視線軌跡を与えたVLMにおける英文読解時の不明単語のZero-shot予測

中村純也 遠藤伶 田中大 衣川和尭 岡田拓也 美野秀弥 河合吉彦
NHK放送技術研究所
{ nakamura.j-hy, endou.r-mm, tanaka.m-oc, kinugawa.k-jg, okada.t-im,
mino.h-gq, kawai.y-lk }@nhk.or.jp

概要

大規模言語モデルは，テキストのみの指示では利用者の意図を十分に捉えられない場合があり，利用者の関心や意図を含む視線軌跡の利用が期待される。近年，視線軌跡を言語モデルへ入力するために専用のエンコーダーを追加する手法が報告されているが，一般の視覚言語モデル（VLM）に対して，視線軌跡に注目した分析は十分でない。本研究では視線軌跡をVLMが直接扱える画像形式に変換した。その上で，英文読解中に利用者が不明と判断した単語の予測タスクで複数のVLMを評価した。その結果，8B級モデルにおける，単語予測の正答率は，1位予測で最大39%，上位5位予測は最大74%を示した。

1 はじめに

大規模言語モデル（LLM）は，画像入力に対応した視覚言語モデル（VLM; Visual Language Model）や，動画や音声などの時系列情報にも対応するマルチモーダルLLM（MLLM）へと進展している[1]。これらは主にテキストプロンプトで指示を受け取るが，対応するモダリティに応じて，視覚や聴覚情報を追加したマルチモーダルプロンプトも扱える。

一方で，テキストプロンプトは数十〜数百語程度が一般的で，人が保持する意図や状況文脈，およびLLMが内部に持つ知識や推論過程に比べ，伝達情報量が小さい。また，利用者は意図を言語化してプロンプト構築する必要があり，時間的・認知的コストも要する。そのため，プロンプトは人とLLMを繋ぐ伝達経路でありながら，意図の取りこぼしや曖昧性を生みだす原因となる。さらに，プロンプトの品質や量，対話の頻度は個人差が大きく，情報伝達の一貫性・安定性の担保も課題である。

この課題を緩和するには，利用者の状態や意図を追加的に観測できるモダリティの導入が有効である。ヒューマンコンピュータインタラクション（HCI）領域では，眼球運動データをロジスティック回帰[2]やサポートベクターマシン[3][4]に適用し，利用者の関心や意図を予測・分類する試みが報告されている。眼球運動計測による視線軌跡は，行動に密接に結びつく暗黙的手がかりであり，高い時間分解能で意図の兆候を捉え得ることから，プロンプトを代替するモダリティとして，MLLMへの応用が期待される。

近年，VLMを視線軌跡に対応させる研究が，Egocentric（自己視点）画像/動画能力やグラウンディング能力の獲得を背景に行われている。ヒートマップ形式の視線データに対応したVLM[5]や，視線軌跡を画像に統合する専用エンコーダーを備えたVLM[6]が報告されている。しかし，これらは専用のエンコーダーの追加と学習を要し，新しい基盤モデルへの追従コストが発生する。また，VLMの能力評価において，視線軌跡に注目した分析は十分に検討されていない。

そこで本研究では，視線軌跡をVLMに直接入力可能な画像表現に変換し，VLMの意図予測能力を評価することを目的とした。評価に際し，疑問の発生による視線滞留が発生しやすい，英文読解タスクを設定し，読解中の視線軌跡と不明点判断ラベルを含む評価データセットを構築する。また，評価データセットから，英文読解中の利用者が「不明」と判断した英単語をVLMに予測させる。同一タスクを人に予測させ，ベースラインとした。

本研究では3つの仮説を設定した。H1. VLMで単語予測タスクをZero-shotで実行したとき，人手評価と比較して同等以上のスコアが得られる。H2. 開発元の異なるVLMを同一の入力方式で横断的に評価できる。H3. 同一系列のVLMでは，パラメータ規模の違いによるスケーリング効果が得られる。

2 方法

2.1 英文読解中の視線計測実験

英文読解中の眼球運動計測と，参加者自身による計測データへのアノテーション実験を実施した。実験で得られたデータから, VLM で評価するためのデータセットを構築した。

2.1.1 参加者・装置

8 名の正常な視力または矯正視力をもつ，日本語を母国語とする成人の実験参加者がインフォームドコンセントに署名し，実験に参加した。実験は NHK 放送技術研究所 心理･生理実験審査委員会の承認を得て実施された。

参加者は椅子に座り，机上のラップトップコンピュータ（ASUS GA503Q, 15.6 型, 2560 [width]×1440 [height]）のキーボードに左手を添え，液晶ディスプレイに提示された記事を読解した。参加者の眼球運動は，視線計測装置（Tobii Pro Nano）で計測され，左右の眼球が注視したディスプレイ上の正規化座標を 1/60 秒間隔で記録した。

2.1.2 刺激提示・タスク

本実験は，英文記事読解タスク・アノテーションタスクで構成した。

英文記事読解タスクでは，コンピュータのスクリーンに，OpenCV を用いて英文記事のタイトルと本文テキスト，記事の切り替えナビゲーションを描画した（図１）。記事は，英語版 Wikipedia の Today’s featured article (2025 年 10 月分) を用いた[7]。参加者へは，記事の内容の理解を目標として読解するよう指示し，読解方法は一任した。また，読解中に疑問点や不明点が生じた時点の全てにおいて，指定したキーを押下し，イベント記録を実施するよう指示した。読解時間の制限は設けなかった。

アノテーションタスクでは，参加者に記事読解中の視線計測データに対し，不明点判断ラベルを作成した。ラベル付け作業は，読解中のイベント記録と視線軌跡を参照しながら，そこで生じた疑問や不明点，読み進みの停滞が発生した際の理由について記録するよう指示した。具体的には，参加者の不明な単語やその具体的な理由の記述などが行われた。また，本実験では，参加者 1 人あたり 7〜12 記事を読解した。

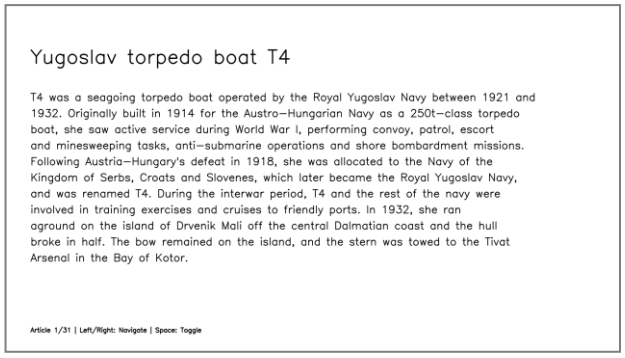

図 1：英文記事読解タスクで提示した画面構成

2.1.3 手続き

実験開始後，実験参加者は英文記事読解タスクを実施しやすいよう，椅子やコンピュータの位置を調整し，視線計測装置のキャリブレーションは 1 記事毎に行った。その後，英文記事読解タスク・アノテーションタスクを順に実施し，これを繰り返した。

2.2 VLM の意図予測能力の評価

2.2.1 視線計測データの変換・形式

本研究では，VLM 間で有効な画像入力形式を比較・評価するため，文書でよく利用される半透明の黄色マーカーを模した描画（図 2-A, Marker 方式），VLM のグラウンディングタスクで利用される, 赤枠バウンディングボックスの描画（図 2-B, Bounding Box 方式），視線軌跡の描画として利用されるヒートマップ形式の描画（図 2-C, Heatmap 方式）の 3 形式を設定した。

Marker 方式と Bounding Box 方式では，VLM での 1 回の推論で画像を 10 枚入力する。10 枚の画像群の作成フローは，①アノテーション記録時点の直前 5 秒間の視線軌跡データを抽出，②これを連続する 500ms 区間に 10 分割, ③各区間の視線軌跡を記事上に描画，④ 1 区間につき 1 枚の画像を書き出し，⑤合計 10 枚作成，とした。

Heatmap 形式は，視線軌跡の累積を表現する方法であるため，5 秒間の累積結果を 1 枚の画像で描画した。

2.2.2 予測タスクの対象抽出と指示設定

利用者が読解時に生じた不明点の予測を行うにあたり，実験参加者が記録したアノテーションから，不明な「単語」を対象として，全 537 件のアノテーションから，100 件をランダムに抽出した。抽出し

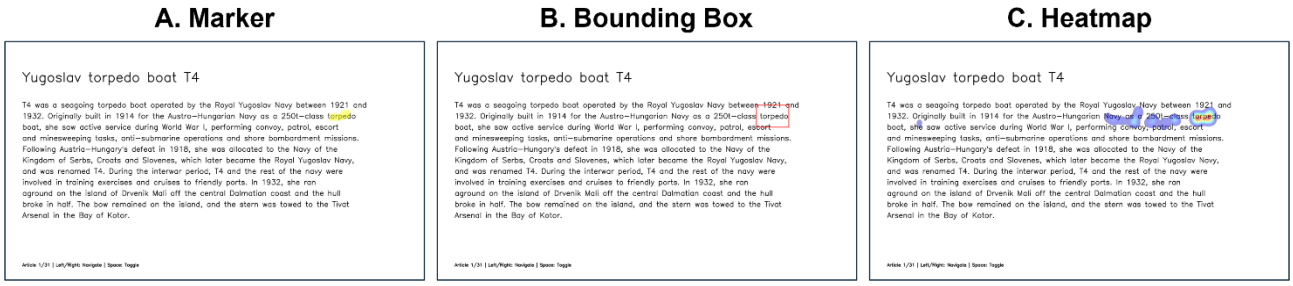

図 2：視線計測データの画像への変換イメージ

た各アノテーションに対し，該当記事の内容と直前 5 秒間から生成した視線軌跡画像を VLM に入力したうえで，利用者の不明語の予測とその理由を信頼度の高い順に 5 件予測するように指示した（付録. Marker 方式の指示プロンプト）。これにより，VLM が記事内容と視線軌跡データを統合的に用いて，利用者が不明と判断した英単語を予測した。

2.2.3 評価対象 VLM・ベースライン

評価対象の VLM は, Qwen/Qwen3-VL-{2B, 4B, 8B, 32B}-Instruct [8], OpenGVLab/InternVL3_5-8B [9], nvidia/NVIDIA-Nemotron-Nano-12B-v2-VL-BF16 [10] を用いた。また，商用モデルとの比較のため，GPT-4o:2024-02-01 (Azure API) [1]を用いた。また，ベースラインとして，同一タスクに対する人の予測を筆者が行い，各アノテーションについて，1〜5 位の英単語を予測した。また，選定理由の記述は実施時間が長大化するため，省略した。

3 結果

VLM で予測した単語の採点には，LLM as a Judge を用いて, 正解アノテーションとの一致を評価した。評価する LLM には，gpt-oss-120b を用いた。評価では，VLM が予測した上位 5 件について，1 位から 5 位予測の一致数と，不一致数をカウントした（付録. LLM as a Judge 実施プロンプト）。なお，人の予測は一致数と不一致数をカウントし，各 VLM の 1 位予測と比較する。パラメータ規模の違いによる比較では半透明の黄色マーカーを模した描画（Marker）の Qwen3-VL を用いた。

3.1 入力形式・モデル比較

VLM で評価した結果，予測 1 位での一致数（Top-1）は，Marker 形式では，Qwen3-VL-8B が 35 件，InternVL-3.5-8B が 32 件，Nemotron-Nano-12B-v2-VL が 21 件, GPT-4o が 30 件, 人手評価が 49 件だった。Bounding Box 形式では，Qwen3-VL-8B が 39 件，InternVL-3.5-8B が 30 件，Nemotron-Nano-12B-v2-VL が 28 件, GPT-4o が 24 件, 人手評価が 45 件だった。Heatmap 形式では, Qwen3-VL-8B が 23 件, InternVL-3.5-8B が 17 件，Nemotron-Nano-12B-v2-VL が 9 件，GPT-4o が 23 件，人手評価が 43 件だった。予測 1 位での比較ではいずれの方式においても人手評価が最も高い結果となった（図 3）。

予測 5 位までの一致数（Top-5）は，Marker 形式では, Qwen3-VL-8B が 62 件, InternVL-3.5-8B が 73 件, Nemotron-Nano-12B-v2-VL が 48 件, GPT-4o が 68 件, 人手評価が 79 件だった。Bounding Box 形式では, Qwen3-VL-8B が 74 件，InternVL-3.5-8B が 63 件, Nemotron-Nano-12B-v2-VL が 54 件, GPT-4o が 79 件, 人手評価が 75 件だった。Heatmap 形式では，Qwen3-VL-8B が 47 件，InternVL-3.5-8B が 56 件，Nemotron-Nano-12B-v2-VL が 25 件，GPT-4o が 54 件，人手評価が 77 件だった。予測 5 位までの一致数では，Bounding Box 形式での GPT-4o の一致数のみ，人手評価を上回った。

3.2 パラメータ規模比較

Qwen3-VL シリーズにおける，パラメータ規模の違いによる能力を評価した結果，2B は，35 項目の予測が一致し，65 項目が不一致であった。4B は，62 項目の予測が一致し, 38 項目が不一致であった。32B は，67 項目の予測が一致し，33 項目が不一致であった（図 4）。モデルにおいては，32B モデルが最良の結果を示したが，Top-1，Top-5 のいずれも人手評価を超えなかった。

4 考察

本研究では，視線軌跡データを VLM に入力可能な画像形式に変換し, VLM が視線データから利用者が不明と判断した英単語を Zero-shot で予測する能力を有しているかを評価した。また，開発元の異な

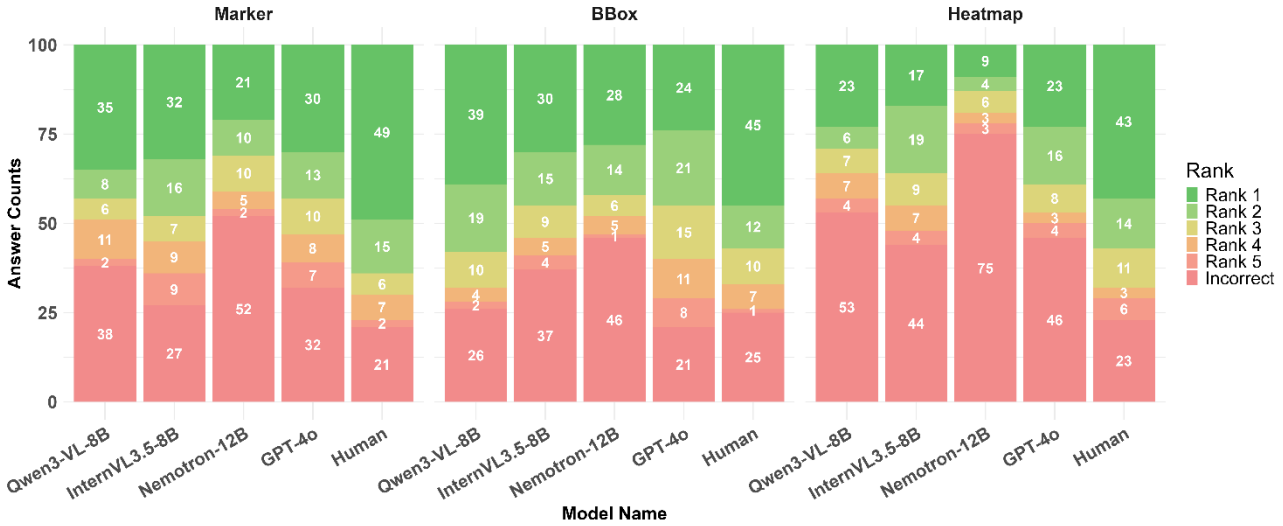

図 3：入力形式・モデル比較結果

る複数 VLM において，同一入力で，能力評価が可能かを評価した。さらに，同一系列モデルにおけるパラメータ規模の比較を行った。

評価の結果，モデル間の比較においては，1 位予測（Top-1）と上位 5 位予測（Top-5）のいずれも，人が最も高い成績であった（図 3）。そのため，仮説 H1 は支持されなかった。一方，全ての VLM に共通して，上位 5 件の予測のうち，1 位の予測が正解と一致する数が最多であった。また，Qwen3-VL，InternVL3.5, GPT-4o の上位 5 件の正解数は過半数を超えた。これは，VLM が視線軌跡の解析と利用者の不明英単語を予測するタスクに対する一定程度の能力を有することを示唆した。これらから，開発元が異なる VLM においても，連番画像形式で表現された，視線軌跡データを分析する能力を持つことが示唆された。そのため，仮説 H2 は支持された。

入力形式においては，Marker 形式と Bounding Box 形式は，Heatmap 形式より高いスコアを示した。Heatmap 形式は 5 秒間の累積データを 1 枚の画像で提示できるが，時系列方向の単位時間毎の変化の情報が欠落していたため，スコアが低下したと考えられる。この結果は，時間的変化を考慮しない静的パターンでの推測は困難である[4]ことを示唆した先行研究と一致した。

パラメータ規模と性能の関係については，2B パラメータモデルで顕著な性能低下が確認された（図 4）。一方で，32B は 8B に対する明確な性能向上は得られず，スケーリング効果 [11] が頭打ちになっていることが示唆された。そのため仮説 H3 は部分的に支持された。

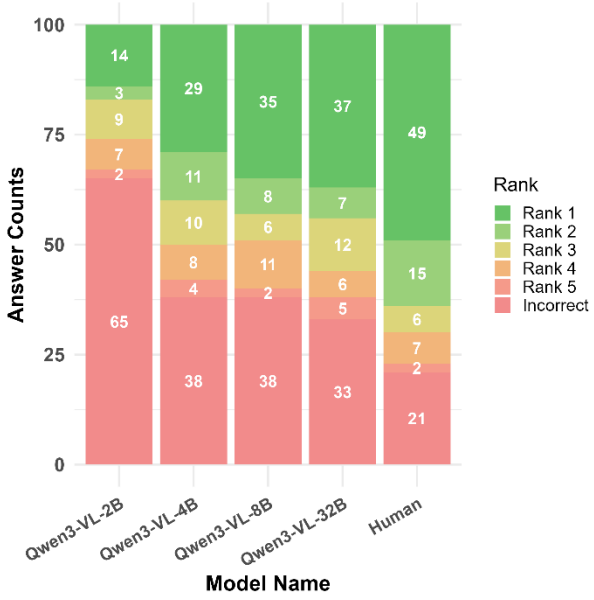

図 4：パラメータ規模比較結果

5 おわりに

本研究では，視線軌跡という特殊なマルチモーダル情報について，VLM の Zero-shot 能力が得られることが示唆され，また，連番画像による時間的変化の情報の提示が有効であることが示唆された。一方で，VLM+専用エンコーダー方式の先行研究[5][6]はモデルが非公開なため，比較できないのは課題である。

今後，より広範囲に利用者の要求を予測するタスクへの拡大や，利用者が助けを必要としているタイミングの検出，より優れた入力形式の検討や，モバイルデバイスでの動作，リアルタイム化を進める。

参考文献

[1] OpenAI: Aaron Hurst, Adam Lerer, Adam P. Goucher et al. GPT-4o System Card. **arXiv preprint arXiv: 2410.21276**. 2024.

[2] Melih Kandemir, Veli-Matti Saarinen, and Samuel Kaski. Inferring object relevance from gaze in dynamic scenes. **Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications**, pp. 105-108, 2010.

[3] Roman Bednarik, Hana Vrzakova, and Muchal Hradis. What do you want to do next: a novel approach for intent prediction in gaze-based interaction. **Proceedings of the symposium on eye tracking research and applications**, pp. 83-90, 2012.

[4] Michelle R. Greene, Tommy Liu, and Jeremy M. Wolfe. Reconsidering Yarbus: A failure to predict observers' task from eye movement patterns. **Vision research**, Vol. 62, pp. 1-8, 2012.

[5] Kun Yan, Lei Ji, Zeyu Wang, Yuntao Wang, Nan Duan, and Shuai Ma. Voila-a: Aligning vision-language models with user's gaze attention. **Advances in Neural Information Processing Systems**, Vol. 37, pp. 1890-1918, 2024.

[6] Sounak Mondal, Naveen Sendhilnathan, Ting Zhang, Yue Liu, Michael Proulx, Michael Louis Iuzzolino, Chuan Qin, and Tanya R. Jonker. Gaze-Language Alignment for Zero-Shot Prediction of Visual Search Targets from Human Gaze Scanpaths. **Proceedings of the IEEE/CVF International Conference on Computer Vision**, pp. 2738-2749, 2025.

[7] Wikipedia:Today's featured article/October 2025 – Wikipedia, (2025-12 閲覧). https://en.wikipedia.org/wiki/Wikipedia:Today%27s_featured_article/October_2025

[8] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen et al. Qwen3-VL Technical Report. **arXiv preprint arXiv: 2511.21631**. 2025.

[9] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu et al. InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency. **arXiv preprint arXiv: 2508.18265**. 2025.

[10] NVIDIA: Amala Sanjay Deshmukh, Kateryna Chumachenko, Tuomas Rintamaki et al. NVIDIA Nemotron Nano V2 VL. **arXiv preprint arXiv: 2511.03929**. 2025.

[11] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. **arXiv preprint arXiv:2001.08361**. 2020.

付録

Marker 方式の指示プロンプト（利用者の注目している単語の推論プロンプト）

```
"system_prompt": "You are an expert in analyzing eye-tracking data and predicting user focus on specific words. Provide precise word-level predictions based on gaze patterns."
"analysis_prompt": "Please analyze the following {segment_count} sequential gaze trajectory images showing eye-tracking data at 500ms intervals over a 5-second period.¥n¥n**Context:**¥n- Each image shows gaze trajectory data for 0.5 second segments¥n- Yellow highlighted areas indicate the user's focus regions¥n¥n**Word Prediction Request:**¥nBased on the gaze patterns across these sequential segments, please predict the specific words that the user is focusing on.¥n¥n**Analysis Guidelines:**¥n- Words near or within yellow highlighted areas should be given higher priority¥n- Words with longer dwell time (appearing in multiple consecutive segments) should be ranked higher¥n- Words appearing in later segments (closer to the end of the 5-second period) indicate sustained attention and should be weighted more heavily¥n- Consider both the proximity to yellow markers and the duration of attention¥n- Prioritize words that maintain focus towards the final segments as they represent deliberate sustained attention¥n¥n**Output Format:**¥nPlease provide your response in the following JSON format only:¥n¥n```json¥n{{¥n  ¥"focused_words¥": [¥n    {{¥n      ¥"rank¥": 1,¥n      ¥"word¥": ¥"example_word_1¥",¥n      ¥"reason¥": ¥"Located within yellow highlighted area across segments 6-9 with sustained fixation toward end¥"¥n    }},¥n    {{¥n      ¥"rank¥": 2,¥n      ¥"word¥": ¥"example_word_2¥",¥n      ¥"reason¥": ¥"Near yellow marker with extended dwell time in final 3 segments (7-9)¥"¥n    }},¥n    {{¥n      ¥"rank¥": 3,¥n      ¥"word¥": ¥"example_word_3¥",¥n      ¥"reason¥": ¥"Adjacent to yellow highlight with attention maintained through segments 5-8¥"¥n    }},¥n    {{¥n      ¥"rank¥": 4,¥n      ¥"word¥": ¥"example_word_4¥",¥n      ¥"reason¥": ¥"Brief fixation near yellow area appearing in middle to late segments (4-6)¥"¥n    }},¥n    {{¥n      ¥"rank¥": 5,¥n      ¥"word¥": ¥"example_word_5¥",¥n      ¥"reason¥": ¥"Peripheral to yellow marker with early attention but fading toward end (segments 1-3)¥"¥n    }}¥n  ],¥n  ¥"analysis_timestamp¥": ¥"2024-12-08T00:00:00Z¥",¥n  ¥"segment_count¥": {segment_count}¥n}}¥n```¥n¥nProvide only valid JSON output with the top 5 words that received the most attention, ranked by focus intensity. For each word, provide a specific reason explaining why you identified it as a focus point, specifically mentioning its relationship to yellow highlighted areas, the duration of attention (dwell time) observed across segments, and the temporal pattern (early, middle, or late segments) which indicates the sustained nature of attention."
```

LLM as a Judge 実施プロンプト[i]

```
"system_prompt": "You are an expert evaluator of eye-tracking word prediction accuracy. Your task is to evaluate how well predicted focus words align with the user's actual question or comment intent. You must respond ONLY in valid JSON format."
"evaluation_prompt": "Please evaluate the predicted focus words based on the user's actual comment/question.¥n¥n**User's Actual Comment/Question:**¥n¥"{comment}¥"¥n¥n**Predicted Focus Words:**¥n{predicted_words}¥n¥n**SIMPLE RANKING EVALUATION:**¥nFor each word, determine if it matches the user's question:¥n- MATCH: The word is relevant to answering the user's question¥n- NO MATCH: The word is not relevant to the question¥n¥n**SCORING SYSTEM (Maximum 10 points):**¥nFind the HIGHEST RANKING word that matches. The final score is:¥n- Rank 1 match = 10 points (best possible score)¥n- Rank 2 match = 8 points¥n- Rank 3 match = 6 points¥n- Rank 4 match = 4 points¥n- Rank 5 match = 2 points¥n- No match = 0 points¥n¥n**EXAMPLE:**¥n- If only Rank 2 word matches → Final Score = 8 points¥n- If Rank 1 and Rank 4 both match → Final Score = 10 points (highest rank)¥n- If Rank 3 and Rank 5 both match → Final Score = 6 points (highest rank)¥n¥n**IMPORTANT: You must respond with ONLY a valid JSON object in the following exact format:**¥n¥n{{¥n  ¥"focus_words_evaluation¥": [¥n    {{¥n      ¥"rank¥": 1,¥n      ¥"word¥": ¥"word_text¥",¥n      ¥"is_match¥": true,¥n      ¥"rank_points¥": 10,¥n      ¥"reasoning¥": ¥"Brief explanation for the match decision¥"¥n    }}¥n  ],¥n  ¥"overall_assessment¥": {{¥n    ¥"final_score¥": 10,¥n    ¥"max_possible_score¥": 10,¥n    ¥"best_matched_rank¥": 1,¥n    ¥"percentage¥": 100.0,¥n    ¥"summary¥": ¥"Brief assessment of prediction quality¥"¥n  }}¥n}}¥n¥nProvide your evaluation in valid JSON format only."
```

実行環境

ハードウェア：JCS VCAR-DSS1301, Ryzen9 9950X3D, DDR5 64GB, RTX PRO 6000 BW WS ED, SN7100 4TB
ソフトウェア：vLLM v0.12.0, torch 2.9.0+cu129, Python 3.12.12, Ubuntu 24.04, WSL2, Windows 11

[i] 本実験の分析では Point は使用していない