# P-QuASAR: A Unified Probabilistic Framework for Holistic Patent Quality Assessment and Refinement

**Xinyuan Song[1]*, Ziyi Ni[2]*, Fred Yang[3]†, Bo Zhang[3], Yijin Wang[4], jane zhang[3]**
[1]Emory University [2]University of Chinese Academy of Science
[3]AIIP Lab [4]Xidian University

## Abstract

Automated assessment of patent quality is increasingly important given the growth of patent filings and the adoption of AI-assisted drafting. Existing methods often rely on modular pipelines or generic detectors, resulting in fragmented decisions and limited integration across quality dimensions. We propose **P-QuASAR** (Patent Quality Assurance via Structured Assessment and Refinement), a unified probabilistic framework that represents patent specifications as *Quality Graphs*. Multiple interdependent quality dimensions—such as regulatory compliance, technical coherence, and figure–text consistency—are jointly modeled using uncertainty-aware Quality Assessment Functions with learned edge potentials. Cross-dimensional evidence propagation via loopy belief propagation enables calibrated defect detection, while *Optimal Intervention Paths* translate inferred quality states into prioritized and actionable refinement recommendations. Evaluated on 500 patents across eight IPC domains against seven state-of-the-art baselines, P-QuASAR achieves substantial improvements: 99.86% balanced accuracy on regulatory compliance, 88.91% on technical coherence, and 94.70% on figure consistency, outperforming the strongest baselines by 3.0%, 9.0%, and 7.1%, respectively. Ablation studies confirm that joint graph reasoning contributes 3.66 points to average performance. When applied for refinement, P-QuASAR reduces average defects in AI-generated patents from 9.04–12.15 to 3.21 per document, surpassing human-authored patents.

## 1 Introduction

The global patent system faces unprecedented scalability challenges. According to the World Intellectual Property Organization (WIPO), over 3.55 million patent applications were filed worldwide in 2023 alone[1], representing a 15.7% increase over the past decade. Each patent specification—the legal and technical document defining an invention's scope—must satisfy stringent requirements across multiple quality dimensions: regulatory compliance with jurisdiction-specific rules, technical accuracy and coherence, and internal consistency between textual descriptions and visual elements. While traditional manual expert review remains thorough, it is increasingly unsustainable given this exponential growth. The average examination pendency at major patent offices now exceeds 24 months, creating bottlenecks that delay innovation protection (Jaffe and De Rassenfosse, 2017).

Concurrently, the proliferation of AI-powered patent drafting tools introduces novel quality risks that exacerbate this challenge (Ji et al., 2023; Yu, 2025; Ni et al., 2025b). Commercial platforms such as MindFlowing, PatSnap, and Eureka now offer automated generation of complete patent specifications. However, as our empirical analysis reveals (Section 4), these AI-generated documents exhibit significantly higher defect rates than human-authored patents across all quality dimensions—averaging 9.04–12.15 defects per document compared to 7.13 for human patents. These defects include technical inaccuracies, regulatory non-compliance, and figure-text misalignments that can undermine patent validity and enforceability. This confluence of rising volume and emerging AI-induced defect patterns creates a pressing need for robust, automated quality assessment frameworks.

Existing automated approaches for patent quality evaluation remain fundamentally inadequate to address this challenge. We identify three categories of prior work, each with distinct limitations:

*Proxy-based methods* rely on external indicators such as citation counts (Hall et al., 2005), patent

*These authors contributed equally to this work.
†**Correspondence:**fredyang@aiiplab.com.

[1]https://www.wipo.int/pressroom/en/articles/2024/article_0015.html, accessed August 2025.



*Findings of the Association for Computational Linguistics: ACL 2026*, pages 11935–11951
July 2-7, 2026 

family size (Trajtenberg, 1990), or commercial value predictions (Harhoff et al., 1999); while useful for portfolio-level analysis, these methods cannot assess the intrinsic, document-specific quality essential for legal validity. *Text analysis methods* employ deep learning for patent text understanding (Wang et al., 2021), including claim generation (Lee and Hsiang, 2020) and similarity measurement (Kim et al., 2022; Wang and Liu, 2024); however, these approaches typically use generic language metrics that fail to capture domain-specific imperatives such as jurisdictional compliance rules or claim-specification alignment. *AI content detectors* such as DetectGPT (Mitchell et al., 2023) and GPTZero provide binary human/AI classification but cannot identify specific deficiencies or provide actionable guidance for quality improvement.

A critical gap thus exists: *no existing framework provides comprehensive, multi-dimensional assessment of patent specifications with the ability to diagnose specific defects and recommend targeted corrections.*

Large Language Models (LLMs) offer a promising foundation to bridge this gap, owing to their cross-domain knowledge and sophisticated reasoning capabilities (Brown et al., 2020). Recent surveys highlight the growing intersection of NLP and patent analysis (Risch et al., 2024), while multimodal LLMs enable integrated analysis of both text and figures (Liu et al., 2024)—a capability particularly relevant for patent specifications containing technical drawings. However, directly applying LLMs to patent quality assessment faces fundamental challenges: (1) the *isolation problem*, where separate models for different quality dimensions cannot share evidence; (2) the *calibration problem*, where heterogeneous signals lack principled fusion mechanisms; and (3) the *actionability problem*, where detection is disconnected from actionable refinement guidance.

To address these challenges, we introduce **P-QuASAR** (**P**atent **Qu**ality **A**ssurance via **S**tructured **A**ssessment and **R**efinement), a unified probabilistic framework that fundamentally reformulates patent quality assessment. The key insight underlying P-QuASAR is that quality dimensions in patent specifications are *not independent*—a figure-text mismatch often signals broader technical description deficiencies, while regulatory noncompliance may indicate systematic drafting errors. By modeling these interdependencies explicitly, we can leverage cross-dimensional evidence to improve detection accuracy and generate coherent refinement strategies.

Our contributions are threefold. *First*, we propose modeling a patent specification as a *Quality Graph* $G = (V, E)$, where nodes represent quality dimensions and edges encode conditional dependencies learned from expert annotations. This formulation enables cross-dimensional evidence propagation via message-passing inference, addressing the isolation problem inherent in modular pipeline approaches. *Second*, we design three specialized uncertainty-aware Quality Assessment Functions (QAFs)—for regulatory compliance (QAF-R), technical coherence (QAF-T), and figure-reference consistency (QAF-F)—each producing calibrated probability estimates with explicit uncertainty quantification. This enables principled Bayesian fusion of heterogeneous quality signals within the Quality Graph. *Third*, we formalize the refinement problem as constrained optimization over atomic editing operators, deriving prioritized intervention sequences (Optimal Intervention Paths) that translate detected defects into actionable corrections. Applied to AI-generated patents, this approach reduces average defects from 9.04–12.15 to 3.21 per document—a 65–74% reduction that surpasses human-authored patent quality.

We evaluate P-QuASAR on a large-scale dataset of 500 patent specifications spanning eight IPC technical domains against seven state-of-the-art baselines. Experimental results demonstrate substantial improvements: 99.86% balanced accuracy on regulatory compliance (+3.0% over best baseline), 88.91% on technical coherence (+9.0%), and 94.70% on figure consistency (+7.1%). Comprehensive ablation studies confirm that the Quality Graph contributes 3.66 percentage points to average performance, validating our core hypothesis that cross-dimensional reasoning improves patent quality assessment.

## 2 Methodology

We propose **P-QuASAR** (Patent Quality Assurance via Structured Assessment and Refinement), a unified probabilistic framework that reformulates patent quality assessment as joint inference over a graphical model. This section first introduces the notation and problem formulation, then details each component of our framework.

### 2.1 Notation and Problem Formulation

Table 1 summarizes the key notation used throughout this paper.

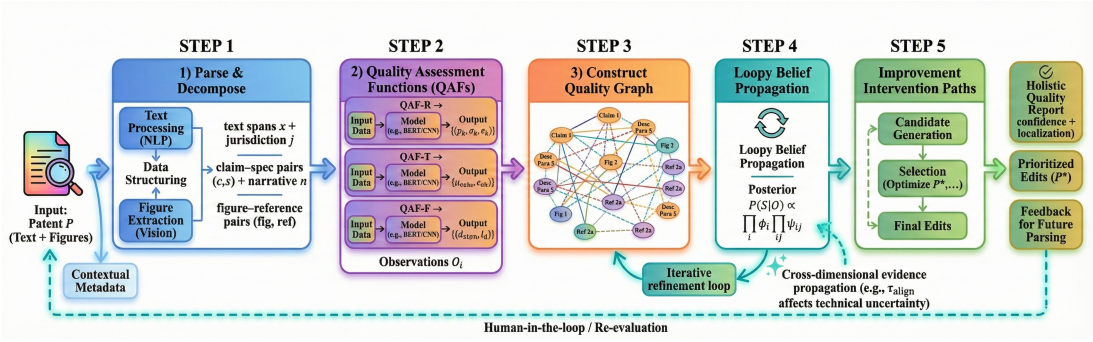


Figure 1: **Algorithmic pipeline of P-QuASAR.** Given a patent specification $P$, the framework (i) parses and decomposes $P$ into text spans, claim–specification pairs, and figure–reference pairs; (ii) applies uncertainty-aware QAF-R/QAF-T/QAF-F to produce noisy observations; (iii) constructs a Quality Graph and performs joint inference with Loopy Belief Propagation to approximate the posterior $P(S \mid O)$; and (iv) derives an Optimal Intervention Path by mapping inferred defects to atomic improvement operators, yielding a holistic quality report and prioritized editing actions.

Table 1: Summary of Notation

| Symbol | Description |
|---|---|
| $\mathcal{P}$ | Patent specification document |
| $G = (V, E)$ | Quality Graph with nodes $V$ and edges $E$ |
| $v_i \in V$ | Quality Dimension (QD) node |
| $e_{ij} \in E$ | Edge encoding dependency between $v_i$ and $v_j$ |
| $S_i \in \{0, 1\}$ | Latent quality state (1 = violation) |
| $\mathbf{S}$ | Vector of all latent states $[S_1, \ldots, S_{\lvert V \rvert}]$ |
| $O_i$ | Observation from QAF for node $v_i$ |
| $\mathbf{O}$ | Vector of all observations |
| $\phi_i(S_i, O_i)$ | Node potential (unary factor) |
| $\psi_{ij}(S_i, S_j)$ | Edge potential (pairwise factor) |
| $\theta_R, \theta_T, \theta_F$ | Parameters for QAF-R, QAF-T, QAF-F |
| $\mathcal{V}^{\text{reg}}$ | Set of regulatory violation types |
| $\mathcal{V}^{\text{tech}}$ | Set of technical defect modes |
| $\mathcal{OP}$ | Set of atomic improvement operators |
| $P^*$ | Optimal Intervention Path |
| $\lambda$ | Cost-utility trade-off parameter |

**Problem Formulation.** Given a patent specification $\mathcal{P}$ containing textual content (title, abstract, claims, detailed description) and visual elements (figures with labels), our goal is threefold: (i) *assess* the document by inferring the posterior distribution $P(\mathbf{S}|\mathcal{P})$ over latent quality states across all dimensions; (ii) *diagnose* specific defect types and their locations within $\mathcal{P}$; and (iii) *refine* the document by deriving an optimal sequence of editing operations to correct identified defects.

## 2.2 Framework Overview

P-QuASAR addresses three core limitations of existing modular approaches: (1) the **isolation problem**, where parallel detection modules cannot share evidence across quality dimensions; (2) the **calibration problem**, where heterogeneous signals lack principled fusion mechanisms; and (3) the **actionability problem**, where detection is disconnected from refinement guidance. Our framework unifies these aspects through the architecture illustrated in Figure 1 and formalized in Algorithm 1.

## 2.3 Quality Graph Construction

The core of P-QuASAR is a probabilistic graphical model representing the patent document. A patent $\mathcal{P}$ is decomposed into a set of textual and graphical elements through document parsing. We define a *Quality Graph* $G = (V, E)$, where each node $v_i \in V$ represents a specific *Quality Dimension* (QD) associated with a document element or a cross-element relationship. Edges $e_{ij} \in E$ encode conditional dependencies between QDs, modeling how a defect in one dimension (e.g., a missing figure reference) probabilistically influences the state of another (e.g., the sufficiency of a technical description).

Each QD node $v_i$ has a latent state variable $S_i \in \{0, 1\}$, where $S_i = 1$ indicates a quality violation. The system infers the posterior distribution $P(\mathbf{S}|\mathcal{P})$ over all latent states given the observed patent content. This inference integrates signals from three specialized yet interconnected *Quality Assessment Functions* (QAFs). The complete inference and refinement procedure is formalized in Algorithm 1.

## 2.4 Quality Assessment Functions (QAFs) with Calibrated Confidence

Each QAF is a parameterized function that analyzes the patent and produces a distribution over potential violations for its assigned QD cluster. Crucially, each QAF estimates its own epistemic uncertainty via Monte Carlo Dropout (Gal and Ghahramani, 2016), enabling confidence-aware evidence fusion. We detail each QAF below.

**QAF-R: Regulatory Compliance.** This function maps a text span $x$ (e.g., a sentence) and a target jurisdiction $j$ to a set of potential violation types $\mathcal{V}^{\text{reg}}$. Based on China National Intellectual Property Administration (CNIPA) examination guidelines, we define 12 regulatory violation types: (R1) improper claim dependency structure; (R2) missing essential technical features in independent claims; (R3) inconsistent terminology between claims and description; (R4) insufficient disclosure of best mode; (R5) abstract exceeding length limits (>300 characters); (R6) missing or improper reference numerals; (R7) non-standard technical term usage; (R8) claim scope broader than disclosure support; (R9) missing unity of invention justification; (R10) improper use of functional language; (R11) incomplete embodiment descriptions; and (R12) format and structural violations.

For each violation type $k \in \mathcal{V}^{\text{reg}}$, QAF-R outputs a calibrated probability and uncertainty estimate:

$$f_{\text{comp}}(x, j; \theta_R) \to \{(p_k, \sigma_k, e_k)\}_{k=1}^{|\mathcal{V}^{\text{reg}}|} \quad (1)$$

where $p_k = \frac{1}{M}\sum_{m=1}^{M} \hat{p}_k^{(m)}$ is the mean probability over $M$ stochastic forward passes, $\sigma_k = \sqrt{\frac{1}{M}\sum_{m=1}^{M}(\hat{p}_k^{(m)} - p_k)^2}$ captures epistemic uncertainty, and $e_k$ is a structured explanation. The function $f_{\text{comp}}$ is implemented via a fine-tuned DeBERTa-large model with dropout rate 0.1, using $M = 10$ forward passes during inference.

**QAF-T: Technical Coherence.** This function evaluates the logical and factual soundness of technical descriptions. It operates over claim-specification pairs $(c, s)$ and technical narratives $n$. QAF-T models coherence via a four-tier risk classification: *High Risk* indicates content containing factual errors, logical contradictions, or fabricated technical details that would invalidate the patent; *Medium Risk* denotes ambiguous descriptions, unsupported assertions, or incomplete technical explanations; *Low Risk* encompasses minor clarity issues or stylistic problems that do not affect technical validity; and *Safe* represents technically sound content with no identified defects.

QAF-T additionally detects specific defect modes

$$\mathcal{V}^{\text{tech}} = \{\text{contradiction}, \text{fabrication}, \text{insufficiency}, \text{ambiguity}\}$$

$$f_{\text{coh}}(c, s, n; \theta_T) \to (\mu_{\text{coh}}, \sigma_{\text{coh}}, \{(p_d, l_d)\}_{d \in \mathcal{V}^{\text{tech}}}) \quad (2)$$

Here, $\mu_{\text{coh}}$ and $\sigma_{\text{coh}}$ parameterize a Gaussian distribution over the coherence score, enabling soft classification into risk tiers. Each defect mode $d$ has a probability $p_d$ and a localization $l_d$ (token span indices). This formulation separates severity (encoded in $\mu_{\text{coh}}$) from defect type, enabling more nuanced quality assessment.

**QAF-F: Figure-Reference Consistency.** This function processes figure-reference pairs $(fig, ref)$. It employs a vision encoder (ViT-L/14) to parse figures and extract visual entities $\mathcal{E}_{\text{fig}}$ (labeled components, reference numerals), and a text encoder to extract referenced entities $\mathcal{E}_{\text{ref}}$ from the specification. Consistency is modeled via optimal bipartite matching using the Hungarian algorithm, with the function outputting a discrepancy set:

$$f_{\text{fig}}(fig, ref; \theta_F) \to (\mathcal{D}_{\text{mis}}, \mathcal{D}_{\text{miss}}, \mathcal{D}_{\text{extra}}, \tau_{\text{align}}) \quad (3)$$

where $\mathcal{D}_{\text{mis}}$ denotes mismatched labels (visual label $\neq$ textual reference), $\mathcal{D}_{\text{miss}}$ captures missing references (visual entities without textual mention), $\mathcal{D}_{\text{extra}}$ identifies extraneous references (textual mentions without visual entities), and $\tau_{\text{align}} \in [0, 1]$ provides an overall alignment confidence score. Cross-figure validation ensures consistency of reference numerals across multiple figures in the same patent.

### 2.5 Joint Reasoning via the Quality Graph

The QAF outputs serve as noisy observations $\mathbf{O}$ for the corresponding QD nodes in the graph $G$. A key innovation is the *learned edge potential* $\psi_{ij}(S_i, S_j)$, which captures the conditional dependency between connected QDs $v_i$ and $v_j$. Unlike hand-crafted rules, these potentials are learned end-to-end from expert annotations.

**Edge Potential Parameterization.** We parameterize edge potentials using a log-linear model:

$$\psi_{ij}(S_i, S_j) = \exp\left(\mathbf{w}_{ij}^{\top} \boldsymbol{\phi}(S_i, S_j)\right) \quad (4)$$

where $\boldsymbol{\phi}(S_i, S_j) = [S_i, S_j, S_i \cdot S_j, |S_i - S_j|]^{\top}$ is a feature vector and $\mathbf{w}_{ij}$ are learned weights. This allows the model to capture both positive correlations (co-occurring defects) and negative correlations (mutually exclusive defects).

**Node Potential Construction.** Node potentials integrate QAF outputs with their uncertainty estimates:

$$\phi_i(S_i, O_i) = \begin{cases} \frac{p_i}{\sigma_i + \epsilon} & \text{if } S_i = 1 \\ \frac{1 - p_i}{\sigma_i + \epsilon} & \text{if } S_i = 0 \end{cases} \quad (5)$$

where $p_i$ and $\sigma_i$ are the probability and uncertainty from the corresponding QAF, and $\epsilon = 10^{-6}$ prevents division by zero. This formulation down-weights evidence from uncertain predictions.

**Inference via Loopy Belief Propagation.** The joint posterior probability is approximated using message-passing inference (Murphy et al., 1999; Yedidia et al., 2003):

$$P(\mathbf{S}|\mathbf{O}) \propto \prod_{v_i \in V} \phi_i(S_i, O_i) \prod_{e_{ij} \in E} \psi_{ij}(S_i, S_j) \quad (6)$$

Following the sum-product algorithm on factor graphs (Kschischang et al., 2001), messages are iteratively updated until convergence (Algorithm 1, Lines 16–23). The joint inference permits evidence to flow across modules: a low alignment confidence $\tau_{\text{align}}$ from QAF-F can increase the posterior probability of technical defects from QAF-T if they share an edge, yielding a more robust overall assessment.

**Complexity Analysis.** The time complexity of P-QuASAR inference is $O(|\mathcal{P}| \cdot d + |V| \cdot |E| \cdot T)$, where $|\mathcal{P}|$ is document length, $d$ is the QAF feature dimension, $|V|$ and $|E|$ are graph size, and $T$ is the number of BP iterations. In practice, the Quality Graph is sparse ($|E| = O(|V|)$) and BP typically converges within $T \leq 10$ iterations, making inference efficient.

### 2.6 Deriving Optimal Intervention Paths

The final stage translates the inferred quality state $\hat{\mathbf{S}}$ and structured defect information from the QAFs into actionable guidance. This bridges the gap between detection and refinement within a unified framework.

**Atomic Improvement Operators.** We define a set of atomic *Improvement Operators* $\mathcal{OP}$, each corresponding to a concrete editing action. Table 2 lists the primary operator types.

Table 2: Atomic Improvement Operators

| Operator | Description |
|---|---|
| `ReplaceTerm`$(t_1, t_2)$ | Replace term $t_1$ with $t_2$ |
| `AddReference`$(fig, lbl)$ | Add label $lbl$ to figure $fig$ |
| `RemoveReference`$(fig, lbl)$ | Remove label from figure |
| `RewriteClause`$(c, c')$ | Rewrite clause $c$ as $c'$ |
| `AddDisclosure`$(sec, txt)$ | Add text to section $sec$ |
| `FixDependency`$(c_i, c_j)$ | Correct claim dependency |
| `AlignTerminology`$(T)$ | Unify terminology set $T$ |

Each defect type from the QAFs is mapped to one or more candidate operators via a learned mapping function $\mathcal{M} : \mathcal{V}^{\text{reg}} \cup \mathcal{V}^{\text{tech}} \cup \mathcal{V}^{\text{fig}} \to 2^{\mathcal{OP}}$.

**Cost and Utility Functions.** We define the operator cost $C(\Phi_k)$ based on edit complexity:

$$C(\Phi_k) = \alpha \cdot \text{EditDist}(\Phi_k) + \beta \cdot \text{Scope}(\Phi_k) \quad (7)$$

where EditDist measures character-level changes and Scope measures the number of affected document sections. The utility function estimates post-intervention quality improvement:

$$U(\hat{\mathbf{S}}, P) = \sum_{i:\hat{S}_i=1} \mathbb{E}[\mathbf{1}[S_i = 0|P]] \cdot w_i \quad (8)$$

where $w_i$ is the severity weight for defect $i$ (High Risk: 3.0, Medium: 2.0, Low: 1.0).

**Optimization via Greedy Search.** Finding the globally optimal intervention path is NP-hard due to operator dependencies. We employ a greedy algorithm with look-ahead:

$$P^* = \arg \max_{P \in \mathcal{P}_{\text{feasible}}} \left[ U(\hat{\mathbf{S}}, P) - \lambda \sum_{\Phi \in P} C(\Phi) \right] \quad (9)$$

where $\mathcal{P}_{\text{feasible}}$ respects operator precedence constraints (e.g., `AddReference` before `AlignTerminology`). The trade-off parameter $\lambda$ is tuned on the validation set. The final output is a prioritized intervention sequence $P^*$, each operator linked to root-cause analysis from the QAFs and joint inference, providing transparent and actionable improvement guidance.

## 3 Experimental Setup

### 3.1 Research Design Overview

We evaluate P-QuASAR on a large-scale dataset of 500 patent documents comprising both human-authored and AI-generated patents across eight technical domains (detailed in Section B.1). Our experiments address four research questions: **RQ1** investigates the detection accuracy of P-QuASAR compared to expert annotations and state-of-the-art baselines. **RQ2** examines how defect patterns vary across patent sections and technical domains. **RQ3** explores quality differences between human-authored, AI-generated, and P-QuASAR-refined patents. **RQ4** analyzes the contribution of each framework component through comprehensive ablation studies.

### 3.2 Evaluation Methodology

#### 3.2.1 Baseline Methods

We compare P-QuASAR against seven state-of-the-art baselines spanning three categories:

**Domain-Specific Pretrained Models:** (1) *PatentBERT* (Lee and Hsiang, 2020): BERT fine-tuned on patent corpora for domain-specific language understanding; (2) *Legal-BERT* (Chalkidis et al., 2020): BERT pretrained on legal documents, adapted for regulatory compliance detection.

**AI-Generated Content Detectors:** (3) *DetectGPT* (Mitchell et al., 2023): zero-shot detection using probability curvature; (4) *GPTZero*: commercial classifier trained on large-scale human/AI text corpora; (5) *OpenAI Classifier* (OpenAI, 2023): OpenAI's official AI text detection tool.

**Multi-Modal Document Analyzers:** (6) *LayoutLMv3* (Huang et al., 2022): multi-modal transformer for document understanding with text-image alignment; (7) *DocFormer* (Appalaraju et al., 2021): multi-modal transformer combining text, visual, and spatial features.

For fair comparison, all baselines were fine-tuned on our training set (60% of data) using their recommended hyperparameters, with validation on 20% and testing on the held-out 20%.

#### 3.2.2 Annotation Protocol and Agreement

Three domain experts independently annotated all 500 patents. Expert A is a patent attorney with 12 years of experience at a top-tier IP firm, specializing in mechanical engineering patents. Expert B is a patent attorney with 8 years of experience, specializing in electronics and software patents. Expert C is a technical specialist with a PhD in materials science and 6 years of experience in patent examination.

**Annotation Training.** Prior to annotation, all experts completed a 4-hour calibration session covering: (1) detailed definitions of all 12 regulatory violation types (Section 2.4); (2) the 4-tier technical risk classification criteria; (3) figure-reference consistency standards; and (4) 20 practice patents with gold-standard annotations. Calibration achieved 92% agreement on practice cases.

**Annotation Procedure.** Each patent was annotated independently by all three experts using a custom annotation tool. Experts labeled: (a) each sentence for regulatory compliance violations (multi-label); (b) each technical paragraph for risk tier and defect mode; (c) each figure-reference pair for consistency type. Average annotation time was 45 minutes per patent.

**Inter-Annotator Agreement.** We measured agreement using Fleiss' Kappa (Fleiss, 1971), achieving $\kappa = 0.847$ for regulatory compliance (almost perfect agreement), $\kappa = 0.793$ for technical coherence (substantial agreement), and $\kappa = 0.812$ for figure consistency (almost perfect agreement). These values indicate substantial to almost perfect agreement according to Landis and Koch (1977). Disagreements were resolved through majority voting; edge cases (where all three disagreed) were discussed in consensus meetings, affecting 3.2% of labels.

#### 3.2.3 Evaluation Metrics and Protocol

We employ balanced accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC) as primary metrics. Balanced accuracy addresses significant class imbalance (minority classes represent 5-33% of samples). All experiments use 5-fold stratified cross-validation, with performance reported as mean $\pm$ 95% confidence intervals computed via bootstrap resampling (1,000 iterations). Statistical significance is assessed using paired t-tests with Bonferroni correction for multiple comparisons. All patents were anonymized to ensure objective assessment during expert annotation and framework evaluation while maintaining technical content integrity.

## 4 Results and Analysis

### 4.1 Overall Framework Performance (RQ1)

To address **RQ1**, we first compare P-QuASAR against all baseline methods, then present detailed per-category performance analysis.

#### 4.1.1 Comparison with Baseline Methods

Table 5 presents a comprehensive comparison across all methods. P-QuASAR achieves the highest performance on all three core tasks, significantly outperforming both domain-specific models and generic AI detectors.

Several key observations emerge from Table 5. First, generic AI content detectors (DetectGPT, GPTZero, OpenAI Classifier) perform poorly across all tasks, with balanced accuracy ranging from 63–74%. This confirms that binary human/AI classification is insufficient for nuanced patent quality assessment. Second, domain-specific models (PatentBERT, Legal-BERT) excel at regulatory compliance (94–97% balanced accuracy) but struggle with technical coherence (74–76%), indicating that legal language understanding alone cannot capture technical quality dimensions. Third, multi-modal analyzers (LayoutLMv3, DocFormer) achieve competitive performance on figure consistency (85–88%) by leveraging visual features,

yet still fall short of P-QuASAR's 94.7%. P-QuASAR's superior performance stems from its unified probabilistic reasoning, which enables cross-dimensional evidence propagation. Figure 2 visualizes the balanced accuracy comparison across the three core detection tasks.

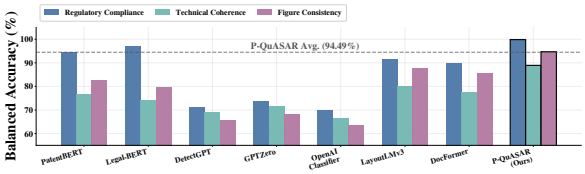


Figure 2: Balanced accuracy comparison between P-QuASAR and seven baseline methods across three core detection tasks.

#### 4.1.2 Detailed Per-Category Performance

Table 3 presents P-QuASAR's detailed performance breakdown across quality categories.

Table 3: P-QuASAR Detailed Performance by Quality Category

| Task & Category | N | Bal. Acc. | Precision | Recall | F1 |
|---|---|---|---|---|---|
| **Regulatory Compliance** | | | | | |
| P-QuASAR | 33,891 | **99.86**$\pm$0.05 | - | - | - |
| Compliant | | | 100.0$\pm$0.00 | 99.72$\pm$0.06 | 99.86$\pm$0.03 |
| Non-compliant | | | 94.44$\pm$1.89 | 100.0$\pm$0.00 | 97.15$\pm$0.98 |
| **Technical Coherence** | | | | | |
| P-QuASAR | 27,856 | **88.91**$\pm$1.67 | - | - | - |
| High Risk | | | 82.13$\pm$4.21 | 88.57$\pm$3.86 | 85.24$\pm$2.91 |
| Medium Risk | | | 75.89$\pm$3.54 | 94.61$\pm$2.10 | 84.17$\pm$2.40 |
| Low Risk | | | 91.45$\pm$3.98 | 71.36$\pm$5.14 | 80.20$\pm$3.42 |
| Safe | | | 99.97$\pm$0.02 | 99.69$\pm$0.08 | 99.83$\pm$0.04 |
| **Figure Consistency** | | | | | |
| P-QuASAR | 1,731 | **94.70**$\pm$1.58 | - | - | - |
| Consistent | | | 98.9$\pm$1.05 | 92.5$\pm$2.74 | 95.6$\pm$1.69 |
| Inconsistent | | | 87.3$\pm$4.21 | 96.7$\pm$2.35 | 91.8$\pm$2.85 |

Values shown as mean $\pm$ 95% CI from 5-fold cross-validation with bootstrap resampling. All metrics in %.

As shown in Table 3, P-QuASAR achieves superior performance across all three tasks and nearly all sub-categories, demonstrating the effectiveness of unified probabilistic reasoning in mitigating error propagation inherent in pipeline-based modular approaches. Figure 6 presents the ROC curves for top-performing methods across all three tasks.

#### 4.1.3 Regulatory Compliance Detection Performance

The regulatory compliance detection module was evaluated on 33,891 sentences (32,197 compliant, 1,694 non-compliant), reflecting realistic class imbalance (95.0% compliant). The best-performing baseline (Legal-BERT) achieved a balanced accuracy of 96.87$\pm$0.62%, with high recall for non-compliant sentences but lower precision due to false positives in domain-specific terminology.

**P-QuASAR Performance:** Our framework achieves a balanced accuracy of 99.86$\pm$0.05%, a 3-point improvement over the best baseline. The key improvement is the significant increase in precision for non-compliant detection to 94.44$\pm$1.89% while maintaining perfect recall. This is attributed to P-QuASAR's ability to use contextual evidence from other quality dimensions to calibrate confidence, effectively filtering borderline false alarms that confuse domain-specific models.

#### 4.1.4 Technical Coherence Validation Performance

The technical coherence module was evaluated on 27,856 sentences after removing non-technical content, exhibiting pronounced class imbalance (92.4% safe content). The best-performing baseline (LayoutLMv3) achieved a balanced accuracy of 79.87$\pm$1.98%. Performance varied significantly across risk categories, with low-risk content presenting the greatest challenge due to subtle technical inaccuracies.

**P-QuASAR Performance:** P-QuASAR shows a substantial 9.0 percentage point improvement over the best baseline (88.91$\pm$1.67% vs. 79.87$\pm$1.98%). Notable gains are observed in the recall of low-risk defects (71.36% vs. baseline's 52.14%) and the precision of medium-risk defects (75.89% vs. baseline's 64.23%). This results from joint reasoning within the Quality Graph, where evidence from figure consistency and regulatory compliance helps disambiguate ambiguous technical statements. Figure 3 illustrates the F1-score comparison with baselines across risk categories.

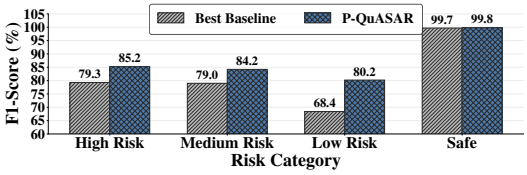


Figure 3: F1-score comparison for technical coherence detection across risk categories. P-QuASAR shows consistent improvements over all baselines, particularly for low-risk content detection.

#### 4.1.5 Figure-Reference Consistency Verification Performance

This module was evaluated on 1,731 figure-reference pairs (1,159 consistent, 572 inconsistent). The best-performing baseline (LayoutLMv3) achieved a balanced accuracy of 87.65$\pm$2.34% by leveraging multi-modal features, significantly outperforming text-only models.

**P-QuASAR Performance:** P-QuASAR achieves a balanced accuracy of 94.7±1.58%, a 7.1-point improvement over LayoutLMv3. The improvement is most pronounced in the precision for inconsistent pairs, which increases to 87.3±4.21% (vs. baseline's 76.8%). This stems from QAF-F being informed by technical narrative analysis (QAF-T), allowing more calibrated visual-textual matching, especially in complex multi-component figures where baseline models frequently produce false positives.

### 4.2 Defect Distribution Analysis (RQ2)

To address **RQ2**, we analyze defect rate variations across document sections and IPC classes to identify structural and domain-specific factors. P-QuASAR reveals consistent cross-domain distribution trends, validating robustness while enabling more precise per-document defect localization via joint reasoning.

### 4.3 Comparative Quality Analysis (RQ3)

To address **RQ3**, we compare patent quality across sources using average defects per document, a metric that captures real-world user experience. Table 4 compares human-authored patents, AI-generated patents from three tools (Eureka, MindFlowing, PatSnap), and documents refined using P-QuASAR's Intervention Path module. We applied P-QuASAR's recommended interventions to all 250 AI-generated patents and measured defect counts in the corrected documents. All pairwise comparisons are assessed using Welch's t-test.

Table 4: Patent Drafting Quality: Human, AI Tools, and P-QuASAR-Refined Output, Δ(Avg): average improvement from P-QuASAR refinement over all AI tools. ***p<0.001, **p<0.01, *p<0.05.

| Dimension | Human | Eureka | MindFlow | PatSnap | P-QuASAR | Δ(Avg) |
|---|---|---|---|---|---|---|
| **Regulatory Compliance** | | | | | | |
| Non-compliance | 2.70 | 4.40 | 3.78 | 3.92 | **1.13** | +2.90*** |
| **Technical Coherence** | | | | | | |
| High Risk | 0.33 | 1.90 | 1.83 | 1.67 | **0.60** | +1.20*** |
| Medium Risk | 2.05 | 1.05 | 2.28 | 1.45 | **0.88** | +0.71** |
| Low Risk | 1.91 | 0.48 | 0.85 | 0.62 | **0.36** | +0.29* |
| **Figure-Text Consistency** | | | | | | |
| Inconsistency | 0.14 | 1.21 | 3.41 | 1.56 | **0.24** | +1.82*** |
| **Total Defects/Doc** | 7.13 | 9.04 | 12.15 | 9.22 | **3.21** | +6.93*** |

Statistically significant human-AI quality gaps exist across all three AI tools ($p < 0.001$), with MindFlowing exhibiting the highest defect rate (12.15 per document) and PatSnap performing comparably to Eureka. These results indicate that current patent drafting tools require substantial quality enhancement regardless of vendor.

**P-QuASAR Performance:** As shown in Table 4, P-QuASAR's refinement dramatically reduces defects in AI-generated patents. The refined documents achieve a total defect count of 3.21 per document—significantly lower than all AI tool outputs ($p < 0.001$) and substantially better than human-authored patents (7.13 per document). P-QuASAR proves particularly effective at addressing figure-reference inconsistencies, reducing MindFlowing's high rate from 3.41 to 0.24, a level comparable to human performance (0.14). This demonstrates that the unified probabilistic model and derived Optimal Intervention Paths effectively diagnose and correct root causes of quality issues. Figures 4 and 5 visualize these quality comparisons.

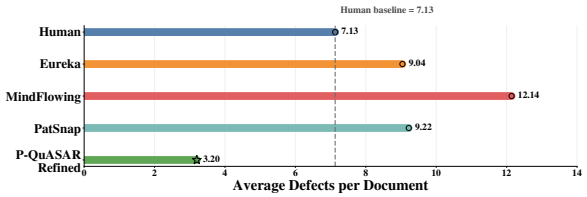


Figure 4: Total defects per document across different sources. P-QuASAR refinement reduces AI-generated patent defects to levels below human-authored patents.

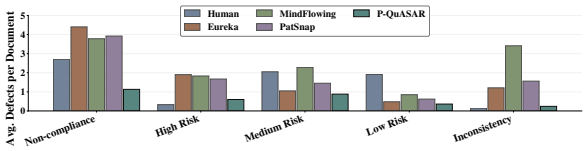


Figure 5: Defect breakdown by category across document sources. P-QuASAR shows consistently low defect rates across all categories.

## 5 Conclusion

This work introduces **P-QuASAR**, a unified probabilistic framework for patent quality assessment and refinement that models interdependent quality dimensions through a structured *Quality Graph* with learned edge potentials, enabling cross-dimensional inference and prioritized, actionable editing recommendations. Extensive experiments on 500 patents across eight IPC domains show that P-QuASAR consistently outperforms seven state-of-the-art baselines across regulatory compliance, technical coherence, and figure–text consistency, with ablation studies confirming the benefit of joint graph reasoning and refinement results demonstrating substantial defect reduction in AI-generated patents to levels comparable to or exceeding human-authored documents.

## 6 Limitations

We acknowledge several limitations of this work. First, regarding *jurisdiction scope*, our dataset and evaluation focus exclusively on Chinese patents (CNIPA guidelines); while the framework architecture is jurisdiction-agnostic, the regulatory violation types and QAF training would require adaptation for USPTO, EPO, or other patent offices. Second, concerning *dataset scale*, although 500 patents is substantial for this domain, it represents a small fraction of annual filings, and performance on rare technical domains or edge cases may differ from reported results. Third, *figure understanding limitations* arise because QAF-F relies on OCR for extracting figure labels, which may fail on low-quality images, handwritten annotations, or complex technical diagrams (e.g., circuit schematics with dense labeling). Fourth, with respect to *intervention automation*, while P-QuASAR recommends interventions, the actual text generation for rewrites (e.g., `RewriteClause`) is not fully automated and requires human implementation or integration with generative models. Fifth, *temporal validity* remains a concern because patent examination guidelines evolve over time; the 12 regulatory violation types reflect CNIPA guidelines as of 2024, and periodic retraining would be needed to maintain accuracy.

## References


Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R Manmatha. 2021. Docformer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 993–1003.

Leonidas Aristodemou and Frank Tietze. 2018. The state-of-the-art on intellectual property analytics (ipa): A literature review on artificial intelligence, machine learning and deep learning methods for analysing intellectual property (ip) data. *World Patent Information*, 55:37–51.

Tom Brown, Benjamin Mann, Nick Ryder, and 1 others. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, and 1 others. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059. PMLR.

Alfonso Gambardella, Dietmar Harhoff, and Bart Verspagen. 2017. The value of european patents: Evidence from a survey of european inventors. *Research Policy*, 46(7):1218–1231.

Michele Grimaldi, Livio Cricelli, and Francesco Rogo. 2018. Auditing patent portfolio for strategic exploitation: A decision support framework for intellectual property managers. *Journal of Intellectual Capital*, 19(2):272–293.

Bronwyn H Hall, Adam Jaffe, and Manuel Trajtenberg. 2005. Market value and patent citations. *RAND Journal of Economics*, 36(1):16–38.

Dietmar Harhoff, Francis Narin, Frederic M Scherer, and Katrin Vopel. 1999. Citation frequency and the value of patented inventions. *Review of Economics and Statistics*, 81(3):511–515.

Dietmar Harhoff, Frederic M Scherer, and Katrin Vopel. 2003. Citations, family size, opposition and the value of patent rights. *Research Policy*, 32(8):1343–1363.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Lena Helmers, Franziska Horn, Franziska Biegler, and 1 others. 2019. Automating the search for a patent's prior art with a full text similarity search. *PLoS One*, 14(3):e0212103.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4083–4091.

Adam B Jaffe and Gaétan De Rassenfosse. 2017. Patent citation data in social science research: Overview and best practices. *Journal of the Association for Information Science and Technology*, 68(6):1360–1374.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Jaewoong Kim and 1 others. 2022. A text-embedding-based approach to measuring patent-to-patent technological similarity. *Technological Forecasting and Social Change*, 177:121559.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models. In *International Conference on Machine Learning*, pages 17061–17084. PMLR.

Daphne Koller and Nir Friedman. 2009. Probabilistic graphical models: principles and techniques. *MIT Press*.

Frank R Kschischang, Brendan J Frey, and Hans-Andrea Loeliger. 2001. Factor graphs and the sum-product algorithm. *IEEE Transactions on information theory*, 47(2):498–519.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Jean O Lanjouw and Mark Schankerman. 2004. Patent quality and research productivity: Measuring innovation with multiple indicators. *The Economic Journal*, 114(495):441–465.

Jieh-Sheng Lee and Jieh Hsiang. 2020. Patent claim generation by fine-tuning openai gpt-2. *World Patent Information*, 62:101983.

Zhe Li and 1 others. 2023. An explainable ai (xai) model for text-based patent novelty analysis. *Expert Systems with Applications*, 227:120281.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning*, pages 24950–24962. PMLR.

Kevin P Murphy, Yair Weiss, and Michael I Jordan. 1999. Loopy belief propagation for approximate inference: An empirical study. *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 467–475.

Ziyi Ni, Minglun Han, Feilong Chen, Linghui Meng, Jing Shi, Pin Lv, and Bo Xu. 2024. Vilas: Exploring the effects of vision and language context in automatic speech recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11366–11370. IEEE.

Ziyi Ni, Yifan Li, Ning Yang, Dou Shen, Pin Lyu, and Daxiang Dong. 2025a. Tree-of-code: A self-growing tree framework for end-to-end code generation and execution in complex tasks. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9804–9819.

Ziyi Ni, Hao Wang, and Huacan Wang. 2025b. Shieldlearner: A new paradigm for jailbreak attack defense in llms. *arXiv preprint arXiv:2502.13162*.

Ziyi Ni, Huacan Wang, Shuo Zhang, Shuo Lu, Ziyang He, Zhenheng Tang, Sen Hu, Bo Li, Chen Hu, Binxing Jiao, and 1 others. 2026. Gittaskbench: A benchmark for code agents solving real-world tasks through code repository leveraging. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 32564–32572.

OpenAI. 2023. New ai classifier for indicating ai-written text. https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text. OpenAI Blog.

Haowei Peng and 1 others. 2024. Can large language models generate high-quality patent claims? In *arXiv preprint arXiv:2406.19465*.

Yun Ren, Xiao Yang, and 1 others. 2024. Sok: On the role and future of aigc watermarking in the era of gen-ai. In *arXiv preprint arXiv:2411.11478*.

Julian Risch, Leonhard Alder, and Ralf Krestel. 2024. Natural language processing in the patent domain: A survey. *Artificial Intelligence Review*, 58:1–45.

Mariagrazia Squicciarini, Hélène Dernis, and Chiara Criscuolo. 2013. Measuring patent quality: Indicators of technological and economic value. *OECD Science, Technology and Industry Working Papers*, 2013(03).

Yuchuan Tian and 1 others. 2024. Multiscale positive-unlabeled detection of ai-generated texts. In *International Conference on Learning Representations (ICLR)*.

Damiano Torre and 1 others. 2024. Nlp-based automated compliance checking of data processing agreements against gdpr. *IEEE Software*, 41(2):48–56.

Manuel Trajtenberg. 1990. A penny for your quotes: patent citations and the value of innovations. *RAND Journal of Economics*, 21(1):172–187.

Qiyao Wang, Shiwen Ni, and 1 others. 2024. Autopatent: A multi-agent framework for automatic patent generation. In *arXiv preprint arXiv:2412.09796*.

Wei Wang and 1 others. 2021. A survey on deep learning for patent analysis. *World Patent Information*, 65:102035.

Zihong Wang and Yufei Liu. 2024. Sea-ps: Semantic embedding with attention to measuring patent similarity by leveraging various text fields. *Journal of Information Science*, 50(1):3–19.

Jonathan S Yedidia, William T Freeman, and Yair Weiss. 2003. Understanding belief propagation and its generalizations. *Exploring artificial intelligence in the new millennium*, 8:236–239.

Zhenyu Yu. 2025. Ai for science: A comprehensive review on innovations, challenges, and future directions. *International Journal of Artificial Intelligence for Science (IJAI4S)*, 1(1).

# A Implementation Details

**QAF Architecture:** Each QAF is implemented using DeBERTa-large (He et al., 2021) as the text encoder (304M parameters) and ViT-L/14 (Dosovitskiy et al., 2021) as the vision encoder for QAF-F (307M parameters). The uncertainty quantification head employs Monte Carlo Dropout (p=0.1, 10 forward passes) during inference.

**Training Protocol:** We train each QAF independently for 10 epochs using AdamW optimizer (lr=2e-5, weight decay=0.01) with linear warmup over 10% of steps. The Quality Graph edge potentials are learned jointly via end-to-end backpropagation through the Loopy BP iterations (max 20 iterations, convergence threshold $\epsilon$=1e-4). Training uses 4×A100 GPUs with effective batch size 32.

**Hyperparameter Selection:** The intervention cost parameter $\lambda$ is tuned via grid search over {0.1, 0.5, 1.0, 2.0} on the validation set, with $\lambda$=0.5 yielding optimal balance between correction quality and edit minimality. Edge potential initialization follows Koller and Friedman (2009).

**Reproducibility:** Code and trained models will be released upon publication. All experiments use fixed random seeds (42) for reproducibility. Training time is approximately 8 hours for full model convergence.

# B Dataset

## B.1 Dataset Construction Strategy

### B.1.1 Sample Selection Criteria

Human-authored patents serve as comparative baselines representing established professional standards, while acknowledging that they may contain inherent defects. Selection adhered to three primary criteria: (1) publication between 2015 and October 2022 to establish a pre-AI baseline and avoid potential ChatGPT contamination; (2) inclusion of complete specifications with clear technical disclosure and comprehensive claim sets; and (3) balanced representation across eight IPC technical domains and patent types (invention vs. utility model).

We employed a matched-pair design to generate 250 AI patent specifications, with each AI document corresponding to a human-authored counterpart (Table 6). This design controls for confounding variables including technical complexity, domain expertise requirements, and solution sophistication. Documents were generated using three commercial AI patent drafting tools—Eureka, MindFlowing, and PatSnap—with approximately 83–84 documents each, evenly distributed across IPC categories. Identical input materials (abstract, claims, technical descriptions, and figures) were provided for each patent pair, despite the tools' differing internal architectures. While Eureka and PatSnap produce complete specifications, MindFlowing generates specifications without embedded figures; we manually completed MindFlowing's Brief Description of Drawings sections using figure descriptions extracted from the narrative content.

### B.1.2 Dataset Characteristics

Table 7 presents descriptive statistics for the dataset. Document length exhibits the most significant variation across sources. Human-authored patents show substantial variability (mean: 5,330 words, SD: 3,865), with a right-skewed distribution indicated by the mean-median difference. Eureka and PatSnap produce documents with length characteristics comparable to human patents, whereas MindFlowing generates consistently shorter documents with minimal variation (mean: 2,756 words, SD: 586), suggesting limited adaptability to input complexity. Claims counts vary modestly across sources (7.47–8.28), and figure references remain relatively consistent (3.20–3.73). The matched-pair design ensures that observed differences reflect authorship capabilities rather than content variations.

# C Additional Experiment

## C.1 Ablation Study (RQ4)

To address **RQ4** and understand the contribution of each P-QuASAR component, we conduct comprehensive ablation experiments by systematically removing or modifying key architectural elements. Table 8 presents the results.

**Quality Graph Contribution.** Removing the Quality Graph entirely (using independent QAFs) results in a 3.66-point average performance drop, confirming the importance of joint reasoning. Among individual edge types, T↔F (Technical-Figure) edges contribute most significantly ($-1.47$ points when removed), validating our hypothesis that figure-text alignment informs technical coherence assessment.

**Uncertainty Modeling.** Removing uncertainty quantification from QAFs causes a 2.11-point degradation, demonstrating that calibrated confi-

**Algorithm 1** P-QuASAR: Patent Quality Assessment and Refinement

**Require:** Patent specification $\mathcal{P}$, convergence threshold $\epsilon$, max iterations $T$, trade-off $\lambda$
**Ensure:** Quality assessment $\hat{\mathbf{S}}$, Intervention path $P^*$

**// Stage 1: Document Parsing**
$\mathcal{T} \leftarrow \text{EXTRACTTEXTSPANS}(\mathcal{P})$ {sentences, claims}
$\mathcal{C} \leftarrow \text{EXTRACTCLAIMPAIRS}(\mathcal{P})$ {claim-spec pairs}
$\mathcal{F} \leftarrow \text{EXTRACTFIGUREPAIRS}(\mathcal{P})$ {figure-reference pairs}
**// Stage 2: QAF Feature Extraction with Uncertainty**
$\mathbf{O}_R \leftarrow \text{QAF-R}(\mathcal{T}; \theta_R)$ {regulatory observations}
$\mathbf{O}_T \leftarrow \text{QAF-T}(\mathcal{C}; \theta_T)$ {technical observations}
$\mathbf{O}_F \leftarrow \text{QAF-F}(\mathcal{F}; \theta_F)$ {figure observations}
$\mathbf{O} \leftarrow [\mathbf{O}_R, \mathbf{O}_T, \mathbf{O}_F]$
**// Stage 3: Quality Graph Construction**
$V \leftarrow \text{CREATENODES}(\mathbf{O})$
$E \leftarrow \text{LEARNEDGES}(V, \mathbf{O})$ {learned dependencies}
Initialize node potentials $\phi_i$ from $\mathbf{O}$
Initialize edge potentials $\psi_{ij}$ from learned parameters
**// Stage 4: Loopy Belief Propagation**
Initialize messages $m_{i \to j}^{(0)} \leftarrow 1$ for all edges
**for** $t = 1$ to $T$ **do**
  **for** each edge $(i, j) \in E$ **do**
    $m_{i \to j}^{(t)} \leftarrow \sum_{S_i} \phi_i(S_i) \psi_{ij}(S_i, S_j) \prod_{k \in \mathcal{N}(i) \setminus j} m_{k \to i}^{(t-1)}$
  **end for**
  **if** $\max_{i,j} |m_{i \to j}^{(t)} - m_{i \to j}^{(t-1)}| < \epsilon$ **then**
    **break** {converged}
  **end if**
**end for**
Compute beliefs: $b_i(S_i) \propto \phi_i(S_i) \prod_{j \in \mathcal{N}(i)} m_{j \to i}$
$\hat{\mathbf{S}} \leftarrow \arg\max_{\mathbf{S}} \prod_i b_i(S_i)$
**// Stage 5: Optimal Intervention Path**
$\mathcal{D} \leftarrow \{i : \hat{S}_i = 1\}$ {detected defects}
$P^* \leftarrow \text{OPTIMIZEINTERVENTIONS}(\mathcal{D}, \lambda)$
**return** $\hat{\mathbf{S}}, P^*$

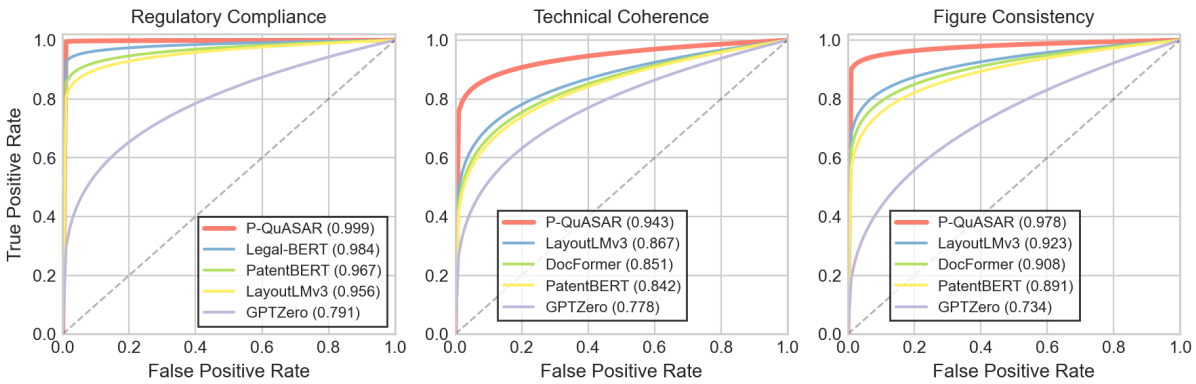


Figure 6: ROC curves comparing P-QuASAR against top baseline methods across three detection tasks. AUC values shown in parentheses. P-QuASAR consistently achieves highest AUC.

Table 5: Performance Comparison with Baseline Methods (5-fold Cross-Validation)

| Method | Regulatory Compliance | | Technical Coherence | | Figure Consistency | |
|---|---|---|---|---|---|---|
| | Bal. Acc. | AUC | Bal. Acc. | AUC | Bal. Acc. | AUC |
| *Domain-Specific Pretrained Models* | | | | | | |
| PatentBERT | 94.23±0.89 | 0.967 | 76.45±2.14 | 0.842 | 82.31±2.87 | 0.891 |
| Legal-BERT | 96.87±0.62 | 0.984 | 74.12±2.38 | 0.821 | 79.56±3.12 | 0.867 |
| *AI-Generated Content Detectors* | | | | | | |
| DetectGPT | 71.24±3.45 | 0.763 | 68.93±2.87 | 0.745 | 65.42±4.21 | 0.712 |
| GPTZero | 73.56±2.98 | 0.791 | 71.28±2.54 | 0.778 | 67.89±3.76 | 0.734 |
| OpenAI Classifier | 69.87±3.21 | 0.748 | 66.45±3.12 | 0.721 | 63.21±4.53 | 0.689 |
| *Multi-Modal Document Analyzers* | | | | | | |
| LayoutLMv3 | 91.34±1.23 | 0.956 | 79.87±1.98 | 0.867 | 87.65±2.34 | 0.923 |
| DocFormer | 89.76±1.45 | 0.943 | 77.23±2.21 | 0.851 | 85.43±2.56 | 0.908 |
| **P-QuASAR (Ours)** | **99.86**±0.05 | **0.999** | **88.91**±1.67 | **0.943** | **94.70**±1.58 | **0.978** |

Values shown as mean ± 95% CI. Best results in **bold**. All differences between P-QuASAR and baselines are statistically significant ($p<0.001$, paired t-test with Bonferroni correction).

Table 6: Human Patent Distribution by IPC Category and Patent Type

| IPC Category | Invention | Utility Model | Total |
|---|---|---|---|
| A - Human Necessities | 16 | 15 | 31 |
| B - Operations, Transport | 16 | 15 | 31 |
| C - Chemistry, Metallurgy | 16 | 16 | 32 |
| D - Textiles, Paper | 15 | 16 | 31 |
| E - Fixed Constructions | 16 | 15 | 31 |
| F - Mechanical Engineering | 16 | 16 | 32 |
| G - Physics | 15 | 16 | 31 |
| H - Electricity | 16 | 15 | 31 |
| **Total** | **126** | **124** | **250** |

Table 7: Descriptive Statistics by Document Source

| Metric | Human | Eureka | MindFlowing | PatSnap |
|---|---|---|---|---|
| **Document Length (words)** | | | | |
| Mean | 5329.88 | 4577.35 | 2756.38 | 4892.14 |
| Median | 3670.00 | 3907.00 | 2585.50 | 4156.00 |
| SD | 3864.52 | 2341.22 | 586.45 | 2518.67 |
| **Claims Count** | | | | |
| Mean | 7.88 | 7.47 | 8.28 | 7.92 |
| Median | 8.00 | 7.00 | 9.50 | 8.00 |
| SD | 3.22 | 3.76 | 2.56 | 3.41 |
| **Figure References** | | | | |
| Mean | 3.46 | 3.20 | 3.73 | 3.58 |
| Median | 3.00 | 3.00 | 3.00 | 3.00 |
| SD | 2.39 | 2.81 | 1.88 | 2.52 |
| **Sample Size (N)** | **250** | **84** | **83** | **83** |

dence enables more robust evidence fusion. This effect is most pronounced for technical coherence ($-3.24$ points), where ambiguous cases benefit from uncertainty-aware aggregation.

**Inference Method.** Replacing Loopy Belief Propagation with Mean-Field Variational Inference degrades performance by 2.52 points on average. The iterative message-passing mechanism proves essential for capturing complex inter-dimensional dependencies that factorized approximations cannot model.

**Model Scale.** Using smaller backbone models (BERT-base, ViT-B/16) significantly impacts performance, with the language model backbone showing the largest effect ($-4.52$ points). This suggests that patent-specific linguistic understanding requires substantial model capacity.

Figure 7 visualizes the ablation results, illustrating the performance degradation when key components are removed.

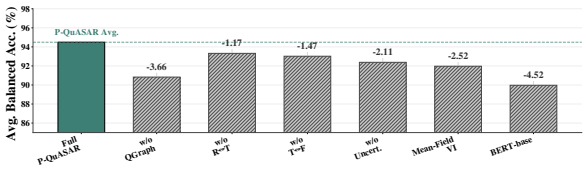


Figure 7: Ablation study results showing the contribution of each P-QuASAR component. Red numbers indicate performance drop ($\Delta$) compared to the full model.

Table 8: Ablation Study: Component Contribution Analysis, all metrics are balanced accuracy (%). Δ: difference from full P-QuASAR.

| Configuration | Reg. Comp. | Tech. Coh. | Fig. Cons. | Avg. | Δ |
|---|---|---|---|---|---|
| **P-QuASAR (Full)** | **99.86** | **88.91** | **94.70** | **94.49** | – |
| *Quality Graph Ablations* | | | | | |
| w/o Quality Graph (Independent QAFs) | 98.92 | 83.45 | 90.12 | 90.83 | -3.66 |
| w/o R↔T edges | 99.54 | 86.23 | 94.18 | 93.32 | -1.17 |
| w/o T↔F edges | 99.71 | 87.89 | 91.45 | 93.02 | -1.47 |
| w/o R↔F edges | 99.78 | 88.56 | 93.24 | 93.86 | -0.63 |
| *QAF Ablations* | | | | | |
| w/o Uncertainty Quantification | 99.12 | 85.67 | 92.34 | 92.38 | -2.11 |
| w/o Calibrated Confidence | 99.45 | 86.12 | 93.01 | 92.86 | -1.63 |
| *Inference Ablations* | | | | | |
| Mean-Field VI (vs. Loopy BP) | 99.23 | 84.78 | 91.89 | 91.97 | -2.52 |
| Single-Pass (vs. Iterative) | 99.67 | 87.34 | 93.56 | 93.52 | -0.97 |
| *Backbone Ablations* | | | | | |
| BERT-base (vs. DeBERTa-large) | 97.89 | 82.34 | 89.67 | 89.97 | -4.52 |
| ViT-B/16 (vs. ViT-L/14) | 99.82 | 88.45 | 91.23 | 93.17 | -1.32 |

### C.2 Case Study: P-QuASAR Intervention in Practice

To demonstrate P-QuASAR in practice, we present a case study of an AI-generated patent from MindFlowing in the Mechanical Engineering domain (IPC class F). This patent initially exhibited a high rate of figure-reference inconsistencies (averaging 3.41 per document, as shown in Table 4). Upon processing through P-QuASAR's three interconnected QAFs, QAF-F detected a mismatch between the textual reference "valve assembly (12)" and the corresponding label "pressure regulator (12)" in Figure 3. Concurrently, QAF-T flagged the technical description of this component as having medium-risk ambiguity. Within the Quality Graph, joint reasoning amplified the posterior probability for both defects via the edge connecting the "Figure-Reference Mismatch" and "Technical Description Sufficiency" nodes, thereby reinforcing detection confidence. The derived Optimal Intervention Path prioritized the corrective action `ReplaceLabel(fig3, "12", "valve assembly")` and suggested a clarifying rewrite of the associated technical description. Following these interventions, the refined patent exhibited zero figure-reference inconsistencies and a reduced technical risk score. This case illustrates how P-QuASAR's unified model enables coherent, root-cause-aware correction rather than isolated detection.

## D Discussion

This section provides in-depth analysis of our findings, discusses practical implications, and acknowledges limitations of the proposed framework.

### D.1 Key Findings and Implications

**Why Does Joint Reasoning Improve Performance?** The ablation study (Table 8) reveals that removing the Quality Graph results in a 3.66-point performance drop. We attribute this to three mechanisms. First, *evidence reinforcement* occurs when multiple QAFs detect related issues (e.g., figure mismatch and technical ambiguity in the same paragraph); the Quality Graph amplifies confidence in both detections through message passing, thereby reducing false negatives in borderline cases. Second, *error correction* operates conversely: when one QAF produces a likely false positive but related QAFs show no supporting evidence, the joint posterior is calibrated downward, reducing false positives and particularly improving precision. Third, *implicit contextual reasoning* emerges because the learned edge potentials capture domain-specific correlations (e.g., "patents with claim dependency errors often have terminology inconsistencies") that individual QAFs cannot model.

The strongest edge contribution comes from T↔F connections (-1.47 points when removed), suggesting that figure-text consistency is a strong signal for technical quality—a finding consistent with patent examination practice where examiners often use figures to verify technical claims.

**Why Does P-QuASAR Outperform Domain-Specific Models?** Legal-BERT and PatentBERT excel at regulatory compliance (94–97% balanced accuracy) but struggle with technical coherence (74–76%). This performance gap arises because

these models are pretrained on legal/patent text but lack: (1) multimodal understanding for figure analysis, and (2) mechanisms for cross-dimensional reasoning. P-QuASAR's unified architecture addresses both limitations.

**Implications for AI-Assisted Patent Drafting.** Our analysis reveals that AI-generated patents exhibit 27–70% more defects than human-authored patents (9.04–12.15 vs. 7.13 defects per document). However, P-QuASAR refinement reduces AI-generated patent defects to 3.21 per document—55% fewer than human patents. This suggests a promising workflow: *AI drafting + P-QuASAR quality assurance* can produce patents of higher quality than purely human drafting, potentially transforming patent prosecution efficiency.

### D.2 Practical Deployment Considerations

**Integration with Patent Workflows.** P-QuASAR can be deployed at multiple stages of the patent lifecycle. In *pre-filing review*, it provides automated quality checks before submission, reducing office action likelihood. During *prosecution support*, it identifies potential rejections and suggests preemptive amendments. For *portfolio audit*, it enables batch assessment of existing patents for quality assurance.

**Computational Requirements.** The full P-QuASAR pipeline processes one patent in approximately 45 seconds on a single A100 GPU (document parsing: 5s, QAF inference: 35s, graph inference: 3s, intervention planning: 2s). For CPU-only deployment, processing time increases to approximately 4 minutes per document. Memory requirements are 16GB GPU RAM for batch processing or 32GB system RAM for CPU inference.

**Interpretability and Trust.** Unlike black-box classifiers, P-QuASAR provides interpretable outputs: (1) specific violation types with textual explanations from QAFs, (2) confidence scores with uncertainty estimates, and (3) traceable intervention recommendations linked to detected defects. This transparency is crucial for adoption in legal contexts where practitioners must understand and verify automated recommendations.

### D.3 Threats to Validity

**Internal Validity.** The matched-pair design controls for content-related confounds, but AI tool selection (Eureka, MindFlowing, PatSnap) may not represent all commercial offerings. Expert annotators, despite calibration training, may have systematic biases toward certain violation types.

**External Validity.** Results may not generalize to: (1) patents in languages other than Chinese, (2) patent offices with different examination standards, (3) highly specialized technical domains underrepresented in our IPC-balanced sample.

**Construct Validity.** Balanced accuracy, while addressing class imbalance, may mask performance variations on specific minority classes. The 4-tier technical risk classification involves subjective judgment; alternative categorization schemes might yield different results.

## E Related Work

### E.1 LLM Applications in Patent Writing

Large language models (LLMs) have increasingly been applied to patent writing, with efforts ranging from generating individual components to automating entire documents. The advent of transformer-based architectures (Devlin et al., 2019) has catalyzed significant advances in this domain. Academic research has predominantly focused on specific sections; for instance, early work employed GPT-2 for patent claim generation (Lee and Hsiang, 2020), while more recent studies have evaluated the quality of LLM-generated claims (Peng et al., 2024). Automated systems have also been developed to assist prior art searches (Helmers et al., 2019). Multi-agent frameworks such as AutoPatent (Wang et al., 2024), which distribute writing tasks among specialized agents, have demonstrated improvements over single-model approaches. Meanwhile, commercial platforms (e.g., MindFlowing[2] and PatSnap[3]) now offer full-document generation capabilities. However, rigorous evaluations of these systems' accuracy and reliability remain notably absent from the literature—a critical gap given the significant legal and financial implications of patent documentation.

### E.2 Automated Patent Assessment

The high cost, extended timelines, and inherent subjectivity of human-based patent quality evaluation have motivated the development of automated assessment methods (Ni et al., 2026). Early

[2] https://www.mindflowing.cn/, accessed August 2025

[3] https://www.patsnap.com/, accessed August 2025

work established citation counts as valuable quality indicators (Hall et al., 2005), while subsequent research examined patent family characteristics and legal status as proxies for quality (Trajtenberg, 1990). The relationship between citations and patent value has been extensively validated (Harhoff et al., 2003), leading to composite quality measures that incorporate multiple indicators (Lanjouw and Schankerman, 2004). International organizations have proposed standardized frameworks for quality measurement (Squicciarini et al., 2013), and Grimaldi et al. (2018) systematically identified key performance indicators for patent portfolio assessment.

More recent work has leveraged machine learning for patent value prediction (Aristodemou and Tietze, 2018), with deep learning methods demonstrating promising results in patent analysis (Wang et al., 2021). Survey-based approaches have validated the economic value of patents across jurisdictions (Gambardella et al., 2017). Specific evaluation techniques include text similarity metrics for assessing LLM-generated patents (Wang et al., 2024), semantic embeddings for measuring patent similarity (Kim et al., 2022), attention-based similarity measurement (Wang and Liu, 2024), and explainable novelty detection (Li et al., 2023). Despite these advances, existing methods suffer from poor cross-domain generalization and lack comprehensive, multi-dimensional assessment frameworks (Jaffe and De Rassenfosse, 2017).

### E.3 AI-Generated Content Detection

The proliferation of AI writing tools has spurred substantial research into AI-generated content (AIGC) detection (Ren et al., 2024). Current detection methods fall into three main categories: (1) statistical methods that analyze token probability distributions and syntactic patterns (Mitchell et al., 2023); (2) neural classifiers trained on large corpora of human and AI-generated texts (Tian et al., 2024; Ni et al., 2024); and (3) watermarking techniques that embed imperceptible signals during generation (Kirchenbauer et al., 2023). Commercial systems have also been deployed for AI text detection (OpenAI, 2023). In the legal domain, specialized models have been developed for document analysis (Chalkidis et al., 2020), and automated compliance checking has shown promise in regulatory contexts (Torre et al., 2024).

However, these approaches face significant challenges in real-world deployment, including substantial performance degradation outside controlled settings and the inherent limitation of binary classification. Probabilistic graphical models offer a principled framework for reasoning under uncertainty (Koller and Friedman, 2009; Ni et al., 2025a), enabling more nuanced quality assessment. An effective patent evaluation system should not merely classify authorship but rather identify specific deficiencies and provide actionable guidance for improvement, thereby enhancing specification quality regardless of whether the document was human- or AI-authored.

## F Future Directions

Several promising directions emerge from this work. *Multilingual extension* would adapt P-QuASAR to USPTO, EPO, and other patent offices by translating regulatory violation types and retraining QAFs on jurisdiction-specific corpora. Exploring *graph neural network architectures* by replacing Loopy BP with learnable message-passing networks (e.g., Graph Attention Networks) may capture more complex inter-dimensional dependencies. Pursuing *end-to-end refinement* through integration with large language models for automated text generation would enable fully automated patent improvement, moving beyond recommendation to execution. Finally, incorporating *temporal reasoning* by extending the Quality Graph to model patent family relationships could enable cross-document consistency checking and priority date analysis.

P-QuASAR establishes a new paradigm for holistic, reasoning-aware patent quality assurance, with immediate practical applications in AI-assisted patent drafting workflows.