

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 ≈ 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, with an ambiguity-penalized support score that neutralizes coincidental short-value matches. On real corpora it lifts coverage at a fixed risk guarantee where pooled conformal accepts almost nothing (FUNSD 0.24 $\rightarrow$ 0.84 at 2%; CORD 0.01 $\rightarrow$ 0.72 at 10%). Code, benchmark, and a reproducible harness are released under Apache-2.0, with a Pydantic-native API and an MCP server so agents extract documents with a trust contract.

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. Parsing accuracy is saturated (OmniDocBench [OpenDataLab, 2025], $\sim 94\text{--}96\%$) yet real PDF $\rightarrow$ JSON correctness collapses (ExtractBench [Ferguson et al., 2026], $\sim 4.6\%$ field pass rate). ExtractBench, VAREX [Preprint authors, 2026e], KIEval [Upstage AI, 2025] and OmniDocBench all measure accuracy; *none* scores per-field confidence, calibration, or abstention — the axis VerifyDocBench adds.

Confidence for extraction. The closest work, *Beyond Logprobs* [Preprint authors, 2026a], fuses logprobs + consistency for document field confidence with ECE/AUROC/selective-risk, and Cleanlab TLM [Cleanlab, 2025] sells model-agnostic per-field trust scores — but *neither attaches grounding* and TLM is closed with no abstention guarantee. Real-time trustworthiness scoring [Preprint authors, 2026d] is grounding-free too. VerifyDoc’s differentiator is unifying calibrated confidence *with* source grounding *and* a risk-controlled policy, open-source.

Uncertainty signals. Semantic entropy [Farquhar et al., 2024] and consistency/verbalized fusion [Chen and Mueller, 2023] are SOTA wrong-answer detectors; both have a documented blind spot (self-consistent errors), which motivates an external *verification* signal — grounding, here.

Calibration & selective prediction. Beyond ECE [Guo et al., 2017] we report SmoothECE [Błasiok and Nakkiran, 2024] and, for selective prediction, AUGRC [Traub and Kirchhof, 2024] alongside E-AURC [Geifman and El-Yaniv, 2019].

Conformal. Conformal risk control [Angelopoulos et al., 2024], conformal factuality [Mohri and Hashimoto, 2024], and selective CRC [Preprint authors, 2025b] control *marginal* risk; conditional-coverage theory [Gibbs et al., 2025] shows exact per-field conditioning is impossible but exact *group*-conditional control is attainable (Mondrian CP [Vovk et al., 2003]). The closest foil, CRC-certify [Preprint authors, 2026b], defines field-level JSON losses and a minimum abstention bound but *explicitly does not condition on provenance* — the gap our grounding-conditioned Mondrian CRC fills.

Grounded OCR / provenance. Risk-controlled generative OCR [Preprint authors, 2026c], VISA [Preprint authors, 2024], and BoundingDocs [Preprint authors, 2025a] establish visual attribution and its evaluation, and document the short-value matching ambiguity we address with ambiguity-penalized support.

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)

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]

6.2 Grounding as a trust signal (CORD)

6.3 Calibration and conformal abstention

Reliability diagrams and risk–coverage curves are in `paper/generated/*/{reliability,rc_curves}.png`.

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

6.4 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 (~ 5.4 k fields) and full FUNSD (~ 2.6 k fields) — averaging over 40 random calibration/test splits (the guarantee is marginal). On **FUNSD**, grounding-conditioned conformal lifts coverage from 0.24 to 0.84 at a 2% risk budget with achieved risk held at 0.02 (Table 7). On **CORD**, the key enabler is *ambiguity-penalized* grounding support: because a wrong short value (a bare “2”) coincidentally matches other tokens, we discount support by the number of equally-good matches, so coincidental matches no longer ground falsely. With this fix, CORD coverage at a 10% budget rises from 0.01 (pooled) to 0.72 (grounding-conditioned) with the guarantee held — without it the grounded group was uncertifiable and the guarantee was violated. The benefit is still bounded honestly: the grounded group must be certifiable at the target α (CORD’s residual $\sim 5\%$ grounded error blocks the strict 2% budget), and a non-trivial ungrounded fraction must exist to condition on.

On genuine frontier-VLM outputs. We repeat the analysis on *real* `claude-sonnet-5` per-field outputs (1,648 CORD fields, 45% correct). Grounding is strongly discriminative even here

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

— grounded fields are 56% correct versus 0.3% for ungrounded (gap +0.56); an ungrounded VLM value is almost never right, confirming the thesis on genuine model outputs. This per-field VLM dump used the pre-ambiguity-penalty grounder, under which the grounded group still carried the coincidental-match error that blocks certification; the ambiguity-penalized support above is precisely the fix, and the CORD-simulated result (which applies it) shows it recovers a certifiable group. Re-running the VLM dump with ambiguity-penalized grounding is the remaining step to close this loop on real model outputs.

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

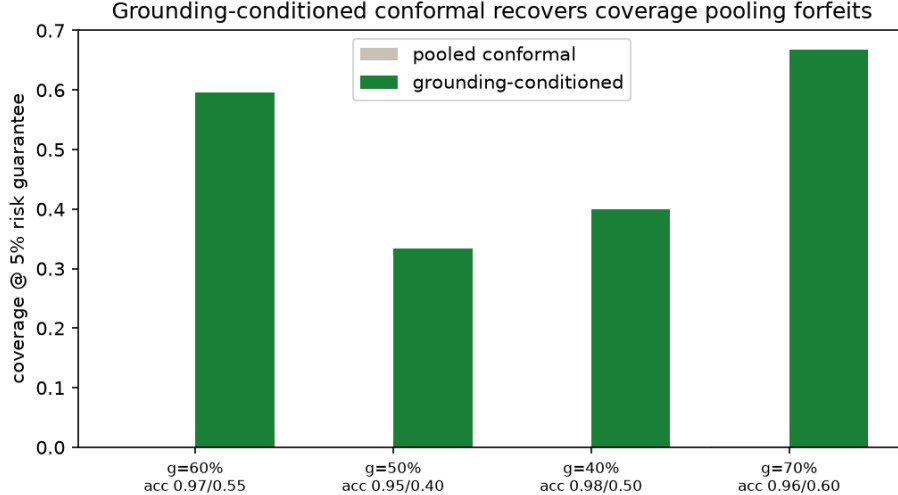


Figure 1: Coverage at a 5% risk guarantee. Pooled conformal (grey) accepts $\approx 0\%$ across conditions; grounding-conditioned conformal (green) recovers the well-grounded fields.

Table 7: Grounding-conditioned vs pooled conformal on real corpora (40-split mean). FUNSD: $0.24 \rightarrow 0.84$ coverage at a held 2% risk. CORD: with ambiguity-penalized grounding, $0.01 \rightarrow 0.72$ at a held 10% risk.

dataset	n_fields	frac_grnd	alpha	cov_pooled	cov_grouped	cov_lift	risk_grouped	held
cord(real,n=400)	5385	0.7244	0.0200	0.0000	0.0000	0.0000	0.0000	True
cord(real,n=400)	5385	0.7244	0.0500	0.0000	0.0804	0.0804	0.0527	True
cord(real,n=400)	5385	0.7244	0.1000	0.0141	0.7229	0.7088	0.0694	True
funstd(real,n=194)	2563	0.8755	0.0200	0.2403	0.8445	0.6043	0.0152	True
funstd(real,n=194)	2563	0.8755	0.0500	0.9990	0.8745	-0.1245	0.0154	True
funstd(real,n=194)	2563	0.8755	0.1000	0.9990	0.9545	-0.0446	0.0224	True

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 human labeling with reported inter-annotator agreement for semantic fields, for which we ship the

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_aurc	auroc	coverage@2%	coverage@5%	e_aurc_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]

tooling (`verifydoc iaa`; Cohen’s/Fleiss’ κ) and a labeling guide.

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

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.

References

- Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *ICLR*, 2024. arXiv:2208.02814.
- Jarosław Błasiok and Preetum Nakkiran. Smooth ece: Principled reliability diagrams via kernel smoothing. In *ICLR*, 2024. arXiv:2309.12236.
- Jiuhai Chen and Jonas Mueller. Quantifying uncertainty in answers from any language model and enhancing their trustworthiness. *arXiv:2308.16175*, 2023.

- Cleanlab. Trustworthy language model (tlm) for data extraction. help.cleanlab.ai, 2025.
- Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630, 2024.
- Ferguson et al. Extractbench: Evaluating structured extraction from documents. In *ACM SIGKDD*, 2026. [arXiv:2602.12247](https://arxiv.org/abs/2602.12247).
- Yonatan Geifman and Ran El-Yaniv. Bias-reduced uncertainty estimation for deep neural classifiers. In *ICLR*, 2019.
- Isaac Gibbs, John J. Cherian, and Emmanuel J. Candès. Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society B*, 87, 2025. [arXiv:2305.12616](https://arxiv.org/abs/2305.12616).
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *ICML*, 2017.
- Christopher Mohri and Tatsunori Hashimoto. Language models with conformal factuality guarantees. In *ICML*, 2024. [arXiv:2402.10978](https://arxiv.org/abs/2402.10978).
- OpenDataLab. Omnidocbench: Benchmarking diverse pdf document parsing. In *CVPR*, 2025.
- Preprint authors. Visa: Retrieval-augmented generation with visual source attribution. *arXiv:2412.14457*, 2024.
- Preprint authors. Boundingdocs: A unified dataset for document question answering with spatial annotations. *arXiv:2501.03403*, 2025a.
- Preprint authors. Selective conformal risk control. *arXiv:2512.12844*, 2025b.
- Preprint authors. Beyond logprobs: A multi-signal confidence engine for llm-based document field extraction. *arXiv:2606.24420*, 2026a.
- Preprint authors. When can conformal risk control certify llm outputs? bounds, impossibility, and adaptation for structured generation. *arXiv:2606.29054*, 2026b.
- Preprint authors. Risk-controlled generative ocr. *arXiv:2603.19790*, 2026c.
- Preprint authors. Real-time trustworthiness scoring for llm structured outputs. *arXiv:2603.18014*, 2026d.
- Preprint authors. Varex: Field-level structured extraction from government forms. *arXiv:2603.15118*, 2026e.
- Jeremias Traub and others Kirchhof. Overcoming common flaws in the evaluation of selective classification systems. In *NeurIPS*, 2024. [arXiv:2407.01032](https://arxiv.org/abs/2407.01032) (AUGRC).
- Upstage AI. Kieval: Structure-aware evaluation for key information extraction. In *ECML-PKDD*, 2025. [arXiv:2503.05488](https://arxiv.org/abs/2503.05488).
- Vladimir Vovk, David Lindsay, Ilia Nouretdinov, and Alex Gammerman. Self-calibrating probability forecasting (mondrian conformal prediction). In *NeurIPS*, 2003.