Can You Trust the Confidence? ConfBench for Vision-Language Models on Document Extraction

Anonymous ACL submission

Abstract

Intelligent document processing (IDP) with vision-language models (VLMs) hinges on confidence scores trustworthy enough to route extractions between automation and human review. Existing document benchmarks are dominated by clean, high-quality samples, leaving low accuracy regions too sparse for calibration assessment. We introduce **ConfBench**, the first calibration-specific benchmark for key information extraction (KIE), built by applying 20 controlled degradation pipelines to a diverse document set, yielding 1,346 variants and 70K+ entity-level evaluations spanning the full accuracy spectrum. We evaluate four proprietary and three open-weight VLMs under verbalized and log-probability confidence estimation methods across three input modalities, and find: (i) OCR+Image modality results in more accurate confidence estimates; (ii) verbalized confidence is reliable for stronger models while log-probability estimates perform better for weaker models; (iii) dense models tend to outperform Mixture-of-Experts (MoE) models and model size does not correlate with confidence quality; (iv) calibration ranges from near-perfect to severely overconfident, therefore post-hoc calibration is needed for most but not all models; and (v) log-probability with first-token aggregation consistently outperforms better than mean-token and margin aggregations. We also introduce *ECARB*, a review-budget metric translating discriminative gains into operational savings. The release of ConfBench will enable systematic study of confidence estimators and calibration methods for trustworthy IDP application deployment.

1 Introduction

Intelligent document processing (IDP) systems powered by multimodal foundation models extract structured information from invoices, contracts, and other enterprise documents at scale (Mandvikar, 2023). Their business value depends on trust: organizations need reliable confidence estimates to route high-certainty extractions to automation while flagging uncertain extractions for human review. Without reliable and calibrated confidence scores, organizations either need to review most documents manually that limits automation, or accept unknown error rates that risk compliance violations and financial losses (Natarajan et al., 2025; Guo et al., 2017; Kim et al., 2022).

While confidence calibration is well studied in image classification and NLP (Guo et al., 2017; Joy et al., 2023; Balanya et al., 2024), IDP lacks both suitable benchmarks and systematic evaluation. Existing document benchmarks focus on clean, high-quality documents that yield predominantly high-accuracy predictions, leaving low- and mid-accuracy regions too sparse for calibration assessment (Townsend et al., 2024; Ouyang et al., 2025). Real world documents, however, exhibit substantial quality variation from scanning artifacts and physical degradation that directly impacts both extraction accuracy and confidence estimation. Although two families of confidence estimation have emerged for LLMs, i.e., verbalized confidence, where models self-report numeric scores (Tian et al., 2023; Xiong et al., 2024), and token-level log-probability methods (Ma et al., 2025), their comparative effectiveness for structured document extraction remains unexplored. No systematic benchmark exists to evaluate confidence calibration for key information extraction in IDP.

We address these limitations through three contributions. First, we introduce **ConfBench**, the first benchmark for evaluating confidence calibration in IDP. We apply 20 controlled degradation pipelines to a verified document subset, generating 1,346 document variants spanning the full accuracy spectrum with over 70K entity-level evaluations after quality filtration. Second, we evaluate seven foundation models, four proprietary and three open-weight, under verbalized confidence (1S-TopK) across three

Table 1: Degradation tiers with representative augmentation pipelines. Accuracy drop is averaged across nine runs (only pipelines with $N > 1{,}000$ entities per run included). See **Appendix A.1** for document samples and **Appendix A.3** for accuracy impact analysis details.

Tier	Count	Examples
High (>10%)	2	`Custom22` (21%): dithering + dot-matrix degrade text regions. `Archetype9` (14%): letterpress, fax artifacts, and high geometric distortion.
Moderate (5–10%)	11	Largest tier by design, covering common real-world degradation. `Custom12` (9%): dirty drum and roller streaks. `Custom18` (8%): rotated scan with page border.
Mild (<5%)	7	`Custom15` (4%): moiré pattern on colored paper. `Custom17` (2%): JPEG compression with subtle noise.

input modalities, establishing the first empirical baselines for VLM confidence calibration on KIE. Third, for the two open-weight models exposing token-level probabilities, we compare three log-probability aggregation methods (first-token, mean-token, and margin) against verbalized confidence, revealing fundamental differences in discriminative ability between the two confidence families. We also introduce **ECARB**, a review-budget metric that translates discriminative ability into operational savings for human-in-the-loop deployments. We offer five findings: (1) OCR+Image is uniformly the strongest input modality for both extraction accuracy and confidence quality, and the gap to image-only is largest for smaller, lower-capability models; (2) on the two open-weight models that admit both families, the better choice depends on capability, verbalized confidence wins for the stronger model (Qwen 3.6-27B) while logprob first-token wins for the weaker model (Gemma 3-12B); (3) for frontier proprietary models, where token-level probabilities are not exposed, verbalized confidence tracks extraction accuracy closely, Claude Opus is nearly well-calibrated without any post-hoc adjustment, indicating that verbalized scores are a viable production signal for the strongest closed models; (4) within a family, confidence quality scales monotonically with capability (Claude: Opus > Sonnet > Haiku), but across families parameter count is a poor predictor and MoE models underperform dense ones of comparable active size; (5) calibration ranges from near-perfect (Claude Opus) to severely overconfident (Gemma 3-12B), so post-hoc calibration remains required for most, but not all models. Among logprob aggregations on open-weight models, first-token consistently dominates margin and mean in its discriminative performance.

2 Related Work

Confidence Calibration The alignment between a model's predicted confidence and its true probability of correctness is a critical requirement for accuracy-sensitive applications, where confidence scores gate routing decisions between automation and human review. Guo et al. (2017) demonstrated that modern neural networks are poorly calibrated and tend toward overconfidence, a miscalibration rooted in optimization objectives that prioritize discriminative performance over probabilistic accuracy. Post-hoc methods such as Temperature Scaling (Guo et al., 2017) and adaptive variants (Joy et al., 2023; Balanya et al., 2024) address this for classification tasks, but do not transfer directly to IDP. For black-box LLMs, Tian et al. (2023) showed that verbalized confidence with multiple answer candidates improves calibration over conditional probabilities. Shrivastava et al. (2024) introduced surrogate confidence modeling for black-box predictions, and Li et al. (2024) showed that multi-answer reflection reduces overconfidence. Recent work comparing verbalization and token-level log-probability confidence methods (Xiong et al., 2024; Ni et al., 2025; Ma et al., 2025) finds that neither approach consistently dominates across tasks (Ni et al., 2025), though token probabilities can be more robust to sampling temperature (Xie et al., 2024a). For VLMs, temperature scaling improves calibration (Tu et al., 2024), and semantic perturbation can address verbalized miscalibration (Zhao et al., 2025). Xie et al. (2024b) survey these approaches comprehensively. While promising for general QA, their application to IDP with multimodal inputs, entity-level predictions, and document degradation remains unexplored.

Document Intelligence Benchmarks Modern benchmarks evaluate optical character recognition (OCR), classification, KIE, visual question

Dataset	Purpose	#p	Color depth	Augmentation
NIST (Dimmick et al., 1992)	f_s	5590	Grayscale	✗
MARG (Long et al., 2005)	f_s	1553	RGB	✗
Tobacco-800 (Zhu and Doermann, 2007)	f_s	800	Grayscale	✗
TAB (Mungmeeprued et al., 2022)	f_s	44.8K	Grayscale	✗
Tobacco-3482 (Kumar and Doermann, 2013)	f_p	3482	Grayscale	✗
RVL-CDIP (Harley et al., 2015)	f_p	400K	Grayscale	✗
DocSplit (Islam et al., 2026)	f_{dp}, f_p, f_d	1.55M	Mixed	✗
RealKIE-FCC-Verified (Amazon AGI Team, 2024)	f_{dp}, f_p, f_d, f_k	383	Mixed	✗
ConfBench	f_{dp}, f_p, f_d, f_k, f_c	6.7K	RGB	✓

Table 2: Comparative statistics of document classification datasets. We present dataset characteristics including intended purpose (f_{dp}: document packet splitting, f_s: form segmentation, f_p: form processing, f_d: form detection, f_k: key information extraction, f_c: confidence estimation), corpus size ($\#p$: pages), and image specifications.

answering (VQA), and table extraction. RealKIE (Townsend et al., 2024) provides five enterprise KIE datasets with PDF documents, OCR output, and text span annotations, though annotation quality varies across the five subsets and may require manual verification for use as ground truth. UniKIE-Bench (Ji et al., 2026) extends to constrained and open-category KIE, and OmniDocBench (Ouyang et al., 2025) targets diverse PDF parsing. These benchmarks focus on accuracy metrics over clean, high-quality documents without real-world degradation artifacts, making them unsuitable for confidence calibration evaluation.

3 Benchmark Design

We build upon the RealKIE-FCC-Verified dataset (Amazon AGI Team, 2024) which is a manually corrected and re-annotated subset of the RealKIE benchmark (Townsend et al., 2024). The original dataset comprises 75 real-world FCC-format invoice documents with manually verified ground truth annotations. Each document contains multiple entity types (invoice number, date, and line items consisting of five sub-fields), yielding a rich set of extraction targets for calibration analysis. The original documents are production-quality scans with minimal degradation, producing predominantly high-confidence, high-accuracy predictions from IDP pipelines. We augment this dataset with controlled synthetic degradation simulating real-world artifacts.

3.1 Artifact Augmentation

To populate the full confidence spectrum, we apply controlled synthetic degradation using Augraphy (Groleau et al., 2023), an open-source document-specific augmentation framework that simulates realistic physical and digital artifacts. We generate 20 augmented variants per original document, yielding 1,500 augmented documents initially. After quality filtration, 1346 documents remain. For sample document visualization by degradation severity, please refer to **Appendix A.1**.

Augmentation Pipeline Design: We design 20 augmentation pipelines, 8 pre-configured Augraphy archetypes and 12 custom pipelines targeting specific degradation patterns. Each pipeline applies sequential noise operations across three processing phases: (1) **Ink-phase**: ink-level degradation (bleed-through, mottling, low ink, dithering), (2) **Paper-phase**: substrate degradation (staining, watermarks, colored paper, moiré patterns, bindings), and (3) **Post-phase**: capture and digitization artifacts (JPEG compression, shadows, lighting gradients, scanner defects, geometric distortion). Configurations for the 20 augmentation pipelines are detailed in the **Appendix A.2**.

OCR-Safe Parameter Selection: We constrain augmentation parameters to preserve OCR readability while introducing measurable quality degradation. Key constraints include: (1) rotation limited to $\pm$ 5° as AWS Textract degrades beyond $\pm$10°, (2) JPEG quality 70–95 because quality $<$ 60 causes severe text degradation, (3) brightness 0.7–1.3 as values outside this range impair character recognition, (4) blur radius $\leq$200 pixels with $<$ 8 iterations, and (5) overlay alpha $<$ 0.3 for text-overlapping effects such as shadows and bleed-through. We select pipeline types and noise strengths through an iterative quality assurance process.

Annotation Process and Quality Assessment. We preserve ground truth annotations from the original RealKIE dataset across all augmented variants, enabling direct comparison of extraction accuracy under varying document quality. Quality assessment ensures that varying pipelines and noise strengths produce a well-distributed spec-

trum of accuracy degradation, and it excludes documents where augmentation renders them unreadable with catastrophic OCR failure. Documents where $> 80\%$ of entity types yield null extraction outputs despite having non-null expected values are flagged as catastrophic processing failures. This signature is characteristic of extreme OCR degradation, e.g., `custom22` dithering pipeline, rendering the document illegible for IDP. Including such documents would artificially inflate calibration error without reflecting genuine model uncertainty. We tune noise-levels to keep the unreadable documents to below $< 5\%$ of the dataset while maintaining instances of high degradation.

Document Degradation Spectrum Validation: Each augmentation pipeline applies noise of varying type and strength across ink, paper, and post-processing phases. **Appendix A.2** details the effect and per-phase noise characteristics, including noise strength, for all 20 pipelines. Stronger noise strength and more degradation phases should manifest as higher entity-extraction accuracy drop, the observable we use to categorize pipelines by severity. To quantify this, we identify the matched subset of documents present in both the original and augmented sets across all nine runs, then compute the per-augmentation accuracy drop. **Appendix A.3** presents the accuracy impact by augmentation type; Table 1 summarizes the three resulting tiers with representative examples.

3.2 Dataset Statistics

Table 2 summarizes the final benchmark composition after generation and quality filtration, and provides a comparison against common existing benchmarks. Our benchmark, **ConfBench**, comprises 1,346 documents. Entity counts range from 70K to 130K depending on the run, as extraction models differ in output efficiency under fixed token budgets and the fraction of null confidence scores varies by estimation method.

Figure 1 confirms the augmentation achieves its intended effect. We compute each document's mean accuracy by averaging its entity-level accuracies, proxying the degree of degradation introduced. The original 75 documents cluster above 0.6, leaving lower accuracy region too sparse for reliable calibration. The augmentation provides statistically sufficient samples across the full spectrum.

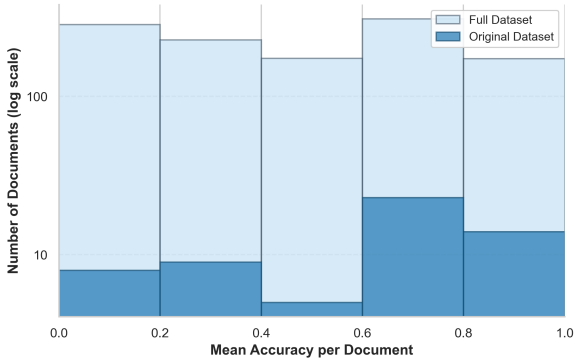


Figure 1: Distribution of per-document mean entity-extraction accuracy (mean over all entity-level binary accuracies per document). (a) Original dataset with 75 documents concentrate above 0.6, leaving low-accuracy region sparsely populated. (b) After augmentation, degraded variants populate the 0.0–0.6 range, enabling calibration assessment across the full accuracy range.

4 Calibration of VLM Models on the Benchmark Dataset

We design a systematic evaluation to study how well VLMs produce calibrated confidence measures for KIE. Our experimental design spans two axes: (1) two confidence estimation approaches (verbalized and logprob), and (2) three input modality configurations, yielding six distinct experimental conditions. We evaluate seven foundation models for the verbalized approach and two open-weight models for the logprob approach, enabling controlled comparison of calibration behavior across model families, estimation methods, and input modalities. **Appendix A.4** details the full workflow from optical character recognition (OCR) through classification, information extraction, confidence estimation, evaluation, and calibration assessment.

4.1 Experimental Setup

We study two complementary approaches to confidence estimation for document entity extraction.

Verbalized confidence elicits self-assessed scores directly from the model via prompt engineering. This approach is model-agnostic: it works with any instruction-following VLM and requires no access to internal model states.

Logprob confidence derives scores from token-level log-probabilities produced during greedy decoding. Because it requires access to the raw probability distribution over the vocabulary, this approach is restricted to open-weight models. It requires no prompt engineering and is complemen-

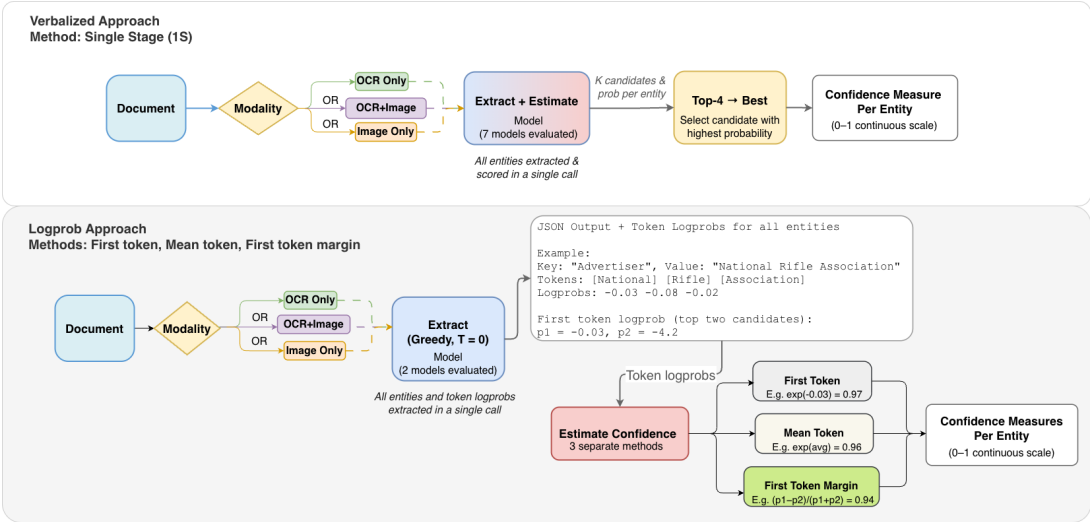


Figure 2: Experimental design overview. **Verbalized:** A single model jointly extracts entities and estimates confidence in one call, producing top candidate guesses with continuous probabilities (0 to 1) per entity; the highest-probability candidate is selected. **Logprob:** A single-stage extraction pass produces structured output via greedy decoding, and confidence is derived from the token-level log-probabilities of each entity.

tary to verbalized confidence: where verbalized scores depend on the model's ability to introspect, logprob scores capture the model's distributional uncertainty directly.

Verbalized method: 1S-TopK. We employ a single-stage design in which extraction and confidence estimation occur in one LLM call (Figure 2, top). Following the TopK Verbalized approach (Tian et al., 2023), the model emits its top four candidate guesses with continuous probabilities per entity, and we select the highest-probability candidate. Following Tian et al. (2023), we prompt for guesses and probabilities only and omit chain-of-thought reasoning. We evaluate seven VLMs spanning multiple capability tiers and providers: Opus 4.6, Sonnet 4.5, and Haiku 4.5; Kimi K2.5; Qwen 3 VL-235B and Qwen 3.6-27B; and Gemma 3-12B.

Logprob methods. Extraction and confidence are obtained from a single greedy decoding pass (T=0): the model emits the JSON output, and we derive confidence from the logprobs of the tokens covering each entity value (Figure 2, bottom). Let $\{\ell_1, \ldots, \ell_K\}$ be the token logprobs spanning a value and $p_1 \geq p_2$ the top-2 probabilities at the first position; we compare three aggregations: *first-token* $\exp(\ell_1)$, *mean token* $\exp(\frac{1}{K}\sum_k \ell_k)$, and *first-token margin* $(p_1 - p_2)/(p_1 + p_2)$. Further details are in Appendix A.5. We evaluate two open-weight VLMs for this approach: Qwen 3.6-27B and Gemma 3-12B.

Input modality configurations: For each strategy, we evaluate three input configurations: (1) **OCR Only**, providing the OCR text extracted by Amazon Textract along with its per-token confidence scores; (2) **OCR + Image**, providing both OCR text (with Textract confidence scores) and the document image; and (3) **Image Only**, providing only the document image. This yields the factorial design: {Verbalized, Logprob} × {OCR Only, OCR+Image, Image Only}.

4.2 Evaluation metrics

We report extraction quality via **Weighted Overall Accuracy (WOA)**, the weighted average of per-field similarity scores across all entity types. For string fields, similarity is the normalized Levenshtein similarity between predicted and ground-truth values; for numeric fields a tolerance-based comparator is used. Each field receives a continuous score between 0 (completely wrong) and 1 (exact match), and WOA is the mean of these scores across all evaluated fields. In this analysis, all entity fields are assigned equal weight. For confidence calibration, we employ four complementary metrics. (1) **AUROC** measures discriminative ability: the area under the ROC curve for separating correct from incorrect predictions (Shrivastava et al., 2024; Xiong et al., 2024), with random at 0.5 and perfect at 1.0. AUROC is scale-invariant. (2) **ECE** measures calibration error: the weighted absolute gap between predicted confidence and observed accu-

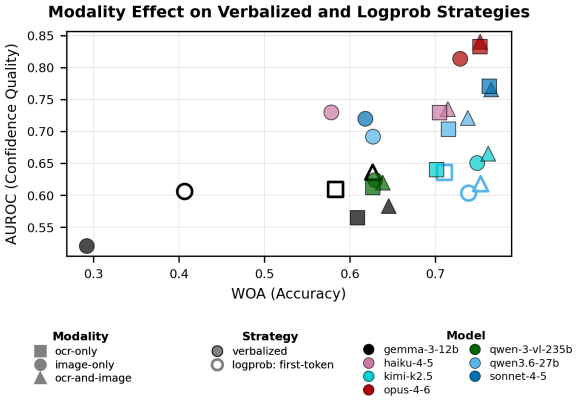


Figure 3: Confidence quality (AUROC) vs. extraction accuracy (WOA) across all three input modalities for both verbalized and logprob first-token strategies. OCR+Image consistently occupies the upper-right region, with OCR-only ranking second and image-only third across all models.

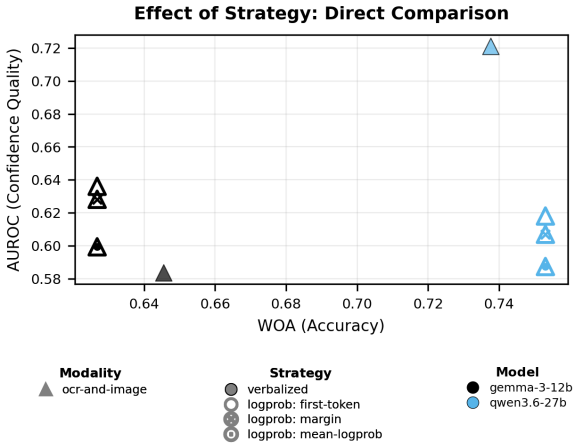


Figure 4: Effect of confidence strategy on AUROC (OCR+Image), restricted to models with all four strategies available. Verbalized outperforms all logprob methods for Qwen 3.6-27B; the pattern reverses for Gemma-3-12B. Among logprob variants, first-token consistently achieves the highest AUROC.

racy, $\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \left|\overline{\text{acc}}(B_m) - \overline{\text{conf}}(B_m)\right|$, computed with adaptive (quantile-based) binning at $M{=}5$ following Shrivastava et al. (2024); ECE does not capture discriminative ability. (3) **Brier Score**, $\text{BS} = \frac{1}{N}\sum_{i=1}^{N}(\text{conf}_i - \text{acc}_i)^2$ (Brier, 1950), penalizes confident incorrect predictions more heavily than confident correct ones and complements AUROC and ECE without binning artifacts. (4) **Error Capture at Review Budget (ECARB).** A major concern for information extraction deployments is the human effort required to catch errors and raise performance to a target quality threshold. We introduce ECARB to measure the lift of confidence-guided review over random sampling. Let predictions sorted by ascending confidence, and let E be the total error count. For a review budget $b \in (0, 1]$, let $k = \min(n, \max(1, \lfloor bn \rfloor))$ be the number of predictions reviewed; reviewing k random predictions catches $(k/n) \cdot E$ errors in expectation:

$$\text{ECARB@}b \;=\; \frac{\sum_{i=1}^{k} \mathbf{1}[\text{acc}_i = 0]}{(k/n) \cdot E} \qquad (1)$$

A value of 1 indicates no lift over random; higher ECARB@b values identify models whose confidence scores deliver the greatest HITL utility.

5 Results and Discussion

The following analysis isolates three factors that influence extraction accuracy and confidence measure: input modality, confidence strategy, and model capability. Across all scatter plots, we use a consistent visual encoding: color denotes model, shape denotes modality (square = OCR, circle = Image, triangle = OCR+Image), and fill style denotes strategy (filled = verbalized, hollow = logprob first-token, $\odot$ = mean-logprob, $\otimes$ = margin).

5.1 Input Modality

Textual OCR text with OCR confidence scores provides a stronger confidence signal than visual input alone, and combining both modalities is uniformly best (Figure 3). The gap between OCR+Image and image-only ranges from marginal in high-capability models to substantial in weaker ones—Gemma 3-12B drops 6 AUROC points while Opus 4.6 drops only 3. OCR-only occupies a middle position: stronger than image-only but weaker than the combined input. This pattern is independent of confidence strategy, motivating OCR+Image as the standard modality for subsequent analysis. The practical impact is reflected in ECARB (Table 3): on OCR+Image, Opus 4.6 captures 2.43× more errors than random at a 30% review budget, the highest gain observed across all configurations.

5.2 Confidence Strategy

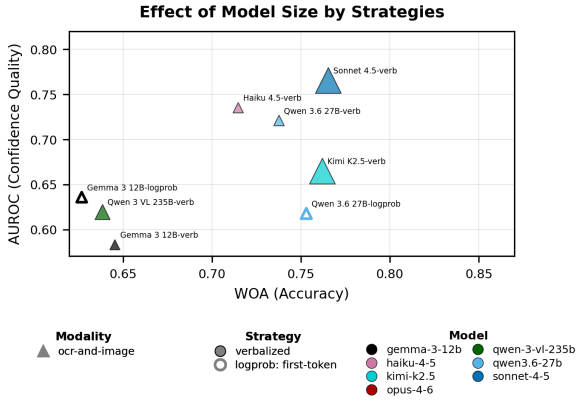
Holding modality constant at OCR+Image, confidence strategy has a substantial effect on AUROC, though its direction depends on model capability (Figure 4). For Qwen 3.6-27B, verbalized confidence yields 0.72 AUROC compared to 0.62 for logprob first-token, a 10-point advantage. The pattern reverses for Gemma 3-12B, where logprob

first-token (0.64) outperforms verbalized (0.58), suggesting that weaker models cannot reliably self-assess but still produce informative logit distributions. Extraction accuracy (WOA) is unaffected by strategy choice, confirming that the observed differences reflect confidence quality rather than task performance. Among logprob variants, first-token consistently outperforms margin and mean aggregation for both models, making it the preferred logprob method. These results indicate that verbalized confidence is the stronger strategy for capable models, while logprob first-token provides a robust alternative that does not require prompt engineering. In terms of HITL utility (Table 3), verbalized Qwen 3.6-27B achieves ECARB=1.93 compared to 1.42 for logprob first-token, confirming that the AUROC advantage translates directly into more errors surfaced per review dollar.

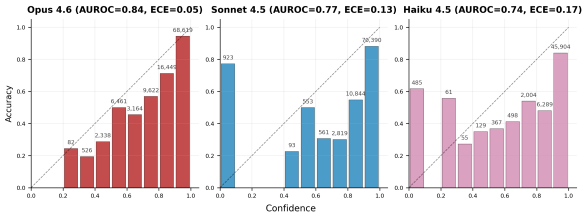
5.3 Model Scale and Family

Holding modality to OCR+Image, model capability has the largest effect of the three factors examined (Figure 5). Within the Claude family, confidence quality scales monotonically with model size: Opus 4.6 achieves the highest AUROC at 0.84, followed by Sonnet 4.5 (0.77) and Haiku 4.5 (0.74). Across families, parameter count is a poor predictor of confidence quality: Qwen 3.6-27B (0.72) surpasses the much larger Qwen 3 VL-235B (0.62) despite having fewer than one-eighth of its parameters, indicating that architecture generation and training methodology are stronger determinants than raw scale. Notably, both MoE models in the evaluation (Kimi K2.5 and Qwen 3 VL-235B) underperform dense models of comparable active parameter count, suggesting that the sparse activation characteristic of MoE architectures may be a limiting factor for confidence calibration. The ECARB metric mirrors this hierarchy: Opus 4.6 achieves 2.43× gain, Sonnet 4.5 reaches 2.24×, and Haiku 4.5 delivers 1.96×, while the weaker models fall below 1.7×, underscoring that model capability directly determines the practical value of confidence-guided review.

The reliability diagrams (Figure 5b) confirm that the AUROC differences reflect genuine calibration quality. Opus 4.6 closely tracks the y=x diagonal across all confidence bins, achieving near-perfect calibration without post-hoc adjustment. Sonnet 4.5 shows mild overconfidence in the 0.6–0.8 range, while Haiku 4.5 exhibits more pronounced deviations. This suggests that larger Claude models



(a) AUROC vs. WOA by model. Marker size is proportional to parameter count. Within the Claude family, confidence quality scales monotonically with size.



(b) Reliability diagrams for the Claude family. Opus 4.6 (left) closely tracks the y=x diagonal (near-perfect calibration). Sonnet 4.5 (center) and Haiku 4.5 (right) show progressively larger overconfident deviations. Bins with <30 samples omitted.

Figure 5: Effect of model size and family on confidence quality (OCR+Image). Calibration degrades gracefully across Claude tiers: opus achieves ECE=0.05 without any post-hoc calibration, while haiku reaches ECE=0.17.

develop stronger introspective ability during training. Full reliability diagrams for all models, modalities, and strategies are provided in Appendix A.6.

5.4 Metrics and Their Implications

Table 3 reveals that different metrics capture distinct aspects of confidence quality, and no single model dominates across all dimensions. AUROC and ECE are complementary: Opus 4.6 achieves the best scores on both (0.84 and 0.05, respectively), reflecting both strong discrimination and well-calibrated confidence. In contrast, Sonnet 4.5 ranks second in AUROC (0.77) but has a higher ECE (0.13), indicating that its confidence scores correctly discriminate between correct versus incorrect predictions, but are not well-calibrated in absolute terms. Task accuracy (WOA) and confidence quality are largely orthogonal: Sonnet 4.5 achieves the highest extraction accuracy (0.77 WOA) while Opus 4.6 leads on all confidence metrics, and Kimi K2.5 achieves competitive accuracy (0.76 WOA) while lagging in AUROC (0.67). Logprob-

Approach	Model	Acc.	F1	AUROC↑	ECE↓	Brier↓	ECARB↑
verbalized	Opus 4.6	0.75	0.42	**0.84**	**0.05**	**0.12**	**2.43**
	Sonnet 4.5	**0.77**	0.51	0.77	0.13	0.16	2.24
	Haiku 4.5	0.71	0.38	0.74	0.17	0.19	1.96
	Qwen 3.6-27B	0.74	0.41	0.72	0.22	0.23	1.93
	Kimi K2.5	0.76	0.46	0.67	0.20	0.22	1.68
	Qwen 3 VL-235B	0.64	0.41	0.62	0.25	0.26	1.53
	Gemma 3-12B	0.65	0.28	0.58	0.31	0.31	1.39
logprob (first-token)	Gemma 3-12B	0.63	0.38	0.64	0.36	0.36	1.45
	Qwen 3.6-27B	0.75	**0.70**	0.62	0.21	0.21	1.41
logprob (margin)	Gemma 3-12B	0.63	0.38	0.63	0.35	0.35	1.43
	Qwen 3.6-27B	0.75	**0.70**	0.61	0.21	0.21	1.40
logprob (mean-token)	Gemma 3-12B	0.63	0.38	0.60	0.36	0.36	1.30
	Qwen 3.6-27B	0.75	**0.70**	0.59	0.21	0.21	1.36

Table 3: Extraction accuracy and confidence calibration metrics on the **OCR+Image** modality, selected as it consistently yields the strongest joint performance in both WOA and AUROC across approaches. In this setting, the model receives both the raw document image and the OCR text extracted by Amazon Textract along with its per-token confidence scores. ECARB reports the error-capture gain at 30% review budget.

based approaches show systematically higher ECE than verbalized methods, despite comparable AUROC in some cases, suggesting that logit distributions are informative for discrimination between correct versus incorrect extractions, but require calibration before use in absolute confidence thresholds. Finally, ECARB provides the most practically relevant signal: Opus 4.6 captures 2.43 times as many errors as a random reviewer at a 30% review budget, while even the weakest logprob approach (mean) yields a 1.30× gain.

Practitioner guidelines: AUROC and ECARB are strongly correlated in our results: better discrimination (AUROC) translates directly into more errors surfaced per unit of review effort (ECARB). For HITL workflows where predictions are reviewed in confidence order, AUROC is the primary selection criterion, with WOA as a secondary factor (higher WOA means fewer total errors to find). ECE becomes relevant only when absolute confidence thresholds are needed, e.g., auto-accepting predictions above 0.9. We recommend selecting models by AUROC first, applying post-hoc calibration if threshold-based routing is required, and using ECARB at the deployment-specific review budget to estimate operational savings.

6 Conclusion and Future Work

This work presents ConfBench and a systematic evaluation of VLM confidence calibration for document KIE. Our results demonstrate that confidence quality is primarily driven by model capability and input modality, with verbalized and logprob approaches serving complementary roles depending on model strength. The proposed ECARB metric bridges the gap between statistical calibration measures and deployment economics: at a 30% review budget, the best configuration surfaces 2.4× more errors than random sampling, while even the weakest yields 1.3×, confirming that confidence-guided review delivers tangible operational value across all evaluated settings.

Limitations and future work. ConfBench currently covers a single document domain (FCC invoices), and expanding to diverse document types, languages, and degradation sources is a priority. Promising directions include combining verbalized and logprob signals into ensemble confidence scores, applying lightweight post-hoc calibration to close the gap for overconfident models, and investigating how calibration quality transfers across document domains without retraining.

Potential Risks: Confidence-guided routing complements rather than replaces human oversight, since calibration established on one distribution may shift under deployment conditions. We view ongoing distribution monitoring and periodic recalibration as productive directions for sustaining trust as ConfBench-informed systems are extended to new domains.

References

Amazon AGI Team. 2024. Realkie-fcc-verified: A verified test set for enterprise key information extraction. https://huggingface.co/datasets/amazon-agi/RealKIE-FCC-Verified. Manually

annotated and corrected subset of the RealKIE benchmark.

Sergio A Balanya, Juan Maronas, and Daniel Ramos. 2024. Adaptive temperature scaling for robust calibration of deep neural networks. *Neural Computing and Applications*, 36(14):8073–8095.

Glenn W. Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3.

DL Dimmick, MD Garris, and CL Wilson. 1992. Nist special database 6: Structured forms database 2. Technical report, National Institute of Standards and Technology.

Alexander Groleau, Kok Wei Chee, Stefan Larson, Samay Maini, and Jonathan Boarman. 2023. Augraphy: A data augmentation library for document images. In *Document Analysis and Recognition – ICDAR 2023*, Lecture Notes in Computer Science, pages 384–401. Springer.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Adam W Harley, Alex Ufkes, and Konstantinos G Derpanis. 2015. Evaluation of deep convolutional nets for document image classification and retrieval. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 991–995. IEEE.

Md Mofijul Islam, Md Sirajus Salekin, Nivedha Balakrishnan, Vincil C Bishop III, Niharika Jain, Spencer Romo, Bob Strahan, Boyi Xie, and Diego A Socolinsky. 2026. Docsplit: A comprehensive benchmark dataset and evaluation approach for document packet recognition and splitting. *Computing Research Repository*, arXiv:2602.15958.

Yifan Ji, Zhipeng Xu, Zhenghao Liu, Zulong Chen, Qian Zhang, Zhibo Yang, Junyang Lin, Yu Gu, Ge Yu, and Maosong Sun. 2026. Unikie-bench: Benchmarking large multimodal models for key information extraction in visual documents. *Computing Research Repository*, arXiv:2602.07038.

Tom Joy, Francesco Pinto, Ser-Nam Lim, Philip HS Torr, and Puneet K Dokania. 2023. Sample-dependent adaptive temperature scaling for improved calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14919–14926.

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 *Computer Vision – ECCV 2022*, Lecture Notes in Computer Science, pages 498–517. Springer.

Jayant Kumar and David Doermann. 2013. Unsupervised classification of structurally similar document images. In *2013 12th International Conference on Document Analysis and Recognition*, pages 1225–1229. IEEE.

Moxin Li, Wenjie Wang, Fuli Feng, Fengbin Zhu, Qifan Wang, and Tat-Seng Chua. 2024. Think twice before trusting: Self-detection for large language models through comprehensive answer reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11858–11875. Association for Computational Linguistics.

L Rodney Long, Sameer K Antani, and George R Thoma. 2005. Image informatics at a national research center. *Computerized Medical Imaging and Graphics*, 29(2-3):171–193.

Huan Ma, Jingdong Chen, Joey Tianyi Zhou, Guangyu Wang, and Changqing Zhang. 2025. Estimating LLM uncertainty with evidence. *Computing Research Repository*, arXiv:2502.00290.

Shreekant Mandvikar. 2023. Augmenting intelligent document processing (idp) workflows with contemporary large language models (llms). *International Journal of Computer Trends and Technology*, 71(10):80–91.

Thisanaporn Mungmeeprued, Yuxin Ma, Nisarg Mehta, and Aldo Lipani. 2022. Tab this folder of documents: page stream segmentation of business documents. In *Proceedings of the 22nd ACM Symposium on Document Engineering*, pages 1–10.

Sriraam Natarajan, Saurabh Mathur, Sahil Sidheekh, Wolfgang Stammer, and Kristian Kersting. 2025. Human-in-the-loop or ai-in-the-loop? automate or collaborate? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 28594–28600.

Shiyu Ni, Keping Bi, Lulu Yu, and Jiafeng Guo. 2025. Are large language models more honest in their probabilistic or verbalized confidence? In *Information Retrieval: 30th China Conference (CCIR 2024)*, Lecture Notes in Computer Science. Springer.

Linke Ouyang, Yuan Qu, Hongbin Zhou, Jiawei Zhu, Rui Zhang, Qunshu Lin, Bin Wang, Zhiyuan Zhao, Man Jiang, Xiaomeng Zhao, and 1 others. 2025. Omnidocbench: Benchmarking diverse pdf document parsing with comprehensive annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24838–24848.

Vaishnavi Shrivastava, Percy Liang, and Ananya Kumar. 2024. Llamas know what GPTs don't show: Surrogate models for selective classification.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence

scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442. Association for Computational Linguistics.

Benjamin Townsend, Madison May, Katherine Mackowiak, and Christopher Wells. 2024. Realkie: Five novel datasets for enterprise key information extraction. *Computing Research Repository*, arXiv:2403.20101.

Weijie Tu, Weijian Deng, Dylan Campbell, Stephen Gould, and Tom Gedeon. 2024. An empirical study into what matters for calibrating vision-language models. *Computing Research Repository*, arXiv:2402.07417.

Johnathan Xie, Annie S. Chen, Yoonho Lee, Eric Mitchell, and Chelsea Finn. 2024a. Calibrating language models with adaptive temperature scaling. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Liangru Xie, Hui Liu, Jingying Zeng, Xianfeng Tang, Yan Han, Chen Luo, Jing Huang, Zhen Li, Suhang Wang, and Qi He. 2024b. A survey of calibration process for black-box llms. *Computing Research Repository*, arXiv:2412.12767.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Yunpu Zhao, Rui Zhang, Junbin Xiao, Ruibo Hou, Jiaming Guo, Zihao Zhang, Yifan Hao, and Yunji Chen. 2025. Object-level verbalized confidence calibration in vision-language models via semantic perturbation. *Computing Research Repository*, arXiv:2504.14848.

Guangyu Zhu and David Doermann. 2007. Automatic document logo detection. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pages 864–868. IEEE.

A Appendix

A.1 Augmentation Quality Examples

Figure 6 illustrates a single source document rendered under three representative pipelines, one from each severity tier defined in Table 1. The same document is used across all four panels so that visual differences are attributable to the augmentation pipeline rather than to underlying document content. Ground truth annotations are preserved across variants (§3.1), making each panel directly comparable in terms of extraction difficulty.

Mild tier (<5% accuracy drop). Panel (b) shows `Custom15`, which applies a moiré pattern over colored paper in the paper phase only (Table 4). The text remains crisp and character strokes are preserved; the dominant artifact is a low-frequency tonal modulation of the background. OCR confidence on body text drops only marginally and extraction accuracy degrades by less than 5%, but the colored cast introduces a measurable shift in document-image appearance that is visible to image-only confidence estimators.

Moderate tier (5 to 10% accuracy drop). Panel (c) shows `Custom12`, a post-phase pipeline combining dirty-drum streaks with roller marks. Vertical and horizontal banding overlay the text region, locally occluding glyphs without rendering them globally illegible. Extraction errors concentrate on entities whose bounding boxes intersect heavy banding (e.g., line-item digits, date separators). This is the largest tier by design: it covers the range of degradation most commonly observed in production scans, where individual artifacts are visible but the document remains machine-readable.

High tier (>10% accuracy drop). Panel (d) shows `Custom22`, which combines ink-phase dithering with a post-phase dot-matrix simulation. Character interiors are broken into dot patterns and inter-character spacing becomes irregular, pushing OCR toward the boundary of the readability envelope defined by our parameter constraints (§3.1): rotation $\leq \pm 5$, JPEG quality ≥ 70, blur radius ≤ 200 px, and overlay alpha < 0.3. Documents whose augmentation pushes more than 80% of entity types to null extraction under these constraints are removed by the unreadability filter (§3.1), so the surviving high-tier documents represent severe but not catastrophic degradation. This regime is where calibration matters most: extraction accuracy is no longer near-perfect, and confidence scores must reliably separate the remaining correct extractions from the now-frequent incorrect ones.

Quality control. Across all three tiers, two properties are preserved by construction. First, ground truth values remain unchanged, so accuracy degradation reflects model robustness rather than annotation drift. Second, the OCR-safe parameter envelope (§3.1) keeps the share of documents removed by the unreadability filter below 5% of the dataset, so the visible degradation in panels (b)–(d) is representative of what enters ConfBench rather than an extreme tail. Per-pipeline configurations underlying these examples are listed in Table 4.

A.2 Augmentation Pipeline Characteristics

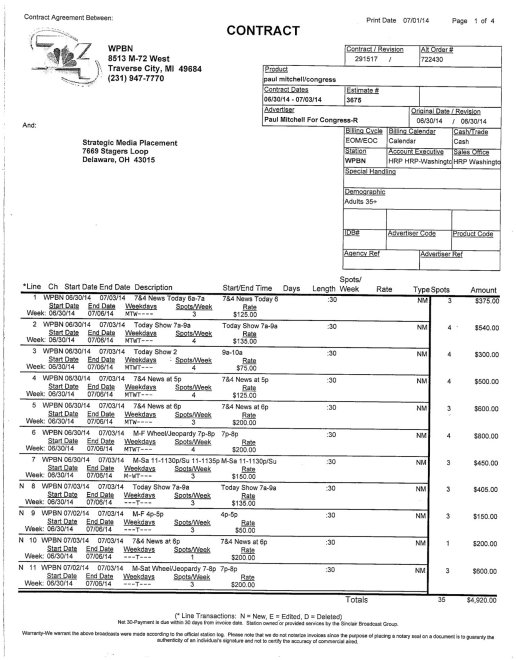
We employ the Augraphy library to generate realistic document degradation patterns across 20 distinct noise pipelines. Each pipeline is structured as a three-phase transformation process that mimics the physical lifecycle of document degradation: *Ink Phase*, *Paper Phase* and *Post Phase*. We utilize eight pre-defined archetype pipelines from the Augraphy library alongside twelve custom-designed pipelines to achieve diverse noise characteristics. Each pipeline is assigned a noise intensity level (High, Medium, or Low) per phase based on the severity and number of augmentations applied, enabling systematic evaluation across varying degradation conditions. This multi-phase approach ensures that our synthetic noisy documents closely replicate real-world scenarios where documents undergo multiple forms of degradation throughout their lifecycle. Table 4 shows which real-world effect does each augmentation pipeline simulate, and the details of the Ink, Paper and Post phases.

A.3 Accuracy Drop by Augmentation Type

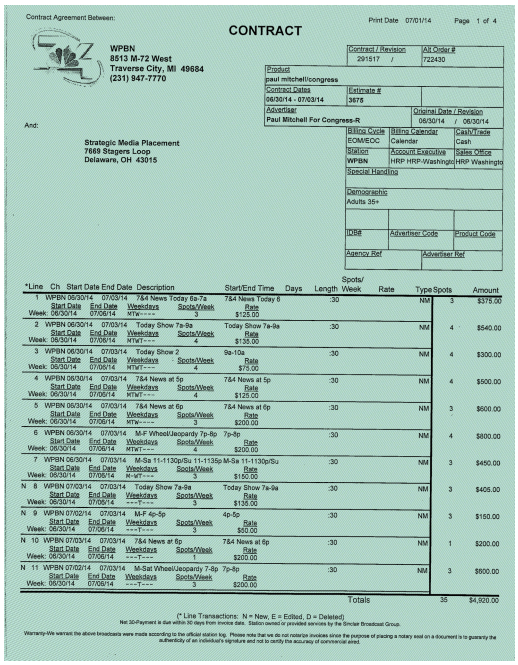
For each augmentation type, we compute the accuracy drop as the difference in mean entity-level extraction accuracy between the original documents and their augmented counterparts, using only the matched subset of documents present in both sets. A higher accuracy drop indicates that the augmentation more severely degrades extraction accuracy. The full per-pipeline accuracy-drop distribution is shown in Figure 7.

A.4 Document Processing Pipeline

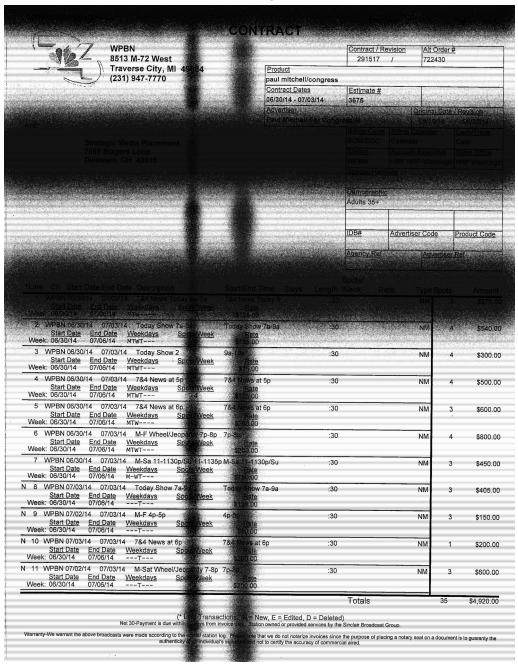
Optical Character Recognition (OCR). We use Amazon Textract with the LAYOUT feature to extract text and structural elements from each document image. Textract returns a word-level output: each token carries a detected string, an axis-aligned



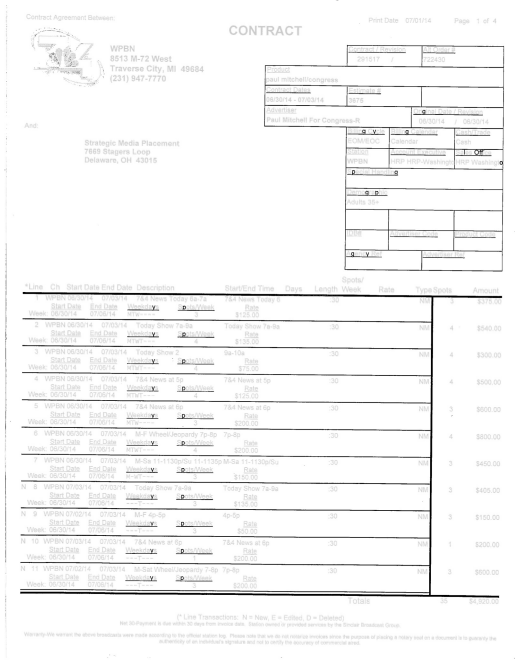
(a) Original



(b) `Custom15` — mild degradation



(c) `Custom12` — moderate degradation



(d) `Custom22` — high degradation

Figure 6: Document `1220d048d2a75d910c52e7f290e32ca0` from our ConfBench dataset under three degradation levels. (a) Original clean scan. (b) `Custom15`: mild degradation (<5% accuracy drop) simulates Moiré pattern on colored paper. (c) `Custom12`: moderate degradation (6–10% accuracy drop) simulates dirty printer drum and rollers. (d) `Custom22`: high degradation (21% accuracy drop) represents dithered print with dot-matrix simulation.

bounding box, a reading-order index, and a per-token OCR confidence score in $[0, 100]$. When the OCR modality is active, we serialize this output as a structured text block of `(word, confidence)` pairs, preserving reading order, and pass it in the user message. For the *OCR-Only* modality the document image is withheld entirely; for *OCR+Image* the structured text block and the document image are provided together; for *Image-Only* the OCR output is withheld and only the image is provided.

Document Classification. The document type is predicted by a VLM, invoked as a separate prior call using both the OCR text and the document image. The predicted class conditions downstream

Table 4: Augmentation pipeline configurations. The *Real-World Effect* column describes the physical scenario each pipeline simulates. Phase columns list noise components with cell color and label indicating severity: **High**, **Medium**, **Low**.

Pipeline	Real-World Effect	Ink Phase	Paper Phase	Post Phase
Default	General multi-artifact degradation	**M:** InkColorSwap, Dithering, InkBleed, Letterpress	**M:** PaperFactory, ColorPaper, WaterMark	**M:** DirtyDrum, Jpeg, BadPhotoCopy, Geometric, Faxify
Archetype2	Marked document with punch holes, photocopier noise and reflected glare	**L:** Markup, InkBleed	**M:** Bindings, PageBorder	**H:** BadPhotoCopy, Geometric, Scribbles, ReflectedLight
Archetype3	Heavy ink bleed with lighting gradients	**M:** InkBleed	**L:** BadPhotoCopy	**M:** LightingGradient, BleedThrough
Archetype4	Cropped fax with punch holes and monochrome conversion	**H:** Geometric, Bindings, InkBleed	**M:** Geometric	**L:** Faxify, BadPhotoCopy
Archetype7	Minimal fax with moderate ink bleed and background noise	**L:** Geometric	**H:** BadPhotoCopy	**M:** Faxify, InkBleed, BadPhotoCopy
Archetype9	Letterpress with fax processing and noisy lines	**M:** Letterpress, Faxify, Dithering	**M:** NoisyLines, BadPhotoCopy	**H:** Geometric
Archetype10	Faded letterpress with minimal photocopier noise	**H:** BadPhotoCopy, Letterpress	—	**M:** Faxify
Archetype11	Aged paper with color shift, shadows, and bleed-through	**L:** LinesDegradation, InkBleed, LowInkLines	**M:** PatternGenerator, ColorPaper	**H:** ColorShift, Jpeg, ShadowCast, BleedThrough
Custom12	Dirty printer drum and rollers	—	—	**H:** DirtyDrum, DirtyRollers
Custom13	Coffee-stained and folded document	—	**M:** Stains	**M:** Folding
Custom14	Heavy ink bleed-through and mottling	**H:** BleedThrough, InkMottling	—	—
Custom15	Moiré pattern on colored paper	—	**M:** Moire, ColorPaper	—
Custom16	Shadow and uneven lighting	—	—	**M:** ShadowCast, LightingGradient
Custom17	Heavy JPEG compression with noise	—	—	**H:** Jpeg, SubtleNoise
Custom18	Rotated scan	**M:** Geometric	—	**L:** PageBorder
Custom19	Binding marks with letterpress imprint	—	**M:** Bindings	**H:** Letterpress
Custom20	Extreme brightness with bad photocopy	**H:** Brightness	**H:** BadPhotoCopy	—
Custom21	Watermarked document with noisy lines	—	**M:** WaterMark	**M:** NoisyLines
Custom22	Dithered print with dot-matrix simulation	**M:** Dithering	—	**M:** DotMatrix
Custom23	Aged paper stock	—	**M:** PaperFactory	**H:** Faxify

extraction by determining which entity schema to apply; for example, the same field name (e.g., *date*) can refer to semantically different entities across document types in a multi-class packet. Because CONFBENCH contains a single document class (FCC invoice), this step always returns the correct class and classification accuracy does not vary across experimental conditions. We include it to faithfully replicate the production IDP architecture, where classification errors in multi-class

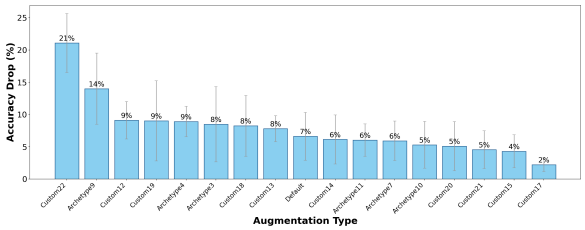


Figure 7: Average accuracy drop in information extraction per augmentation type, averaged across nine runs. All runs use Sonnet 4 as the entity extractor. Error bars show standard deviation across runs. 17 of 20 augmentation types have $N > 1{,}000$ entities per run and mean accuracy drop exceeding the cross-run standard deviation. We exclude the 3 augmentation types where run-to-run variability obscures a consistent degradation signal. Types are sorted by decreasing severity.

packets would propagate to extraction by applying the wrong schema.

Entity Extraction and Confidence Estimation. We extract entity values according to a predefined schema of target types: six document-level fields (`Agency`, `Advertiser`, `GrossTotal`, `PaymentTerms`, `AgencyCommission`, `NetAmountDue`) and a variable-length `LineItems` array, each entry containing five sub-fields (`StartDate`, `EndDate`, `Days`, `Description`, `Rate`).

How extraction and confidence estimation are coupled depends on the strategy. For **verbalized Approach** (1S-TopK method), both operations are performed jointly in a single LLM call: the model emits its top-four candidate values with associated probabilities for every entity simultaneously, and the highest-probability candidate is selected as the extraction output. For **1S-Logprob**, extraction is performed in a single greedy-decoding pass that produces a structured JSON answer; confidence scores are then derived post-hoc from the token-level log-probabilities of the tokens spanning each entity value, without any additional model call (see Appendix A.5 for aggregation details).

Evaluation. We score extracted values using STICKLER,[1] an open-source structured-extraction evaluation library that requires field-level ground truth annotations. Each field uses a comparator suited to its type: string fields by character-level similarity, numeric fields by proximity within a 1%tolerance, and date fields by semantic calendar equivalence so that surface differences in formatting count as correct. `LineItems` are matched to predictions by maximising total similarity across all line items before per-field scoring.

Each entity receives a continuous similarity score in $[0, 1]$ reflecting how closely the predicted value matches ground truth. These scores are averaged across all entities to produce WOA, our headline extraction accuracy metric. For calibration, each score is converted to a binary correctness label using a per-field acceptance threshold, and paired with the entity's confidence score. STICKLER computes all reported metrics directly from these pairs: extraction quality (WOA, F1, precision, recall) and confidence calibration (AUROC, ECE, Brier Score, and ECARB at multiple review budgets).

A.5 Confidence Estimation and Inference Details

Verbalized 1S-TopK protocol. Following Tian et al. (2023), we prompt the model to produce its top-four candidate values, each with a continuous probability in $[0, 1]$, for every target entity in a single JSON call. The candidate with the highest probability is selected as the extraction output; its associated probability is retained as the confidence score and later used to assess calibration. The rationale is that forcing the model to consider alternatives reduces overconfidence compared to a single top-1 prediction; we omit chain-of-thought reasoning, as Tian et al. (2023) found it does not improve calibration further. Because extraction and confidence estimation are coupled in the same call, entity-level accuracy varies across models, unlike logprob methods where extraction is always decoupled from confidence scoring.

Aggregation methods for 1S-Logprob. Let $\{\ell_1, \ldots, \ell_K\}$ be the token log-probabilities of the K generated tokens covering an entity value. Separately, let $p_1 \geq p_2$ denote the top-2 probabilities over the *full vocabulary* at the first token position, that is, the probability of the chosen first token and the probability of the next-best alternative at that same position, not the log-probs of the first and second tokens in the sequence. We compare three ways of summarizing these into a single confidence score.

- **Mean-token** $\exp\left(\frac{1}{K}\sum_k \ell_k\right)$ is the geometric mean probability across all tokens in the value, capturing overall generation confidence.

[1] https://github.com/awslabs/stickler/tree/local/conf-plus-date

- **First-token** $\exp(\ell_1)$ uses only the probability of the first token, reflecting how decisively the model committed to this value at the point of first generation.
- **Margin** $\frac{p_1 - p_2}{p_1 + p_2}$ is the normalized gap between the top-2 vocabulary candidates at the first token position, measuring how unambiguously the model chose its answer over the next best alternative.

Token budgets. For verbalized 1S-TopK, output token limits follow the model provider. Anthropic models (Claude Opus, Sonnet, Haiku) are capped at 40K tokens, reflecting Bedrock's higher output limit for those models. All open-weight models are capped at 8K tokens, matching the Bedrock output limit for those providers; self-hosted models (Qwen 3.6-27B, Gemma 3-12B) use the same 8K cap via `max_new_tokens` for consistency. Because the model emits four candidates per entity across all fields in a single response, entity counts can vary across models due to differences in output verbosity under these caps. For 1S-Logprob, self-hosted models use `max_new_tokens` $= 8192$, which is sufficient because the prompt requests a single compact JSON answer rather than multiple candidates.

Image resize. For 1S-TopK, we resize document images to 1200×1000 pixels, preserving aspect ratio, to give verbalized models sufficient visual resolution for fine-grained text recognition. For 1S-Logprob we set `max_pixels` $= 800 \times 800 = 640{,}000$ in the Qwen processor. This is a total pixel budget: the processor dynamically rescales the image so that width $\times$ height does not exceed this limit while preserving the original aspect ratio, rather than forcing a fixed output shape. This reduces memory footprint during greedy decoding without materially affecting extraction quality on single-page invoice documents.

Generation hyperparameters for 1S-Logprob. We run greedy decoding with temperature $T{=}0$ and `top_p` $= 0.1$ on Qwen 3.6-27B and Gemma 3-12B, capturing the top-10 logprobs at each generation step from the raw pre-temperature logits via `torch.log_softmax`. Chain-of-thought is disabled (`enable_thinking=False` for Qwen 3.6; Gemma 3 does not have a CoT toggle) so logprobs cover only the JSON answer.

Value-token alignment. Given the generated token sequence and its decoded text, we (1) locate the field key `"FieldName"` in the reconstructed string, (2) find the value substring after the key, (3) map each character offset to its source token index via a precomputed character-to-token map, and (4) deduplicate and order the resulting indices. For repeated keys inside the `LineItems` array, we anchor on the n-th occurrence of `"LineItemStartDate"` to disambiguate the target line item, then search forward for the requested sub-field.

Prompt adaptation across providers. Prompts are reformatted to match each provider's expected schema (system / user roles, JSON-mode flags, image-token placement) while preserving the underlying instructions, output schema, and field descriptions, so that differences across models reflect model capability rather than prompt engineering.

A.6 Reliability Diagrams

Figures 8, 9, and 10 present reliability diagrams for all evaluated configurations. While AUROC captures discriminative performance (the ability to separate correct from incorrect predictions), reliability diagrams reveal how well confidence scores are aligned with actual accuracy. The y=x diagonal represents ideal calibration; points below it indicate overconfidence, while points above indicate underconfidence. Notably, a model can have high AUROC (good discrimination) but poor calibration, or vice versa. Each panel plots accuracy (y-axis) against confidence (x-axis). Bins with fewer than 30 samples are omitted. Numbers above bars indicate sample counts per bin.

Key observations. (1) *Modality effect on calibration:* Image-only consistently yields the worst ECE across all models, while OCR+Image produces similar or better calibration than OCR-only, confirming that textual OCR signals improve not only discriminative power but also absolute calibration alignment. (2) *Verbalized overconfidence:* For the two models studied in direct strategy comparison (Gemma 3-12B and Qwen 3.6-27B), verbalized confidence concentrates the vast majority of predictions in the highest confidence bin (0.9–1.0). ECE is therefore dominated by the accuracy of this single bin. (3) *Mean-token follows y=x better than first-token:* Despite first-token achieving higher AUROC, mean-token logprob produces bars that track the diagonal more closely across mid-range bins. The similar ECE values arise because first-token's dominant high-confidence bin happens to be well-calibrated, masking its deviations elsewhere.

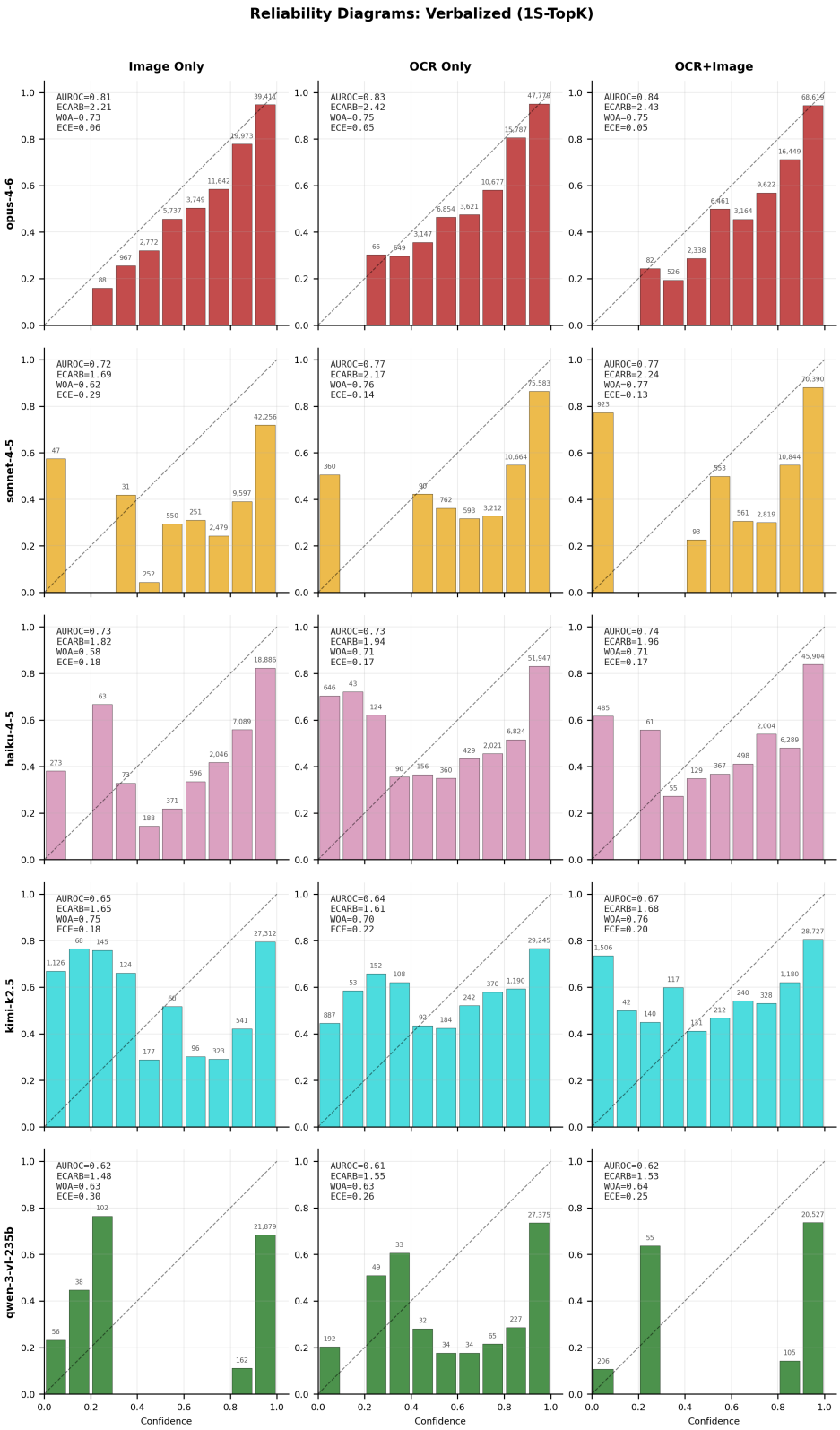


Figure 8: Reliability diagrams for all verbalized (1S-TopK) configurations. Rows: models ordered by AUROC (top = best). Columns: Image Only (left), OCR Only (center), OCR+Image (right). Opus 4.6 closely tracks the diagonal across all modalities, while weaker models show progressively larger overconfident deviations. The dominant high-confidence bin (0.9–1.0) contains the majority of predictions for all models.

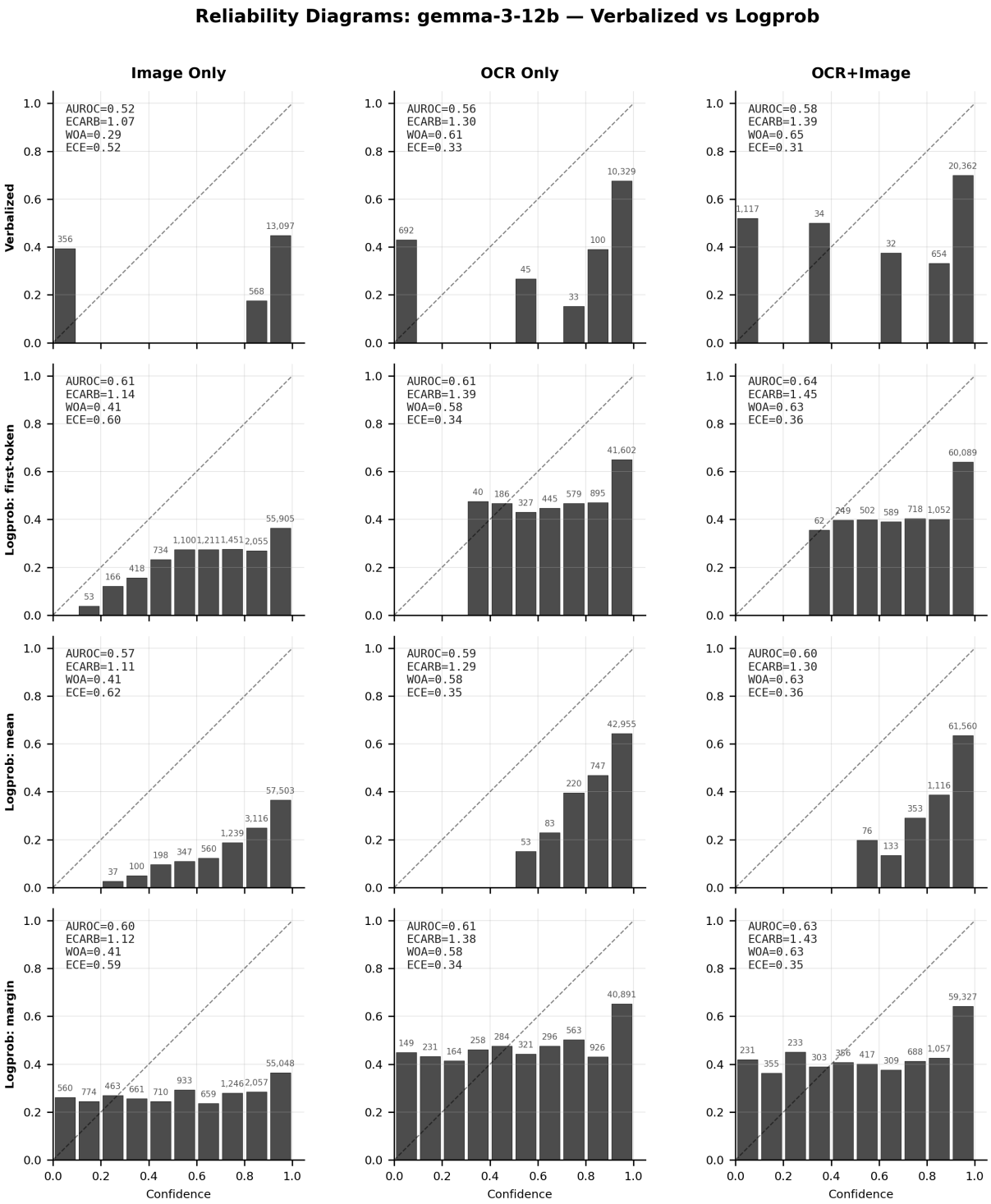


Figure 9: Reliability diagrams for Gemma 3-12B: verbalized vs logprob strategies. Rows (top to bottom): verbalized, first-token, mean, margin. Columns: Image Only (left), OCR Only (center), OCR+Image (right). First-token achieves the best AUROC. Both verbalized and first-token methods concentrate predictions in high-confidence bins. Mean-token tracks the y=x diagonal more faithfully. Margin shows the least alignment between the confidence measure and the corresponding accuracy.

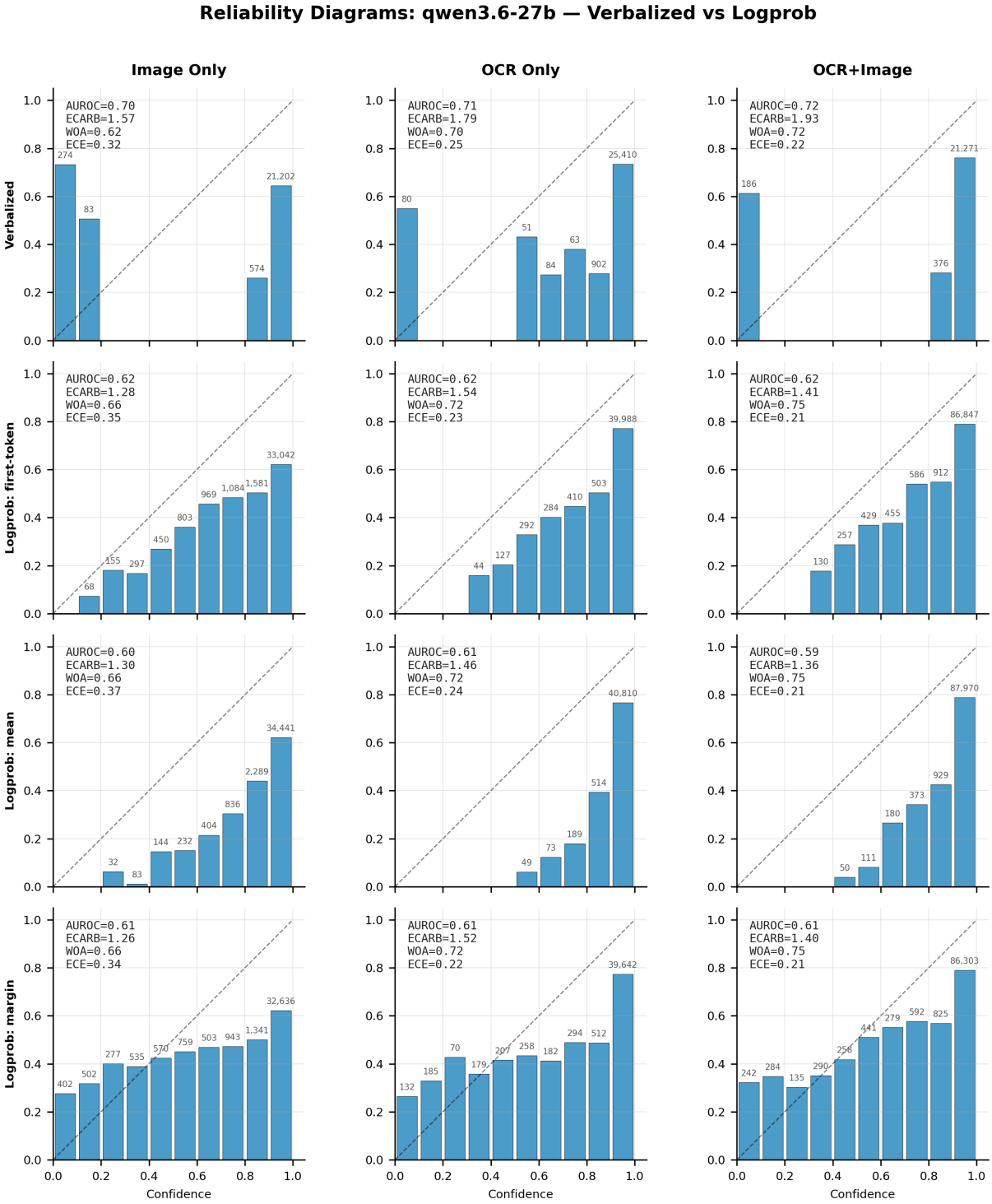


Figure 10: Reliability diagrams for Qwen 3.6-27B: verbalized vs logprob strategies. Rows (top to bottom): verbalized, first-token, mean, margin. Columns: Image Only (left), OCR Only (center), OCR+Image (right). Verbalized achieves the highest AUROC with extreme bimodal distribution. Among logprob methods, first-token and mean show similar ECE but mean tracks the diagonal more closely than the other two methods.