

Article

EviCal: Evidence-Grounded Consistency Calibration for Content-Level Multimodal Labeling

Xiaofeng Zhang ^{1,*}, Baoli Han ¹, Yufeng Yuan ¹, Guangyao Zhu ¹, Huibo Song ¹, Weixing Qiu ¹ and Li Ni ²

¹ State Grid Zhejiang Electric Power Co., Ltd., Shaoxing Power Supply Company, Shaoxing 312021, China; hanbaoli_sgps@163.com (B.H.); 1261411522@qq.com (Y.Y.); zhuguangyao_sgps@163.com (G.Z.); songhuibo@outlook.com (H.S.); qiuweixing@qq.com (W.Q.)

² State Grid Zhejiang Electric Power Co., Ltd., Shengzhou Power Supply Company, Shengzhou 312499, China; 157107378@qq.com

* Correspondence: xf_zhang1241@163.com

Abstract

Power system testing and inspection documents are multimodal and highly structured, making content-level audit labeling challenging due to scattered evidence and cross-component dependencies. We propose EviCal, an evidence-grounded consistency calibration framework under a predefined label space. EviCal decomposes documents into atomic units (text segments, table rows, and figure captions), grounds each label to minimal supporting evidence via label-aware semantic focusing, calibrates local decisions against global causal and logical constraints imposed on symbolic intermediate states, and produces explicit confidence estimates. Experiments on two real-world power-system datasets show that EviCal achieves up to 93.97% accuracy and 81.22 F1, and attains a human score of up to 4.58/5, outperforming strong multimodal baselines and delivering more reliable, interpretable audit predictions.

Keywords: multimodal document understanding; evidence grounding; consistency calibration; content-level labeling; confidence estimation; power system auditing

1. Introduction

Power system testing and inspection documents play a critical role in engineering auditing by recording equipment conditions, test parameters, operational findings, and remedial actions [1]. Because these documents are routinely produced at scale during operation and maintenance, their accurate and consistent analysis is essential for system reliability and regulatory compliance. Such documents are inherently multimodal and structured, comprising natural language descriptions, tabular records, and illustrative figures [2,3]. Their heterogeneous content units, domain-specific expressions, and explicit logical dependencies across document components make reliable and interpretable content-level audit labeling particularly challenging.

Document understanding has recently shifted from traditional encoder-based models [2,4] toward generative Large Multimodal Models (LMMs). By reducing reliance on intermediate processing pipelines, modern end-to-end approaches directly map raw multimodal inputs to structured sequences or free-form conclusions [5]. These models effectively combine visual, textual, and layout cues. Leveraging the reasoning capabilities of large language models (LLMs), they can handle diverse document formats and capture long-range dependencies across textual and visual content [6,7].



Received: 20 March 2026

Revised: 22 May 2026

Accepted: 29 May 2026

Published: 11 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Despite their success, existing generative document understanding approaches exhibit fundamental limitations when applied to audit-oriented, multimodal, and structured document analysis. First, they often struggle to maintain decision stability and inference efficiency. By relying on end-to-end autoregressive generation, these methods may overlook the structured nature of audit labels and the explicit logical dependencies among document components, leading to inefficient inference and predictions that are sensitive to local or spurious cues. Second, they usually lack explicit and traceable evidence grounding. As a result, generated audit conclusions may deviate semantically from the source content, while providing limited transparency regarding which text segments, table entries, or figures support a given decision. Moreover, the absence of explicit confidence estimation further hinders reliable risk assessment, limiting its practical applicability in engineering auditing scenarios.

To address these challenges, we propose EviCal, an evidence-grounded consistency calibration framework for Content-level Multimodal Labeling. Instead of directly generating audit conclusions, EviCal formulates document analysis as a structured decision-making process under a predefined label space. The framework explicitly grounds each audit label to supporting document content, refines local decisions through global causal and logical consistency calibration, and provides an explicit confidence estimate for each output. By decoupling content understanding from global reasoning and restricting LLM involvement to structured intermediate representations, EviCal improves decision stability and reliability while avoiding uncontrolled generative inference.

Our main contributions are summarized as follows:

- We propose EviCal, an evidence-grounded consistency calibration framework for content-level multimodal labeling in audit-oriented document analysis, enabling structured decision-making under predefined label constraints.
- We introduce a novel consistency calibration mechanism that integrates explicit evidence grounding with global causal and logical verification, improving decision stability, interpretability, and reliability without relying on end-to-end generative inference.
- We evaluate EviCal on two real-world power-system audit datasets, where EviCal achieves up to 93.97% accuracy and 81.22 F1, and attains a human score of up to 4.58/5, outperforming strong multimodal baselines.

2. Related Work

2.1. Document Understanding and Extraction

Document information extraction aims to identify and retrieve key elements from unstructured or semi-structured documents, with a central challenge lying in the integration of textual semantics and visual layout cues.

Early document pre-trained models mainly relied on encoder-based architectures and the Masked Language Modeling (MLM) objective. Huang et al. introduced LayoutLMv3 [2], which unifies text and image masking during pretraining and demonstrates the benefit of incorporating visual modality into document representation learning. Li et al. further extended this line of work with StructuralLM [4] by introducing cell-level representations and structural pretraining objectives, enabling better modeling of fine-grained spatial relations in forms and tables. Despite their success, these Bidirectional Encoder Representations from Transformers (BERT)-style approaches still depend heavily on external Optical Character Recognition (OCR) systems and are less suitable for end-to-end generation-based information extraction.

Large generative models have further advanced end-to-end document understanding with document-oriented multimodal language models (Doc-MLLMs). Kim et al. proposed Donut [8], eliminating OCR preprocessing and directly generating structured sequences

from document images for OCR-free extraction. However, scaling such end-to-end frameworks to high-resolution and dense-layout documents remains challenging. To strengthen layout-aware reasoning while avoiding the expense of visual encoders, Wang et al. presented DocLLM [5], which incorporates bounding-box coordinates as a spatial modality and introduces disentangled spatial attention to improve text-layout alignment and spatial reasoning in dense document scenarios.

For hybrid documents containing charts and complex visuals, recent studies emphasize stronger fine-grained perception and better vision-language alignment. Liu et al. proposed TextMonkey [7], leveraging Shifted Window Attention and a multi-scale visual encoder to enhance small-text reading under high-resolution document inputs. Ye et al. designed mPLUG-DocOwl [9], improving OCR-free document understanding via modular connectors and unified instruction tuning over diverse document-related tasks, including charts and tables. In addition, Zhang et al. explored unified vision-language instruction tuning with LLaVAR [10], while Tang et al. introduced UDOP [6] as a universal document foundation model that unifies vision, text, and layout under a sequence-to-sequence pretraining and prompting paradigm.

Nevertheless, these methods remain limited when handling long domain-specific documents because they often emphasize single-page alignment, struggle with cross-page dependencies, and cannot reliably fuse unstructured text, tables, and diagrams within a unified semantic space. These limitations may lead to missing key information or semantic misalignment across modalities.

2.2. AI for Power-System Document Understanding and Labeling

In recent years, in addition to advances in general document understanding, artificial intelligence techniques have increasingly been applied to the understanding, labeling, and structured processing of domain-specific texts and operational records in the power sector. Existing studies in this area mainly focus on text labeling and information extraction for power-domain documents, defect-oriented knowledge extraction and knowledge organization, as well as visual recognition and multimodal modeling for inspection scenarios.

A prominent line of research focuses on text labeling and information extraction in the power domain. For example, Luo et al. proposed a federated named entity recognition model for power grids, which improves recognition performance through collaborative training across multiple platforms while preserving data privacy by avoiding the direct sharing of raw data [11]. Tang et al. leveraged a Universal Information Extraction (UIE) framework to support both named entity recognition and event extraction from power grid outage scheduling documents [12]. Furthermore, Meng et al. focused on power violation description texts and introduced a word-character fusion mechanism combined with multi-head attention to improve the recognition of domain-specific terms and entity boundaries [13]. Finally, Chen et al. explored the use of LLMs for the automatic extraction of bibliographic metadata from power grid standard documents, promoting the transformation of unstructured standard documents into structured metadata [14].

Another active research direction addresses defect-text knowledge extraction and knowledge organization. For example, Yao et al. proposed a joint extraction model based on BERT and an entity-enhanced graph convolutional network. This model addresses the issues of nested entities and unclear boundaries in transformer defect texts, and builds a domain knowledge graph for intelligent troubleshooting [15]. Xiong et al. presented a defect identification method for secondary equipment that combines knowledge graphs with Bayesian networks. By converting unstructured defect texts into structured graph nodes, this method achieves accurate probabilistic reasoning for complex defect phenomena and multi-level causes [16]. Liu et al. designed a deep learning framework based on

ALBERT-BiLSTM-Attention-CRF. By integrating a power domain-specific dictionary for multi-source feature embedding, it effectively separates key knowledge from data noise in fault texts [17]. Additionally, recent work by Liu et al. [18] developed a framework for constructing and completing an automation equipment defect knowledge graph based on RoBERTa and a knowledge graph attention network. This enables the efficient extraction, storage, and visual analysis of defect knowledge for power grid regulation equipment [18].

In parallel, a growing body of work concentrates on visual recognition and multi-modal modeling for power inspection scenarios. To effectively leverage heterogeneous sensing modalities, Zhang et al. proposed a lightweight infrared-visible fusion framework, demonstrating that the integration of thermal and visual information significantly enhances robustness in object detection tasks [19]. Similarly, Guo et al. introduced a decoupled scene-equipment fusion strategy for substation imagery, which improves cross-modal representation learning by independently modeling global scene context and equipment-specific thermal-texture characteristics [20]. Beyond low-level sensor fusion, recent studies have increasingly explored vision-language models (VLMs) to incorporate semantic knowledge into defect detection pipelines. For instance, the MK-DETR approach leverages the CLIP model to inject textual priors into the detection process, thereby improving fault identification accuracy in transmission line inspection [21]. Building upon this paradigm, Zhong et al. proposed PowerGPT, a domain-specific multimodal large language model that unifies perception, localization, and domain-aware reasoning within a single framework, supported by large-scale instruction tuning and a dedicated evaluation benchmark [22]. Furthermore, Wang et al. developed Power-LLaVA, a conversational vision-language assistant for transmission line inspection, which achieves robust multimodal understanding and reliable human-machine interaction through a two-stage training strategy [23].

Overall, existing studies provide an important foundation for content understanding and labeling in the power domain, with methodological paradigms gradually evolving from traditional supervised learning to LLM-based and multimodal large-model approaches. However, most prior work still focuses on text-sequence-level or image-level tasks, or on relatively coarse-grained knowledge understanding and analysis. Content-level labeling for multimodal and structured power system testing and inspection documents therefore remains underexplored, particularly with respect to evidence alignment, consistency calibration, and confidence estimation.

2.3. Causal and Logical Reasoning in LLMs

LLMs demonstrate strong fluency in natural language generation, but their reasoning behavior is still largely driven by statistical correlations, which limits their ability to perform rigorous logical and causal reasoning.

To enhance LLMs' reasoning capabilities, Wei et al. proposed Chain of Thought (CoT) [24], which improves performance on complex problems by explicitly generating intermediate reasoning steps. Building on this idea, Yao et al. introduced Tree of Thoughts [25], which formulates reasoning as a tree-structured search process that enables backtracking and exploration across multiple reasoning paths. To address the instability caused by randomness in the reasoning process, Wang et al. proposed Self-Consistency [26], which improves robustness by sampling multiple reasoning paths and selecting results through majority voting.

However, correlation is not equivalent to causation. Chi et al. argued that LLMs' causal reasoning is often shallow and primarily derived from causal knowledge embedded in model parameters. They reported significant performance drops on newly constructed "fresh" benchmarks, especially in counterfactual settings [27]. To incorporate causal constraints, Kiciman et al. benchmarked LLMs on causal tasks such as causal discovery and

counterfactual reasoning, and discussed combining LLMs with formal causal tools to improve reliability [28]. Meanwhile, Jin et al. introduced the CLADDER benchmark, which evaluates causal reasoning ability through formalized causal query tasks, and provides a standardized benchmark for causal evaluation of LLMs [29].

In terms of integrating structured knowledge and enforcing logical constraints, Pan et al. proposed a framework that collaborates LLMs with knowledge graphs as external resources, using structured knowledge to support reasoning and generation [30]. Yang et al. introduced Verifier-Guided Search [31], which generates natural language proofs to improve the logical rigor of generated content. Liu et al. systematically reviewed the intersection of causal inference and LLMs and emphasized the importance of introducing structured causal knowledge when applying LLMs to causal reasoning tasks [32].

Although techniques such as CoT improve general reasoning, existing logic enhancement methods are primarily designed for mathematical or commonsense settings. They are therefore difficult to adapt directly to the strict “phenomenon–cause–solution” causal chains required in engineering domains. Moreover, current approaches mainly rely on prompt engineering or post hoc reranking and lack explicit mechanisms for enforcing causal consistency during generation, which can result in fluent but logically inconsistent outputs, such as mismatches between remedial actions and defect types.

2.4. Hallucination Mitigation and Trustworthy Auditing

In high-risk applications, the factuality and auditability of generated content are critical for real-world deployment, while hallucination remains a significant obstacle.

To evaluate and mitigate hallucinations, Gao et al. provided a comprehensive survey of retrieval-augmented generation (RAG) [33], which improves the accuracy and credibility of model outputs by incorporating external knowledge bases. Nakano et al. developed WebGPT and trained the model to generate answers with explicit citations collected during web browsing, facilitating human verification of factual accuracy [34]. Furthermore, Gao et al. proposed the ALCE benchmark [35], the first benchmark for automatic evaluation of LLM-generated text with citations. They introduced reproducible protocols and metrics that highlight the importance of citation quality.

For finer-grained factuality verification, Min et al. proposed FActScore [36], which decomposes long-form text into atomic facts and verifies each point individually, enabling quantitative auditing of generated content. Manakul et al. introduced SelfCheckGPT [37], a zero-resource method that detects hallucinations by measuring consistency across multiple stochastically sampled outputs, reducing reliance on labeled data. Kuhn et al. approached this problem from the perspective of semantic uncertainty by grouping semantically equivalent generations to estimate model confidence [38].

To enable closed-loop human-AI collaborative auditing, Ouyang et al. demonstrated the effectiveness of reinforcement learning from human feedback (RLHF) in InstructGPT [39]. Subsequently, Wang et al. proposed the Shepherd model [40], a critic model tuned to critique responses and suggest targeted refinements. Gao et al. introduced the RARR system [41], which performs post-generation retrieval, attribution, and revision to better align each statement with supporting evidence. Huang et al. further summarized recent progress on hallucination detection and mitigation in a survey, and discussed future directions toward more robust and trustworthy LLM systems [42].

Despite progress in citation-based generation and uncertainty estimation, gaps remain in fine-grained auditing for engineering documents. Existing evidence localization methods are often limited to paragraph- or webpage-level references and often do not trace claims back to minimal evidence units, such as specific table rows or figure captions, as required in engineering auditing settings. Moreover, many current auditing mechanisms are static and

one-directional, lacking a closed-loop design that incorporates human review outcomes as feedback signals. As a result, expert feedback is not sufficiently used to calibrate label confidence and iteratively improve model performance. To provide a clearer summary of the related work, Table 1 compares representative technique categories discussed in Sections 2.1–2.4 with the proposed EviCal framework.

Table 1. Comparison of major related work categories and the proposed EviCal framework.

Method Category	Representative Focus	Main Limitation	Gap Addressed by EviCal
Document understanding methods [2,4,5,7,8]	Layout-aware document parsing and multimodal representation.	Task-agnostic outputs with limited audit-level consistency control.	Atomic evidence grounding linked to predefined audit labels.
Power-domain labeling methods [11–13,15,22,23]	Domain-specific entity extraction, defect analysis, and inspection understanding.	Mostly text/image-level or task-specific processing.	Content-level multimodal labeling with structured audit decisions.
LLM reasoning methods [24–26,28,29]	Intermediate reasoning, verification, and structured constraints.	Often prompt-based or post hoc, with weak audit-specific control.	Symbolic audit-state calibration under predefined causal/logical rules.
Trustworthy auditing methods [33–38]	Retrieval, citation, uncertainty estimation, and factual verification.	Coarse evidence units and limited engineering-audit granularity.	Confidence-aware auditing grounded in sentences, table rows, and figure captions.

3. Methodology

We propose EviCal, an evidence-grounded consistency calibration framework for Content-level Multimodal Labeling in audit-oriented document analysis, as illustrated in Figure 1. Under a predefined label space, EviCal predicts label-specific audit decisions with explicit evidence grounding and global consistency calibration. The framework consists of three sequential modules: (i) Multimodal Document Understanding for evidence-aware local prediction, (ii) Causal and Logical Verification for global consistency calibration under predefined constraints, and (iii) Confidence Quantification for reliability estimation.

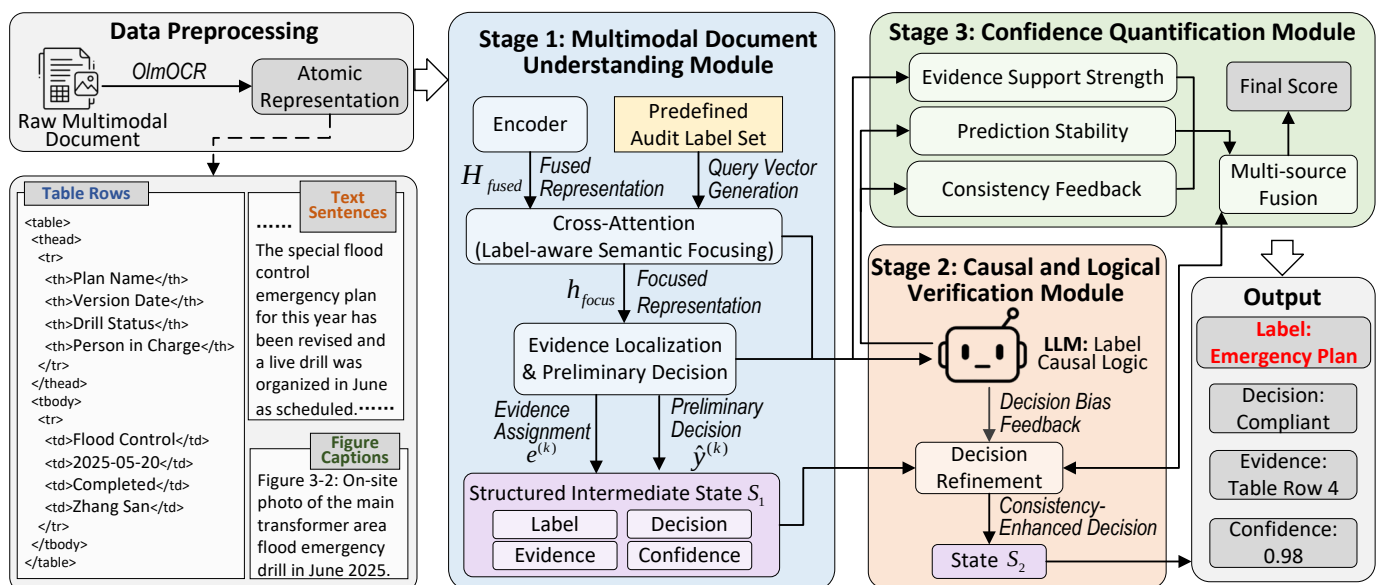


Figure 1. Framework overview of EviCal for content-level multimodal labeling. Arrows indicate the processing flow among modules, and colors distinguish document inputs, intermediate representations, and calibrated decision outputs.

3.1. Task Definition

We formalize Content-level Multimodal Labeling as follows. Given a multimodal document $D = \{T, G, I\}$, where $T = \{t_j\}$ denotes textual segments (e.g., sentences or

phrases), $G = \{g_k\}$ denotes table rows, and $I = \{i_m\}$ denotes figure captions, we define the candidate content units as $U = T \cup G \cup I$. Let $\mathcal{L} = \{l_1, l_2, \dots, l_K\}$ be a predefined set of audit-oriented labels. The goal is to determine which units are relevant to each label and to predict a label-specific audit decision within the constrained label space.

For each activated label l_k , the model outputs a tuple consisting of: (1) an audit decision, (2) explicit supporting evidence grounded in U , and (3) a confidence score indicating prediction reliability.

3.2. Multimodal Document Understanding Module

The multimodal document understanding module serves as the first stage of the proposed framework. Rather than producing final audit conclusions, it aims to construct a structured, evidence-driven intermediate reasoning state from raw multimodal audit documents. This design enables subsequent causal and logical verification modules to operate without directly accessing the original document, thereby decoupling local semantic understanding from global consistency reasoning.

3.2.1. Atomic Representation and Cross-Modal Encoding

Given a multimodal document $D = \{T, G, I\}$, where T , G , and I denote textual segments, table rows, and figure captions respectively, we first decompose the document into the set of minimal semantic units U :

$$U = \{u_n\}_{n=1}^N = T \cup G \cup I. \quad (1)$$

Each semantic unit u_n is treated as a potential atomic evidence unit. To enable unified modeling across modalities, all semantic units are projected into a shared semantic space using a common encoder, yielding contextualized vector representations $\mathbf{h}_n$:

$$\mathbf{h}_n \in \mathbb{R}^d, \quad n = 1, \dots, N, \quad (2)$$

where $\mathbf{h}_n$ denotes the semantic representation of the n -th unit and d is the hidden dimension. Collectively, these representations form the fused document representation

$$H_{\text{fused}} = \{\mathbf{h}_n\}_{n=1}^N, \quad (3)$$

which serves as the semantic foundation for subsequent label-conditioned reasoning.

3.2.2. Label-Aware Semantic Focusing

Let $\mathcal{L} = \{l_1, l_2, \dots, l_K\}$ denote the predefined set of audit labels. Since different labels attend to distinct semantic patterns within the same document, we introduce a label-aware semantic focusing mechanism to construct label-conditional document representations.

Specifically, each audit label l_k is associated with a learnable query vector $\mathbf{q}_k \in \mathbb{R}^d$. Given the fused document representation H_{fused} , we compute a label-specific focused representation via cross-attention:

$$\mathbf{h}_k^{\text{focus}} = \text{Attn}(\mathbf{q}_k, H_{\text{fused}}), \quad (4)$$

where $\mathbf{h}_k^{\text{focus}}$ denotes the label-conditioned semantic representation that captures the document content most relevant to audit label l_k . This mechanism allows the model to form multiple, mutually independent semantic views over the same document, alleviating semantic interference in multi-label audit scenarios.

3.2.3. Evidence Assignment and Intermediate Decision State Construction

Based on the label-specific focused representation, the model explicitly associates each audit label with supporting semantic units to identify minimal evidence. For each label l_k and semantic unit u_n , a relevance score is computed as

$$s_{k,n} = \text{score}(\mathbf{h}_k^{\text{focus}}, \mathbf{h}_n), \quad (5)$$

where $s_{k,n}$ measures the semantic relevance between label l_k and unit u_n . These scores are normalized to obtain an evidence assignment distribution:

$$P(e_k = n) = \frac{\exp(s_{k,n})}{\sum_{j=1}^N \exp(s_{k,j})}, \quad (6)$$

where e_k denotes the selected minimal evidence unit for label l_k . The semantic unit with the highest probability is chosen as the evidence supporting the corresponding label.

Let $\mathbf{h}_k^{\text{evid}} = \mathbf{h}_{e_k}$ denote the representation of the selected evidence unit. Conditioned on both the focused representation and the evidence representation, the model produces a preliminary audit decision:

$$\hat{y}_k = f_{\text{cls}}(\mathbf{h}_k^{\text{focus}}, \mathbf{h}_k^{\text{evid}}), \quad (7)$$

where $\hat{y}_k$ represents a local, evidence-conditioned audit judgment for label l_k , rather than a globally consistent conclusion.

Finally, the multimodal document understanding module outputs a structured intermediate reasoning state:

$$\mathcal{S}_1 = \left\{ (l_k, \hat{y}_k, e_k, s_k, \mathbf{h}_k^{\text{focus}}, \mathbf{h}_k^{\text{evid}}) \right\}_{k=1}^K, \quad (8)$$

where $\mathcal{S}_1$ denotes the intermediate reasoning state produced by this module, and s_k represents the confidence signal associated with the evidence assignment for label l_k (e.g., $s_k = \max_n P(e_k = n)$ or the relevance score of the selected unit). This structured state serves as the input to subsequent causal and logical verification modules.

3.3. Causal and Logical Verification Module

The causal and logical verification module constitutes the second stage of the proposed framework. Its objective is to perform global consistency calibration over preliminary audit decisions by explicitly incorporating causal dependencies and logical constraints among audit labels. Importantly, this module operates purely on the structured intermediate reasoning state produced by the previous stage, without accessing the original document content, thereby avoiding semantic hallucination and preserving evidence traceability.

3.3.1. Structured Audit State Projection

The input to this module is the intermediate reasoning state $\mathcal{S}_1$ produced by the multimodal document understanding module.

To prevent LLMs from directly interacting with raw document content, we provide the LLM with a reduced symbolic projection of $\mathcal{S}_1$:

$$\mathcal{S}_1^{\text{LLM}} = \{(l_k, \hat{y}_k, e_k, s_k)\}_{k=1}^K. \quad (9)$$

This projected view preserves label identities, preliminary decisions, evidence bindings, and confidence signals, while excluding all unstructured textual or visual inputs. As a result, the LLM is constrained to reason solely over structured audit states and predefined logical relations.

3.3.2. LLM-Based Global Consistency Feedback

Conditioned on the structured audit state $\mathcal{S}_1^{\text{LLM}}$ and predefined causal and logical constraints among audit labels, the LLM performs a global consistency assessment and outputs a label-wise calibration signal:

$$\mathbf{b} = \{b_k\}_{k=1}^K, \quad b_k \in \mathbb{R}. \quad (10)$$

Here, b_k represents a global consistency feedback signal for audit label l_k , indicating whether the current preliminary decision $\hat{y}_k$ deviates from the overall causal or logical context implied by other labels. A positive value suggests that the current decision may be overly conservative, while a negative value indicates a potentially overly aggressive judgment. Notably, the LLM does not generate or modify audit decisions directly; instead, it provides directional calibration signals to guide subsequent decision refinement.

3.3.3. Consistency-Calibrated Decision Refinement

The calibration signals produced by the LLM are used to refine preliminary audit decisions in a controlled and interpretable manner. Specifically, for each audit label l_k , the final decision is obtained through a consistency-calibrated decision function:

$$\tilde{y}_k = f_{\text{cal}}(\mathbf{v}_k, b_k). \quad (11)$$

In this formulation, $\mathbf{v}_k$ denotes the joint semantic representation associated with label l_k , constructed from the focused representation $\mathbf{h}_k^{\text{focus}}$ and the evidence representation $\mathbf{h}_k^{\text{evid}}$ (e.g., concatenation or a learned fusion), and b_k serves as a conditioning signal for global consistency calibration. The calibration function $f_{\text{cal}}(\cdot)$ is designed such that when $b_k = 0$, it reduces to the original evidence-conditioned decision function, ensuring that the LLM feedback does not override local model predictions but adjusts the decision boundary in a principled manner.

The output of this module is a consistency-enhanced audit state:

$$\mathcal{S}_2 = \{(l_k, \tilde{y}_k, e_k, s_k, \mathbf{v}_k)\}_{k=1}^K, \quad (12)$$

where preliminary decisions have been globally calibrated while evidence bindings remain unchanged. This structured state serves as the input to the subsequent confidence quantification module.

3.4. Confidence Quantification Module

The confidence quantification module is the final stage of the proposed framework. Its purpose is to explicitly assess the reliability and uncertainty of each audit decision by aggregating multiple complementary signals derived from evidence grounding, prediction stability, and global consistency calibration. Instead of producing additional audit judgments, this module maps the consistency-enhanced audit state to a normalized confidence score, enabling risk-aware and interpretable audit outputs.

The confidence quantification module takes the consistency-enhanced audit state $\mathcal{S}_2$ as input. For each audit label l_k , we define a confidence score $s_k^{\text{conf}} \in [0, 1]$ as a weighted aggregation of three complementary confidence components:

$$s_k^{\text{conf}} = \sigma(w_1 C_k^{\text{evid}} + w_2 C_k^{\text{pred}} + w_3 C_k^{\text{llm}}), \quad (13)$$

where $\sigma(\cdot)$ denotes the sigmoid function, and w_1, w_2, w_3 are learnable scalar weights. Each component captures a distinct aspect of audit reliability, as detailed below.

3.4.1. Evidence Support Strength

The evidence support strength measures how concentrated the model's attention is when grounding an audit label to document evidence. Intuitively, a decision supported by a small number of highly relevant evidence units is considered more reliable than one relying on diffuse or ambiguous evidence.

Based on the evidence assignment distribution $P(e_k = n)$ produced by the multi-modal document understanding module, we define the evidence support strength using normalized entropy:

$$C_k^{\text{evid}} = 1 - \frac{1}{\log N} \sum_{n=1}^N P(e_k = n) \log P(e_k = n), \quad (14)$$

where N denotes the total number of semantic units in the document. This formulation yields higher confidence values when evidence attribution is more focused.

3.4.2. Prediction Stability

Prediction stability reflects the certainty of the calibrated audit decision. Let $\tilde{\mathbf{z}}_k$ denote the logits associated with the consistency-calibrated decision $\tilde{y}_k$, and let

$$\mathbf{p}_k = \text{Softmax}(\tilde{\mathbf{z}}_k) \quad (15)$$

be the corresponding predictive probability distribution. We quantify prediction stability as the maximum class probability:

$$C_k^{\text{pred}} = \max \mathbf{p}_k. \quad (16)$$

This measure captures the sharpness of the decision boundary after global consistency calibration.

3.4.3. Consistency Feedback Reliability

In the LLM-in-the-loop causal and logical verification module, each audit label l_k receives a structured consistency feedback signal in the form

$$\mathcal{F}_k = (r_k, c_k), \quad (17)$$

where $r_k \in \{\text{agree}, \text{disagree}, \text{uncertain}\}$ indicates the qualitative consistency assessment, and $c_k \in [0, 1]$ denotes the associated confidence level. This corresponds to the scalar calibration signal b_k in Section 3.3.2 via $b_k = c_k$, $-c_k$, or 0 when r_k is agree, disagree, or uncertain, respectively.

Based on this feedback, we define the consistency reliability component as:

$$C_k^{\text{llm}} = \begin{cases} +c_k, & r_k = \text{agree}, \\ 0, & r_k = \text{uncertain}, \\ -c_k, & r_k = \text{disagree}. \end{cases} \quad (18)$$

This design ensures that the LLM contributes directional consistency information without directly modifying audit decisions, while explicitly encoding uncertainty when the feedback is inconclusive.

Through the integration of these complementary signals, the confidence quantification module produces audit outputs that are not only evidence-grounded and globally consistent, but also accompanied by explicit and interpretable reliability estimates.

3.5. Training Objective

The final output of EviCal consists of consistency-calibrated audit results with explicit evidence grounding and confidence estimation. Based on document content, the system returns a set of output tuples:

$$\text{Out} = \left\{ \left(l_k, \tilde{y}_k, e_k, s_k^{\text{conf}} \right) \right\}_{k=1}^K, \quad (19)$$

where $k = 1, \dots, K$ indexes the label set $\mathcal{L}$, l_k denotes an audit-oriented label from the predefined label set $\mathcal{L}$, $\tilde{y}_k$ denotes the final audit decision refined through global consistency calibration, e_k is the traceable evidence unit grounded in document content, and $s_k^{\text{conf}} \in [0, 1]$ represents the confidence score.

The proposed framework is trained using supervised audit annotations, where supervision is applied to the final consistency-calibrated audit decisions. Let $y_k \in \{1, \dots, C\}$ denote the ground-truth audit class for label l_k , where C is the number of audit decision types, and let $\tilde{\mathbf{z}}_k \in \mathbb{R}^C$ denote the logits associated with the final prediction $\tilde{y}_k$ produced by the causal and logical verification module. The training objective is defined as:

$$\mathcal{L} = - \sum_{k=1}^K \log \frac{\exp(\tilde{z}_{k,y_k})}{\sum_{c=1}^C \exp(\tilde{z}_{k,c})} + \lambda \mathcal{R}, \quad (20)$$

where $\tilde{z}_{k,c}$ denotes the c -th element of the logit vector $\tilde{\mathbf{z}}_k$, $\lambda \geq 0$ is a balancing coefficient, and $\mathcal{R}$ denotes an auxiliary regularization term (e.g., evidence alignment or confidence calibration loss).

4. Experiments

4.1. Dataset Construction

4.1.1. Data Acquisition and Multimodal Preprocessing

To build a high-fidelity evaluation benchmark covering full power business scenarios, we collaborated with a provincial electric power company to collect massive desensitized unstructured documents from its internal Office Automation (OA) system, Production Management System (PMS), and Operation Management System (OMS). We divided them into two datasets with distinct characteristics:

General-Doc: This dataset focuses on general knowledge at the enterprise operation level. The data sources mainly include various management measures issued by the company, installation and maintenance manuals for software systems, annual compliance audit reports, and scanned copies of procurement contracts. These documents usually have standardized official formats, containing a large amount of administrative logic regarding process approval, permission management, and asset allocation.

Domain-Doc: This dataset focuses on the core domain knowledge of the power system. The data sources cover on-site disposal schemes for substations of different voltage levels, emergency plans for optical cable communication systems, daily operation logs from dispatch centers, and complex equipment fault analysis reports. These documents contain dense electrical parameters, topology diagrams, and strict operating procedures, placing extremely high demands on the model's domain reasoning ability.

In the data preprocessing stage, given the complex structures existing in the original Portable Document Format (PDF) and Word documents, traditional OCR tools struggle to preserve the semantic structure. We introduced the olmOCR [43] tool for deep parsing. This tool converts document pages into high-precision Markdown sequences via a visual encoder. It not only accurately restores text paragraphs but also converts complex nested

tables into standard HTML structures, while preserving image placeholders and their contextual positions. Subsequently, we developed automated cleaning scripts to remove meaningless symbols such as watermarks and page breaks, and segmented the content using natural paragraphs, independent images, and independent tables as the minimum semantic units. Each segmented unit was assigned a unique metadata index, achieving precise traceability from data fragments to the original documents.

4.1.2. Taxonomy and Human-AI Collaborative Annotation

To achieve a deep understanding of power documents, we referred to relevant standards in the electric power industry and formulated an annotation system containing dual-layer dimensions.

Audit Topics: We defined 11 specific audit topics which serve as the semantic anchors for evidence seeking, covering core dimensions from administrative compliance to production dispatch.

Audit Decisions: We introduced a verification mechanism requiring annotators to evaluate the content validity. This is divided into “Support” (evidence is sufficient and conclusive), “Oppose” (evidence has logical contradictions), and “Details Missing” (content is relevant but key parameters are missing).

The annotation process combined model-assisted pre-annotation with expert review. First, the GPT-5.2 model was used as a base agent to generate preliminary labels and confidence scores for large-scale data units through designed prompts. Subsequently, we formed an expert group consisting of three senior engineers, each with more than five years of experience in power grid operation and inspection. The experts conducted secondary verification for samples with model confidence below 0.6 and samples marked as “Details Missing” by the model. After three rounds of iterative cleaning, we constructed a high-quality domain-specific corpus for the electric power field. The details of these two datasets are shown in Table 2.

Table 2. Details of statistics in the General-Doc and Domain-Doc datasets. Kappa denotes inter-annotator agreement during corpus construction.

Dataset	Raw Documents	Processed Units	Key Categories	Kappa
General-Doc	109	18,192	Software Asset List, Compliance	0.68
Domain-Doc	96	24,328	Emergency Plan, Fault Handling	0.62

4.2. Baselines

Qwen2.5-VL [44] is a leading open-source multimodal model that is extensively optimized for long-document understanding and visual reasoning tasks. It natively supports complex document images, videos, and long-form texts, and captures fine-grained document details through an enhanced visual encoder. Notably, Qwen2.5-VL can directly interpret complex document layouts without relying on explicit OCR pipelines.

DeepSeek-V3 [45] adopts a Mixture-of-Experts architecture, exhibiting strong language reasoning and contextual modeling capabilities. For document-centric tasks, the model leverages pre-trained knowledge to infer implicit semantics within documents. This enables robust performance in logical reasoning and complex instruction following.

ChatGPT-4o [46] demonstrates advanced cross-modal alignment and multimodal fusion capabilities. It effectively integrates visual and textual signals to establish deep semantic associations across heterogeneous document components, achieving strong performance in long-context reasoning and information synthesis.

AVIR [47] focuses on adaptive visual information retrieval and understanding by dynamically adjusting visual attention weights to emphasize salient image features within documents. In multimodal auditing scenarios, AVIR is commonly used to improve alignment between visual content and corresponding captions, thereby enhancing the quality of visual evidence extraction.

Doc-CoB [48] is a method specifically designed for modeling the layout of business and engineering documents. It emphasizes leveraging the logical topological structure of documents to assist semantic prediction. By incorporating layout prior knowledge into the feature encoding process, it effectively addresses the issue of semantic fragmentation in traditional models when dealing with non-linear layouts.

ColPali [49] introduces a novel late interaction mechanism that performs document representation and matching directly on the outputs of the visual encoder. Unlike OCR-dependent approaches, ColPali preserves intrinsic visual semantics, including mathematical formulas, charts, and typographic emphasis, thereby enabling efficient retrieval while maintaining the original contextual relationships within document content.

4.3. Implementation Details

The proposed EviCal framework is implemented using PyTorch 2.9.0 and trained on two NVIDIA A5000 GPUs. In the Multimodal Document Understanding Module, a BERT encoder serves as the foundational semantic feature extractor. We utilize the olmOCR tool to parse raw multimodal documents into atomic semantic units, including text segments, table rows, and figure captions, which are subsequently projected into a unified shared latent space. To facilitate precise evidence retrieval, we employ a label-aware semantic focusing mechanism with $M = 4$ attention heads to capture distinct semantic views corresponding to 11 predefined audit topics.

For model optimization, the network is fine-tuned using the Adam optimizer with a learning rate of 2×10^{-5} and a weight decay of 1×10^{-5} over 20 training epochs. The Causal and Logical Verification Module utilizes the Qwen2.5-7B large language model as the primary reasoning engine to perform global consistency assessments on projected symbolic audit states. In our experimental setup, the logical calibration strength β is fixed at 1.0 to maintain an optimal equilibrium between local evidentiary strength and global logical constraints, ensuring that the model's outputs satisfy the rigorous transparency and stability requirements of engineering auditing.

4.4. Evaluation Metrics

To further evaluate the interpretability, logical consistency, and practical utility of the EviCal framework beyond standard classification metrics (Accuracy (Acc) and F1-score (F1)), we conduct a rigorous human evaluation. This assessment focuses on the quality of evidence grounding and the reliability of the decision-making process, which are critical in real-world engineering auditing scenarios.

Three experts with professional backgrounds in power system auditing were invited to perform a double-blind review of the outputs from EviCal and other representative baselines. The final comprehensive Human Score (S_{human}) is calculated as the arithmetic mean across all experts and dimensions:

$$S_{human} = \frac{1}{N \cdot M \cdot |V|} \sum_{i=1}^N \sum_{j=1}^M \sum_{v \in V} s_{i,j,v} \quad (21)$$

where $s_{i,j,v} \in \{1, \dots, 5\}$ denotes the score on a 5-point Likert scale (1-Poor to 5-Excellent) assigned to the i -th sample by the j -th expert on metric $v \in V$, where V represents the set of four fine-grained metrics: (1) Evidence Grounding Accuracy (EGA), measur-

ing the precision of e_k in supporting the decision; (2) Logical Consistency Score (LCS), assessing the logical derivation between e_k and stance $\tilde{y}_k$ to detect hallucinations; (3) Confidence Calibration Alignment (CCA), evaluating the consistency between s_k^{conf} and human-perceived decision difficulty; and (4) Audit Utility (AU), quantifying the system's value in reducing manual workload. To ensure the reliability of these ratings, we calculated Fleiss' Kappa to measure inter-rater agreement, achieving a score of 0.83, which indicates substantial agreement among the experts.

4.5. Main Results

We compared EviCal with current mainstream LLMs and advanced document understanding methods. Table 3 summarizes the results on both datasets.

Table 3. Main results (%) on General-Doc and Domain-Doc datasets. S_{human} denotes the comprehensive human evaluation score (1–5 scale).

Model	General-Doc			Domain-Doc		
	Acc	F1	S_{human}	Acc	F1	S_{human}
Qwen2.5-VL	89.34	72.47	3.42	82.18	69.41	2.85
DeepSeek-V3	88.24	74.32	3.55	80.56	70.12	2.94
ChatGPT-4o	90.55	77.12	3.82	83.12	72.45	3.15
AVIR	88.76	75.28	3.95	84.15	74.32	3.48
Doc-CoB	90.92	77.61	4.12	85.67	76.15	3.76
ColPali	85.10	71.04	3.31	79.45	68.10	2.92
EviCal (Ours)	93.97	81.22	4.58	88.71	79.52	4.42

Tip: Bold values indicate better performance.

The results show that general-purpose LLMs, including Qwen2.5-VL, DeepSeek-V3, and ChatGPT-4o, possess strong semantic understanding capabilities. However, when applied to the content-level labeling task defined in this work, they reveal certain inherent limitations. This arises primarily because general-purpose models typically employ open-ended autoregressive generation mechanisms, whereas audit tasks demand precise decision-making within a predefined label space. This paradigm mismatch makes it difficult for these models to consistently maintain alignment between conclusions and supporting evidence when processing power-system documents with stringent structural constraints and domain-specific terminology. Furthermore, without explicit logical consistency calibration, such models often treat audit labels as independent classification tasks. Consequently, when processing documents with implicit logical dependencies, they frequently encounter misattributed evidence or logical discontinuities, resulting in significantly lower S_{human} compared to specialized methods.

Document-focused methods such as AVIR, Doc-CoB, and ColPali exhibit more stable predictive performance by incorporating layout information or multimodal fusion strategies. Doc-CoB achieves strong baseline results through its detailed modeling of document layout structures. However, these methods are primarily designed for localized semantic extraction or single-point classification tasks. When tasks require collaborative reasoning across multiple audit labels and explicit evidence binding, end-to-end mapping mechanisms often struggle to maintain global consistency for labels with prerequisite dependencies. The variations observed in S_{human} further indicate that, without explicit logical constraints, enhanced feature representations alone are insufficient to fully capture the logical dependencies among audit labels.

EviCal addresses these limitations through an audit-oriented architecture that separates perception from reasoning. The label-aware semantic focusing mechanism enables the model to construct dedicated semantic views for each audit label across heterogeneous

multimodal units, thereby supporting precise evidence anchoring. The causal and logical verification module provides meta-level reasoning aligned with professional audit judgment by using the global logical capabilities of LLMs to evaluate and reconcile preliminary conclusions for consistency.

Empirically, EviCal achieves the strongest performance on the Domain-Doc dataset, particularly with respect to S_{human} . This indicates that for highly procedural and logically rigorous industrial inspection documents, incorporating evidence-chain-based logical calibration holds greater practical value than pure semantic modeling. Furthermore, the multidimensional confidence quantification mechanism not only improves predictive performance but also provides reliability assurance for final audit decisions, thereby supporting risk-aware deployment in real-world engineering settings.

4.6. Ablation Study

To investigate the specific contributions of each core component within the EviCal framework, we designed and evaluated three variants. Table 4 presents F1 and S_{human} for each variant across two datasets.

Table 4. Ablation study on General-Doc and Domain-Doc datasets. Variants A, B, and C denote the removal of the global consistency calibration, label-aware semantic focusing, and atomic semantic decomposition, respectively. S_{human} is the arithmetic mean of expert ratings.

Variant	General-Doc		Domain-Doc	
	F1	S_{human}	F1	S_{human}
(A) Without Causal Verification	79.15	3.82	74.95	3.24
(B) Without Semantic Focusing	72.33	4.31	69.23	4.15
(C) Without Atomic Representation	71.04	4.12	67.85	3.98
EviCal (Ours)	81.22	4.58	79.52	4.42

Tip: Bold values indicate better performance.

Experimental results show that removing the causal verification module (Variant A) leads to a moderate decrease in F1, accompanied by a much larger drop in S_{human} , especially on the Domain-Doc dataset, where the score decreases from 4.42 to 3.24. This discrepancy indicates that although the model can still preserve local semantic accuracy, the lack of global logical constraints makes its conclusions less aligned with professional audit reasoning. As a result, the causal verification module plays an essential role in maintaining logical coherence, which is critical for the credibility of the system in professional audit scenarios.

For Variant B, F1 declines sharply, demonstrating that label-conditioned grounding is crucial for handling long and noisy multimodal reports. Without this semantic focusing mechanism, the model struggles to correctly associate evidence with target labels. Interestingly, Variant B achieves higher S_{human} than Variant A. This suggests that when logical consistency is preserved, the model's conclusions remain more interpretable and acceptable to human experts, even if predictive accuracy is compromised.

Variant C shows the worst performance across all evaluation metrics. This result confirms that fine-grained decomposition of text segments, table rows, and image captions forms the foundation of the proposed framework. In the absence of atomic units, the model's perception and reasoning processes are fundamentally impaired, creating a bottleneck that constrains the effectiveness of all subsequent modules.

These results indicate a complementary interaction among the three components. Atomic Representation and Semantic Focusing primarily contribute to accurate information perception and extraction, while Causal Verification introduces the logical consistency needed for higher-level reasoning and integration. Through this combination, EviCal is

able to achieve strong empirical performance while meeting the reliability requirements expected in industrial auditing scenarios.

The ablation results can be further interpreted together with the granularity sensitivity analysis in Figure 2 to clarify the respective roles of algorithmic calibration and atomic-level preprocessing. Replacing atomic decomposition with block-level or page-level granularity reduces F1 by 7.74 and 16.10 points on General-Doc (81.22 vs. 73.48 vs. 65.12), respectively, confirming that fine-grained preprocessing contributes substantially. Meanwhile, the ablation results isolate the algorithmic contribution. Removing causal verification reduces F1 by only 2.07 points but causes a disproportionately large drop in S_{human} (4.58 $\rightarrow$ 3.82), while removing semantic focusing causes an 8.89-point F1 decline. These findings show that the two components contribute in complementary ways: preprocessing enables precise evidence localization, whereas the EviCal algorithm provides structured reasoning and logical consistency verification that are not captured by preprocessing alone.

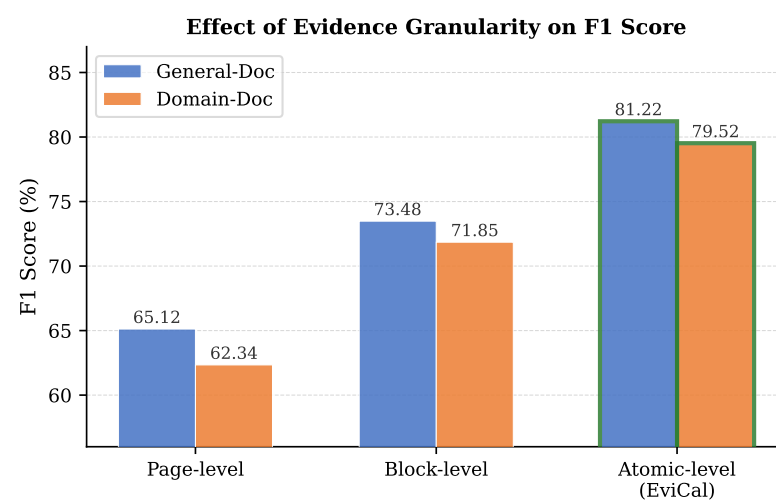


Figure 2. Sensitivity analysis of evidence granularity on F1 score. Green-edged bars mark the atomic-level setting used by EviCal.

4.7. Impact of Backbone LLMs

To investigate the dependency of the causal and logical verification module on the reasoning capabilities of underlying large language models (LLMs), we selected several representative mainstream LLMs as verification engines for comparative experiments. The experimental results, as shown in Table 5, reveal varying performance across different models within this module. Among them, DeepSeek-V3 and Qwen2.5-7B achieved relatively stable human ratings across both datasets. In scenarios involving complex audit rules and multiple constraints, DeepSeek-V3 demonstrated superior consistency in reasoning, aligning well with the structured state projection mechanism proposed in Section 3.3.2.

Table 5. Comparison of different LLMs as the calibration backbone on the S_{human} metric.

Dataset	ChatGPT-4o	DeepSeek-V3	Llama-3-8B	Qwen2.5-7B (Ours)
General-Doc	4.45	4.62	3.92	4.58
Domain-Doc	4.28	4.45	3.64	4.42

Tip: Bold values indicate better performance.

Notably, while ChatGPT-4o maintains high overall performance, its advantage in this verification task is less pronounced. This aligns with EviCal's design philosophy. Since the verification phase processes abstracted, reduced symbolic representations rather than raw multimodal content, the inference process emphasizes logical consistency and causal

coherence. Under this configuration, models with robust logical modeling capabilities are better suited to meet task requirements.

In contrast, Llama-3-8B exhibits a noticeable performance decline when processing the more specialized Domain-Doc dataset. This indicates that the consistency calibration process imposes certain demands on a model's reasoning capabilities. When model scale or logical prior knowledge is insufficient, its ability to model complex causal relationships among specialized audit metrics becomes constrained. Overall, while EviCal's architecture design does not rely on specific underlying models, its effectiveness can still benefit from large language models with strong logical reasoning capabilities, thereby further enhancing system performance in audit reliability.

4.8. Risk-Coverage Analysis

In high-stakes settings such as power system auditing, predictive accuracy alone is insufficient; models must also provide reliable estimates of their own uncertainty. To assess this aspect, we conduct a risk-coverage analysis to examine how well EviCal distinguishes between reliable predictions and potentially erroneous outputs.

The results summarized in Figure 3 show that prediction accuracy consistently increases as coverage is reduced (i.e., as the selection threshold tightens) by excluding samples with lower confidence scores. This trend is especially prominent in the specialized Domain-Doc dataset, confirming that EviCal's confidence scores serve as a robust proxy for decision reliability. Such sensitivity is driven by the internal composite confidence mechanism: whenever evidence support is insufficient or logical inconsistencies arise during verification, the confidence metric drops accordingly. By filtering these low-confidence samples, the system effectively isolates a high-certainty subset characterized by stronger logical coherence and evidentiary grounding.

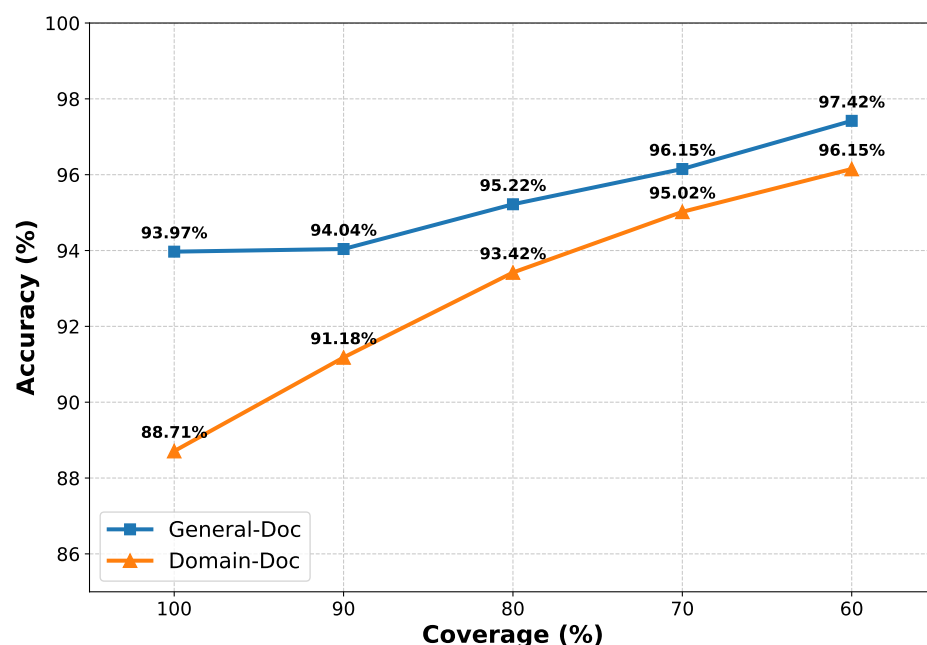


Figure 3. Prediction accuracy of the system on the General-Doc and Domain-Doc datasets under different coverage settings.

This risk-aware capability establishes a functional safety valve for industrial deployments. By routing low-confidence outputs to human experts while automating high-certainty tasks, EviCal balances operational efficiency with the rigorous safety standards required in engineering environments. This strategic integration of selective prediction en-

sures that the system remains trustworthy even when encountering ambiguous or complex document content.

4.9. Effect of Semantic Granularity

Document modeling granularity is a core factor affecting the accuracy of multimodal evidence attribution. It also directly informs whether EviCal's performance gain is mainly attributable to preprocessing or to the proposed algorithm, as discussed in Section 4.6. Therefore, we compared three decomposition strategies with varying levels of granularity.

Page-level: Treats the entire page as a single evidence unit.

Block-level: Physically slices the document using fixed-length sliding windows.

Atomic-level: Decomposes the document into minimal semantic units such as sentences, table rows, and captions based on its semantic structure.

Figure 2 presents the results. Coarse-grained modeling has clear limitations in this task. Because page-level strategies operate over broad evidence units, they often mix critical evidence with irrelevant background information, increasing noise and impairing the model's ability to identify core facts. Although block-level strategies reduce the size of retrieval units, their fixed-length physical segmentation can disrupt semantic continuity, for example by truncating cross-row tables or long sentences, which makes it difficult for models to obtain complete contextual information. In contrast, atomic-level modeling demonstrates superior stability in heterogeneous document scenarios. This approach divides candidate units into minimal semantic units, such as sentences, table rows, or captions, according to the document's natural semantic structure. It therefore preserves the semantic integrity of evidence while reducing noise, facilitating precise evidence localization and providing consistent input for subsequent logical reasoning.

4.10. Hyperparameter Sensitivity Study

4.10.1. Effect of Calibration Strength

To analyze the impact of the calibration strength parameter β on the system's reasoning behavior within the causal and logical verification module, we adjusted the calibration strength while keeping atomic representation and semantic focusing configurations fixed. The consistent performance across multiple document comprehension datasets under different calibration strengths is illustrated in Figure 4a. The results indicate that the calibration strength plays a crucial role in regulating the stability and reliability of the overall reasoning process.

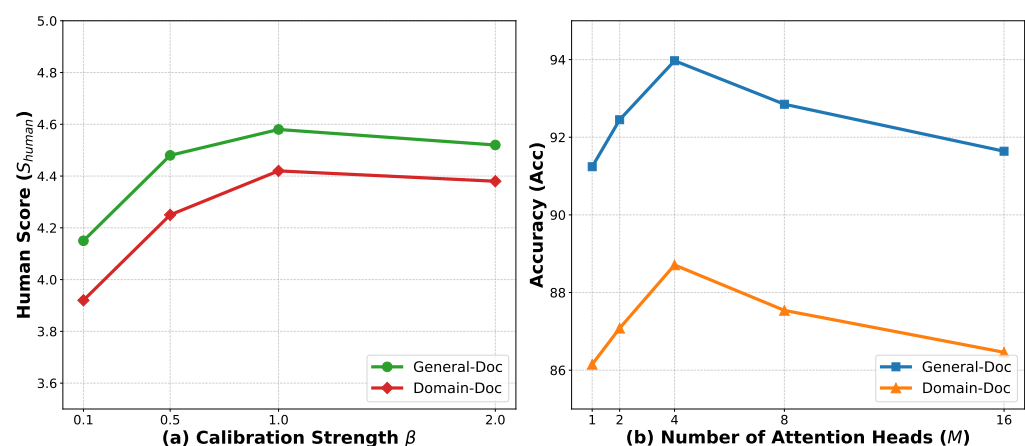


Figure 4. Impact of Calibration Strength β and Attention Head Count M Variations on System Consistency Performance.

When calibration constraints are weak, the model primarily relies on local evidence for judgments, making the reasoning process more susceptible to noisy information or insufficient evidence. As calibration strength progressively increases, the system demonstrates higher consistency in multi-evidence integration and conflict resolution, indicating that causal and logical constraints effectively regulate the reasoning process. However, when calibration strength is further heightened, overly stringent constraints suppress the model’s utilization of marginal yet plausible evidence, causing reasoning strategies to become overly conservative and adversely affecting overall performance.

4.10.2. Effect of Attention Head Count M

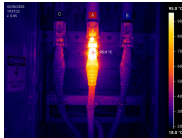
To assess the impact of multi-head attention on the semantic focus module, we conducted a sensitivity analysis on the number of attention heads M . The performance variation under different values of M is illustrated in Figure 4b. Experimental results indicate that within a certain range, increasing the number of attention heads enhances semantic focus effectiveness. When the number of attention heads is low, the model’s representational capacity is relatively constrained when processing structurally complex and semantically diverse document content, making it difficult to capture multiple key semantic clues simultaneously. As the number of attention heads increases, the model can model multiple semantic subspaces in parallel, thereby more effectively distinguishing and focusing on key information at different levels.

However, when the number of attention heads increases further, the performance improvement trend gradually weakens and even declines in some cases. This indicates that excessive attention heads may introduce redundant representations, leading to scattered semantic focus and thereby weakening the overall focusing effect. Overall, the effectiveness of semantic focusing mechanisms does not simply depend on increasing model complexity but requires striking a reasonable balance between representational capability and structural constraints.

4.11. Error Analysis

To better understand the limitations of EviCal on real-world industrial multimodal documents, we analyze representative failure cases in Table 6. The two cases illustrate different error sources: logical contradiction caused by insufficient domain commonsense when reconciling the main text and image captions, and structural fragmentation caused by complex table layouts.

Table 6. Failure case analysis of our EviCal framework when processing the power-system audits.

Multimodal Document	Ground Truth	Prediction Details																														
[Text] 对220 kV 某变电站10 kV 开关柜进行红外测温普查。记录显示：A相触头及母排接头处均未发现明显异常温升现象。 (Infrared thermometry survey conducted on a 10 kV switchgear at a 220 kV substation. Records show no abnormal temperature rise at Phase A contacts and busbar joints.) [Image] 红外热像图显示A相设备呈现亮白色高温耀斑。(The infrared thermogram shows Phase A equipment exhibiting a bright white high-temperature flare.)	 Oppose (Logical conflict between the main text and image caption)	Target Label: O&M Management Process Model Decision: Support Extracted Evidence: Text segment Confidence: 0.91																														
[Text] 本批次《PMS 3.0 系统授权清单》中，包含特殊运维协议（SLA等级为T1）的组件共有4 项，具体详见下表标记。 (In this batch of the PMS 3.0 System Authorization List, there are 4 components with special O&M agreements (SLA level T1), as marked in the table below.) [Table] 系统核心组件授权与维保清单：(System Core Component Authorization and Maintenance List:)	<table border="1"><thead><tr><th>序号</th><th>设备名称</th><th>系统功能描述</th><th>授权等级</th><th>授权期限</th><th>授权日期</th></tr></thead><tbody><tr><td>1</td><td>母线保护装置</td><td>母线故障检测与隔离</td><td>标准</td><td>2026/12/31</td><td></td></tr><tr><td>2</td><td>继电保护装置</td><td>电力系统故障检测与隔离</td><td>T1</td><td>2026/12/31</td><td></td></tr><tr><td>3</td><td>变电站监控系统</td><td>变电站运行状态监控</td><td>标准</td><td>2026/12/31</td><td></td></tr><tr><td>4</td><td>调度管理系统</td><td>电力系统调度指挥</td><td>T1</td><td>2026/12/31</td><td></td></tr></tbody></table> Support (Original table exactly matches the 4 items)	序号	设备名称	系统功能描述	授权等级	授权期限	授权日期	1	母线保护装置	母线故障检测与隔离	标准	2026/12/31		2	继电保护装置	电力系统故障检测与隔离	T1	2026/12/31		3	变电站监控系统	变电站运行状态监控	标准	2026/12/31		4	调度管理系统	电力系统调度指挥	T1	2026/12/31		Target Label: Software Asset List Model Decision: Details Missing Extracted Evidence: Misaligned table header fragments Confidence: 0.48
序号	设备名称	系统功能描述	授权等级	授权期限	授权日期																											
1	母线保护装置	母线故障检测与隔离	标准	2026/12/31																												
2	继电保护装置	电力系统故障检测与隔离	T1	2026/12/31																												
3	变电站监控系统	变电站运行状态监控	标准	2026/12/31																												
4	调度管理系统	电力系统调度指挥	T1	2026/12/31																												

In Case 1, the main text states that no abnormal temperature rise is observed, whereas the associated image caption explicitly describes a high-temperature flare on the relevant equipment. The model assigns excessive weight to the main text and fails to resolve the logical conflict between the document text and the image caption, leading to an incorrect “Support” decision. In Case 2, the system needs to verify whether the table contains four T1-level components. However, the nested table structure causes header misalignment and context fragmentation during atomic representation extraction. As a result, the model cannot adequately preserve the dependency between table entries and their parent headers, and outputs “Details Missing”. The relatively low confidence score of 0.48 indicates that the confidence quantification module can still reflect uncertainty when document parsing reaches its limits.

5. Discussion

Although EviCal is evaluated on power-system audit documents, its core architecture is not inherently tied to this domain. The core components, namely atomic multimodal decomposition, label-aware semantic focusing, symbolic state projection, consistency calibration, and confidence estimation, do not rely on power-system-specific assumptions. Domain-specific knowledge is confined to three configuration components: (1) the audit label space $\mathcal{L}$; (2) the evidence taxonomy that guides atomic decomposition; and (3) the causal and logical constraint templates used in the verification module. These components can be replaced independently without modifying the underlying learning architecture. Adapting EviCal to a new domain, such as medical report auditing, legal document review, or financial compliance checking, primarily requires redefining these three components with domain-expert input, while the attention-based grounding mechanism and confidence quantification module transfer directly.

The current implementation also has several boundaries worth noting. The consistency verification module relies on predefined rule templates rather than a formal domain knowledge graph, which limits the expressiveness of cross-label reasoning for highly complex constraint structures. Additionally, the symbolic state projection discards raw document content before the LLM reasoning stage, improving traceability but potentially leaving some cross-modal conflicts unresolved when resolution requires deep domain expertise beyond the symbolic representation. Finally, practical adaptation to a new domain still requires domain expert involvement to define the label space and constraint rules, introducing annotation costs that future work may address through low-resource transfer strategies.

6. Conclusions

In this paper, we propose EviCal, an evidence-grounded consistency calibration framework for content-level multimodal labeling in audit-oriented power-system documents. EviCal decomposes documents into atomic units, grounds each label decision to explicit evidence, and calibrates local predictions with global causal and logical verification on symbolic intermediate states. It further provides confidence estimates by combining evidence strength, prediction stability, and consistency feedback reliability, enabling risk-aware auditing. Experiments on two real-world datasets show that EviCal consistently outperforms strong multimodal baselines, achieving up to 93.97% accuracy and 81.22 F1 with a human score of 4.58/5. In future work, we will extend EviCal to broader document types and explore tighter integration of domain rules and human feedback for more scalable and robust industrial deployment.

Author Contributions: Conceptualization, X.Z., B.H., H.S. and L.N.; Methodology, X.Z. and B.H.; Software, B.H., Y.Y., H.S. and W.Q.; Validation, Y.Y. and G.Z.; Formal analysis, Y.Y. and W.Q.; Investigation, G.Z. and L.N.; Resources, B.H., G.Z. and W.Q.; Data curation, X.Z. and H.S.; Writing—original draft, X.Z., H.S. and W.Q.; Writing—review & editing, X.Z., G.Z., H.S., W.Q. and L.N.; Visualization, B.H., Y.Y. and L.N.; Supervision, Y.Y., H.S. and L.N.; Project administration, B.H. and G.Z.; Funding acquisition, Y.Y., W.Q. and L.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Science and Technology Project of State Grid Zhejiang Electric Power Co., Ltd., grant number 5211SX250007.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to data security and confidentiality requirements of the electric power industry.

Conflicts of Interest: Authors Xiaofeng Zhang, Baoli Han, Yufeng Yuan, Guangyao Zhu, Huibo Song, and Weixing Qiu were employed by State Grid Zhejiang Electric Power Co., Ltd., Shaoxing Power Supply Company. Author Li Ni was employed by State Grid Zhejiang Electric Power Co., Ltd., Shengzhou Power Supply Company. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Alvarez-Alvarado, M.S.; Donaldson, D.L.; Recalde, A.A.; Noriega, H.H.; Khan, Z.A.; Velasquez, W.; Rodríguez-Gallegos, C.D. Power System Reliability and Maintenance Evolution: A Critical Review and Future Perspectives. *IEEE Access* **2022**, *10*, 51922–51950. [\[CrossRef\]](#)
2. Huang, Y.; Lv, T.; Cui, L.; Lu, Y.; Wei, F. LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking. In *Proceedings of the 30th ACM International Conference on Multimedia MM '22*; Association for Computing Machinery: New York, NY, USA, 2022; pp. 4083–4091. [\[CrossRef\]](#)
3. Appalaraju, S.; Jasani, B.; Kota, B.U.; Xie, Y.; Manmatha, R. DocFormer: End-to-End Transformer for Document Understanding. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, 10–17 October 2021; pp. 973–983. [\[CrossRef\]](#)
4. Li, C.; Bi, B.; Yan, M.; Wang, W.; Huang, S.; Huang, F.; Si, L. StructuralLM: Structural Pre-training for Form Understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, Online, 1–6 August 2021.
5. Wang, D.; Raman, N.; Sibue, M.; Ma, Z.; Babkin, P.; Kaur, S.; Pei, Y.; Nourbakhsh, A.; Liu, X. Docllm: A layout-aware generative language model for multimodal document understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Bangkok, Thailand, 11–16 August 2024; pp. 8529–8548.
6. Tang, Z.; Yang, Z.; Wang, G.; Fang, Y.; Liu, Y.; Zhu, C.; Zeng, M.; Zhang, C.; Bansal, M. Unifying vision, text, and layout for universal document processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, BC, Canada, 18–22 June 2023; pp. 19254–19264.
7. Liu, Y.; Yang, B.; Liu, Q.; Li, Z.; Ma, Z.; Zhang, S.; Bai, X. Textmonkey: An ocr-free large multimodal model for understanding document. *arXiv* **2024**, arXiv:2403.04473. [\[CrossRef\]](#)
8. Kim, G.; Hong, T.; Yim, M.; Nam, J.; Park, J.; Yim, J.; Hwang, W.; Yun, S.; Han, D.; Park, S. OCR-free Document Understanding Transformer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, Tel Aviv, Israel, 23–27 October 2022.
9. Ye, J.; Hu, A.; Xu, H.; Ye, Q.; Yan, M.; Dan, Y.; Zhao, C.; Xu, G.; Li, C.; Tian, J.; et al. mplug-docowl: Modularized multimodal large language model for document understanding. *arXiv* **2023**, arXiv:2307.02499.
10. Zhang, Y.; Zhang, R.; Gu, J.; Zhou, Y.; Lipka, N.; Yang, D.; Sun, T. Llavav: Enhanced visual instruction tuning for text-rich image understanding. *arXiv* **2023**, arXiv:2306.17107.
11. Luo, J.; Yao, S.; Zhao, C.; Xu, J.; Feng, J. A federated named entity recognition model with explicit relation for power grid. *Comput. Mater. Contin.* **2023**, *75*, 4207. [\[CrossRef\]](#)
12. Tang, W.; Zhang, Y.; Mao, X.; Shan, M.; Lv, K.; Sun, X.; Ding, Z. Enhanced Named Entity Recognition and Event Extraction for Power Grid Outage Scheduling Using a Universal Information Extraction Framework. *Energies* **2025**, *18*, 3617. [\[CrossRef\]](#)

13. Meng, L.; Wang, Y.; Huang, Y.; Ma, D.; Zhu, X.; Zhang, S. A Named Entity Recognition Model for Chinese Electricity Violation Descriptions Based on Word-Character Fusion and Multi-Head Attention Mechanisms. *Energies* **2025**, *18*, 401. [\[CrossRef\]](#)
14. Chen, G.; Xie, W.; Liu, Y.; Yuan, X.; Zhao, L. Systematically modeling and extracting bibliographic metadata of power grid standard documents with LLMs. *Inf. Res. Int. Electron. J.* **2025**, *30*, 654–665. [\[CrossRef\]](#)
15. Yao, J.; She, Y.; Shi, K. Domain Knowledge Graph Construction and Troubleshooting Method for Intelligent Diagnosis of Power Transformer Defects. In Proceedings of the 2025 IEEE 7th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), Chongqing, China, 5–7 December 2025; Volume 7, pp. 1205–1213. [\[CrossRef\]](#)
16. Xiong, J.; Yang, P.; Chen, B.; Chen, Z. Defect Identification Method of Power Grid Secondary Equipment Based on Coordination of Knowledge Graph and Bayesian Network Fusion. *Energy Eng.* **2026**, *123*, 1. [\[CrossRef\]](#)
17. Liu, P.; Tian, B.; Liu, X.; Gu, S.; Yan, L.; Bullock, L.; Ma, C.; Liu, Y.; Zhang, W. Construction of Power Fault Knowledge Graph Based on Deep Learning. *Appl. Sci.* **2022**, *12*, 6993. [\[CrossRef\]](#)
18. Liu, W.; Gu, Y.; Zeng, Z.; Qi, D.; Li, D.; Luo, Y.; Li, Q.; Wei, S. Automated Equipment Defect Knowledge Graph Construction for Power Grid Regulation. *Electronics* **2024**, *13*, 4430. [\[CrossRef\]](#)
19. Zhang, L.; Kuang, J.; Teng, Y.; Xiang, S.; Li, L.; Zhou, Y. A Lightweight Infrared and Visible Light Multimodal Fusion Method for Object Detection in Power Inspection. *Processes* **2025**, *13*, 2720. [\[CrossRef\]](#)
20. Guo, Y.; He, Y.; Zhang, K.; Zhang, T.; Lin, Y.; Wang, Z.; Chen, H. A decoupled scene-equipment fusion method for power substation equipment detection. *Knowl.-Based Syst.* **2025**, *331*, 114792. [\[CrossRef\]](#)
21. Zhang, K.; Zheng, Z.; Wang, J.; Yang, J.; Xiao, Y. Multimodal knowledge-guided method for power transmission line fault detection using a vision-language model. *Int. J. Electr. Power Energy Syst.* **2026**, *176*, 111696. [\[CrossRef\]](#)
22. Zhong, Y.; Luo, P.; Yan, Y.; Jia, T.; Qi, D. PowerGPT: A multimodal foundation model for power inspection. *Appl. Soft Comput.* **2025**, *186*, 113939. [\[CrossRef\]](#)
23. Wang, J.; Li, M.; Luo, H.; Zhu, J.; Yang, A.; Rong, M.; Wang, X. Power-LLaVA: Large Language and Vision Assistant for Power Transmission Line Inspection. *arXiv* **2024**, arXiv:2407.19178.
24. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q.V.; Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 24824–24837.
25. Yao, S.; Yu, D.; Zhao, J.; Shafraan, I.; Griffiths, T.; Cao, Y.; Narasimhan, K. Tree of thoughts: Deliberate problem solving with large language models. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 11809–11822.
26. Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S.; Chowdhery, A.; Zhou, D. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In Proceedings of the International Conference on Learning Representations (ICLR), Kigali, Rwanda, 1–5 May 2023.
27. Chi, H.; Li, H.; Yang, W.; Liu, F.; Lan, L.; Ren, X.; Liu, T.; Han, B. Unveiling causal reasoning in large language models: Reality or mirage? *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 96640–96670.
28. Kiciman, E.; Ness, R.; Sharma, A.; Tan, C. Causal reasoning and large language models: Opening a new frontier for causality. *Trans. Mach. Learn. Res.* **2023**. Available online: <https://par.nsf.gov/biblio/10574854> (accessed on 19 March 2026)
29. Jin, Z.; Chen, Y.; Leeb, F.; Gresele, L.; Kamal, O.; Lyu, Z.; Blin, K.; Gonzalez Adauto, F.; Kleiman-Weiner, M.; Sachan, M.; et al. Cladder: Assessing causal reasoning in language models. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 31038–31065.
30. Pan, S.; Luo, L.; Wang, Y.; Chen, C.; Wang, J.; Wu, X. Unifying large language models and knowledge graphs: A roadmap. *IEEE Trans. Knowl. Data Eng.* **2024**, *36*, 3580–3599. [\[CrossRef\]](#)
31. Yang, K.; Deng, J.; Chen, D. Generating Natural Language Proofs with Verifier-Guided Search. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022; Association for Computational Linguistics (ACL): Stroudsburg, PA, USA, 2022; pp. 89–105.
32. Liu, X.; Xu, P.; Wu, J.; Yuan, J.; Yang, Y.; Zhou, Y.; Liu, F.; Guan, T.; Wang, H.; Yu, T.; et al. Large language models and causal inference in collaboration: A comprehensive survey. In Findings of the Association for Computational Linguistics: NAACL 2025; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 7668–7684.
33. Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; Wang, H.; Wang, H. Retrieval-augmented generation for large language models: A survey. *arXiv* **2023**, arXiv:2312.10997.
34. Nakano, R.; Hilton, J.; Balaji, S.; Wu, J.; Ouyang, L.; Kim, C.; Hesse, C.; Jain, S.; Kosaraju, V.; Saunders, W.; et al. WebGPT: Browser-assisted question-answering with human feedback. *arXiv* **2021**, arXiv:2112.09332.
35. Gao, T.; Yen, H.; Yu, J.; Chen, D. Enabling Large Language Models to Generate Text with Citations. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023.
36. Min, S.; Krishna, K.; Lyu, X.; Lewis, M.; Yih, W.t.; Koh, P.; Iyyer, M.; Zettlemoyer, L.; Hajishirzi, H. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–10 December 2023; pp. 12076–12100.

37. Manakul, P.; Liusie, A.; Gales, M. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–10 December 2023; pp. 9004–9017.
38. Kuhn, L.; Gal, Y.; Farquhar, S. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Large Language Models. In Proceedings of the International Conference on Learning Representations (ICLR), Kigali, Rwanda, 1–5 May 2023.
39. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 27730–27744.
40. Wang, T.; Yu, P.; Tan, X.E.; O’Brien, S.; Pasunuru, R.; Dwivedi-Yu, J.; Golovneva, O.; Zettlemoyer, L.; Fazel-Zarandi, M.; Celikyilmaz, A. Shepherd: A critic for language model generation. *arXiv* **2023**, arXiv:2308.04592. [[CrossRef](#)]
41. Gao, L.; Dai, Z.; Pasupat, P.; Chen, A.; Chaganty, A.T.; Fan, Y.; Zhao, V.; Lao, N.; Lee, H.; Juan, D.C.; et al. Rarr: Researching and revising what language models say, using language models. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Toronto, ON, Canada, 9–14 July 2023; pp. 16477–16508.
42. Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.* **2025**, *43*, 1–55. [[CrossRef](#)]
43. Poznanski, J.; Borchardt, J.; Dunkelberger, J.; Huff, R.; Lin, D.; Rangapur, A.; Wilhelm, C.; Lo, K.; Soldaini, L. olmOCR: Unlocking Trillions of Tokens in PDFs with Vision Language Models. *arXiv* **2025**, arXiv:2502.18443. Available online: <http://arxiv.org/abs/2502.18443> (accessed on 19 March 2026). [[CrossRef](#)]
44. Team, Q. Qwen2.5-VL Technical Report. *arXiv* **2025**, arXiv:2502.13923v1. [[CrossRef](#)]
45. DeepSeek-AI. DeepSeek-V3 Technical Report. *arXiv* **2024**, arXiv:2412.19437.
46. OpenAI. Hello GPT-4o. 2024. Available online: <https://openai.com/index/hello-gpt-4o> (accessed on 19 March 2026).
47. Li, Z.; Li, Y.; Kang, L.; Karatzas, D.; Ma, W. AVIR: Adaptive Visual In-Document Retrieval for Efficient Multi-Page Document Question Answering. In *Proceedings of the 7th ACM International Conference on Multimedia in Asia*; Association for Computing Machinery: New York, NY, USA, 2025; pp. 1–7.
48. Mo, Y.; Shao, Z.; Ye, K.; Mao, X.; Zhang, B.; Xing, H.; Ye, P.; Huang, G.; Chen, K.; Huan, Z.; et al. Doc-CoB: Enhancing Multi-Modal Document Understanding with Visual Chain-of-Boxes Reasoning. *arXiv* **2025**, arXiv:2505.18603.
49. Faysse, M.; Sibille, H.; Wu, T.; Omrani, B.; Viaud, G.; Hudelot, C.; Colombo, P. ColPali: Efficient Document Retrieval with Vision Language Models. In *Proceedings of the International Conference on Learning Representations*; Yue, Y., Garg, A., Peng, N., Sha, F., Yu, R., Eds.; ICLR: Appleton, WI, USA, 2025; Volume 2025, pp. 61424–61449.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.