BizDocVQA

実世界ビジネス帳票に対する根拠付きVQAデータセットの提案

NLP 2026

Takahiro Kubo

Developer Relations Engineer
Amazon Web Services Japan G.K.

Agenda

1. はじめに
2. BizDocVQA データセット
3. モデル評価
4. ANLS × IoU 乖離分析
5. ハルシネーションリスク評価
6. おわりに

はじめに

BizDocVQA とは?

日本語レシートに対する根拠付き VQA データセット

データセット概要

- 実世界の日本語レシート **116 枚**
- 経理実務に基づく **624 件** の QA ペア
- 各 QA に **回答の根拠領域(BBox)** を付与

2 軸評価

- **ANLS**: テキストとしての回答精度
- **IoU**: 回答根拠の空間的正確さ

なぜ 2 軸が必要か?

- ビジネス上の評価では、テキストの正しさに加え**根拠**が重要
- 読取の正確性は ANLS / 回答根拠の位置を IoU で評価

📊 テキスト精度 + 空間的根拠の 2 軸で VLM の文書理解能力を評価

背景 1：正答 = 根拠が正しいわけではない

ANLS メトリクスの盲点

- 回答テキストの編集距離のみで評価 → 根拠性（groundedness）を考慮しない
- ハルシネーションでも字面が近ければ高スコアになる
- モデルが「**どの領域を見て回答したか**」を検証する手段がない

先行研究でも空間的理解の低さが判明

- **BBox-DocVQA** (Yu et al., 2025): エビデンス BBox を付与した初の大規模データセット
  - Qwen2.5VL-72B でも平均 IoU が **40% 未満**
  - 正しい回答を出力しながら無関係な領域を参照するケースが多い
- **Nourbakhsh et al.** (NAACL Findings 2025)
  - ANLS の根拠性欠如を指摘

背景 2：実帳票かつ VQA のデータセットが希少

整形された PDF 文書の VQA、実画像だが VQA でない状況

データセット	言語	文書タイプ	根拠領域	VQA	規模
BizDocVQA	**日本語**	**レシート**	**回答根拠BBox**	○	**624 QA**
DocVQA (2021)	英語	一般文書	なし	○	50K QA
SROIE (2019)	英語	レシート	テキストBBox	×	1K画像
CORD (2019)	インドネシア語	レシート	ボックス	×	数千枚
JDocQA (2024)	日本語	PDF文書	ページ＋BBox	○	11.6K QA
BBox-DocVQA (2025)	英語	学術論文	エビデンスBBox	○	32K QA
OCRBench v2 (2025)	英中	多種	IoU	○	10K QA

BizDocVQA データセット

BizDocVQA データセットの詳細

レシート画像に対する質問・回答・根拠を収録したデータセット。

アノテーション構成

- **レシート画像**: 実世界で撮影された日本語レシート
- **質問**: 経費精算等で必要な 7 種類のフィールド
- **回答**: テキスト形式
- **根拠領域**: 正規化座標 [x0, y0, x1, y1]（0〜1）

データ仕様

項目	内容
アノテーション数	624 件
画像数	116 枚
言語	日本語
文書タイプ	レシート
質問タイプ	抽出型
評価指標	ANLS + IoU
ライセンス	CC BY-SA 4.0

質問タイプと評価指標

情報不在時は空文字列 + 画像全体の BBox を付与し「該当なし」の判定能力も評価できるように。

7 種類の質問フィールド

1. **日付** — 取引日
2. **登録番号** — インボイス制度の T+13桁
3. **合計金額（10%対象）** — 標準税率
4. **合計金額（8%対象）** — 軽減税率
5. **購入品目** — 商品名
6. **ポイント** — ポイント情報
7. **伝票番号** — レシート番号

評価指標

ANLS（テキスト精度）

- Average Normalized Levenshtein Similarity
- OCR 誤差を許容しつつ回答精度を測定

IoU（空間的根拠精度）

- Intersection over Union
- 予測 BBox と正解領域の重なり度合い

モデルによる評価

Bedrock 上の 8 モデルを統一条件で評価

共通のプロンプトで回答と根拠領域 (BBox) を同時に生成させ、ANLS と IoU で比較

モデル	規模
Opus 4.6	非公開（大規模）
Sonnet 4.6	非公開（中規模）
Sonnet 4.5	非公開（中規模）
Haiku 4.5	非公開（軽量）
Qwen3 VL 235B	235B (A22B)
Nova Pro	非公開
Gemma 3 27B	27B
Nemotron 12B	12B

評価設定

- **データ**: 624 件のアノテーション
- **温度**: 0.1
- **プロンプト**: 共通（回答 + BBox 生成）
- **不在指示**: 該当情報なし → 空文字列を返却

テキスト精度と根拠領域の特定には大きな差がある

上位モデルが ANLS 0.76 以上を記録する一方、IoU は最高でも 0.31 — 下位モデルは 0.01 以下

モデル	ANLS	IoU
Opus 4.6	**0.9037**	**0.3144**
Sonnet 4.6	0.8724	0.2953
Sonnet 4.5	0.8171	0.1629
Qwen3 VL 235B	0.7598	0.1826
Haiku 4.5	0.7009	0.1220
Gemma 3 27B	0.6149	0.0390
Nemotron 12B	0.5803	0.0174
Nova Pro	0.4590	0.0097

Opus 4.6 の精度は先行研究に沿う

- BBox-DocVQA で報告された最高精度 (~0.40) と同等
- 唯一、 IoU > 0.3 を達成

ANLS と IoU の非対称性

- ANLS の差は最大 0.44pt
- IoU の差は最大 0.30pt
- IoU はどのモデルにとっても難しい

情報不在のケースではハルシネーションが多い

質問タイプ	ANLS（平均）	難易度
日付	0.981	★☆☆☆☆
登録番号	0.938	★★☆☆☆
合計金額（10%対象）	0.627	★★★☆☆
購入品目	0.603	★★★☆☆
合計金額（8%対象）	0.508	★★★★☆
ポイント	0.451	★★★★☆
伝票番号	0.185	★★★★★

難易度の背景

※ プロンプトでは存在しない場合回答しないよう指示

- 高難易度の項目は欠損が多い : 軽減税率の 8% やポイント・伝票番号
- モデルが「聞かれたら何か答える」挙動を学習している可能性

回答精度 x 根拠領域の乖離分析

正答の約 57%~ が誤った空間領域を参照している

ANLS ≥ 0.5 を正答 (C=Correct)、IoU ≥ 0.3 を根拠正確(G=Grounded)として 4 象限に分類
最大 99% が根拠不明

モデル	C+G	C+UG	W+G	W+UG
Opus 4.6	**39.4%**	52.7%	0.3%	7.5%
Sonnet 4.6	28.7%	60.6%	4.5%	6.3%
Qwen3 235B	17.5%	61.1%	2.4%	19.1%
Sonnet 4.5	16.8%	66.8%	1.0%	15.4%
Haiku 4.5	13.5%	61.1%	2.1%	23.4%
Gemma 27B	1.9%	62.3%	0.5%	35.3%
Nemotron 12B	0.3%	59.9%	1.1%	38.6%
Nova Pro	0.2%	47.9%	0.3%	51.6%

C+UG が支配的

- 全モデルで C+UG（正答だが根拠不正確）が最大カテゴリ
- Opus 4.6 でも C+G は 39.4% にとどまる

凡例

- C+G = 正答＋根拠正確（Correct + Grounded）
- C+UG = 正答＋根拠不正確（Correct + Ungrounded）
- W+G = 誤答＋根拠正確（Wrong + Grounded）
- W+UG = 誤答＋根拠不正確（Wrong + Ungrounded）

根拠が画像内にあっても 72% ~ が根拠不正確

情報不在サンプルを除外しても Ungrounded 率は改善せず、空間的理解の課題は本質的

Correct+Grounded 率

- Opus 4.6: **39.4%**
- Gemma 3 27B: **1.9%**
- Nemotron 12B: **0.3%**
- Nova Pro: **0.2%**

根拠があるの場合の Ungrounded 率

- Opus 4.6: **72.3%**
- Sonnet 4.5: **93.3%**
- Nemotron 12B: **100%**
- Nova Pro: **100%**

💡 情報不在サンプルの取り扱いがベンチマーク設計の重要な考慮事項

正答でも根拠が正しいとは限らない①

ANLS はどちらも 1.0 で区別できないが、根拠領域は全く異なる

Opus 4.6：IoU 0.32 ✅

Gemma 27B：IoU 0.0 ❌

正答でも根拠が正しいとは限らない②

登録番号(T1010401090638)は双方正答だが、根拠領域が異なる。下記は Qwen が双方成功

Qwen3 235B：IoU 0.34 ✅

Nova Pro：IoU 0.0 ❌

ハルシネーションリスク評価

回答がない場合、ハルシネーション率が高くなる

Opus がハルシネーション率 14% と正確に回答可否を判断する一方、90% を超えるモデルも。
VLM が「常に回答を生成する」ことに強いバイアスを持つことがうかがえる。

モデル	ハルシネーション率
Claude Opus 4.6	**13.7%**
Claude Sonnet 4.6	32.4%
Claude Sonnet 4.5	46.0%
Qwen3 VL 235B	67.6%
Claude Haiku 4.5	92.1%
Gemma 3 27B	98.6%
Nemotron 12B	98.6%
Nova Pro	99.3%

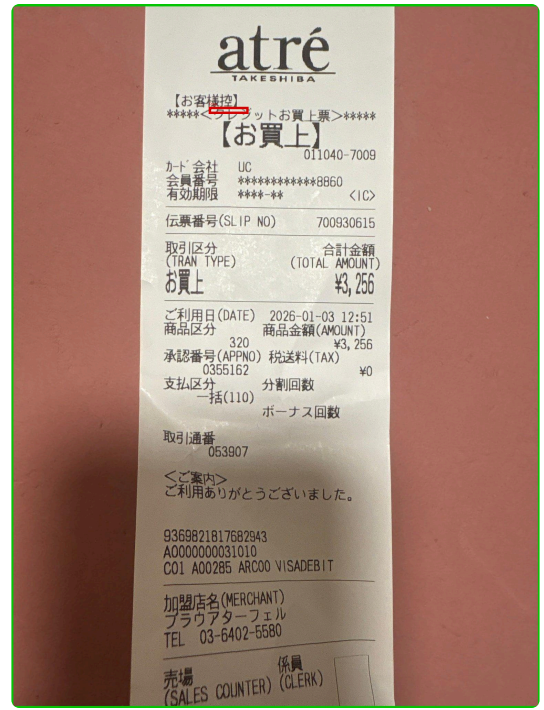
見落とし率は低い（False Negative）

- ほぼ全モデルで **2.1% 以下**
- 例外: Nemotron Nano 12B（7.0%）

存在しない登録番号を「T+数字列」で生成

登録番号パターン（T+13桁）を認識しており、その事前知識を基に"それらしい"値を生成している。

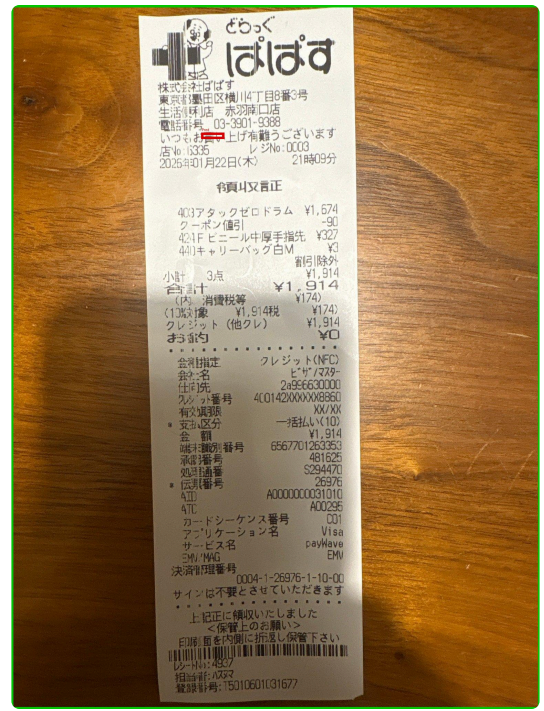
モデル	予測	判定
Opus 4.6	（空文字列）	✅
Sonnet 4.6	（空文字列）	✅
Sonnet 4.5	（空文字列）	✅
Qwen3 235B	T011040-7009	❌
Haiku 4.5	011040-7009	❌
Nova Pro	T000930615	❌
Gemma 27B	T00930615	❌
Nemotron 12B	936982181…	❌



10% 対象の金額を 8% 対象に誤帰属

全品目が標準税率（10%）対象で 8% 金額は存在しない場合、複数モデルが 10% の金額「1914」を回答。

モデル	予測	判定
Opus 4.6	（空文字列）	✅
Sonnet 4.5	1914	❌
Qwen3 235B	1914	❌
Haiku 4.5	1914	❌
Nova Pro	1,914	❌
Gemma 27B	914	❌
Nemotron 12B	1914	❌



登録番号をそのまま伝票番号として回答

伝票番号はないが「登録」番号はあるケースで、5 モデルがフィールドを混同。

モデル	予測	判定
Opus 4.6	（空文字列）	✅
Nova Pro	T901060103…	❌ 混同
Gemma 27B	T901060103…	❌ 混同
Nemotron 12B	T901060103…	❌ 混同
Qwen3 235B	T901060103…	❌ 混同
Haiku 4.5	T901060103…	❌ 混同
Sonnet 4.5	4129-4937	❌ 誤認
Sonnet 4.6	4129-4997	❌ 誤認

パターン生成と誤帰属が実務上大きな影響を与える

パターン生成

学習済み知識から存在しない値を生成

- 「T」で始まる番号のハルシネーション
- 登録番号パターンの過学習

頭文字や番号体系で入力チェックをしていた場合、チェックを回避し存在しない値が入力される可能性がある

誤帰属

実在する数値を異なるフィールドに帰属

- 10% 金額 → 8% 金額に誤帰属
- 登録番号 → 伝票番号に誤帰属

演算での誤りに直結

おわりに

まとめ

BizDocVQA の貢献

1. **初の根拠付き帳票 VQA データセット**
   - 実世界日本語レシート 116 枚・624 件
   - ANLS + IoU の 2 軸評価
2. **8 モデルの体系的評価**
   - Amazon Bedrock 上で統一条件
   - テキスト精度と空間的根拠の乖離を定量化
3. **ハルシネーション分析**
   - 情報不在時の応答パターンを類型化
   - 実務リスクの定量的評価

主要な知見

🔍 正答サンプルの **約 80%** が誤った空間領域を参照

⚠️ テキスト精度の高さは空間的理解を**保証しない**

💡 VLM は「常に回答を生成する」**強いバイアス**を持つ

📊 ハルシネーション耐性はモデル選定の**重要な評価軸**

今後の展望

データセットの拡張と自動構築パイプラインの適用

スケーリング計画

- **対象拡大**: 請求書・領収書など他のビジネス帳票
- **多言語化**: 日本語以外の帳票を集約
- **規模**: 数千件規模のデータセットへ発展

自動構築パイプライン

- BBox-DocVQA の Segment–Judge–and–Generate
  - SAM による領域セグメンテーション
  - VLM による意味的判定
  - 高性能 VLM による QA 自動生成
- 手動アノテーションのガイドラインをシード情報として活用

統合型ベンチマークへの貢献

- OCRBench v2 等への組み込み
- 日本語ビジネス文書ドメインの拡張
- より正確なモデル性能評価に貢献

データセット公開

🤗 Hugging Face にて公開中

`icoxfog417/biz-doc-vqa`

https://huggingface.co/datasets/icoxfog417/biz-doc-vqa

参考文献

- Mathew, M., Karatzas, D., Jawahar, C.V. (2021). DocVQA: A Dataset for VQA on Document Images. *WACV 2021*.
- Huang, Z. et al. (2019). ICDAR2019 Competition on Scanned Receipt OCR and Information Extraction. *ICDAR 2019*.
- Park, S. et al. (2019). CORD: A Consolidated Receipt Dataset for Post-OCR Parsing. *NeurIPS 2019 Workshop*.
- Yu, W. et al. (2025). BBox-DocVQA: Bounding-Box-Grounded Dataset for Document VQA. *arXiv:2511.15090*.
- Nourbakhsh, A. et al. (2025). Where is this coming from? Making groundedness count in Document VQA. *NAACL Findings 2025*.
- Onami, E. et al. (2024). JDocQA: Japanese Document QA Dataset. *LREC-COLING 2024*.
- Fujitake, M. (2024). JaPOC: Japanese Post-OCR Correction Benchmark. *arXiv:2409.19948*.
- Liu, Y. et al. (2024). OCRBench: On the Hidden Mystery of OCR in Large Multimodal Models. *SCIS*.
- Fu, L. et al. (2025). OCRBench v2: Improved Benchmark for LMMs on Visual Text. *arXiv:2501.00321*.

ありがとうございました

BizDocVQA: https://huggingface.co/datasets/icoxfog417/biz-doc-vqa