すれば誤りであると判断できる選択肢。かつ、漫画中に登場するエンティティ[2)]や、起こっているものごとに言及しているもの。

誤答肢 2 (medium)　漫画なしで見れば問いに対する答えとして尤もらしいが、漫画の文脈を考慮すれば誤りであると判断できる選択肢。かつ、漫画中に登場しないエンティティや、起こっていないものごとに言及しているもの。

誤答肢 3 (easy)　漫画なしで見ても、問いに対する

1)　REASON と INTENT の違いについて。例えば、A さんの「ふざけるな」という発話に関して、「B さんが A さんを侮辱した」は理由ではあるが意図ではない。反対に、「B さんを非難すること」は同じ発話の意図ではあるが理由ではない。

2)　例えばキャラクター、建造物、持ち物など。

表 1　設問カテゴリの定義と問題例。

カテゴリ	定義	例
REASON	対象吹き出しにおける発話・行為・現象の理由として正しいものを選択する。	男の子がこのように発話している理由として、最も適切なものはどれですか？
INTENT	対象吹き出しにおける発話・行為の意図・目的として正しいものを選択する。	彩ちゃんがこのようにぴかるに言った目的として、最も適切なものを 1 つ選びなさい。
FEELING	対象吹き出しの場面において、明示されていない心情の説明として正しいものを選ぶ。	この場面におけるぴかるの心情として最も適切なものを選びなさい。
REFERENCE	対象吹き出しにおける指示詞の対象として正しいものを選ぶ。	「あれ」の指すものとして最も適切なものを選びなさい。
ELLIPSIS	対象吹き出しにおいて生じている格省略の補完として正しいものを選ぶ。	なにが「だんだん遠くなってる」のですか？
INDIRECT	対象吹き出しにおける修辞的・婉曲的な発話の直截的な意味として正しいものを選ぶ。	この発話の直截的な意味として、最も適切なものを選びなさい。
IMPLICATURE	対象吹き出しの発話やその場面が暗示することとして正しいものを選ぶ。	この発話が暗示することとして最も適切なものを選びなさい。

表 2　構築したベンチマークの統計。

		作品 1	作品 2	合計
対象ページ数 (見開き)		47	59	106
含まれる発話数		591	811	1,402
問題数	REASON	17	11	28
	INTENT	5	4	9
	FEELING	13	5	18
	REFERENCE	5	6	11
	ELLIPSIS	5	13	28
	INDIRECT	5	5	10
	IMPLICATURE	1	6	7
	合計	**51**	**50**	**101**

答えとして尤もらしさが低く、誤答肢であると判断しうる選択肢。

3.3　構築したベンチマークの統計

『ぴかる★元気です！』(作品 1、© 栗城祥子) と『学園ノイズ』(作品 2、© オオシマヒロユキ) の 2 作品をコーパスとしてベンチマークを構築した。各作品のジャンルは、恋愛とバトルである。コーパスとベンチマークの統計を表 2 に示す。また、作成した各設問と選択肢の長さを付録の表 5 に示す。

4　VLLM の評価実験

4.1　実験設定

出題形式　3.2 節で作成した設問と選択肢を、VLLM に入力した。プロンプトを付録 §B.1 に示す。推論時の VLLM の `temperature` は 0、`top_p` は 1、`top_k` は 0、`seed` は 42 に設定した。なお、選択肢の提示順はシャッフルした。

表 3　実験結果。w/ manga は漫画画像を文脈として与えた場合、w/o manga は与えなかった場合の正解率 (%)。All 列は、作品 1/2 の正解率を問題数で重み付けしたうえで平均を計算した値。Human は人間が解いた際の正解率。

Model		作品 1	作品 2	All
Qwen3-32B	w/ manga	50.98	60.00	**55.45**
	w/o manga	49.02	34.00	**41.58**
Llama-4-17Bx16E	w/ manga	41.18	42.00	**41.59**
	w/o manga	41.18	26.00	**33.67**
Llama-4-18Bx128E	w/ manga	47.06	40.00	**42.58**
	w/o manga	52.94	36.00	**44.55**
GPT-5	w/ manga	82.35	58.00	**70.30**
	w/o manga	64.71	40.00	**52.48**
Human 1	w/ manga	96.07	88.00	**92.07**
Human 2	w/ manga	94.11	90.00	**92.08**

また、n ページ目の発話についての問題を入力する際には、n ページ目までの全ての漫画画像 (`.jpg`) をモデルに入力した。ただし、通常のサイズで漫画を入力した場合にはモデルのコンテキスト長を超過するため実験を行うことができないので、元 PDF の 0.5 倍の解像度の画像を使用した。解像度の参考画像を付録 §B.2 に示す。また、比較のため、文脈となる漫画画像を与えない設定でも実験を行った。

対象モデル　オープンモデルと商用モデルで実験を行い、正解率を記録した。オープンモデルとして Qwen3-VL-32B-Instruct、Llama-4-Scout-17B-16E-Instruct、Llama-4-Maverick-17B-128E-Instruct-FP8 を、商用モデルとして gpt-5-2025-08-07 を対象とした。

また、比較のために同一の問題を人間の被験者にも出題した。被験者は日本語を母語とし日本で生育した 50 代の男女 1 名ずつの計 2 名であり、いずれも対象作品は未読であった。漫画画像と問題をいず

表 4　設問カテゴリ別の正解率 (%)。w/は漫画画像を入力した場合、w/o は入力しなかった場合の正解率を示す。

モデル	REASON		INTENT		FEELING		INDIRECT		REFERENCE		IMPLICATURE		ELLIPSIS	
	w/	w/o	w/	w/o	w/	w/o	w/	w/o	w/	w/o	w/	w/o	w/	w/o
Qwen3-32B	60.7	42.9	44.4	33.3	61.1	44.4	50.0	30.0	54.5	72.7	42.9	42.9	55.6	27.8
Llama-4-17Bx16E	32.1	35.7	33.3	44.4	61.1	50.0	30.0	10.0	45.5	27.3	42.9	28.6	44.4	27.8
Llama-4-18Bx128E	46.4	35.7	55.6	55.6	38.9	61.1	10.0	20.0	63.6	63.6	57.1	28.6	38.9	44.4
GPT-5	71.4	57.1	55.6	22.2	83.3	55.6	40.0	50.0	72.7	72.7	57.1	28.6	83.3	55.6

れも紙面で配布し[3)]、手書きでの回答を求めた。

4.2　結果

実験の結果 (表 3)、構築したベンチマークにおいて人間は 90%程度正解できるのに対して、VLLM は最新の大規模モデル (GPT-5) であっても 70%程度の正解率しか達成できなかった。このことは、マルチモーダルかつ長い文脈を扱う語用論的推論能力に関して、現状の VLLM と人間には依然としてギャップが存在することを示唆している。

作品 1 では漫画画像を文脈として与えた場合と与えない場合でのモデル正解率の差が小さかったのに対し、作品 2 では同様の事象は起こらなかった。どちらの作品でも、漫画画像を与えない場合の正解率はチャンスレート (25.00%) よりも高かった。これには、誤答肢 3 (easy) など容易に誤答と見抜ける誤答肢を設けていることや、設問・選択肢作成時のくせがアーティファクトとして残存し、モデルに画像なしでも正解させる手がかりを与えていることなど、複数の原因が考えられる。

4.3　分析

設問カテゴリ別の正解率を表 4 に示す。多くのモデルで特に正解率が低いのは INDIRECT であった。GPT-5 の回答を観察したところ、このカテゴリの問題においては、発話の言外の意味を捉えず字義通りな解釈を選択したり、言外の意味を捉えたような解釈ではあるが文脈にそぐわない選択肢を誤って選択する傾向が見られた。

また、問題の位置 (何ページ目に関する問題か) によって正解率に変動が出るかを分析するため、問題番号によって全問題を 5 つのセグメントに分割し、それぞれの正解率を集計した。結果を図 2,3 に示す。セグメントごとの正解率について、作品・モデルを横断する一貫した傾向は見て取ることができず、問題の位置によって正解率が変動するわけで

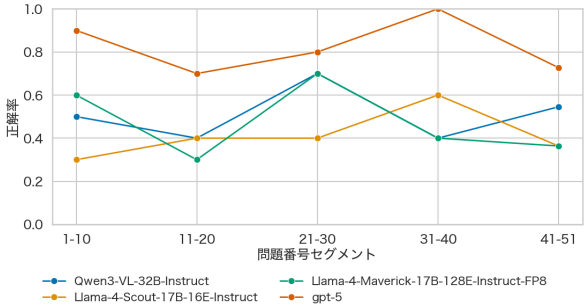

図 2　作品 1 のセグメント別正解率。

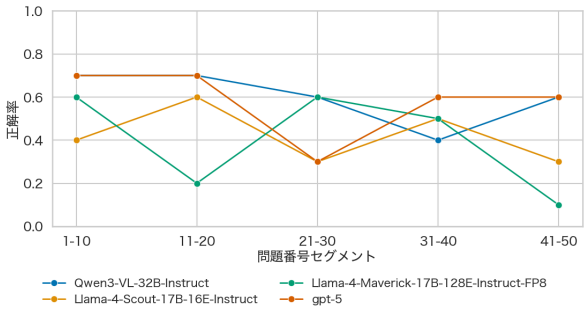

図 3　作品 2 のセグメント別正解率。

はないことが分かった。重要な点として、この結果は、VLLM の低い正解率の原因が長文脈の処理能力のみではないことを示唆している。長文脈の処理能力の低さのみが誤答の原因となるならば、入力される画像枚数の少ない前半のセグメントほど、正解率が高くなるはずであるためである。

5　おわりに

本研究では、マルチモーダルかつ長い文脈の処理が求められる語用論的推論ベンチマークを構築し、VLLM と人間の正解率の評価実験を行った。実験の結果、最新の VLLM であっても人間の正解率には劣後することが示された。今後は、ベンチマークの問題数を増やすとともに、構築したベンチマークを用いて VLLM の語用論的推論能力やその内部機序について分析することを予定している。

3)　問題用紙の画像を付録 §B.3 に示す。

謝辞

本研究の一部は、JST PRESTO JPMJPR24TC の支援を受けた。また、本研究は公益財団法人トヨタ財団 2024 年度特定課題 先端技術と共創する新たな人間社会 (D24-ST-0023) の支援を部分的に受けた。

参考文献

[1] Paul Grice. **Studies in the Way of Words**. Harvard University Press, 1989.

[2] S. C. Levinson. **Pragmatics**. Cambridge University Press, 1983.

[3] Bolei Ma, Yuting Li, Wei Zhou, Ziwei Gong, Yang Janet Liu, Katja Jasinskaja, Annemarie Friedrich, Julia Hirschberg, Frauke Kreuter, and Barbara Plank. Pragmatics in the era of large language models: A survey on datasets, evaluation, opportunities and challenges. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 8679–8696, Vienna, Austria, July 2025. Association for Computational Linguistics.

[4] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. **Multimedia Tools and Applications**, Vol. 76, No. 20, pp. 21811–21838, 2017.

[5] Kiyoharu Aizawa, Azuma Fujimoto, Atsushi Otsubo, Toru Ogawa, Yusuke Matsui, Koki Tsubota, and Hikaru Ikuta. Building a manga dataset "manga109" with annotations for multimedia applications. **IEEE MultiMedia**, Vol. 27, No. 2, pp. 8–18, 2020.

[6] Robyn Carston. **Thoughts and Utterances: The Pragmatics of Explicit Communication**. Blackwell Publishing Ltd., 2002.

[7] Shengguang Wu, Shusheng Yang, Zhenglun Chen, and Qi Su. Rethinking pragmatics in large language models: Towards open-ended evaluation and preference tuning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 22583–22599, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[8] Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. The goldilocks of pragmatic understanding: fine-tuning strategy matters for implicature resolution by LLMs. In **Proceedings of the 37th International Conference on Neural Information Processing Systems**, NIPS '23. Curran Associates Inc., 2024.

[9] Settaluri Sravanthi, Meet Doshi, Pavan Tankala, Rudra Murthy, Raj Dabre, and Pushpak Bhattacharyya. PUB: A pragmatics understanding benchmark for assessing LLMs' pragmatics capabilities. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Findings of the Association for Computational Linguistics ACL 2024**, pp. 12075–12097. Association for Computational Linguistics, August 2024.

[10] Zae Myung Kim, David E. Taylor, and Dongyeop Kang. "is the pope catholic?" applying chain-of-thought reasoning to understanding conversational implicatures, 2023.

[11] Akhila Yerukola, Saujas Vaduguru, Daniel Fried, and Maarten Sap. Is the pope catholic? yes, the pope is catholic. generative evaluation of non-literal intent resolution in LLMs. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**, pp. 265–275. Association for Computational Linguistics, August 2024.

[12] Dan Sperber and Deirdre Wilson. **Relevance: Communication and Cognition (2nd Edition)**. Blackwell, 1995.

[13] Takuma Sato, Seiya Kawano, and Koichiro Yoshino. Pragmatic theories enhance understanding of implied meanings in llms, 2025.

[14] Md Tahmid Rahman Laskar, Sawsan Alqahtani, M Saiful Bari, Mizanur Rahman, Mohammad Abdullah Matin Khan, Haidar Khan, Israt Jahan, Amran Bhuiyan, Chee Wei Tan, Md Rizwan Parvez, Enamul Hoque, Shafiq Joty, and Jimmy Huang. A systematic survey and critical review on evaluating large language models: Challenges, limitations, and recommendations. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 13785–13816, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[15] Junya Takayama, Tomoyuki Kajiwara, and Yuki Arase. DIRECT: Direct and indirect responses in conversational text corpus. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, **Findings of the Association for Computational Linguistics: EMNLP 2021**, pp. 1980–1989. Association for Computational Linguistics, November 2021.

[16] Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. A fine-grained comparison of pragmatic language understanding in humans and language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 4194–4213, Toronto, Canada, July 2023. Association for Computational Linguistics.

[17] Hengli Li, Song-Chun Zhu, and Zilong Zheng. DiPlomat: a dialogue dataset for situated pragmatic reasoning. In **Proceedings of the 37th International Conference on Neural Information Processing Systems**, NIPS '23. Curran Associates Inc., 2024.

[18] Shisen Yue, Siyuan Song, Xinyuan Cheng, and Hai Hu. Do large language models understand conversational implicature – a case study with a chinese sitcom, 2024.

[19] Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 8690–8705. Association for Computational Linguistics, July 2020.

[20] Zilong Zheng, Shuwen Qiu, Lifeng Fan, Yixin Zhu, and Song-Chun Zhu. GRICE: A grammar-based dataset for recovering implicature and conversational rEasoning. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, **Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021**, pp. 2074–2085. Association for Computational Linguistics, August 2021.

[21] Chenhao Zhang, Xi Feng, Yuelin Bai, Xeron Du, Jinchang Hou, Kaixin Deng, Guangzeng Han, Qinrui Li, Bingli Wang, Jiaheng Liu, Xingwei Qu, Yifei Zhang, Qixuan Zhao, Yiming Liang, Ziqiang Liu, Feiteng Fang, Min Yang, Wenhao Huang, Chenghua Lin, Ge Zhang, and Shiwen Ni. Can MLLMs understand the deep implication behind Chinese images? In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 14369–14402, Vienna, Austria, July 2025. Association for Computational Linguistics.

[22] Jeonghun Baek, Kazuki Egashira, Shota Onohara, Atsuyuki Miyai, Yuki Imajuku, Hikaru Ikuta, and Kiyoharu Aizawa. Mangavqa and mangalmm: A benchmark and specialized model for multimodal manga understanding, 2025.

[23] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to Japanese morphological analysis. In Dekang Lin and Dekai Wu, editors, **Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing**, pp. 230–237, Barcelona, Spain, July 2004. Association for Computational Linguistics.

表 5　各設問と選択肢の長さ。トークン数は、MeCab[23]による分かち書ちに基づいて数えた。

	設問	正解	hard	medium	easy
平均文字数	31.8	28.7	29.3	27.1	24.4
平均トークン数	20.4	18.6	19.1	17.4	15.2

図 4　0.5 倍の解像度の漫画画像。© 栗城祥子

A　ベンチマークの統計

表 5 に、作成した各設問と選択肢の長さを示す。

B　実験設定の詳細

B.1　VLLM へのプロンプト

```
与えられた漫画の{page_num}ページ目の
「{utterance}」という発話についての問題です。
回答選択肢の番号（A/B/C/D)だけを出力し、他のこ
とは一切出力しないようにしてください。

問題: {question}

選択肢
A. {option_1}
B. {option_2}
C. {option_3}
D. {option_4}
```

B.2　画像解像度の例

0.5 倍と 1.0 倍の解像度の漫画画像の例を図 4,5 に示す。0.5 倍の解像度であっても、文字や漫画表現の視認が大きく妨げられるわけではないことがわかる。

B.3　人間回答時の問題用紙

人間によるベンチマークへの回答の際に配布した紙面の元画像を図 6 に示す。

図 5　1.0 倍の解像度の漫画画像。© 栗城祥子

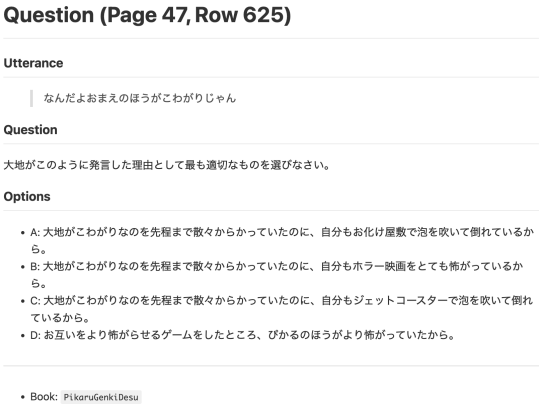

図 6　人間による回答実験においては、このような問題と漫画画像を紙面に印刷したうえで被験者に配布し、手書きでの回答を求めた。