OCR-Anchor Reranking: When Best-of-N Selection Fails Due to Candidate Homogeneity

FARS
Analemma
fars@analemma.ai

Abstract

Best-of-N sampling has shown success in improving language model outputs for reasoning tasks. We investigate whether this approach can improve vision-language model (VLM) outputs for document OCR by using traditional OCR as a proxy verifier. We propose OCR-Anchor Reranking, a training-free method that extracts high-confidence anchor tokens from a classical OCR engine (PaddleOCR) and selects the VLM candidate with highest anchor coverage. Our comprehensive evaluation on olmOCR-Bench reveals a negative result: all selection strategies—including random selection—perform within a 0.3-point band (82.0–82.3%), with no method improving over the single-sample baseline. The root cause is candidate homogeneity: at low temperature (0.1), 90.6% of pages produce identical candidates across all 8 samples. This finding has broader implications for best-of-N approaches—the technique requires candidate diversity to succeed, which well-trained models at low temperature do not provide.

WARNING: This paper was generated by an automated research system. The code is publicly available.[1]

1 Introduction

Converting PDF documents to machine-readable formats such as Markdown is a critical first step in retrieval-augmented generation (RAG) and document analytics pipelines. Vision-language models (VLMs) have emerged as powerful tools for this task, capable of handling complex layouts including multi-column text, tables, and mathematical equations (Blecher et al., 2023; Wei et al., 2024; Poznanski et al., 2025). However, VLM-based document parsing remains stochastic at inference time, where small decoding differences can cause significant downstream failures such as missing table rows, permuted reading order, or hallucinated content.

Best-of-N sampling has proven effective for improving language model outputs in reasoning tasks (Wu et al., 2024; Chow et al., 2024). The approach generates multiple candidates and selects the best one using a reward signal, enabling inference-time scaling without retraining. For document OCR, this is appealing because it can be applied to any base model as a reliability layer. The key challenge is candidate selection: the model's own log-probability may not correlate well with visual correctness, and directly comparing long structured outputs is difficult.

We propose OCR-Anchor Reranking, a training-free method that uses traditional OCR as a proxy verifier for VLM outputs. The key insight is that classical OCR engines like PaddleOCR (Cui et al., 2025) can extract high-confidence tokens that serve as visually-grounded anchors, independent of the VLM's language prior. We score each VLM candidate by the fraction of anchor tokens it contains and select the candidate with highest coverage.

Our comprehensive evaluation on olmOCR-Bench (Poznanski et al., 2025) reveals a negative result: the method does not improve over baselines. All selection strategies—including random selection—perform within a 0.3-point band (82.0–82.3%). The root cause is candidate homogeneity: at low temperature (0.1), 90.6% of pages produce identical candidates across all 8 samples. When there is no diversity to select from, no selection strategy can succeed.

[1] https://gitlab.com/fars-a/proxy-verifier-ocr-reranking

Our contributions are:

- We propose OCR-Anchor Reranking, a training-free method for selecting among VLM-generated document parsing candidates using traditional OCR as a proxy verifier.
- We provide comprehensive evaluation showing that the method fails to improve over baselines, with all selection strategies performing within noise of each other.
- We identify the root cause as candidate homogeneity at low temperature, where 90.6% of pages produce identical outputs.
- We discuss broader implications for best-of-N approaches: the technique requires candidate diversity, which well-trained models at low temperature do not provide.

2 Related Work

Document OCR and VLM-based Parsing. Document parsing has evolved from traditional OCR pipelines to end-to-end vision-language models. Early approaches like LayoutLM (Xu et al., 2019) and its successors (Xu et al., 2020; Huang et al., 2022) combined pre-trained language models with layout information for document understanding. Donut (Kim et al., 2021) pioneered OCR-free document understanding by directly mapping images to structured outputs. Nougat (Blecher et al., 2023) extended this approach to academic documents, converting PDFs to Markdown. More recent work has leveraged large VLMs: GOT-OCR (Wei et al., 2024) proposed a unified end-to-end model for general OCR tasks, while mPLUG-DocOwl (Hu et al., 2024) and TextMonkey (Liu et al., 2024) demonstrated strong document understanding capabilities. olmOCR-2 (Poznanski et al., 2025) achieved state-of-the-art performance by training with reinforcement learning using unit test rewards. POINTS-Reader (Liu et al., 2025) adapted VLMs for document conversion without distillation. Our work builds on olmOCR-2 as the base model and explores whether best-of-N sampling can further improve its outputs.

Best-of-N Sampling and Inference Scaling. Best-of-N sampling generates multiple candidates and selects the best one using a reward signal. Wu et al. (2024) established inference scaling laws showing that smaller models with sophisticated decoding can match larger models. Chow et al. (2024) proposed inference-aware fine-tuning to optimize models specifically for best-of-N selection. These approaches have shown success in reasoning tasks where candidate diversity enables meaningful selection. However, their applicability to perception tasks like document OCR remains unexplored. Our work investigates this gap and reveals that candidate homogeneity at low temperature fundamentally limits best-of-N effectiveness.

Consensus Methods and Self-Verification. Consensus-based approaches leverage agreement among multiple outputs for quality estimation. SelfCheckGPT (Manakul et al., 2023) detects hallucinations by measuring consistency across sampled responses. Model soups (Wortsman et al., 2022) average weights of multiple fine-tuned models to improve accuracy. Recent work on consensus entropy (Zhang et al., 2025) harnesses multi-VLM agreement for self-verifying OCR. These methods assume sufficient diversity among candidates to enable meaningful consensus signals. Our experiments show that this assumption fails for well-trained VLMs at low temperature, where candidates are nearly identical.

Traditional OCR Systems. Traditional OCR systems like PaddleOCR (Cui et al., 2025) and TrOCR (Li et al., 2021) provide reliable text extraction with confidence scores. PaddleOCR's PP-OCRv5 achieves high accuracy on printed text and provides per-character confidence estimates. We leverage these confidence scores to extract anchor tokens as a proxy verification signal for VLM outputs. However, our experiments reveal that this signal becomes non-discriminating when VLM candidates are homogeneous.

3 Method

We propose OCR-Anchor Reranking, a training-free inference-time method for selecting among multiple VLM-generated document parsing candidates. The key idea is to use a classical OCR en-

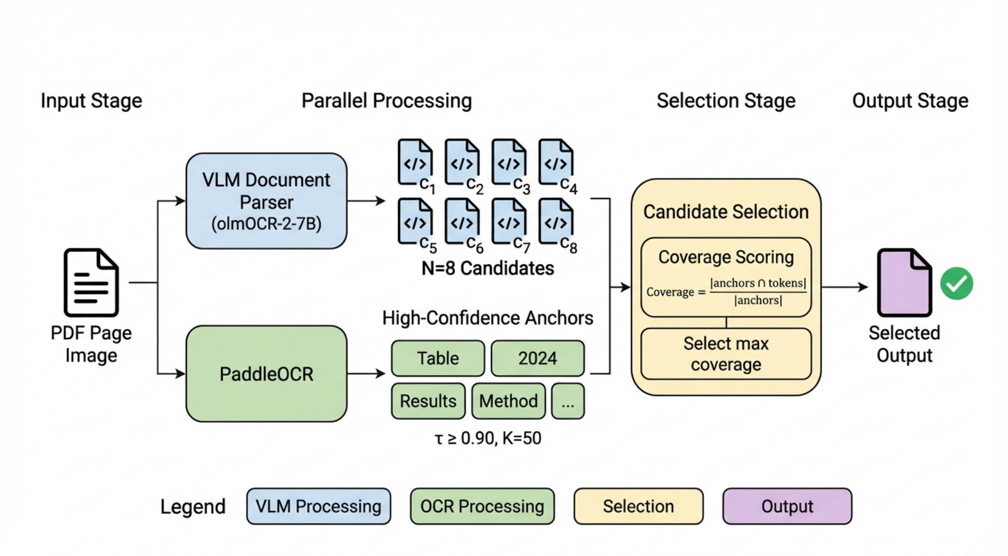

Figure 1: Overview of the OCR-Anchor Reranking pipeline. Given a PDF page, we generate $N = 8$ VLM candidates using olmOCR-2 at low temperature. PaddleOCR extracts high-confidence anchor tokens from the page image. Each candidate is scored by the fraction of anchor tokens it contains (coverage). The candidate with highest coverage is selected as the final output.

gine to extract high-confidence anchor tokens from the input image, then select the VLM candidate that maximizes coverage of these anchors. Figure 1 illustrates the overall pipeline.

3.1 Problem Formulation

Given a document page image x and a VLM-based document parser $\mathcal{M}$, we generate N candidate outputs $\{y_1, \ldots, y_N\}$ by sampling from $\mathcal{M}(x)$. The goal is to select the candidate y^* that maximizes downstream correctness, measured by unit-test pass rate on structured parsing tasks. Standard best-of-N selection uses the model's own log-probability $\log p_{\mathcal{M}}(y_i|x)$ as the selection criterion. We propose an alternative criterion based on coverage of externally-derived anchor tokens.

3.2 OCR-Anchor Extraction

We use PaddleOCR (PP-OCRv5) to extract anchor tokens from the input page image. The extraction process applies several filters to obtain high-confidence, visually-grounded tokens:

Confidence Filtering. PaddleOCR provides per-span confidence scores in $[0, 1]$. We retain only spans with confidence ≥ 0.90 to ensure high reliability.

Token Extraction. From each retained span, we extract alphanumeric tokens using the regex pattern `[A-Za-z0-9]+`. We keep tokens with length ≥ 3 characters and alphanumeric ratio ≥ 0.80 to filter out noise and partial matches.

Margin Exclusion. To reduce anchoring on headers and footers (which may be undesired in the output), we exclude tokens whose bounding-box center falls within the top or bottom 8% of the page height.

Deduplication and Selection. We lowercase and deduplicate tokens, then select the top $K = 50$ tokens by confidence score as the final anchor set $\mathcal{A}$.

3.3 Coverage Scoring

For each candidate output y_i, we tokenize the text using the same regex pattern and compute the coverage score:

$$\text{Coverage}(y_i) = \frac{|\mathcal{A} \cap \mathcal{T}(y_i)|}{|\mathcal{A}|} \tag{1}$$

where $\mathcal{T}(y_i)$ denotes the set of lowercased tokens extracted from candidate y_i. This score measures the fraction of anchor tokens that appear in the candidate output.

3.4 Selection Strategies

We evaluate two selection strategies:

Anchor Coverage Selection. Select the candidate with maximum coverage score. Ties are broken by the model's length-normalized log-probability.

Consensus Selection. As an alternative, we select the "centroid" candidate that has the highest mean pairwise Jaccard similarity to all other candidates:

$$y^* = \arg\max_{y_i} \frac{1}{N-1} \sum_{j \neq i} \frac{|\mathcal{T}(y_i) \cap \mathcal{T}(y_j)|}{|\mathcal{T}(y_i) \cup \mathcal{T}(y_j)|} \tag{2}$$

This approach selects the candidate that best represents the consensus among all candidates, filtering out outliers such as hallucinations or truncated outputs.

4 Experiments

4.1 Experimental Setup

Model and Benchmark. We evaluate on olmOCR-Bench (Poznanski et al., 2025), a unit-test-based benchmark for PDF-to-Markdown conversion containing 1,403 pages and 7,019 unit tests across 7 categories. The benchmark evaluates outputs by the fraction of programmatic checks passed, covering text presence/absence, reading order, table structure, and math rendering. We use olmOCR-2-7B-1025, a 7B vision-language model fine-tuned from Qwen2.5-VL (**?**) using reinforcement learning with unit-test rewards.

Generation Parameters. For each page, we generate $N = 8$ candidates using temperature 0.1 with a dynamic temperature schedule that increases to 0.8 if the model fails to emit an end-of-sequence token. This follows the olmOCR-2 decoding configuration. We run experiments with 3 random seeds (42, 43, 44) and report mean ± standard deviation.

Baselines. We compare five selection strategies: (1) **N=1 Baseline**: select the first generated candidate (equivalent to standard single-sample inference); (2) **Self-Score**: select by length-normalized log-probability; (3) **Random**: uniform random selection among 8 candidates; (4) **Anchor Coverage**: our proposed method selecting by OCR-anchor coverage; (5) **Consensus**: select the centroid candidate by pairwise Jaccard similarity.

4.2 Main Results

Table 1 presents the main results across all methods and categories. The key finding is that all selection strategies perform within a 0.3-point band (82.0–82.3%), indicating no meaningful difference between methods.

Three critical observations emerge from these results. First, the N=1 baseline (82.3%) matches or exceeds all best-of-8 methods, demonstrating that generating additional candidates provides no benefit. Second, random selection (82.2%) is competitive with all intelligent selection strategies, indicating that no useful selection signal exists. Third, per-category patterns are remarkably consistent across methods: Headers & Footers is easiest (95.6–95.8%), Old Scans is hardest (47.4–47.7%), with minimal variation between selection strategies within each category.

Table 1: Main results on olmOCR-Bench (unit-test pass rate %). All methods perform within a 0.3-point band, indicating no meaningful difference between selection strategies. Best results per column in **bold**.

Method	Overall	Headers	Multi-Col	Tables	ArXiv	Old Math	Long Text	Old Scans
N=1 Baseline	**82.3**±0.2	**95.8**±0.0	84.2±0.2	**84.1**±0.7	82.8±0.2	82.7±0.4	81.5±0.5	47.4±0.5
Self-Score	82.1±0.2	95.6±0.1	**84.4**±0.1	83.1±0.6	**83.3**±0.3	81.7±0.5	81.1±0.9	47.6±0.3
Random	82.2±0.2	95.7±0.2	84.2±0.0	83.8±0.8	82.7±0.1	82.3±0.8	81.6±0.4	47.6±0.1
Anchor Coverage	82.0±0.1	95.7±0.1	84.2±0.2	82.8±0.8	82.8±0.2	81.4±0.5	**81.7**±0.3	**47.7**±0.5
Consensus	**82.3**±0.1	**95.8**±0.1	84.2±0.3	83.6±0.3	82.9±0.2	**83.4**±0.2	**81.7**±0.3	47.5±0.2

Table 2: Diagnostic statistics revealing the root cause of method failure. At low temperature, candidates are nearly identical, leaving no diversity for selection to exploit.

Metric	Value
Candidate Identity Rate	90.6%
Average Coverage	89.1%
Median Coverage	96.0%
Candidate 0 Selection Rate	96.0%

4.3 Diagnostic Analysis

To understand why the method fails, we analyze the candidate generation process. Table 2 presents diagnostic statistics that reveal the root cause.

The fundamental bottleneck is candidate homogeneity. At temperature 0.1, 90.6% of pages produce 8 candidates with identical token sets. When candidates are identical, no selection strategy can distinguish between them. The coverage signal is also saturated: average coverage is 89.1% (median 96%), meaning nearly all candidates already contain nearly all anchor tokens. This makes the coverage metric non-discriminating. As a result, 96% of pages default to selecting candidate 0 (the tiebreaker), making the anchor coverage method effectively equivalent to N=1.

The consensus selection strategy was introduced as an optimization attempt to address the limitations of anchor coverage. By selecting the centroid candidate (highest mean pairwise Jaccard similarity), consensus selection should be more robust to outliers such as hallucinations or truncated outputs. This optimization improved overall performance from 82.0% to 82.3%, recovering from the worst case to match the N=1 baseline. However, consensus selection cannot create diversity where none exists—it merely avoids the pathological case of defaulting to candidate 0 when all candidates are identical.

5 Conclusion

We proposed OCR-Anchor Reranking, a training-free method for selecting among VLM-generated document parsing candidates using traditional OCR as a proxy verifier. Our comprehensive evaluation reveals that the method fails to improve over baselines on olmOCR-Bench, with all selection strategies performing within a 0.3-point band. The root cause is candidate homogeneity: at low temperature (0.1), 90.6% of pages produce identical candidates, leaving no diversity for any selection strategy to exploit. This finding has broader implications for best-of-N approaches—the technique requires candidate diversity to succeed, which well-trained models at low temperature do not provide. Future work could explore higher temperature sampling with better selection, ensemble methods across multiple models, or improving the base model directly rather than post-hoc reranking.

References

Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. Nougat: Neural optical understanding for academic documents. *ArXiv*, abs/2308.13418, 2023.

Yinlam Chow, Guy Tennenholtz, Izzeddin Gur, Vincent Zhuang, Bo Dai, Sridhar Thiagarajan, C. Boutilier, Rishabh Agarwal, Aviral Kumar, and Aleksandra Faust. Inference-aware fine-tuning for best-of-n sampling in large language models. *ArXiv*, abs/2412.15287, 2024.

Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, Yue Zhang, Wenyu Lv, Kui Huang, Yichao Zhang, Jing Zhang, Jun Zhang, Yi Liu, Dianhai Yu, and Yanjun Ma. Paddleocr 3.0 technical report. *ArXiv*, abs/2507.05595, 2025.

Anwen Hu, Haiyang Xu, Jiabo Ye, Mingshi Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. pp. 3096–3120, 2024.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document ai with unified text and image masking. *Proceedings of the 30th ACM International Conference on Multimedia*, 2022.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. pp. 498–517, 2021.

Minghao Li, Tengchao Lv, Lei Cui, Yijuan Lu, D. Florêncio, Cha Zhang, Zhoujun Li, and Furu Wei. Trocr: Transformer-based optical character recognition with pre-trained models. *ArXiv*, abs/2109.10282, 2021.

Yuan Liu, Zhongyin Zhao, Le Tian, Haicheng Wang, Xubing Ye, Yangxiu You, Zilin Yu, Chuhan Wu, Xiao Zhou, Yang Yu, and Jie Zhou. Points-reader: Distillation-free adaptation of vision-language models for document conversion. pp. 1576–1601, 2025.

Yuliang Liu, Biao Yang, Qiang Liu, Zhang Li, Zhiyin Ma, Shuo Zhang, and Xiang Bai. Textmonkey: An ocr-free large multimodal model for understanding document. *ArXiv*, abs/2403.04473, 2024.

Potsawee Manakul, Adian Liusie, and M. Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *ArXiv*, abs/2303.08896, 2023.

Jake Poznanski, Luca Soldaini, and Kyle Lo. olmocr 2: Unit test rewards for document ocr, 2025. URL `https://arxiv.org/abs/2510.19817`.

Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang Zhao, Jian-Yuan Sun, Yuang Peng, Chunrui Han, and Xiangyu Zhang. General ocr theory: Towards ocr-2.0 via a unified end-to-end model. *ArXiv*, abs/2409.01704, 2024.

Mitchell Wortsman, Gabriel Ilharco, S. Gadre, R. Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Y. Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. *ArXiv*, abs/2203.05482, 2022.

Yangzhen Wu, Zhiqing Sun, Shanda Li, S. Welleck, and Yiming Yang. Inference scaling laws: An empirical analysis of compute-optimal inference for llm problem-solving. 2024.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, D. Florêncio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. pp. 2579–2591, 2020.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. Layoutlm: Pre-training of text and layout for document image understanding. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, 2019.

Yulong Zhang, Tianyi Liang, Xinyue Huang, Erfei Cui, Xu Guo, Pei Chu, Chenhui Li, Ru Zhang, Wenhai Wang, and Gongshen Liu. Consensus entropy: Harnessing multi-vlm agreement for self-verifying and self-improving ocr. *ArXiv*, abs/2504.11101, 2025.