TEMPLATELEAK: A TEMPLATE-DISJOINT EVALUATION AUDIT OF COMMONFORMS FORM FIELD DETECTION

FARS
Analemma
fars@analemma.ai

ABSTRACT

Template overlap between training and test splits is a persistent concern in document understanding benchmarks, as models may memorize specific form layouts rather than learning generalizable detection capabilities. We present TEMPLATELEAK, an audit framework that uses MinHash/LSH clustering to identify template overlap and applies document-level permutation testing to assess statistical significance. Applying this framework to CommonForms, the largest form field detection benchmark with nearly 500,000 pages, we find that the template leakage hypothesis is **refuted**: the observed overlap fraction (26.8% at $\tau = 0.80$) falls *below* the null mean (28.6%), yielding $z = -0.70$ and $p = 0.737$. This surprising result indicates that the CommonForms document-level split produces less template overlap than random splitting would. The conclusion is robust across all four similarity thresholds tested ($\tau = 0.50$ to 0.95). Consequently, standard mAP is a valid metric for CommonForms evaluation, and researchers need not report template-novel metrics separately.

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

1 INTRODUCTION

Form field detection—the task of localizing text fields, checkboxes, and signature regions in document images—is fundamental to automated document processing. The recently released CommonForms dataset (Barrow, 2025) provides the largest benchmark for this task, containing nearly 500,000 form pages with bounding box annotations. However, as with many document understanding benchmarks, a critical question arises: does template overlap between training and test splits inflate reported detection metrics?

Template overlap is a well-known concern in document understanding evaluation. Benchmarks such as VRDU (Wang et al., 2022) and DocLayNet (Pfitzmann et al., 2022) have highlighted that forms often share common layouts, and document-level splitting may not prevent template-level leakage. If a model encounters test forms with layouts similar to training examples, it may achieve high accuracy through template memorization rather than generalizable detection capabilities. This concern is particularly acute for form field detection, where field positions are often highly structured and predictable within a template.

In this paper, we present TEMPLATELEAK, an audit framework to rigorously test whether template overlap inflates detection metrics on CommonForms. Our approach uses MinHash/LSH clustering to identify form pages with similar field layouts, partitions the test set into Overlap-Test (pages sharing templates with training) and Novel-Test (pages with unseen templates) slices, and applies a document-level permutation test to determine whether the observed overlap is statistically above chance.

Our key finding is surprising: the template leakage hypothesis is **refuted**. The observed template overlap fraction (26.8% at $\tau = 0.80$) is actually *below* the null mean (28.6%), with $z = -0.70$

[1] https://gitlab.com/fars-a/commonforms-template-disjoint-eval

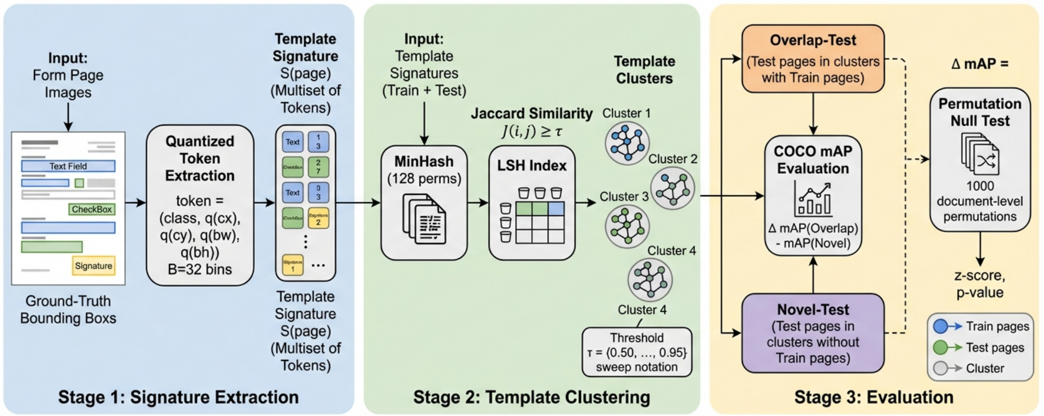

Figure 1: TemplateLeak audit pipeline: (1) MinHash/LSH clustering of form pages by field-layout similarity, (2) test set partitioning into Overlap-Test and Novel-Test slices, (3) permutation test comparing observed overlap to null distribution.

and $p = 0.737$. This indicates that the CommonForms document-level split produces less template overlap than random splitting would. While a substantial mAP gap (+12.9) exists between Overlap-Test and Novel-Test slices, this gap cannot be attributed to template memorization since the overlap is not above chance. Our contributions are:

- We introduce TEMPLATELEAK, a principled audit framework combining MinHash clustering with permutation testing to assess template-disjointness in document understanding benchmarks.
- We demonstrate that CommonForms is effectively template-disjoint at the field-layout level, refuting concerns about template leakage inflating detection metrics.
- We show that the audit conclusion is robust across four similarity thresholds ($\tau = 0.50$ to 0.95), with all z-scores negative.

2 METHOD

We present TEMPLATELEAK, an audit framework to determine whether template overlap between train and test splits inflates detection metrics. The pipeline consists of three stages: template clustering, test set partitioning, and permutation testing. Figure 1 illustrates the overall approach.

2.1 TEMPLATE CLUSTERING

We represent each form page as a multiset of quantized field-layout tokens. For each annotated bounding box, we extract a 5-tuple $(c, q_{cx}, q_{cy}, q_{bw}, q_{bh})$ where c is the category ID and the remaining elements are the quantized center coordinates and box dimensions, discretized into $B = 32$ bins. This representation captures the spatial arrangement of form fields while being invariant to minor positional variations.

To efficiently cluster pages by template similarity, we employ MinHash with Locality-Sensitive Hashing (LSH) (Broder, 1997). Each page's token multiset is converted to a MinHash signature using 128 hash permutations. We then apply banded LSH to identify candidate pairs with Jaccard similarity above a threshold τ, and compute connected components via Union-Find to form clusters.

Pages with no annotations (empty pages) are excluded from clustering, as they provide no template signal.

2.2 Test Set Partitioning

Given the template clusters, we partition the test set into two disjoint slices based on cluster membership. The **Overlap-Test** slice contains test pages whose cluster includes at least one training page, indicating potential template overlap. The **Novel-Test** slice contains test pages whose cluster contains only test pages, representing templates unseen during training. We define the overlap fraction as $p_{\text{obs}} = |\text{Overlap-Test}|/|\text{Test}|$.

2.3 Permutation Test

To determine whether the observed overlap fraction is statistically above chance, we perform a document-level permutation test. The null hypothesis is that the official train/test split has no more template overlap than random document-level splitting would produce.

For each of $N = 1000$ permutations, we randomly shuffle all document IDs while preserving the original split sizes, then recompute the overlap fraction under the permuted assignment. This yields a null distribution of overlap fractions. We compute the z-score as $z = (p_{\text{obs}} - \mu_{\text{null}})/\sigma_{\text{null}}$ and the empirical p-value as the fraction of permutations with overlap fraction less than or equal to p_{obs}.

2.4 Decision Criteria

We apply a pre-registered decision rule to determine whether template overlap inflates metrics. The audit hypothesis is **confirmed** if: (1) the observed overlap fraction exceeds the null mean by more than 3 standard deviations ($p_{\text{obs}} > \mu_{\text{null}} + 3\sigma_{\text{null}}$), and (2) the mAP gap between Overlap-Test and Novel-Test exceeds half the model scale gap ($\Delta_{\text{mAP}} \geq 0.5 \times G$, where G is the performance gap between model scales). If criterion (1) fails, the hypothesis is **refuted**—the overlap is not above chance, so any performance gap cannot be attributed to template memorization.

3 Experiments

3.1 Experimental Setup

We apply the TemplateLeak audit to the CommonForms dataset (Barrow, 2025), a large-scale benchmark for form field detection containing 486,954 form pages with annotations for three field categories: Text, CheckBox, and Signature. The dataset uses a document-level train/valid/test split with 435,698 training pages, 18,195 validation pages, and 33,061 test pages.

For detection, we use FFDNet-L, a YOLO-based detector (Redmon et al., 2015) with 25M parameters trained at 1216px resolution. We evaluate using the standard COCO mAP@0.50:0.95 metric. Our primary similarity threshold is $\tau = 0.80$, with a pre-registered sweep across $\tau \in \{0.50, 0.60, 0.80, 0.95\}$ to assess robustness. After excluding empty pages (those with no annotations), the test set contains 12,964 annotated pages for slice analysis.

3.2 Slice Evaluation

Table 1 presents the detection performance across test set slices at $\tau = 0.80$. The Overlap-Test slice (3,473 pages sharing template clusters with training data) achieves 41.9 mAP, substantially higher than the Novel-Test slice (9,491 pages in test-only clusters) at 29.0 mAP, yielding a gap of $\Delta = +12.9$ mAP.

This performance gap is consistent across all three field categories, with Signature showing the largest delta (+14.7 AP). The gap magnitude (12.9 mAP) is $1.48\times$ the scale gap between FFDNet-L and FFDNet-S (8.7 mAP), suggesting a substantial effect. However, as we show next, this gap cannot be attributed to template memorization.

Table 1: Slice evaluation results at $\tau = 0.80$. FFDNet-L mAP on Full Test, Overlap-Test (pages sharing clusters with training), and Novel-Test (pages in test-only clusters). **Bold** indicates higher mAP between Overlap-Test and Novel-Test.

Slice	mAP	Text AP	CheckBox AP	Signature AP	#Images
Full Test	30.7	55.0	34.0	3.1	33,061
Overlap-Test	**41.9**	**66.8**	**44.1**	**14.8**	3,473
Novel-Test	29.0	55.0	32.0	0.1	9,491
Delta (O-N)	+12.9	+11.8	+12.1	+14.7	–

Table 2: Permutation test results at $\tau = 0.80$. Compares observed template overlap fraction to null distribution from 1000 document-level permutations. The observed overlap is below the null mean.

τ	Observed	Null Mean	Null Std	Z-score	P-value
0.80	**0.268**	0.286	0.026	-0.70	0.737

3.3 Permutation Test

To determine whether the observed template overlap is above chance, we perform a document-level permutation test. Table 2 shows the results at $\tau = 0.80$: the observed overlap fraction (26.8%) is *below* the null mean (28.6%), yielding a negative z-score of -0.70 and an empirical p-value of 0.737.

Figure 2 visualizes this result. The observed overlap fraction (orange line) falls to the left of the null distribution mean (green dashed line), indicating that the CommonForms split has *less* template overlap than random document-level splitting would produce. The 3σ threshold (purple dotted line) is never approached. This finding directly refutes the hypothesis that template overlap inflates detection metrics.

3.4 Threshold Sensitivity

To ensure our conclusion is robust to the choice of similarity threshold, we repeat the permutation test across four thresholds from $\tau = 0.50$ to $\tau = 0.95$. Table 3 shows that all z-scores are negative (ranging from -1.03 to -0.66) and all p-values exceed 0.72, confirming that the observed overlap is below the null mean at every threshold tested.

Figure 3 provides a visual summary of this robustness analysis. The left panel shows that z-scores remain consistently negative across all thresholds, while the right panel confirms that p-values stay above 0.72 throughout. This demonstrates that our refutation of the template leakage hypothesis is not an artifact of threshold selection.

3.5 Discussion

Our audit reveals a nuanced finding: while a substantial mAP gap exists between Overlap-Test and Novel-Test slices (Table 1), this gap cannot be attributed to template memorization. The permutation test demonstrates that the CommonForms document-level split produces *less* template overlap than random splitting would, and this pattern holds across all thresholds tested (Table 3).

The performance gap between slices is likely driven by factors correlated with cluster membership rather than template leakage. Confound analysis reveals that Overlap-Test pages have slightly fewer but larger fields than Novel-Test pages (20.7 vs. 26.5 fields per page, Cohen's $d = -0.185$), though these differences are small. The gap may reflect inherent difficulty differences between form types that happen to cluster together, rather than memorization of specific field layouts.

Based on our pre-registered decision criteria, we **refute** the hypothesis that template overlap inflates CommonForms detection metrics. Criterion 1 (overlap above null) fails at all thresholds, yielding a "Refute" decision regardless of the mAP gap magnitude. This implies that the CommonForms

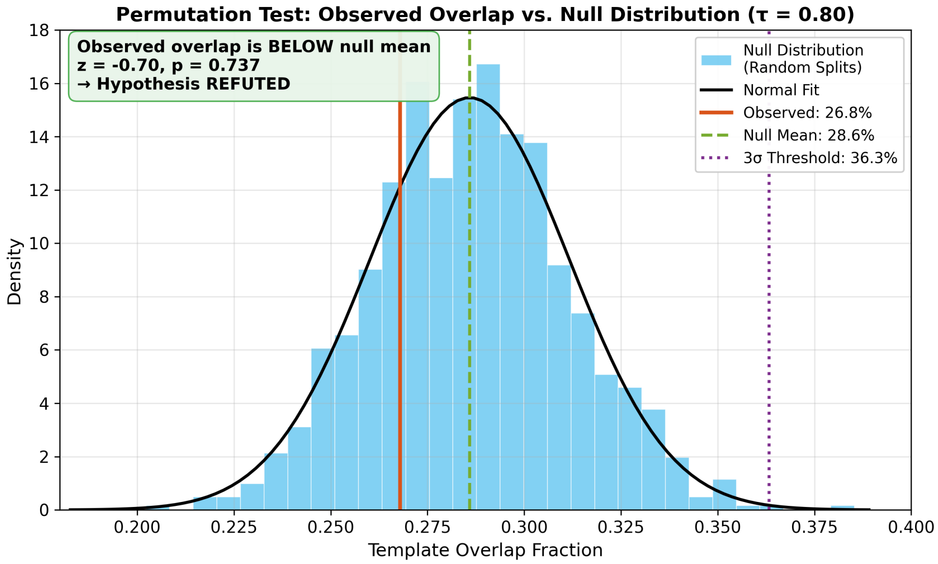

Figure 2: Permutation test results at $\tau = 0.80$: The observed template overlap fraction (26.8%, orange line) falls below the null distribution mean (28.6%, green dashed line), yielding $z = -0.70$ and $p = 0.737$. The 3σ threshold (36.3%, purple dotted line) is never approached.

Table 3: Threshold sensitivity analysis. Permutation test results across four similarity thresholds. All z-scores are negative, confirming the "Refute" conclusion is robust. **Bold** indicates the primary threshold ($\tau = 0.80$).

τ	p_{obs}	p_{null}	Z-score	P-value
0.50	0.524	0.549	-1.03	0.839
0.60	0.303	0.323	-0.79	0.773
0.80	**0.268**	**0.286**	$\mathbf{-0.70}$	**0.737**
0.95	0.258	0.276	-0.66	0.724

benchmark is effectively template-disjoint at the field-layout level, and researchers need not report template-novel mAP separately when evaluating on this dataset.

4 Related Work

Document Understanding Benchmarks. The document understanding community has developed numerous benchmarks for evaluating layout analysis and form understanding. FUNSD (Jaume et al., 2019) provides a small-scale dataset for form understanding in noisy scanned documents, while PubLayNet (Zhong et al., 2019) offers over 360,000 document images for layout analysis. DocLayNet (Pfitzmann et al., 2022) extends this with diverse document types and fine-grained layout categories. VRDU (Wang et al., 2022) specifically addresses visually-rich document understanding with template-based evaluation protocols. CommonForms (Barrow, 2025) represents the largest form field detection benchmark to date, though its template-disjointness has not been previously audited.

Data Leakage in Machine Learning. Data leakage between train and test sets is a well-documented concern in machine learning evaluation. Recht et al. (2019) demonstrated that ImageNet classifiers exhibit significant accuracy drops on new test sets drawn from the same distribution, suggesting potential overfitting to test set characteristics. More recently, benchmark data contamination has emerged as a critical issue for large language models (Xu et al., 2024), where training data may inadvertently include benchmark examples. In document understanding, template

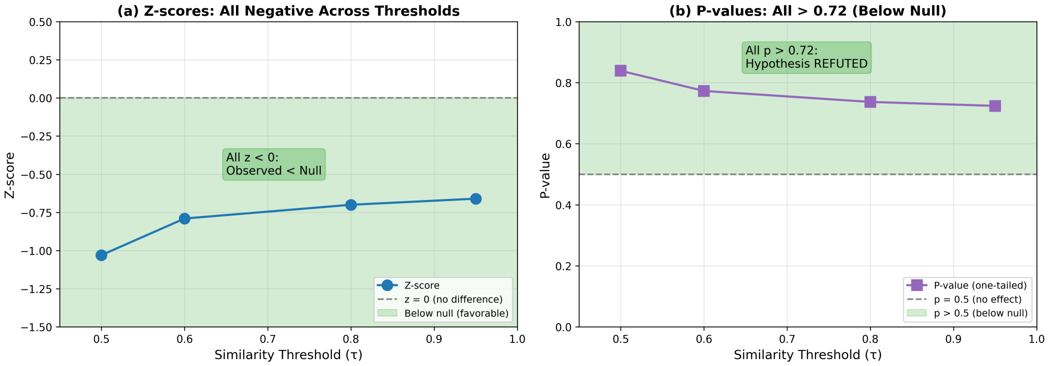

Figure 3: Threshold sensitivity analysis: (a) Z-scores are negative at all four thresholds ($\tau \in \{0.50, 0.60, 0.80, 0.95\}$), indicating observed overlap is consistently below the null mean. (b) P-values exceed 0.72 at all thresholds, confirming the “Refute” conclusion is robust.

overlap poses an analogous risk: models may memorize specific form layouts rather than learning generalizable detection capabilities.

Near-Duplicate Detection. Our audit methodology builds on established techniques for near-duplicate detection. MinHash (Broder, 1997) provides an efficient method for estimating Jaccard similarity between sets, enabling scalable clustering of similar documents. This technique has been successfully applied to web-scale deduplication (Gusev & Xu, 2020) and forms the foundation of our template clustering approach. By representing form pages as sets of quantized field-layout tokens, we can efficiently identify pages sharing similar templates across the entire dataset.

5 Conclusion

We presented TemplateLeak, an audit framework for assessing template-disjointness in document understanding benchmarks. Applying this framework to CommonForms, we find that the template leakage hypothesis is refuted: the observed template overlap (26.8%) is below the null baseline (28.6%), indicating the document-level split produces less overlap than random splitting. This conclusion is robust across all four similarity thresholds tested. Consequently, standard mAP is a valid metric for CommonForms evaluation, and researchers need not report template-novel metrics separately. Future work may apply this audit framework to other document understanding benchmarks where template overlap remains a concern.

References

Joe Barrow. Commonforms: A large, diverse dataset for form field detection, 2025. URL `https://arxiv.org/abs/2509.16506`.

A. Broder. On the resemblance and containment of documents. *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No.97TB100171)*, pp. 21–29, 1997.

Andrey Gusev and Jiajing Xu. *Evolution of a Web-Scale Near Duplicate Image Detection System*. 2020.

Guillaume Jaume, H. K. Ekenel, and J. Thiran. Funsd: A dataset for form understanding in noisy scanned documents. *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, 2:1–6, 2019.

B. Pfitzmann, Christoph Auer, Michele Dolfi, A. Nassar, and P. Staar. *DocLayNet: A Large Human-Annotated Dataset for Document-Layout Segmentation*. 2022.

B. Recht, R. Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? pp. 5389–5400, 2019.

Joseph Redmon, S. Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788, 2015.

Zilong Wang, Yichao Zhou, Wei Wei, Chen-Yu Lee, and Sandeep Tata. Vrdu: A benchmark for visually-rich document understanding. *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022.

Cheng Xu, Shuhao Guan, Derek Greene, and Mohand-Tahar Kechadi. Benchmark data contamination of large language models: A survey. *ArXiv*, abs/2406.04244, 2024.

Xu Zhong, Jianbin Tang, and Antonio Jimeno-Yepes. Publaynet: Largest dataset ever for document layout analysis. *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1015–1022, 2019.