

Adaptive Cost-Efficient Evaluation for Reliable Patent Claim Generation

Yongmin Yoo, Qiongkai Xu, Longbing Cao

Frontier AI Research Centre, Macquarie University

School of Computing, FSE, Macquarie University

yooyongmin91@gmail.com {qiongkai.xu, longbing.cao}@mq.edu.au

Abstract

Automated validation of patent claims demands near-zero error tolerance, yet existing methods are trapped in a rigidity-resource dilemma: lightweight encoders miss long-range legal dependencies, while exhaustive LLM verification is prohibitively costly at scale. Critically, naive confidence-based cascades cannot resolve this trade-off, as binary validity scores conflate heterogeneous error types that demand different reasoning depths. We introduce Adaptive Cost-efficient Evaluation (ACE), which instead exploits the categorical structure of patent errors for uncertainty-aware routing. A fine-tuned encoder projects each claim into a $K+1$ distribution over legal error types; its predictive entropy determines whether the claim is resolved immediately or escalated to an expert LLM that performs a schema-constrained Chain-of-Patent-Thought (CoPT) protocol grounded in 35 U.S.C. standards. On our newly constructed 40k-claim benchmark, ACE outperforms a fine-tuned 70B-parameter LLM while cutting cost by 78%, and its routing threshold transfers to real USPTO rejections without re-calibration, cutting inference time by 60%. Beyond patents, this suggests that uncertainty itself can serve as a generalizable routing signal for high-stakes verification tasks with heterogeneous error structures.

1 Introduction

Validating generated patent claims poses a uniquely high-stakes challenge in natural language processing. Unlike general prose, patent claims simultaneously function as technical descriptions and legally binding definitions of intellectual property rights. In this domain, a single word choice error, antecedent inconsistency, or logical contradiction can render a claim legally defective, with significant financial and legal repercussions (USPTO, 2024; Faber and Kowalski, 2023). Consequently, automating the validation of generated patent claims, which we focus on

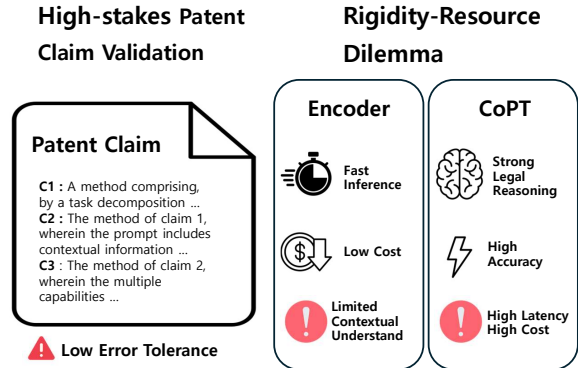


Figure 1: High-stakes patent claim validation requires low error tolerance while demanding cost efficiency. Adaptive Cost-efficient Evaluation (ACE) leverages the throughput of encoder models and the inference accuracy of Chain-of-Patent-Thought (CoPT) reasoning.

here as structural validity under 35 U.S.C. §112(b), requires rigorous adherence to legal and structural standards, such as antecedent consistency and logical soundness, while far exceeding the approaches for evaluating standard text generation. This growing demand is mirrored in a rapidly expanding body of patent-specific NLP research spanning both claim generation (Yoo et al., 2026b) and patent evaluation (Yoo et al., 2026c), yet, as we detail next, the cost-accuracy trade-off central to large-scale legal validation remains unresolved.

With these stringent requirements, existing text generation evaluation methodologies fail to reconcile validation accuracy with computational scalability. Reference-based metrics, including BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and BERTScore (Zhang* et al., 2020), rely on surface-level n-gram overlap or semantic embedding similarity. These metrics are not designed to diagnose structural flaws, such as antecedent basis violations or circular dependencies, which are critical for legal validity (USPTO, 2024). Conversely, while the LLM-as-a-Judge approach offers stronger reasoning capabilities, utilizing LLMs as unconstrained evaluators introduces a critical con-

fluct between industrial-scale throughput and the high reliability required in legal review. First, the extreme computational demand of large-scale models creates a significant scalability bottleneck for exhaustive verification of millions of patent claims. Second, even when resources are available, the inherent stochasticity of generative systems without structured legal grounding often produces plausible but incorrect justifications. Consequently, this combination of substantial computational overhead and probabilistic instability falls short of the reliability and transparency requirements of legal review, crystallizing into a rigidity-resource dilemma: lightweight methods lack legal rigor, while rigorous methods lack scalability.

To mitigate the rigidity-resource dilemma, checklist-based evaluations offer a transparent alternative by decomposing complex legal validity criteria into granular verification items, thereby ensuring the interpretability needed in legal review (Ribeiro et al., 2020; Lee et al., 2025). However, current implementations typically enforce a static level of scrutiny across all instances, regardless of their difficulty. This introduces a redundant computational burden, as unambiguous claims consume the same resources as complex borderline cases. Consequently, an adaptive mechanism is needed that intelligently modulates the depth of reasoning, reserving exhaustive checklist verification for cases exhibiting high intrinsic uncertainty.

To bridge this gap, we propose the **Adaptive Cost-efficient (patent) Evaluation (ACE)** framework. The key insight is that the predictive entropy of a lightweight encoder can effectively identify which claims require expensive LLM verification. ACE engineers a fine-tuned PatentBERT-based encoder as an uncertainty-aware gatekeeper whose predictive entropy serves as a runtime routing signal, escalating only high-uncertainty claims to an expert LLM instruction-tuned for legal reasoning. The expert then executes a Chain of Patent Thought (CoPT) protocol, performing structured statutory mapping to produce legally grounded verdicts. By combining efficient screening with selective deep verification, ACE decouples reasoning depth from evaluation cost. To support reproducibility, we publicly release the trained ACE Gatekeeper checkpoint, the QLoRA adapter weights for the Expert LLM, and the datasets.¹

Our contributions are summarized as follows:

- **System-Level Innovation:** ACE as an entropy-driven hybrid framework decouples reasoning depth from evaluation cost, surpassing competitive baselines including a 70B-parameter LLM while reducing operational costs and inference time by 78%.
- **Diagnostic Insight:** We identify limitations of encoder-based models in tracking long-range dependencies (e.g., *Antecedent Basis*), then ACE bridges this gap via selective generative auditing. Analysis on real prosecution data further reveals that cross-claim multi-hop references constitute the dominant residual failure mode.
- **Reasoning Protocol:** We introduce the Chain of Patent Thought (CoPT), a schema-constrained protocol that structures reasoning around 35 U.S.C. standards, reducing inference latency by 42% (11.85s $\rightarrow$ 6.88s) while producing actionable diagnostic outputs for legal practitioners.
- **Domain Datasets:** We introduce ACE-40k, a 40,000-claim dataset with hierarchical labels grounded in MPEP error categories for training and primary evaluation. We additionally present ACE-Real112b, a corpus of 204 genuine USPTO §112(b) rejections that serves as an out-of-domain stress test, on which ACE’s routing mechanism transfers without re-calibration.

2 Related Work

2.1 General Evaluation and Scalability Barriers

Standard metrics, ranging from lexical n-gram overlaps (Papineni et al., 2002; Lin, 2004; Banerjee and Lavie, 2005) to embedding-based measures (Zhang* et al., 2020; Zhao et al., 2019; Selam et al., 2020), exhibit a fundamental semantic-legal gap in patent contexts. Patent validity hinges on rigid structural dependencies; high semantic overlap routinely masks fatal legal flaws such as antecedent basis violations or term-of-art misuse (USPTO, 2024). The LLM-as-a-Judge approach (Zheng et al., 2023; Liu et al., 2023; Fu et al., 2024) and checklist-driven decompositions (Min et al., 2023; Lee et al., 2025) offer stronger diagnostic capacity, yet apply uniform computational depth to every instance, rendering

¹<https://github.com/Yongmin-Yoo/ACE>

exhaustive validation of millions of claims economically unsustainable.

Adaptive computation has been explored through early-exit decoding (Schuster et al., 2022), LLM cascading (Chen et al., 2024), and selective prediction that abstains or defers based on confidence signals (Varshney et al., 2022). However, these methods route on binary or scalar confidence, which is poorly calibrated when the underlying task involves multiple heterogeneous failure modes of differing legal severity. ACE instead trains its Gatekeeper on a K+1 categorical taxonomy over legal error types; while the resulting routing signal correlates strongly with a simpler binary-entropy variant computed from the same distribution (Appendix H), the K+1 formulation additionally yields per-instance error-type diagnostics that a binary confidence score cannot provide, information ACE exploits both for interpretability and as auxiliary supervision for the downstream expert LLM (§4.1).

2.2 Domain-Specific Constraints

Domain-adapted encoders such as PatentBERT (Lee and Hsiang, 2020) and evaluation frameworks including PatentScore (Yoo et al., 2025), PatDEVAL (Yoo and Pan, 2026) and PEDANTIC (Knappich et al., 2025) address patent-specific syntax and legal conventions. However, PatentScore condenses structural flaws into scalar scores without actionable diagnostic rationale, while PEDANTIC restricts its scope to definiteness, leaving broader defects such as circular dependencies unaddressed. Crucially, both process trivial and complex claims with identical depth. ACE integrates an uncertainty-driven gatekeeper with a selective LLM evaluator, reserving exhaustive statutory verification for the minority of claims that genuinely require it.

3 Dataset Construction

Category	Description	Count
Pass	Published patent	20,000
Fail: Antecedent	Lack of antecedent basis	4,000
Fail: Dependency	Invalid cross-reference	4,000
Fail: Logical	Impossible causality or circularity	4,000
Fail: Ambiguity	Indefinite terminology	4,000
Fail: Syntax	Structural format violations	4,000
Total	Balanced Hierarchically Labeled Dataset	40,000

Table 1: Statistics of the constructed ACE dataset. We enforce a strictly balanced distribution of 1:1 between valid and invalid samples.

To facilitate robust automated validation and

fine-grained error diagnosis for patent claims, we constructed a large-scale, balanced patent claim dataset with hierarchical labels, comprising authentic patent claims and synthetically corrupted counterparts. The corrupted samples were generated using a rule-driven synthetic error injection method powered by multiple LLMs, a strategy specifically adopted to prevent bias towards a single model’s generation pattern. Distinct from prior studies predicated on subjective quality metrics, our dataset aligns directly with the standards of the Manual of Patent Examining Procedure (USPTO, 2024), centering on the binary validity of a claim and the specific rationale for rejection. Detailed statistics and representative examples of the constructed samples are provided in Appendix A.

3.1 Data Composition and Error Injection

The dataset is derived from legally granted patents within the USPTO Bulk Data Storage System, ensuring that all positive samples are drawn from legally granted patents. To synthesize negative samples, we employed a *rule-driven synthetic negative sampling* approach. For each valid anchor claim, we systematically injected targeted errors corresponding to five distinct rejection categories defined in patent law.

Crucially, to prevent model bias towards the majority class and ensure rigorous evaluation, we strictly balanced the distribution between valid and invalid samples. The final dataset consists of 40,000 samples, split evenly between valid claims labeled as Pass and invalid claims assigned to specific error categories labeled as Fail. This ratio of 1:1 ensures that the performance of the model reflects genuine discriminatory capability rather than a heuristic reliance on class priors. The error taxonomy is rigorously grounded in specific MPEP sections, targeting five key categories: Antecedent Basis, Dependency, Logical, Ambiguity, and Syntax. Detailed definitions for these violations and the corresponding generation prompts are provided in Appendices B and C.

We adopt synthetic injection over real Office Action rejections for three reasons: it yields class-balanced, claim-level labels across five fine-grained error types that skewed, sparsely-labeled real rejections cannot provide at scale; it produces valid/invalid minimal pairs from the same anchor claim, giving the model a contrastive signal isolating the injected defect itself; and it supports our intended

train-on-synthetic, evaluate-on-real design, whose validity is confirmed empirically in §5.1, where the threshold calibrated on synthetic data transfers to genuine USPTO rejections without re-calibration.

3.2 Strategic Alignment and Utility

Our error taxonomy is strategically aligned with known generative failure modes: Antecedent and Dependency errors serve as proxies for hallucinations (Liu et al., 2024), while Logical and Ambiguity errors target causal inconsistency (Kiciman et al., 2024) and vagueness (Atanasova et al., 2023). To support the ACE framework, samples are annotated with hierarchical labels (binary validity and fine-grained categories), a structure whose reliability was verified through a rigorous annotation audit. Specifically, two PhD-level patent domain experts independently evaluated a random sample of 500 instances, evenly split between the two generators (Qwen3-32B and Gemma-3-27B-IT), on both binary validity and 6-way hierarchical error labels. The audit yielded a raw agreement of 99.2% and a Cohen’s κ of 0.98 with no significant difference in agreement rate between generators, confirming near-perfect inter-annotator consistency and the absence of detectable model-specific artifacts. This dual-granularity is critical for our architecture: the binary labels train the efficient Gatekeeper to detect general inadmissibility, as described later. Simultaneously, the fine-grained categories facilitate uncertainty quantification via entropy calculation and provide fine-grained supervision for the LLM, as detailed subsequently.

4 Methodology

We propose the Adaptive Cost-efficient Evaluation (ACE) framework, designed to reconcile computational scalability with legal reliability. As illustrated in Figure 2, ACE integrates a high-throughput PatentBERT gatekeeper with a reasoning-intensive LLM via uncertainty-driven routing. The framework operates in two phases: (1) *domain-adaptive training* to optimize both the uncertainty-aware encoder and the instruction-tuned evaluator, and (2) *adaptive inference*, where high-entropy claims are dynamically escalated to the expert LLM, effectively decoupling evaluation cost from reasoning depth.

4.1 Domain-Adaptive Training Strategy

The ACE framework jointly optimizes two specialized components, ensuring that each model is tailored to its distinct role in the triage pipeline; the specific training hyperparameters for both components are detailed in Appendix D.

Stage 1: Gatekeeper via Unified Multitask Learning We engineer the Gatekeeper using PatentBERT (Lee and Hsiang, 2020) to process claims with high throughput. Instead of employing decoupled heads for binary and multiclass tasks, which can lead to feature misalignment, we propose a Unified Categorical Projection. Specifically, we project the latent representation h into a single distribution over $K + 1$ classes, encompassing one valid state ($c = 0$) and K specific error types ($c = 1, \dots, K$),

$$p_{\text{dist}} = \text{softmax}(W_e h + b_e) \in \mathbb{R}^{K+1}. \quad (1)$$

To optimize this granular output space, we employ a standard cross-entropy objective:

$$\mathcal{L}_{\text{GK}} = - \sum_{c=0}^K y_c \log p_{\text{dist}}^{(c)}. \quad (2)$$

By training directly on this fine-grained taxonomy, the encoder learns a shared representation where validity is intrinsically modeled as the *absence* of structural defects. This probabilistic granularity is essential for the robust entropy calculation utilized in the subsequent inference stage.

Stage 2: Expert Evaluator via Instruction Tuning To resolve complex statutory ambiguities (e.g., 35 U.S.C. §112), we fine-tune Llama-3-70B-Instruct. We adhere to computational constraints (single NVIDIA H100) via Quantized Low-Rank Adaptation (QLoRA). The backbone is frozen in 4-bit NormalFloat (NF4), and low-rank adapters ($r = 16, \alpha = 32$) are injected into all linear projection layers using *bfloat16* precision. The model is optimized via Supervised Fine-Tuning (SFT) employing a response-only masking strategy, where the training loss is computed exclusively on the generated completion tokens, to prevent catastrophic forgetting of general reasoning capabilities while enhancing domain adherence.

4.2 Adaptive Inference Framework

We deploy the trained models within a dynamic routing framework designed to decouple reasoning depth from evaluation cost.

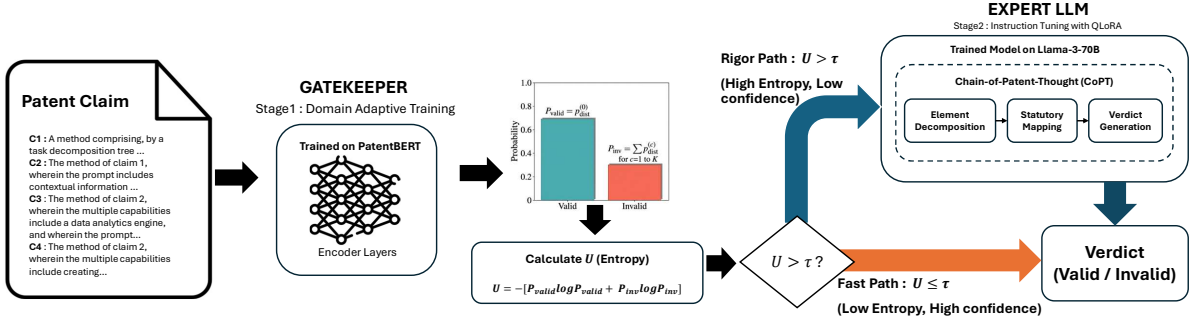


Figure 2: The ACE framework. An uncertainty-aware Gatekeeper routes patent claims based on predictive entropy: confident predictions ($U \leq \tau$) are resolved via the Fast Path, while ambiguous claims ($U > \tau$) are escalated to an instruction-tuned Expert LLM that performs Chain-of-Patent-Thought (CoPT) reasoning for statutory verification.

Stage 1: Uncertainty Quantification & Threshold Calibration To assess instance-level difficulty, we first reconstruct the global validity verdict from the Gatekeeper’s fine-grained output. Given the distribution p_{dist} , we define the probability of a claim being invalid (P_{inv}) by marginalizing over all specific error subtypes ($c = 1 \dots K$):

$$P_{\text{valid}} = p_{\text{dist}}^{(0)}, \quad P_{\text{inv}} = \sum_{c=1}^K p_{\text{dist}}^{(c)}. \quad (3)$$

Using these aggregated probabilities, we calculate the predictive entropy U as a runtime proxy for reliability:

$$U = -[P_{\text{valid}} \log P_{\text{valid}} + P_{\text{inv}} \log P_{\text{inv}}]. \quad (4)$$

While a simpler binary-entropy signal computed directly from $\{P_{\text{valid}}, P_{\text{inv}}\}$ yields nearly identical routing decisions to the full $K+1$ categorical entropy over p_{dist} (Appendix H), we retain the $K+1$ training objective because its value lies primarily in the representation it induces rather than in the routing signal itself: the encoder is forced to internalize a fine-grained error taxonomy, which supplies the specific error category associated with each escalation decision and provides auxiliary supervision for the Stage 2 expert. To rigorously determine the optimal routing threshold τ , we utilize a *Risk-Coverage Trade-off Analysis*. We identify the optimal escalation rate γ^* by maximizing the trade-off between the Gatekeeper’s F1-score $\mathcal{F}(\gamma)$ on retained samples and the escalation ratio γ ,

$$\gamma^* = \arg \max_{\gamma \in [0,1]} (\mathcal{F}(\gamma) - \lambda \cdot \gamma), \quad (5)$$

where λ governs the trade-off preference between accuracy and computational overhead. The

threshold τ is then calibrated to the uncertainty value at the γ^* -th percentile.

Stage 2: Selective Escalation via CoPT The final validity decision $\hat{y}$ is governed by a resource-aware hard-switching policy based on the calibrated threshold τ . For the Fast Path, the binary prediction p_{bin} is derived as $p_{\text{bin}} = \arg\max(P_{\text{valid}}, P_{\text{inv}})$:

$$\hat{y} = \begin{cases} p_{\text{bin}} & \text{if } U \leq \tau \quad (\text{Fast Path}) \\ p_{\text{LLM}} & \text{if } U > \tau \quad (\text{Rigorous Path}) \end{cases} \quad (6)$$

Claims located in the high-entropy region ($U > \tau$) are escalated to the Supervised Fine-tuned Expert LLM to execute the Chain of Patent Thought (CoPT) protocol. We reformulate the validity assessment not as a simple classification task, but as a conditional sequential generation problem. The expert model is mandated to first construct a reasoning chain R , encompassing element decomposition and statutory mapping against 35 U.S.C. standards, before deducing the verdict y . The comprehensive prompt templates and in-context examples utilized for this diagnostic process are detailed in Appendix F. Mathematically, we optimize and select the best reasoning chain and verdict, $\hat{R}$ and $\hat{y}$ by maximizing the joint likelihood:

$$\arg \max_{R, y} \underbrace{P(y | R, x)}_{\text{Verdict}} \cdot \underbrace{P(R | x)}_{\text{Reasoning}}. \quad (7)$$

By explicitly factorizing the probability, this formulation enforces a strict causal dependency where the final verdict y is conditioned on the diagnostic evidence R . In practice, this joint objective is approximated via autoregressive decoding, where the model first generates R token-by-token and then produces y conditioned on the completed reasoning chain. This effectively mitigates ungrounded

hallucinations, ensuring that the decision resolves complex structural ambiguities (e.g., Antecedent Basis) that the Gatekeeper’s encoder architecture fails to capture.

5 Experiments

We evaluate ACE on the ACE-40k dataset (8:1:1 split), benchmarking against general encoders (e.g., Longformer), domain-specific encoders (e.g., PatentBERT), and large language models (Llama-3-70B-Instruct). To validate the stability of our uncertainty estimation, we additionally perform stratified 5-fold cross-validation on the Stage 1 Gatekeeper. All models are evaluated on a single NVIDIA H100 80GB GPU. We report Macro-F1 and estimated operational cost (\$3.00/hr); detailed configurations are in Appendix D.

5.1 Main Performance Analysis

Table 2 presents the evaluation of the ACE framework against baseline methods. The ACE Hybrid model achieves the best F1 among the evaluated methods at 94.95%, surpassing competitive baselines while reducing operational costs to \$1,247 per million claims, a 78% reduction compared to the standalone Expert LLM (\$5,733).

Efficacy and Efficiency of CoPT. The standalone expert model (SFT w/ CoPT) not only achieves high F1 (94.76%) but also demonstrates significantly lower latency (6.88s) compared to the SFT Zero-shot baseline (11.85s). This counter-intuitive efficiency gain stems from our CoPT protocol imposing a strict JSON schema on the output, which curbs uncontrolled verbosity and optimizes both reasoning quality and inference speed.

Breaking the Performance Ceiling. Without task-specific fine-tuning, the domain-specific encoder PatentBERT plateaus at 76.59% F1, confirming that domain-adaptive pre-training alone is insufficient for capturing the long-range dependencies required in patent claim validation. Even after fine-tuning (91.58% F1), the encoder cannot match the Expert LLM (94.76%), indicating a fundamental reasoning gap that requires the generative capabilities integrated by ACE.

Discriminative Robustness (AUC). The Stage 1 Gatekeeper achieves an AUC of 97.16%, contrasting sharply with vanilla PatentBERT (51.16%) and Legal-BERT (46.30%), which perform near random chance. While the lexical baseline (TF-IDF)

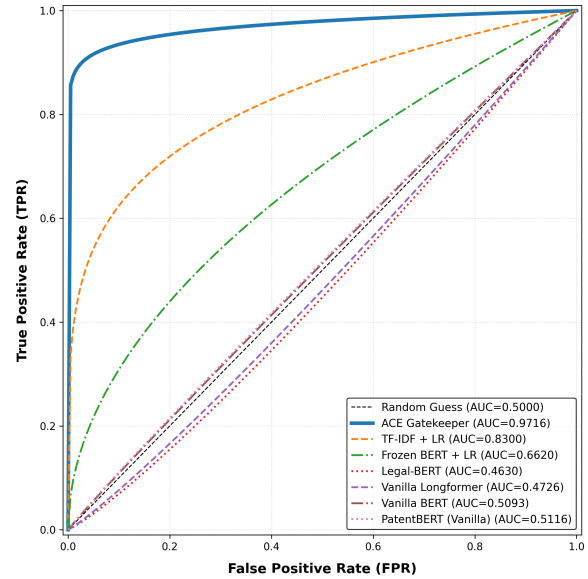


Figure 3: ROC curves for claim validity discrimination. The Gatekeeper (AUC=0.9716) substantially outperforms baselines.

reaches 83.00%, it fails to capture the complex logical dependencies that our Gatekeeper masters. This high AUC validates the entropy-based routing logic: the model’s confidence scores provide a reliable signal for determining which claims require expensive Stage 2 reasoning.

Real World Dataset Construction. From the Harvard USPTO Patent Dataset (Suzgun et al., 2023), we extract patent applications filed in 2016 that received at least one §112(b) indefiniteness rejection in their Office Actions, yielding 204 evaluation instances (one claim per application). Because all samples are authentic rejections (positive-only), we report recall as the primary transfer metric. No thresholds or parameters were re-tuned on this set; the escalation ratio calibrated on ACE-40k is applied without modification. To additionally measure precision and false-positive behavior, which a positive-only set cannot support, we extend ACE-Real112b with 204 genuine, human-authored USPTO claims from the same 2016 HUPD filings that received no §112(b) rejection, forming a class-balanced set of 408 claims. The same routing threshold is applied to this extended set without any re-tuning.

Results and Analysis. As presented in Table 3, the ACE Hybrid achieves 66.2% recall, matching the unrouted Expert LLM (65.2%) while reducing inference time by over 60%. Compared to Patent-

Model / Method	Acc. (%)	F1 (%)	AUC (%)	Latency (s)	Est. Cost (\$)/M
<i>Sanity Check & General Baselines</i>					
Vanilla BERT	49.60	66.25	50.93	0.12	\$100
Vanilla Longformer	75.52	73.11	47.26	0.42	\$350
Frozen BERT + Logistic Regression	69.27	68.12	66.20	0.08	\$67
TF-IDF + Logistic Regression	73.38	72.23	83.00	< 0.01	< \$1
<i>Domain-Specific Experts</i>					
Legal-BERT	69.70	68.86	46.30	0.12	\$100
PatentBERT	77.25	76.59	51.16	0.12	\$100
PatentBERT Fine-tuned	91.65	91.58	97.16	0.12	\$100
<i>Large Language Models (Llama-3-70B)</i>					
Base w/ Zero-shot	70.10	67.75	-	7.31	\$6,092
Base w/ Few-shot	60.00	34.85	-	3.32	\$2,767
SFT w/ Zero-shot	94.60	94.63	-	11.85	\$9,875
SFT w/ CoPT (Expert LLM Only)	94.80	94.76	-	6.88	\$5,733
<i>Proposed Framework</i>					
ACE Gatekeeper (Stage 1 Only)	91.65	91.58	97.16	0.12	\$100
ACE Hybrid (Stage 1 + 2)	95.05	94.95	-	1.50	\$1,247

Table 2: Main Performance Results. Domain-specific baselines use pre-trained representations without task-specific fine-tuning. ACE Hybrid processes 1M claims at \$1,247, a 78% reduction versus the standalone expert LLM (\$5,733). Cost estimated at \$3.00/hr for NVIDIA H100. AUC is reported only for single-model classifier.

Model	Recall (%)	Time (s)
Llama-3-8B (zero-shot)	13.2	103.1
Llama-3-70B (zero-shot)	8.3	355.3
PatentBERT Standalone	56.4	7.2
Llama-3-70B + SFT + CoPT	65.2	6,820.9
ACE Hybrid (Ours)	66.2	2,674.8

Table 3: Results on ACE-Real112b (204 real USPTO §112(b) rejections). All samples are genuine Office Action rejections; recall is reported as the transfer metric. No parameters were re-tuned on this set.

BERT standalone, the hybrid framework provides a +9.8 percentage point improvement, confirming that entropy-based routing effectively channels difficult cases to generative reasoning. Notably, zero-shot LLMs (Llama-3-8B: 13.2%, Llama-3-70B: 8.3%) fail substantially on real prosecution data, reaffirming that domain-specific fine-tuning and structured prompting are indispensable for §112(b) detection. The fact that the escalation threshold calibrated on synthetic data transfers directly to real USPTO rejections without re-calibration confirms the distributional robustness of the Gatekeeper’s uncertainty estimation. A detailed threshold sensitivity analysis across escalation ratios from 10% to 50% is provided in Appendix I, confirming a smooth and monotonic recall-cost trade-off.

Model	Precision (%)	FPR (%)
Always-invalid baseline	50.0	100.0
ACE Hybrid (Ours)	93.75	4.41

Table 4: Precision and false-positive rate on the class-balanced ACE-Real112b extension (204 real invalid + 204 real valid USPTO claims, 408 total), evaluated without re-calibration. The always-invalid baseline row illustrates the trivial upper bound on recall that a degenerate classifier would achieve at the cost of maximal FPR.

Precision and False-Positive Analysis. Recall alone cannot rule out a degenerate always-invalid classifier, which would trivially achieve 100% recall at the cost of 100% FPR and 50% precision. On the class-balanced extension of ACE-Real112b, ACE Hybrid attains a precision of 93.75% and a false-positive rate of 4.41% (Table 4), correctly classifying 195 of the 204 real valid claims; recall on the invalid subset remains unchanged (66.2% vs. 66.2% on the positive-only set), confirming that the reported recall is not an artifact of indiscriminate rejection. Because the added valid claims are authentic USPTO grants rather than LLM-generated text, this result also weighs against the hypothesis that the Gatekeeper’s real-data performance merely reflects generator-specific artifacts inherited from ACE-40k’s synthetic negatives.

To assess whether the small recall advantage of ACE Hybrid over the standalone Expert LLM (66.2% vs. 65.2%) is statistically meaningful, we compute 10,000-resample bootstrap confidence intervals: 66.2% [59.8%, 72.5%] for ACE Hybrid versus 65.2% [58.3%, 71.6%] for the Expert LLM. The intervals overlap almost completely, indicating the two systems are statistically indistinguishable in recall; we therefore characterize ACE as *matching*, rather than surpassing, Expert-LLM recall while requiring 60% less inference time. As this evaluation still draws on a single statutory category (§112(b)) and a moderate sample size (408 claims), we present these real-world results as evidence of promising synthetic-to-real transfer rather than a definitive measure of deployment-ready reliability; broader validation across additional defect types and statutory provisions remains an important direction for future work (see Limitations).

5.2 Fine-grained Error Analysis

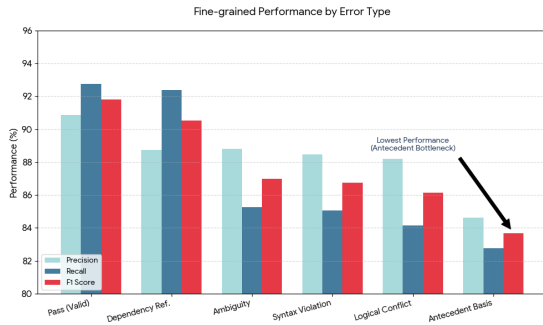


Figure 4: Fine-grained Gatekeeper performance across error types. *Antecedent Basis* records the lowest recall, confirming the long-range tracking bottleneck.

As shown in Figure 4, the Gatekeeper excels at detecting *Dependency Reference* errors (92.37% recall), which manifest as explicit pattern mismatches easily captured by self-attention, yet records the lowest recall on *Antecedent Basis* errors (82.77%), confirming that tracking entity lifecycles (e.g., “a widget” → “the widget”) across long sequences exceeds the semantic consistency limits of encoder-based models. This ~17% miss rate constitutes a significant legal risk that ACE mitigates through selective escalation: the CoPT protocol explicitly identifies such violations via element parsing (e.g., “Step 2: ‘the sensor’ lacks a proper antecedent basis”), as illustrated in Appendix G.

Real-World Error Patterns. Analysis of the 69 missed cases on ACE-Real112b (Table 5) reveals

Failure Pattern	Count	%
Cross-claim dependency / Nested antecedent	38	55.1
Claim-set semantic scope ambiguity	17	24.6
Examiner-specific subjective judgment	14	20.3

Table 5: Breakdown of 69 missed cases on ACE-Real112b.

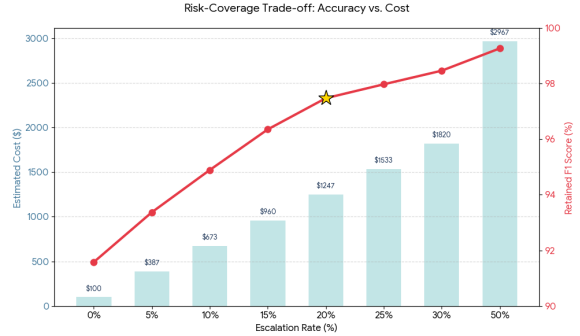


Figure 5: Risk-Coverage Trade-off. Retained F1 and operational cost as a function of escalation rate.

that the antecedent bottleneck persists as the dominant failure source: 38 errors (55.1%) stem from cross-claim dependencies where elements introduced in an independent claim propagate through chains of dependent claims, forming multi-hop references spanning the entire claim set. The remaining 31 errors divide into claim-set semantic scope ambiguity (17 cases, 24.6%), where validity requires access to the full specification beyond claim text, and examiner-specific subjective judgment (14 cases, 20.3%), where rejection rests on interpretive ambiguity rather than explicit structural defects. The former delineates a principled input-scope boundary, while the latter reflects inherent label indeterminacy, together suggesting that the gap between synthetic (94.95% F1) and real-world (66.2% recall) performance stems primarily from a qualitative shift in error complexity, from intra-claim single errors to multi-hop cross-claim dependencies (55.1% of failures), rather than a generalization failure of the routing mechanism. We further validated ACE’s stability through stratified 5-fold cross-validation ($\sigma_{\text{AURC}} < 0.002$) and a simulated 9:1 industrial distribution, which reduces the escalation rate to 8.2% and yields ~90% cost savings, as detailed in Appendix J.

5.3 Ablation Study: Gatekeeper Efficacy and Sensitivity

We conduct a Risk-Coverage analysis to determine how effectively the predictive entropy identifies

error-prone samples and calibrate the optimal uncertainty threshold (τ).

Sensitivity and Optimal Trade-off. As visualized in Figure 5, the efficacy of ACE hinges on the Gatekeeper’s ability to concentrate errors within the high-uncertainty bracket. The Retained F1 curve measures the Gatekeeper’s standalone performance exclusively on the subset of claims not routed to the LLM. The steep ascent of this curve confirms that predictive entropy is a robust proxy for error probability. Filtering out just the top 15% of uncertain samples results in a dramatic performance surge in the retained subset. Notably, the system achieves a sweet spot at an escalation rate of 20% (marked by the star), where the Gatekeeper maintains near-perfect reliability (97.90% F1) on the remaining 80% of the data. This implies that the vast majority of patent claims contain clear structures that do not require expensive reasoning. By offloading only this specific 20% slice of ambiguous cases ($\tau \approx 0.0762$) to the Stage 2 LLM, the ACE framework maximizes resource efficiency while ensuring that the encoder-processed majority maintains legal-grade precision.

5.4 Ablation Study: Component Analysis of the Expert LLM

While the Gatekeeper ensures efficiency, the reliability of ACE is ultimately anchored in the Stage 2 Expert LLM. To isolate the contributions of our training strategies, we evaluate the impact of Supervised Fine-Tuning (SFT) and the Chain of Patent Thought (CoPT) protocol.

Impact of Domain-Specific SFT. As evidenced in Table 2, the Llama-3-70B base model yields 67.75% F1 in the zero-shot setting. Introducing two few-shot demonstrations (one Pass, one Fail) degrades performance further to 34.85%, as patent claims routinely exceed 2,000 tokens and even minimal demonstrations consume a disproportionate share of the context, biasing the model toward the examples rather than the target claim. This confirms that in-context learning is ill-suited for long-document legal tasks. In sharp contrast, SFT boosts F1 to 94.63%, a +26.88% gain confirming that parametric knowledge internalization is essential for patent validation. Model scale is equally critical: a Llama-3-8B expert integrated into ACE Hybrid achieves only 89.3% F1, failing to surpass the Gatekeeper alone (91.58% F1), as the smaller model

lacks sufficient reasoning capacity to correctly resolve the escalated high-entropy samples.

Necessity and Efficiency of the CoPT Protocol.

Although CoPT yields a modest +0.13% F1 increment, its primary contribution is operational rather than metric-driven: in patent prosecution, a single unverified error can incur substantial litigation costs. CoPT transforms the model into a transparent auditor by externalizing reasoning through element parsing and 35 U.S.C. statutory mapping, providing patent attorneys with actionable grounds for rejection. Furthermore, CoPT delivers a counterintuitive efficiency gain: by imposing a strict JSON schema, it reduces inference time from 11.85s (unstructured SFT) to 6.88s, curtailing uncontrolled verbosity while maintaining legal robustness. We emphasize that CoPT is not presented as a mechanism for improving raw classification accuracy, for which its gain is marginal; its role is to convert an opaque verdict into an auditable, statute-grounded rationale that patent practitioners can verify, with the latency reduction arising incidentally from output-length control rather than from improved legal reasoning per se. A representative output is provided in Appendix G.

6 Conclusion

We proposed ACE to resolve the rigidity-resource dilemma in patent claim validation. By integrating an uncertainty-aware gatekeeper with an expert LLM executing a schema-constrained CoPT protocol, ACE decouples reasoning depth from evaluation cost, achieving the best F1 among evaluated methods (94.95%) while reducing costs by 78% and per-claim latency by 42%. The entropy-based routing threshold transfers to real USPTO prosecution data without re-calibration, matching the recall of a standalone expert LLM while reducing inference time by 60%, and substantially outperforming lighter baselines such as PatentBERT and zero-shot LLMs. Beyond patents, treating model uncertainty as a routing signal offers a generalizable blueprint for high-stakes domains requiring adaptive verification depth.

Limitations

This work targets structural validity verification under the definiteness requirement of 35 U.S.C. §112(b), and this scope is bounded along two orthogonal axes. First, along the statutory axis,

the legal adequacy of patent claims spans additional requirements, including novelty (§102), non-obviousness (§103), written description and enablement (§112(a)), and patent-eligible subject matter (§101); each entails a fundamentally different reasoning regime, as §102/§103 demand external comparison against prior art documents, §112(a) requires inference of support relationships between the full specification and the claims, and §101 involves abstract jurisprudential judgment under frameworks such as the Alice/Mayo two-step test, none of which reduce to the claim-internal structural consistency verification addressed herein. Second, along the documentary axis, ACE operates exclusively over claim text, whereas §112(b) determinations in practice often depend on term definitions in the specification, referential consistency with the drawings, and new matter issues arising from amendments during prosecution history; extending toward an integrated verification system encompassing the full application document, as explored for description generation in FlowPlan-G2P (Pan and Yoo, 2026), constitutes a separate research direction.

ACE is trained on ACE-40k, constructed via rule-based synthetic error injection. This approach enables controlled experimentation over rejection types codified in the MPEP; however, real examination environments include borderline cases in which examiner interpretive judgment is involved in §112(b) rejections. This reflects the fact that our error taxonomy is designed to target explicit structural defects, and that the domain of interpretive indeterminacy, where inter-examiner disagreement arises, lies outside its intended scope. The residual error analysis on ACE-Real112b empirically corroborates this boundary: 20.3% of undetected cases are attributable to subjective interpretation rather than explicit structural flaws. We also note that ACE-Real112b, even extended to 408 claims with matched valid samples (§5.1), is drawn entirely from a single statutory category and a single filing year; while it demonstrates promising synthetic-to-real transfer, it does not by itself establish generalization across the broader diversity of naturally occurring claim defects or other statutory provisions, and broader real-world validation remains an important direction for future work.

Additionally, the current routing mechanism relies on an empirically calibrated threshold rather than a learned decision policy. Replacing this

with a cost-aware routing function that jointly optimizes accuracy and computational budget remains future work. A promising direction to address the Antecedent Basis bottleneck identified in Section 5 is to replace the Gatekeeper’s flat encoder with a dependency-graph-aware architecture such as PHAGE (Yoo et al., 2026d), which explicitly models typed claim-to-claim relations; integrating such structural priors into the entropy estimation step could reduce escalation on cross-claim reference errors without sacrificing throughput. While the entropy-driven routing architecture is domain-agnostic in principle, the CoPT checklist is tailored to U.S. patent law; adapting to other jurisdictions (e.g., EPO Article 84) requires remapping the statutory checklist, which remains unexplored.

Finally, undetected errors may propagate beyond litigation risk into downstream economic decisions such as patent valuation (Yoo et al., 2026a), suggesting that future work should quantify the real-world cost of residual failure modes across the full lifecycle of patent prosecution and commercialization.

Ethical Considerations

During the preparation of this work, the author(s) utilized generative AI to refine linguistic clarity and support the creation of certain diagrams. The author(s) carefully reviewed all outputs and maintain full responsibility for the intellectual content and originality of the final paper.

References

- Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. 2023. [Faithfulness tests for natural language explanations](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 283–294, Toronto, Canada. Association for Computational Linguistics.
- Satanjeev Banerjee and Alon Lavie. 2005. [METEOR: An automatic metric for MT evaluation with improved correlation with human judgments](#). In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.
- Lingjiao Chen, Matei Zaharia, and James Zou. 2024. [FrugalGPT: How to use large language models while](#)

- reducing cost and improving performance. *Transactions on Machine Learning Research*. Featured Certification.
- Robert C. Faber and Thomas J. Kowalski. 2023. *Faber & Kowalski on Mechanics of Patent Claim Drafting*, 8th edition. Practising Law Institute.
- Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. [GPTScore: Evaluate as you desire](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576, Mexico City, Mexico. Association for Computational Linguistics.
- Emre Kiciman, Robert Ness, Amit Sharma, and Chenhao Tan. 2024. [Causal reasoning and large language models: Opening a new frontier for causality](#). *Transactions on Machine Learning Research*. Featured Certification.
- Valentin Knappich, Annemarie Friedrich, Anna Häty, and Simon Razniewski. 2025. [Pedantic: A dataset for the automatic examination of definiteness in patent claims](#).
- Jieh-Sheng Lee and Jieh Hsiang. 2020. [Patent classification by fine-tuning bert language model](#). *World Patent Information*, 61:101965.
- Yukyung Lee, JoongHoon Kim, Jaehee Kim, Hyowon Cho, Jaewook Kang, Pilsung Kang, and Najoung Kim. 2025. [CheckEval: A reliable LLM-as-a-judge framework for evaluating text generation using checklists](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15771–15798, Suzhou, China. Association for Computational Linguistics.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. [Lost in the middle: How language models use long contexts](#). *Transactions of the Association for Computational Linguistics*, 12:157–173.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. [G-eval: NLG evaluation using gpt-4 with better human alignment](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.
- Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. [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*, pages 12076–12100, Singapore. Association for Computational Linguistics.
- Kris W Pan and Yongmin Yoo. 2026. [Flowplan-g2p: A structured generation framework for transforming scientific papers into patent descriptions](#).
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. [Beyond accuracy: Behavioral testing of NLP models with CheckList](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.
- Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Q. Tran, Yi Tay, and Donald Metzler. 2022. [Confident adaptive language modeling](#). In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS ’22*, Red Hook, NY, USA. Curran Associates Inc.
- Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. [BLEURT: Learning robust metrics for text generation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.
- Mirac Suzgun, Luke Melas-Kyriazi, Suproteem K. Sarkar, Scott Duke Kominers, and Stuart M. Shieber. 2023. [The harvard uspto patent dataset: a large-scale, well-structured, and multi-purpose corpus of patent applications](#). In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS ’23*, Red Hook, NY, USA. Curran Associates Inc.
- USPTO. 2024. [Manual of Patent Examining Procedure \(MPEP\), Section 2173: Claims Must Particularly Point Out and Distinctly Claim the Invention](#). 9th Edition, Revision 01.2024.
- Neeraj Varshney, Swaroop Mishra, and Chitta Baral. 2022. [Investigating selective prediction approaches across several tasks in IID, OOD, and adversarial settings](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1995–2002, Dublin, Ireland. Association for Computational Linguistics.
- Yongmin Yoo, Seungwoo Kim, and Jingjiang Liu. 2026a. [Era: Aligning semantic models with revealed economic preference for real-time and explainable patent valuation](#). *Information Processing & Management*, 63(7, Part B):104898.

Yongmin Yoo and Kris W Pan. 2026. [Pat-deval: Chain-of-legal-thought evaluation for patent description](#).

Yongmin Yoo, Zhangkai Wu, and Longbing Cao. 2026b. [Pointer-augmented autoregressive generation of patent claims with joint topology and content decoding](#).

Yongmin Yoo, Qionghai Xu, and Longbing Cao. 2025. [PatentScore: Multi-dimensional evaluation of LLM-generated patent claims](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30727–30746, Suzhou, China. Association for Computational Linguistics.

Yongmin Yoo, Qionghai Xu, and Longbing Cao. 2026c. [PatentMind: A multi-aspect reasoning graph for patent similarity evaluation](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 14947–14964, San Diego, California, United States. Association for Computational Linguistics.

Yongmin Yoo, Qionghai Xu, Zhangkai Wu, and Longbing Cao. 2026d. [Heterogeneous dependency graph-guided attention for patent representation learning](#).

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with bert](#). In *International Conference on Learning Representations*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. [MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS ’23*, Red Hook, NY, USA. Curran Associates Inc.

A Representative Data Samples

Table 6 presents detailed examples of the constructed dataset. Each invalid sample contains a targeted error injected based on the MPEP guidelines, such as Antecedent Basis violations or Logical conflicts. Detailed dataset statistics are provided in Table 1 of the main text.

B Error Taxonomy Definitions

The error taxonomy employed in the ACE dataset is rigorously grounded in specific MPEP sections.

Label	Claim Text	Error Type
Pass	A semiconductor device comprising a substrate, a gate structure formed over the substrate, and a spacer adjacent to gate structure.	None
Fail	A semiconductor device comprising the gate structure formed over the substrate, wherein the substrate is configured to receive the gate structure of claim 1.	Antecedent
Fail	The system of claim 1, wherein the processor of claim 3 further executes the module of claim 7.	Dependency
Fail	A battery including a charging unit that generates electrical energy from the discharged battery cell.	Logical

Table 6: Representative samples from the ACE dataset. Invalid claims contain single targeted errors injected based on MPEP guidelines.

For each category, we define the violation and provide a minimal example below:

1. **Antecedent Basis (MPEP §2173.05(e))**: Referencing an element using a definite article without prior introduction.
Ex: "The sensor..." appearing without a prior "a sensor".
2. **Dependency (MPEP §608.01(n))**: Invalid claim references, including forward referencing or circular dependencies.
Ex: Claim 2 citing "The device of claim 5" (where claim 5 is subsequent).
3. **Logical (35 U.S.C. §112(b))**: Internal contradictions or physically impossible functional relationships.
Ex: "A transparent layer made of opaque metal".
4. **Ambiguity (MPEP §2173.05(b))**: Use of subjective or undefined degree terms that obscure the scope of the claim.
Ex: "Heating to a substantially high temperature".
5. **Syntax (35 U.S.C. 112)**: Violations of formal grammatical structure or required patent formatting.
Ex: Missing the transitional phrase "comprising" or incorrect punctuation.

C Synthetic Data Generation Details

To construct the ACE dataset, we employed a rule-driven synthetic error injection method using Large Language Models (LLMs), specifically

Qwen/Qwen3-32B and Gemma-3-27B-IT. To ensure the injected errors accurately reflect real-world patent prosecution challenges, we utilized Few-Shot Prompting, providing the models with specific examples of MPEP violations.

C.1 Error Injection Prompts with In-Context Examples

The following templates demonstrate the instructions provided to the generator. Each prompt includes a definition of the error based on the Manual of Patent Examining Procedure (MPEP) and a minimal example to guide the generation process.

1. Antecedent Basis Error Prompt (MPEP §2173.05(e))

ROLE: Senior Patent Prosecutor
TASK: Rewrite the provided claim set to introduce 5-15 **Antecedent Basis Errors**.
DEFINITION: An antecedent error occurs when a claim element is introduced with a definite article (e.g., "the lever", "said lever") without having been previously introduced with an indefinite article (e.g., "a lever").
EXAMPLE:
 [Original] A device comprising a sensor; wherein the sensor captures data.
 [Error] A device comprising a processor; wherein the sensor captures data. ("the sensor" lacks a precedent "a sensor")
REQUIREMENTS:
 1. Preserve claim structure and numbering.
 2. Replace valid references with unmatched elements (e.g., "the handle" when no handle exists).
 3. Output ONLY the revised claims.
TARGET CLAIMS: {claim_text}

2. Dependency Error Prompt (MPEP §608.01(n))

ROLE: Senior Patent Prosecutor
TASK: Rewrite the claim set to introduce 5-15 **Dependency Errors**.
DEFINITION: A dependency error involves incorrect cross-referencing, such as referencing a non-existent claim or a future claim (forward referencing).
EXAMPLE:
 [Original] The method of claim 1, further comprising...
 [Error] The method of claim 8, further comprising... (If claim 8 does not exist or is subsequent).
REQUIREMENTS:
 1. Modify the "claim X" references to invalid numbers.
 2. Create circular dependencies (e.g., Claim 2 cites Claim 3, Claim 3 cites Claim 2).
 3. Output ONLY the revised claims.
TARGET CLAIMS: {claim_text}

3. Logical Error Prompt (35 U.S.C. §112(b))

ROLE: Senior Patent Prosecutor
TASK: Rewrite the claim set to introduce 5-15 **Logical Errors**.
DEFINITION: Logical errors include internal contradictions, physical impossibilities, or inconsistent element properties.
EXAMPLE:
 [Original] A transparent glass layer acting as a window.
 [Error] A transparent glass layer made of opaque steel. (Contradiction: transparent vs. opaque steel).
REQUIREMENTS:
 1. Insert conflicting adjectives or impossible functional relationships.
 2. Do not alter the grammatical structure, only the semantic logic.
 3. Output ONLY the revised claims.
TARGET CLAIMS: {claim_text}

4. Ambiguity Error Prompt (MPEP §2173.05(b))

ROLE: Senior Patent Prosecutor
TASK: Rewrite the claim set to introduce 5-15 **Ambiguity (Indefiniteness) Errors**.
DEFINITION: Claims must be definite. Errors involve subjective terms of degree without metric definitions.
EXAMPLE:
 [Original] Heating the water to 100 degrees Celsius.
 [Error] Heating the water to a substantially high temperature. ("Substantially high" is indefinite).
REQUIREMENTS:
 1. Replace precise numerical values with vague terms (e.g., "about", "approximately", "strong", "large").
 2. Output ONLY the revised claims.
TARGET CLAIMS: {claim_text}

5. Syntax Error Prompt (35 U.S.C. 112)

ROLE: Senior Patent Prosecutor
TASK: Rewrite the claim set to introduce 5-15 **Syntax Errors**.
DEFINITION: Violations of formal claim formatting, punctuation, or missing transitional phrases.
EXAMPLE:
 [Original] A system comprising: a processor; and a memory.
 [Error] A system includes a processor a memory (Missing "comprising", missing semicolons/commas).
REQUIREMENTS:
 1. Remove or corrupt transitional phrases (comprising, consisting of).
 2. Delete necessary punctuation (semicolons, periods).
 3. Output ONLY the revised claims.
TARGET CLAIMS: {claim_text}

D Training and Experimental Details

All experiments were implemented using PyTorch on a single NVIDIA H100 80GB GPU environ-

ment. This section provides the complete hyperparameter configurations for both components of the ACE framework.

D.1 Gatekeeper Model (Stage 1)

Table 7 details the hyperparameters for the Gatekeeper. The objective was to ensure high-throughput screening while maintaining robust error detection capabilities.

Table 7: Hyperparameters for Gatekeeper Training

Parameter	Value	Rationale
Base Model	PatentBERT-base	Domain-Adapted Efficiency
Max Sequence Length	512	Standard BERT Limit
Batch Size	32	Stable Gradient Estimation
Learning Rate	2×10^{-5}	Optimal for Fine-tuning
Epochs	5	Sufficient Convergence
Weight Decay	0.01	L2 Regularization

D.2 Expert LLM Evaluator (Stage 2)

For the LLM Evaluator, we employed Parameter-Efficient Fine-Tuning (PEFT) using QLoRA to adapt the model to the legal domain while minimizing memory overhead. Table 8 summarizes the configuration.

Table 8: Hyperparameters for Llama-3 Evaluator Training (QLoRA)

Parameter	Value	Rationale
Base Model	Meta-Llama-3-70B-Instruct	Instruction-Following Capability
Quantization	4-bit (NF4)	Memory Efficiency (QLoRA)
LoRA Rank (r)	16	Parameter Efficiency
LoRA Alpha (α)	32	Scaling Factor ($2 \times r$)
LoRA Dropout	0.05	Regularization
Learning Rate	2×10^{-4}	Optimal for QLoRA Fine-tuning
Optimizer	Paged AdamW (32-bit)	Memory Optimization
Precision	Bfloat16 (BF16)	H100 Mixed Precision Stability
Batch Size	1 (per device)	Restricted by VRAM (Context 4096)
Gradient Accumulation	24	Effective Batch Size = 24

D.3 Cost Estimation Protocol

To provide a realistic assessment of industrial scalability, we define the *Estimated Cost* (\$) metric. This serves as a standardized proxy for the operational expenditure (OPEX) of deploying the model at the scale of a national patent office. The cost is calculated as:

$$\text{Cost} = \frac{\text{Total Inference Time (hrs)}}{1M \text{ Claims}} \times \text{Hourly Rate} \quad (8)$$

We assume a standard cloud compute rate of \$3.00 per hour for an on-demand NVIDIA H100 instance. Latency measurements include tokenization, model inference, and detokenization overheads, averaged over the entire test set.

E Fine-grained Performance Analysis

Table 9 presents the detailed performance metrics of the Gatekeeper across five distinct error categories. The results reveal a clear performance dichotomy: while the model demonstrates robust detection in semantic tasks (Ambiguity, Logical), it exhibits significant limitations in structural verification tasks (Antecedent, Syntax). This structural deficiency empirically justifies the necessity of the Stage 2 LLM Evaluator for handling complex long-range dependencies.

Table 9: Fine-grained Performance of the Gatekeeper (Test Set). The model struggles with structural errors (Antecedent, Syntax), highlighting the specific areas where the LLM Evaluator is most needed.

Error Type	Recall	F1-Score	Nature
Semantic Reasoning Tasks			
Ambiguity	90.75%	0.8985	Semantic Interpretation
Logical	86.45%	0.8645	Inferential Reasoning
Structural Verification Tasks			
Dependency	92.37%	0.9130	Claim Hierarchy
Antecedent	82.77%	0.8141	Long-Range Reference
Syntax	78.35%	0.8340	Grammatical Structure

F Prompt Specifications

We provide the exact prompt templates used for our experiments. To ensure reproducibility, all prompts were formatted using the Llama-3 specific chat template structure (e.g., `<|start_header_id|>system...`).

F.1 Standard SFT Prompt (Baseline)

For the **SFT Only** setting, we aligned the inference prompt exactly with the training data format to maximize the effect of the fine-tuned adapter. The model is instructed to function as a "Senior Patent Attorney" and output a concise verdict without reasoning, mirroring its training objective. The exact prompt template is provided in Table 10.

F.2 Chain of Patent Thought (CoPT) Protocol

To enable the SFT model to perform deep reasoning despite being trained on short answers, we utilized a specific **System Prompt Override** strategy. The detailed prompt structure is presented in Table 11.

F.2.1 System Instruction Design

The system prompt maintains the training trigger phrase ("You are a Senior Patent Attorney") to activate the fine-tuned weights. However, we explicitly

Role	Content
System	You are a Senior Patent Attorney. Your task is to evaluate the validity of a patent claim based on specific error types. If ANY error exists, output {"verdict": "Fail"}. If NO error, output {"verdict": "Pass"}.
User	Claim: {claim_text}
Assistant	(Model Generation Target)

Table 10: The Standard SFT prompt template. This strictly follows the format used during the supervised fine-tuning stage to prevent distribution shift during inference.

injected a [Special Instruction] tag to override the default behavior, mandating the model to follow a step-by-step reasoning chain before concluding.

F.2.2 In-Context Demonstrations (2-Shot)

We employed a 2-shot In-Context Learning strategy to demonstrate the expected JSON structure.

- **Shot 1 (Pass):** Demonstrates valid element parsing and statutory compliance.
- **Shot 2 (Fail):** Demonstrates how to identify an Antecedent Basis error.

These examples specifically teach the model to parse elements (Step 1) and check for statutory compliance (Step 2) before rendering a verdict.

Use of AI Assistants. AI assistants were used only for limited writing and language-editing support, as well as preparation of submission materials. All technical content, methodology, implementation, experimental results, and final verification were performed and checked by the authors.

G CoPT Output Example

Table 12 presents a representative output of the Chain of Patent Thought (CoPT) protocol on a real USPTO §112(b) rejection case (Application No. 14902826). The structured three-step reasoning enables transparent auditability by explicitly identifying the defective claim elements and citing the applicable legal basis.

H Additional Baseline Experiments

To further validate the design choices of the ACE framework, we report three additional experiments:

an ablation with a smaller expert model, an evaluation of a proprietary LLM, and an ablation isolating the contribution of the K+1 categorical formulation from the routing signal itself.

Smaller Expert Model (8B Ablation). We fine-tuned Llama-3-8B-Instruct under the same SFT protocol used for the 70B expert and evaluated it on ACE-40k. As shown in Table 13, the 8B model achieves 83.5% F1 as a standalone system (\$667/M). When integrated as the Stage 2 expert within the ACE Hybrid framework, it yields 89.3% F1 (\$233/M). Critically, this falls below the Stage 1 Gatekeeper alone (91.58% F1), indicating that the 8B model lacks sufficient reasoning capacity to correctly resolve the high-entropy samples routed to it. Instead of complementing the Gatekeeper, the smaller expert overrides correct predictions with incorrect verdicts, empirically confirming that the escalated subset contains genuinely complex legal ambiguities requiring 70B-class reasoning depth.

Proprietary LLM Baseline (GPT-4o). To provide a proprietary reference point, we evaluated GPT-4o in a zero-shot setting on ACE-Real112b (204 real USPTO §112(b) rejections). GPT-4o achieves only 18.0% recall, substantially below ACE Hybrid (66.2%) and even the standalone Gatekeeper (56.4%). This result demonstrates that neither model scale nor proprietary pretraining is sufficient for reliable §112(b) detection without domain-specific fine-tuning and structured prompting.

Binary vs. K+1 Entropy Routing. Sections 2.1 and 4.2 note that the K+1 categorical formulation’s contribution to *routing* is limited, since a simpler binary-entropy signal achieves nearly identical behavior. We isolate this effect with two controlled comparisons under identical data splits and training budgets. First, we train a variant of the Gatekeeper with a binary (valid/invalid) classification head in place of the K+1 categorical head, holding all other hyperparameters fixed. This binary-only model achieves an AUC of 96.87% on invalid detection, closely matching the K+1 model’s 97.16% (Table 14), indicating that the K+1 taxonomy is not required to achieve strong discriminative performance at the valid/invalid decision boundary. Second, we compare routing decisions produced by the two-term binary entropy used in the main text against a routing signal derived from the full K+1 entropy computed directly over $p_{\text{dist}} \in \mathbb{R}^{K+1}$. The two entropy values are strongly correlated (Spear-

Role	Content
System	<p>You are a Senior Patent Attorney.</p> <p>Your task is to evaluate the validity of a patent claim based on specific error types.</p> <p>[Special Instruction]: To ensure strict adherence to 35 U.S.C. § 112, you must explicitly document your reasoning process using the Chain of Patent Thought (CoPT) protocol before concluding.</p> <ol style="list-style-type: none"> 1. [Element Parsing]: Identify preamble, transition, and body elements. 2. [Statutory Compliance]: Check for Antecedent Basis errors, Indefinite terms, and Logical Contradictions. 3. [Verdict]: Determine if the claim passes or fails based strictly on the above. <p>Output Requirement:</p> <p>You must return a JSON object containing a “reasoning” field (for the step-by-step analysis) and a “verdict” field (“Pass” or “Fail”).</p> <p>Format: {"reasoning": "Step 1... Step 2...", "verdict": "Pass/Fail"}</p>
User (Shot 1)	Claim: “A semiconductor device comprising a substrate, a gate structure formed over the substrate, and a spacer adjacent to the gate structure.”
Assistant (Shot 1)	{"reasoning": "Step 1 [Element Parsing]: The claim includes a substrate, a gate structure, and a spacer. Step 2 [Statutory Compliance]: 'the substrate' and 'the gate structure' have clear antecedent bases from the preamble. No indefinite terms or logical conflicts found. Step 3 [Verdict]: The claim meets all 35 U.S.C. § 112 requirements.", "verdict": "Pass"}
User (Shot 2)	Claim: “A device comprising a processor; wherein the sensor captures data based on input.”
Assistant (Shot 2)	{"reasoning": "Step 1 [Element Parsing]: The claim introduces a processor but explicitly recites 'the sensor' in the body. Step 2 [Statutory Compliance]: The term 'the sensor' lacks a proper antecedent basis (i.e., 'a sensor' was never introduced). This is a violation of 35 U.S.C. § 112(b). Step 3 [Verdict]: The claim is invalid due to lack of antecedent basis.", "verdict": "Fail"}
User (Target)	Claim: “{claim_text}”

Table 11: The full prompt template used for the Stage 2 Specialist LLM (SFT + CoPT). The [Special Instruction] tag and 2-shot examples are designed to override the model’s training bias towards short responses, enforcing a detailed legal analysis path.

man $\rho = 0.977$), and the resulting escalation decisions are nearly identical: at the 20% escalation operating point used throughout the paper, the two variants yield retained Macro-F1 of 97.85% and 97.90%, respectively (Table 14). Taken together, these results indicate that the K+1 formulation and the binary-entropy signal are largely interchangeable for routing itself; we retain K+1 training in ACE because it is a strict superset of the binary formulation in terms of information, preserving routing performance while additionally producing the fine-grained error-category labels used for diagnostic reporting (§5.3) and as auxiliary supervision for the Stage 2 expert (§4.1).

I Threshold Sensitivity Analysis

To assess the robustness of the uncertainty-based routing under varying operational budgets, we evaluate ACE on ACE-Real112b across multiple esca-

lation ratios. The threshold calibrated on ACE-40k is applied without re-tuning; only the percentile cutoff is varied.

As shown in Table 15, increasing the escalation ratio yields consistent recall improvements with proportional cost increases. The absence of abrupt performance transitions around the default operating point (20%) confirms that the entropy-based routing provides a stable and predictable control surface for practitioners to adjust based on their cost-recall requirements.

J Robustness Analysis

To ensure the reliability of ACE in practical applications, we evaluated its robustness from two complementary perspectives: internal stability via cross-validation and ecological validity under industrial-scale class imbalance. These two analyses target distinct sources of concern. The first addresses

Component	Content
Input Claim (Excerpt)	1. (canceled) 2. The workpiece made according to the method of claim 8, further comprising an intermediate layer containing TiAlN provided between the TixSi1-xN layer and the substrate body of the workpiece. (...) 5. The workpiece made according to the method of claim 8, further comprising an intermediate layer containing CryAl1-yN provided between the TixSi1-xN layer and the substrate body of the workpiece (...) (...) 8. A method for coating a workpiece having a coating comprising at least one TixSi1-xN layer, wherein $x \leq 0.85$ (...), the method comprising: using a sputtering process to produce the TixSi1-xN layer (...)
CoPT Output	{"reasoning": "Step 1 [Element Parsing]: Parsed claim elements across claims 2, 5, and 8. Step 2 [Statutory Compliance]: Antecedent Basis error identified: Claim 2 refers to claim 8, but claim 8 is not yet defined at that point; Claim 5 refers to claim 8 under the same condition. No other Definiteness errors found. Step 3 [Verdict]: Fail due to Antecedent Basis errors.", "verdict": "Fail"}
Ground Truth	Fail (§112(b) rejection in USPTO Office Action)

Table 12: Example of CoPT protocol output on a real USPTO §112(b) case. The model identifies forward-reference dependencies (claims 2 and 5 referencing the later-defined claim 8) through structured step-by-step reasoning, providing actionable grounds for the rejection verdict.

Configuration	F1 (%)	Cost (\$/M)
Llama-3-8B SFT (Standalone)	83.5	\$667
ACE Hybrid w/ 8B Expert	89.3	\$233
ACE Gatekeeper (Stage 1 Only)	91.58	\$100
ACE Hybrid w/ 70B Expert	94.95	\$1,247

Table 13: Effect of expert model scale on ACE Hybrid performance (ACE-40k test set). The 8B expert degrades below the Gatekeeper alone, confirming that 70B-class reasoning is necessary for the escalated high-uncertainty subset.

Configuration	AUC	Retained F1 @ 20%
Binary-only Gatekeeper	96.87%	–
K+1 Gatekeeper (main)	97.16%	–
Binary-entropy routing	–	97.85%
K+1-entropy routing	–	97.90%
Spearman ρ (binary vs. K+1 entropy) = 0.977		

Table 14: Binary vs. K+1 formulations for discrimination (top) and routing (bottom). The K+1 formulation does not meaningfully change routing outcomes; its value lies in the error-type diagnostics it additionally provides.

whether our reported thresholds and performance figures are an artifact of a single, fortuitously favorable train-test split, while the second addresses whether the balanced experimental setting used throughout the main text generalizes to the far more skewed distributions encountered in real patent office pipelines.

Internal Stability (Cross-Validation). To confirm that our uncertainty-based routing is not an artifact of a specific data split, we evaluated the Gatekeeper using stratified 5-fold cross-validation, partitioning the ACE-40k dataset such that the proportion of each error category (Antecedent, Dependency, Logical, Ambiguity, and Syntax) is preserved within every fold. For each fold, we independently retrain the Gatekeeper, recompute the predictive entropy over the held-out validation samples, and re-derive the Risk-Coverage curve from

which the optimal escalation rate is calibrated. This procedure allows us to treat γ^* and the associated AURC not as single point estimates but as distributions over five independent realizations of the train-test split, providing a more rigorous test of stability than a single held-out evaluation could offer. The results show high stability in the uncertainty estimation, with the Area Under the Risk-Coverage Curve (AURC) exhibiting negligible standard deviation across folds ($\sigma < 0.002$), indicating that the encoder consistently orders claims by difficulty regardless of which subset of the data it was trained on. The optimal escalation rate (γ^*) also converged consistently around 20% ($\pm 1.5\%$) across all folds, with no fold producing an outlier threshold that would suggest sensitivity to a particular partition of the data. This confirms that the sweet spot identified in Section 5.3 is an intrinsic property of the

Escal Ratio	Stage-2 Samples	Recall	Cost (\$/M)
10%	20	60.8	\$673
20%	40	66.2	\$1,247
30%	61	72.1	\$1,820
40%	81	78.9	\$2,393
50%	101	86.3	\$2,967

Table 15: Threshold sensitivity on ACE-Real112b. The recall-cost trade-off is smooth and monotonic across escalation ratios.

Gatekeeper’s discriminative capability, that is, a structural characteristic of how predictive entropy separates easy from hard claims in this domain, rather than a fluctuation dependent on data distribution or a byproduct of a particular random seed.

Ecological Validity (Industrial Scenarios). The primary experiments in Section 5.1 used a balanced dataset (1:1 Pass/Fail ratio), a design choice made to prevent the Gatekeeper from exploiting class priors and to ensure that reported performance reflects genuine discriminative capability rather than majority-class bias. Real-world patent prosecution, however, is characterized by a highly skewed distribution, since the overwhelming majority of claims drafted by experienced patent attorneys are structurally sound, and only a minority trigger substantive §112(b) objections. A framework validated exclusively under an artificially balanced regime therefore leaves open the question of whether its cost and latency advantages persist once deployed on the class distribution actually encountered in practice. To assess this ecological validity, we simulated a skewed test set with a 9:1 ratio (Pass:Fail), constructed by resampling the existing ACE-40k test partition to reflect this proportion while keeping the underlying claims and labels unchanged, thereby isolating the effect of class prevalence from any confound introduced by new data. This setting reflects a scenario where most incoming claims are structurally well-formed and only a small fraction warrant escalation. Under this setting, the Gatekeeper routes the vast majority of valid claims through the Fast Path, since well-formed claims tend to produce low-entropy, high-confidence predictions, dropping the LLM escalation rate from 20.3% (balanced) to 8.2% (skewed). This reduction arises directly from the composition of the escalated subset: because the number of genuinely ambiguous or defective claims does not change with the resampling, a lower proportion of Fail-

labeled claims in the overall pool translates into fewer claims crossing the uncertainty threshold τ , even though τ itself is held fixed. The estimated operational cost thus falls to $\sim$ \$600 per million claims, a $\sim$ 90% reduction compared to the standalone Expert LLM, confirming that ACE is a cost-optimized solution suited for high-volume industrial pipelines. This result is particularly relevant for deployment at the scale of a national patent office, where the cost savings from selective escalation compound favorably as the underlying claim population becomes increasingly dominated by routine, well-formed submissions, meaning that the efficiency gains reported under the balanced setting in the main text should be interpreted as a conservative lower bound on the savings achievable in realistic operating conditions.