Omni-JDocVQA: 多種多様な文書を含んだ日本語視覚文書理解ベンチマークの構築

梶川 怜恩 ♣,‡ 中山 功太 ‡ 小田 悠介 ‡ 神田 峻介 ◇ 赤部 晃一 ◇ 二宮 崇 ♣ 岡崎 直観 ‡,♡
♣ 愛媛大学 ‡ 国立情報学研究所 大規模言語モデル研究開発センター
◇ シェルパ・アンド・カンパニー株式会社 ♡ 東京科学大学
{reon@ai.cs., ninomiya.takashi.mk@}ehime-u.ac.jp {nakayama, odashi}@nii.ac.jp
{shunsuke.kanda, koichi.akabe}@cierpa.co.jp okazaki@c.titech.ac.jp

概要

既存の日本語視覚文書理解ベンチマークは，収集範囲の限定，評価文書のラベリング不足，および文書内に必ず正解が存在することを前提とした設問設計という 3 つの課題を抱えており，現実的な文書理解モデルの適用条件下での評価が困難である．本研究は，広範な文書種別で構成された新しい日本語視覚文書理解ベンチマーク（Omni-JDocVQA）を提案する．本ベンチマークは，網羅的に収集された文書に対する新たに設計された文書ラベリングに加え，実際の利用シーンを想定した現実的な QA タスクを特徴とする．公開されている 15 種類の日本語・多言語の視覚言語モデルを Omni-JDocVQA で評価し，モデルの汎用性と課題を明らかにする．

1 はじめに

視覚言語モデル（Vision Language Model; VLM）の発展に伴い，文書画像を直接理解する視覚文書理解 [1] の研究が進んでいる．その中核をなすタスクが，Document Visual Question Answering（DocVQA）[2–10] である．英語圏においては，実世界の文書の多様性や複雑性を反映するため，多様なドメインを対象とした DUDE [6] や，長文読解に特化した MMLongBench-Doc [7] など，高度なベンチマークが提案されている．しかし，日本語圏におけるベンチマーク整備 [8–10] は，英語圏の進展と比べて発展途上であり，以下の 3 つの課題を抱えている．

日本語文書の網羅性の欠如　既存ベンチマークは，官公庁の公開文書 [8] やスライド形式 [9] など特定の情報源・形式に偏っており，多様なドメインやフォーマットを網羅できていない．DocVQA タスクにおける VLM の汎用的な能力を測定するためには，広範な情報源から抽出された多様な文書群による評価が不可欠である．

文書の種類に応じたラベルの欠如　既存ベンチマークの多くは，評価対象の文書に対してドメインやフォーマットの分類が行われていない．そのため，評価結果が単一のスコアとして算出されるに留まり，「モデルがどの分野の文書を得意とし，どの形式で失敗しやすいのか」という詳細なエラー分析が不可能である．多角的な分析を通じてモデルの改善指針を得るためには，文書の特性を精緻に捉えるラベル体系を定義・付与することで，評価の粒度を高める必要がある．

QA アノテーションの不自然さ　既存データセットの主要な作問方法は，回答が記載された文書を作問者が熟読した上で，文書に関連する質問を作成する方式が一般的である．一方で，実世界で QA システムを利用する状況では，ユーザの置かれた状況に基づく質問が先に存在し，それを解決するために文書を調査するのであり，データセットの作問方法とは質問と文書の因果関係が反転している．したがって，モデルの真の実用性を評価するためには，このような因果関係まで想定した自然な質問を設計することが求められる．

これらの課題を解決すべく，本研究では新しい日本語視覚文書理解ベンチマーク「Omni-JDocVQA」を設計する．まず，分析対象となる文書として広範なドメインとフォーマットを網羅し，これらを横断するラベル体系を定義・付与することで，評価対象のモデルに対する精密な分析を可能とする．質問作成プロセスにおいては「ユーザの利用コンテキスト」を先に定義し，文書内容を既知としない状況下での作問を行うことで，実際の利用に即した推論能力を要求する VQA タスクとした．本ベンチマーク

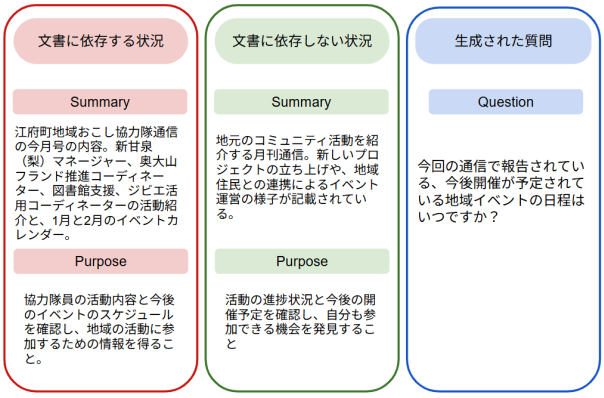

図 1　収集した文書から文書に直接依存しない質問を生成

は特定の文書種別に依存しない汎用的な視覚文書理解能力の評価基盤を提供する．また本ベンチマークを用いた評価を通じて，現行の VQA モデルの限界を明らかにする．

2　Omni-JDocVQA

本節ではベンチマークタスクの構築手順および各タスクの評価方法について説明する．

2.1　文書収集

本ベンチマークは，実世界に存在する多様な文書から構築することを目標とする．この実現のため，Common Crawl から抽出された大規模な PDF ファイル群である CCpdf [11] を起点とし，ここから日本語で記述されたファイルを無差別に抽出することで作問対象となる PDF 集合を構築した[1)]．PDF データから Qwen3-VL[2)] [12] による OCR でテキストを抽出し，fasttext-langdetect[3)] [13, 14] により抽出テキストが日本語である確率が 0.9 より高いと判定された PDF を日本語 PDF とした．最終的に 70,347 件のファイルが日本語 PDF であると判定された．

2.2　文書ラベリング

本ベンチマークでは，文書種別を**ドメイン**（学術・医療など文書の対象分野），**フォーマット**（レポート・マニュアルなど文書の記述形式）の 2 分類で定義し，文書ごとにこれらの分類をラベリングする．これによりモデルの文書理解能力を分析する際，全体的な性能の分析だけでなく，分類を限定した分析が可能となる．一つの文書に複数のラベルを許容するマルチラベル形式とし，いずれの定義にも該当しない文書には「その他」ラベルを付与する．ラベリングには Qwen3-VL を使用した．文書ラベルの詳細を付録 A の表 2 に掲載する．

2.3　QA アノテーション

文書サンプリング　QA 対象の文書を選別するため，以下の手順でサンプリングを行う．まず，埋め込みモデル ruri-v3-310m[4)] [15] を用いて文書の先頭 3 ページをベクトル化する．次に，得られたベクトル表現に対して K-means 法 [16] によるクラスタリングを行い，各クラスタの中心点に最も近い文書を抽出する．これにより，各ラベル内における文書の多様性を保ちつつ，典型的な文書構造を持つ文書をサンプリングする．

状況生成　文書の内容を把握していないユーザが，特定の目的を持って文書を参照する状況を想定し，質問を生成する．この要件を満たすため，本研究では以下の 2 段階の生成プロセスを導入する．まず，Qwen3-VL を用いて文書から「概要」と「閲覧目的」の 2 要素を「その文書を参照する状況」として生成する．この時点では生成された状況が文書に依存した内容である可能性が高いため，更に Qwen3[5)] [17] を用いて文書内容と無関係な類似の状況へと変換することで，状況としての枠組みは維持しつつ、特定の文書への依存を排除する．

質問生成　gpt-oss-120b[6)] [18] を使用し，前節で生成した状況 1 つにつき 3 件の質問を生成する．質問の生成にあたっては，「明確かつ具体的であること」「文書を読まないと回答不能であること」等を条件付けする（付録 B）．

文書と質問の再紐付け　生成した質問は文書集合からは独立しており，回答に使用する文書を再度紐付ける必要がある．ここでは ruri-v3-310m を用いて質問から文書へのベクトル検索を行い，検索上位 3 件の文書を質問に対する関連文書として紐付ける．

解答例の作成　紐付けられた質問と文書のペアに対し，人手による解答例のアノテーションを行う．回答内容は文書内に記載されている客観的事実のみを根拠とし，日本語の自然な文章で構成することを要求事項とした．回答文に加え，解答の根拠となっ

1）　収集した PDF ファイルのうち，暗号化やデータ破損により読み込みができない PDF ファイルは予め除外した．
2）　https://huggingface.co/Qwen/Qwen3-VL-30B-A3B-Instruct
3）　https://pypi.org/project/fasttext-langdetect/
4）　https://huggingface.co/cl-nagoya/ruri-v3-310m
5）　https://huggingface.co/Qwen/Qwen3-30B-A3B
6）　https://huggingface.co/openai/gpt-oss-120b

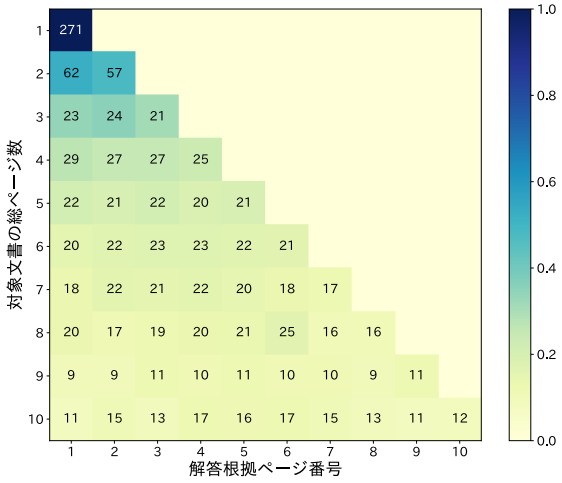

図 2　文書内における解答根拠箇所の分布

たページ番号を併せて付与する．なお人手アノテーションの作業負荷を考慮し，対象文書は 10 ページ以内のものに限定した．

作業者が回答を導き出せない場合は単に「回答できません」とするのではなく，その理由を自由記述で付記する[7]．特に，文書内に回答根拠が存在せず，質問と文書の関連性も認められない事例については「解答根拠なし」として個別にラベリングを行った．回答不能例はオープンドメイン QA では必然的に発生する問題であり，このような事例を意図的に包含することで「文書から必ず回答可能である」という運用上不都合なバイアスをベンチマーク全体から排除することができる．最終的に，本工程を経て構築された QA セットは合計 717 件となった．

2.4　評価方法

モデルの回答精度を LLM-as-a-Judge [19] により評価する．本評価では入力として「質問」「正解の解答例」「モデルの生成した回答」の 3 つ組を評価者モデルに提示し、質問の文脈において回答が正解とどの程度一致しているかを 5 段階のリッカート尺度により判定する[8]．前述のように本ベンチマークには回答不能例を含むため，回答が生成できない場合にその理由を生成することも要求する．

7）　著者らの知る限り，回答不能理由をアノテーションした視覚文書理解ベンチマークは本研究が初である．
8）　詳細は付録 B を参照されたい．

2.5　統計

文書ラベルごとの事例数を図 3 に示す．「その他」を除く全てのラベルにおいて少なくとも 50 件程度の事例が収録されている．なお「解答根拠なし」は 155 件（全体の 21%）であった．

解答根拠となるページ番号の分布を表すヒートマップを図 2 に示す．各文書の最大ページ数（縦軸）に対する解答根拠ページ（横軸）の出現頻度を見ると，特定のページに集中することなく，概ね均一に分散しており，最大ページ数（各行）における分散の平均値は 2.85 であった．このように解答根拠が特定のページに偏らず分散していることは，ベンチマークとして文書の局所的な特徴のみ参照すれば回答できることを許さず，モデルが文書全体を正しく読み取れているかを評価できることを示唆する．

3　評価実験

3.1　実験設定

本ベンチマークを用い，2023 年から 2025 年に公開された日本語・多言語の VLM を評価する．評価するモデルは全て HuggingFace Hub[9] より取得した．推論時の温度パラメータは 0.7 とし最大トークン数は各モデルの最大値[10] に設定した．シード値を変えながら 3 回の評価スコアを得て、その平均を報告する．評価尺度として，LLM-as-a-Judge を利用し，評価者モデルとして gpt-5.1-2025-11-13 を用いた．評価時の温度パラメータ, シード値はともに 0 とした．

3.2　評価結果

評価結果を図 3 に示す．平均スコア (overall) では GPT5.1 や Qwen3-VL などのが 4 点前後のスコアを達成しているのに対し，日本語モデルは 1 から 2 点台に留まっており，日本語モデルにおける視覚文書理解能力には依然として改善の余地が認められる．

日本語モデルの中で比較的高性能である karakuri-vl-32b-instruct [20] について分類レベルでスコアを観察すると，ドメインではビジネスで高い性能 (3.2) を示す一方，技術 (2.8) や学術 (2.5) では相対的に低いスコアに留まっている．フォーマットではウェブサイト (3.6) やスライド (3.5) に対しては高い性能を示すものの，創作物 (2.6) やマニュアル (2.5) では

9）　https://huggingface.co/models
10）　位置埋め込みの最大長の値を参照した．

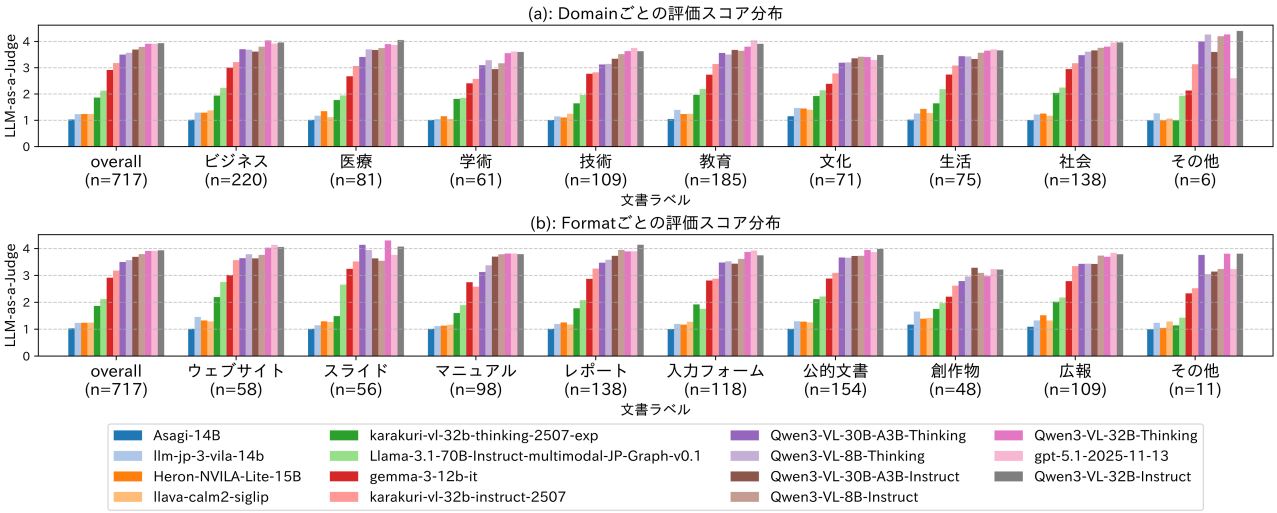

図 3　文書ラベル別の平均評価スコアの分布: データ件数 n を併記する．

表 1　回答不能問題の例（Qwen: Qwen3-VL-8B-Instruct，LLM-jp: llm-jp-3-vila-14b）

質問文	解答例	回答文
2 つの報告書で評価項目として共通している項目は何ですか？	提示された文書は 1 つのみで、また報告書ではなく検定試験の自己評価シートであるため解答できません。	Qwen: 文書には「検定試験の自己評価シート」という 1 つの報告書のみが記載されています。質問で言及されている「2 つの報告書」は文書内に存在しないため、評価項目の共通点を特定することはできません。 → **5.0** LLM-jp: ありません。 → **1.0**
この研究の目的は何ですか？	提示された文書は、コンテンツマッチ広告を提案するための文書であるため解答できません。	LLM-jp: この研究の目的は、人間の脳がどのように音やイメージを知覚するかを示すことです。 → **1.0**

相対的に向上の余地が認められる．当該モデルと gemma-3-12b-it を比較するとマニュアルのみスコアが逆転している一方でウェブサイトや広報では優位性が高いなど，本ベンチマークによりモデルの詳細な特徴を比較検討することが可能である．

3.3　回答不能問題の分析

表 1 に示す回答不能問題に対する挙動を検討する．最初の質問について，Qwen3-VL-8B-Instruct は回答不能である理由を的確に説明している一方，llm-jp-3-vila-14b [21] は簡潔な否定応答のみに留まっている．次の事例では，llm-jp-3-vila-14b の応答は実際には文書と無関係である．回答不能問題ではこのようなハルシネーションが誘発されやすいと考えられ，今後より詳細に分析を行う予定である．

3.4　複数ページの処理能力

特に日本語モデルにおいて，入力ページ数の増加に伴い記号の羅列や単語の反復といった，破綻した回答が確認された．これは入力画像数の増大による視覚トークン数の増加に対し，モデルの処理能力が追いついていないことが原因と考えられる [22]．Qwen3-VL では上述のような出力の崩壊は確認されなかったが，これは同モデルが数十ページ規模の文書画像を処理可能なモデル [12] として調整されているためである．このように多量の視覚トークンを扱うモデルの構築は，日本語 VLM 開発における今後の課題である．

4　おわりに

本研究では，日本語の視覚文書理解能力を測る Omni-JDocVQA を提案した．本ベンチマークを用いた評価により，現実的な QA に基づく汎用的な性能評価を可能とする．一方で，データセット自体の正当性に課題が残る．具体的には，評価スコアの妥当性や，解答例の網羅性不足による誤判定の有無を検証できていない．今後は，これらの精査を通じてベンチマークの信頼性を確立させる必要がある．

謝辞

本研究成果は，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．本ベンチマークのアノテーションは株式会社バオバブに実施いただきました．

参考文献

[1] Yihao Ding, Siwen Luo, Yue Dai, Yanbei Jiang, Zechuan Li, Geoffrey Martin, and Yifan Peng. A Survey on MLLM-based Visually Rich Document Understanding: Methods, Challenges, and Emerging Trends, 2025.

[2] Minesh Mathew, Dimosthenis Karatzas, and C.V. Jawahar. DocVQA: A Dataset for VQA on Document Images. In **Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)**, pp. 2200–2209, January 2021.

[3] Minesh Mathew, Viraj Bagal, Rubèn Pérez Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V Jawahar. InfographicVQA. **arXiv:2104.12756**, 2021.

[4] Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. SlideVQA: A Dataset for Document Visual Question Answering on Multiple Images. In **AAAI**, 2023.

[5] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning. In **Findings of the Association for Computational Linguistics: ACL 2022**, pp. 2263–2279, Dublin, Ireland, 2022.

[6] Jordy Van Landeghem, Rubèn Pérez Tito, Łukasz Borchmann, Michal Pietruszka, Pawel J'oziak, Rafal Powalski, Dawid Jurkiewicz, Mickaël Coustaty, Bertrand Ackaert, Ernest Valveny, Matthew B. Blaschko, Sien Moens, and Tomasz Stanislawek. Document Understanding Dataset and Evaluation (DUDE). **2023 IEEE/CVF International Conference on Computer Vision (ICCV)**, pp. 19471–19483, 2023.

[7] Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, Pan Zhang, Liangming Pan, Yu-Gang Jiang, Jiaqi Wang, Yixin Cao, and Aixin Sun. MMLONGBENCH-DOC: benchmarking long-context document understanding with visualizations. In **Proceedings of the 38th International Conference on Neural Information Processing Systems**, 2024.

[8] Eri Onami, Shuhei Kurita, Taiki Miyanishi, and Taro Watanabe. JDocQA: Japanese Document Question Answering Dataset for Generative Language Models. In **Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)**, pp. 9503–9514, 2024.

[9] Stockmark Inc. BusinessSlideVQA, 2025. https://github.com/stockmarkteam/business-slide-questions.

[10] Ricoh Co., Ltd. JGraphQA, 2025. https://huggingface.co/datasets/r-g2-2024/JGraphQA.

[11] Michał Turski, Tomasz Stanisławek, Karol Kaczmarek, Paweł Dyda, and Filip Graliński. CCpdf: Building a High Quality Corpus for Visually Rich Documents from Web Crawl Data. In **Document Analysis and Recognition - ICDAR 2023: 17th International Conference, San José, CA, USA, August 21–26, 2023, Proceedings, Part III**, p. 348–365, 2023.

[12] Qwen Team. Qwen3-Vl Technical Report. 2025.

[13] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of Tricks for Efficient Text Classification. **arXiv preprint arXiv:1607.01759**, 2016.

[14] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. FastText.zip: Compressing text classification models. **arXiv preprint arXiv:1612.03651**, 2016.

[15] Hayato Tsukagoshi and Ryohei Sasano. Ruri: Japanese General Text Embeddings, 2024.

[16] S. Lloyd. Least squares quantization in PCM. **IEEE Transactions on Information Theory**, Vol. 28, pp. 129–137, 1982.

[17] Qwen Team. Qwen3 Technical Report, 2025.

[18] OpenAI. gpt-oss-120b & gpt-oss-20b Model Card, 2025.

[19] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge With MT-bench and Chatbot Arena. In **Proceedings of the 37th International Conference on Neural Information Processing Systems**, 2023.

[20] カラクリ株式会社. カラクリ、日本企業初の Computer-Using Agent「KARAKURI VL」を公開 - コンピュータ操作を完全自動化できる AI エージェント, 2025. https://karakuri.ai/news/GENIAC.

[21] 笹川慶人, 前田航希, 杉浦一瑳, 栗田修平, 岡崎直観, 河原大輔. LLM-jp-3 VILA: 日本語マルチモーダルデータセット及び強力な日本語マルチモーダルモデルの構築. 言語処理学会第 31 回年次大会発表論文集, pp. 1185–1190, 2025.

[22] MMLongBench: Benchmarking Long-Context Vision-Language Models Effectively and Thoroughly, author=Zhaowei Wang and Wenhao Yu and Xiyu Ren and Jipeng Zhang and Yu Zhao and Rohit Saxena and Liang Cheng and Ginny Wong and Simon See and Pasquale Minervini and Yangqiu Song and Mark Steedman. In **The 39th (2025) Annual Conference on Neural Information Processing Systems**, 2025.

表 2　文書ラベルの定義

ドメイン		フォーマット	
ラベル	説明	ラベル	説明
学術	学問研究や専門知識	レポート	調査・分析・活動の結果の報告
技術	科学的知識の応用や製品開発	公的文書	公的機関が発行する法的拘束や手続き
ビジネス	企業活動、市場、金融等	入力フォーム	情報を記入・選択するための文書
医療	医学、健康、治療、美容等	マニュアル	使い方や手順、利用方法の説明
文化	芸術、文学、歴史などの創作物	広報	商品・サービス・イベントなどをの告知等
社会	公共システムや社会課題等	創作物	物語や知識の提供、芸術表現
教育	学習活動や教育制度等	スライド	口頭発表やプレゼンテーションの補助資料
生活	日常生活、趣味、消費活動	ウェブサイト	ウェブ上に公開されることを前提とした文書
その他	上記ラベルに該当しない文書	その他	上記ラベルに該当しない文書

付録

A　文書ラベル

文書ラベルの定義を表 2 に示す.

B　プロンプト

本研究で使用したプロンプトを以下に示す.

テキスト抽出用プロンプト

以下の文書画像のテキスト、表、数式を改変せずに Markdown 形式で転記してください。
転記した内容以外は生成しないでください。

文書ラベリング（domain）用プロンプト

あなたは文書の種類を分類するアノテーターです。
次に示す文書に適切なラベル（domain）を、必ず 1 つ以上選んでください。

domain（文書に記載されている内容分野を表す）
- 学術：学問研究や専門知識に関する分野
- 技術：科学的知識の応用や製品・システム開発に関する分野
- ビジネス：企業活動、市場、金融、商業に関する分野
- 医療：医学、健康、治療、美容に関する分野
- 文化：芸術、文学、歴史、宗教など人間の創作物に関する分野
- 社会：社会のルール、公共システム、社会全体の課題に関する分野
- 教育：学習・教育活動や教育制度に関する分野
- 生活：日常生活、趣味、家庭、消費活動に関する分野
- その他：上記ラベルに該当しない分野

条件：
- ラベルは必ず 1 つ以上選んでください。
- ラベル以外は出力しないでください。

文書ラベリング（format）用プロンプト

あなたは文書の種類を分類するアノテーターです。
次に示す文書に適切なラベル（format）を、必ず 1 つ以上選んでください。

format（文書の形式や目的を表す）
- レポート：調査、分析、活動の結果を報告する文書
- 公的文書：公的機関が発行する法的拘束や手続きを目的とする文書
- 入力フォーム：情報を記入・選択することを目的とした文書
- マニュアル：使い方や手順、利用方法を説明するための文書
- 広報：商品、サービス、イベントなどを広く知らせるための文書
- 創作物：物語や知識の提供、芸術表現を目的とした文書
- スライド：口頭発表やプレゼンテーションの補助を目的とした文書
- ウェブサイト：ウェブ上に公開されることを前提とした文書
- その他：上記ラベルに該当しない文書

条件：
- ラベルは必ず 1 つ以上選んでください。
- ラベル以外は出力しないでください。

質問生成用プロンプト

あなたは文書に書かれた内容から必要な回答から得ることを目的とした質問を作成するアノテーターです。
これから提供する文書を参照する状況に基づいて、質問を 3 件作成してください。
ただし、以下の条件に従ってください。

1. 質問は明確で具体的な内容で構成してください
2. 誰がいつ質問しても同じ回答が得られる質問にしてください
3. 質問および回答には、専門知識を必要としない内容にしてください
4. 文書から答えが得られる想定で質問を作成してください
5. 客観的な事実を求める質問を作成してください
6. 文書を読まないと回答できない質問を作成してください
7. 固有名詞は避けて作成してください
8. 流暢で自然な日本語で作成してください
9. 状況から上記の条件を満たす質問が作成できない場合は、'None' とだけ答えてください

評価用プロンプト

文書に対して質問応答を行うモデルの回答を評価してください。

評価の対象:
評価対象として、以下の 3 つ組が与えられます。
- 質問: 文書の内容に関する質問
- 正解: 文書に基づいた、正確かつ必要十分な回答
- 回答: 評価対象のモデルが生成した回答

回答の条件:
評価対象の回答は、以下の条件で生成されたものである。
1. 回答は、文書に書かれている客観的な事実のみをもとに生成すること。
2. 回答は、日本語の文章で生成すること。
3. 文書の内容から回答が生成できない場合は、回答できない理由を生成すること。

スコアの基準:
質問の文脈において、回答がどれだけ正解と意味的に一致しているかを基準とし、以下の 5 段階で採点してください。
5: 優秀 - 正解と完全に一致している。
4: 良い - 正解と回答に軽微な違いがある。
3: 普通 - 正解と回答に明確な違いがある。
2: 悪い - 正解と回答は大きく異なっている。
1: 非常に悪い - 正解と解答が完全に異なっているまたは、回答が日本語を主に書かれてない場合や、正解として回答できない理由が必要であるにもかかわらず、その理由が回答に明記されていない場合。

出力形式:
- スコア（1〜5 の整数）のみを出力してください。
- 説明は一切追加しないこと。

質問: {Question}
正解: {Answer}
回答: {Prediction}
スコア: