Towards a Metrology of Exhaustiveness in Document Analysis:

A Systemic Framework for Layout Completeness Assessment

Gabriel Zo-Hasina Rasatavohary*
Aquantic / ZONOVA Research, France
ORCID 0009-0006-2770-1474

February 2026

Document type	Position Paper (no experimental results)
Author	Gabriel Zo-Hasina Rasatavohary
ORCID	0009-0006-2770-1474
Affiliation	Aquantic / ZONOVA Research, France
Programme	MatrixAI — Intelligent Document Processing R&D
Date	February 2026
Preprint server	OSF Preprints
Primary field	Computer Vision (cs.CV), Digital Libraries (cs.DL)
Secondary	Artificial Intelligence (cs.AI), Information Retrieval (cs.IR)
Peer review	Pre-submission review by Prof. Emeritus Ioan Roxin (Univ. Franche-Comté)
License (paper)	CC BY-NC-ND 4.0 — © 2026 Aquantic / ZONOVA Research
License (code)	Non-Commercial Research License v1.0 — see https://github.com/rasata/dla_cci
Code	https://github.com/rasata/dla_cci
Core concept	Completeness Confidence Index (CCI)
Keywords	Document Layout Analysis, Completeness Assessment, Uncertainty Quantification, Unknown Unknowns, Layout Metrology, Process Certification, Residual Entropy, Ghost OCR, Conformal Prediction

Abstract

Current evaluation paradigms in Document Layout Analysis (DLA) overwhelmingly focus on measuring the quality of *detected* elements through metrics such as Intersection over Union (IoU), mean Average Precision (mAP), and F1-score. While these metrics assess detection accuracy, they remain structurally silent on a critical question: *what has been missed?* In industrial contexts where documents feed safety-critical processes—energy infrastructure maintenance, pharmaceutical compliance, financial auditing, legal contract

*ZONOVA Research / MatrixAI Programme — Intelligent Document Processing R&D. Contact: zo@research.zonova.io

analysis—an omitted layout element can propagate significant downstream consequences that far exceed the cost of a misclassification. This paper argues that the field requires a fundamental shift from *detection performance* to *completeness certification*. We introduce the **Completeness Confidence Index (CCI)**, a conceptual framework that aggregates three independent proof vectors—*residual signal analysis*, *structural coherence validation*, and *cross-modal redundancy*—to estimate the probability that a layout analysis has captured all semantically relevant regions of a document. We formalize the notion of *informative void*, drawing on epistemic uncertainty quantification, conformal prediction theory, and probabilistic document grammars. Rather than presenting experimental results, this position paper establishes the theoretical foundations and formalizes the research agenda, calling for the creation of an "Omission Challenge" benchmark and for process-dependent calibration of completeness metrics. We argue that as AI-driven document analysis becomes pervasive in industrial pipelines, neutralizing uncertainty about what remains undetected is not merely an academic concern but an operational imperative.

Keywords: Document Layout Analysis, Completeness Assessment, Uncertainty Quantification, Unknown Unknowns, Layout Metrology, Process Certification, Residual Entropy.

1 Introduction

1.1 The Presence Bias in Document AI

The remarkable progress of deep learning in Document Layout Analysis (DLA) has been measured almost exclusively through metrics inherited from object detection: mean Average Precision (mAP), Intersection over Union (IoU), precision, recall, and F1-score [Everingham et al., 2010, Lin et al., 2014]. These metrics share a fundamental structural property: they evaluate the quality of *what has been detected*. A model that achieves 95% mAP on PubLayNet [Zhong et al., 2019] is praised for its detection prowess, yet the metric is agnostic to whether the remaining 5% represents benign noise or a systematically missed document structure.

We term this phenomenon the **presence bias**: the tendency of evaluation frameworks to reward the correct identification of present elements while providing no formal mechanism to quantify confidence about *absent* ones. As Pfitzmann et al. [2022] have shown through large-scale annotation studies, current evaluation metrics for DLA suffer from fundamental limitations—including significant inter-annotator disagreement—that mask important failure modes. The LED benchmark [Heo et al., 2025] has taken an important step by *diagnosing* structural errors, categorizing how models fail when they do detect elements—but even this diagnostic framework does not address the question of what was never detected in the first place.

This gap is not merely theoretical. It reflects a deeper epistemological asymmetry: **measuring what is present is a bounded problem; estimating what is absent is an open one**.

In practice, industrial practitioners have developed ad hoc compensatory strategies: double extraction (running two independent pipelines and comparing outputs), systematic human verification (manual page-by-page review), and heuristic post-processing rules tailored to specific document types. While these approaches provide a degree of safety, they are fundamentally unscalable, lack formal guarantees, and offer no quantitative measure of the residual risk of

omission. The CCI framework proposed in this paper aims to fill precisely this gap: providing a principled, measurable, and process-calibrated alternative to these informal practices.

1.2 Industrial Stakes: When Omission Becomes Risk

The urgency of this problem is amplified by the rapid industrialization of Document AI. Intelligent document processing pipelines are now embedded in workflows where the consequences of omission are not abstract:

- **Energy and nuclear infrastructure.** Maintenance procedures, technical specifications, and safety protocols are processed through DLA pipelines. A missed safety warning, an omitted procedural step, or an undetected annotation in a technical drawing can propagate through maintenance workflows with potentially catastrophic consequences [Leveson, 2011].
- **Pharmaceutical and medical compliance.** Drug approval dossiers, clinical trial reports, and patient records undergo automated layout analysis for regulatory review. Missing a contraindication table or a dosage specification is not a metric degradation—it is a patient safety failure.
- **Legal and financial document processing.** Contract analysis, insurance claim processing, and financial auditing rely on complete extraction. An omitted clause, a missed liability table, or an undetected footnote with material conditions can result in significant financial and legal exposure.
- **Archival and cultural heritage.** As demonstrated by Beyene and Dancy [2025], the analysis of historical document collections without ground truth introduces specific completeness challenges, where the absence of supervision compounds the absence of detection guarantees.

In each of these domains, the **industrial process** downstream of document analysis imposes specific requirements on completeness that current metrics cannot express. The same layout analysis result may be acceptable for a bibliometric survey but dangerously incomplete for a nuclear safety audit. This observation leads to a central thesis of this paper: *completeness is not an intrinsic property of a detection output, but a process-relative judgment that must be calibrated to the stakes of the downstream application.*

The European AI Act [European Parliament and Council, 2024] reinforces this perspective by establishing risk-based classification of AI systems, explicitly requiring that high-risk AI applications—which include many document processing scenarios in regulated industries—demonstrate appropriate levels of accuracy, robustness, and transparency. We argue that **completeness assurance** is an implicit but critical dimension of these requirements that the field has not yet adequately addressed.

1.3 AI Ubiquity and the Imperative to Neutralize Uncertainty

Artificial intelligence is no longer a research curiosity deployed in controlled laboratory settings. It pervades industrial document chains at scale: from invoice processing in small enterprises to regulatory compliance pipelines in multinational organizations. This ubiquity creates

a compounding effect: a systematic blind spot in layout analysis, replicated across millions of documents, becomes a systemic risk.

The problem is that **uncertainty about omissions is invisible by design**. A user reviewing a layout analysis output sees what was detected. The undetected regions are, by definition, invisible—they appear as blank space, indistinguishable from genuinely empty areas. Unlike a misclassification, which produces a visible (and potentially questionable) result, an omission produces *nothing*. This asymmetry between visible errors and invisible omissions is the fundamental challenge that motivates this work.

We argue that it is **essential to quantify and bound the uncertainty** inherent in AI-driven document analysis. Not by eliminating uncertainty—which is epistemologically impossible—but by making it *measurable, bounded, and actionable*. The framework proposed in this paper is a step toward that goal.

1.4 Contributions and Paper Organization

This paper makes the following contributions:

1. We formalize the **problem of layout completeness** as distinct from layout detection accuracy, and argue for its industrial necessity (Section 2).
2. We propose an **ontology of document completeness** organized around three abstraction layers and a taxonomy of unknowns adapted from epistemic uncertainty theory (Section 3).
3. We introduce the **Completeness Confidence Index (CCI)**, a framework that aggregates three independent proof vectors to estimate completeness probability, with mathematical formalization (Section 4).
4. We provide a **concrete algorithmic specification and reference implementation** of Ghost OCR, with an experiment protocol for empirical validation (Section 5).
5. We discuss the implications of **process-dependent risk calibration** and the path from detection to certification (Section 6).
6. We formalize a **feedback loop** in which evidence of omission is reinjected into the layout detection pipeline to propose corrections, and analyze the **model access requirements** (white-box vs. black-box) that this entails (Section 6.4).
7. We call for the creation of an **"Omission Challenge" benchmark** specifically designed to evaluate completeness rather than detection accuracy (Section 7).

2 Related Work and Limitations of Current Metrics

2.1 Standard Metrics and Their Blind Spots

The evaluation of DLA models has been dominated by adaptations of object detection metrics. The mAP metric, inherited from PASCAL VOC [Everingham et al., 2010] and COCO [Lin et al., 2014], computes average precision across IoU thresholds for each element category. While mAP

elegantly captures the trade-off between precision and recall for detected elements, it harbors a critical limitation: **recall is computed against a ground truth that is itself assumed to be complete**.

This creates a circular dependency: the metric can only measure omissions relative to a reference that is presumed exhaustive. Yet, as Pfitzmann et al. [2022] demonstrate through the DocLayNet dataset, real-world ground truth annotations are neither complete nor consistent. Their study of 80,863 manually annotated pages reveals substantial inter-annotator variability, ambiguous layout boundaries, and inherent subjectivity in what constitutes a "layout element"—all of which conspire to make the ground truth itself an uncertain reference. Metric scores can vary significantly depending on annotation choices, calling into question the reliability of comparative evaluations.

Furthermore, precision-recall curves treat all errors symmetrically: a false positive and a false negative have equal weight. In the completeness framework we propose, this symmetry is fundamentally broken—**the cost of a false negative (omission) is process-dependent and often asymmetric with respect to false positives**.

2.2 Error Diagnosis and Robustness

Recent work has begun to move beyond aggregate metrics toward structured error analysis. The LED benchmark [Heo et al., 2025] introduces a diagnostic framework that categorizes structural errors in layout analysis: boundary errors, classification confusion, fragmentation, and merging artifacts. This taxonomy is valuable because it reveals that "incorrect detection" is not monolithic—different structural errors have different downstream consequences.

However, LED focuses on *diagnosing errors in what was detected*, not on *detecting what was missed*. Its error categories implicitly assume that the model has produced some output for every relevant region. The case where a region is simply absent from the output—the pure omission—falls outside the diagnostic framework.

The RoDLA benchmark [Chen et al., 2024] addresses a complementary concern: model robustness under perturbation. By testing DLA models against corrupted inputs (noise, blur, geometric distortions), RoDLA reveals that state-of-the-art models can suffer dramatic performance degradation under conditions that are common in real-world document processing (poor scans, mobile captures, aged documents). Critically, Chen et al. [2024] show that robustness failures are not uniform—certain element types are more vulnerable to specific perturbations, and these vulnerabilities are poorly predicted by standard mAP scores.

This finding directly supports our argument: a model may achieve high mAP on clean data while systematically omitting elements under the degraded conditions that characterize industrial document workflows. **Robustness failures are, in essence, conditional omission patterns that current metrics fail to surface.**

2.3 Evaluation Without Ground Truth

A particularly relevant line of research explores evaluation when complete ground truth is unavailable. Beyene and Dancy [2025] tackle the challenge of processing historical Black digital

archives, where no annotated ground truth exists. Their approach to unsupervised evaluation—using cross-modal signals between layout detection and OCR output—prefigures one of the proof vectors we formalize in the CCI framework: **cross-modal redundancy as a proxy for completeness**.

The insight is powerful: if the layout model claims a region is empty but the OCR engine detects text in that region, there is evidence of an omission. Conversely, if both modalities agree on emptiness, confidence in completeness increases. This cross-validation principle, which Beyene and Dancy [2025] deploy pragmatically, can be elevated to a formal component of a completeness framework.

2.4 Uncertainty Quantification and Unknown Unknowns

The broader machine learning community has developed substantial tools for uncertainty quantification. Bayesian approaches [Gal and Ghahramani, 2016], deep ensembles [Lakshminarayanan et al., 2017], and conformal prediction [Angelopoulos and Bates, 2021] provide mathematically grounded frameworks for expressing model uncertainty.

Angelopoulos and Bates [2021] introduce conformal prediction as a distribution-free method for constructing prediction sets with guaranteed coverage. Applied to DLA, conformal prediction could provide *coverage guarantees* over layout elements: rather than producing a single set of detections, the model would produce a set of *plausible completions*, with statistical guarantees that the true layout falls within this set with probability $1 - \alpha$.

Liang et al. [2023] directly address the detection of unknown objects in computer vision, introducing a Generalized Object Confidence (GOC) score that quantifies the likelihood that a detected region corresponds to a known versus unknown category. Their key insight—that models can be trained to recognize the *boundaries of their own knowledge*—is foundational to the residual signal vector of our CCI framework, where we extend the principle from unknown object *detection* to unknown object *absence estimation*.

Finally, Wang et al. [2020] demonstrate that explicitly accounting for *unannotated* relationships—elements that exist but are absent from the training signal—significantly improves structured prediction in scene graph generation. Their work reveals that presuming unannotated entity pairs as "not related" introduces systematic false negatives that bias models toward under-detection. We extend this principle to document layout: **modeling the expected absence of content in certain regions, rather than treating it as confirmed emptiness, strengthens the completeness assessment of detected regions**.

3 Ontology of Document Completeness

3.1 Three Abstraction Layers

We propose that document completeness must be evaluated across three distinct abstraction layers, each providing independent evidence about the exhaustiveness of the analysis.

Geometric Layer (Bounding Boxes). This is the layer at which most DLA models operate. It concerns the spatial coverage of detected regions: are all areas of the document page that

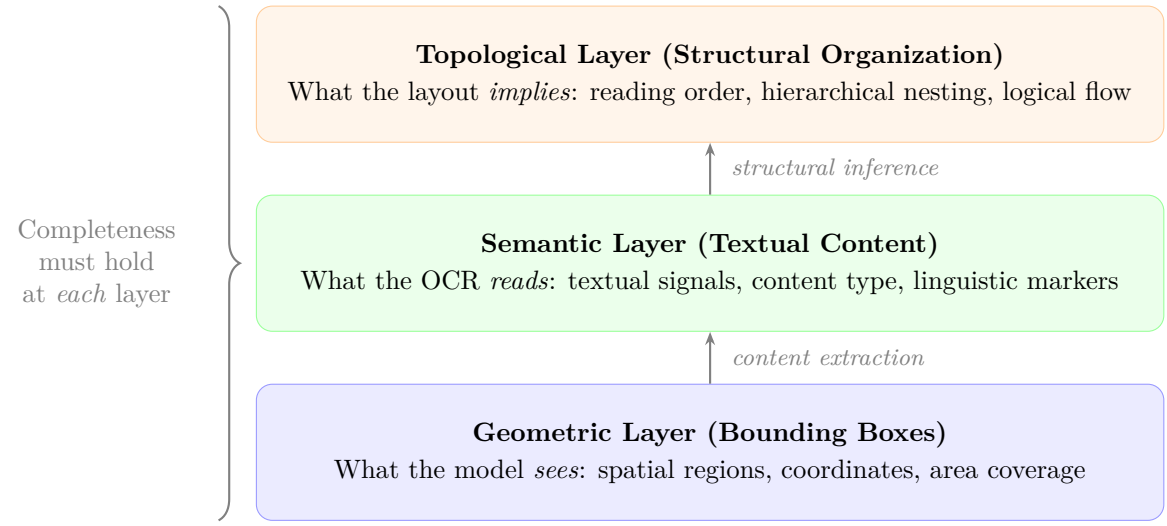

Figure 1: Three abstraction layers of document completeness. A document may be geometrically complete (all regions detected) but semantically incomplete (OCR missed text within a region) or topologically incomplete (reading order is broken). True completeness requires convergence across all three layers.

contain meaningful content covered by at least one bounding box? The geometric layer answers the question: *what does the model see?*

A critical operational parameter at this layer is the **geometric tolerance** δ. Because text characters have finite spatial extent and OCR engines report glyph positions at the pixel level, a bounding box that misses a text line by even a few pixels will fail to "capture" it. Through empirical observation across representative document corpora—spanning typographic conventions from dense academic papers to scanned industrial manuals—we adopt a tolerance of $\delta = 10$ pixels as the operational threshold for text-level completeness. This value reflects the typical half-height of a text line at standard scanning resolutions (200–300 DPI): a bounding box whose boundary lies within 10 pixels of a glyph center is considered to cover that glyph, while a gap exceeding δ constitutes a potential omission. Formally, we define the **effective coverage region** of a detected bounding box R_i as its δ-erosion:

$$R_i^{-\delta} = \{x \in R_i : d(x, \partial R_i) \geq \delta\} \tag{1}$$

where $d(x, \partial R_i)$ is the distance from point x to the boundary of R_i. Text falling within R_i but outside $R_i^{-\delta}$—i.e., within the δ-margin of the bounding box boundary—is in a **boundary uncertainty zone** where geometric coverage is ambiguous. This erosion-based formalization ensures that completeness assessment accounts for the finite precision of bounding box coordinates relative to the spatial granularity of textual content.

Semantic Layer (Textual Content). Beyond spatial coverage, the semantic layer concerns the content within detected regions. An OCR engine operating on detected regions produces textual output that can be compared against signals from the raw document image. The semantic layer answers: *what does the OCR read, and does it account for all readable content?*

Topological Layer (Structural Organization). The topological layer concerns the logical structure of the document: reading order, hierarchical nesting (sections, subsections), cross-references, and the overall narrative flow. As Wang et al. [2024] demonstrate, reading order prediction is tightly coupled with layout analysis, and structural coherence provides independent evidence about completeness. The topological layer answers: *does the detected structure form a coherent, complete whole?*

A critical insight is that **completeness at one layer does not guarantee completeness at another**. A document may be geometrically complete (all regions detected) but topologically incomplete (the detected regions do not form a coherent structure because a connecting element is missing). This multi-layer perspective is essential for robust completeness assessment.

3.2 Taxonomy of the Unknown in Document Analysis

Adapting the epistemological framework formalized by Pawson et al. [2011] for evidence-based policy—and widely known through its earlier popularization [Rumsfeld, 2011]—we propose a four-quadrant taxonomy of knowledge states specific to document layout analysis:

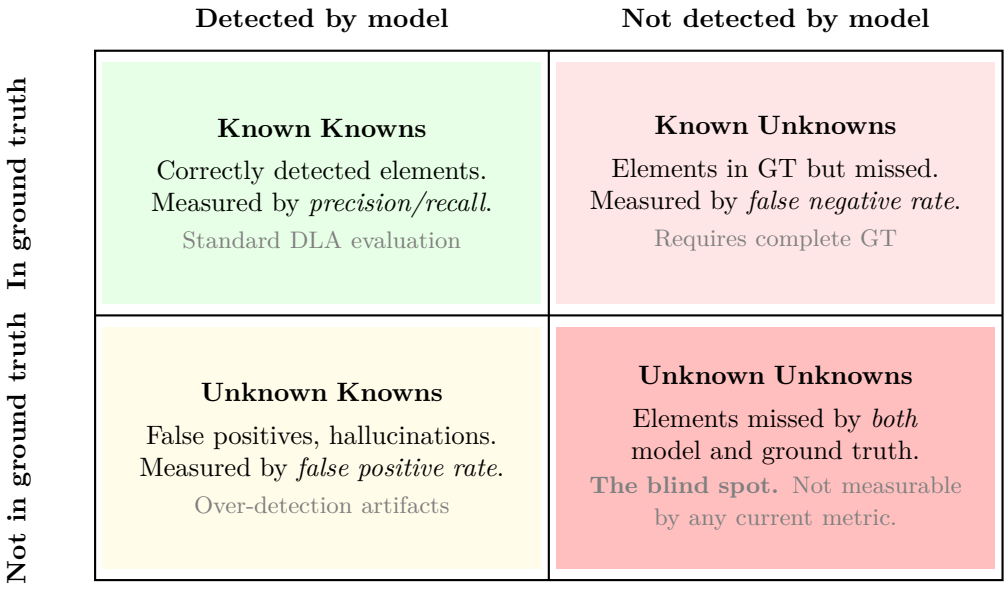

Figure 2: Taxonomy of knowledge states in document layout analysis. Standard metrics operate in the "Known Knowns" quadrant. The CCI framework aims to provide evidence about the "Unknown Unknowns" quadrant—the structurally invisible omissions.

Definition 1 (Unknown Unknowns in DLA)**.** *Let D be a document, $\mathcal{G}$ the (unknown) set of all semantically relevant regions in D, $\hat{\mathcal{G}}$ the annotated ground truth, and $\mathcal{M}$ the model output. The set of* ***unknown unknowns*** *is:*

$$\mathcal{U}^2 = \mathcal{G} \setminus (\hat{\mathcal{G}} \cup \mathcal{M})$$

These are regions that are semantically relevant but appear in neither the ground truth nor the model output. By definition, $|\mathcal{U}^2|$ cannot be directly measured—it can only be estimated *through indirect evidence.*

The existence of unknown unknowns is not hypothetical. Annotation studies consistently show that different annotators produce different ground truths for the same document, and that complex layouts (multi-column, nested tables, marginal notes) are particularly prone to incomplete annotation [Pfitzmann et al., 2022]. The unknown unknowns are not exotic edge cases—they are a structural feature of the DLA evaluation paradigm.

3.3 Theory of the Informative Void

We introduce the concept of the **informative void**: the principle that the regions of a document page *not* assigned to any detected element carry information about the completeness of the detection.

Definition 2 (Residual Space)**.** *Let Ω denote the document page domain and $\mathcal{M} = \{R_1, R_2, \ldots, R_n\}$ the set of detected regions. The **residual space** is:*

$$\Omega_{res} = \Omega \setminus \bigcup_{i=1}^{n} R_i$$

The residual space is the set of all points on the page not covered by any detected bounding box.

*To account for the finite spatial precision required for text-level analysis, we further define the **conservative residual space**:*

$$\Omega_{res}^{+\delta} = \Omega \setminus \bigcup_{i=1}^{n} R_i^{-\delta}$$

where $R_i^{-\delta}$ is the δ-eroded bounding box (Equation 1) with $\delta = 10$ px. The conservative residual space expands the residual by including the boundary uncertainty zones of all detected regions, providing a more stringent basis for completeness assessment. In all subsequent analysis, Ghost OCR and residual entropy computations operate on $\Omega_{res}^{+\delta}$ rather than Ω_{res} to ensure that text near bounding box boundaries is not falsely assumed to be captured.

The key insight is that Ω_{res} is not uniformly uninformative. It can be decomposed into:

1. **Structurally expected void**: margins, inter-element spacing, column gutters. These are regions where the absence of content is predicted by the document's typographic grammar.
2. **Ambiguous void**: regions where content could plausibly exist but is not detected. The informational status of these regions is uncertain.
3. **Anomalous void**: regions where structural or textual signals suggest content should be present but none was detected. These are the primary candidates for unknown unknowns.

The CCI framework (Section 4) formalizes the analysis of Ω_{res} through residual entropy computation, cross-modal interrogation, and structural constraint checking to distinguish between these three categories.

4 The Completeness Confidence Index (CCI)

4.1 Framework Overview

The Completeness Confidence Index is a composite score $\mathrm{CCI}(D, \mathcal{M}) \in [0, 1]$ that estimates the probability that a layout analysis $\mathcal{M}$ of document D has captured all semantically relevant regions. It aggregates three independent **proof vectors**, each providing complementary evidence about completeness:

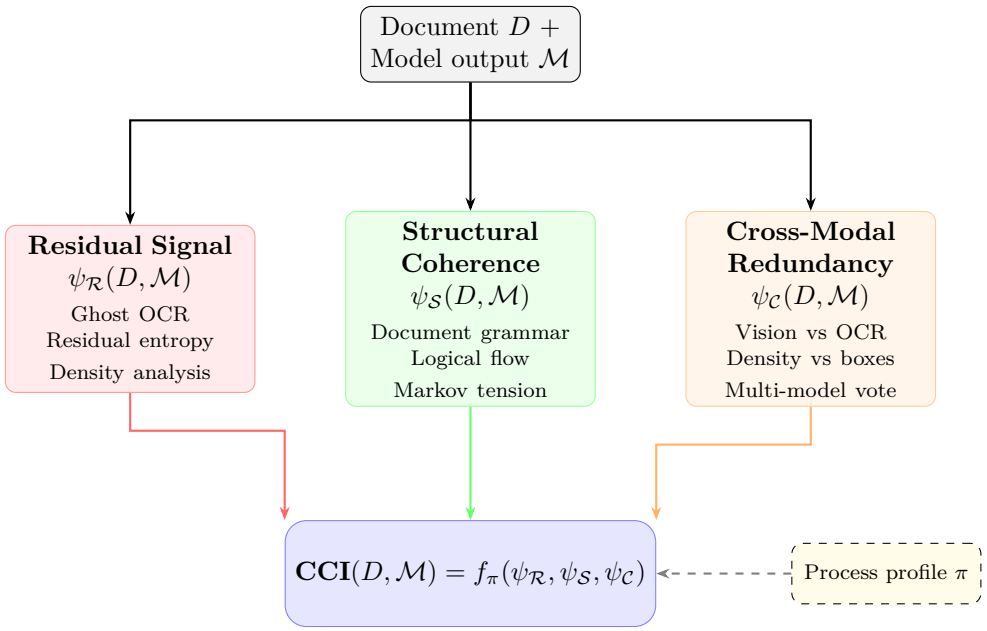

Figure 3: Architecture of the Completeness Confidence Index (CCI). Three independent proof vectors are aggregated through a process-calibrated function f_π that weights each vector according to the downstream industrial requirements.

Definition 3 (Completeness Confidence Index)**.** *Given a document D, a model output $\mathcal{M}$, and a process profile π, the CCI is defined as:*

$$CCI(D, \mathcal{M}; \pi) = f_\pi(\psi_{\mathcal{R}}(D, \mathcal{M}),\ \psi_{\mathcal{S}}(D, \mathcal{M}),\ \psi_{\mathcal{C}}(D, \mathcal{M})) \tag{2}$$

where $\psi_{\mathcal{R}} \in [0, 1]$ is the residual signal score, $\psi_{\mathcal{S}} \in [0, 1]$ is the structural coherence score, $\psi_{\mathcal{C}} \in [0, 1]$ is the cross-modal redundancy score, and $f_\pi : [0, 1]^3 \to [0, 1]$ is a monotone aggregation function parameterized by the process profile π.

The process profile π encodes the downstream application's tolerance for incompleteness. A high-criticality process (e.g., nuclear safety documentation) would assign greater weight to the structural coherence vector, while a bibliometric analysis might weight cross-modal redundancy more heavily. We discuss process calibration in Section 4.5.

A natural instantiation of f_π is the weighted geometric mean:

$$\mathrm{CCI}(D, \mathcal{M}; \pi) = \psi_{\mathcal{R}}^{\alpha_\pi} \cdot \psi_{\mathcal{S}}^{\beta_\pi} \cdot \psi_{\mathcal{C}}^{\gamma_\pi} \tag{3}$$

where $\alpha_\pi + \beta_\pi + \gamma_\pi = 1$ and $\alpha_\pi, \beta_\pi, \gamma_\pi > 0$. The geometric mean ensures that a near-zero

score on any single vector drives the overall CCI toward zero—a desirable property, since strong evidence of omission from *any* source should reduce completeness confidence regardless of the other vectors.

Justification and alternatives. The choice of aggregation function is not unique. Several alternatives exhibit the same "veto" property:

- **Product t-norm** (Schweizer–Sklar family): $T_P(a, b) = a \cdot b$. Equivalent to the unweighted geometric mean when applied to three arguments. The geometric mean generalizes this to weighted exponents, providing process-dependent flexibility.
- **Minimum t-norm** (Łukasiewicz): $T_{\min}(a, b, c) = \min(a, b, c)$. Maximally conservative—the CCI equals its weakest vector. While appealing for safety-critical applications, this discards all information from the other two vectors.
- **Copula-based aggregation**: Clayton or Frank copulas can model non-trivial dependency structures between the vectors, accommodating scenarios where correlated failures reduce the effective information gain (see Section 6.5). However, copula calibration requires empirical data on joint vector distributions that is not yet available.

We adopt the weighted geometric mean as the default for three reasons: (i) it provides a smooth interpolation between the product and minimum t-norms depending on the weight distribution; (ii) its monotonicity in each argument is analytically transparent; and (iii) it requires only three interpretable parameters $(\alpha_\pi, \beta_\pi, \gamma_\pi)$ that map directly to process priorities. Future empirical work should compare these alternatives on real document corpora to determine whether the additional modeling flexibility of copulas justifies their calibration cost.

4.2 Proof Vector 1: Residual Signal Analysis ($\psi_{\mathcal{R}}$)

The residual signal vector quantifies the informational content of the residual space Ω_{res} (Definition 2). The guiding principle is: *if the residual space contains signals that indicate the presence of undetected content, completeness confidence should decrease.*

Ghost OCR. We propose applying OCR to the conservative residual space—a technique we term **Ghost OCR**. Any text detected in $\Omega_{\text{res}}^{+\delta}$ constitutes direct evidence of omission: there exists readable content that no detected region reliably covers, accounting for the $\delta = 10$ px geometric tolerance.

Let $T_{\text{ghost}} = \text{OCR}(\Omega_{\text{res}}^{+\delta})$ denote the text extracted from the conservative residual space. We define:

$$\rho_{\text{ghost}}(D, \mathcal{M}) = 1 - \frac{|T_{\text{ghost}}|}{|T_{\text{ghost}}| + |T_{\mathcal{M}}|} \tag{4}$$

where $|T_{\text{ghost}}|$ is the volume of ghost text (e.g., character count) and $|T_{\mathcal{M}}|$ is the total text detected within the δ-eroded model output regions $\{R_i^{-\delta}\}$. When $\rho_{\text{ghost}} = 1$, no ghost text was found; as $\rho_{\text{ghost}} \to 0$, the volume of unaccounted text approaches that of detected text. The use of $\Omega_{\text{res}}^{+\delta}$ rather than Ω_{res} ensures that text glyphs located near bounding box edges—within the 10-pixel boundary uncertainty zone—are treated as potentially uncaptured, preventing false confidence from imprecise box coordinates.

Residual Entropy. Following the intuition of unknown object detection [Liang et al., 2023], we compute the entropy of visual features in Ω_{res}. A truly empty region (background, margins) should have low entropy—uniform color, no texture variation. Regions with high visual entropy in Ω_{res} suggest the presence of undetected content.

Let $p(x)$ denote the pixel-level feature distribution in Ω_{res}. The **residual entropy** is:

$$H_{\text{res}}(D, \mathcal{M}) = -\sum_{x \in \Omega_{\text{res}}} p(x) \log p(x) \tag{5}$$

This can be normalized by the maximum possible entropy to yield a score in $[0, 1]$:

$$\rho_{\text{entropy}}(D, \mathcal{M}) = 1 - \frac{H_{\text{res}}(D, \mathcal{M})}{H_{\max}(\Omega_{\text{res}})} \tag{6}$$

The residual vector aggregates these signals:

$$\psi_{\mathcal{R}}(D, \mathcal{M}) = \rho_{\text{ghost}} \cdot \rho_{\text{entropy}} \tag{7}$$

Connection to Physics-Informed Models. The residual analysis is conceptually analogous to physics-informed neural networks (PINNs) [Raissi et al., 2019], where physical laws constrain the solution space. In our framework, the "physics" of documents—typographic conventions, ink distribution patterns, spatial regularity—provide constraints that the residual space should satisfy if the detection is complete. Violations of these constraints (e.g., high-entropy patches in expected margin areas) signal potential omissions.

4.3 Proof Vector 2: Structural Coherence ($\psi_{\mathcal{S}}$)

The structural coherence vector evaluates whether the detected layout elements form a logically complete and consistent structure according to the document's expected grammar.

Probabilistic Document Grammar. We model the expected structure of a document class as a probabilistic context-free grammar (PCFG) [Manning and Schütze, 1999]:

$$G = (N, \Sigma, P, S) \tag{8}$$

where N is the set of non-terminal symbols (document sections, logical blocks), Σ is the set of terminal symbols (layout element types: title, paragraph, table, figure, caption, header, footer, page number, etc.), P is the set of probabilistic production rules, and S is the start symbol (document).

For example, a typical academic paper might have the production rules:

$$\begin{aligned} \text{Document} &\to \text{Header}\ \ \text{Body}\ \ \text{Footer} \quad [0.95] \\ \text{Body} &\to \text{Title}\ \ \text{Abstract}\ \ \text{Sections} \quad [0.90] \\ \text{Section} &\to \text{SectionTitle}\ \ \text{Paragraphs}\ \ \text{Figures}^* \quad [0.85] \end{aligned}$$

The structural coherence score measures how well the detected elements can be parsed by

this grammar:

$$\psi_{\mathcal{S}}(D, \mathcal{M}) = P(\mathcal{M} \text{ is a valid derivation of } G) \tag{9}$$

Beyond context-free grammars. We acknowledge that PCFGs may be insufficient for complex industrial documents (multi-column technical manuals, regulatory dossiers with deeply nested cross-references) whose structure is not context-free. Two extensions merit consideration. First, **graph grammars** can model non-sequential structural dependencies—e.g., a figure caption referencing a table in a different column, or cross-page continuations—that escape the sequential parsing assumption. Second, **neural structure prediction** models, such as the transformer-based approach of Wang et al. [2024] that jointly predicts layout regions and reading order, offer a data-driven alternative to handcrafted rules. In this setting, the structural coherence score $\psi_{\mathcal{S}}$ would be derived from the model's own confidence in the predicted structure, rather than from explicit grammar compliance. We retain the PCFG formulation as the simplest instantiation that illustrates the principle, while noting that operational deployment would likely require these richer structural models.

Markov Chain Model of Logical Flow. At a finer granularity, we model the sequential flow of layout elements as a first-order Markov chain. Let $e_1, e_2, \ldots, e_k$ be the sequence of detected elements in reading order [Wang et al., 2024]. The transition probability $P(e_{i+1}|e_i)$ captures expected sequential patterns. For instance, $P(\text{Caption}|\text{Figure})$ is expected to be high, while $P(\text{PageNumber}|\text{Title})$ should be near zero in the middle of a document.

A **structural tension** at position i occurs when:

$$\tau_i = -\log P(e_{i+1}|e_i) \tag{10}$$

is anomalously high, suggesting that a mediating element between e_i and e_{i+1} may have been omitted (Figure 4).

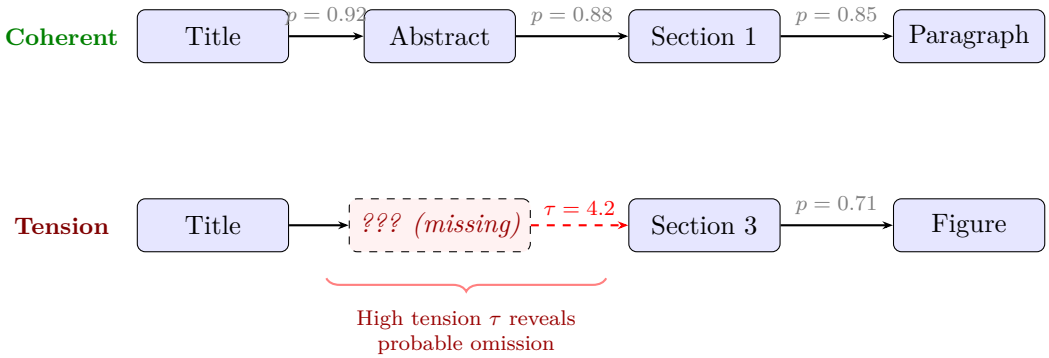

Figure 4: Logical flow analysis. Top: a coherent sequence with high transition probabilities. Bottom: a structural tension at the gap between "Title" and "Section 3" signals that intermediate elements (Abstract, Sections 1–2) are likely missing.

The global structural coherence score can be defined as:

$$\psi_{\mathcal{S}}(D, \mathcal{M}) = \exp\left(-\frac{1}{k-1}\sum_{i=1}^{k-1}\max(0, \tau_i - \bar{\tau})\right) \tag{11}$$

where $\bar{\tau}$ is the expected tension under a complete layout, acting as a baseline. This formulation ensures that the score is close to 1 when all transitions are expected and decreases exponentially with anomalous tensions.

4.4 Proof Vector 3: Cross-Modal Redundancy ($\psi_{\mathcal{C}}$)

The cross-modal redundancy vector exploits the fact that a document can be analyzed through multiple independent modalities, each providing a distinct "view" of the content. Disagreements between modalities signal potential omissions.

Modalities. We consider three primary modalities:

1. **Visual (V)**: the layout detection model's output, operating on the document image.
2. **Textual (T)**: OCR output, which detects readable text independently of layout structure.
3. **Density (D)**: pixel-level density analysis (ink coverage, gradient magnitude), which detects the presence of visual content without semantic interpretation.

Cross-Modal Agreement. For each region r of the document page, we define binary presence indicators. To ensure consistent spatial alignment across modalities, all region comparisons apply the geometric tolerance $\delta = 10$ px: visual layout regions are δ-eroded before comparison, and OCR glyph positions must fall within the eroded region to count as "covered." Formally:

$$v(r) = \mathbb{1}[r \in \mathcal{M}_V^{-\delta}], \quad t(r) = \mathbb{1}[\text{OCR}(r) \neq \emptyset], \quad d(r) = \mathbb{1}[\text{density}(r) > \theta_d] \tag{12}$$

The **agreement score** for a region is:

$$a(r) = \frac{v(r) + t(r) + d(r)}{3} \tag{13}$$

A region where $a(r) = 1$ (all modalities agree on presence) or $a(r) = 0$ (all agree on absence) provides strong evidence. A region where modalities disagree ($0 < a(r) < 1$) is a candidate for omission or hallucination.

The cross-modal redundancy score aggregates over the entire page:

$$\psi_{\mathcal{C}}(D, \mathcal{M}) = 1 - \frac{|\{r : a(r) \in (0, 1)\}|}{|\{r : a(r) > 0\}|} \tag{14}$$

This measures the fraction of content-bearing regions on which all modalities agree, penalizing disagreements.

Connection to Panoptic Segmentation. The multi-modal approach has conceptual parallels with document panoptic segmentation [Cao et al., 2021], which aims to assign every pixel to either a "thing" (discrete layout element) or "stuff" (background). The panoptic quality metric already requires accounting for all pixels, but treats unassigned regions as background by default. Our framework challenges this default: **an unassigned region is not necessarily background; it may be an undetected "thing"**.

4.5 Index Aggregation and Process Calibration

Independence Assumption. A key design principle of the CCI is that the three proof vectors are **intentionally designed to be as independent as possible**. The residual vector operates on pixel-level signals in undetected regions; the structural vector operates on the sequence and grammar of detected elements; the cross-modal vector operates on agreement between distinct analysis pipelines. This independence ensures that the CCI is robust: a failure or bias in one vector does not automatically corrupt the others.

Process Profiles. We define a **process profile** $\pi = (\alpha_\pi, \beta_\pi, \gamma_\pi, \theta_\pi)$ where:

- $\alpha_\pi, \beta_\pi, \gamma_\pi$ are the vector weights (summing to 1),
- $\theta_\pi \in [0, 1]$ is the **completeness acceptance threshold**: the minimum CCI value for the document to be considered "certified complete" for this process.

Table 1: Illustrative process profiles showing how the CCI framework adapts to different industrial contexts. Weights reflect the relative importance of each proof vector for the given application. These values are hypothetical and intended to demonstrate the framework's configurability; empirical calibration on domain-specific corpora would be required to establish operational profiles.

Process	α_π (Residual)	β_π (Structural)	γ_π (Cross-modal)	θ_π (Threshold)
Nuclear safety docs	0.30	0.45	0.25	0.97
Pharmaceutical dossiers	0.25	0.40	0.35	0.95
Legal contract review	0.35	0.30	0.35	0.93
Financial auditing	0.30	0.35	0.35	0.92
Bibliometric analysis	0.40	0.25	0.35	0.80
Archival digitization	0.45	0.20	0.35	0.75

Table 1 illustrates how different industrial contexts naturally lead to different CCI configurations. Nuclear safety documentation demands the highest threshold ($\theta_\pi = 0.97$) and heavily weights structural coherence, because a missing procedural step in a safety protocol is detectable primarily through violations of the expected document grammar. Legal contract review, by contrast, weights the residual vector more heavily, because the risk lies in textual content (fine-print clauses, footnotes) that may exist outside detected regions.

Conformal Calibration. Following Angelopoulos and Bates [2021], the CCI can be endowed with statistical guarantees through conformal prediction. Given a calibration set of documents with known completeness labels, conformal calibration produces a threshold $\hat{\theta}$ such that:

$$P\Big(\text{CCI}(D, \mathcal{M}; \pi) \geq \hat{\theta} \implies D \text{ is complete}\Big) \geq 1 - \alpha \tag{15}$$

for a user-specified confidence level $1 - \alpha$. This transforms the CCI from a heuristic score into a **statistically grounded certification tool**.

5 Ghost OCR: Algorithm and Proof of Feasibility

While this paper is primarily a position paper, we provide here a concrete algorithmic specification of the Ghost OCR procedure and a reference implementation to demonstrate its feasibility.

5.1 Algorithm

Algorithm 1 describes the complete Ghost OCR pipeline, from input document to residual signal score $\psi_{\mathcal{R}}$.

Algorithm 1 Ghost OCR — Residual Signal Analysis

Require: Document page image I of size $H \times W$, detected bounding boxes $\mathcal{M} = \{R_1, \ldots, R_n\}$, geometric tolerance δ, OCR engine $\mathcal{O}$
Ensure: Residual signal score $\psi_{\mathcal{R}} \in [0, 1]$, ghost text T_{ghost}

Phase 1: Conservative residual mask

$M \leftarrow$ matrix of ones, size $H \times W$ ▷ 255 = residual
for each $R_i = (x_1, y_1, x_2, y_2) \in \mathcal{M}$ **do**
  $R_i^{-\delta} \leftarrow (x_1 + \delta,\ y_1 + \delta,\ x_2 - \delta,\ y_2 - \delta)$ ▷ δ-erosion
  $M[R_i^{-\delta}] \leftarrow 0$ ▷ Mark as covered
end for

Phase 2: Ghost OCR

$I_{\text{masked}} \leftarrow$ copy of I
$I_{\text{masked}}[M = 0] \leftarrow$ white ▷ Blank covered regions
$T_{\text{ghost}} \leftarrow \mathcal{O}(I_{\text{masked}})$ ▷ OCR on residual only
$n_{\text{ghost}} \leftarrow |\text{chars}(T_{\text{ghost}})|$

Phase 3: Residual entropy

$\mathbf{p} \leftarrow \text{histogram}(I_{\text{gray}}[M = 1], \text{bins} = 256)$ ▷ Pixel distribution in residual
$H_{\text{res}} \leftarrow -\sum_i p_i \log_2 p_i$
$\rho_{\text{entropy}} \leftarrow 1 - H_{\text{res}} / \log_2(256)$

Phase 4: Score computation

$n_{\text{det}} \leftarrow \text{estimate\_chars}(\mathcal{M})$ ▷ Estimated chars in detected regions
$\rho_{\text{ghost}} \leftarrow 1 - n_{\text{ghost}} / (n_{\text{ghost}} + n_{\text{det}})$
$\psi_{\mathcal{R}} \leftarrow \rho_{\text{ghost}} \times \rho_{\text{entropy}}$
return $\psi_{\mathcal{R}}$, T_{ghost}

5.2 Reference Implementation

We provide an open-source Python implementation of Algorithm 1, available at:

https://github.com/rasata/dla_cci

The implementation uses Tesseract 4+ as the OCR engine, OpenCV for image processing, and NumPy for entropy computation. The complete pipeline—residual mask construction, ghost OCR, entropy computation, and score aggregation—is implemented in approximately 300 lines of Python and requires no GPU. On a commodity laptop (Apple M-series, no GPU), a single page at 200 DPI is processed in 2–4 seconds, dominated by the Tesseract OCR pass. The repository also includes a `-remove` option to simulate controlled omissions, enabling the experiment protocol described below. The code is released under a non-commercial research license; commercial use requires written authorization from the authors.

5.3 Illustrative Analysis

To demonstrate that Ghost OCR can detect real omissions, we describe a simple experiment protocol applicable to any publicly available DLA dataset (e.g., PubLayNet [Zhong et al., 2019], DocLayNet [Pfitzmann et al., 2022]):

1. **Select** k document pages with known complete ground truth annotations.
2. **Simulate omissions** by deliberately removing m ground truth bounding boxes from each page, creating controlled "unknown unknowns."
3. **Run Ghost OCR** on the degraded layout (with m boxes removed). The ghost text T_{ghost} should recover text from the removed regions.
4. **Measure detection rate**: what fraction of the deliberately omitted regions are flagged by ghost text? What is the false positive rate (ghost text in genuinely empty areas)?

This protocol directly tests the core claim of the CCI framework: that residual signal analysis can provide evidence of omission. Even a modest detection rate (e.g., recovering 60–80% of deliberately omitted text regions) would validate the principle that the informative void carries actionable signal. We note that the reference implementation is fully functional and ready for such evaluation; we leave the systematic benchmark to a dedicated experimental paper, consistent with the position paper scope of this work.

Expected outcomes. Based on the properties of the algorithm, we anticipate the following behavior:

- **High sensitivity for text-bearing regions**: any omitted region containing readable text (paragraphs, captions, headers, footnotes) should produce a non-zero T_{ghost}, since Tesseract will detect the characters.
- **Lower sensitivity for non-text elements**: omitted figures, decorative elements, or separator lines may not produce ghost text but should still elevate ρ_{entropy} through their visual complexity.
- **Graceful degradation under noise**: on degraded scans, both the original layout model and Ghost OCR will struggle, but the residual entropy signal should remain informative as long as ink is visible.

6 Discussion: From Detection to Certification

6.1 Process-Dependent Risk and the Cost of Omission

The CCI framework embodies a fundamental conceptual shift: from evaluating *model performance* to evaluating *process fitness*. A model with 90% mAP may be perfectly adequate for one industrial process and dangerously insufficient for another—not because the model changed, but because the *consequences of the remaining 10%* are radically different.

This process-dependent perspective aligns with the risk-based approach of the European AI Act [European Parliament and Council, 2024], which classifies AI systems not by their technical accuracy but by the potential impact of their failures. In this framework, the relevant question is not "how good is the model?" but "**how confident can we be that this specific document, analyzed by this specific model, is complete enough for this specific process?**"

The industrial implications are significant:

- **Safety-critical processes** (energy, nuclear, aerospace) require near-certain completeness. An omitted safety procedure in a nuclear maintenance document is not a "missed detection"—it is a potential incident vector. In these contexts, the CCI must approach 1, and any unresolved structural tension should trigger human review.
- **Regulatory compliance processes** (pharmaceutical, financial, legal) require demonstrable completeness. Auditors and regulators need evidence that the AI system has "seen everything." The CCI framework provides this evidence in a structured, auditable form.
- **Operational processes** (invoice processing, mail sorting, customer service) tolerate higher omission rates but benefit from completeness awareness for exception handling and quality control.

The cost of omission can be formalized. Let $C_{\text{omit}}(e, \pi)$ be the cost of omitting element e in process π. The **expected omission risk** is:

$$\mathcal{R}(D, \mathcal{M}; \pi) = \sum_{e \in \mathcal{U}^2} P(e \text{ exists}) \cdot C_{\text{omit}}(e, \pi) \tag{16}$$

While $P(e \text{ exists})$ is unknown for true unknown unknowns, the CCI framework provides an upper bound by decomposing the evidence across its three vectors. A low CCI implies a higher upper bound on the expected omission risk, triggering appropriate mitigation actions.

6.2 Reliability vs. Performance: A Necessary Trade-off

The introduction of completeness assessment introduces a tension with the traditional performance optimization paradigm. A model optimized purely for mAP may learn to suppress low-confidence detections to improve precision—but each suppressed detection is a potential omission. The CCI framework makes this trade-off explicit: **a model that detects "too much" (lower precision) but leaves less residual signal may achieve a higher CCI than a conservative model with higher precision but more anomalous voids**.

This observation suggests that the training objectives for DLA models should be reconsidered in the context of completeness. The loss function should not only penalize false positives but should also penalize the creation of anomalous voids in the residual space. We leave the design of such training objectives to future work, but note that this represents a significant departure from current practice.

6.3 Towards AI-Driven Document Certification

The ultimate vision of the CCI framework is to enable **document certification**: a formal statement that a given document has been analyzed with quantifiable confidence in the completeness of the result. This is analogous to certification in other engineering domains: a bridge is not merely tested but *certified* to withstand specified loads; a pharmaceutical process is not merely monitored but *validated* to produce consistent outputs.

Document certification through CCI would involve:

1. **Process specification**: defining the completeness requirements (π) for the downstream application.
2. **Multi-vector analysis**: computing $\psi_{\mathcal{R}}$, $\psi_{\mathcal{S}}$, $\psi_{\mathcal{C}}$ for the document.
3. **Threshold evaluation**: comparing $\mathrm{CCI}(D, \mathcal{M}; \pi)$ against θ_π.
4. **Certification or escalation**: if $\mathrm{CCI} \geq \theta_\pi$, the document is certified; otherwise, it is flagged for human review with specific indicators of where completeness evidence is weakest.

This workflow transforms the DLA pipeline from a "best-effort detection" system into a "guaranteed-completeness-or-escalate" system—a fundamental shift that is necessary for the responsible deployment of AI in critical document workflows.

6.4 Closing the Loop: From Completeness Assessment to Layout Correction

The CCI framework, as described so far, operates as a *post-hoc* assessment: it evaluates the completeness of an existing layout analysis but does not modify it. However, a natural and powerful extension is to **feed the evidence of omission back into the processing chain** to propose corrections and refinements to the original layout.

The feedback principle. When the CCI identifies anomalous residual signals—ghost text in $\Omega_{\text{res}}^{+\delta}$, structural tensions in the Markov flow, cross-modal disagreements—these signals do not merely lower a confidence score. They carry *spatially localized, semantically typed* information about what is likely missing and where. A ghost OCR hit in a specific page region suggests a missing text block at those coordinates. A structural tension between a section title and a page number suggests missing body content. A cross-modal disagreement where density analysis detects ink but the layout model reports nothing suggests an undetected figure or table.

This evidence can be reformulated as **candidate layout proposals**: hypothetical bounding boxes, with tentative type labels, that would resolve the identified anomalies. These proposals can then be reinjected into the document analysis pipeline—either as additional detections to be validated, or as attention-guiding signals for a second-pass analysis (Figure 5).

The model access requirement. This feedback loop, however, introduces a **fundamental architectural dependency**: to propose meaningful layout adjustments, the correction mechanism must have access to the *algorithm* $\mathcal{A}$ and the *learned parameters* ϕ of the model that produced the original layout $\mathcal{M} = \mathcal{A}_\phi(D)$. Without this access, the feedback loop is limited to

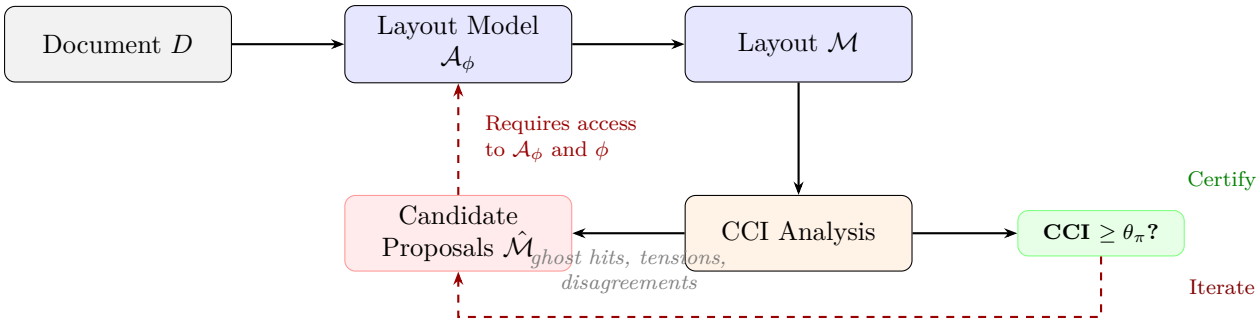

Figure 5: Feedback loop architecture. Evidence of omission gathered by the CCI is transformed into candidate layout proposals $\hat{\mathcal{M}}$ that are fed back into the layout model $\mathcal{A}_\phi$ for refinement. This closed loop iterates until CCI $\geq \theta_\pi$ or a maximum iteration count is reached. Critically, the feedback path (dashed arrows) requires access to the model's algorithm $\mathcal{A}$ and its learned parameters ϕ.

naïve strategies—e.g., inserting raw bounding boxes around ghost OCR hits—that ignore the model's internal representation and classification logic.

Access to $\mathcal{A}_\phi$ enables several qualitatively superior correction strategies:

1. **Guided re-inference.** The candidate regions identified by the CCI can be used to define *regions of interest* (ROIs) that are submitted to the model for targeted re-analysis. By constraining the model's attention to specific areas where omission evidence is strongest, guided re-inference achieves higher detection sensitivity than a full-page second pass, while limiting computational cost.

2. **Threshold relaxation.** Many DLA models apply a confidence threshold to suppress low-scoring detections. If the CCI identifies anomalous residual in a specific region, the model can be re-run on that region with a *relaxed* confidence threshold, recovering detections that were suppressed during the initial pass. This directly addresses the reliability-vs-performance trade-off discussed in the previous section.

3. **Feature-level injection.** At the deepest level, the candidate proposals can be injected into the model's intermediate feature maps as attention biases or anchor priors, nudging the detection head toward regions that the CCI has flagged. This requires white-box access to the model architecture but offers the most precise correction mechanism.

4. **Iterative convergence.** The feedback loop can be iterated: $\mathcal{M}^{(0)} \to \text{CCI} \to \hat{\mathcal{M}}^{(1)} \to \mathcal{A}_\phi(\hat{\mathcal{M}}^{(1)}) \to \mathcal{M}^{(1)} \to \text{CCI} \to \cdots$, until the CCI stabilizes above the process threshold θ_π or a maximum iteration count is reached. Convergence of this loop is not guaranteed in general but can be encouraged through monotonicity constraints on the CCI across iterations.

The black-box barrier. In practice, many industrial DLA deployments rely on third-party or proprietary models for which $\mathcal{A}$ and ϕ are not accessible. Cloud-based document processing APIs (e.g., commercial OCR and layout services) return layout outputs without exposing model internals. In such **black-box** settings, the CCI remains fully operational as a completeness *assessment* tool, but the feedback loop is constrained to external strategies only: submitting

cropped sub-images of flagged regions as independent queries, or ensembling the proprietary model's output with a secondary open-source model for cross-validation.

This observation has implications for the design of industrial document processing architectures. If completeness certification is a requirement—as we argue it should be for safety-critical processes—then **the choice of a DLA model is not merely a performance decision but an architectural one**: open models that permit feedback-driven correction offer a structurally stronger path to certification than black-box services that limit the pipeline to single-pass assessment.

6.5 Limitations and Open Questions

We acknowledge several limitations of the current framework:

- **The bootstrap problem.** The structural coherence vector requires a document grammar G, which must be defined or learned for each document class. For novel or heterogeneous document types, this grammar may be unavailable or unreliable.
- **The calibration problem.** Conformal calibration of the CCI requires a calibration set with known completeness labels. Creating such labels is itself a completeness problem, introducing a potential circularity.
- **The irreducible unknown.** No framework can guarantee detection of *all* unknown unknowns. A sufficiently unusual layout element, invisible to all modalities, will escape the CCI. The framework provides probabilistic bounds, not certainties.
- **Sensitivity to geometric tolerance δ.** The choice of $\delta = 10$ px as the boundary uncertainty margin is grounded in typical scanning resolutions and typographic dimensions, but it is not universal. Documents scanned at lower resolutions (e.g., 72 DPI) may require a proportionally smaller δ, while high-resolution scans (600 DPI) or documents with very small fonts may demand a larger tolerance. An adaptive δ calibrated to the document's effective DPI and font size distribution is a natural extension. A principled approach would define δ in physical units (e.g., $\delta \approx 0.5$ mm) and convert to pixels based on the document's actual DPI: $\delta_{\text{px}} = \delta_{\text{mm}} \times \text{DPI}/25.4$. At 200 DPI this yields ≈ 4 px; at 300 DPI, ≈ 6 px. Our default of 10 px at 200–300 DPI is deliberately conservative; a formal sensitivity analysis on representative corpora is needed to determine the optimal trade-off between over-counting boundary artifacts and missing genuine boundary omissions.
- **Computational cost.** The multi-vector analysis adds non-trivial computational overhead to the DLA pipeline. To give a rough order of magnitude: the residual mask computation is $O(n)$ in the number of detected regions (negligible). Ghost OCR requires one additional Tesseract pass over the residual image, adding approximately 1–3 seconds per page at 200–300 DPI on commodity hardware—comparable to the initial layout detection itself, yielding an overhead factor of approximately $\times 2$. Residual entropy computation is $O(|\Omega_{\text{res}}|)$ in the number of residual pixels and is typically sub-second. Cross-modal redundancy requires running a density analysis pass (< 1 s per page) in addition to the already-available layout and OCR outputs. In total, the CCI pipeline adds an estimated $\times 2$–$\times 3$ overhead relative

to a single-pass DLA pipeline. For real-time applications, this cost may be prohibitive; however, for batch-mode industrial processing where completeness certification is the goal, this overhead is likely acceptable. Selective application—computing the full CCI only for pages where a fast preliminary check (e.g., residual area ratio above a threshold) suggests potential omissions—can further reduce the amortized cost.

- **Inter-vector independence and correlated failures.** While the three vectors are designed to be independent, in practice they may share implicit biases that reduce the effective information gain from aggregation. The most important failure mode is *correlated degradation*: when a document is severely degraded (low-quality scan, heavy noise, poor contrast), all three modalities are simultaneously weakened. Ghost OCR produces more noise on degraded images, entropy analysis becomes unreliable when background texture is non-uniform, and structural coherence suffers because the layout model itself detects fewer elements. In such scenarios, the CCI may yield a misleadingly high score (all three vectors report "no anomaly") precisely because the degradation has rendered the evidence-gathering mechanisms themselves unreliable. This correlation of failures suggests that the CCI should be supplemented with an explicit *input quality assessment* step—estimating scan quality, resolution adequacy, and noise level—that modulates the confidence placed in the CCI score itself. A low input quality score should widen the uncertainty bounds on the CCI, even if the three vectors individually report high values.

- **Model access for feedback.** The correction feedback loop (Section 6.4) requires white-box or at minimum gray-box access to the layout detection model. In many industrial deployments, DLA is consumed as a black-box service, precluding the most effective correction strategies. The extent to which external-only feedback (cropped re-queries, ensemble augmentation) can approximate white-box correction remains an open empirical question.

7 Conclusion and Perspectives

7.1 Summary

This paper has argued that the field of Document Layout Analysis faces a fundamental blind spot: the inability to quantify confidence in the *completeness* of a layout analysis, as distinct from its *accuracy*. We have shown that this blind spot has concrete industrial consequences, particularly in safety-critical domains where AI-driven document processing is increasingly pervasive.

We have introduced the **Completeness Confidence Index (CCI)**, a framework that addresses this gap through three complementary proof vectors: residual signal analysis (interrogating the void left by detection), structural coherence validation (checking the logical grammar of the document), and cross-modal redundancy (exploiting agreement between independent analysis modalities). The framework is designed to be process-calibrated, reflecting the reality that completeness requirements depend on the downstream application.

7.2 Call for an "Omission Challenge" Benchmark

To advance this research agenda, we call for the creation of a dedicated benchmark—the **Omission Challenge**—specifically designed to evaluate completeness assessment methods. Such a benchmark would require:

1. **Documents with deliberate omissions**: ground truths from which specific elements have been systematically removed, creating controlled unknown unknowns.
2. **Process-specific evaluation tracks**: different tracks for different industrial contexts (safety-critical, regulatory, operational), each with appropriate evaluation criteria.
3. **Multi-annotator ground truths**: multiple independent annotations of the same documents, making the annotation uncertainty itself part of the evaluation.
4. **Degraded input variants**: following Chen et al. [2024], documents under realistic degradation conditions (poor scans, noise, partial occlusion) where omission is most likely.

This benchmark would complement existing DLA benchmarks (PubLayNet [Zhong et al., 2019], DocBank [Li et al., 2020]) by shifting the evaluation focus from "how well do you detect what is there?" to "**how confident are you that nothing is missing?**"

7.3 Future Research Directions

We identify several promising directions for future work:

- **Learning document grammars**: automated extraction of probabilistic document grammars from large corpora, reducing the dependence on manually specified rules for the structural coherence vector.
- **CCI-aware training**: designing loss functions that directly optimize for completeness (minimizing residual entropy and structural tension) in addition to detection accuracy.
- **Active completeness verification**: using the CCI to guide iterative refinement—regions where the CCI identifies low confidence could trigger targeted re-analysis with specialized models. The feedback loop formalized in Section 6.4 provides the architectural basis for this; future work should investigate convergence guarantees, optimal iteration strategies, and the minimal level of model access (white-box, gray-box, black-box) required to achieve a target CCI improvement per iteration.
- **Cross-document completeness**: extending the framework from single-page analysis to multi-page documents, where structural expectations span across pages.
- **Human-AI collaboration**: designing interfaces that present CCI information to human reviewers, enabling efficient targeted verification of low-confidence regions rather than exhaustive manual review.
- **Formal safety guarantees**: connecting the CCI to formal verification methods to provide provable bounds on omission probability under specified assumptions, suitable for the highest-criticality applications.

As artificial intelligence becomes the default tool for document analysis across industries, the question is no longer whether AI can detect layout elements—it demonstrably can, with impressive accuracy. The question that now demands our attention is whether we can *certify* that nothing was missed. The Completeness Confidence Index is a first step toward answering that question with mathematical rigor and industrial responsibility.

Acknowledgments

The author thanks **Prof. Emeritus Ioan Roxin** (Université de Franche-Comté, ORCID: 0000-0002-4143-4177) for reviewing this manuscript prior to preprint submission. His experience in multimedia information systems and knowledge organization contributed useful suggestions that strengthened the paper. The revisions are summarized in the peer review statement below.

Pre-Submission Peer Review Statement

This manuscript was independently reviewed by Prof. Emeritus Ioan Roxin (Université de Franche-Comté) prior to preprint submission. In the interest of transparency and scientific rigor, we summarize the reviewer's main observations and the corresponding revisions.

Reviewer assessment. The reviewer considered the paper a well-structured theoretical contribution that addresses a legitimate gap in the DLA evaluation landscape. The core contributions—Ghost OCR, the CCI tripartite architecture, the informative void taxonomy, and process-dependent calibration—were assessed as original. The reviewer noted the clarity of the exposition and the substance of the limitations section, while identifying several areas where the argument could be strengthened.

Modifications requested and implemented. The reviewer identified several areas for improvement. We list each point and the action taken:

1. **Absence of empirical validation, even minimal** (*critical*).
   Action: Added Section 5 ("Ghost OCR: Algorithm and Proof of Feasibility") containing a formal algorithmic specification (Algorithm 1), a reference Python implementation (`src/ghost_ocr_demo.py`), and a reproducible experiment protocol for validation on public DLA datasets.

2. **Insufficient discussion of the independence assumption** between the three proof vectors.
   Action: Expanded the limitation on inter-vector independence (Section 6.5) with a detailed analysis of correlated degradation failures and a proposal for an explicit input quality assessment step.

3. **Insufficient justification of the geometric mean** as aggregation function.
   Action: Added a comparative discussion (Section 4.5) of alternative aggregation functions—product t-norm, minimum t-norm, and copula-based aggregation—with a three-point justification for the geometric mean choice.

4. **PCFG under-dimensioned** for complex industrial documents.
   Action: Added a paragraph in Section 4.3 discussing graph grammars and neural structure prediction (transformer-based approaches) as extensions beyond context-free grammars.

5. **Sensitivity of the $\delta = 10\,\mathbf{px}$ parameter** not analyzed.
   Action: Expanded the δ sensitivity discussion in Section 6.5 with a principled DPI-based conversion formula and concrete values across scanning resolutions.

6. **No discussion of existing ad hoc industrial practices.**
   Action: Added a paragraph in the Introduction (Section 1) discussing double extraction, systematic human verification, and heuristic post-processing as existing practices that the CCI framework aims to supersede.

7. **Computational complexity not quantified.**
   Action: Added order-of-magnitude estimates in Section 6.5: $\times 2$–$\times 3$ overhead relative to single-pass DLA, with per-component breakdowns and a selective application strategy.

8. **Tone too emphatic** in several passages.
   Action: Moderated the formulations identified by the reviewer ("catastrophic" $\rightarrow$ "significant," "primordial to neutralize" $\rightarrow$ "essential to quantify and bound").

9. **Rumsfeld reference potentially divisive** in academic context.
   Action: Reordered the citation to foreground the academic source [Pawson et al., 2011] and relegate the popularization [Rumsfeld, 2011] to a secondary mention.

10. **"How to Cite" section unusual and presumptuous** for a preprint.
    Action: Section removed entirely.

Points acknowledged but deferred. The reviewer recommended a full empirical benchmark on PubLayNet or DocLayNet. While we provide a reference implementation and experiment protocol (Section 5), we defer the systematic experimental evaluation to a dedicated follow-up paper, consistent with the position paper scope of this work.

The author thanks Prof. Emeritus Roxin for his constructive feedback.

References

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Fitsum Sileshi Beyene and Christopher L. Dancy. Layout-aware OCR for black digital archives with unsupervised evaluation. *arXiv preprint arXiv:2509.13236*, 2025.

Ruoyu Cao, Hongliang Li, Guanghui Zhou, and Ping Luo. Towards document panoptic segmentation with pinpoint accuracy: Method and evaluation. In *International Conference on Document Analysis and Recognition (ICDAR)*, volume 12822 of *Lecture Notes in Computer Science*. Springer, 2021. doi: 10.1007/978-3-030-86331-9_1.

Yufan Chen, Jiaming Zhang, Kunyu Peng, Junwei Zheng, Ruiping Liu, Philip Torr, and Rainer Stiefelhagen. RoDLA: Benchmarking the robustness of document layout analysis models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. arXiv:2403.14442.

European Parliament and Council. Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (AI Act). *Official Journal of the European Union*, 2024. Entry into force: 1 August 2024. Public legislation, freely citable.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, volume 48, pages 1050–1059, 2016.

Inbum Heo, Taewook Hwang, Jeesu Jung, and Sangkeun Jung. LED benchmark: Diagnosing structural layout errors for document layout analysis. *arXiv preprint arXiv:2507.23295*, 2025. License: CC BY 4.0.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Nancy G. Leveson. *Engineering a Safer World: Systems Thinking Applied to Safety*. MIT Press, 2011. Open Access.

Minghao Li, Yiheng Xu, Lei Cui, Shaohan Huang, Furu Wei, Zhoujun Li, and Ming Zhou. DocBank: A benchmark dataset for document layout analysis. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pages 949–960, 2020.

Wenteng Liang, Feng Xue, Yihao Liu, Guofeng Zhong, and Anlong Ming. Unknown sniffer for object detection: Don't turn a blind eye to unknown objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. arXiv:2303.13769. First public benchmark for unknown object detection; introduces Generalized Object Confidence (GOC) score.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755, 2014.

Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

Ray Pawson, Geoff Wong, and Lesley Owen. Known knowns, known unknowns, unknown unknowns: The predicament of evidence-based policy. *American Journal of Evaluation*, 32(4): 518–546, 2011. doi: 10.1177/1098214011403831.

Birgit Pfitzmann, Christoph Auer, Michele Dolfi, Ahmed S. Nassar, and Peter W. J. Staar. DocLayNet: A large human-annotated dataset for document-layout segmentation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3743–3751, 2022. arXiv:2206.01062. Includes analysis of inter-annotator agreement and evaluation metric limitations for DLA.

Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.

Donald Rumsfeld. *Known and Unknown: A Memoir*. Sentinel (Penguin), 2011. The "known knowns" taxonomy originates from a Department of Defense press conference, February 12, 2002.

Jiawei Wang, Kai Hu, and Qiang Huo. DLAFormer: An end-to-end transformer for document layout analysis. In *International Conference on Document Analysis and Recognition (ICDAR)*, 2024. arXiv:2405.11757. Unifies text region detection, logical role classification, and reading order prediction.

Tzu-Jui Julius Wang, Selen Pehlivan, and Jorma Laaksonen. Tackling the unannotated: Scene graph generation with bias-reduced models. In *British Machine Vision Conference (BMVC)*, 2020. arXiv:2008.07832. Addresses missing/unannotated relationships in scene graphs and annotation false negatives.

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