Knowing when a model is wrong: field-level confidence for AI document extraction

25 November 2026 · 8 min read · blogs.iictp.org/vlm-field-confidence-token-logprobs-financial-document-extraction-explained

Short answer: an arXiv paper submitted on 23 November 2026 by Varad Srivastava and Victor Casas Hernandez shows how to attach a confidence score to each field a vision-language model extracts from scanned financial documents, using the token log-probabilities the model already produces. On two benchmarks the minimum token probability per field detected errors with AUROC of 0.74 to 0.91, against 0.53 to 0.56 for the model's own stated confidence. The paper was accepted at the AI4Finance Summit at ICAIF 2026 and says an extended version will follow.

TL;DR The short version

- A paper shows how to give each field a vision-language model extracts from scanned financial documents a confidence score, using token log-probabilities the model already produces.
- Minimum token probability detected errors far better than the model's own stated confidence, with AUROC of 0.74 to 0.91 against 0.53 to 0.56.
- The tests used receipts and NDAs with small training sets, but the approach suits accountants' review-by-exception habit.

Read online, listen to it, or change the text size: https://blogs.iictp.org/vlm-field-confidence-token-logprobs-financial-document-extraction-explained

What problem does it address?

Models can extract structured fields (totals, dates, parties) from receipts and contracts, but give no signal of when a field is wrong. For accounting work, where an unnoticed error flows into a ledger, a reliable "check this one" flag is more valuable than an average accuracy figure.

How it works

- **Token log-probabilities.** The model's output already carries a probability for each token.
- **Token-to-span alignment.** The authors map those probabilities onto each extracted value inside the structured JSON, so each field gets its own score.
- **Minimum token probability.** The lowest token probability in a field is the confidence signal.
- **Confidence-guided curriculum.** The uncertain fields become training examples for fine-tuning with Group Relative Policy Optimisation (GRPO), using only training-split ground truth.

What is new compared with earlier approaches

The common alternative is to ask the model how confident it is. The paper reports that this “verbalised” confidence barely separates right from wrong, whereas probabilities from the output do.

Results reported by the source

- Datasets: CORD v2 receipts and Kleister NDA legal contracts, with three VLM families.
- Error detection AUROC 0.74 to 0.91 for minimum token probability, against 0.53 to 0.56 for verbalised confidence.
- GRPO fine-tuning improved accuracy from 78.5% to 87.5% on CORD (67 examples) and from 70.7% to 75.1% on multi-page NDA contracts (259 examples).
- On Kleister NDA, error-detection AUROC rose from 0.70 to 0.85 and AUARC from 0.78 to 0.92.
- A frontier model used as a pseudo-label judge agreed with ground truth on about 92% of rewards.

Limitations and open questions

- The training sets are very small (67 and 259 examples) and the benchmarks are receipts and NDAs, not invoices, bank statements or ledgers.
- The authors say an extended version with a revised experimental protocol will follow, so numbers may change.
- It needs access to token probabilities, which not every hosted model API exposes. Sources were read on 25 November 2026. Products, rules and standards change, so check the linked originals before you rely on any detail.

How to apply the idea in a finance workflow

1. Extract fields into JSON and keep the token probabilities where your model API provides them.
2. Score each field by its minimum token probability.
3. Route low-confidence fields to a person and let high-confidence fields pass with sampling checks.
4. Track the share of routed fields and the errors found, and tune the threshold on your own documents.

What it means for accountants and finance operations

Automation that tells you which fields to check fits how accountants already work: review by exception. A threshold needs calibrating on your own document types, and a human must remain responsible for what is posted. See Filing-evidence: small models reading SEC filings for a related evidence-first approach.

Frequently asked questions

- **What is field-level confidence?** A score for each extracted value, rather than for the whole document.
- **How is it computed?** From the minimum token probability inside each field’s span.
- **How good is it?** AUROC 0.74 to 0.91 on the two benchmarks, per the authors.
- **Is it peer reviewed?** It is an arXiv paper accepted at an ICAIF workshop-style summit; the authors say an extended version follows.

Related reading

- Filing-evidence: small LLMs reading SEC filings
- Auditable financial knowledge graphs
- AI Practitioner certification

> Related credential: AI Practitioner. Explore the exam facts and syllabus.

Sources and further reading

1. Srivastava V, Casas Hernandez V. Could This BE Any More Wrong? Teaching VLMs to Know What They Don't Know in Document Extraction. arXiv 2611.30119, 23 Nov 2026 (accepted at the AI4Finance Summit, ICAIF 2026). arxiv.org/abs/2611.30119

Published by IICTP, an independent technical credential body, on 25 November 2026. The current version of this post is at https://blogs.iictp.org/vlm-field-confidence-token-logprobs-financial-document-extraction-explained. More guides: blogs.iictp.org.