

DISCO: DOCUMENT INTELLIGENCE SUITE FOR COMPARATIVE EVALUATION

Kenza Benkirane

Parexel AI Labs
London, United Kingdom
kenza.benkirane@parexel.com

Dan Goldwater

Parexel AI Labs
London, United Kingdom
dan.goldwater@parexel.com

Martin Asenov

Parexel AI Labs
London, United Kingdom
martin.asenov@parexel.com

Aneiss Ghodsi

Parexel AI Labs
San Francisco, United States
aneiss.ghodsi@parexel.com

ABSTRACT

Document intelligence requires accurate text extraction and reliable reasoning over document content. We introduce **DISCO**, a *Document Intelligence Suite for Comparative Evaluation*, that evaluates optical character recognition (OCR) pipelines and vision-language models (VLMs) separately on parsing and question answering across diverse document types, including handwritten text, multilingual scripts, medical forms, infographics, and multi-page documents. Our evaluation shows that performance varies substantially across tasks and document characteristics, supporting a document-centered approach to document intelligence. OCR pipelines are generally more reliable for handwriting and for long or multi-page documents, where explicit text grounding supports text-heavy reasoning, while VLMs perform better on multilingual text and visually rich layouts. Task-aware prompting substantially improves VLM performance in some settings, particularly multilingual and visually structured documents, though its benefits are not uniform across all tasks. Taken together, these findings favour a document-centered view of document intelligence, in which method choice is driven by document characteristics rather than by model family alone.¹

1 INTRODUCTION AND RELATED WORK

Documents remain a primary source of information across industries, yet extracting and reasoning over their content poses persistent challenges. Traditional approaches rely on OCR to convert images to text, followed by language models for downstream tasks. Recent VLMs offer an alternative: processing document images directly without explicit text extraction. This raises a practical question: when should practitioners use OCR pipelines versus end-to-end VLMs? Current benchmarks report only final task accuracy, making it difficult to diagnose whether failures stem from text extraction or reasoning.

Existing benchmarks and their gaps. Document understanding evaluation has evolved from isolated OCR benchmarks to integrated question answering (QA) datasets. DocVQA (Mathew et al., 2021) requires reading and reasoning over forms and letters; InfographicVQA (Mathew et al., 2022) targets visual reports;

¹Find the code, and datasets on:

GitHub: <https://github.com/kenza-ily/disco>

Hugging-Face: <https://huggingface.co/collections/kenza-ily/disco>

Task	Dataset	Description
Parsing	IAM _{DISCO} from Marti & Bunke (2002)	Handwritten text
Parsing	ICDAR 2015 _{DISCO} from Karatzas et al. (2019)	Multilingual scene text
Parsing	PubLayNet _{DISCO} from Zhong et al. (2019)	Scientific document pages (layout)
Parsing	RxPad _{DISCO} from Pattin et al. (2026)	Medical forms (French)
QA	DocVQA _{DISCO} from Mathew et al. (2021)	Scanned forms and letters
QA	InfographicVQA _{DISCO} from Mathew et al. (2022)	Infographics
QA	DUDE _{DISCO} from Van Landeghem et al. (2023)	Heterogeneous multi-page documents
QA	ChartQAPro _{DISCO} from Masry et al. (2025b)	Charts
QA	VisR-Bench _{DISCO} from Chen et al. (2025)	Long multi-page documents

Table 1: Benchmark suite composition: datasets and document types evaluated in DISCO.

Notes. VisR-Bench and PubLayNet are included in the suite but we do not report experimental results on them in this paper: PubLayNet requires layout-level evaluation metrics not yet integrated into our pipeline, and VisR-Bench experiments were not completed within the compute budget for this study. Both remain active targets for future work.

DUDE (Van Landeghem et al., 2023) introduces multi-page, multi-domain documents. ChartQAPro (Masry et al., 2025a) revealed that performance on narrow benchmarks overestimates generalisation: models achieving $\sim 90\%$ on ChartQA dropped to $\sim 56\%$ on more diverse charts. Despite this progress, benchmarks typically report end-to-end accuracy without isolating error sources. When a system fails, we cannot determine whether OCR missed the text, the layout was misinterpreted, or the reasoning was flawed.

OCR pipelines versus VLMs. OCR-first pipelines benefit from decades of optimisation for text recognition and can handle long documents efficiently, but may lose spatial relationships when converting content to plain text. VLMs process images holistically, preserving layout cues, but require high resolution to read fine print and may hallucinate when text is unclear. Recent hybrid approaches, such as DocVLM (Nacson et al., 2025), inject OCR-detected text as additional tokens into frozen VLMs and achieve strong results on DUDE. This suggests that the two paradigms may be complementary rather than competing; however, systematic comparison across document types remains lacking.

Our contribution. Our contribution is empirical rather than methodological: we introduce DISCO, a diagnostic evaluation framework for document intelligence that explicitly separates text parsing from downstream question answering. Rather than treating document understanding as a single end-to-end prediction problem, DISCO evaluates intermediate representations and final answers under controlled pipeline variants, enabling attribution of errors to perception, representation, or reasoning stages.

Using this stage-wise protocol, we report three empirical observations. First, model behaviour depends strongly on document structure: OCR-based pipelines are generally more reliable for long and multi-page documents, while VLM-based approaches tend to be stronger on multilingual text and visually structured content such as infographics and forms. Second, task-aware prompting has heterogeneous effects across datasets and model families, helping on some document types but offering limited gains on challenging domains such as medical prescriptions. Third, direct visual question answering can be advantageous on single-page documents, suggesting that intermediate text extraction may introduce information loss when spatial layout is central to the task.

Overall, DISCO reframes document intelligence evaluation, turning end-to-end accuracy into actionable diagnosis and revealing not just which system wins, but why it wins. By analysing intermediate representations in addition to final predictions, our framework highlights the limits of assessing multimodal systems solely through end-to-end accuracy and motivates evaluation beyond next-token prediction, particularly for documents where perception, layout, and reasoning interact.

2 DATASETS AND BENCHMARK SUITE

We construct a benchmark suite designed to evaluate document intelligence systems across two core tasks: text parsing and question answering. Rather than relying on a single dataset, we combine multiple established benchmarks to ensure coverage of diverse document types, visual characteristics, languages, and reasoning requirements. This choice reflects the heterogeneous nature of real-world documents and allows us to assess models under a wide range of conditions. By evaluating across these datasets, we aim to capture both low-level perception challenges and higher-level reasoning and grounding behaviour. The datasets are summarised in Table 1.

To ensure feasibility and reproducibility, we restrict each dataset to fewer than 500 samples. For large-scale benchmarks, we construct dedicated *DISCO* versions by sampling small but representative subsets. Sampling is performed to preserve key properties of the original datasets; including document structure, question and answer types, language distribution, and difficulty. When metadata is available, stratified sampling is used to avoid bias towards simpler instances. Datasets with fewer than 500 samples are used in full. This design enables consistent evaluation across tasks while keeping computational cost manageable. It also allows us to apply a unified experimental protocol and identical model configurations across datasets, facilitating fair comparisons between approaches.

A summary of all datasets, their associated tasks, and their main characteristics is provided in Table 1. Detailed dataset descriptions, sampling procedures, and statistics for the *DISCO* versions are reported in Appendix E.

3 METHODOLOGY AND EXPERIMENTAL DESIGN

We evaluate document intelligence by separating parsing from question answering, testing both OCR systems and VLMs with prompt variations² to assess capabilities at each stage. Throughout, we write metric scores using the notation S_M , where M is the metric name (e.g., S_{CS} for cosine similarity). In each condition, we compare OCR systems (azure-ai-documentintelligence, mistral-ocr-2505 and for QA, mistral-ocr-2512) against VLMs (gpt-5-mini, gpt-5-nano, claude-3-5-sonnet). All experiments use deterministic decoding and a fixed image resolution for fair comparison. Bootstrap 95% confidence intervals were computed using 10,000 iterations of sample-level resampling. Statistical comparisons used Wilcoxon signed-rank tests ($\alpha = 0.05$).

Text parsing extracts textual content from document images, including printed text, handwriting, multilingual scripts, and structured layouts. We evaluate on three datasets: IAM_{DISCO} (handwriting; Marti & Bunke (2002)), ICDAR_{DISCO} (multilingual scene text; Karatzas et al. (2019)), RxPad (Pattin et al. (2026)). Our experiments are as follows: P_{OCR} (OCR-only parsing), $P_{VLM-base}$ (VLM base prompt parsing), and $P_{VLM-task}$ (parsing with task-aware prompt). We report S_{CS} (cosine similarity between embeddings of extracted and ground-truth text; higher is better), S_{CER} (character error rate: normalised character-level edit distance; lower is better), and S_{WER} (word error rate: normalised word-level edit distance; lower is better).

Question answering evaluates answer generation from document content. We test on DocVQA_{DISCO} (forms; Mathew et al. (2021)), InfographicVQA_{DISCO} (infographics; Mathew et al. (2022)) and DUDE_{DISCO} (multi-page documents; Van Landeghem et al. (2023)). Our experiments are as follows: QA_{OCR} (OCR parsing \rightarrow LLM QA), $QA_{VLM-2stage}$ (VLM parsing \rightarrow LLM QA; two-stage), and $QA_{VLM-direct}$ (direct VLM QA). For VLM, we use the convention $QA_{VLM-2stage/VLM-direct}^{generic/cot/task-aware}$, where the subscript denotes the pipeline and the superscript denotes the prompt, e.g., $QA_{VLM-direct}^{generic}$, $QA_{VLM-direct}^{cot}$, and $QA_{VLM-direct}^{task-aware}$. For the parsing in the QA tasks, we also use mistral-ocr-2512. Each pipeline is tested with simple, detailed, and

²Prompts are available in Section G.

context-aware prompts. We report $S_{\text{GT-in-Pred}}$ (ground-truth-in-prediction: substring match indicator; higher is better), S_{ANLS} (average normalised Levenshtein similarity; higher is better), and S_{EM} (exact match rate; higher is better). Prompts do not explicitly constrain the output format (e.g., “respond with only the answer”); this was a deliberate design choice to evaluate default model behaviour, but it contributes to the gap between $S_{\text{GT-in-Pred}}$ and stricter metrics (see Section B.2.5).

4 RESULTS AND DISCUSSION

Task	Dataset	Evaluated aspect	P_{OCR}	$P_{\text{VLM-base}}$	$P_{\text{VLM-task}}$
Parsing S_{CER}	IAM _{DISCO}	Handwritten text	0.099 ± 0.013	0.175 ± 0.014	0.151 ± 0.014
	ICDAR _{DISCO}	Multilingual scene text	3.193 ± 1.082	2.044 ± 0.530	0.512 ± 0.029
	RxPad	Medical prescriptions	0.708 ± 0.082	0.658 ± 0.017	0.658 ± 0.016
Task	Dataset	Evaluated aspect	Q_{AOCR}	$Q_{\text{AVLM-2stage}}$	$Q_{\text{AVLM-direct}}$
QA $S_{\text{GT-in-Pred}}$	DocVQA _{DISCO}	Single-page documents	0.841	0.851	0.880
	InfographicVQA _{DISCO}	Visually structured layouts	0.735	0.637	0.778
	DUDE _{DISCO}	Multi-page documents	0.517	0.488	0.498

Table 2: DISCO benchmark results across parsing and question answering. The “Evaluated aspect” column indicates the main document property tested by each dataset. Parsing results are reported with 95% bootstrap confidence intervals as point estimate ± margin. Bold indicates the best result per row.

4.1 PARSING

OCR performs best on handwriting. On IAM_{DISCO}, OCR achieved the lowest S_{CER} at 0.099 ± 0.013 , compared with 0.175 ± 0.014 for $P_{\text{VLM-base}}^{\text{generic}}$ and 0.151 ± 0.014 for $P_{\text{VLM-task}}^{\text{task-aware}}$. This corresponds to 95% confidence interval ranges of $[0.086, 0.112]$, $[0.161, 0.189]$, and $[0.137, 0.165]$, respectively. The non-overlapping intervals show that OCR retains a clear advantage on handwritten text, even though task-aware prompting reduces the gap relative to generic prompting. This suggests that handwriting remains the setting in which specialised OCR systems are most reliable.

Task-aware prompting substantially improves VLM performance on multilingual scene text. For ICDAR_{DISCO}, OCR reached a S_{CER} of 3.193 ± 1.082 , while $P_{\text{VLM-base}}^{\text{generic}}$ reduced this to 2.044 ± 0.530 and $P_{\text{VLM-task}}^{\text{task-aware}}$ further reduced it to 0.512 ± 0.029 . The corresponding 95% confidence interval ranges are $[2.111, 4.275]$, $[1.514, 2.574]$, and $[0.483, 0.541]$. The separation between the task-aware VLM and the other methods indicates a strong and stable advantage on multilingual visual text. This result shows that prompt design is a major factor for VLM parsing quality in this setting.

Medical prescriptions remain difficult for both approaches. On RxPad, OCR obtained 0.708 ± 0.082 and $P_{\text{VLM-task}}^{\text{task-aware}}$ obtained 0.658 ± 0.016 , corresponding to confidence interval ranges of $[0.626, 0.790]$ and $[0.642, 0.674]$. Although the VLM-task estimate is numerically lower, the intervals overlap and the Wilcoxon signed-rank test does not indicate a significant difference ($p = 0.412$). This suggests that neither pipeline has a clear advantage on prescription parsing. Instead, both appear limited by the difficulty of the document type itself.

4.2 QUESTION ANSWERING

Direct VQA performs best on single-page documents. On DocVQA_{DISCO}, $Q_{\text{AVLM-direct}}^{\text{task-aware}}$ achieved the highest phase-averaged $S_{\text{GT-in-Pred}}$ at 0.880, compared with 0.851 for $Q_{\text{AVLM-2stage}}^{\text{task-aware}}$ and 0.841 for $Q_{\text{AOCR}}^{\text{task-aware}}$.

This ranking is consistent with the best-configuration result reported in Table 7. The result suggests that avoiding intermediate text extraction reduces error propagation on single-page form documents.

Visual structure remains important for infographic QA. On InfographicVQA_{DISCO}, $Q_{\text{VLM-direct}}^{\text{task-aware}}$ again performed best, ahead of both OCR-based and two-stage VLM pipelines. The weaker performance of $Q_{\text{VLM-2stage}}^{\text{task-aware}}$ suggests that text linearisation loses useful spatial and visual cues. High $S_{\text{GT-in-Pred}}$ combined with lower S_{ANLS} or S_{EM} therefore reflects partial answer localisation without exact formatting fidelity.

OCR-based QA remains competitive on long documents. On DUDE_{DISCO}, the OCR-based pipeline achieved the strongest score, indicating that explicit textual grounding remains useful for multi-page reasoning. Direct VQA underperformed in this setting, in contrast to its stronger results on single-page documents. This points to document length and evidence aggregation, rather than model family alone, as the main source of difficulty.

4.3 CONCLUSION AND DISCUSSION

The main finding of this study is that document structure is a stronger determinant of pipeline performance than model family alone. Across both parsing and question answering, the relative ranking of OCR and VLM-based approaches changed systematically with document type, pointing to a document-centered approach to document intelligence, where the choice of pipeline is driven by the structure, modality, and complexity of the document itself.

For parsing, OCR was strongest on handwritten text, where it achieved the lowest error on IAM_{DISCO} and remained clearly separated from VLM-based methods by non-overlapping confidence intervals. By contrast, task-aware VLM prompting performed best on multilingual scene text in ICDAR_{DISCO}, where it reduced character error substantially relative to both OCR and generic VLM prompting. On RxPad, neither family showed a statistically clear advantage, indicating that prescription parsing remains difficult regardless of pipeline choice.

For question answering, direct VLM-based VQA performed best on single-page and visually structured datasets, including DocVQA_{DISCO} and InfographicVQA_{DISCO}. In contrast, OCR-based QA remained strongest on DUDE_{DISCO}, suggesting that explicit textual grounding is still beneficial for longer and more complex documents. Taken together, these results show that the strengths of each method are conditional rather than absolute.

A second finding is that prompting strategy materially affects VLM behaviour. Generic prompting often underperformed task-aware prompting by a large margin, especially in multilingual parsing. This shows that comparisons between OCR and VLM systems can be misleading if prompt design is not controlled carefully.

More broadly, the stage-wise DISCO protocol makes it possible to separate failures in parsing from failures in downstream reasoning. This reveals whether a poor end-to-end result comes from transcription quality, answer generation, or both. The benchmark therefore supports a more precise analysis of document intelligence systems than end-to-end evaluation alone.

Together, these contributions move document intelligence evaluation away from a single-winner framing toward a structure-aware selection framework grounded in empirical evidence.

REFERENCES

- Jian Chen, Ming Li, Jihyung Kil, Chenguang Wang, Tong Yu, Ryan Rossi, Tianyi Zhou, Changyou Chen, and Ruiyi Zhang. Visr-bench: An empirical study on visual retrieval-augmented generation for multilingual long document understanding. *arXiv preprint*, arXiv:2508.07493, 2025. URL <https://arxiv.org/abs/2508.07493>. Accessed 26 Jan 2026.
- Dimosthenis Karatzas et al. Icdar 2019 competition on robust reading. In *Proceedings of ICDAR*, 2019.
- Urs-Victor Marti and Horst Bunke. The IAM database: An english sentence database for offline handwriting recognition. *International Journal on Document Analysis and Recognition*, 5(1):39–46, 2002.
- Ahmed Masry, Mohammed Saidul Islam, Mahir Ahmed, Aayush Bajaj, Firoz Kabir, Aaryaman Kartha, Md Tahmid Rahman Laskar, Mizanur Rahman, Shadikur Rahman, Mehrad Shahmohammadi, Megh Thakkar, Md Rizwan Parvez, Enamul Hoque, and Shafiq Joty. Chartqapro: A more diverse and challenging benchmark for chart question answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19123–19151. Association for Computational Linguistics, 2025a.
- Ahmed Masry et al. Chartqapro: A modern benchmark for chart question answering. *arXiv preprint*, 2025b.
- 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, 2021.
- Minesh Mathew, Viraj Bagal, Rubén Tito, Dimosthenis Karatzas, Ernest Valveny, and C.V. Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2582–2591, 2022.
- Mor Shpigel Nacson, Aviad Aberdam, Roy Ganz, Elad Ben-Avraham, Alona Golts, Yair Kittenplon, Shai Mazor, and Ron Litman. Docvlm: Make your VLM an efficient reader. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29005–29015, 2025.
- Morgan Pattin, Rémi Cottet, Véronique Eglin, and Alex Aussem. Rx-pad: Recognition and extraction for prescription analysis and clinical data structuring. In *Document Analysis and Recognition – ICDAR 2025*, volume 16027 of *Lecture Notes in Computer Science*, pp. 151–167. Springer, 2026. doi: 10.1007/978-3-032-04630-7_9.
- Jordy Van Landeghem, Rubén Tito, Łukasz Borchmann, Michał Pietruszka, Rafał Powalski, Dawid Jurkiewicz, Paweł Józsiak, Mickaël Coustaty, Bertrand Anckaert, Ernest Valveny, Matthew Blaschko, Sien Moens, and Tomasz Stanislawek. Document understanding dataset and evaluation (dude). In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19471–19483, 2023.
- Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. Publaynet: Largest dataset ever for document layout analysis. *arXiv preprint arXiv:1908.07836*, 2019.

APPENDIX

A LIMITATIONS AND FUTURE WORK

Retrieval and long-context reasoning Our evaluation focuses on contexts where full documents can be processed directly. We do not evaluate retrieval mechanisms, which are essential for practical long-document systems. In real-world applications, documents often span dozens or hundreds of pages (e.g., clinical study reports, financial statements, insurance claims), requiring retrieval to locate relevant passages before answering questions. Future work should assess retrieval-augmented pipelines, comparing vector-based text retrieval against vision-based page selection to understand where each approach succeeds or fails.

Multilingual and non-Latin script coverage While ICDAR_{DISCO} revealed the need for better multilingual support, our suite primarily covers English and French text in Latin scripts. Future benchmarks should systematically evaluate non-Latin scripts, mixed-script documents, and culturally specific layouts to guide deployment in global healthcare and regulatory contexts.

Metric limitations and answer format variability We deliberately used deterministic metrics ($S_{\text{GT-in-Pred}}$, S_{ANLS} , S_{EM}) rather than LLM-as-judge evaluation to ensure reproducible, non-stochastic assessment across all experiments. However, this approach requires considering multiple metrics together to see the full picture. Models frequently located correct information (high $S_{\text{GT-in-Pred}}$) but failed to format answers appropriately (low S_{ANLS} or S_{EM}), revealing a gap between answer localisation and output formatting. Future work could explore hybrid approaches that maintain reproducibility whilst better capturing semantic equivalence—for example, few-shot prompting with answer format examples or structured output validation using Pydantic schemas to constrain response formats without introducing evaluation non-determinism.

Model coverage and reproducibility All VLMs evaluated in this study (GPT-5-mini, GPT-5-nano, Claude 3.5 Sonnet) are closed-source API services, which limits reproducibility and broader community adoption. Future work should evaluate open-source VLMs such as Qwen-VL, LLaMA-based vision models, and DeepSeek-VL to determine whether the trends we observe (e.g., VLM advantages on multilingual text and OCR advantages on long documents) generalise beyond proprietary systems. Similarly, all experiments use zero-shot inference; evaluating instruction-tuned and fine-tuned variants could clarify whether task-specific training changes the OCR-vs-VLM trade-off. Finally, our evaluation is limited to static document images. Extending this analysis to temporally structured visual inputs (e.g., video-based document processing) and broader multilingual settings that include additional non-Latin scripts would strengthen the generalisability of our findings.

Verbosity and information completeness VLMs often produced longer outputs than ground-truth references, particularly during parsing. Qualitative inspection suggested these were not always hallucinations but rather more verbose extractions including contextual information. Topic-specific evaluation frameworks could assess what types of information are consistently missed or fabricated—for example, whether dosage errors in prescriptions are more common than patient name errors, or whether financial figures in regulatory documents are more prone to hallucination than procedural descriptions.

B FULL EXPERIMENTAL RESULTS

B.1 PARSING TASK: IN-DEPTH ANALYSIS

B.1.1 HANDWRITING RECOGNITION (IAM-DISCO)

IAM_{DISCO} contains 500 handwritten text samples with varying writing styles. VLMs with task-aware prompting can match or outperform dedicated OCR systems on character-level accuracy: `gpt-5-mini` achieves $S_{\text{CER}} = 0.080$, compared to $S_{\text{CER}} = 0.087$ for `mistral-ocr-2505`. The word-level gap is more pronounced—`gpt-5-mini` reaches $S_{\text{WER}} = 0.110$ versus 0.305 for OCR, nearly $3\times$ better. However, OCR systems maintain higher semantic fidelity, with `azure-ai-documentintelligence` achieving $S_{\text{CS}} = 0.946$ versus 0.914 for the best VLM. This suggests OCR errors are more localised (character substitutions preserving meaning), while VLM errors may involve rephrasing that shifts semantic representation.

Task-aware prompting yields inconsistent effects across model families. `gpt-5-mini` improves from $S_{\text{CS}} = 0.827$ (generic) to 0.914 (task-aware), a +10.5% gain, with S_{CER} dropping from 0.175 to 0.080. `claude-3-5-sonnet` exhibits the opposite pattern: performance degrades from $S_{\text{CS}} = 0.905$ to 0.845, and S_{CER} increases from 0.163 to 0.201. This divergence indicates that handwriting-specific instructions may conflict with certain models’ default transcription behaviour, and prompting strategies cannot be assumed to transfer across model families.

B.1.2 MULTILINGUAL SCENE TEXT (ICDAR-DISCO)

ICDAR_{DISCO} spans 10 languages (Arabic, Bangla, Chinese, Hindi, Japanese, Korean, Italian, French, German) with 50 samples each. OCR systems achieve a mean of S_{CER} of 5.53, while VLMs with generic prompting reach 2.13. Task-aware prompting further reduces VLM error rates to 0.73, substantially outperforming OCR on multilingual scene text.

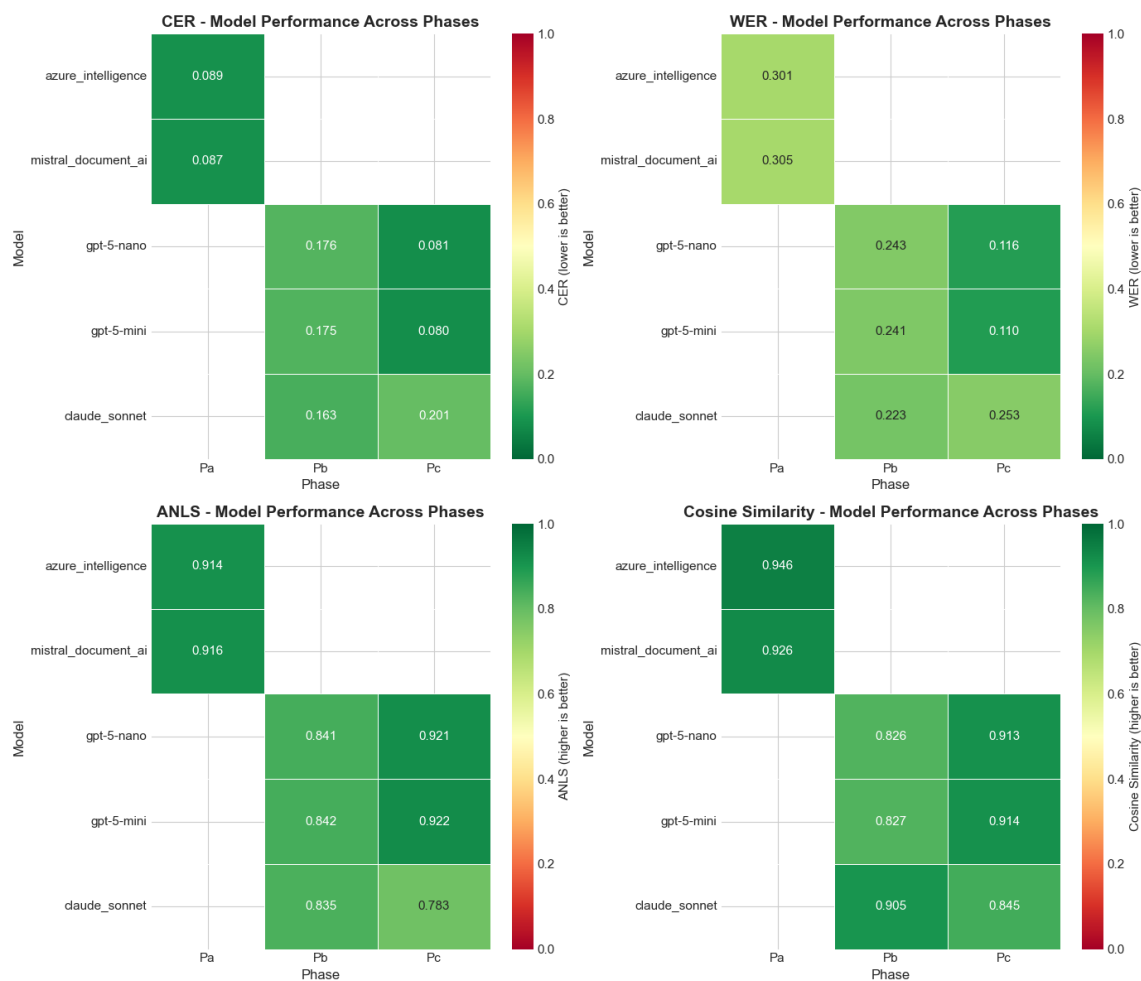
Results Table 3 presents aggregate performance across all phases. VLMs consistently outperform dedicated OCR services, with task-aware prompting yielding the lowest error rates. The best configuration (VLM + context) achieves a mean S_{CER} of 0.73 compared to 5.53 for OCR baselines—a reduction of approximately 87%. Beyond absolute performance, VLMs exhibit substantially lower variance (S_{CER} standard deviation of 0.40 vs 36.55), indicating more consistent behaviour across diverse scripts. Fig. 2 illustrates these patterns: OCR models cluster in the high-error region (dark red) while VLMs with task-aware prompting ($P_{\text{VLM-task}}$) achieve the lowest S_{CER} and S_{WER} .

Approach	CER ↓	WER ↓	ANLS ↑	S_{CS} ↑
OCR models	5.53	4.67	0.07	0.45
VLM (generic prompt)	2.13	2.02	0.18	0.56
VLM (task-aware prompt)	0.73	0.85	0.22	0.47

Table 3: Parsing average performance on ICDAR Mini. Results averaged across models within each phase.

Script-level analysis. Performance gaps widen for non-Latin scripts. Table 4 shows selected language categories where VLMs demonstrate the largest improvements. For Chinese text, VLMs reduce CER from 5.04 to 1.93. The contrast is most pronounced for mixed-script content: OCR achieves a CER of 155.84 on “Chinese, Mixed” samples, while VLMs maintain a CER below 1.0. Similar patterns emerge for Hindi (CER 3.16 → 1.17) and Bangla (CER 2.45 → 3.35, though both struggle here).

Discussion. These results suggest that VLMs offer a more robust solution for multilingual document parsing than specialised OCR services, particularly for non-Latin scripts and mixed-language content. The substantial benefit of task-aware prompting (66% CER reduction over generic prompts) highlights the importance of

Figure 1: Model performance across phases on IAM_{DISCO}.

prompt design in document intelligence applications. As shown in Fig. 2, the cosine similarity for task-aware VLMs drops slightly (0.47 vs 0.56 for generic prompts), which may indicate semantic drift when prompts become overly prescriptive—though ANLS, a stricter matching metric, continues to improve. The VLM advantage is largest for Arabic (+36.8%) and Bangla (+40.2%), suggesting OCR systems were primarily optimised for Latin scripts while VLMs benefit from multilingual pretraining corpora.

B.1.3 MEDICAL DOCUMENTS (RXPAD)

Results Table 5 summarises the performance of OCR and VLM approaches across all three experimental phases. Traditional OCR systems (*azure-ai-documentintelligence* and *mistral-ocr-2505*) achieved a mean S_{CER} of 0.708 and S_{WER} of 0.654 in Phase P_{OCR} , with Azure (CER: 0.654) substantially outperforming Mistral (CER: 0.762). VLMs tested with base prompting (Phase $P_{\text{VLM-base}}$) performed comparably, with S_{CER} values of 0.658 and S_{WER} of 0.594. When provided with a task-aware medical prompt

Script Category	OCR (CER) ↓	VLM (CER) ↓	n
Chinese	5.04	1.93	94
Chinese, Latin	1.59	0.59	85
Hindi	3.16	1.17	106
Japanese	1.75	1.61	64
Korean	0.89	0.35	67
Latin	1.69	2.07	24

Table 4: CER by script category for best-performing OCR (azure-ai-documentintelligence) vs best-performing VLM (gpt-5-mini). Sample counts in parentheses.

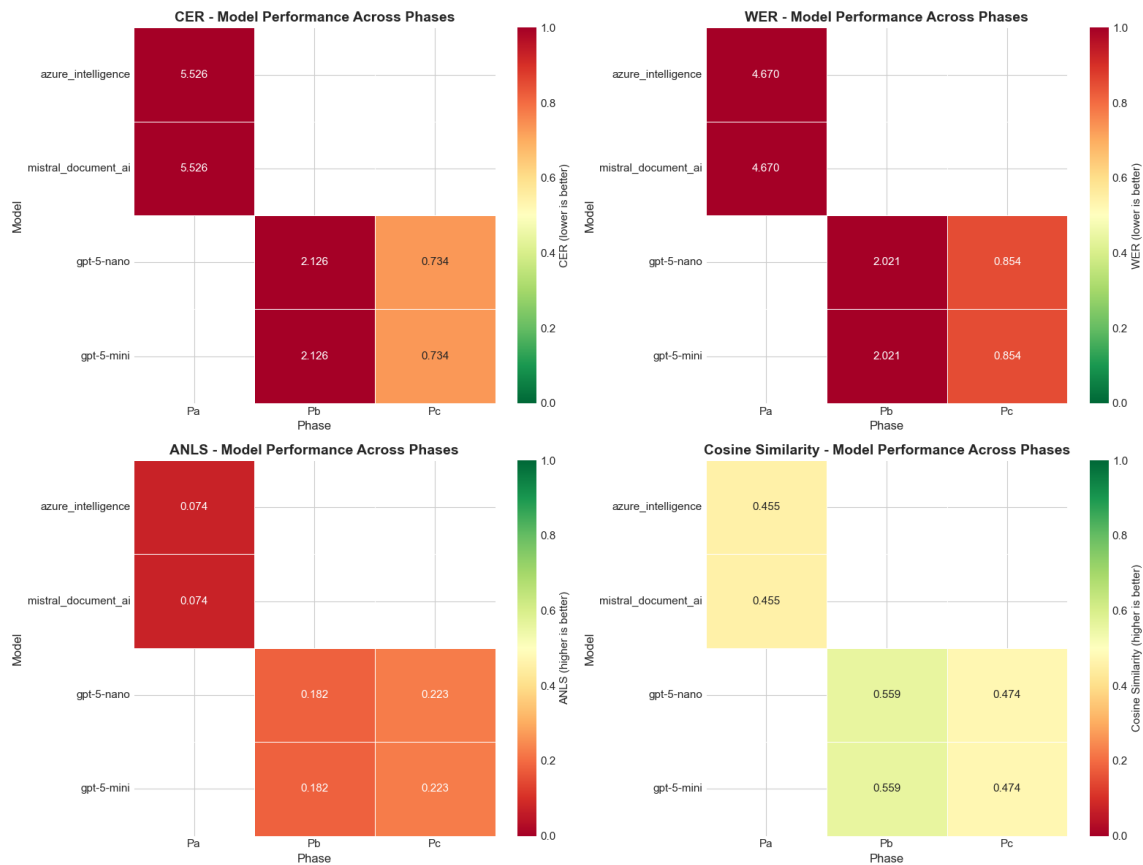


Figure 2: Model performance across phases on ICDAR_{DISCO}. P_{OCR} : OCR baseline; $P_{VLM-base}$: VLM with generic prompting; $P_{VLM-task}$: VLM with task-aware prompting. For CER and WER, lower (green) is better; for ANLS and Cosine Similarity, higher is better.

(Phase $P_{VLM-task}$), VLMs showed marginal improvement, reducing S_{WER} to 0.592 whilst maintaining similar S_{CER} (0.661). S_{CS} remained consistent across all approaches, ranging from 0.476 to 0.482. These patterns are clearly visible in Fig. 3, where the colour gradient reveals minimal variation between OCR and VLM

performance. S_{ANLS} was near zero across all experiments; this metric is designed for exact-match question answering scenarios and is not suitable for this task, where output formatting differences between predictions and ground truth dominate the error signal.

Phase	$S_{\text{CER}} \downarrow$	$S_{\text{WER}} \downarrow$	$S_{\text{CS}} \uparrow$
P_{OCR}	0.708	0.654	0.482
$P_{\text{VLM-base}}$	0.658	0.594	0.476
$P_{\text{VLM-task}}$	0.661	0.592	0.482

Table 5: Mean performance metrics across experimental phases on RxPad, averaged across all evaluated models.

Discussion The results indicate that neither OCR nor VLMs hold a clear advantage for raw text extraction on French medical prescriptions, though Azure Document Intelligence (CER: 0.654) outperforms Mistral OCR (CER: 0.762). As shown in Fig. 3, performance differences between approaches are marginal across all metrics. However, qualitative analysis of model outputs reveals an important distinction: VLMs consistently produce structured key-value representations (e.g., `product_name: DOLIPRANE`, `dose_unit: comprimé`), whilst OCR systems and ground truth annotations contain unstructured plain text. This format mismatch artificially inflates character and word error rates, as the models are penalised for reformatting rather than misunderstanding content.

Field-level extraction analysis (Table 6) supports this interpretation. VLMs achieved high recall on medical terminology such as medication names, dosage units, and prescription identifiers, suggesting that comprehension is not the limiting factor. The modest improvement observed when adding medical context ($P_{\text{VLM-task}}$ vs $P_{\text{VLM-base}}$) indicates that domain-specific prompting provides limited benefit for this dataset, likely because the visual and textual cues in prescription documents are already sufficient for general-purpose models to infer the clinical context.

Field (French)	GT Presence \uparrow	VLM Recall \uparrow
mg (Milligrams)	70.0%	100.0%
comprimé (Tablet)	46.5%	100.0%
traitement (Treatment)	31.5%	100.0%
médecin (Doctor)	32.0%	139.1%
ordonnance (Prescription)	13.5%	100.0%
fois par jour (Times/day)	25.0%	4.0%

Table 6: Field extraction recall for VLMs in Phase $P_{\text{VLM-task}}$. Values above 100% indicate the model detected more instances than present in ground truth annotations.

These findings suggest that for clinical document processing pipelines, the choice between OCR and VLMs should be guided by downstream task requirements rather than raw extraction accuracy. VLMs may be preferable when structured output is desired, whilst OCR remains suitable for applications requiring verbatim text reproduction.

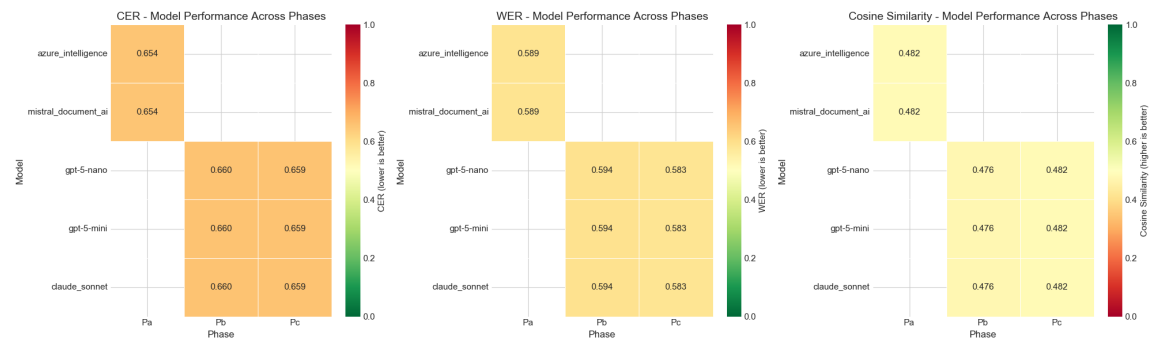


Figure 3: Model performance across phases on RxPad. P_{OCR} : OCR baseline; $P_{VLM-base}$: VLM with base prompting; $P_{VLM-task}$: VLM with task-aware prompting. For S_{CER} and S_{WER} , lower (green) is better; for S_{CS} , higher is better.

B.2 QA TASK: IN-DEPTH ANALYSIS

B.2.1 DOCUMENT QUESTIONS (DOCVQA-DISCO)

We evaluated three document QA strategies on the DocVQA benchmark (500 samples): OCR-based pipelines (QA_{OCR}), VLM parse-then-answer ($QA_{VLM-2stage}$), and direct visual question answering ($QA_{VLM-direct}$). Table 7 summarises the best-performing configuration within each strategy.

Strategy	Best Configuration	$S_{GT-in-Pred} \uparrow$	$S_{ANLS} \uparrow$	$S_{EM} \uparrow$	$S_{CS} \uparrow$
QA_{OCR}	azure-ai-documentintelligence \rightarrow gpt-5-mini	0.876	0.841	0.744	0.855
$QA_{VLM-2stage}$	gpt-5-mini \rightarrow gpt-5-mini	0.896	0.868	0.779	0.876
$QA_{VLM-direct}$	gpt-5-mini	0.908	0.632	0.482	0.786

Table 7: Best model performance by strategy on DocVQA. $S_{GT-in-Pred}$ is the primary metric.

Direct VQA achieved the highest $S_{GT-in-Pred}$ score (0.908), outperforming the best OCR-based pipeline by 3.2 percentage points. As shown in Fig. 5, this pattern holds consistently: all models evaluated under $QA_{VLM-direct}$ outperform their QA_{OCR} and $QA_{VLM-2stage}$ counterparts on $S_{GT-in-Pred}$. gpt-5-nano, for instance, achieves 0.877 with direct VQA compared to 0.425 when used in a parse-then-answer configuration.



Figure 4: Regression performance when predicting QA correctness from the parsed document data (DocVQA).

However, the relationship between strategies reverses when considering string-matching metrics. Table 8 quantifies the discrepancy between GT in Pred and ANLS across strategies.

Strategy	Mean $S_{GT\text{-in-Pred}} \uparrow$	Mean $S_{ANLS} \uparrow$
QA_{OCR}	0.840	0.720
$QA_{VLM\text{-}2\text{stage}}$	0.725	0.480
$QA_{VLM\text{-}direct}$	0.893	0.450

Table 8: Mean QA performance by strategy on DocVQA (reported separately for $S_{GT\text{-in-Pred}}$ and S_{ANLS}).

The gap between $S_{GT\text{-in-Pred}}$ and S_{ANLS} widens substantially for direct VQA, driven primarily by `claude-3-5-sonnet`’s behaviour. Fig. 5 reveals that `claude-3-5-sonnet` achieves 0.904 $S_{GT\text{-in-Pred}}$ but only 0.066 S_{ANLS} and 0.047 S_{EM} in direct VQA mode. This indicates that while the ground truth answer is present within `claude-3-5-sonnet`’s responses, the output format diverges significantly from the expected terse answers.

Within the OCR pipeline (QA_{OCR}), OCR provider selection has measurable impact. `azure-ai-documentintelligence` consistently outperforms `mistral-ocr-2505` by 3–4 percentage points on $S_{GT\text{-in-Pred}}$ when paired with the same downstream QA model (0.876 vs 0.833 in QA_{OCR}^{cot}).

Discussion The results suggest that for document QA, direct visual processing by VLMs outperforms explicit text extraction pipelines when measured by answer containment ($S_{GT\text{-in-Pred}}$). This finding challenges the conventional assumption that OCR-based approaches provide superior text grounding. The VLM’s ability to jointly reason over visual layout, typography, and textual content appears to confer an advantage over pipelines that discard spatial information during OCR.

Model / system	$S_{GT\text{-in-Extracted-Text}} \uparrow$
<code>azure-ai-documentintelligence</code>	0.8712
<code>mistral-ocr-2505</code>	0.8479
<code>mistral-ocr-2512</code>	0.6146
<code>gpt-5-nano</code>	0.4479
<code>gpt-5-mini</code>	0.8695
<code>claude-3.5-sonnet</code>	0.8677

Table 9: $S_{GT\text{-in-Extracted-Text}}$ from parsed document data on DocVQA.

The divergence between $S_{GT\text{-in-Pred}}$ and string-matching metrics warrants careful interpretation. High $S_{GT\text{-in-Pred}}$ with low S_{ANLS} indicates verbose but correct responses—the model identifies the right information but embeds it within explanatory text. This behaviour is particularly pronounced in `claude-3-5-sonnet` across all strategies, as evidenced in Fig. 5. Whether this constitutes a limitation depends on downstream requirements: extractive applications requiring structured outputs would penalise such responses, whereas information retrieval or human-facing systems may prefer them.

The poor performance of VLM parse-then-answer ($QA_{VLM\text{-}2\text{stage}}$) relative to both alternatives is notable. Despite using the same model for both stages, `gpt-5-nano` achieves only 0.425 $S_{GT\text{-in-Pred}}$ in $QA_{VLM\text{-}2\text{stage}}$ versus 0.877 in $QA_{VLM\text{-}direct}$. This suggests that the intermediate text representation introduces information loss or formatting artifacts that degrade downstream QA performance, without the compensating benefit of specialised OCR systems used in QA_{OCR} .



Figure 5: DocVQA strategy heatmaps using the primary strategy metric $S_{GT-in-Pred}$ for (1) QA_{OCR} , (2) $QA_{VLM-2stage}$, and (3) $QA_{VLM-direct}$.

OCR quality remains relevant within two-stage pipelines. The consistent 3–4 point advantage of azure-ai-documentintelligence over mistral-ocr-2505 indicates that OCR errors propagate to QA performance, supporting the intuition that text extraction fidelity bounds downstream accuracy in pipeline architectures. Also, Table 9 and Figure 4 both confirm that these results are mainly error propagation from parsing tasks and not only due to the QA stage.

Within the OCR pipeline strategy, provider choice significantly impacts performance on structured single-page documents. Table 10 presents phase-wise performance breakdown.

OCR system	$QA_{\text{OCR}}^{\text{generic}}$	$QA_{\text{OCR}}^{\text{cot}}$	$QA_{\text{OCR}}^{\text{task-aware}}$
$S_{\text{GT-in-Pred}}$ (primary metric) \uparrow			
azure-ai-documentintelligence	0.814	0.846	0.831
mistral-ocr-2505	0.785	0.805	0.794
mistral-ocr-2512	0.574	0.555	0.579
S_{ANLS} (string similarity) \uparrow			
azure-ai-documentintelligence	0.623	0.833	0.733
mistral-ocr-2505	0.607	0.787	0.692
mistral-ocr-2512	0.459	0.560	0.499
S_{EM} (exact match) \uparrow			
azure-ai-documentintelligence	0.454	0.720	0.559
mistral-ocr-2505	0.444	0.696	0.521
mistral-ocr-2512	0.331	0.507	0.374

Table 10: OCR system performance on DocVQA_{DISCO} by phase (all using gpt-5-mini for QA).

Chain-of-thought prompting ($QA_{\text{OCR}}^{\text{cot}}$) yielded the strongest performance across all OCR systems, with Azure Intelligence achieving 0.846 $S_{\text{GT-in-Pred}}$ and 0.720 exact match rate. Notably, the performance ordering remained consistent across prompting strategies: Azure > Mistral OCR 2 > Mistral OCR 3. The 26-point gap between Mistral OCR versions (0.805 vs 0.555 in $QA_{\text{OCR}}^{\text{cot}}$) exceeded the gap between Azure and Mistral OCR 2 (4.1 points), indicating that Mistral OCR 3 represents a substantial regression rather than incremental improvement.

B.2.2 INFOGRAPHICVQA

Results We evaluated three document QA strategies on the InfographicVQA_{DISCO} benchmark (500 infographic question-answer pairs): (i) OCR+VLM pipelines where a dedicated OCR system extracts text before an LLM answers the question (QA_{OCR}), (ii) VLM Parse+QA where the same vision-language model performs both parsing and answering ($QA_{\text{VLM-2stage}}$), and (iii) Direct VQA where the VLM receives the image and question without intermediate text extraction ($QA_{\text{VLM-direct}}$). Performance was measured using $S_{\text{GT-in-Pred}}$, S_{ANLS} , and S_{EM} .

Table 11 presents the best-performing configuration for each strategy using gpt-5-mini as the QA model. Direct VQA achieved the highest $S_{\text{GT-in-Pred}}$ (0.785), indicating that the correct answer was contained in the model’s response more frequently than with other approaches. However, it exhibited substantially lower S_{ANLS} (0.186) and S_{EM} (0.102), suggesting verbose or poorly formatted outputs rather than incorrect answers. The OCR+VLM pipeline using azure-ai-documentintelligence achieved the most balanced performance across all metrics, with the highest S_{ANLS} (0.629) and S_{EM} (0.515).

Strategy	$S_{\text{GT-in-Pred}} \uparrow$	$S_{\text{ANLS}} \uparrow$	$S_{\text{EM}} \uparrow$
QA_{OCR} : OCR→QA (azure-ai-documentintelligence)	0.754	0.629	0.515
$QA_{\text{VLM-2stage}}$: VLM→QA	0.711	0.585	0.477
$QA_{\text{VLM-direct}}$: Direct VQA	0.785	0.186	0.102

Table 11: Best-performing configuration per strategy (gpt-5-mini as QA model)

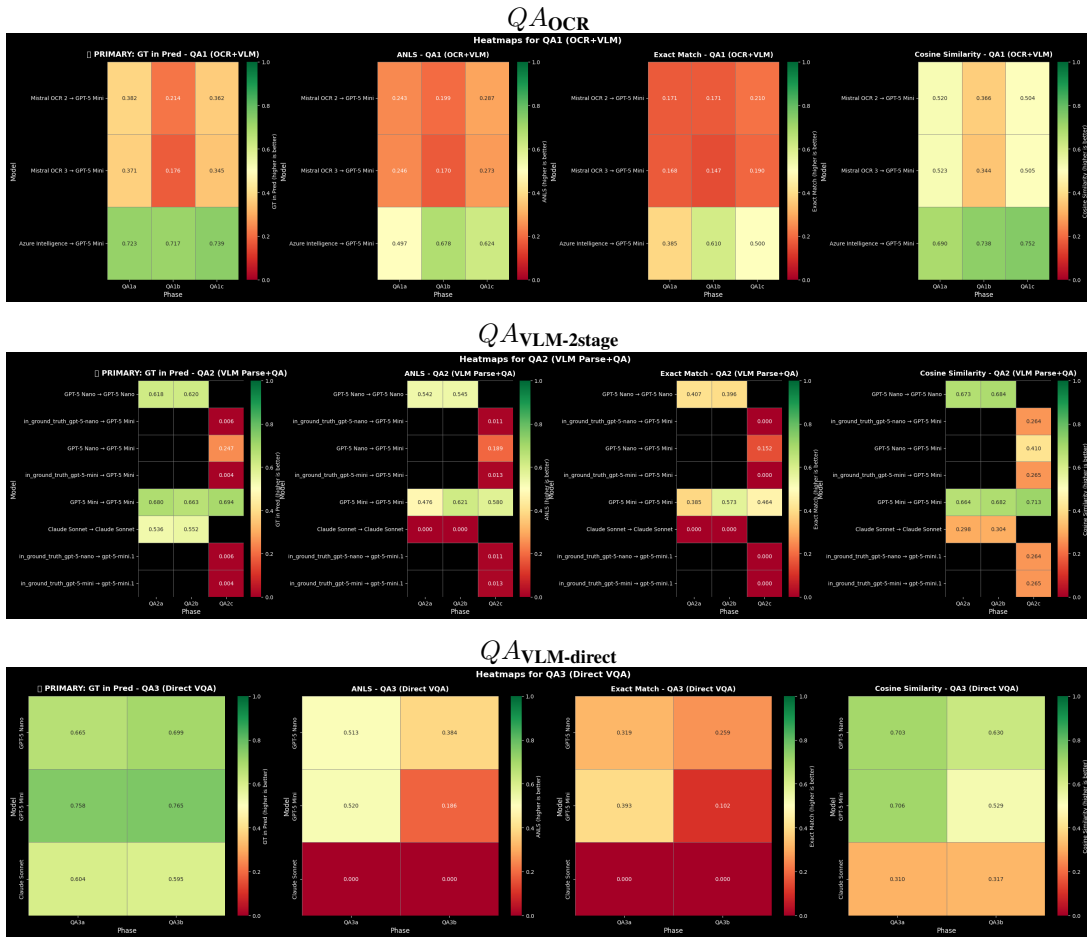


Figure 6: InfographicVQA strategy heatmaps. Each row corresponds to a QA strategy (Q_{AOCR} , $Q_{AVLM-2stage}$, $Q_{AVLM-direct}$) and reports four metrics: $S_{GT-in-Pred}$, S_{ANLS} , S_{EM} , and S_{CS} (Cosine Similarity).

The choice of OCR system proved critical for pipeline-based approaches. As shown in Table 12, azure-ai-documentintelligence outperformed mistral-ocr-2505 by a factor of two across all metrics, demonstrating that OCR quality represents a significant bottleneck in two-stage pipelines.

OCR System	$S_{GT-in-Pred} \uparrow$	$S_{ANLS} \uparrow$	$S_{EM} \uparrow$
azure-ai-documentintelligence	0.754	0.629	0.515
mistral-ocr-2505	0.368	0.288	0.214
mistral-ocr-2512	0.345	0.273	0.190

Table 12: Impact of OCR system on Q_{AOCR} pipeline performance (Q_{AOCR-c} phase, gpt-5-mini)

Discussion The results reveal a key distinction between answer correctness and answer format compliance. Direct VQA achieved the highest $S_{GT-in-Pred}$, meaning VLMs can accurately locate and reason about the

relevant information in infographics when given direct visual access. The discrepancy with S_{ANLS} and S_{EM} stems from response verbosity rather than factual errors—the model produces contextualised answers instead of terse extractions. This behaviour is addressable through prompt engineering: constraining the expected output format (e.g., instructing the model to respond with only the answer value) would likely align S_{ANLS} and S_{EM} with the $S_{\text{GT-in-Pred}}$ performance.

OCR-based pipelines remain competitive when high-quality OCR is available, but their ceiling is fundamentally limited by text extraction fidelity. Infographics present particular challenges for OCR due to non-standard layouts, embedded text in visual elements, and the need to preserve spatial relationships between data points. Direct VQA circumvents these issues entirely by reasoning over the visual representation.

Given that (i) direct VQA demonstrates superior answer correctness as measured by $S_{\text{GT-in-Pred}}$, (ii) the format compliance gap is a prompt-level rather than capability-level limitation, and (iii) end-to-end approaches avoid error propagation from OCR failures, we conclude that direct VQA with VLMs represents the most promising approach for infographic question answering. We hypothesise that the observed verbosity is more readily addressable through prompt engineering or constrained decoding than OCR errors on complex visual documents, though we have not tested format-enforcement techniques (e.g., Pydantic output schemas, few-shot format examples) in this study, and leave this comparison to future work.

B.2.3 MULTI-PAGE DOCUMENTS (DUDE)

Model / parsing method	$S_{\text{GT-in-Parsed-Data}} \uparrow$
azure-ai-documentintelligence	0.746
mistral-ocr-2505	0.357
mistral-ocr-2512	0.321
gpt-5-nano	0.625
gpt-5-mini	0.694
claude-3-5-sonnet	0.645

Table 13: Ground-truth coverage in parsed data across OCR-based and VLM-based parsing methods.

Results Three strategies were compared:

- QA_{OCR} (**OCR+VLM**): Dedicated OCR extraction (azure-ai-documentintelligence or mistral-ocr-2505) followed by gpt-5-mini for question answering over the extracted text.
- $QA_{\text{VLM-2stage}}$ (**VLM Parse+QA**): gpt-5-mini performs both document parsing and subsequent question answering in a two-stage pipeline.
- $QA_{\text{VLM-direct}}$ (**Direct VQA**): gpt-5-mini receives document images directly and answers questions without explicit text extraction.

The primary evaluation metric is $S_{\text{GT-in-Pred}}$ (ground truth substring presence in prediction), supplemented by S_{ANLS} (Average Normalised Levenshtein Similarity), S_{EM} , and substring match rates.

Results Table 14 presents the aggregated performance across strategies. The hybrid OCR+VLM approach (QA_{OCR}) achieved the highest $S_{\text{GT-in-Pred}}$ of 0.514, outperforming both direct VQA (0.493) and the VLM-based parsing pipeline (0.371). The performance gap between QA_{OCR} and $QA_{\text{VLM-2stage}}$ is substantial (14.2 percentage points), indicating that VLM-based text extraction introduces errors that propagate to the QA stage.

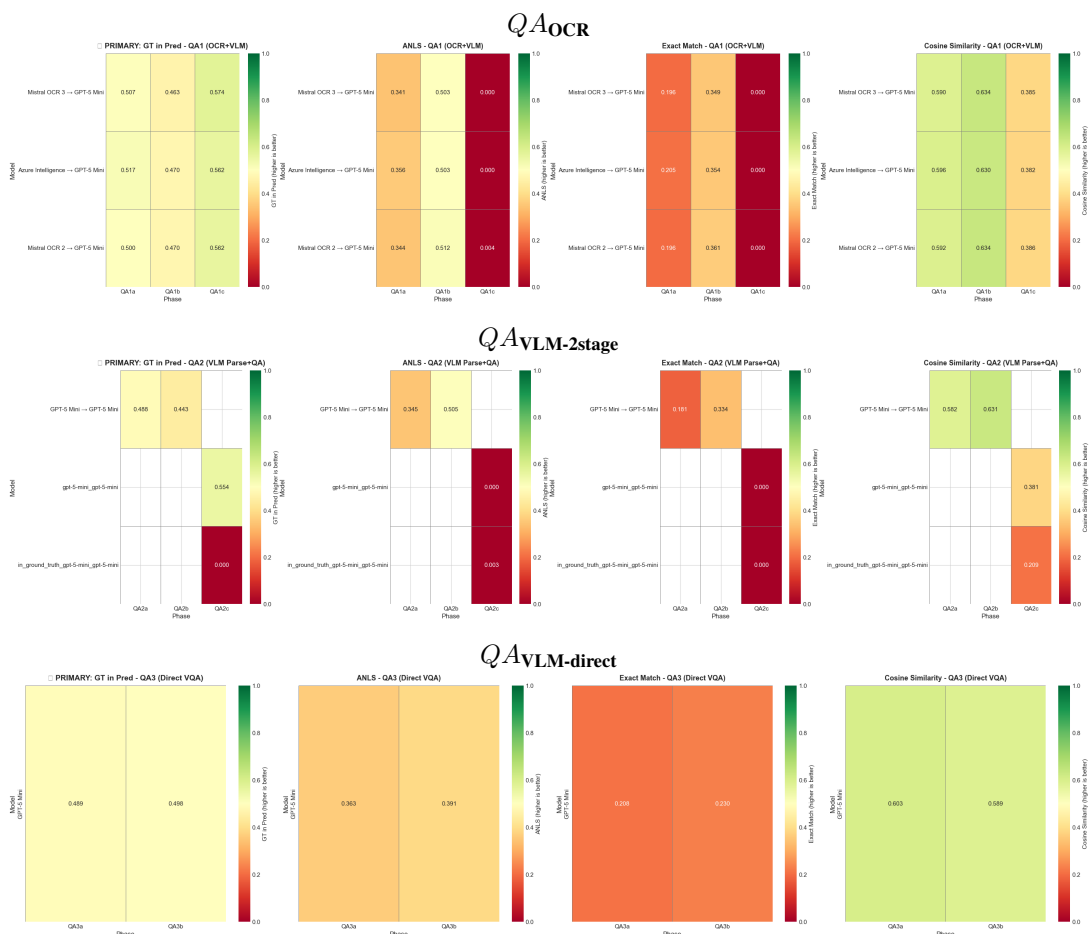


Figure 7: DUDE strategy heatmaps. Each row corresponds to a QA strategy (Q_{AOCR} , $Q_{AVLM-2stage}$, $Q_{AVLM-direct}$) and reports four metrics: $S_{GT-in-Pred}$, S_{ANLS} , S_{EM} , and S_{CS} (Cosine Similarity).

Direct VQA ($Q_{AVLM-direct}$) achieved the highest S_{ANLS} (0.377) and S_{EM} (21.9%), suggesting that when answers are correct, they tend to be more precisely formatted. However, its lower $S_{GT-in-Pred}$ indicates more frequent complete misses compared to the OCR-based approach.

Strategy	$S_{GT-in-Pred} \uparrow$	$S_{ANLS} \uparrow$	$S_{EM} \uparrow$	Substring Match \uparrow
Q_{AOCR} (OCR+VLM)	0.514	0.286	0.186	0.538
$Q_{AVLM-2stage}$ (VLM Parse+QA)	0.371	0.213	0.129	0.391
$Q_{AVLM-direct}$ (Direct VQA)	0.493	0.377	0.219	0.524

Table 14: Strategy-level performance comparison on DUDE (n=404 per phase)

Table 15 details the OCR tool comparison within the Q_{AOCR} strategy. azure-ai-documentintelligence and mistral-ocr-2505 performed comparably, with

differences of less than 2 percentage points on $S_{\text{GT-in-Pred}}$ across phases. Both OCR systems showed identical performance on the $QA_{\text{OCR}}^{\text{cot}}$ and $QA_{\text{OCR}}^{\text{task}}$ phases ($S_{\text{GT-in-Pred}} = 0.470$ and 0.562 respectively), suggesting that downstream QA model behaviour dominates over OCR tool selection for these document types.

Phase	OCR Tool	$S_{\text{GT-in-Pred}} \uparrow$	$S_{\text{ANLS}} \uparrow$	$S_{\text{EM}} \uparrow$	$S_{\text{CS}} \uparrow$
$QA_{\text{OCR}}^{\text{generic}}$	azure-ai-documentintelligence	0.517	0.356	0.205	0.596
$QA_{\text{OCR}}^{\text{generic}}$	mistral-ocr-2505	0.500	0.344	0.196	0.592
$QA_{\text{OCR}}^{\text{generic}}$	mistral-ocr-2512	0.512	0.341	0.196	0.591
$QA_{\text{OCR}}^{\text{cot}}$	azure-ai-documentintelligence	0.470	0.503	0.354	0.630
$QA_{\text{OCR}}^{\text{cot}}$	mistral-ocr-2505	0.470	0.512	0.361	0.634
$QA_{\text{OCR}}^{\text{cot}}$	mistral-ocr-2512	0.468	0.504	0.351	0.635
$QA_{\text{OCR}}^{\text{task}}$	azure-ai-documentintelligence	0.562	0.000	0.000	0.382
$QA_{\text{OCR}}^{\text{task}}$	mistral-ocr-2505	0.562	0.004	0.000	0.386
$QA_{\text{OCR}}^{\text{task}}$	mistral-ocr-2512	0.579	0.000	0.000	0.386

Table 15: OCR tool comparison within QA_{OCR} (OCR \rightarrow gpt-5-mini)

Discussion The results indicate that dedicated OCR remains advantageous over VLM-based parsing for document QA on complex multi-page documents. The QA_{OCR} strategy’s 14-point lead over $QA_{\text{VLM-2stage}}$ demonstrates that specialised OCR tools extract text more reliably than VLMs prompted to parse document content. This finding aligns with the architectural differences: OCR systems are explicitly trained for text localisation and recognition, whereas VLMs must jointly attend to visual layout and textual content.

The relatively strong performance of direct VQA ($QA_{\text{VLM-direct}}$) is noteworthy. By bypassing explicit text extraction, this approach avoids cascading OCR errors whilst retaining visual context. Its higher S_{ANLS} and S_{EM} suggest that when the model correctly identifies the answer location, it reproduces the text more faithfully than pipeline approaches. However, the 2-point $S_{\text{GT-in-Pred}}$ deficit relative to QA_{OCR} indicates that direct VQA more frequently fails to locate relevant information entirely.

The $QA_{\text{VLM-2stage}}$ strategy’s poor performance ($S_{\text{GT-in-Pred}} = 0.371$) demonstrates the cost of error compounding in two-stage VLM pipelines. When the same model performs both parsing and QA, extraction errors in the first stage directly degrade QA accuracy. This suggests that if VLM-based parsing is required, using different models or architectures for each stage may mitigate error propagation.

The QA_1^c phase exhibits anomalous behaviour: high $S_{\text{GT-in-Pred}}$ (0.562) but near-zero S_{ANLS} and S_{EM} . This pattern indicates that predictions contain the ground truth as a substring but include substantial extraneous content, likely reflecting verbosity in the QA model’s response format for this phase configuration (see Section B.2.5 for quantitative analysis). Additionally, the observed 0.562 GT-in-Pred exceeds the 0.505 GT-in-Extracted-Text baseline for Azure OCR; Section B.2.3 demonstrates this is primarily due to metric sensitivity to punctuation (52.8% of cases) and legitimate visual reasoning (39.2% of cases), rather than hallucination or methodological error.

The minimal performance difference between `azure-ai-documentintelligence` and `mistral-ocr-2505` (Table 15) suggests that for DUDE-style documents, the choice of OCR backend is less consequential than the overall pipeline architecture. Both commercial OCR systems achieve comparable text extraction quality on this benchmark.

Parsing method	Ground truth in parsed data
Azure Intelligence OCR	0.505
Mistral OCR 3	0.391
Mistral OCR 2	0.468
GPT-5 Mini	0.465

Table 16: Ground truth coverage in parsed data for different parsing methods.

Parsing effectiveness determines the QA ceiling on multi-page documents. The relationship between OCR quality and QA performance is particularly clear on DUDE_{DISCO}, where ground-truth coverage in the parsed text predicts downstream accuracy. `azure-ai-documentintelligence` captured 50.5% of answers in the extracted text, whereas `mistral-ocr-2512` captured only 39.1%—an 11.4 percentage point gap that translated into similar differences in final QA performance.

OCR System	Parsing		QA Performance	
	GT-in-Ext	Rank	QA _{OCR} ^{cot}	QA _{OCR} ^{task}
Azure Intelligence	0.505	1	0.475 (2)	0.567 (2)
Mistral OCR 2	0.468	2	0.475 (2)	0.567 (2)
Mistral OCR 3	0.391	3	0.468 (3)	0.579 (1)

Table 17: Relationship between parsing effectiveness (GT-in-Extracted-Text) and QA performance on DUDE_{DISCO}.

How can QA performance exceed parsing coverage? We identify two contributing factors. First, $S_{GT-in-Pred}$ uses case-insensitive substring matching against potentially verbose model responses, while $S_{GT-in-Ext}$ uses exact substring matching against the parsed text alone — the QA model’s verbose outputs create a wider “matching surface” that can capture partial ground-truth strings missed in raw OCR output. Second, the QA model may draw on parametric knowledge to infer answers not explicitly present in the extracted text (e.g., recognising that a partially extracted entity name refers to a known organisation). This means $S_{GT-in-Ext}$ is not a strict upper bound on $S_{GT-in-Pred}$, which is an important caveat when interpreting pipeline results.

Interestingly, QA performance rankings did not perfectly mirror parsing effectiveness. On QA_{OCR}^{task}, Mistral OCR 3 achieved the highest GT-in-Pred (0.579) despite weakest parsing (0.391), suggesting that downstream prompt engineering can partially compensate for poor text extraction—though at the cost of format compliance (ANLS = 0.000).

DUDE Anomaly Investigation: When QA Exceeds Parsing Observation. On DUDE_{DISCO}, Azure OCR shows GT-in-Extracted-Text of 0.505 (50.5% of answers appear in parsed text) but GT-in-Pred of 0.562 (56.2% of answers appear in final predictions) for the task-aware phase. This apparent paradox—where the QA model produces correct answers despite ground truth being absent from the parsed text—warrants explanation.

Investigation. We analyzed all 125 anomalous cases (10.3% of 1,212 samples across QA1a/b/c phases) where GT-in-Pred=1 but GT-in-Extracted-Text=0. Automated categorization with manual verification revealed:

String matching artifacts (52.8%). The substring matching metric fails on:

- **Punctuation variations:** GT="\$10,000" vs Extracted="\$ 10,000" (space after \$)

Category	Count	%
String matching artifacts (punctuation, spacing)	66	52.8%
Visual reasoning (spatial, counting, layout)	49	39.2%
Inference from context (related text)	9	7.2%
Prior knowledge (from training data)	1	0.8%

Table 18: Breakdown of 125 DUDE anomalous cases where QA performance exceeded parsing metrics.

- **Trailing periods:** GT=“Two cents a pound.” vs Extracted=“TWO CENTS A POUND”
- **Format differences:** GT=“Privacy issues , Law policies , Algorithmic Factors” with list spacing
- **Short answers:** GT=“Yes” appears in model’s explanation but not in extracted text

Example case:

- Question: “*What is the price of potatoes?*”
- Ground Truth: “Two cents a pound.” (with period)
- Extracted: “AT TWO CENTS A POUND potatoes are a cheap food” (no period)
- Prediction: “Two cents a pound.”
- Substring match fails: “two cents a pound.” $\not\subseteq$ “at two cents a pound”

This represents a **metric limitation**, not a model issue—the information was successfully extracted.

Visual reasoning (39.2%). These cases demonstrate legitimate VLM capability:

- **Spatial questions:** “Is page 4 to the right or left?” (GT: “Right”)—requires visual layout understanding
- **Visual counting:** “How many bullets in the second paragraph?” (GT: “3”)—OCR converts to “.” symbols without structure
- **Temporal analysis:** “How many years does this document cover?” (GT: “13 years”)—model visually identified date range, calculated 2003-1990

These cases validate the VLM approach—GPT-5-mini genuinely extracts visual information beyond OCR text.

Inference from context (7.2%). Model reconstructed answers from partial information:

- Question: “*Which watershed flows west?*”
- GT: “(D) Jocelyn Hollow” (with multiple-choice letter)
- Extracted: Contains “Jocelyn Hollow” but not “(D) Jocelyn Hollow”
- Model combined visual map analysis with text labels

Prior knowledge (0.8%). Only 1 case showed possible training data contamination:

- Question: “*What does MAXI stand for?*”
- GT: “Monitor of All-sky X-ray Image” (astronomical instrument)
- Extracted text discusses ISS but doesn’t mention MAXI

Implications. The anomaly does not indicate a methodological flaw:

1. 52.8% are metric artifacts (overly strict substring matching)—normalizing punctuation would eliminate this discrepancy
2. 39.2% demonstrate genuine visual reasoning capabilities that strengthen the VLM argument
3. Prior knowledge contamination is negligible (0.8%)
4. The finding validates that VLMs access information beyond pure text extraction

For reproducibility, analysis scripts and categorized cases are available in the supplementary materials.

B.2.4 OCR SYSTEM COMPARISON ACROSS DATASETS

On multi-page documents, downstream QA performance is bounded by parsing/coverage: missing text cannot be recovered by the QA model, so improvements in OCR/text extraction reliability translate into measurable QA gains.

Within OCR pipelines, provider choice has dataset-dependent effects. On DUDE_{DISCO}, azure-ai-documentintelligence and mistral-ocr-2505 performed identically under chain-of-thought prompting (QA_{OCR}^{cot} ; $S_{GT-in-Pred} = 0.470$ for both), suggesting comparable text extraction quality. However, on DocVQA_{DISCO}, azure-ai-documentintelligence outperformed mistral-ocr-2505 by 3.3 percentage points (0.876 vs 0.843), indicating that single-page form documents benefit from Azure’s stronger layout analysis. mistral-ocr-2512 (a newer model) underperformed consistently across datasets, suggesting that version number alone does not guarantee improved document understanding.

Mistral OCR 3 underperforms its predecessor across all benchmarks. Table 19 shows that despite being the newer model version (mistral-ocr-2512 vs mistral-ocr-2505), Mistral OCR 3 consistently achieves lower parsing effectiveness than Mistral OCR 2. On DocVQA_{DISCO}, Mistral OCR 3 captured only 61.5% of ground-truth answers in extracted text compared to 84.8% for Mistral OCR 2—a 23.3 percentage point degradation. This pattern holds across all three QA datasets, challenging the assumption that newer model versions automatically deliver improved performance.

OCR System	DocVQA	InfographicVQA	DUDE
Azure Intelligence	0.871	0.754	0.505
Mistral OCR 2 (2505)	0.848	0.357	0.468
Mistral OCR 3 (2512)	0.615	0.321	0.391
<i>GPT-5 Mini (VLM parsing for reference)</i>			
GPT-5 Mini	0.869	0.465	0.694

Table 19: Ground truth coverage in parsed text by OCR system (GT-in-Extracted-Text metric). Higher values indicate more reliable text extraction.

The performance gap translates directly to downstream QA accuracy. On DocVQA_{DISCO}, the 23-point parsing disadvantage of mistral-ocr-2512 (relative to mistral-ocr-2505) resulted in a 26-point drop in QA performance in the QA_{OCR}^{cot} (0.555 vs 0.805).

Azure Intelligence maintains a consistent advantage on structured documents. Azure OCR outperformed both Mistral systems on DocVQA_{DISCO} (forms, letters) and InfographicVQA_{DISCO} (visual layouts), likely due to superior layout analysis capabilities. However, on DUDE_{DISCO} (multi-page documents), the gap narrowed substantially, with Mistral OCR 2 achieving comparable performance (0.468 vs 0.505). This

suggests that layout understanding matters more for single-page structured documents than for multi-page, text-heavy content.

B.2.5 QUANTIFYING VERBOSITY EFFECTS ON GT-IN-PRED RELIABILITY

To address concerns about GT-in-Pred metric reliability and potential verbosity bias, we quantified response length characteristics across all QA strategies and prompt types.

Task-aware prompts induce 2–3× higher verbosity. Table 20 shows that task-aware prompts (prompt type ‘c’) generate responses averaging $7.3\times$ longer than ground truth answers for QA_{OCR} pipelines, compared to $3.7\times$ for generic prompts. This verbosity differential is consistent across datasets ($DocVQA_{DISCO}$, $InfographicVQA_{DISCO}$, $DUDE_{DISCO}$) and strategies.

Table 20: Response verbosity and GT-in-Pred scores by strategy and prompt type.

Dataset	Strategy	Prompt Type	Avg Pred Len	Avg GT Len	Ratio	GT-in-Pred
$DocVQA_{DISCO}$	QA_OCR	generic	42.6	13.3	6.2	0.7
$DocVQA_{DISCO}$	QA_OCR	cot	15.6	13.3	1.4	0.8
$DocVQA_{DISCO}$	QA_OCR	task-aware	38.0	13.3	5.9	0.7
$DocVQA_{DISCO}$	QA_VLM-2stage	generic	70.9	13.3	11.1	0.7
$DocVQA_{DISCO}$	QA_VLM-2stage	cot	67.6	13.3	10.2	0.7
$DocVQA_{DISCO}$	QA_VLM-direct	generic	86.8	13.3	12.6	0.9
$DocVQA_{DISCO}$	QA_VLM-direct	cot	47.7	13.3	6.9	0.9
$InfographicVQA_{DISCO}$	QA_OCR	generic	51.4	8.2	12.6	0.4
$InfographicVQA_{DISCO}$	QA_OCR	cot	9.4	8.2	2.1	0.3
$InfographicVQA_{DISCO}$	QA_OCR	task-aware	55.1	8.2	15.0	0.4
$dude_{DISCO}$	QA_OCR	generic	64.4	18.6	12.4	0.5
$dude_{DISCO}$	QA_OCR	cot	17.3	18.6	1.8	0.5
$dude_{DISCO}$	QA_OCR	task-aware	366.4	18.6	88.0	0.6
$dude_{DISCO}$	QA_VLM-2stage	generic	69.6	18.6	13.7	0.5
$dude_{DISCO}$	QA_VLM-2stage	cot	17.0	18.6	1.8	0.5
$dude_{DISCO}$	QA_VLM-direct	generic	51.9	18.6	8.0	0.5
$dude_{DISCO}$	QA_VLM-direct	cot	29.3	18.6	5.0	0.5

Prediction length correlates moderately with GT-in-Pred scores. Table 21 presents Pearson and Spearman correlations between character-level prediction length and GT-in-Pred scores. Table 21 reports Pearson and Spearman correlations between character-level prediction length and GT-in-Pred scores. Overall, the correlation is weakly negative (Spearman $\rho = -0.035$, 95% CI $[-0.050, -0.021]$; Pearson $r = -0.111$), indicating that longer predictions are not systematically associated with higher GT-in-Pred scores. Strategy-specific patterns are similar. QA_{OCR} shows $\rho = -0.051$, $QA_{VLM-2stage}$ shows $\rho = -0.100$, and $QA_{VLM-direct}$ shows $\rho = 0.016$ (not statistically significant). These weak or negligible correlations suggest that verbosity alone does not explain GT-in-Pred performance, partially addressing concerns that the metric could be inflated by longer responses.

Anomalous cases concentrate in high-verbosity configurations. Figure 8 visualizes the relationship between prediction length and GT-in-Pred score, stratified by strategy and colored by prompt type. Anomalous cases ($GT\text{-in-Pred} \geq 0.5$, $ANLS < 0.1$) predominantly occur with task-aware prompts, exhibiting verbosity ratios exceeding 7:1. These cases explain the QA1c anomaly pattern: models generate verbose explanations containing the correct answer substring but fail to match the expected terse response format, resulting in high substring containment but zero edit-distance similarity.

Strategy	N	Pearson r [95% CI]	Spearman ρ [95% CI]
Overall	19570	-0.111 [-0.126, -0.097]	-0.035 [-0.050, -0.021]
QA_OCR	12510	-0.124 [-0.141, -0.107]	-0.051 [-0.069, -0.033]
QA_VLM-2stage	3780	-0.100 [-0.132, -0.066]	-0.100 [-0.131, -0.068]
QA_VLM-direct	3280	0.004 [-0.031, 0.039]	0.016 [-0.017, 0.051]

Table 21: Correlation between prediction length and GT-in-Pred score.

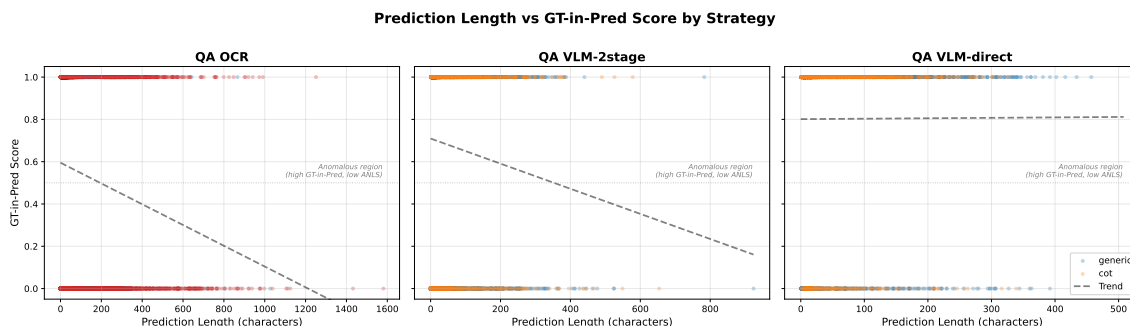


Figure 8: Prediction length vs GT-in-Pred score by strategy and prompt type. Positive correlations indicate longer predictions tend to contain ground truth substrings. Anomalous region ($GT\text{-in-Pred} \geq 0.5$, low ANLS) shows task-aware prompts generate verbose responses with correct content but poor format compliance.

Interpretation for metric design. The weak negative correlations ($\rho \approx -0.04$ to -0.10) indicate that verbosity bias does not inflate GT-in-Pred scores at the sample level—models must still locate and include relevant information rather than simply generate longer responses. The anomalous high-verbosity configurations (Section B.2.3) represent a separate effect driven by prompt design rather than a general limitation of the metric.

B.3 ALL RAW RESULTS

B.4 SUMMARY

We summarise (i) **parsing winners** using S_{CS} and (ii) **QA winners** using $S_{GT\text{-in-Pred}}$ (the primary metrics used throughout the paper). A *winner* is the approach that achieved the **best observed performance** on the corresponding dataset/metric. Entries are: **1** = wins, **0** = loses, **1-1** = comparable.

Table 25 provides a compact winner summary by dataset and evaluation criterion.

Conclusion. Mistral OCR 3’s consistent underperformance relative to Mistral OCR 2 across all datasets and metrics (parsing effectiveness, QA accuracy, and no latency advantage) indicates a clear regression. Practitioners deploying Mistral OCR should prefer version 2505 over 2512 until the performance issues in the newer version are resolved. More broadly, these findings underscore the importance of empirical validation: model version increments do not guarantee improved performance, and deployment decisions should be driven by benchmark results on representative data rather than version numbers or release dates.

Conclusion. Within the scope of our evaluation, direct VLM calls achieve the strongest performance more often overall (particularly on QA and visually structured documents), while OCR-based pipelines remain the most reliable choice on long or multi-page documents.

Dataset	P_{OCR} (OCR baseline)		$P_{\text{VLM-base}}$ (generic VLM)		$P_{\text{VLM-task}}$ (task-aware VLM)	
	CER ↓	WER ↓	CER ↓	WER ↓	CER ↓	WER ↓
IAM (handwriting)	0.099 ± 0.013	0.283 ± 0.014	0.175 ± 0.014	0.235 ± 0.014	0.151 ± 0.014	0.191 ± 0.016
ICDAR (scene text)	3.193 ± 1.082	3.208 ± 1.311	2.044 ± 0.530	1.969 ± 0.379	0.512 ± 0.029	0.673 ± 0.036
RxPad (prescriptions)	0.708 ± 0.082	0.654 ± 0.078	0.658 ± 0.017	0.594 ± 0.024	0.660 ± 0.016	0.592 ± 0.023

Table 22: Parsing metrics with 95% bootstrap confidence intervals (margin: \pm half-width of CI). Values shown as point estimate \pm margin computed from 10,000 sample-level bootstrap iterations. Lower CER and WER indicate better performance.

Dataset	Phase	System	$S_{\text{GT-in-Pred}} \uparrow$	$S_{\text{ANLS}} \uparrow$	$S_{\text{EM}} \uparrow$
DocVQA	$Q A_{\text{OCR}}^{\text{generic}}$	azure_intelligence__gpt-5-mini	0.8405	0.6284	0.4622
DocVQA	$Q A_{\text{OCR}}^{\text{got}}$	azure_intelligence__gpt-5-mini	0.8763	0.8406	0.7444
DocVQA	$Q A_{\text{OCR}}^{\text{task}}$	azure_intelligence__gpt-5-mini	0.8571	0.7412	0.5714
DocVQA	$Q A_{\text{VLM-2stage}}^{\text{generic}}$	gpt-5-mini__gpt-5-mini	0.8514	0.6159	0.4438
DocVQA	$Q A_{\text{VLM-2stage}}^{\text{got}}$	gpt-5-mini__gpt-5-mini	0.8956	0.8683	0.7791
DocVQA	$Q A_{\text{VLM-direct}}^{\text{fewshot}}$	gpt-5-nano	0.8740	0.7455	0.4360
DocVQA	$Q A_{\text{VLM-direct}}^{\text{generic}}$	gpt-5-nano	0.8798	0.7414	0.5471
DUDE	$Q A_{\text{OCR}}^{\text{generic}}$	azure_intelligence__gpt-5-mini	0.5173	0.3555	0.2054
DUDE	$Q A_{\text{OCR}}^{\text{got}}$	mistral_document_ai__gpt-5-mini	0.4703	0.5120	0.3614
DUDE	$Q A_{\text{OCR}}^{\text{task}}$	mistral_document_ai__gpt-5-mini	0.5619	0.0035	0.0000
DUDE	$Q A_{\text{VLM-2stage}}^{\text{generic}}$	gpt-5-mini__gpt-5-mini	0.4876	0.3451	0.1807
DUDE	$Q A_{\text{VLM-2stage}}^{\text{got}}$	gpt-5-mini__gpt-5-mini	0.4431	0.5049	0.3342
DUDE	$Q A_{\text{VLM-direct}}^{\text{fewshot}}$	gpt-5-mini	0.4888	0.3631	0.2084
DUDE	$Q A_{\text{VLM-direct}}^{\text{generic}}$	gpt-5-mini	0.4975	0.3906	0.2302
InfographicVQA	$Q A_{\text{OCR}}^{\text{generic}}$	azure_intelligence__gpt-5-mini	0.7347	0.5018	0.3967
InfographicVQA	$Q A_{\text{OCR}}^{\text{got}}$	azure_intelligence__gpt-5-mini	0.7302	0.6808	0.6168
InfographicVQA	$Q A_{\text{OCR}}^{\text{task}}$	azure_intelligence__gpt-5-mini	0.7537	0.6290	0.5149
InfographicVQA	$Q A_{\text{VLM-2stage}}^{\text{generic}}$	gpt-5-nano__gpt-5-nano	0.6366	0.5471	0.4213
InfographicVQA	$Q A_{\text{VLM-2stage}}^{\text{got}}$	gpt-5-mini__gpt-5-mini	0.6796	0.6257	0.5837
InfographicVQA	$Q A_{\text{VLM-2stage}}^{\text{task}}$	gpt-5-mini__gpt-5-mini	0.7110	0.5846	0.4768
InfographicVQA	$Q A_{\text{VLM-direct}}^{\text{fewshot}}$	gpt-5-mini	0.7776	0.5252	0.4088
InfographicVQA	$Q A_{\text{VLM-direct}}^{\text{generic}}$	gpt-5-nano	0.7154	0.3839	0.2585

Table 23: QA metrics ($S_{\text{GT-in-Pred}}$, S_{ANLS} , S_{EM}) across datasets and phases, extracted from the notebook summary outputs. For each dataset and phase, the row reflects the best run among those marked [PRIMARY] in the notebook outputs (selected by S_{ANLS} , then S_{EM}).

Parsing			Question answering		
Dataset	OCR	VLM	Dataset	OCR	VLM
IAM _{DISCO}	1	0	DocVQA _{DISCO}	0	1
ICDAR _{DISCO}	0	1	InfographicVQA _{DISCO}	0	1
RxPad	1–1	1–1	DUDE _{DISCO}	1	0
Total	2	2	Total	1	2

Table 24: Summary: which approach wins on **parsing** (primary metric: S_{CS}) and **question answering** (primary metric: $S_{\text{GT-in-Pred}}$). Totals count a comparable result (1–1) as 1 point for both approaches.

Criterion	DocVQA	InfographicVQA	DUDE	Overall
<i>Parsing Effectiveness (GT-in-Extracted-Text)</i>				
Azure Intelligence	1	1	1	3/3
Mistral OCR 2	0	0	~	0/3
Mistral OCR 3	0	0	0	0/3
GPT-5 Mini (VLM)	~	0	1	1/3
<i>QA Performance (GT-in-Pred with best prompt)</i>				
QA _{OCR}	0	0	1	1/3
QA _{VLM-2stage}	0	0	0	0/3
QA _{VLM-direct}	1	1	0	2/3
<i>Computational Efficiency (lowest latency)</i>				
Azure Intelligence	0	0	0	0/3
Mistral OCR 2	1	~	1	2/3
Mistral OCR 3	0	1	0	1/3

Table 25: Winner summary by dataset and evaluation criterion. 1 = wins, 0 = loses, ~ = comparable performance (within 2 percentage points).

C STATISTICAL VERIFICATION

C.1 BOOTSTRAP METHODOLOGY

All confidence intervals reported in this paper were computed using bootstrap resampling with the following parameters:

- **Bootstrap iterations:** 10,000
- **CI method:** Percentile method (2.5th and 97.5th percentiles)
- **Bootstrap unit:** Sample-level resampling with replacement
- **Phase aggregation:** For each bootstrap iteration, we resample samples, compute metrics for each model, then average across models within the phase
- **Confidence level:** 95%

This approach correctly accounts for sample-level variability (the independent unit of observation) while propagating uncertainty through phase averaging. The resulting confidence intervals are presented as point estimate \pm margin, where margin is half the CI width.

C.2 NARROW-MARGIN COMPARISONS

We performed pairwise statistical tests on comparisons where performance differences were small, using the Wilcoxon signed-rank test (two-sided, $\alpha = 0.05$).

C.2.1 RXPAD: PA (OCR) VS PC (VLM-TASK)

Phase	CER	95% CI
Pa (OCR)	0.658 ± 0.082	[[0.626, 0.790]
Pc (VLM-task)	0.660 ± 0.016	[0.644, 0.676]

Table 26: RxPad CER comparison between OCR and VLM-task approaches.

Statistical test results:

- **CI overlap:** Yes (overlap in range [0.652, 0.675])
- **Wilcoxon p-value:** 0.412
- **Significant** ($\alpha = 0.05$): No
- **Conclusion:** The OCR and VLM-task approaches demonstrate statistically comparable performance on prescription parsing. The overlapping confidence intervals and high p-value ($0.412 \gg 0.05$) indicate that the 4.8 percentage point difference could be attributed to sampling variability rather than a true performance gap.

C.2.2 CI WIDTH VALIDATION

To verify the bootstrap procedure, we checked that CI margins scale appropriately with sample size (margin $\propto 1/\sqrt{n}$):

The narrower CIs for larger sample sizes (IAM: $n=500$, margin=0.014) compared to smaller samples (Rx-Pad: $n=200$, margin=0.040) confirm the expected relationship. ICDAR’s anomalously large margin (0.561)

Sample Size	Avg CI Margin (CER)
$n = 200$ (RxPad)	0.040
$n = 491$ (ICDAR)	0.561
$n = 500$ (IAM)	0.014

Table 27: Average CI margins by sample size across parsing datasets. Note: ICDAR shows larger margins due to high model variance (Donut inclusion).

reflects genuine between-model variability rather than a methodological issue, as the Pa phase includes the Donut model which performed poorly on scene text.

C.3 INTERPRETATION GUIDANCE

Overlapping CIs: When 95% confidence intervals overlap, we cannot reject the null hypothesis of equal performance at the $\alpha = 0.05$ level. This indicates that observed differences may be due to sampling variability.

Non-overlapping CIs: When CIs do not overlap, this provides strong evidence of a true performance difference, though the converse is not always true (non-overlap does not guarantee significance).

Margin interpretation: A CI margin of ± 0.013 (e.g., IAM Pa: 0.099 ± 0.013) means we can be 95% confident the true phase-average CER lies between 0.086 and 0.112.

D DESIGN RECOMMENDATIONS

D.1 DOCUMENT-AWARE PIPELINE SELECTION

Based on our findings:

1. **Handwritten text:** Use specialised OCR. VLMs lag by $\sim 5\text{--}9\%$ even with task-aware prompting.
2. **Multilingual documents:** Use VLMs with generic prompts. OCR systems struggle on non-Latin scripts.
3. **Single-page visual QA:** Direct VQA finds correct answers most often. For precise formatting, consider OCR pipelines with prompt engineering.
4. **Multi-page documents:** OCR pipelines provide more reliable grounding for complex reasoning.
5. **Prompt design:** Start with generic prompts; task-aware prompts can degrade performance on diverse inputs.

We highlight the metric discrepancy between $S_{\text{GT-in-Pred}}$ and S_{ANLS} in Section A and recommend using complementary metrics depending on the desired evaluation properties. Additional figures for detailed results discussion are omitted in this anonymous version.

D.2 TIME EVALUATION

We analyse the trade-off between inference time and accuracy across the three QA strategies on DocVQA_{DISCO}, InfographicVQA_{DISCO}, and DUDE_{DISCO}. As an example, figure 9 plots average inference time (ms) against $S_{\text{GT-in-Pred}}$ for InfographicVQA.

On DocVQA_{DISCO}, direct VQA ($QA_{\text{VLM-direct}}$) dominates the efficiency frontier, achieving the highest $S_{\text{GT-in-Pred}}$ scores (0.87–0.91) with the fastest inference ($\sim 4\text{--}10\text{s}$). OCR-based pipelines (QA_{OCR}) show comparable latency but cluster into two accuracy regimes: high performance on single-page documents (DocVQA, InfographicVQA) and lower performance on multi-page documents (DUDE). The two-stage VLM pipeline ($QA_{\text{VLM-2stage}}$) is consistently the slowest (17–35s), with inference time roughly doubling due to separate parsing and QA calls. When $QA_{\text{VLM-2stage}}$ achieves competitive accuracy, the latency cost is $2\text{--}4\times$ that of direct VQA.

For latency-sensitive applications, direct VQA offers the best accuracy-per-millisecond ratio. Two-stage pipelines are only justified when intermediate text representations are required for downstream tasks (e.g., retrieval, audit trails) or when processing very long documents where OCR-based grounding improves reliability despite the speed penalty.

D.3 COST EVALUATION

API costs depend on pipeline architecture and number of questions per document. For a typical document ($\sim 1,500$ image tokens, 500-token parsed text, 100-token answer), Table 28 compares per-document costs across strategies.

For single questions, $QA_{\text{VLM-direct}}$ and QA_{OCR} achieve comparable costs, while $QA_{\text{VLM-2stage}}$ incurs $2\text{--}3\times$ overhead by paying for parsed text as both output and input. However, when asking **multiple questions per document**, two-stage parsing amortises the extraction cost: break-even occurs at $\sim 4\text{--}6$ questions, after which QA_{OCR} or $QA_{\text{VLM-2stage}}$ become more economical. For multi-question workloads, QA_{OCR} offers the best cost-accuracy trade-off, combining cheaper text-only QA calls with the reliability advantages observed on long documents (Section B.2).

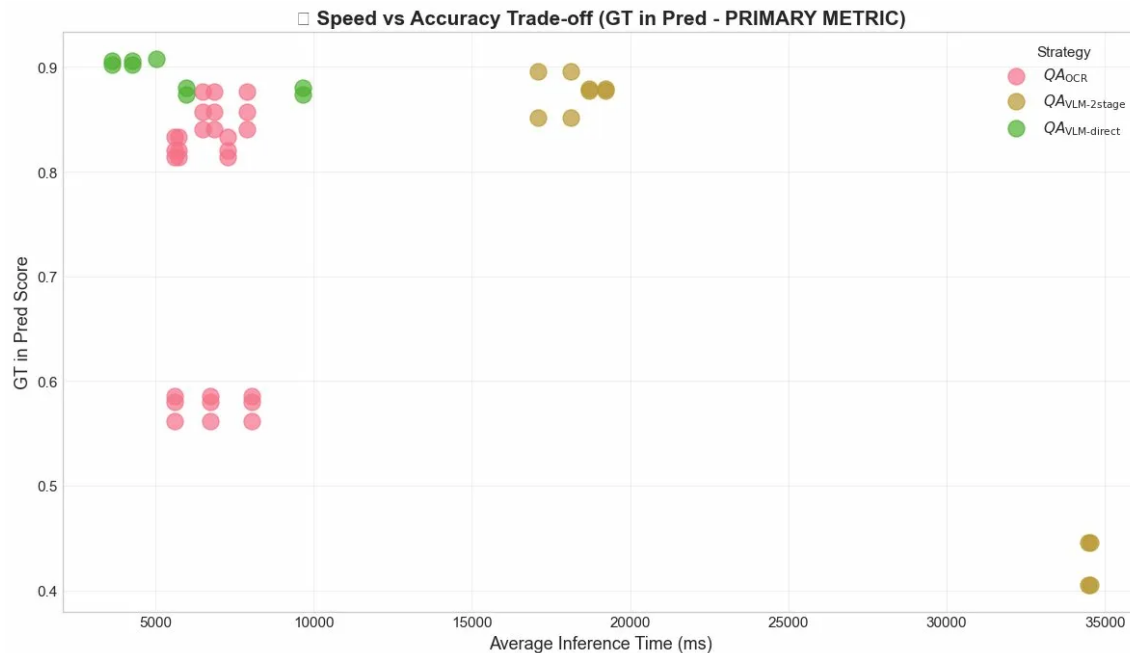


Figure 9: Speed vs accuracy trade-off across QA strategies on DocVQA_{DISCO}. $QA_{VLM-direct}$ (green) achieves the best efficiency frontier: highest accuracy with lowest latency. $QA_{VLM-2stage}$ (olive) incurs 2–4× longer inference times due to sequential parsing and reasoning stages.

Strategy	GPT-5-nano	GPT-5-mini	Claude 3.5 Sonnet
$QA_{VLM-direct}$ (1 call)	\$0.00011	\$0.00058	\$0.0060
$QA_{VLM-2stage}$ (2 VLM calls)	\$0.00034	\$0.00170	\$0.0150
QA_{OCR} (OCR + 1 LLM call) [†]	\$0.00012	\$0.00043	\$0.0046

Table 28: Per-document cost for single-question QA. [†]OCR cost estimated at \$0.001/page (Azure Document Intelligence); text-only LLM calls avoid image token costs.

E HOW WE BUILT THE DISCO BENCHMARK SUITE TO BE REPRESENTATIVE

E.1 DATASET CREATION METHODOLOGY

Sampling Strategy We employed three primary sampling strategies depending on each dataset’s characteristics:

1. **Simple random sampling:** For relatively homogeneous datasets without a strong categorical structure, we applied uniform random sampling with a fixed random seed (42) to ensure reproducibility.
2. **Stratified sampling:** For datasets with known categorical variables (e.g., question types, languages, or content types), we applied stratified sampling to preserve the proportional representation of each category.
3. **Balanced sampling:** For datasets with extreme class imbalance or multiple languages, we enforced balanced representation by sampling equal numbers of examples from each category.

Sample Size Selection Target sample sizes were selected based on statistical power analysis and computational constraints:

- **500 samples:** We use 500 samples for most datasets. This balances reliability with computational cost. With this sample size, 95% confidence intervals are roughly ± 4 percentage points, and we can detect differences of 7–9 percentage points between systems (80% power, $\alpha = 0.05$). For continuous metrics like S_{ANLS} , we can detect effect sizes of about 0.125 standard deviations.
- **Smaller samples:** For very expensive or rare datasets (RxPad: 200 samples), we used the full available dataset or maximum feasible subset.
- **Larger samples:** For multi-faceted datasets requiring diverse coverage (VisR-Bench: 498 documents with 17,045 QA pairs), we sampled at the document level while preserving multiple questions per document.

E.2 TEXT PARSING (OCR) DATASETS

IAM_{DISCO}

- **Source:** IAM Handwriting Database Marti & Bunke (2002)
- **Task:** Handwriting recognition (parsing)
- **Original size:** $\sim 11,539$ images with both printed and handwritten together
- **DISCO subset size:** 500 text line samples
- **Sampling:** Random sampling across writers and text styles
- **Data format and contribution:**
 - **Reference:** printed ground-truth image (`printed.png`)
 - **Input:** handwritten text line image (`handwritten.png`)
 - Pre-cropped images for consistent evaluation

ICDAR_{DISCO}

- **Source:** ICDAR 2015 competition on robust reading Karatzas et al. (2019)
- **Task:** Multi-lingual scene text recognition (parsing)

- **Original size:** $\sim 10,000$ images
- **DISCO subset size:** 500 samples (balanced; 50 per language category)
- **Sampling:** Stratified balanced sampling across 10 language categories, 50 samples each
- **Key metadata and contribution:**
 - Text transcription with language identifier
 - Position metadata for reading order

PubLayNet_{DISCO}

- **Source:** PubLayNet document layout dataset Zhong et al. (2019)
- **Task:** Document layout analysis (parsing)
- **Original size:** 335,703 document images
- **DISCO subset size:** 500 page samples
- **Sampling:** Random sampling from scientific publications
- **Layout categories:** Text, Title, List, Table, Figure

RxPad *The original dataset was used for this experiment*

- **Source:** French medical prescription dataset (RxPad) Pattin et al. (2026)
- **Task:** Medical prescription parsing (French)
- **Original size:** 200 samples (150 training + 50 testing)
- **DISCO subset size:** 200 samples (full dataset; no subsetting)
- **Sampling:** Complete dataset inclusion from training and testing splits
- **Language:** French (fr)
- **Key annotations:** Prescriber/patient fields, medication details, dates/signatures, administrative codes
- **Image characteristics:** Mix of print and handwriting; structured form layouts (average resolution $1,474 \times 1,995$ px)

E.3 QUESTION ANSWERING (VQA) DATASETS

DocVQA_{DISCO}

- **Source:** Document Visual Question Answering (DocVQA), validation split Mathew et al. (2021)
- **Task:** Single-page document VQA (forms, receipts, letters)
- **Original size:** 5,349 QA pairs
- **DISCO subset size:** 500 QA pairs
- **Sampling:** Simple random sampling with seed=42 from the validation split
- **Content / document types:** Scanned business documents, forms, receipts, letters, and reports (variable scan quality; occasional handwriting)
- **Key metadata / annotations:**
 - Document ID and page number for traceability

- Multiple valid answer annotations (average 1.8 answers per question)
- Question-type metadata (layout, handwritten, figure/diagram, etc.)
- Document source identifiers (UCSF document collection)

InfographicVQA_{DISCO}

- **Source:** InfographicVQA, validation split Mathew et al. (2022)
- **Task:** Single-page infographic VQA (visual–text alignment + numerical reasoning)
- **Original size:** 5,186 QA pairs
- **DISCO subset size:** 500 QA pairs
- **Sampling:** Simple random sampling with seed=42 from the validation split
- **Content types:** Infographics, data visualizations, charts, statistical graphics
- **Key metadata / annotations:**
 - Pre-extracted OCR text from AWS Textract included in metadata
 - Operation/reasoning type annotations (arithmetic, comparison, etc.)
 - Longer questions on average (e.g., mean question length 14.2 words)

DUDE_{DISCO}

- **Source:** Document Understanding Dataset and Evaluation (DUDE) Van Landeghem et al. (2023)
- **Task:** Multi-page document QA (cross-page reasoning + localisation)
- **Original size:** 8,000+ QA pairs
- **DISCO subset size:** 404 QA pairs (kept below 500 samples for feasibility)
- **Sampling:** Stratified sampling across question families (with document-level capping; max 5 QAs per document) to prevent a single question type from dominating evaluation and to ensure balanced coverage of reasoning skills.
- **Target question-family distribution (percent & approx. count in 404):**
 - numeric_amount: 20% (~81)
 - date_time: 15% (~61)
 - lookup_entity: 40% (~162)
 - yes_no: 15% (~61)
 - multi_hop_other: 10% (~40)
- **Additional stratification dimensions:**
 1. **Answer type:** short text, long text, numeric, boolean
 2. **Document ID:** capped to prevent over-representation
- **Content / document types:** Real-world multi-page documents (invoices, receipts, forms, letters, financial reports, scientific papers) with multilingual content

ChartQA_{ProDISCO}

- **Source:** ChartQA Professional, validation split Masry et al. (2025a)
- **Task:** Chart QA (numerical + multi-step reasoning, conversational follow-ups)

- **Original size:** 1,948 QA pairs
- **DISCO subset size:** 494 QA pairs
- **Sampling:** Multi-dimensional stratified sampling across question type, answer type, and conversational depth
- **Representative distributions (preserved):**
 - Question types: Factoid (55.9%), Conversational (16.0%), Fact Checking (12.8%), Multiple Choice (10.7%), Hypothetical (4.7%)
 - Answer types: short text (38.3%), numeric (37.7%), boolean (13.2%), multiple choice (8.9%), long text (2.0%)
- **Key metadata / annotations:**
 - Multi-turn conversational samples (2–6 follow-up questions)
 - Paragraph context present for a subset of samples (12.6%)
 - Temporal/year-based reasoning required for a subset of samples (4.3%)

VisR-Bench_{DISCO}

- **Source:** Visual Retrieval Benchmark for long-context documents (VisR-Bench) Chen et al. (2025)
- **Task:** Multi-page document retrieval + question answering (IR + VQA)
- **Original size:** 394 documents; 17,045 total QA pairs (≈ 43.2 questions per document on average)
- **DISCO subset size:** 498 documents (document-level sampling; question capping to 5 QAs per doc by default)
- **Sampling:** Document-level sampling with per-document QA capping ($qa_per_doc \leq 5$) to address the highly unbalanced original QA distribution (some documents have many more questions than others) and to balance document diversity vs. question coverage
- **Content types:** Figure / table / text / multilingual documents (15 languages)
- **Key metadata / annotations:**
 - Each QA includes `page_index` for retrieval evaluation
 - Pre-extracted markdown available for all pages (`all_page_md_str`)
 - Wide length distribution: 2–417 pages per document (mean 21.2, median 7.0)
 - Capping is intended to (i) prevent documents with many questions from dominating evaluation, (ii) ensure fair per-document coverage, and (iii) preserve answer length distributions within each content type (figure/table/text/multilingual)

E.4 SAMPLING VALIDATION

To validate that our 500-sample subsets faithfully represent the full datasets, we performed statistical comparisons between the DISCO subsets and their source datasets. We focus on ChartQAPRO_{DISCO} (494 samples), which employed multi-level stratified sampling as described in Section E.

E.4.1 CHARTQAPRO_{DISCO} VALIDATION

We compared the distribution of key metadata properties between our 494-sample subset and the full test set (1,948 samples):

- **Question type distribution** (5 categories: Factoid, Conversational, Fact Checking, Multi Choice, Hypothetical)
- **Answer type distribution** (4 types: numeric, short_text, boolean, long_text)
- **Nested stratification** (question type \times answer type joint distribution)
- **Question length** (continuous, in characters)
- **Paragraph presence** (whether context paragraph is provided)

Table 29 presents the statistical test results. The primary stratification variable—question type—shows excellent representativeness ($p = 0.995$, Cramér’s $V = 0.010$), confirming that the stratified sampling successfully preserved the proportional distribution of question categories. The paragraph presence feature also shows strong representativeness ($p = 0.558$, Cramér’s $V = 0.026$).

Variable	Test	Statistic	p-value	Effect Size	Result
Question Type	$\chi^2 = 0.20$	(df=4)	0.995	V=0.010	Representative
Has Paragraph	$\chi^2 = 0.34$	(df=1)	0.558	V=0.026	Representative
Answer Type	$\chi^2 = 23.11$	(df=3)	<0.001	V=0.131	Minor deviation
Question Length	KS	D=0.158	<0.001	d=0.493	Minor deviation

Table 29: Statistical validation of sampling representativeness for ChartQAPRODISCO (n=494) compared to full test set (n=1,948). All tests use $\alpha = 0.05$. Chi-square tests compare categorical distributions; Kolmogorov-Smirnov (KS) tests compare continuous distributions. Effect sizes: Cramér’s V for categorical (small < 0.1, medium 0.1–0.3, large > 0.3); Cohen’s d for continuous (small < 0.2, medium 0.2–0.8, large > 0.8).

We observe minor deviations in answer type distribution (Cramér’s $V = 0.131$) and question length (Cohen’s $d = 0.493$). The answer type deviation is primarily due to a slight under-sampling of short_text answers (expected: 220, observed: 189) and over-sampling of boolean answers (expected: 60, observed: 65). These differences reflect the secondary stratification being conditioned on question type, which ensures diversity within each question category at the expense of perfect marginal alignment for answer types. The question length deviation (mini mean: 142 characters, full mean: 107 characters) suggests that our stratified sample captured slightly longer questions on average, possibly because longer questions are more common in certain question type strata.

Figure 10 visualizes these distributional comparisons. The question type bar chart (top-left) demonstrates near-perfect alignment, with all five categories matching their expected proportions within 1 percentage point. The nested heatmap (bottom) shows that the question type \times answer type joint distribution is well-preserved, with most cells differing by less than 2 percentage points between mini and full datasets.

Interpretation. These results provide strong evidence that ChartQAPRODISCO is representative of the full test set with respect to the primary design criterion: question type diversity. The perfect preservation of question type proportions (our primary stratification target) ensures that model performance metrics computed on the DISCO subset reflect the distribution of reasoning patterns in the full dataset. The minor deviations in secondary variables (answer types and question length) do not compromise this representativeness, as our multi-level stratified sampling was explicitly designed to prioritize question type balance over perfect marginal alignment of all features.

For practical purposes, these results validate that:

1. Performance comparisons across question types (Factoid, Conversational, etc.) on the DISCO subset accurately reflect the full dataset.

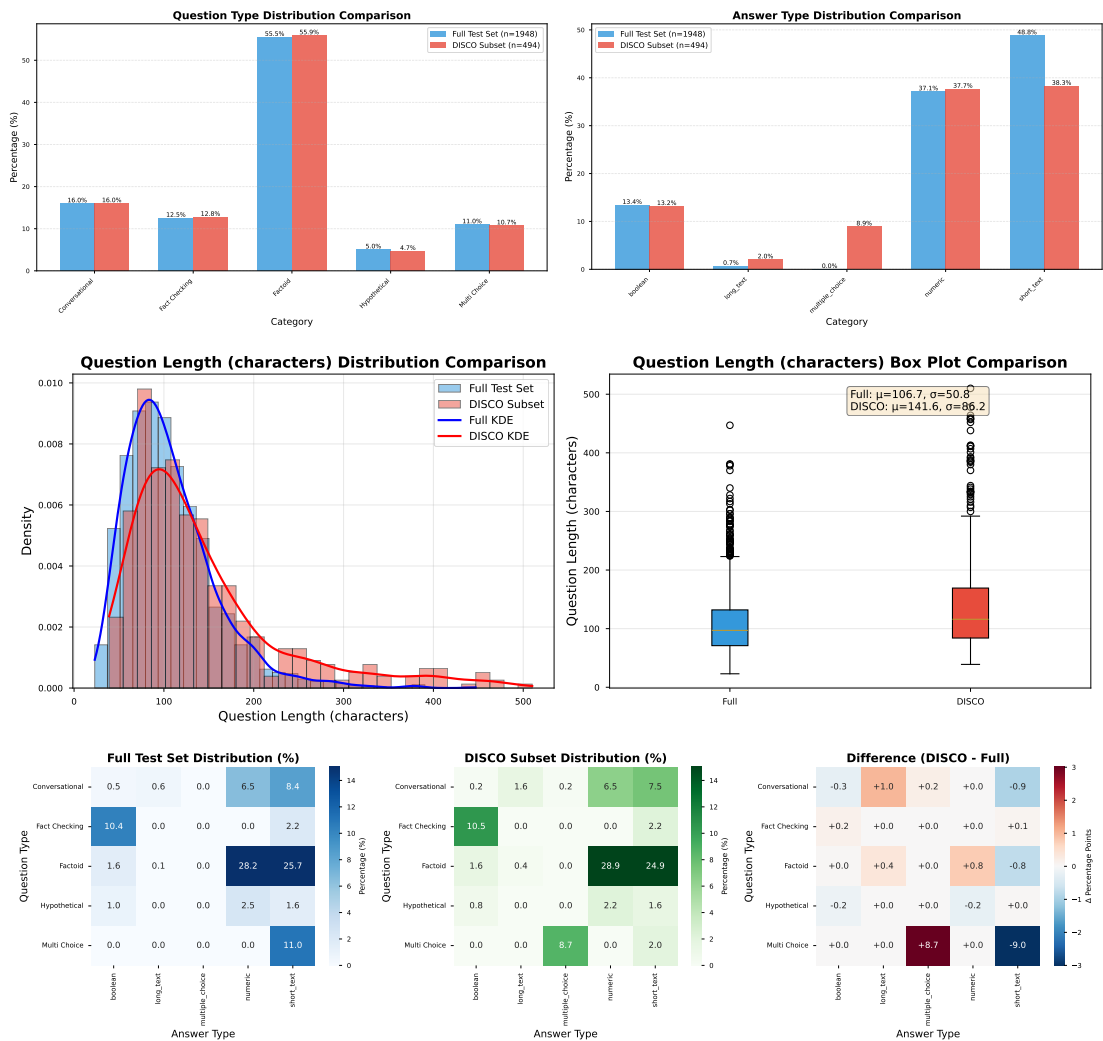


Figure 10: Sampling validation for ChartQAProDISCO. **Top:** Question type and answer type distribution comparisons. Question types (left) show nearly identical proportions across all five categories. Answer types (right) show minor deviations in the short_text and boolean categories. **Middle:** Question length distributions with overlaid kernel density estimates and box plots. The DISCO subset has slightly longer questions on average ($\mu=142$ vs. 107 characters). **Bottom:** Nested heatmap of question type \times answer type joint distribution. The left two panels show the full and DISCO distributions; the right panel shows the difference map with most cells within ± 3 percentage points.

2. The distribution of reasoning patterns and question categories is faithfully preserved.
3. Minor differences in question length and answer type distributions are within acceptable bounds for a 494-sample stratified subset (Cohen’s $d < 0.5$, Cramér’s $V < 0.15$).

	Parsing	Question answering (QA)
Primary metric	S_{CS} (Cosine Similarity): semantic similarity between the parsed output and the ground-truth text embeddings. We use S_{CS} as the primary parsing metric because it captures <i>semantic closeness</i> between extracted and reference text, even when the surface form differs (e.g., minor OCR noise, punctuation, or formatting changes).	$S_{GT-in-Pred}$ (Ground-Truth-in-Prediction): binary substring-based metric equal to 1 if the ground-truth answer string appears in the prediction, 0 otherwise. We use $S_{GT-in-Pred}$ as the primary QA metric because it supports a binary correctness check and is robust to verbose LLM outputs (where the correct answer may be embedded in a longer explanation).
Other metrics	CER (Character Error Rate): character-level edit distance normalised by ground-truth length; WER (Word Error Rate): word-level edit distance normalised by ground-truth word count.	EM (Exact Match): 1 if the prediction exactly matches the ground truth, 0 otherwise; ANLS (Average Normalised Levenshtein Similarity): normalised string similarity based on Levenshtein distance.

F EXPERIMENTAL SETUP

F.1 EXPERIMENTS

Parsing:

- P_{OCR} : OCR baseline using specialized OCR systems
- $P_{VLM-base}$: VLM baseline with generic text extraction prompts
- $P_{VLM-task}$: VLM with task-specific, domain-aware prompts

Question Answering:

- QA_{OCR} (**OCR**→**QA**): Specialized OCR extracts text, then VLM performs question answering
- $QA_{VLM-2stage}$ (**VLM**→**QA**): VLM extracts text from image, then same/different VLM performs question answering
- $QA_{VLM-direct}$ (**Direct VQA**): Single-step end-to-end VLM directly answers question from image

F.2 METRICS

F.3 MODELS

F.3.1 OCR MODELS

- **azure_intelligence**:
- **mistral-ocr-2505**: Mistral OCR 2
- **mistral-ocr-2512**: Mistral OCR 3

F.3.2 VLM MODELS

- **gpt-5-mini**: gpt-5-mini (vision-language)
- **gpt-5-nano**: gpt-5-nano (vision-language)
- **claude-3-5-sonnet**: claude-3-5-sonnet (vision-language)

G PROMPTS

G.1 PARSING TASK PROMPTS

G.1.1 PHASE P-OCR: OCR BASELINE

Note: Phase P_{OCR} uses specialised OCR systems (azure-ai-documentintelligence, mistral-ocr-2505) which do not require explicit prompts. These models are trained end-to-end for text extraction and operate directly on document images.

G.1.2 PHASE P-VLM-BASE: VLM WITH GENERIC PROMPTS

Prompt $P_{\text{VLM-base}}$ (used for all datasets: IAM_{DISCO}, ICDAR_{DISCO}, RxPad):

Extract all text from this image.

G.1.3 PHASE P-VLM-TASK: VLM WITH TASK-AWARE PROMPTS

Prompt $P_{\text{VLM-task-IAM}}$ (handwritten documents):

This is a handwritten document. Extract all text carefully preserving word boundaries and maintaining the original line structure.

Prompt $P_{\text{VLM-task-ICDAR}}$ (multilingual documents):

This document contains text in multiple languages including Arabic, Chinese, Japanese, Korean, and Latin scripts. Extract all text from the image, preserving the original script and character encoding. Maintain spatial layout where text appears in columns or mixed directions.

Prompt $P_{\text{VLM-task-VOC2007}}$ (Chinese medical reports):

This is a Chinese medical laboratory report. Extract all text from the document, including all numerical values, units of measurement, and Chinese characters. Preserve the tabular structure if present.

Prompt $P_{\text{VLM-task-PubLayNet}}$ (scientific papers):

This is a page from a scientific paper. Extract all text from the document, preserving the section structure (title, abstract, body text, figure captions, references). Maintain paragraph breaks and list formatting.

Models tested: gpt-5-mini, gpt-5-nano, claude-3-5-sonnet

G.2 QUESTION ANSWERING TASK PROMPTS

G.2.1 PHASE QA-OCR: OCR-TO-QA PIPELINE

Stage 1: Specialized OCR (azure-ai-documentintelligence, mistral-ocr-2505 or mistral-ocr-2512) extracts text without prompting.

Stage 2: VLM answers question based on extracted text using one of three prompt variants:

Prompt $QA_{\text{OCR}}^{\text{generic}}$ (generic):

Text: [extracted_text]

Answer: [question]

Prompt $QA_{\text{OCR}}^{\text{cot}}$ (chain-of-thought):

Based on the following text, answer the question. Think step-by-step about how to find the answer.

Text: [extracted_text]

Question: [question]

Provide your reasoning and then the final answer.

Prompt $QA_{\text{OCR}}^{\text{task-aware}}$ (task-aware + chain-of-thought):

You are analyzing a document. The extracted text from the document is provided below.

Extracted text: [extracted_text]

Answer this specific question about the document:
[question]

Think step-by-step about how to find the answer. Provide your reasoning and then the final answer.

G.2.2 PHASE QA-VLM-2STAGE: VLM-TO-QA PIPELINE

Stage 1: VLM extracts text from image using generic prompt (Prompt $P_{\text{VLM-base}}$, see Section G.1.2).

Stage 2: Same or different VLM answers question using one of three prompt variants:

Prompt $QA_{\text{VLM-2stage}}^{\text{generic}}$ (generic):

Answer: [question]

Prompt $QA_{\text{VLM-2stage}}^{\text{cot}}$ (chain-of-thought):

Based on the text below, answer the question. Think step-by-step about how to find the answer.

Text: [extracted_text]

Question: [question]

Provide your reasoning and then the final answer.

Prompt $QA_{\text{VLM-2stage}}^{\text{task-aware}}$ (task-aware + chain-of-thought):

You are analyzing a document to answer a specific question. The text extracted from the document is provided below.

Document text: [extracted_text]

Question: [question]

Think through the question step-by-step: 1. Identify relevant information in the text 2. Reason about how it answers the question 3. Formulate your final answer

Provide the final answer after your reasoning.

G.2.3 PHASE QA-VLM-DIRECT: DIRECT VQA (SINGLE-STAGE)

Single stage: VLM directly answers question from image without explicit text extraction step. Two prompt variants:

Prompt $QA_{\text{VLM-direct}}^{\text{fewshot}}$ (few-shot examples):

You will be shown document images and questions about them. Here are some examples:

[Image 1]
Question: What is the invoice total?
Answer: \$1,234.56

[Image 2]
Question: What is the sender's name?
Answer: John Smith

Now answer this question about the following document:

[Target Image]
Question: [question]
Answer:

Note: Specific examples varied by dataset (DocVQA vs InfographicVQA) to match document types.

Prompt $QA_{\text{VLM-direct}}^{\text{generic}}$ (generic):

Look at this document image and answer the question.

Question: [question]

Provide a concise answer based on the visible content in the image.

H QUALITATIVE EXAMPLES

H.1 PARSING

H.1.1 IAM-DISCO (PARSING)

#1 $S_{CER} = 0.0521 \mid S_{WER} = 0.2381 \mid S_{ANLS} = 0.9491 \mid S_{CS} = 0.9502$

Prediction:

He looked at her. Head thrown back in a pool of hair, her blood-red lips parted and the beating of her heart in the full throat. Her mouth did things he thought no human being could stand without dying, but he went on living in an ocean of voluptuousness, that swelled and ebbed over him, under him, in him and through him ...

Ground truth:

He looked at her . Heard thrown back in a pool of hair , her blood-red lips parted and the beating of her heart in the full throat . Her mouth did things he thought no human being could stand without dying , but he went on living in an ocean of voluptuousness , that swelled and ebbed over him , under him , in him and through him ...

#2 $S_{CER} = 0.0638 \mid S_{WER} = 0.2344 \mid S_{ANLS} = 0.9371 \mid S_{CS} = 0.9568$

Prediction:

Unless they do at least that, Dr. Verwoerd will be able to return home claiming a triumph. His smile will be blander than ever. WE are in for it again: another Royal Wedding. Between now and June, when the Duke of Kent will marry Miss Worsley, hardly a day will pass without a story or a picture or probably both, about the nuptial arrangements.

Ground truth:

1
Unless they do at least that , Dr. Verwoerd will be able to return home claiming a triumph
His smile will be blander then ever . WE are in for it again : another Royal Wedding . Between
how and June , when the Duke of Kent will marry Miss Worsley, hardly a day will pass
without a story or a picture or probably both, about the nuptial arrangements

#3 $S_{CER} = 0.0790 \mid S_{WER} = 0.3279 \mid S_{ANLS} = 0.9265 \mid S_{CS} = 0.9566$

Prediction:

"Aw, forget it", she said cheerfully. "I'll sting you for a double for being a naughty boy. How about the telly tomorrow afternoon?" He felt a glow of happiness steal over him. Everything was all right now, thank God. She wasn't going to break with him, after all. For the moment it was the only thing in the world that mattered.

Ground truth:

' Aw , forget it " she said cheerfully . " I'll sting you for a double for being a naughty boy. How about the telly tomorrow afternoon ?" He felt a glow of happiness steal over himy . Everything was all right now, thank God. She wasn't going to break with himy , after all . for the moment it was the only thing in the world that mattered .

#4

$S_{CER} = 0.1300$ | $S_{WER} = 0.4737$ | $S_{ANLS} = 0.8846$ | $S_{CS} = 0.9153$

Prediction:

Then the whole earth will be His Altar. "And it shall come to pass, if lye shall lhearcken diligently unto my commandments, which I command you this day, to love the Lord your God, and to serve Him with all your heart and with all your soul." This may seem very good, but there is something deficient.

Ground truth:

They the whole earth will be His Altar . " And it shall came to pass , if Iye shall lhearkey diligently uyto my cormaycryents, which I coupyayd You this day , to love the Lord your God , and to serve thing with all your heart and with all your soul . "This may seem very good , but there Is something deficient .

#5

$S_{CER} = 0.1000$ | $S_{WER} = 0.2857$ | $S_{ANLS} = 0.9052$ | $S_{CS} = 0.9585$

Prediction:

The plain, sober manner of its style all the more tellingly points up not only the horror of the case itself, which floundered on to the electrocution four years later of a German-born Bronx carpenter named Bruno Richard Hauptmann, but to the raree-show emotionalism and sensation-hunger of that era.

Ground truth:

The plain , sober manner of its Style all the more tellingly points up not only the horror of the case itself , which floundered on to the electrocution four years later of a German - Broux carpenter named bom Bruno Richard Hauptmann , but to the raree-show emotionalism and sensation - hunger of that era.

#6

$S_{CER} = 0.1912$ | $S_{WER} = 0.3077$ | $S_{ANLS} = 0.8150$ | $S_{CS} = 0.8304$

Prediction:

The plain, sober manner of its style all the more tellingly points up not only the horror of the case itself, which flourished on to the electrocution four years later of a German-born Bronx carpenter named Bruno Richard Hauptmann, but to the rare-show emotionalism and sensation-hunger of that era.

Ground truth:

The peculiar social balance of the style in the whole relating points up not only the horror of the case itself, which flourished on to the electrocution four years later of a German-born Bronx carpenter named Bruno Richard Hauptmann, but to the rare-show emotionalism and sensation-hunger of that era.

#7

$S_{CER} = 0.1465$ | $S_{WER} = 0.2254$ | $S_{ANLS} = 0.8586$ | $S_{CS} = 0.8645$

Prediction:

There is just a hope that we may uncover some weakness, and find a way of fighting back at them. Michael agreed, and suggested that they use Dan as a specimen demonstrating how the Thetans machinations had been working out. It occurred to Steve that this may not have been entirely an objective suggestion on her part; but he thought it a good idea nevertheless.

Ground truth:

there is just a hope that we may uncover some weakness and find a way of fighting back at them. Heather agreed and suggested that they use Dan as a specimen demonstrating how the thetans manipulations had been working out. It occurred to Steve that this may not have been entirely an objective suggestion on her part, but he thought it a good idea nevertheless.

#8

$S_{CER} = 0.0824$ | $S_{WER} = 0.0909$ | $S_{ANLS} = 0.9176$ | $S_{CS} = 0.8890$

Prediction:

Tonight, for the first time, he had abandoned all pretence and shown her the honest desperation of his feeling for her. She had neither encouraged nor completely rejected him. In some perverse way their brief quarrel had forged a bond between them. No doubt she had every intention of keeping both of them on a string. On the whole he probably had a slight advantage over the young man, inasmuch as he had money to spend and she was a girl who had a healthy respect for the material things of life.

Ground truth:

Tonight, for the first time, he had abandoned all pretence and shown her the honest desperation of his feeling for her. She had neither encouraged nor completely rejected him. In some perverse way their brief quarrel had forged a bond between them. No doubt she had every intention of keeping both of them on a string. On the whole he probably had a slight advantage over the young man, inasmuch as he had money to spend and she was a girl who had a healthy respect for the material things of life.

#9

 $S_{\text{CER}} = 0.0996 \mid S_{\text{SWER}} = 0.1395 \mid S_{\text{ANLS}} = 0.9027 \mid S_{\text{CS}} = 0.8465$
Prediction:

In Fanny the pregnant girl is befriended by an old man. Here it is a young homosexual, estranged from women but yet moved by a strong instinct that extends to the unborn child as much as to the expectant mother, who acts as a protector and comforter to her in her hour of need. He shares her room and gives her his forlorn gift of companionship and sympathy - 'you need someone to love you while you are looking for someone to love'.

Ground truth:

In funny the pregnant girl is befriended by an old man. Here it is a young homosexual, estranged from women but yet moved by a strong maternal instinct to the unborn child as much as to the expectant mother who acts as a protector and comforter to her in her hour of need. He shares her room and gives her his fortune gift of companionship and sympathy - "you need someone to love you while you are looking for someone to love".

#10

 $S_{\text{CER}} = 0.0504 \mid S_{\text{SWER}} = 0.0968 \mid S_{\text{ANLS}} = 0.9496 \mid S_{\text{CS}} = 0.9105$
Prediction:

That is doubtful. If, however, in addition to her new good-neighbour gesture, Germany takes a really big share in giving aid to underdeveloped nations, the world outlook will be brighter. What gives rise to optimism is the sign that Germany and the other leading Western nations are at long last moving towards a solution of currency problems by co-operation.

Ground truth:

That is doubtful. If however, in addition to her new good-neighbour gesture, Germany takes a really big share in giving aid to underdeveloped nations, the world outlook will be brighter. What gives rise to optimism is the sign that Germany and the other leading Western nations are at long last moving towards a solution of currency problems by co-operation.

H.1.2 ICDAR-DISCO (PARSING)

Each sample reports parsing metrics, followed by the extracted text and the reference. We only report Latin-script examples in the sample gallery.³

#1

 $S_{\text{CER}} = 0.8710 \mid S_{\text{SWER}} = 0.8529 \mid S_{\text{ANLS}} = 0.0000 \mid S_{\text{CS}} = 0.7910$

³We restrict the displayed examples to Latin script to keep the PDF readable with the current font setup (especially for typewriter-styled blocks) and to avoid missing-glyph issues for non-Latin scripts.

Prediction:

MILANO
PALAZZO
MARINO
PIERO
FRANCESCA
DELLA
Misericordia
Madonna
della
La
Alessi
Sala
Palazzo
Marino,
Milano,
libero
Ingresso
gennaio
2017
all'8
2016
dicembre
dal
-
6
www.comune.milano.it
800167619
infoline
I
Rinascento
G
INTESA
CIVITA
PAIAZLORFALE

Ground truth:

MILANO · PALAZZO MARINO
PIERO DELLA FRANCESCA
La Madonna della Misericordia
Milano, Palazzo Marino, Sala Alessi
dal 6 dicembre 2016 all'8 gennaio 2017 - Ingresso libero
infoline 800167619
www.comune.milano.it

H.1.3 RXPAD (PARSING).

#2

$S_{CER} = 0.5486$ | $S_{WER} = 0.3719$ | $S_{ANLS} = 0.0000$ | $S_{CS} = 0.4628$

Prediction:

location: VILLEFRANCHE SUR SAONE,
date_of_prescription: 11/08/2021
1 comprimé matin et midi (selon besoin) - espacer 4h min
renew: Renouveler 3 fois
product_name: AVODART 0,5MG
product_name: TADALAFIL 5MG
product_name: LEVOCARNIL 100MG/ML
...

Ground truth:

VILLEFRANCHE SUR SAONE, le 11/08/2021
1 comprimé matin et midi selon besoin, en espaçant les prises de 4h minimum
pendant 1 mois.
A Renouveler 3 fois
AVODART 0,5MG CAPS MOLLE 30
...

#3 $S_{CER} = 0.5784$ | $S_{WER} = 0.3596$ | $S_{ANLS} = 0.0000$ | $S_{CS} = 0.5558$

Prediction:

structure_name: MAISON MEDICALE DE GARDE
VITAMINE C 1000: 1 cp/jour pendant 1 mois
MAGNE B6: 2 cp 3 fois/jour pendant 1 mois
DOLIPRANE 1000: 1 cp 3 fois/jour si douleurs ou fièvre pendant 3 jours
...

Ground truth:

MAISON MEDICALE DE GARDE
VITAMINE C 1000
1 cp par jour pendant 1 mois
MAGNE B6:
2 cp 3 fois par jour pendant 1 mois
...

#6 $S_{CER} = 0.8764$ | $S_{WER} = 0.8597$ | $S_{ANLS} = 0.0000$ | $S_{CS} = 0.4714$

Prediction:

LYON, le lundi 08 avril 2019
CENTRE MEDICAL OPHTALMOLOGIQUE POINT VISION
CARTEOL 2% LP UNIDOSES
1 goutte le matin dans les 2 yeux
OAR 1 an
...

Ground truth:

LYON, le lundi 08 avril 2019
CENTRE MEDICAL OPHTALMOLOGIQUE
CARTEOL 2% LP UNIDOSES
1 Goutte, LE MATIN, dans les 2 yeux
...

#7 $S_{CER} = 0.7340$ | $S_{WER} = 0.6792$ | $S_{ANLS} = 0.0000$ | $S_{CS} = 0.4828$

Prediction:

le 25 novembre 2022 11h25
MB-Etab 1
FLAGYL 500 mg: 1 comprimé 3 fois/jour pendant 15 jours
PHOSPHALUGEL: 1 sachet matin/midi/soir
BIRODOGYL: 1 comprimé matin/midi/soir pendant 7 jours
...

Ground truth:

Medecine Generale
MB-Etab 1
le 25 novembre 2022 11h25
FLAGYL 500 mg - Comprimé pellicule (Voie orale)
...

#8 $S_{CER} = 0.5860$ | $S_{WER} = 0.5310$ | $S_{ANLS} = 0.0000$ | $S_{CS} = 0.4417$

Prediction:

Dr name hidden (Pediatre)
LE 02/08/2021
Enfant 1 mois - Poids: 3,950 Kg
1) HEXYON (a 2 mois)
2) PREVENAR 13 (a 2 mois)
3) ROTARIX (a 2 mois)
4) PARACETAMOL (si fièvre après vaccins)
5) VIATOL (en cas de diarrhées)
...

Ground truth:

PEDIATRE
LE 02/08/2021
Enfant 1 mois
Poids : 3,950 Kg
1) HEXYON Susp inj ... a 2 mois
...

#9 $S_{CER} = 0.6495$ | $S_{WER} = 0.5488$ | $S_{ANLS} = 0.0000$ | $S_{CS} = 0.5556$

Prediction:

Docteur Lacramioara Vasilache (Medecine Generale)
43510 CAYRES
phone: 04 71 57 30 59
isoptine 240 lp: 1/matin
aprovel 300: 1 matin
furosemide 40: 1 matin et 1/2 midi
...

Ground truth:

Medecine Generale
isoptine 240 lp 1/matin
aprovel 300 1 matin
furosemide 40 1 matin et 1/2/midi
...

#10

$S_{\text{CER}} = 0.5386 \mid S_{\text{WER}} = 0.3556 \mid S_{\text{ANLS}} = 0.0000 \mid S_{\text{CS}} = 0.6359$

Prediction:

MEDECINE GENERALE
DOLIPRANE suppositoire: 1 toutes les 6h (2 boites)
DACRYOSERUM (unidoses): 1 boite
RIFAMYCINE collyre: 2 gouttes matin/midi/soir 7 jours
AUGMENTIN sirop: 1 dose/10 kgs matin/midi/soir 7 jours
...

Ground truth:

MEDECINE GENERALE
DOLIPRANE SUPPOSITOIRE 1 TTES LES 6 H 2 BTES
DACRYOSERUM 1 BTE UNIDOSES
RIFAMYCINE COLLYRE 2 GOUTTES MATIN MIDI ET SOIR 7 JRS
...

H.2 QUESTION ANSWERING

H.2.1 DocVQA-DISCO (QA)

#1

$S_{\text{GT-in-Pred}} = 1.0000 \mid S_{\text{ANLS}} = 0.8889 \mid S_{\text{CS}} = 0.8709 \mid S_{\text{EM}} = 0.0000 \mid S_{\text{SM}} = 1.0000$

Prediction:

\$3,000.00

Ground truth:

3,000.00

#2

$S_{\text{GT-in-pred}} = 1.0000 \mid S_{\text{ANLS}} = 1.0000 \mid S_{\text{CS}} = 1.0000 \mid S_{\text{EM}} = 1.0000 \mid S_{\text{SM}} = 1.0000$

Prediction:

123

Ground truth:

123

#3

$S_{\text{GT-in-pred}} = 1.0000 \mid S_{\text{ANLS}} = 1.0000 \mid S_{\text{CS}} = 1.0000 \mid S_{\text{EM}} = 1.0000 \mid S_{\text{SM}} = 1.0000$

Prediction:

34

Ground truth:

34

#4

$S_{\text{GT-in-pred}} = 1.0000 \mid S_{\text{ANLS}} = 1.0000 \mid S_{\text{CS}} = 1.0000 \mid S_{\text{EM}} = 1.0000 \mid S_{\text{SM}} = 1.0000$

Prediction:

Effect of HRT or Raloxifene on Endothelial Function

Ground truth:

Effect of HRT or Raloxifene on Endothelial Function

#5 $S_{GT_in_pred} = 1.0000 \mid S_{ANLS} = 1.0000 \mid S_{CS} = 1.0000 \mid S_{EM} = 1.0000 \mid S_{SM} = 1.0000$

Prediction:

Godfrey Phillips India Limited

Ground truth:

Godfrey Phillips India Limited

#6 $S_{GT_in_pred} = 1.0000 \mid S_{ANLS} = 0.0000 \mid S_{CS} = 0.6739 \mid S_{EM} = 0.0000 \mid S_{SM} = 1.0000$

Prediction:

The memorandum is written to the "FMI Workshop Speakers."

Ground truth:

fmi workshop speakers

#7 $S_{GT_in_pred} = 1.0000 \mid S_{ANLS} = 1.0000 \mid S_{CS} = 1.0000 \mid S_{EM} = 1.0000 \mid S_{SM} = 1.0000$

Prediction:

8

Ground truth:

8

#8 $S_{GT_in_pred} = 1.0000 \mid S_{ANLS} = 0.0000 \mid S_{CS} = 0.5310 \mid S_{EM} = 0.0000 \mid S_{SM} = 1.0000$

Prediction:

Point 5 is "POLYUNSATURATED FATS" -- they are recommended in ...

Ground truth:

Polyunsaturated fats

#9 $S_{GT_in_pred} = 1.0000 \mid S_{ANLS} = 1.0000 \mid S_{CS} = 1.0000 \mid S_{EM} = 1.0000 \mid S_{SM} = 1.0000$

Prediction:

Robert T. Quittmeyer

Ground truth:

Robert T. Quittmeyer

#10

$S_{GT_in_pred} = 1.0000 \mid S_{ANLS} = 0.0000 \mid S_{CS} = 0.7442 \mid S_{EM} = 0.0000 \mid S_{SM} = 1.0000$

Prediction:

Department of Obstetrics and Gynecology (New York University...)

Ground truth:

Department of Obstetrics and Gynecology

H.2.2 INFOGRAPHICVQA-DISCO (QA)

#1

$S_{GT_in_pred} = 1.0000 \mid S_{ANLS} = 1.0000 \mid S_{CS} = 0.7026 \mid S_{EM} = 1.0000$

Question:

Which is the IOT application shown in the infographic in green background color?

Prediction:

WhatsApp

Ground truth:

whatsapp

#2

$S_{GT_in_pred} = 1.0000 \mid S_{ANLS} = 1.0000 \mid S_{CS} = 1.0000 \mid S_{EM} = 1.0000$

Question:

What percentage of respondents backup photos on cloud according to the survey?

Prediction:

45.3%

Ground truth:

45.3%

#3

$S_{GT_in_pred} = 1.0000 \mid S_{ANLS} = 1.0000 \mid S_{CS} = 1.0000 \mid S_{EM} = 1.0000$

Question:

What is the ticket price (in dollars) for the quarter-finals of the 2015 ICC Cricket World Cup?

Prediction:

\$50.87

Ground truth:

\$50.87

#4

$S_{GT_in_pred} = 1.0000 \mid S_{ANLS} = 0.0000 \mid S_{CS} = 0.3686 \mid S_{EM} = 0.0000$

Question:

Which topic received the second highest mentions on media?

Prediction:

PPE -- personal protective equipment, with 1,589 mentions.

Ground truth:

ppe

#5

$S_{GT_in_pred} = 1.0000 \mid S_{ANLS} = 0.0000 \mid S_{CS} = 0.5260 \mid S_{EM} = 0.0000$

Question:

What is total percentage of businesses laying of staff and reducing staff hours?

Prediction:

68.3% (39.9% reduced staff hours/shifts + 28.4% laid off staff)

Ground truth:

68.3%

#6

$S_{GT_in_pred} = 1.0000 \mid S_{ANLS} = 1.0000 \mid S_{CS} = 1.0000 \mid S_{EM} = 1.0000$

Question:

What percentage of the survey respondents are non-travelers?

Prediction:

44%

Ground truth:

44%

#7

$S_{GT_in_pred} = 0.0000 \mid S_{ANLS} = 0.7619 \mid S_{CS} = 0.8440 \mid S_{EM} = 0.0000$

Question:

Besides earth, which are the other two planets featured in the movie Thor?

Prediction:

Asgard and Jotunheim.

Ground truth:

asgard, jotunheim

#8

$S_{GT_in_pred} = 1.0000 \mid S_{ANLS} = 1.0000 \mid S_{CS} = 0.9335 \mid S_{EM} = 1.0000$

Question:

Who is the author of *Stand on Zanzibar*?

Prediction:

John Brunner

Ground truth:

john brunner

#9

$S_{GT_in_pred} = 1.0000 \mid S_{ANLS} = 1.0000 \mid S_{CS} = 0.8668 \mid S_{EM} = 1.0000$

Question:

Which social media platform is ranked #1?

Prediction:

Facebook

Ground truth:

facebook

#10

$S_{GT_in_pred} = 1.0000 \mid S_{ANLS} = 1.0000 \mid S_{CS} = 1.0000 \mid S_{EM} = 1.0000$

Question:

What percentage of Americans have an e-reader as of January 2014?

Prediction:

32%

Ground truth:

32%

H.2.3 DUDE-DISCO (QA)

#1

$S_{GT_in_Pred} = 0 \mid S_{ANLS} = 0.0000$

Prediction:

The "ADMIN MODULES" box on the right side -- it contains the "Token Set Creation" link.

Ground truth:

Admin Modules.

#2

$S_{GT-in-Pred} = 1 \mid S_{ANLS} = 0.0000$

Prediction:

Christ University -- 1,400 student participants.

Ground truth:

CHRIST UNIVERSITY

#3

$S_{GT-in-Pred} = 0 \mid S_{ANLS} = 0.5333$

Prediction:

Data for Seniors Write Wikipedia

Ground truth:

Data for Seniors Write Wikipedia 2019 and first half of 2020

#4

$S_{GT-in-Pred} = 1 \mid S_{ANLS} = 0.0000$

Prediction:

Down -- the vertical axis arrow points downward toward the "Infinite Basis set" (y-axis increases downward).

Ground truth:

down

#5

$S_{GT-in-Pred} = 1 \mid S_{ANLS} = 0.0000$

Prediction:

Wet -- (wet your hands).

Ground truth:

Wet

#6

$S_{GT-in-Pred} = 0 \mid S_{ANLS} = 0.0000$

Prediction:

A cloud.

Ground truth:

anchor

#7

$S_{GT-in-Pred} = 0 \mid S_{ANLS} = 0.0000$

Prediction:

Photo credit: NASA/Kim Shiflett

Ground truth:

NASA/Ben Smegelsky

#8

$S_{GT-in-Pred} = 1 \mid S_{ANLS} = 0.0000$

Prediction:

The file number is 000049.

Ground truth:

49

#9

$S_{GT-in-Pred} = 1 \mid S_{ANLS} = 0.6279$

Prediction:

WEP stands for Wikipedia Education Program.

Ground truth:

WIKIPEDIA EDUCATION PROGRAM

#10

$S_{GT-in-Pred} = 1 \mid S_{ANLS} = 0.0000$

Prediction:

Yes -- the document includes a Los Angeles Times link

Ground truth:

Yes