Out of the Box: Zero-Shot Vision-Language Models for Redaction Detection and Page-Stream Segmentation

Özgür Ateş
University of Amsterdam
Amsterdam, The Netherlands
o.ates@student.uva.nl

David Graus
University of Amsterdam
Amsterdam, The Netherlands
d.p.graus@uva.nl

Abstract

Large collections of Dutch government documents are processed by OCR systems that extract plain text while discarding layout structure. Traditional OCR also struggles with visual redaction bars and cannot reliably split long page streams into separate documents. This paper investigates whether publicly available, modern Vision–Language Models (VLMs) can address these limitations out of the box, and offer a unified approach to layout-preserving OCR, automated redaction detection, and page-stream segmentation (PSS).

We evaluate Nanonets OCR-S in a zero-shot setting and introduce a dual-pass inference framework: a *text pass* that transcribes the page and tags redacted spans, and a *visual pass* that counts redaction bars directly from the page image. We compare three different redaction-detection inference pipelines, with cross-modal gating methods that aim to combine strengths from both modalities. Our baseline achieves the most balanced behavior (text F1 = 0.384, visual F1 = 0.542). Our first refined pipeline (V1) achieves a visual F1 of 0.574, a modest but real improvement over a count-level mean baseline (F1 = 0.479), though the comparison to task-specific models is limited by evaluation methodology: our VLM pipelines are evaluated on count accuracy only, not on redaction bar localization. For PSS, a refined prompt reaches F1 = 0.513 on the OpenPSS-LONG split, substantially outperforming a stratified random baseline (F1 ≈ 0.238), illustrating the strong influence of prompt design.

Compared to task-specific models (Mask R-CNN for redaction detection, which achieves F1 ≈ 0.95, and a multimodal ensemble for PSS, which achieves F1 ≈ 0.85), our zero-shot VLM approach is less accurate, but does generalize well across tasks within a single, general-purpose model. Our results indicate that VLMs can effectively reason over textual and visual features in ways that traditional OCR cannot, although performance is constrained by prompt sensitivity, and operational feasibility is limited by computational cost.

CCS Concepts

• **Computing methodologies** → **Computer vision tasks**; *Scene understanding*; • **Information systems** → **Document management and text processing**.

AI&OG @ ICAIL, Singapore

Keywords

Vision-Language Models, OCR, Document Understanding, FOIA, Redaction Detection, Open Government, Page-Stream Segmentation

ACM Reference Format:
Özgür Ateş and David Graus. 2026. Out of the Box: Zero-Shot Vision-Language Models for Redaction Detection and Page-Stream Segmentation. In *The First Workshop on Artificial Intelligence & Open Government at the 21st International Conference on Artificial Intelligence and Law, June 8, 2026, Singapore*. ACM, New York, NY, USA, 7 pages.

1 Introduction

The ongoing digitization of government and historical archives leaves large volumes of information locked inside scanned paper documents and image-based PDFs; providing transparent access to these government records has been identified as a major challenge in information retrieval [11]. And while concrete steps have been taken in increasing its accessibility [13], these approaches rely on traditional Optical Character Recognition (OCR), which is known to discard structural and spatial context that gives these documents meaning. This loss of layout limits accessibility, searchability, and automated analysis, as relationships between headers, footnotes, and tables are not preserved [8].

These limitations are especially critical in the public and legal domain. Records released under the Dutch Open Government Act (Woo), for example, frequently contain sensitive information such as personally identifiable information (PII), which is manually obscured by black or grey bars before release. Detecting these redactions reliably is necessary for transparency, privacy compliance, and machine-readability. In addition, documents released in response to Freedom of Information (FOIA) requests, typically end up with several document scans collated in a single PDF file; accurately identifying document boundaries, a task known as Page-Stream Segmentation (PSS), is equally important for information retrieval and record management.

Traditional OCR tools such as Tesseract [9] treat each page independently, ignoring visual cues that signal redactions or document boundaries. Current government digitization pipelines may therefore rely on multiple specialized components: OCR for text, computer vision for redaction detection, and heuristic models for segmentation; this type of fragmented setup can be inefficient, complex, and difficult to maintain.

Recent progress in Vision-Language Models (VLMs) offers a potential solution. Trained on large datasets of text paired with images, VLMs can reason over both modalities simultaneously, capturing textual content and layout structure holistically. Recent work has

demonstrated the potential of VLMs for OCR on historical data, including governmental records [7], suggesting that general-purpose vision-language capabilities can transfer to specialized archival domains. This paper explores how well modern and publicly available VLMs can generalize to Open Government-specific tasks beyond plain text extraction, specifically: automated redaction detection and PSS. In addition, we compare their performance with traditional OCR and task-specific computer vision systems.

This paper makes the following contributions:

- We propose a VLM-powered redaction detection framework that separates redaction detection into text and visual modality passes, and introduce inference pipelines with cross-modal gating strategies.
- We provide the first zero-shot VLM evaluation on two Dutch Open Government benchmarks, redaction detection and page-stream segmentation, enabling direct comparison to task-specific supervised methods.
- Our results show that zero-shot VLMs are surprisingly capable on specialized open government document tasks without in-context learning or task-specific training, with zero-shot performance meaningfully exceeding count-level and stratified random baselines. We also find that performance is primarily affected by prompt design, a relevant finding for practitioners deploying VLMs in document processing pipelines.

2 Related work

2.1 Traditional OCR and its limitations

Modern OCR engines such as Tesseract [9] leverage deep learning to achieve high accuracy on clean, well-formatted documents [1]. However, their sequential pipeline design (preprocessing → text detection → recognition → restructuring) inherently discards spatial and contextual information. Critical elements such as headers, tables, and footnotes lose their structural relationships, and visual features such as redaction bars or document boundary cues are entirely ignored.

2.2 Redaction detection

van Heusden et al. [14] developed neural image-segmentation approaches (Mask R-CNN and Mask2Former) for detecting redacted text, reporting precision 0.96 and recall 0.94 on pages containing redactions, dropping only slightly (precision 0.90, recall 0.92) when pages without redactions were included. This demonstrates the robustness of neural approaches over traditional pipelines. Integrating redaction detection into a unified OCR framework via VLMs remains an open challenge.

2.3 Page-stream segmentation

van Heusden et al. [12] introduced the OpenPSS benchmark for PSS using real Dutch government material. The full OpenPSS-LONG corpus contains 110 multi-document streams comprising 24,181 individual documents and 89,491 pages. The best-performing model was a late ensemble of Dutch BERT and EfficientNet, which achieved a weighted document-level F1 (Panoptic Quality, PQ) of 0.83 and a precision/recall of 0.87/0.83 on boundary detection. Single-modality models performed worse (Dutch BERT: 0.77, EfficientNet: 0.73), confirming the value of combining text and visual features. In this paper we evaluate on the official LONG *test split*, a smaller subset described in Section 4.1.

2.4 VLMs for document understanding

VLMs integrate a vision encoder and a language model through a shared embedding space [6]. The vision encoder converts image patches into dense feature tokens interpreted jointly with textual prompts, enabling cross-modal reasoning over layout, objects, and semantics. For redaction detection, the joint attention mechanism can associate dark rectangular regions with linguistic cues such as `[REDACTED]`. Models such as LayoutLM [15] have demonstrated that jointly encoding text and layout structure substantially improves document understanding, motivating the shift toward multimodal approaches for archival documents. The DocVQA benchmark [4] established document visual question answering as a standard evaluation setting for multimodal document understanding, demonstrating that VLMs can reason over both the textual content and visual layout of scanned documents. For PSS, VLMs can exploit textual transitions and visual cues across consecutive pages—capabilities that traditional OCR entirely lacks.

3 Methodology

3.1 Model selection

To select an appropriate model for our pipeline, we conducted a comparative evaluation of publicly available VLMs before settling on a final candidate. The evaluation spanned compact models (SmolVLM [3]), end-to-end document parsers (Donut [2]), multilingual models (Mistral OCR [5]), and dedicated OCR models (Nanonets OCR-S [10]).

Each model was assessed on our target task of structured text extraction from Dutch government documents containing redactions. SmolVLM exhibited hallucinations on long Dutch pages. Donut detected some visual redactions but produced inconsistent text extraction. Mistral OCR was not stably accessible during the evaluation period. **Nanonets OCR-S** consistently produced the most usable structured output in zero- and few-shot settings and was therefore selected as the primary model. It is built on the Qwen2.5 architecture with approximately 3 billion parameters, balancing capability with computational tractability.

3.2 Redaction detection

Deploying a VLM for redaction detection requires resolving a fundamental prompt design tension: a single prompt asked to simultaneously transcribe text and count redactions tends to hallucinate extra `[REDACTED]` tags or miss faint grey bars. We therefore propose our VL detection framework that separates these concerns across two independent *modality passes*, whose outputs are subsequently reconciled through a cross-modal gating method.

3.2.1 Dual modality passes. We explicitly leverage the multi-modal nature of VLMs by proposing a dual-pass framework, that combines both modalities through two passes. We illustrate this dual pass framework in the top half of Figure 1.

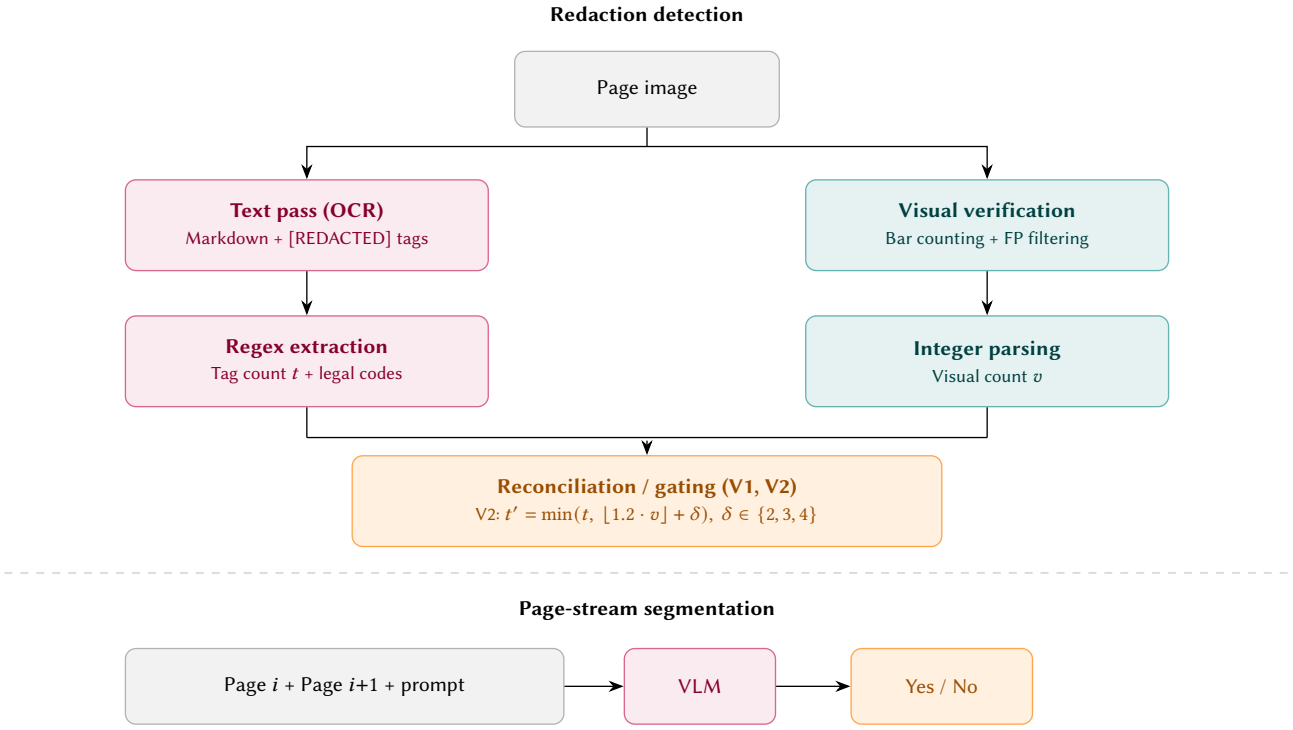

Figure 1: Architecture of the dual-pass redaction-detection pipeline (top) and the page-stream segmentation pipeline (bottom). For redaction detection, the page image is processed independently by a text pass (OCR with redaction tagging) and a visual verification pass (bar counting with false-positive filtering); the two counts are reconciled by the gating step used in V1 and V2. The formula shown is the V2 variant; V1 uses the simpler form $t' = \min(t, v + \delta)$ with $\delta = 2$. For PSS, two consecutive pages and a structured boundary prompt are fed jointly to the VLM, which returns a binary boundary decision.

Text pass: A role-based prompt instructs the model to act as a careful archivist, producing Markdown output that preserves document layout and inserts `[REDACTED]` exactly where bars cover text. Post-processing with regular expressions counts tags and extracts nearby legal references (e.g. Woo article "5.1.2.e"). This pass is sensitive to contextual cues but prone to overcounting: the model can repeat `[REDACTED]` tags, or misinterpret layout elements such as table borders or signature lines as redactions, producing unrealistically high counts.

Visual pass: A short prompt presents the same page image and asks only: *"How many redaction bars are visible on this page?"*, explicitly excluding table borders, underlines, and decorative elements. The answer is parsed as an integer. This pass is conservative and less affected by textual noise, but can undercount on dense pages where individual bars appear small in the full-page view, and the model struggles to attend to all regions simultaneously (e.g., when redacting multiple names in an email's To: field with a large number of recipients).

These two passes are complementary by design: the text pass captures subtle language and contextual cues, while the visual pass anchors predictions in the actual count of dark rectangles.

3.2.2 Inference pipelines. The dual-pass approach may result in different predictions between passes; naively reporting both counts leaves any discrepancy unresolved, and at the same time they may inform each other. We address this through cross-modal gating strategies, that ensure the visual modality can be used to constrain and ground the text pass, correcting for its tendency to overcount, while preserving its sensitivity to contextual cues.

Baseline. First, we report the text- and visual-pass counts independently, without reconciliation. This serves as a reference point against which our contributions are evaluated.

V1. Introduces *text–visual gating* ($t' = \min(t, v + \delta)$, $\delta = 2$), which limits the text-pass count with the visual-pass count as an upper bound, while allowing a small margin δ to accommodate redactions that the visual pass may have missed due to ambiguity or partial visibility. This prevents cases where the text pass may return, e.g., 40 redactions, while the visual pass confidently reports only 5. Next, to handle high-density pages (i.e., pages with many, typically small, redaction bars), where the model fails to detect individual bars at full-page resolution, we apply *dense-page tiling* and split the pages with a predicted count ≥ 15 into a 2×2 grid with 50-pixel overlap, allowing the model to focus on smaller regions, effectively increasing resolution where it matters. Finally, in V1 we apply *stricter prompting* to further reduce noise, e.g., through explicitly informing the model of page borders, table grid lines, underlines, or signature lines as potential false positives.

V2. Addresses three failure modes that persisted in V1: page-level variations in the gating margin (δ), false positives on visually

rich pages that have no redactions at all, and double-counting of redactions that span multiple tiles. We address the first by replacing the fixed margin δ with an *adaptive* variant ($t' = \min(t,\ \alpha \cdot v + \delta)$, $\alpha \approx 1.2$, $\delta \in \{2, 3, 4\}$). Here, pages with many small bars can tolerate a slightly larger margin, while keeping sparse pages more tightly constrained. Next, a lightweight *false-positive filter* scans for connected dark regions of sufficiently low pixel intensity; if none are found and the text pass contains no legal codes (that refer to redaction grounds), the visual count is forced to zero, reducing false positives on pages containing only text, logos, or photographs. Finally, we refine tiling with cross-boundary deduplication, to avoid double-counting redactions that span multiple tiles. The scaling factor $\alpha = 1.2$ was chosen manually as a heuristic tolerance to allow slight overcounting by the text-based pass relative to the visual count; only δ was systematically varied in the current experiments.

3.3 Page-stream segmentation

Prior work on PSS already established that multimodal approaches consistently outperform uni-modal methods [12]. Our approach operationalizes this insight using VLMs: rather than ensembling separate text and image models, we exploit the fact that Nanonets OCR-S jointly processes both modalities in a single inference pass, without the need for separate model pipelines. Our approach is illustrated in the bottom half of Figure 1.

Concretely, the model receives pairs of consecutive page images and is prompted to determine whether the second page begins a new document. The prompt provides explicit boundary cues drawn from the structural characteristics of Dutch government correspondence: signatures at the bottom of page i; changes in date, subject line, recipient, or sender; resets in page numbering or document identifiers; and clear layout shifts. The model outputs a binary decision (BOUNDARY: YES/NO), a confidence score, and a short natural-language justification.

Prompt design proved critical, here too. An initial neutral prompt defaulted almost always to "continuation," consistent with the known class imbalance in PSS datasets [12]. We then refined the prompt, by explicitly prioritizing boundary indicators, and breaking ties towards predicting boundaries, which substantially improved recall, analogous to the recall-oriented tuning strategies reported in the PSS literature. See Figure 2 for the refined PSS prompt.

```
SYSTEM: You are a document boundary detection
expert. Be sensitive to subtle differences that
mark new documents, especially signatures and
date changes.

USER: [Page i image] [Page i+1 image]
Determine if Page 2 starts a NEW document
(BOUNDARY: YES) or continues Page 1
(BOUNDARY: NO).

CHECK THESE IN ORDER OF IMPORTANCE:
 1. Signatures / sign-offs on Page 1
 2. Document IDs / reference numbers (different)
 3. Dates (different, even by 1 day)
 4. Page numbers (restart to "1")
 5. Recipients / addresses (different)
 6. Letter headers (new salutation)
 7. Subject lines (different)
 8. Sender names (different)

SAME-DOCUMENT INDICATORS:
 - Sequential page numbers
 - Mid-sentence continuation
 - "Continued..." text

DEFAULT: When uncertain and any indicator above
is present (especially signatures, dates, IDs),
prefer BOUNDARY: YES.

RESPOND EXACTLY:
 BOUNDARY: YES or NO
 CONFIDENCE: 0-100
 REASON: which indicator drove the decision
```

Figure 2: Refined PSS prompt used with Nanonets OCR-S. The initial neutral prompt omitted the priority-ordered cue list and the uncertainty default, causing the model to almost always predict continuation.

4 Experimental setup

All experiments were conducted on the Snellius HPC cluster (SURF), using a single NVIDIA H100 (80 GB VRAM) with 8 vCPUs and 32 GB of system RAM. The software stack comprised Python 3.11.3, PyTorch 2.x with CUDA, and the Hugging Face `transformers` and `accelerate` libraries. All pages were rasterised to 300 DPI PNG using PyMuPDF, and inference ran in bfloat16 precision with temperature 0.0 and deterministic decoding. The text pass used `max_new_tokens` = 3500; the visual pass used a substantially smaller limit, reflecting the simpler integer output expected.

4.1 Datasets

For both redaction detection and page-stream segmentation tasks we turn to publicly available, domain-specific datasets of Dutch Open Government data, allowing direct comparison to state of the art.

4.1.1 Redaction detection. We evaluate on the Neural Image Segmentation for Redacted Text Detection dataset [14], comprising 1,464 page images with 11,351 manually annotated redaction bars spanning black, grey, colored, and bordered variants. Evaluation uses the combined Classic (284 images) and Extended (439 images) test splits, yielding 723 pages with 7,172 annotated redaction bars. Pages with no redactions are retained as hard negatives to assess false-positive behavior.

4.1.2 Page-stream segmentation. For PSS we use the official test split of the OpenPSS-LONG dataset [12], a large-scale benchmark of Dutch government documents described in Section 2.3. The test split contains 197 multi-page document streams and 1,024 page images, yielding 827 page transitions for evaluation. Of these, 197 (23.8%) are true document boundaries and 630 are continuations; reflecting the class imbalance characteristic of realistic PSS settings, where boundaries are sparse by nature.

4.2 Evaluation

Since our V1 and V2 approaches affect the performance of the visual and textual passes in different ways (e.g., tiling affects the predictions of the visual pass, and the gating method affects predictions

of the textual pass), we report F1, precision and recall across both modalities (i.e., the text and visual passes described in §3.2.1), and each of the three inference pipelines (§3.2.2).

We compute these metrics based on redaction bar counts per page: we prompt the VLM to count redaction bars, and compare the predicted count per page against the ground-truth count, without localizing individual bars. We adopt this count-level evaluation because it matches the signal the VLM actually produces in our framework: the text pass emits `[REDACTED]` tags that we count via pattern matching, and the visual pass emits a single integer; in neither case does the model return bounding boxes or pixel masks. This contrasts with van Heusden et al. [14], who evaluate at the *instance* level, using spatial pixel-level overlap between predicted and annotated bars. Such a localization-based evaluation is in principle possible by prompting the VLM to emit pixel coordinates or polygons, and matching them against ground-truth regions, but this constitutes a fundamentally different evaluation setup (localization rather than counting). We therefore treat the comparison to van Heusden et al. [14] as useful context for situating zero-shot VLM performance, but not as a one-to-one comparison.

5 Results

5.1 Redaction Detection

In Table 1 we report performance of our three pipelines across textual and visual modalities, reporting Precision, Recall, and F1-scores.

Table 1: Redaction detection results across three pipelines.

	Text-based			Visual		
Pipeline	F1	Prec	Rec	F1	Prec	Rec
Baseline	**0.384**	**0.416**	**0.511**	0.542	**0.680**	**0.765**
V1	0.350	0.384	0.348	**0.574**	0.663	0.605
V2	0.322	0.397	0.308	0.264	0.270	0.263

Our baseline approach produced the most balanced behavior. V1 improved visual F1 marginally, at the expense of text-based performance. V2's aggressive heuristics lowered both text and visual F1, often suppressing genuine detections. However, subset analysis revealed that V2 was highly effective at eliminating false positives on clean pages (visual F1 $\approx$ 0.99 on pages without redactions), but performed poorly on grey and black bars. Dense pages ($\geq$15 redactions) were most challenging; the baseline handled them more consistently because tiling-and-gating interactions in V1 and V2 caused undercounting or double-counting.

A closer look at the precision–recall tradeoff across pipelines reveals an interesting asymmetry. For the text pass, recall drops substantially from baseline to V2 (from 0.511 to 0.308) while precision remains largely stable (from 0.416 to 0.397), suggesting that the cross-modal gating suppresses both true positives alongside the false positives we were trying to prevent. This is consistent with the gating mechanism's design: capping the text count at $v + \delta$ removes redactions from the top of the predicted count, with no way to distinguish real detections from hallucinated ones.

For the visual pass the picture is more symmetric: both precision and recall decline from V1 to V2 similarly (precision from 0.663 to 0.270, recall from 0.605 to 0.263), reflecting the more aggressive intervention of V2's false-positive filter, which can suppress the visual count entirely on pages it misclassifies as redaction-free.

Taken together, this suggests that the heuristics are effective at reducing overcounting in the text pass, but do so at a recall cost that is not compensated with higher precision.

5.1.1 Subset analysis. Looking at subsets in the dataset reveals substantial variation across pipeline configurations, that are lost in our aggregated metrics. On the Classic subset, V2's heuristics proved too restrictive. Here, text-based F1 dropped to approximately 0.145, indicating frequent suppression of legitimate detections. On the Extended subset, the models performed comparatively much better, perhaps as more heterogeneous layouts provide clearer visual cues, suggesting V2 does better with heterogeneous pages than uniform.

Redaction type also mattered: V2 was highly effective on clean pages, nearly eliminating false positives (text F1 $\approx$ 0.96, visual F1 $\approx$ 0.99), but performed poorly on grey and black bars, where its filtering heuristics removed subtle or low-contrast redactions that the baseline preserved. Colored redactions were detected more reliably across all pipelines, suggesting the model responds to saturation contrast rather than a semantic understanding of redaction as a concept.

Finally, redaction density was the most challenging factor: on dense pages ($\geq$15 bars), the baseline showed the most consistent performance, while V2's text-based F1 fell to approximately 0.07, as tiling-and-gating interactions caused severe undercounting.

Table 2: Comparison of our best-performing method to the Mask R-CNN method, alongside count-level random baseline fit on the training split.

Method	**Training**	F1	Prec	Rec
Mask R-CNN [14]	Task-specific	**0.950**	**0.960**	**0.940**
VLM V1 (visual)	Zero-shot	0.574	0.663	0.605
Random (mean, predict 8)	—	0.479	0.536	0.433
Random (median, predict 2)	—	0.240	0.717	0.145

The gap between the supervised Mask R-CNN model (F1 = 0.95) and our best zero-shot VLM pipeline (F1 = 0.574) should be interpreted with care: the two metrics are not directly comparable. For Mask R-CNN, a detection is only correct if its bounding box spatially overlaps with an annotated bar in the ground-truth dataset; our VLM pipelines are prompted to count the number of bars per page, and are hence evaluated on count accuracy, instead of localization of redaction bars. These two setups therefore represent different tasks (instance-level localization vs. page-level counting), and the gap conflates this difference in evaluation setup with the zero-shot versus supervised distinction; we include the comparison as context rather than as a direct comparison. To further contextualize our count-based evaluation, we include two trivial baseline predictors, that are fit on the 1,025 unique training pages of the classic and extended subsets, i.e., they always predict the rounded training mean (8) or median (2).

Since the metric only evaluates the predicted number of redaction bars per page, and ignores their location, a constant predictor close to the training mean already achieves F1 = 0.479. V1's 0.574 F1-score therefore represents a modest but real improvement over a distribution-matching guess, and notably does so by considering actual page content, rather than distribution matching; this represents a qualitative difference that count-based evaluation cannot fully capture.

5.2 Page-Stream Segmentation

Next, we report precision, recall, and F1-scores over the boundary class of our PSS approach, simple baselines, and the task-specific approach, in addition to accuracy, which is less informative as we'll discuss.

Table 3: PSS results on the OpenPSS-LONG test split. Prior work figures are boundary-detection precision/recall/F1 from Van Heusden et al. (the 0.83 cited in their paper is a different document-level weighted metric; 0.85 is the boundary F1 used here for comparison).

Method	Accuracy	Prec	Rec	F1
Majority (always continuation)	**0.762**	0.000	0.000	0.000
Random (stratified)	—	0.238	0.238	0.238
Initial prompt	0.754	0.000	0.000	0.000
Improved prompt	0.752	0.496	0.532	0.513
Van Heusden et al. [12]	—	**0.870**	**0.830**	**0.850**

Our initial neutral prompt almost never predicted a boundary, yielding F1 = 0.000 despite high accuracy, clearly reflecting the class imbalance and emphasizing how accuracy is unsuitable as a metric for evaluating PSS. This method also approaches the simple majority class baseline. Our refined prompt then produced a balanced prediction distribution and achieved F1 = 0.513, a dramatic improvement from a small phrasing change. This yields a substantial improvement over our random baseline, which predicts a boundary with probability equal to the observed boundary rate (23.8%), yielding an expected F1 = 0.238 under stratified random prediction. False negatives arose when two consecutive pages shared similar layout; false positives were triggered by signatures or abrupt layout shifts, which the refined prompt treats as boundary indicators.

6 Discussion

The results from our experiments highlight 2 recurring themes. First, *prompt design is decisive*: a small phrasing change in the PSS prompt shifted F1 from 0 (from predicting no boundaries) to 0.513, and the prompt phrasing for redaction detection determined whether the model over- or under-counted redactions. Systematic prompt optimization or meta-learning over prompts is therefore a promising direction for future work.

Second, *general-purpose VLMs are surprisingly capable out of the box*: without any task-specific training or fine-tuning, Nanonets OCR-S achieves meaningful performance on two specialized archival tasks involving Dutch government documents, that were not part of its training objective. The gap with supervised methods remains substantial, but the zero-shot results suggest that careful prompting alone can unlock a significant fraction of the headroom available without labeled data.

In addition, our results show that for redaction detection, a count-level mean baseline already achieves F1 = 0.479, simply because many pages have a similar number of bars. Our VLM (F1 = 0.574) improves on this by responding to actual page content rather than distribution-matching, a qualitative difference that count-based evaluation is too coarse to fully capture. For PSS, the picture is clearer: our refined prompt (F1 = 0.513) substantially outperforms a stratified random baseline (F1 ≈ 0.238), suggesting that the model is genuinely learning to exploit boundary cues rather than exploiting class priors.

Finally, one finding that is not immediately clear from our experimental results: scalability is a practical concern. Redaction detection requires dozens of seconds per page; PSS evaluation required approximately 1.2 seconds per transition. This means large-scale processing of government archives would require parallelization, batching, or model quantization to become operationally feasible.

Despite the above limitations and constraints, VLMs offer fundamentally different capabilities compared to traditional OCR approaches. VLMs can jointly reason about visual structure and textual content, and, next to addressing the two tasks studied in this paper, can produce layout-preserving Markdown output, and overall handle multiple document tasks within a single, general-purpose model, thereby avoiding fragmented pipelines of separate specialized tools.

Our results indicate that what drives VLM behavior on these tasks is not a single prompt string but the surrounding *instruction context*: role framing, task description, formatting constraints, and explicit cues about what does and does not count as a redaction bar or a document boundary. We therefore view instruction-context design as a first-class methodological factor for applying VLMs to open government document processing tasks, on par with model choice or post-processing heuristics, and one that should be reported and varied systematically in future work on zero-shot document understanding.

7 Conclusion

This paper examines to what extent a modern VLM, Nanonets OCR-S in a zero-shot setting, can address the limitations of traditional OCR for Dutch government documents, specifically for redaction detection and page-stream segmentation. The baseline redaction pipeline achieved an F1 of 0.542 based on visual processing; more elaborate heuristic variants did not consistently improve performance. For PSS, a refined prompt yielded F1 = 0.513 on the OpenPSS-LONG split. Both results fall substantially below their task-specific supervised counterparts (Mask R-CNN F1 ≈ 0.95; multimodal PSS ensemble boundary F1 ≈ 0.85), but the zero-shot VLM approach requires no task-specific training, and addresses multiple tasks within a single architecture.

Future work could explore few-shot learning, parameter-efficient fine-tuning (e.g., LoRA), improved multimodal fusion, systematic prompt optimization, and cross-lingual and generalization across sources (e.g., governmental organizations) or document types (e.g., emails vs. reports). Advances in model efficiency are needed before

full-scale open government processing becomes computationally feasible.

Acknowledgments

David Graus is partly funded by ICAI (AI for Open Government Lab). Views expressed in this paper are not necessarily shared or endorsed by those funding the research.

References

[1] AIMultiple. 2025. OCR Accuracy: How to Evaluate OCR Solutions. https://aimultiple.com/ocr-accuracy. Accessed: 2025-08-31.

[2] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. OCR-free Document Understanding Transformer. In *European Conference on Computer Vision*. Springer, 498–517.

[3] Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, et al. 2025. SmolVLM: Redefining Small and Efficient Multimodal Models. *arXiv preprint arXiv:2504.05299* (2025).

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

[5] Mistral AI Team. 2025. Mistral OCR: Introducing the World's Best Document Understanding API. https://mistral.ai/news/mistral-ocr. Accessed: 2025-11-10.

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning Transferable Visual Models from Natural Language Supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.

[7] Sina Semnani, Han Zhang, Xinyan He, Merve Tekgurler, and Monica Lam. 2025. CHURRO: Making History Readable with an Open-Weight Large Vision-Language Model for High-Accuracy, Low-Cost Historical Text Recognition. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, Suzhou, China, 34777–34824. doi:10.18653/v1/2025.emnlp-main.1763

[8] Gregory Slager and Maarten Marx. 2026. WCAG Compliance of Open Government Documents. In *New Trends in Theory and Practice of Digital Libraries*, Wolf-Tilo Balke, Koraljka Golub, Yannis Manolopoulos, Kostas Stefanidis, Zheying Zhang, Trond Aalberg, and Paolo Manghi (Eds.). Springer Nature Switzerland, Cham, 176–184.

[9] Ray Smith. 2007. An Overview of the Tesseract OCR Engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, Vol. 2. IEEE, 629–633. Source code: https://github.com/tesseract-ocr/tesseract.

[10] Souvik Mandal. 2025. Nanonets OCR Small OCR with semantic understanding: Go beyond just text. https://nanonets.com/research/nanonets-ocr-s/. Accessed: 2025-11-10.

[11] Johanne R. Trippas, J. Shane Culpepper, Mohammad Aliannejadi, James Allan, Enrique Amigó, Jaime Arguello, Leif Azzopardi, Peter Bailey, Jamie Callan, Rob Capra, Nick Craswell, Bruce Croft, Jeff Dalton, Gianluca Demartini, Laura Dietz, Zhicheng Dou, Carsten Eickhoff, Michael Ekstrand, Nicola Ferro, Norbert Fuhr, Dorota Glowacka, Faegheh Hasibi, Danula Hettiachchi, Rosie Jones, Jaap Kamps, Noriko Kando, Sarvnaz Karimi, Makoto P. Kato, Bevan Koopman, Yiqun Liu, Chenglong Ma, Joel Mackenzie, Maria Maistro, Jiaxin Mao, Dana McKay, Bhaskar Mitra, Stefano Mizzaro, Alistair Moffat, Josiane Mothe, Iadh Ounis, Lida Rashidi, Yongli Ren, Mark Sanderson, Rodrygo Santos, Falk Scholer, Chirag Shah, Laurianne Sitbon, Ian Soboroff, Damiano Spina, Paul Thomas, Julián Urbano, Arjen de Vries, Ryen White, Abby Yuan, Hamed Zamani, Oleg Zendel, Min Zhang, Shengyao Zhuang, Justin Zobel, and Guido Zuccon. 2025. Report from the 4th Strategic Workshop on Information Retrieval in Lorne (SWIRL 2025). *SIGIR Forum* 59, 1 (Oct. 2025), 1–68. doi:10.1145/3769733.3769739

[12] Ruben van Heusden, Jaap Kamps, and Maarten Marx. 2024. OpenPSS: An Open Page Stream Segmentation Benchmark. In *International Conference on Theory and Practice of Digital Libraries (TPDL)*. Springer, 413–429.

[13] Ruben van Heusden, Maik Larooij, Jaap Kamps, and Maarten Marx. 2025. A collection of FAIR Dutch Freedom of Information Act documents. *Scientific Data* 12, 1 (May 2025), 795. doi:10.1038/s41597-025-05052-2

[14] Ruben van Heusden, Kaj Meijer, and Maarten Marx. 2025. Redacted Text Detection Using Neural Image Segmentation Methods. *International Journal on Document Analysis and Recognition (IJDAR)* (2025), 597–607.

[15] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. LayoutLM: Pre-training of Text and Layout for Document Image Understanding. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3394486.3403172