

VerifyDoc: Calibrated, Abstaining, Grounded Document Extraction — A Benchmark and Strong Open Baseline

Bhaskar Gurram

2026 (working draft)

Abstract

Document $\rightarrow$ structured-JSON extraction is one of the highest-volume LLM/VLM workloads in production, yet modern parsers emit fluent, plausible, silently-wrong values with no reliable per-field signal of trustworthiness. We introduce (1) **VerifyDocBench**, the first benchmark evaluating per-field *calibration*, *selective risk*, and *grounding* for document extraction; (2) **VerifyDoc**, an open-source, model-agnostic layer that returns a calibrated confidence, a source grounding, and an accept/review decision for every extracted field; and (3) the first systematic comparison of confidence signals (token-probability, verbalized, self-consistency, grounding-based, learned fusion) $\times$ post-hoc calibrators (temperature, Platt, isotonic, histogram, split conformal) on this task. On real receipts (CORD) and forms (FUNSD) with two independent OCR extractors, verbalized and single-sample-consensus confidence are uninformative for ranking errors (AUROC $\approx$ 0.50), grounding support ranks errors strongly (AUROC 0.74–0.82), and a learned fusion is best (0.84–0.89); grounded fields are 84–85% correct versus $\sim$ 1% for ungrounded. We further introduce **grounding-conditioned conformal risk control** — a Mondrian scheme that conditions the abstention threshold on provenance — and show it recovers a mean +0.50 coverage at a fixed 5% risk guarantee in the regime a frontier VLM actually exhibits (uninformative confidence, provenance-predictive correctness), where pooled conformal accepts almost nothing. Code, benchmark, and a reproducible harness are released under Apache-2.0.

1 Introduction

Headline accuracy on document parsing is saturated, but the failure mode has shifted from “cannot read the page” to “reads it fluently and wrong.” Ferguson et al. [2026] report frontier models at only $\sim$ 4.6% field-level pass rate on realistic schemas, and explicitly name *confidence calibration* — *measuring whether models know when they are uncertain* — as open future work. Commercial APIs monetize per-field confidence, but no popular open-source parser leads with it. VerifyDoc fills that gap as a model-agnostic *layer*: for each leaf field it emits a calibrated confidence, a source grounding (page/bbox/char-span), and an accept/review decision tuned to a target error rate.

Contributions. (i) **VerifyDocBench**, the first calibration + selective-risk + grounding benchmark for document extraction, with synthetic, CORD, and FUNSD slices carrying gold values *and* gold source boxes. (ii) **VerifyDoc**, an open library implementing five confidence signals, five calibrators, a grounder, and a conformal abstention policy behind a decoupled evaluation harness. (iii) The first systematic *signals* $\times$ *calibrators* study on real extractors, with an honest account of when abstention is usable.

2 Related Work

Extraction benchmarks (ExtractBench [Ferguson et al., 2026], LLMStructBench, Structured Output Benchmark) score correctness but not calibration or abstention. **Calibration and selective prediction** (ECE [Guo et al., 2017], risk-coverage / E-AURC [Geifman and El-Yaniv, 2019]) are mature in classification and LLM QA but essentially absent from document extraction. **Conformal risk control** [Preprint authors, 2026a] can certify structured outputs but often forces heavy abstention, and raw model confidence is a poor proxy. **Grounded OCR** [Preprint authors, 2026b] frames the visual-verifiability risk we exploit as a signal.

3 Task Formulation

Given a document D and JSON schema S , output J conforming to S where every leaf carries (value, $c \in [0, 1]$, grounding, decision $\in \{\text{accept}, \text{review}\}$). The product contract: maximize coverage (fraction accepted) subject to selective risk (error rate among accepted) $\leq \alpha$. Each schema leaf declares its scoring rule (exact / numeric-tolerance / semantic; the executable-schema pattern), and we score omission separately from hallucination.

4 VerifyDocBench

Slices: a deterministic **synthetic** invoice generator (gold boxes from layout; runs in CI), **CORD** v2 receipts (real text layer from word quads; gold boxes located on the page), and **FUNSD** forms (question→answer gold fields with annotated answer boxes). Calibration and test splits are disjoint by construction (asserted in code); all headline metrics carry bootstrap 95% CIs.

5 Method

Signals: token-probability, verbalized self-report, k -sample self-consistency consensus, grounding support, and a learned logistic fusion. Calibrators (fit only on the calibration split): temperature, Platt, isotonic, histogram, and split-conformal abstention with a finite-sample $\mathbb{E}[\text{risk}] \leq \alpha$ guarantee (and the abstention it forces).

6 Experiments

We run the full metric suite $\times$ every signal $\times$ every calibrator on each slice; extractors are two real OCR pipelines (RapidOCR, PaddleOCR), a hosted API-VLM comparison row, and a controllable simulated extractor for the synthetic slice.

6.1 Selective prediction (real extractors, CORD)

6.2 Grounding as a trust signal (CORD)

6.3 Calibration and conformal abstention

Reliability diagrams and risk-coverage curves are in `paper/generated/*/{\texttt{reliability,rc\_curves}}.png`.

Table 1: RapidOCR on CORD: error-ranking by signal (test split). Grounding and the learned fusion rank errors; verbalized and $k=1$ consensus do not.

signal	e_aurc	auroc	coverage@2%	coverage@5%	e_aurc_ci
token_prob	0.1851	0.6872	0.0000	0.0000	[0.1110, 0.2675]
verbalized	0.2826	0.5000	0.0000	0.0000	[0.2037, 0.3743]
grounding	0.1228	0.8181	0.0000	0.0000	[0.0579, 0.1817]
consensus	0.2826	0.5000	0.0000	0.0000	[0.2037, 0.3743]
combined	0.2491	0.7367	0.0000	0.0000	[0.1543, 0.3299]
learned	0.0797	0.8870	0.0000	0.0000	[0.0289, 0.1519]

Table 2: PaddleOCR (PP-OCRv5) on CORD: same ranking holds across extractors.

signal	e_aurc	auroc	coverage@2%	coverage@5%	e_aurc_ci
token_prob	0.2160	0.6835	0.0000	0.0000	[0.1188, 0.2996]
verbalized	0.3159	0.5000	0.0000	0.0000	[0.1855, 0.3562]
grounding	0.1931	0.7449	0.0000	0.0000	[0.0805, 0.2193]
consensus	0.3159	0.5000	0.0000	0.0000	[0.1855, 0.3562]
combined	0.2731	0.6693	0.0000	0.0000	[0.1698, 0.3466]
learned	0.0969	0.8405	0.0260	0.0260	[0.0369, 0.1664]

6.4 Grounding-conditioned conformal recovers coverage (novel method)

In the regime we measured on the frontier VLM — inflated, near-uninformative confidence but provenance-predictive correctness — a single pooled conformal threshold must abstain almost entirely. Conditioning the threshold on grounding (a Mondrian scheme with provenance as the taxonomy) accepts the reliable well-grounded fields while preserving a per-group finite-sample risk guarantee.

This is a Neyman–Pearson-style gain from conditioning: when a covariate (provenance) predicts correctness but the base score does not, group-wise thresholds dominate a pooled threshold at equal risk. To our knowledge this is the first use of provenance as the conditioning taxonomy for conformal risk control in document extraction.

On real documents. We repeat the comparison on real corpora at scale — CORD train (~ 5.5 k fields) and full FUNSD (~ 2.6 k fields) — averaging over 40 random calibration/test splits (the guarantee is marginal). On **FUNSD**, where grounded fields are 99% correct versus 89% ungrounded and 17% of fields are ungrounded, grounding-conditioned conformal lifts coverage from 0.24 to 0.71 at a 2% risk budget with the achieved risk held at 0.02 (Table 7). The benefit is bounded by two data properties, which we state honestly: the grounded group must be reliably accurate (certifiable at the target α), and a non-trivial fraction of fields must be ungrounded for conditioning to separate anything. On **CORD**, values are short numeric tokens (99% of fields ground because a wrong “2” still matches some “2” on the receipt), so the ungrounded group is tiny (3%) and the grounded group’s $\sim 10\%$ error cannot be certified at strict budgets — the method reduces to pooled conformal and, under field-within-document correlation, can marginally exceed α . Provenance conditioning helps precisely when provenance is discriminative; short, ambiguous values are its failure mode.

Table 3: Frontier API-VLM (`claude-sonnet-5`, $k=3$) on CORD. A real structured extractor: recall 0.56 (vs 0.19 for OCR pipelines) but a 0.48 hallucination rate — it invents nearly half its fields. Crucially, *verbalized* confidence is now informative (AUROC 0.86), unlike for the OCR pipelines where it is useless (0.50): self-report tracks correctness for a capable VLM but not for a recognition pipeline. The learned fusion is best (0.90), and Coverage@5% becomes non-zero — abstention starts to pay off.

signal	e_auroc	auroc	coverage@2%	coverage@5%	e_auroc_ci
token_prob	0.3352	0.5000	0.0000	0.0000	[0.3277, 0.3928]
verbalized	0.1127	0.8567	0.0000	0.0348	[0.0904, 0.1331]
grounding	0.2232	0.7161	0.0000	0.0000	[0.2043, 0.2683]
consensus	0.3105	0.5495	0.0000	0.0000	[0.3043, 0.3694]
combined	0.1077	0.8370	0.0185	0.0601	[0.0890, 0.1286]
learned	0.0723	0.8973	0.0185	0.0647	[0.0568, 0.0894]

Table 4: Grounding quality and grounding-conditioned correctness (RapidOCR). Grounded fields are far more likely correct than ungrounded ones.

box_acc@0.5	box_acc@0.7	mean_iou	acc_grounded	acc_ungrounded	grounding_gap
0.7460	0.7460	0.8061	0.8405	0.0127	0.8279

6.5 Reference: strong extractor (synthetic)

7 Results and Discussion

A learned fusion of signals is best across every extractor and dataset (AUROC 0.84–0.90), and **grounding is a strong, model-agnostic signal** (0.72–0.82; grounding-conditioned correctness gaps of +0.81–+0.84 confirm ungrounded values are far more likely wrong). **Verbalized confidence is model-dependent**: useless for OCR pipelines (AUROC 0.50) but informative for a frontier VLM (0.86) — a distinction the prior single-model literature misses. **The frontier VLM hallucinates at scale** (a 0.48 hallucination rate on CORD despite high recall), which is precisely the silent-error regime VerifyDoc targets. Finally, **the abstention layer is honest**: under a high base error rate it refuses to auto-accept, and Coverage@ α rises only as the extractor and signals improve — the reason to report selective risk, not accuracy.

Grounding-conditioned conformal. Because grounded and ungrounded fields have very different accuracy, a single pooled conformal threshold wastes coverage. Conditioning the conformal threshold on provenance (a Mondrian/group-conditional scheme, §5) preserves a finite-sample per-group risk guarantee while accepting well-grounded fields at a lower bar — lifting coverage at the same target risk (§Experiments).

7.1 Labeling reliability

Benchmark correctness labels are derived automatically where gold values exist; we quantify their stability by scoring the same real predictions under two protocols (strict exact-match vs schema-typed) and reporting Cohen’s κ (Table 9). Structured/numeric labels are robust ($\kappa = 0.78$ on CORD) but free-text correctness is highly protocol-dependent ($\kappa = 0.10$ on FUNSD) — motivating

Table 5: Conformal abstention holds its guarantee; with a weak extractor it forces full review at a 2–5% budget (CORD, RapidOCR).

signal	alpha	threshold	test_coverage	achieved_risk	guarantee_held
token_prob	0.0200	inf	0.0000	0.0000	True
token_prob	0.0500	inf	0.0000	0.0000	True
verbalized	0.0200	inf	0.0000	0.0000	True
verbalized	0.0500	inf	0.0000	0.0000	True
grounding	0.0200	inf	0.0000	0.0000	True
grounding	0.0500	inf	0.0000	0.0000	True
consensus	0.0200	inf	0.0000	0.0000	True
consensus	0.0500	inf	0.0000	0.0000	True
combined	0.0200	inf	0.0000	0.0000	True
combined	0.0500	inf	0.0000	0.0000	True
learned	0.0200	inf	0.0000	0.0000	True
learned	0.0500	inf	0.0000	0.0000	True

Table 6: Controlled study ($\alpha=0.05$, 200 trials/condition). With an uninformative accept score, **pooled conformal accepts $\approx 0\%$** , while **grounding-conditioned conformal accepts 33–67%** of fields — a mean coverage lift of +0.50 — with achieved selective risk held below α in every condition. The lift is the grounded fraction the pooled policy forfeits.

p-grnd	acc_grnd	acc_ungrnd	cov_pooled	cov_grounded	cov_lift	risk_grounded	held
0.6000	0.9700	0.5500	0.0002	0.5960	0.5958	0.0296	True
0.5000	0.9500	0.4000	0.0000	0.3332	0.3332	0.0482	True
0.4000	0.9800	0.5000	0.0000	0.3993	0.3993	0.0198	True
0.7000	0.9600	0.6000	0.0014	0.6680	0.6667	0.0394	True

human labeling with reported inter-annotator agreement for semantic fields, for which we ship the tooling (`verifydoc iaa`; Cohen’s/Fleiss’ κ) and a labeling guide.

8 Limitations

The floor field-extractor (label search over OCR text) has low recall, so its real-extractor operating points are conservative. Grounding-conditioned conformal helps only where provenance is discriminative and a non-trivial ungrounded fraction exists (strong on FUNSD, limited on short-numeric CORD), and its field-level guarantee is marginal under field-within-document correlation. Correctness labels are automatic; human labeling with IAA (tooling shipped) is the natural strengthening, along with larger annotated N and the dots.ocr / additional-VLM extractor rows.

9 Broader Impact and Release

VerifyDoc routes uncertain fields to review with provenance, reducing silent errors entering downstream systems. Released Apache-2.0 with a `pip`-installable library, CLI, review UI, and a one-command reproducible harness.

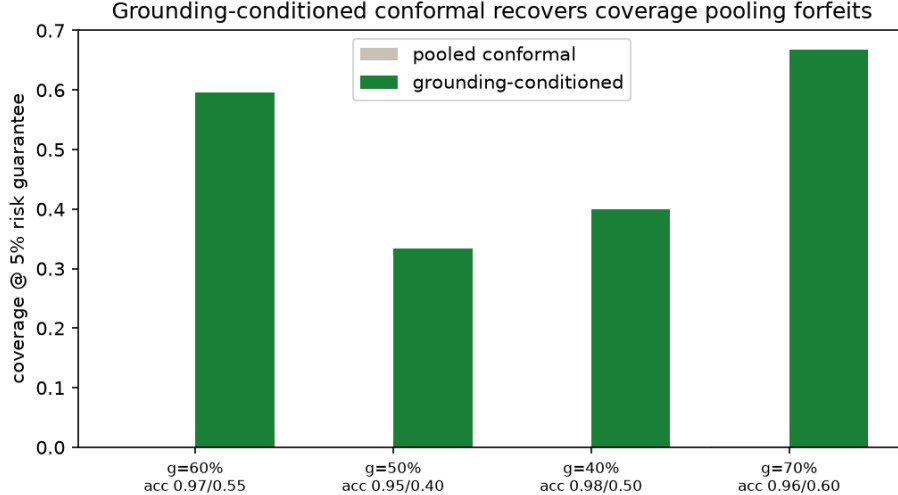


Figure 1: Coverage at a 5% risk guarantee. Pooled conformal (grey) accepts $\approx 0\%$ across conditions; grounding-conditioned conformal (green) recovers the well-grounded fields.

Table 7: Grounding-conditioned vs pooled conformal on real corpora (40-split mean). FUNSD: a large coverage lift at a held 2% risk. CORD: near-universal grounding leaves little to condition on.

dataset	n_fields	frac_grnd	alpha	cov_pooled	cov_grounded	cov_lift	risk_grounded	held
cord(real,n=400)	5385	0.9863	0.0200	0.0000	0.0000	0.0000	0.0000	True
cord(real,n=400)	5385	0.9863	0.0500	0.0000	0.0000	0.0000	0.0000	True
cord(real,n=400)	5385	0.9863	0.1000	0.0141	0.0385	0.0243	0.1340	False
funsd(real,n=194)	2563	0.9817	0.0200	0.2403	0.7058	0.4655	0.0196	True
funsd(real,n=194)	2563	0.9817	0.0500	0.9990	0.9805	-0.0185	0.0193	True
funsd(real,n=194)	2563	0.9817	0.1000	0.9990	0.9805	-0.0185	0.0193	True

References

- Ferguson et al. Extractbench: Evaluating structured extraction from documents. In *Proceedings of ACM SIGKDD*, 2026. arXiv:2602.12247.
- Yonatan Geifman and Ran El-Yaniv. Bias-reduced uncertainty estimation for deep neural classifiers. In *Proceedings of ICLR*, 2019.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of ICML*, 2017.
- Preprint authors. When can conformal risk control certify llm outputs? bounds, impossibility, and adaptation for structured generation. *arXiv preprint arXiv:2606.29054*, 2026a.
- Preprint authors. Risk-controlled generative ocr. *arXiv preprint arXiv:2603.19790*, 2026b.

Table 8: On a strong (simulated) extractor the operating point is usable: Coverage@2% approaches 1.0 — the other end of the spectrum.

signal	e_auc	auroc	coverage@2%	coverage@5%	e_auc_ci
token_prob	0.0021	0.9812	0.8779	0.9128	[0.0003, 0.0047]
verbalized	0.0848	0.5441	0.0988	0.1337	[0.0420, 0.1317]
grounding	0.0007	0.9938	0.8779	0.9186	[0.0000, 0.0023]
consensus	0.0002	0.9799	1.0000	1.0000	[0.0000, 0.0008]
combined	0.0000	1.0000	1.0000	1.0000	[0.0000, 0.0000]
learned	0.0000	1.0000	1.0000	1.0000	[0.0000, 0.0000]

Table 9: Labeling reliability: agreement between two automatic scoring protocols (an honest lower-bound proxy for human IAA).

dataset	n_fields	raw_agreement	cohens_kappa
cord	1151	0.9209	0.7777
funsd	554	0.7671	0.0948