PATRE: A FULL-STAGE OFFICE ACTION AND REBUTTAL GENERATION BENCHMARK FOR PATENT EXAMINATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Patent examination is a complex, multi-stage process requiring both technical expertise and legal reasoning, increasingly challenged by rising application volumes. Prior benchmarks predominantly view patent examination as discriminative classification or static extraction, failing to capture its inherently interactive and iterative nature, similar to the peer review and rebuttal process in academic publishing. In this paper, we introduce **PatRe**, the first benchmark that models the full patent examination lifecycle, including Office Action generation and applicant rebuttal. PatRe comprises 480 real-world cases and supports both oracle and retrieval-simulated evaluation settings. Our benchmark reframes patent examination as a dynamic, multi-turn process of justification and response. Extensive experiments across various LLMs reveal critical insights into model performance, including differences between proprietary and open-source models, as well as task asymmetries between examiner analysis and applicant-side rebuttal. These findings highlight both the current limitations of LLMs in modeling complex, real-world legal reasoning and technical novelty judgment in patent examination.

1 INTRODUCTION

Patent examination ensures that applications meet statutory requirements such as novelty and non-obviousness, while growing application volumes and technical and legal complexity place increasing pressure on examiners. For example, in 2025, the United States Patent and Trademark Office (USPTO) received 475,223 patent applications, with 837,928 unexamined applications and a first-action pendency of 20.5 months. Examiners must assess applications against prior art under statutory requirements outlined in the MPEP (USPTO, 2020).

Researchers have made significant efforts to leverage AI in assisting the patent examination process. HUPD (Suzgun et al., 2023) formulates patent acceptance prediction as binary classification, while PANORAMA (Lim et al., 2025) further evaluates rejection reasons and prior-art conflicts. However, these tasks remain discriminative, providing limited analysis of the reasoning behind examination decisions.

In the patent examination practice, an Office Action (OA) is not a one-time event. Applicants can submit a rebuttal to the examiner's OA in hopes of obtaining a grant until a final decision is reached, similar to the discussion and rebuttal process in peer review of academic papers (Zhang et al., 2025; Li et al., 2026). However, prior work focuses only on reviewing the initial version of a patent application, overlooking the multi-turn interaction between the examiner and applicant and the evolution of subsequent patent versions. Additionally, all these works rely on acceptance or statute accuracy as the metric, lacking a fine-grained analysis of the correctness of the examination suggestions.

In this work, we focus on the entire patent examination lifecycle and introduce **PatRe**, the first full-stage benchmark of **Pat**ent Office Actions and **Re**buttals generation, as illustrated in Figure 1. It primarily includes two types of tasks: **(I)** Office Action (OA) Generation, which requires the model to produce formal examination reports by analyzing patent claims against potential prior art. We further distinguish three OA settings: *Directly Prompting*, which provides no references; *Reference-Oracle Generation*, which provides oracle references; and *Retrieval-Simulated Generation*, which augments oracle references with retrieved distractors to evaluate evidence selection under retrieval noise. **(II)**

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

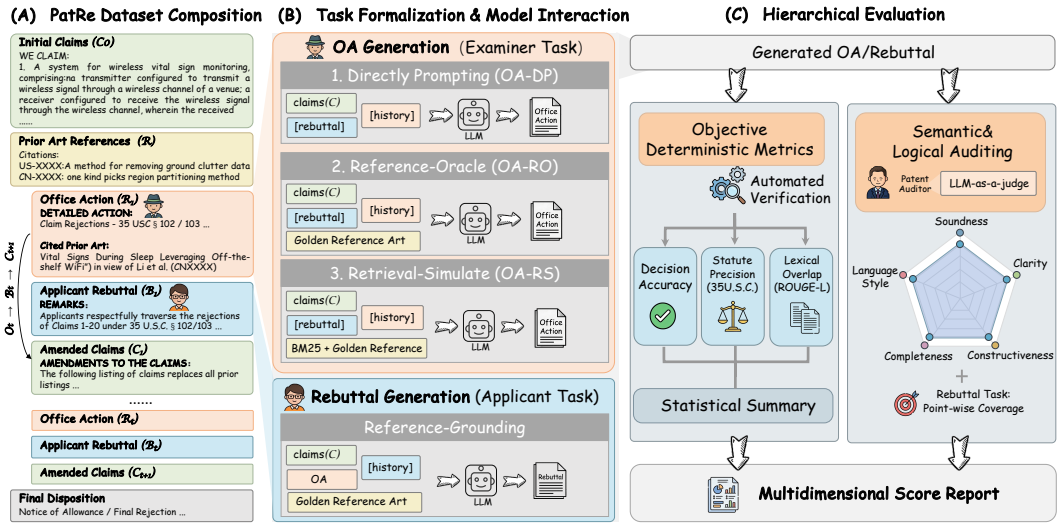


Figure 1: The overall framework of the PatRe benchmark.

Rebuttal Generation, which assesses the model's capacity to simulate applicant responses. Given an examiner's OA, the model must generate legal and technical remarks to contest specific rejection grounds and provide persuasive arguments to overcome the cited prior art, focusing on the logical consistency and legal validity of the defense. **Our main contributions are as follows:**

- We introduce the first full-stage patent examination benchmark, **PatRe**, which focuses on the entire lifecycle of multi-turn Office Action and rebuttal generation. It contains 480 recent patent examination records and covers diverse IPC fields and legal attributes, including final Office Action decisions, intermediate rejection types, and examiner-cited reference patents.
- Moving beyond classification and static extraction, we view patent examination as a dynamic process of justification between the examiner and the applicant. Notably, given the novelty assessment requirements in Office Action generation, we evaluate LLMs under varying levels of cited reference exposure and noise, aligning with realistic examination procedures.
- We conduct extensive experiments on a range of LLMs and provide in-depth insights, including the gap between proprietary and open-source models, the asymmetry between proactive examination (OA) and reactive advocacy (rebuttal), and broader analytical findings.

2 Related Work

Binary Patent Classification and Static Justification Extraction. Early work primarily views patent examination as a classification task. HUPD (Suzgun et al., 2023) utilizes BERT-based models for acceptance prediction, while IPBench (Wang et al., 2025) extends this to modern LLMs. Additionally, PILOT-Bench (Jang et al., 2025) aligns patent board decisions with the IRAC framework. However, these works remain restricted to post-hoc classification and treat legal reasoning as a static annotation problem. They fail to capture the proactive drafting logic and lack the modeling of multi-turn document generation. To move toward an explainable examination, recent studies have targeted specific legal statutes. PANORAMA (Lim et al., 2025) introduced rejection reason identification, while PEDANTIC (Knappich et al., 2025a) focused on 35 U.S.C. 112 (b) by performing justification extraction from OAs. Although they provide valuable insights, they remain static single-stage analyses. They fail to capture the complexity of full OA generation or support multi-turn legal justification generation between examiners and applicants.

Claim Revision and Patent Drafting. Another research direction of patent examination investigates how patent claims evolve over time. PatentEdits (Lee et al., 2024) and Patent-CR (Jiang et al., 2025) aligned initial applications with granted versions to study claim revisions. While these datasets capture the results of the prosecution process, they primarily focus on static version alignment, omitting

the explicit examiner-applicant discussion that fundamentally drives these revisions. Additionally, Pap2Pat (Knappich et al., 2025b) and AutoPatent (Wang et al., 2024b) explored synthesizing patents, potentially increasing applications and intensifying the need for efficient patent examination. Recent work on academic peer review (Jin et al., 2024; Li et al., 2026) and rebuttals (Ma et al., 2026; He et al., 2026) has explored the simulation of iterative scientific communication. However, the patent examination process lacks such benchmarks, which demands stricter adherence to laws like the MPEP (USPTO, 2020). As shown in Table 1, our PatRe bridges this gap by providing the first full-stage benchmark for multi-turn generation of OAs and rebuttals, enabling the entire examination lifecycle modeling.

Table 1: Comparison of PatRe with related patent datasets and benchmarks. **Statute**: Explicit legal basis. **Evolution**: Claim versioning. **Adversarial**: Multi-turn interaction.

Category	Dataset	Year	Task	Statute	Evolution	Adversarial	Full-stage
Classification & Extraction	HUPD (Suzgun et al., 2023)	2023	Discriminative	No	No	No	No
	IPEval (Wang et al., 2024a)	2024	Discriminative	**Yes**	No	No	No
	IPBench (Wang et al., 2025)	2025	Discriminative	No	No	No	No
	PILOT-Bench (Jang et al., 2025)	2025	Discriminative	**Yes**	No	No	No
	PANORAMA (Lim et al., 2025)	2025	Discriminative	**Yes**	No	No	No
	PEDANTIC (Knappich et al., 2025a)	2025	Discriminative	Partial	No	No	No
Revision & Drafting	MOZIP (Ni et al., 2024)	2024	Discriminative	No	No	No	No
	PatentEdits (Lee et al., 2024)	2024	Generative	No	**Yes**	No	Partial
	Patent-CR (Jiang et al., 2025)	2025	Generative	No	**Yes**	No	Partial
Ours	**PatRe**	**2026**	**Generative**	**Yes**	**Yes**	**Yes**	**Yes**

3 PatRe Benchmark

3.1 Task Taxonomy and Formalization

As shown in Figure 1, we conceptualize the patent examination process as a multi-turn strategic interaction between an Examiner (E) and an Applicant (A). Where $\mathcal{D}$ denotes the complete examination history of a given patent, and $|\mathcal{D}|$ represents the number of discussion rounds for current patent. At each examination turn t, the process is grounded in the current version of claims $\mathcal{C}_t$ and the provided prior art $\mathcal{R}$. The examiner first issues an Office Action (OA) $\mathcal{O}_t$ by evaluating $\mathcal{C}_t$ against $\mathcal{R}$ to identify legal and technical defects. Subsequently, the applicant responds with a rebuttal $\mathcal{B}_t$, which provides arguments to contest the rejections or justifies further amendments to the claims $\mathcal{C}_t \rightarrow \mathcal{C}_{t+1}$. We simulate this entire process by introducing two primary types of tasks, as detailed below:

Task 1: Office Action Generation. The objective of OA generation is to evaluate model's ability to simulate the examiner's decision-making process. Given the current version of claims $\mathcal{C}_t$ and the potentially preceding rebuttal $\mathcal{B}_{t-1}$ (when $t > 1$), the model is instructed to generate a formal OA $\mathcal{O}$. We formalize this under three settings with varying levels of information guidance:

- *Directly Prompting* **(OA-DP)**: We leverage a zero-shot prompting setting to instruct the model to generate the Office Action $\mathcal{O}$ by relying solely on its internal parameters and pre-trained knowledge, without access to any specific external prior art.
- *Reference-Oracle Generation* **(OA-RO)**: We provide the model with an oracle reference set $\mathcal{R}_{\text{oracle}}$, consisting of the ground-truth references cited by the examiner as well as the references cited by the applicant and considered during examination. The model must autonomously select the most relevant references from $\mathcal{R}_{\text{oracle}}$ to construct legal justifications for the Office Action $\mathcal{O}$. This subtask evaluates the model's performance under the most comprehensive information setting.
- *Retrieval-Simulated Generation* **(OA-RS)**: The model is provided with a noisy candidate pool $\mathcal{R}_{noise}$ containing oracle references and retrieved distractors. It must identify novelty-relevant references from the noise to generate the Office Action $\mathcal{O}$.

Task 2: Applicant Rebuttal Generation. This task simulates the responsive phase of patent examination. Given the current Office Action $\mathcal{O}_t$ and the associated prior art $\mathcal{R}$ at turn t, the model

must generate a rebuttal $\mathcal{B}_t$. Unlike procedural legal filings, we focus on the substantive argumentation required to overcome the examiner's objections. Formally, we model this as $P(\mathcal{B}_t|\mathcal{C}_t, \mathcal{O}_t, \mathcal{R})$, which requires the model to perform a tripartite alignment: ***(i)*** grounding legal arguments in the specific rejection grounds of $\mathcal{O}_t$, ***(ii)*** contrasting the technical features of $\mathcal{C}_t$ against $\mathcal{R}$, and ***(iii)*** maintaining logical consistency with the intended scope of the invention.

These tasks collectively establish the PatRe for evaluating multi-dimensional technical and legal reasoning within the patent domain. Where PatRe-OA challenges the model's statutory interpretability by requiring the mapping of claim features to prior art disclosures under legal constraints. Conversely, PatRe-Rebuttal assesses adversariality through the model's proficiency in synthesizing counter-arguments that adhere to both the technical scope and the MPEP legal framework.

3.2 Evaluation Metric Design

To provide a comprehensive assessment of the generated Office Action and rebuttal documents, we establish a hierarchical evaluation framework that moves beyond surface-level linguistic similarity to capture legal and technical nuances, including two levels: (I) Deterministic metrics for objective verification; (II) LLM-as-a-Judge metrics for deep semantic and logical auditing.

I: Objective Deterministic Metrics. To ensure the factual correctness of the generated OA and rebuttal, we implement an objective metric suite that measures alignment with expert-verified label. **(1) Statutory and Decision Alignment**, the accuracy of the legal basis and final decision. We compute *Decision Accuracy* as a binary indicator of whether the predicted OA decision matches the label. Where more fine-grained *Statute Precision* measures the precision of the invoked 35 U.S.C. statutes, *i.e.*

$$\text{Statute Precision} = \frac{|S_{\text{pred}} \cap S_{\text{gt}}|}{|S_{\text{pred}}|} \tag{1}$$

where S_{pred} and S_{gt} denote the sets of statutes cited in the generated and ground-truth legal documents, respectively. **(2) Lexical Overlap**, which adopts Rouge-L (Lin, 2004) to measure the sequential alignment between generated and ground-truth texts.

II: Semantic and Logical Auditing (LLM-as-a-Judge). To evaluate the semantic and logical quality of generated OA and rebuttal, we employ `Gemini-3.1-Flash-Lite` as a patent auditor, following (Wang et al., 2025). Each document is scored on a 1-10 scale across five dimensions: **(1) Soundness**, which evaluates the technical and legal soundness of the generated texts; **(2) Clarity**, which focuses on the legal readability, logical coherence, and specific structure; **(3) Constructiveness**, which emphasizes the actionability of the model response; In OA generation, it measures the usefulness of examiner guidance; in rebuttal generation, it reflects the strength of counterarguments. **(4) Completeness**, which focuses on the utility of the feedback; **(5) Language Style**, which focuses on adherence to the legal style and procedural conventions of Office Action and rebuttal drafting. Especially for rebuttal generation task, we introduce the **Point-wise Coverage**, which evaluates the responsive rate to atomic OA rejection points, providing a semantic measure of defense thoroughness.

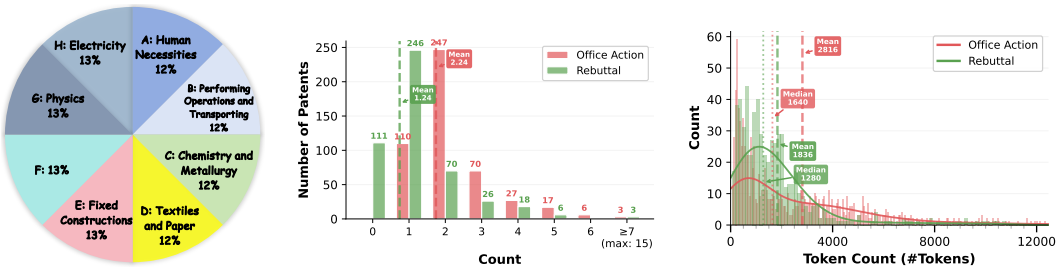
3.3 Data Collection and Processing

To construct the PatRe benchmark, we develop a reproducible data collection pipeline to extract the longitudinal examination history of patents from the USPTO public database. Unlike prior datasets, which typically capture final version of granted patents, we focus on reconstructing the complete trajectory of a patent application by recording the full history sequence of examiner-applicant interactions. For each patent record, we collect the full-stage correspondence starting from the initial filing, including the verbatim text of all OAs, the corresponding applicant rebuttals, the iterative versions of claims at each stage ($\mathcal{C} \rightarrow \mathcal{C}'$), and the complete reference list cited by examiners.

To ensure the high fidelity of PatRe, we apply automated filtering to remove noisy or incomplete records, followed by manual verification by three PhD students specializing in Intellectual Property. They verify claims, OAs, rebuttals, amendments, and cited references along the original USPTO prosecution chronology, with inconsistencies resolved through discussion. Finally, all personally identifiable information is redacted.

3.4 Dataset Statistics

PatRe comprises 480 complete prosecution trajectories across all eight IPC sections (A–H), yielding 1,075 Office Action and 596 rebuttal instances (1,671 in total). Each trajectory preserves OAs, applicant rebuttals, claim revisions, and legal metadata including rejection types and cited references. OA-DP, OA-RO, and OA-RS share the same 1,075 OA instances and differ only in evidence access. As an evaluation benchmark, PatRe uses no train/dev/test split; all models are evaluated on the same fixed set of cases. Figure 2 shows (a) IPC coverage, (b) OA and rebuttal rounds, and (c) document lengths; additional statistics on rejection types, OA types, and cited references are provided in Appendix A.



(a) Distribution of IPC sections.

(b) Distribution of Office Action and rebuttal round count.

(c) Distribution of Office Action and rebuttal document length.

Figure 2: Distributions across IPC sections, rounds count, and length. IPC Section F covers Mechanical Engineering, Lighting, Heating, Weapons, and Blasting.

4 Experiment

4.1 Experimental Setup

Evaluated Models. We benchmark different LLMs covering a broad range of sizes, architectures and families, with model details provided in Appendix B. These include commercial proprietary models such as GPT series (Singh et al., 2025) (GPT-5-mini and GPT-4o-mini), Gemini series (Team, 2023) (Gemini-2.5-Flash) and DeepSeek series (Liu et al., 2025) (DeepSeek-V3.2). We also include open-source models ranging from 8B to 70B, including LLaMA series (Grattafiori et al., 2024), Qwen3.5 series (Qwen Team, 2026) and Gemma3 series (Gemma Team, 2025) models.

Implementation Details. Proprietary models are evaluated via official APIs, while open-source models are run with vLLM (Kwon et al., 2023) on 8 NVIDIA A800 GPUs. For OA-RS, BM25 constructs the default distractor pool, with BGE-M3 additionally used for retrieval robustness analysis. We set each model's maximum output length to its context limit and use temperature $= 0$ throughout. `Gemini-3.1-Flash-Lite` serves as the primary LLM judge; we additionally evaluate OA-RO outputs with `GPT-5.4` and `Claude-Opus-4.8` to assess judge robustness. Prompts and model details, including API costs, are provided in Appendix E and Appendix B, respectively. Additional labels, including rejection types and citations, are extracted using regular expressions.

4.2 Main Results

***Observation* 1: Proprietary models, especially GPT-5-mini, demonstrate consistently superior performance across both Office Action and Rebuttal generation tasks.** As shown in Table 2, `GPT-5-mini` achieves the highest Decision Accuracy in both OA-DP (51.4%) and OA-RO (50.0%), while is the second best in the OA-RS setting (52.7%). This performance extends to rebuttal generation (Table 4), where it reports a Point-wise Coverage of 90.5% and a Soundness score of 8.71. Notably, the performance disparity between proprietary and open-source models remains relatively narrow in the structured decisional logic required for Office Action tasks. However, a more pronounced gap emerges in rebuttal generation tasks, where proprietary models exhibit a distinct advantage in the technical precision and global logical alignment necessary for effective adversarial reasoning. This

Table 2: Office Action generation results. *Dec.*, *Stat.*, *R-L*, and *Ovr.* denote Decision Accuracy, Statute Precision, Rouge-L, and the primary LLM judge's average score, respectively.

Model	OA-DP				OA-RO				OA-RS			
	Dec.	Stat.	R-L	Ovr.	Dec.	Stat.	R-L	Ovr.	Dec.	Stat.	R-L	Ovr.
Proprietary Models												
GPT-5-mini	**51.4**	45.1	14.8	**4.56**	**50.0**	49.2	17.4	**4.89**	52.7	42.5	15.4	**5.39**
Gemini-2.5-Flash	50.0	46.6	16.4	4.26	46.4	**56.3**	**20.5**	4.36	**52.8**	**51.5**	**20.5**	4.37
DeepSeek-V3.2	47.6	**47.1**	**17.3**	4.26	49.7	55.4	19.9	4.34	42.2	47.0	18.7	3.50
GPT-4o-mini	24.4	44.8	14.0	3.75	43.3	55.1	15.7	3.61	36.3	48.9	16.4	3.59
Open-Source Models												
LLaMA3.1-8B-it	41.4	42.5	**19.3**	3.12	45.6	46.2	**24.2**	3.07	39.9	41.0	**23.9**	3.19
Qwen3.5-9B	41.7	38.7	17.7	4.07	44.3	48.7	20.0	4.11	43.4	43.1	18.8	4.11
Gemma3-12B-it	45.1	42.7	15.3	3.68	43.4	52.6	19.6	3.59	41.6	42.6	19.8	3.51
Gemma3-27B-it	39.5	43.9	14.8	3.75	45.4	**57.7**	18.0	3.65	38.1	**49.8**	18.9	3.68
Qwen3.5-27B	**48.8**	43.7	18.3	**4.35**	**47.6**	56.3	21.4	**4.37**	**50.5**	47.7	20.0	**4.30**
LLaMA3.3-70B-it	10.2	**46.3**	15.1	3.45	9.7	54.7	16.6	3.40	21.8	48.9	17.6	3.50

Table 3: Primary LLM judge scores for Office Action generation (1–10). *Sou.*, *Cla.*, *Con.*, *Com.*, and *Sty.* denote Soundness, Clarity, Constructiveness, Completeness, and Language Style, respectively.

Model	OA-DP					OA-RO					OA-RS				
	Sou.	Cla.	Con.	Com.	Sty.	Sou.	Cla.	Con.	Com.	Sty.	Sou.	Cla.	Con.	Com.	Sty.
Proprietary Models															
GPT-5-mini	**2.46**	**7.67**	**2.16**	2.28	**8.22**	**3.07**	**7.77**	**2.65**	**3.14**	**7.84**	**3.43**	**7.89**	**3.09**	**3.86**	**8.69**
Gemini-2.5-Flash	2.38	7.41	1.67	2.30	7.52	2.60	7.21	1.79	2.60	7.60	2.54	7.24	1.74	2.74	7.61
DeepSeek-V3.2	2.26	7.48	1.73	**2.33**	7.48	2.39	7.43	1.86	2.61	7.42	2.18	7.39	1.66	2.48	7.38
GPT-4o-mini	2.06	6.81	1.34	1.84	6.74	2.05	6.32	1.45	2.02	6.21	1.86	6.43	1.37	1.94	6.33
Open-Source Models															
LLaMA3.1-8B-it	1.79	5.44	1.26	1.91	5.22	1.78	5.26	1.24	1.98	5.09	1.80	5.45	1.25	2.10	5.35
Qwen3.5-9B	1.98	7.28	1.65	2.17	7.27	2.26	6.94	1.84	2.58	6.92	2.22	6.99	1.78	2.60	6.97
Gemma3-12B-it	1.87	6.66	1.40	1.86	6.61	1.98	6.34	1.36	2.00	6.27	1.76	6.35	1.27	1.89	6.29
Gemma3-27B-it	1.80	6.99	1.35	1.66	6.96	2.01	6.51	1.34	1.95	6.45	1.82	6.76	1.27	1.84	6.72
Qwen3.5-27B	**2.31**	**7.53**	**1.94**	**2.45**	**7.53**	**2.56**	**7.19**	**2.09**	**2.83**	**7.18**	**2.41**	**7.19**	**1.94**	**2.77**	**7.17**
LLaMA3.3-70B-it	1.95	6.21	1.31	1.75	6.04	1.91	6.08	1.28	1.80	5.95	1.91	6.27	1.27	1.86	6.17

suggests that while open-source models are becoming increasingly viable for categorical patentability determinations, a functional bottleneck persists in their ability to handle the complex linguistic and logical demands of applicant-examiner discourse.

***Observation* 2: The evolution across OA generation settings underscores a performance divergence between statutory citation and substantive adjudication.** The transition across three OA settings reveals how information guidance impacts models differently. Specifically, while the OA-RO setting acts as an upper bound for Statute Precision (Stat.) due to the availability of oracle references, it does not consistently improve Decision Accuracy (Dec.). For instance, `Gemini-2.5-Flash` achieves its peak Stat. in OA-RO, yet its Dec. falls below its zero-shot performance in OA-DP. In OA-RS setting, the top-tier models demonstrate the ability to filter noise and maintain decisional stability. These observations indicate that, while external evidence strengthens formal legal alignment, it does not inherently reinforce the logical consistency required for accurate patentability determinations.

***Observation* 3: Models exhibit a significant performance decoupling across various LLM-as-a-Judge dimensions, particularly between surface language style and internal logic.** We report the detailed LLM-as-a-judge scores across five dimensions in Table 3 and Table 4. Models perform well in language Style and Clarity, but lag far behind in Soundness, Constructiveness, and Completeness.

Table 4: Results for Rebuttal Generation. Where *Cov.* denotes Point-wise Coverage. The LLM-as-a-judge quality dimensions, scored on a 1–10 scale, include Soundness, Clarity, Constructiveness (Construct.), Completeness (Complete.), and Language Style (Lang.-Sty.).

Model	Cov.	Rouge-L	Soundness	Clarity	Construct.	Complete.	Lang.-Sty.	Average
Proprietary Models								
GPT-5-mini	**90.5**	14.9	**8.71**	**9.62**	**8.78**	**9.12**	**9.65**	**9.18**
Gemini-2.5-Flash	81.5	16.4	7.77	8.80	8.06	8.12	8.96	8.34
DeepSeek-V3.2	79.1	17.9	7.66	9.06	8.06	7.94	9.15	8.37
GPT-4o-mini	27.8	**21.1**	2.82	6.43	3.44	3.01	6.80	4.50
Open-Source Models								
LLaMA3.1-8B-it	29.5	18.4	2.75	4.63	3.08	3.16	4.94	3.71
Qwen3.5-9B	63.4	17.7	6.09	8.02	6.71	6.33	8.28	7.09
Gemma3-12B-it	30.5	21.8	3.12	6.19	3.70	3.21	6.67	4.58
Gemma3-27B-it	39.9	19.5	4.05	6.96	4.74	4.17	7.42	5.47
Qwen3.5-27B	**79.2**	17.5	**7.61**	**8.87**	**8.05**	**7.93**	**8.99**	**8.29**
LLaMA3.3-70B-it	30.9	**21.9**	3.08	6.22	3.65	3.37	6.65	4.59

This confirms a pronounced discrepancy between linguistic form and legal content, where a professional style on surface masks a logic flaw in technical adjudication. While this dimensional asymmetry persists across the entire examination lifecycle, its magnitude changes with tasks.

In Figure 3, when models transition from proactive examination (OA) to reactive defense (Rebuttal), Soundness and Constructiveness more than double, while other dimensions also show marked improvements. This suggests that the constraints of legal reasoning are partially mitigated when models respond to explicit grounds. But they perform substantially better as responders than as proactive examiners. Overall, while professionalism in style is relatively mature, logical reasoning and analysis in Soundness and Completeness remain insufficient in patent examination.

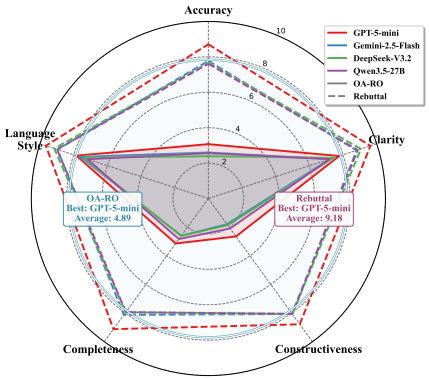


Figure 3: LLM-as-a-judge scores across 5 dimensions for generated OA-RO and rebuttal.

4.3 Analysis

Finding 1: LLMs exhibit far greater proficiency in *reactive defense* than in *proactive problem discovery*. As illustrated in Figure 4, a significant performance gap exists between the models' roles as applicants and examiners. Surprisingly, almost all models achieve a high *average LLM-as-a-Judge score* during rebuttal generation, but notably lower in OA generation. For example, while `DeepSeek-V3.2` and `Qwen3.5-27B` struggle to exceed an overall score of 4.5 in various OA settings, they consistently achieve scores above 8.0 in rebuttals. This performance gap indicates a fundamental difference in task complexity: issuing an OA requires proactive problem discovery, where the model must act as an impartial arbitrator to identify statutory defects within dense claims. Current models naturally aligned with the role of patent attorney, excelling at persuasive advocacy, but struggle with the rigorous, impartial logical deduction required of an examiner.

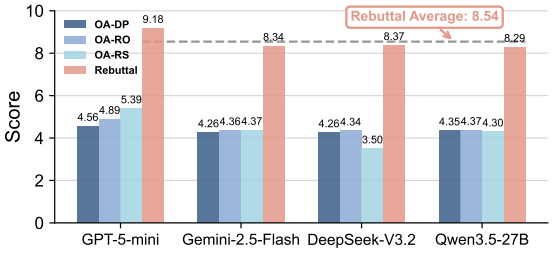


Figure 4: Office Action vs. Rebuttal generation performance (average LLM-as-a-judge score) of LLMs.

Finding 2: A pervasive hyper-critical bias leads to exceptionally high false-rejection rates in Allowance cases. Our quantitative error analysis in Figure 5, reveals that when processing applications

that should be granted, models frequently hallucinate non-existent conflicts with prior art to justify a rejection. Especially for `LLaMA-3.3-70B-it`, which presents a paradoxical performance profile: while it maintains competitive *Statute Precision* (54.7%), its Decision Accuracy suffers a catastrophic deficit, falling as low as 9.7%. Specifically, 37.47% of the errors in `LLaMA-3.3-70B-it` stem from the premature classification of non-final cases as final rejections. This suggests that the primary bottleneck for the LLaMA series appears to be decision-boundary calibration rather than statute matching quality alone. This fault-finding behavior reflects an immaturity in legal balance, as models seem to equate rigorous examination with a guaranteed rejection. This bias is particularly detrimental in the patent domain, as it indicates the model's inability to recognize when an invention has successfully overcome the threshold of patentability. Instead of performing a neutral assessment, models tend to over-scrutinize valid applications, fabricating technical defects to maintain a stance of rejection.

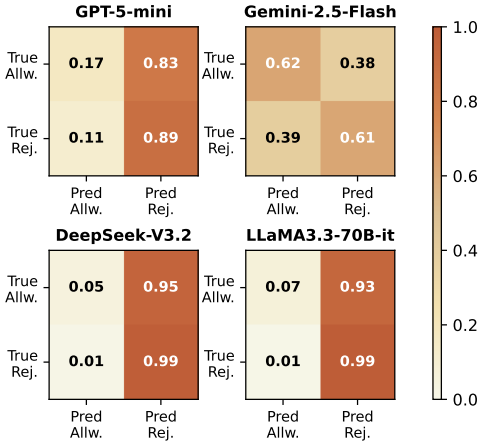


Figure 5: Confusion matrix heatmap for allowance (Allw.) vs. rejection (Rej.).

Finding 3: LLMs exhibit severe logical inconsistency and over-extension in applying conceptually demanding legal statutes. As illustrated in Figure 6, we categorize statutory errors into two primary modes: ***statutory fabrication*** (False Positives, FP), where the model predicts a violation that is not present and ***statutory omission*** (False Negatives, FN), where the model fails to identify a violation that is actually present. For 35 U.S.C. §101 (Patent Eligibility), models suffer from a **dual failure**: they exhibit the highest fabrication rate at 72.8% while failing to identify 48.8% of actual eligibility defects. Similarly, §102 (Novelty) demonstrates significant instability, with high fabrication (FP: 47.5%) and omission (FN: 40.3%). This indicates unstable legal reasoning in LLMs, causing hallucinated oversights. In contrast, for statutes §103 (Obviousness) and §112 (Definiteness), the error shifts toward aggressive **over-enforcement**. While models are less likely to miss a valid rejection (FN: 7.6% and 19.9% respectively), they frequently over-extend these rules. Notably, §112 exhibits the second-highest fabrication rate at 60.0%, while §103 shows a similar over-extension at 47.6%. These findings suggest that while LLMs are prone to over-identifying violations, their ability to accurately distinguish between statutory and non-statutory matter is profoundly insufficient.

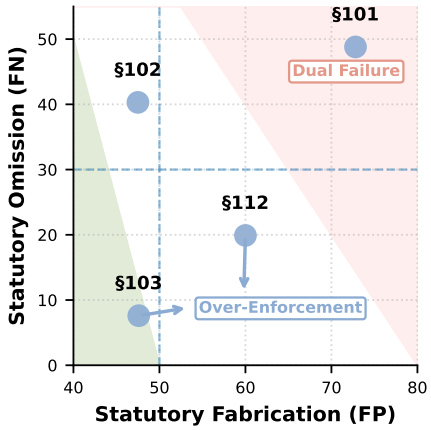


Figure 6: Statutory error mode landscape: Fabrication vs. Omission.

Finding 4: LLMs struggle to select novelty-relevant evidence under noisy retrieval. We evaluate Reference Citation Accuracy (RCA):

$$\mathrm{RCA} = \frac{|R_{\text{pred}} \cap R_{\text{valid}}|}{|R_{\text{pred}}|}, \tag{2}$$

where R_{pred} and R_{valid} are the predicted and valid references. As shown in Figure 7, oracle references substantially improve RCA over no references, but RCA drops when they are mixed with retrieved distractors. Importantly, retrieval in OA-RS is used only to introduce distractors and increase the difficulty of evidence selection, without altering the core experimental setting. Thus, the observed degradation reflects the difficulty

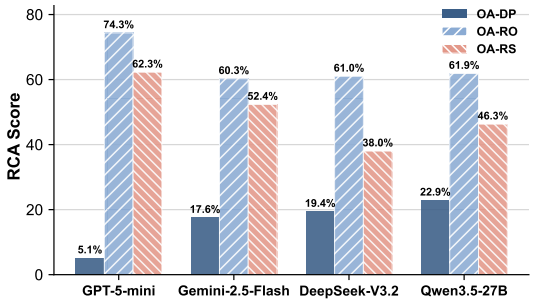


Figure 7: RCA across OA settings.

of selecting novelty-relevant evidence under noise rather than retrieval performance itself. Replacing BM25 with the dense retriever BGE-M3 while keeping other settings unchanged further reduces RCA from 62.3% to 39.17% for GPT-5-mini and from 52.4% to 28.05% for Gemini-2.5-Flash, while DeepSeek-V3.2 remains comparable (38.00% vs. 38.56%). These results further show that more plausible distractors increase the difficulty of evidence selection in OA-RS. See Appendix C for detailed settings and full results.

Finding 5: Lexical metrics exhibit a significant performance decoupling from substantive legal validity. Lexical overlap correlates weakly with both Decision Accuracy (Kendall's $\tau = 0.0258$) and LLM-as-a-judge scores ($\tau = 0.1440$), indicating that high Rouge-L does not necessarily reflect legal correctness or reasoning quality. We further conduct human evaluation on OA-RO and rebuttal generation using three experts under the same criteria as the LLM judge. As shown in Table 5, Rouge-L aligns less with human judgments than LLM-based evaluation, while human experts show substantial agreement (Pearson's $r = 0.7285$; dimension-wise results in Appendix C). To assess judge robustness, we additionally evaluate OA-RO outputs with GPT-5.4 and Claude-Opus-4.8. We provide the detailed scores from the three LLM judges in Table 14. The three judges show strong agreement ($\tau/\rho/r = 0.881/0.956/0.972$ on average), and their ensemble correlates well with human judgments ($0.619/0.786/0.845$), with no systematic same-family preference (Appendix C). These results suggest that LLM-based evaluation better captures substantive patent-examination quality than lexical metrics and is robust across different judge models.

Table 5: Correlation (↑) with human judgments (p-value ↓).

Metric	Kendall τ	Pearson r	Spearman ρ
Decision Accuracy	**0.4490 (0.0000)**	0.4863 (0.0000)	0.5412 (0.0000)
Statute Precision	0.2852 (0.0040)	0.2766 (0.0053)	0.2108 (0.0059)
Rouge-L	0.1351 (0.0095)	0.1672 (0.0293)	0.1931 (0.0117)
LLM-as-a-judge	0.4480 (0.0000)	**0.6808 (0.0000)**	**0.6231 (0.0000)**

4.4 Case Study

To further investigate LLM performance in patent examination, we analyze representative cases across different task settings, including OA-RO, OA-RS, OA-DP, and rebuttal generation. Table 17–18 (See Appendix D) present examples across various stages of the patent examination lifecycle to highlight the gap between professional drafting and model capabilities. These examples reveal that in OA generation, models often struggle with substantive functional mapping and tend to adopt templated legal discourse to mask a lack of actual technical ambiguity, whereas they demonstrate notable reactive proficiency in Rebuttals by effectively identifying flaws in rejections. Overall, these cases demonstrate that current LLMs excel at mimicking the linguistic form of patent examination but struggle with the rigorous, evidence-based reasoning required for precise prior art analysis and statutory evaluation.

5 Conclusion

We introduce PatRe, the first full-stage benchmark for patent Office Action and rebuttal generation, modeling patent examination as a dynamic, multi-turn process of justification and response beyond binary classification and static extraction. Extensive experiments reveal that current LLMs remain insufficient as independent examination systems and show a notable gap between proprietary and open-source models, while highlighting their potential to assist examiners and applicants. PatRe currently focuses on English, text-only USPTO prosecution records and contains relatively few rare rejection types, reflecting their natural distribution. Future work will extend PatRe to larger-scale, multimodal, multilingual, and cross-jurisdictional settings, which require adapting jurisdiction-specific legal standards and procedural metadata while preserving the core multi-turn formulation.

AI Use Statement

Generative AI tools were used to assist with language polishing, translation, and code refinement and debugging. All research ideas, methodological choices, experimental settings, analyses, and conclusions were developed and verified by the authors. We manually reviewed all AI-assisted outputs for correctness and take full responsibility for the final content.

REFERENCES

Gemma Team. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Zhitao He, Zongwei LYU, and Yi R. Fung. Dancing in chains: Strategic persuasion in academic rebuttal via theory of mind. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=zkCZDdeS9s.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Yehoon Jang, Chaewon Lee, Hyun-seok Min, and Sungchul Choi. Pilot-bench: A benchmark for legal reasoning in the patent domain with irac-aligned classification tasks. In *Proceedings of the Natural Legal Language Processing Workshop 2025*, pp. 240–280, 2025.

Lekang Jiang, Pascal A Scherz, and Stefan Goetz. Patent-cr: A dataset for patent claim revision. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2300–2314, 2025.

Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. Agentreview: Exploring peer review dynamics with llm agents. In *Conference on Empirical Methods in Natural Language Processing*, 2024. URL https://api.semanticscholar.org/CorpusID:270562353.

Valentin Knappich, Annemarie Friedrich, Anna Hätty, and Simon Razniewski. Pedantic: A dataset for the automatic examination of definiteness in patent claims. *arXiv preprint arXiv:2505.21342*, 2025a.

Valentin Knappich, Anna Hätty, Simon Razniewski, and Annemarie Friedrich. PAP2PAT: Benchmarking outline-guided long-text patent generation with patent-paper pairs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 9524–9554, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.496. URL https://aclanthology.org/2025.findings-acl.496/.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Haotong Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. *Proceedings of the 29th Symposium on Operating Systems Principles*, 2023. URL https://api.semanticscholar.org/CorpusID:261697361.

Ryan Lee, Alexander Spangher, and Xuezhe Ma. Patentedits: Framing patent novelty as textual entailment. *arXiv preprint arXiv:2411.13477*, 2024.

Shuaimin Li, Liyang Fan, Yufang Lin, Zeyang Li, Xian Wei, Shiwen Ni, Hamid Alinejad-Rokny, and Min Yang. Automatic paper reviewing with heterogeneous graph reasoning over llm-simulated reviewer-author debates. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 31717–31725, 2026.

Hyunseung Lim, Sooyohn Nam, Sungmin Na, Ji Yong Cho, June Yong Yang, Hyungyu Shin, Yoonjoo Lee, Juho Kim, Moontae Lee, and Hwajung Hong. PANORAMA: A dataset and benchmarks capturing decision trails and rationales in patent examination. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=JWewCpdjjq.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*, 2025.

Qianli Ma, Chang Guo, Zhiheng Tian, Siyu Wang, Jipeng Xiao, Yuanhao Yue, and Zhipeng Zhang. Paper2rebuttal: A multi-agent framework for transparent author response assistance. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 46134–46166, 2026.

Shiwen Ni, Minghuan Tan, Yuelin Bai, Fuqiang Niu, Min Yang, Bowen Zhang, Ruifeng Xu, Xiaojun Chen, Chengming Li, and Xiping Hu. Mozip: A multilingual benchmark to evaluate large language models in intellectual property. In Nicoletta Calzolari, Min-Yen Kan, Véronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pp. 11658–11668. ELRA and ICCL, 2024. URL https://aclanthology.org/2024.lrec-main.1018.

Qwen Team. Qwen3.5: Accelerating productivity with native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Mirac Suzgun, Luke Melas-Kyriazi, Suproteem Sarkar, Scott D Kominers, and Stuart Shieber. The harvard uspto patent dataset: A large-scale, well-structured, and multi-purpose corpus of patent applications. *Advances in neural information processing systems*, 36:57908–57946, 2023.

Gemini Team. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

USPTO. Manual of patent examining procedure. pp. 4 v. (loose–leaf), 2020. URL http://tind.wipo.int/record/41638. This resource was extracted from USPTO.gov.

Qiyao Wang, Jianguo Huang, Shule Lu, Yuan Lin, Kan Xu, Liang Yang, and Hongfei Lin. Ipeval: A bilingual intellectual property agency consultation evaluation benchmark for large language models. *arXiv preprint arXiv:2406.12386*, 2024a.

Qiyao Wang, Shiwen Ni, Huaren Liu, Shule Lu, Guhong Chen, Xi Feng, Chi Wei, Qiang Qu, Hamid Alinejad-Rokny, Yuan Lin, et al. Autopatent: A multi-agent framework for automatic patent generation. *arXiv preprint arXiv:2412.09796*, 2024b.

Qiyao Wang, Guhong Chen, Hongbo Wang, Huaren Liu, Minghui Zhu, Zhifei Qin, Linwei Li, Yilin Yue, Shiqiang Wang, Jiayan Li, et al. Ipbench: Benchmarking the knowledge of large language models in intellectual property. *arXiv preprint arXiv:2504.15524*, 2025. URL https://arxiv.org/abs/2504.15524.

Daoze Zhang, Zhijian Bao, Sihang Du, Zhiyi Zhao, Kuangling Zhang, Dezheng Bao, and Yang Yang. Re^2: A consistency-ensured dataset for full-stage peer review and multi-turn rebuttal discussions. *arXiv preprint arXiv:2505.07920*, 2025. URL https://arxiv.org/abs/2505.07920.

A More Details about Data Statistics

Preprocessing and Filtering. We automatically filter records with excessive document noise, missing essential prosecution documents, parsing failures, or incomplete and inconsistent metadata that prevent reliable reconstruction of the examination trajectory. The retained records are aligned according to the original USPTO prosecution chronology, matching claims, Office Actions, applicant responses, and claim amendments by their temporal order, and are subsequently subjected to manual quality verification.

Data Quality Control. Three PhD students specializing in Intellectual Property manually verify the completeness and consistency of claims, Office Actions, applicant rebuttals, claim amendments, and cited references. Disagreements are resolved through discussion among the annotators.

We collect patent examination records from the public USPTO Open Data Portal (ODP)[1], which provides publicly accessible patent examination data under public-domain open-data terms. All patents in our dataset were published after 2024, following their Notice of Allowance (NOA) dates or Final Decision dates. Given the legal nature of patent examination, we report the distribution of rejection types, including §103 (Obviousness), §112 (Written Description/Enablement), §102 (Lack of Novelty), §101 (Patent-Eligible Subject Matter), and double patenting (DP), as summarized in Table 6. Notably, multiple rejection types may appear within a single case.

Table 6: Distribution of rejection types.

Rejection Type	**Count**	**Percentage**
103 (Obviousness)	415	40.53%
112 (Written Description/Enablement)	230	22.46%
102 (Lack of Novelty)	202	19.73%
DP (Double Patenting)	122	11.91%
101 (Patent-Eligible Subject Matter)	55	5.37%
Total	1024	100.00%

In the Office Action generation task, a complete patent examination history may involve multiple Office Actions. Accordingly, we report the distribution of Office Action types, as summarized in Table 7, covering four categories: Notice of Allowance, Non-Final Rejection, Final Rejection, and Ex Parte Quayle Action.

Table 7: Distribution of Office Action (OA) types.

OA Type	**Count**	**Percentage (%)**
Notice of Allowance	479	44.56
Non-Final Rejection	435	40.47
Final Rejection	152	14.14
Ex Parte Quayle Action	9	0.84
Total	1075	100.00

Beyond legal labels, to reflect the novelty requirements in patent examination, we analyze the distribution of cited references within each patent record, as shown in Table 8.

B More Details about Evaluated Models

We summarize the evaluated models, context and output limits, and access methods in Table 9. Proprietary models and DeepSeek-V3.2 are evaluated via official APIs, while open-source models up to 70B are deployed with vLLM (Kwon et al., 2023) on 8 NVIDIA A800 GPUs. We use temperature $= 0.0$ throughout, with model-specific context and output limits summarized in Table 9. API costs are reported in Table 10.

[1] https://data.uspto.gov/home

Table 8: Citation vs. discussion statistics across categories. "Avg Cited" and "Avg Discussed" denote the average numbers of cited and explicitly discussed references per patent record, respectively. "Difference" is computed as cited minus discussed, and "Ratio" represents the proportion of discussed references relative to cited ones.

Category	Avg Cited	Avg Discussed	Difference	Ratio (%)
A	11.33	9.60	1.73	84.71%
B	8.82	6.15	2.67	69.75%
C	2.88	5.77	-2.88	200.00%
D	9.32	4.32	5.00	46.33%
E	9.75	4.72	5.03	48.38%
F	8.92	3.72	5.20	41.68%
G	9.70	9.52	0.18	98.11%
H	8.08	6.57	1.52	81.24%
All	8.60	6.30	2.30	73.78%

Table 9: The overview of evaluated models.

Model	Size	Max Context	Max Output	Access
GPT-5-mini (Singh et al., 2025)	–	400K	128K	OpenAI API
GPT-4o-mini (Hurst et al., 2024)	–	128K	16K	OpenAI API
Gemini-2.5-Flash (Team, 2023)	–	1M	64K	Google API
DeepSeek-V3.2 (Liu et al., 2025)	671B	128K	64K	DeepSeek API
Llama3.1-8B-Instruct (Grattafiori et al., 2024)	8B	128K	–	Weights
Qwen3.5 (Qwen Team, 2026)	9/27B	1M	64K	Weights
Gemma3-Instruct (Gemma Team, 2025)	9/27B	128K	–	Weights
Llama3.3-70B-Instruct (Grattafiori et al., 2024)	70B	128K	–	Weights

Table 10: Cost of different models (accessed via API).

Model	GPT-5-mini	GPT-4o-mini	Gemini-2.5-Flash	DeepSeek-V3.2
Cost (USD)	37.85	11.08	14.13	23.49

C More Results

Retrieval Robustness Analysis. The default OA-RS setting uses BM25 to construct a noisy candidate pool by augmenting the oracle references with retrieved distractors. OA-RO provides the clean oracle-only condition, corresponding to the evidence available under ideal retrieval, whereas OA-RS introduces retrieved distractors around the same oracle evidence. Thus, the difference between OA-RO and OA-RS forms a controlled ablation over reference noise, designed to evaluate reference selection and evidence-grounded reasoning rather than retrieval quality.

To examine whether this finding depends on the choice of retriever, we additionally reconstruct OA-RS using BGE-M3, a stronger dense retriever, while keeping the oracle evidence and all other experimental settings unchanged. Compared with BM25, BGE-M3 retrieves semantically closer distractors, creating a more challenging evidence-selection setting.

As shown in Table 11, RCA decreases substantially when oracle references are mixed with distractors, despite the relevant evidence remaining available in every candidate pool. Moreover, replacing BM25 with BGE-M3 further reduces RCA for GPT-5-mini and Gemini-2.5-Flash, while DeepSeek-V3.2 remains approximately unchanged. This indicates that semantically similar distractors make novelty-relevant reference selection more difficult for current LLMs.

These results support the interpretation of OA-RS as a controlled evidence-selection-under-noise setting rather than a benchmark of end-to-end patent retrieval. We note that real patent examination involves richer search mechanisms, including citation networks, patent families, iterative query reformulation, and reranking; evaluating such end-to-end retrieval pipelines is beyond the current scope of PatRe.

Table 11: Reference Citation Accuracy (RCA, %) under different evidence conditions. OA-RO contains clean oracle evidence, while BM25 and BGE-M3 augment the same oracle evidence with retrieved distractors. Δ denotes the RCA change from BM25 to BGE-M3.

Model	OA-RO	BM25	BGE-M3	Δ
GPT-5-mini	74.30	62.30	39.17	-23.13
Gemini-2.5-Flash	60.30	52.40	28.05	-24.35
DeepSeek-V3.2	61.00	38.00	38.56	+0.56

Human Evaluation Setup. We employ three PhD students specializing in Intellectual Property (IP), who are trained and have academic backgrounds in IP, equipping them with foundational knowledge of both the technical and legal aspects of IP. All human evaluators are provided with the same evaluation criteria as those in Figure 10 and Figure 11. We sample 100 generated OAs and 100 generated rebuttals across 7 models and all IPC sections, and evaluate them under a blind setting, where the evaluators are unaware of which model generated each instance.

Table 12: Inter-rater agreement between three human experts across five evaluation dimensions, measured using Pearson's r, Spearman's ρ, and Kendall's τ.

Dimension	Pearson r	Spearman ρ	Kendall τ
Accuracy	0.6061	0.6545	0.5380
Clarity	0.6899	0.6863	0.5718
Completeness	0.6858	0.7259	0.5819
Constructiveness	0.5929	0.6621	0.5486
Language Style	0.4676	0.4609	0.3787
Average Score	0.7285	0.7715	0.5861

Detailed Human Evaluation Result. We provide detailed human evaluation results across five dimensions for generated Office Actions and rebuttals, as shown in Table 13. Notably, we provide detailed inter-rater agreement among three human experts across five dimensions, as shown in Table 12, demonstrating strong consistency across these dimensions.

Table 13: Human evaluation results for OA-RO and Rebuttal generation. Scores are on a 1–10 scale across five dimensions: Soundness (Sound.), Clarity (Clar.), Constructiveness (Const.), Completeness (Comp.), and Language Style (Lang.).

Model	OA-RO Generation						Rebuttal Generation					
	Sound.	Clar.	Const.	Comp.	Lang.	Avg.	Sound.	Clar.	Const.	Comp.	Lang.	Avg.
Proprietary Models												
GPT-5-mini	2.42	4.50	2.69	2.69	5.65	3.59	**5.94**	6.75	**6.69**	**7.06**	**7.81**	**6.85**
Gemini-2.5-Flash	2.36	4.86	2.78	3.11	5.31	3.68	5.75	**7.21**	6.04	6.92	7.42	6.67
DeepSeek-V3.2	**2.71**	**5.71**	**3.32**	**3.37**	**6.55**	**4.33**	4.00	6.58	4.00	4.75	5.92	5.05
GPT-4o-mini	1.58	3.96	1.85	1.96	5.46	2.97	3.38	6.31	3.38	4.31	6.63	4.80
Open-Source Models												
LLaMA3.1-8B-it	1.55	3.77	1.45	1.73	4.55	2.61	3.92	5.67	4.17	5.33	6.58	5.13
Gemma3-12B-it	**1.79**	**4.68**	1.79	2.07	5.50	**3.16**	4.12	5.92	4.08	4.85	6.71	5.16
Gemma3-27B-it	1.71	4.21	**1.96**	**2.11**	**5.71**	3.14	**4.61**	**6.96**	**4.71**	**6.18**	**7.36**	**5.96**

Multi-Judge Robustness. To assess the robustness of LLM-as-a-judge evaluation, we re-evaluate all OA-RO outputs using `GPT-5.4` and `Claude-Opus-4.8`, in addition to the primary Gemini judge. All judges use the same evaluation criteria. Table 14 reports the detailed scores across judges.

Despite differences in absolute score calibration, the three judges show broadly consistent model-level performance. Table 15 further reports their agreement and alignment with human evaluation.

Table 14: OA-RO evaluation results from three LLM judges on a 1–10 scale.

Model	Gemini						GPT-5.4						Claude-Opus-4.8					
	Sou.	Cla.	Con.	Com.	Sty.	Avg.	Sou.	Cla.	Con.	Com.	Sty.	Avg.	Sou.	Cla.	Con.	Com.	Sty.	Avg.
GPT-5-mini	3.07	7.77	2.65	3.14	7.84	4.89	2.18	7.68	3.19	2.37	7.68	4.62	2.92	7.58	4.50	3.54	7.53	5.21
Gemini-2.5-Flash	2.60	7.21	1.79	2.60	7.60	4.36	2.24	6.81	2.22	1.99	6.79	4.01	2.82	6.80	3.03	2.86	6.51	4.40
DeepSeek-V3.2	2.39	7.43	1.86	2.61	7.42	4.34	2.42	7.14	2.53	2.47	7.14	4.34	3.02	7.50	3.76	3.82	7.28	5.07
GPT-4o-mini	2.05	6.32	1.45	2.02	6.21	3.61	2.33	6.19	2.15	1.56	6.18	3.68	2.45	6.04	2.65	2.59	5.74	3.89
LLaMA3.1-8B-it	1.78	5.26	1.24	1.98	5.09	3.07	2.23	5.08	1.39	1.69	5.07	3.09	2.05	4.12	1.50	2.15	4.04	2.77
Qwen3.5-9B	2.26	6.94	1.84	2.58	6.92	4.11	2.15	6.64	2.53	2.32	6.63	4.05	2.65	6.39	3.39	3.36	6.07	4.37
Gemma3-12B-it	1.98	6.34	1.36	2.00	6.27	3.59	2.31	6.20	2.15	1.60	6.20	3.69	2.35	5.93	2.65	2.51	5.62	3.81
Gemma3-27B-it	2.01	6.51	1.34	1.95	6.45	3.65	2.45	6.37	1.93	1.56	6.36	3.73	2.52	6.14	2.46	2.65	5.78	3.91
Qwen3.5-27B	2.56	7.19	2.09	2.83	7.18	4.37	2.42	6.81	2.74	2.38	6.80	4.23	2.80	6.74	3.83	3.52	6.48	4.67
LLaMA3.3-70B-it	1.91	6.08	1.28	1.80	5.95	3.40	2.46	6.19	1.89	1.33	6.18	3.61	2.42	5.72	2.45	2.29	5.46	3.67

Table 15: Agreement across LLM judges and with human evaluation.

Comparison	Kendall's τ	Spearman's ρ	Pearson's r
Gemini vs. GPT-5.4	0.822	0.927	0.969
Gemini vs. Claude	0.911	0.964	0.954
GPT-5.4 vs. Claude	0.911	0.976	0.992
Average (inter-judge)	0.881	0.956	0.972
3-Judge Ensemble vs. Human	0.619	0.786	0.845

Table 16: Analysis of potential same-family preference.

Model	**Own Judge**	**Gemini**	**GPT-5.4**	**Claude**	**Own–Others**
Gemini-2.5-Flash	Gemini	4.36	4.01	4.40	+0.15
GPT-5-mini	GPT-5.4	4.89	4.62	5.21	-0.43
GPT-4o-mini	GPT-5.4	3.61	3.68	3.89	-0.07

The three judges exhibit strong agreement, and their ensemble also aligns well with human evaluation. We further observe no systematic same-family score inflation, supporting the robustness of our LLM-as-a-judge evaluation.

D Detailed Case Study

To further investigate the LLMs' performance in patent examination, we analyze specific cases across different task settings. We provide detailed generated Office Action cases, including OA-RO (Table 18), OA-RS (Table 19) and OA-DP (Table 20) settings, as well as a generated rebuttal case (Table 17).

Across different settings, we observe consistent limitations in model performance on patent examination tasks. In the OA-RO setting, even with access to oracle references, models can identify relevant prior art but often fail to perform rigorous claim-to-art mapping, instead reducing complex technical distinctions to superficial similarities. In the OA-RS setting, although the correct reference may be present, models tend to incorporate irrelevant documents into combinations, leading to unsupported mappings and unstable novelty judgments. In the OA-DP setting, when external evidence is unavailable, models frequently produce incorrect rejections that are not grounded in the claim text or prior art. In contrast, in rebuttal generation, models show stronger performance by effectively responding to given rejections and constructing coherent counter-arguments. Overall, these results suggest that while external evidence improves citation accuracy, the core reasoning required for precise claim-to-prior-art analysis remains a major bottleneck, and current models perform better in reactive argumentation than in proactive examination.

Table 17: Example of generated rebuttal.

Task: *Rebuttal Generation*

Input:

- **Claim 1, 2, 3, ...** (1. A pharmaceutical composition comprising ...).
- **Oracle Art (Cited in OA):** Ref US11117934 (A pharmaceutical composition comprising:) a myxoma virus (MYXV) comprising a genetic modification at its M153 locus and a transgene encoding a cytokine or a cell matrix protein...) and ...
- **Reference OA:**DETAILED ACTION Notice of Pre-AIA or AIA Status The present application, filed on or after March 16, 2013, is being examined under the first inventor to file provisions of the AIA...

Output:

RESPONSE TO OFFICE ACTION
AMENDMENT AND REMARKS
I. Status of the Claims...
IV. Section 102 Rejections
The Examiner rejects claims 1-20 over claims 1-4, 8, 16-23, 26, 27 of U.S. Patent No. 11,117,934 (the "'934 Patent") under 35 U.S.C. § 102. Applicant respectfully disagrees. The '934 Patent broadly describes myxoma viruses (MYXV) having enhanced anti-cancer activity, wherein the MYXV is genetically engineered to attenuate an activity or expression level of its M153 protein, and to express a non-viral molecule, wherein the non-viral molecule is a cytokine or a cell matrix protein. While the '934 Patent discloses similar components, it fails to teach the specific genetic modification at the M153 locus and the combination of a disrupted M153 gene with a transgene encoding a cytokine or cell matrix protein as recited in the currently amended claim 1...
Conclusion...
REFINED CLAIMS...

E PROMPTS

Office Action Generation Prompts. As shown in Figure 8, we provide the detailed prompts for Office Action generation across three OA task settings.

Rebuttal Generation Prompts. As shown in Figure 9, we provide the detailed prompts of rebuttal generation task.

LLM-as-a-judge Evaluation Prompts. We provide the detailed LLM-as-a-judge evaluation prompts used for assessing generated Office Actions (Figure 10) and rebuttals (Figure 11).

F LIMITATIONS

While PatRe represents a meaningful step toward systematic evaluation of LLMs in patent examination and legal-technical reasoning, several limitations remain. PatRe is currently constructed from real-world USPTO prosecution records and thus primarily reflects the legal standards, examination procedures, and documentation practices of the U.S. patent system. The publicly accessible USPTO prosecution histories provide a reliable basis for systematically reconstructing multi-turn examiner–applicant interactions. Extending PatRe across jurisdictions involves more than incorporating additional data, as patent systems differ not only in their legal standards, examination procedures, and documentation practices, but also in the accessibility and organization of prosecution records. In some jurisdictions, examination materials may be subject to different access conditions or require case-specific retrieval procedures, making large-scale reconstruction of comparable multi-turn prosecution trajectories more challenging. Developing multilingual and cross-jurisdictional extensions while preserving a comparable multi-turn evaluation framework is therefore an important direction for future work.

Additionally, PatRe currently focuses on textual modalities and does not explicitly incorporate patent drawings or figures, which may contain important technical information not fully captured in textual descriptions. Extending the benchmark to incorporate multimodal information represents another important direction for future work.

Table 18: Example of generated Office Action under Reference-Oracle Generation (OA-RO) setting.

Task: *Reference-Oracle Generation* (OA-RO)

[Initial] Input:

- **CLAIMS:**
  WHAT IS CLAIMED IS: 1. A system for just-in-time (JIT) game development and gameplay, the system comprising:a game engine: an oracle comprising one or more Al models [... Claims 2-30 omitted for brevity ...]
- **Oracle Art:**
  Ref US20190362589: 1. A system for coincident gambling hybrid gaming having coincident gambling events[...]
  Ref US20240390801A1: A game server for providing updates related to an online game [... Other pertinent references omitted ...]
- **Task summary:**
  Act as a USPTO examiner. Given current claims and an oracle reference set, evaluate statutory rejections including but not limited to 35 U.S.C. 101/102/103/112 and Double Patenting (DP). Select pertinent references, perform element-by-element mapping, and justify reference selection. Adhere to MPEP standards and avoid disclaimers.

[Initial] Output:

DETAILED ACTION
OPENING STATEMENT
The present application is examined for compliance with 35 U.S.C. § 101, 102, 103, and 112. It is determined that claims 1-30 are not allowable in their present form.
CLAIM REJECTIONS - 35 U.S.C. 103
Claims 1-30 are rejected under 35 U.S.C. § 103 as being unpatentable over [Ref US20190362589] in combination with [Ref US20240390801A1].
Ref US20190362589 discloses a system for coincident gambling hybrid gaming having coincident gambling events and game events provided on a computing device. Ref US20240390801A1 discloses a game server for providing updates related to an online game, wherein the game server is configured to sandbox potentially malicious event-driven code.[...]
Specifically, the following limitations are met: [... Detailed element-by-element mapping for Claims 1-30 omitted for brevity ...]
CONCLUSION: Claims 1-30 are rejected as being unpatentable under 35 U.S.C. § 103 over Ref US20190362589 in combination with Ref US20240390801A1. The applicant could respond to these rejections with arguments or amendments that overcome the references.

[Intermediate] Input:

- **CLAIMS:**
  1. (Currently Amended) A system for just-in-time (JIT) game development and gameplay, the system comprising:a game engine; oracle comprising one or more Al models [... Claims 2-30 omitted for brevity ...]
- **Oracle Art:**
  Ref US20240390801A1: A game server for providing updates related to an online game, wherein the game server is configured to sandbox potentially malicious event-driven code [... Other pertinent references omitted for brevity ...]
- **rebuttal:**
  REMARKS In response to the non-final Office Action dated March 26, 2025, Applicant amends claims 1-7, 11, 19-27, 29, and 30 [...]
- **Examination History:**
  Previous Round Office Action: The present application, filed on or after March 16, 2013, is being examined under the first inventor to file provisions of the AIA [...]
- **Task summary:**
  Act as a USPTO examiner to conduct a formal re-examination by evaluating amended claims and applicant rebuttals against the established prosecution history. The process requires a critical selection of prior art to address specific traversals, applying Strict Statute Gating to determine whether to maintain or withdraw previous rejections across all applicable statutory grounds. Each argument must be assessed as Persuasive or Unpersuasive with technical reasoning grounded in a rigid Scope Lock, ensuring the final report acknowledges claim amendments, identifies any new matter violations, and presents updated rejections within a realistic, structured USPTO-style action.

[Intermediate] Output:

DETAILED ACTION
The application is currently pending. Amendments have been made to claims 1-7, 11, 19-27, 29, and 30. Claims 1 and 26 are in independent form.
RESPONSE TO ARGUMENTS
The applicant's arguments regarding 35 U.S.C. § 112 rejections for claims 4 and 7 are persuasive. [...] The applicant's arguments regarding 35 U.S.C. §§ 102 and 103 rejections are unpersuasive. While the applicant attempts to distinguish the claimed invention from Ignatchenko, the asserted reference, in combination with Arnone, renders the claims unpatentable. The applicant's attempts to characterize the claimed features as "notorious well-known" do not negate the teachings of the cited references.
WITHDRAWN REJECTIONS
Claims 4 and 7 – 35 U.S.C. 112
CLAIM REJECTIONS - 35 U.S.C. 102
Claims 1-10, 23-28, and 30 are rejected under 35 U.S.C. 102(a)(1) as being anticipated by Ignatchenko (US 20240390801 A1). [...] The claimed features are all disclosed within Ignatchenko.
CLAIM REJECTIONS - 35 U.S.C. 103
Claims 11-13, 21, 22, and 29 are rejected under 35 U.S.C. 103 as being unpatentable over Ignatchenko (US 20240390801 A1) in combination with Arnone (US 20190362589 A1). [...] It would have been obvious to one of ordinary skill in the art to combine these references to achieve the claimed invention.
[... Rejections for Claims 14-20 and formal Sections 112, Double Patenting omitted for brevity ...]
CONCLUSION
Claims 1-10, 11-13, 14-20, 21, 22, 23-28, 29, and 30 are currently rejected under 35 U.S.C. 102 and 103. Claims 1 and 26 are in condition for allowance, pending the applicant's response to the rejections.

Table 19: Example of generated Office Action under Retrieval-Simulated Generation (OA-RS) setting.

Task: *Retrieval-Simulated Generation* (OA-RS)

[Initial] Input:

- **CLAIMS:**
  1. A system for at least partially filling an orchestra pit or extending a stage, comprising [... Claims 2-12 omitted for brevity ...]
- **Candidate Pool:**
  – **Oracle Art: Ref US4332116**: 1. A building structure formed from prefabricated building components, the plan view of which is a polygon havinga central support column [... Other pertinent references omitted for brevity ...]
  – **BM25 Set: Ref US8117787B2**: A length-adjustable construction support, comprising[...] [... Other pertinent references omitted for brevity ...]
- **Task summary:**
  Act as a USPTO examiner. Given claims and a noisy pool of 10 candidates (BM25 + Oracle), evaluate patentability across various statutory grounds (e.g., 102, 103, 112). Distinguish pertinent prior art from irrelevant noise. Ground all technical reasoning in the provided inputs and generate a formal Office Action report.

[Initial] Output:

DETAILED ACTION
OPENING STATEMENT
This is an examination of claims 1-20 of your application. A review of the claims and prior art has revealed certain objections. The following analysis details these objections and provides a basis for rejection.
CLAIM REJECTIONS - 35 U.S.C. 103
Claims 1-20 are rejected under 35 U.S.C. § 103 as unpatentable over [US4332116] in combination with [US8117787B2]. US4332116 discloses a building structure formed from prefabricated building components, including a frame assembly with a leg.[...] US8117787B2 discloses a length-adjustable construction support comprising a first pipe, a second pipe, and an inner stopper coupled to the outer circumference of the second pipe. [...]
[... Detailed element-by-element mapping for Claims 1-20 omitted for brevity ...]
CONCLUSION
Claims 1-20 are rejected as unpatentable under 35 U.S.C. § 103 and 35 U.S.C. § 112(a). Applicant is invited to amend the claims to overcome these rejections. A reasoned explanation of any amendments should be provided.

[Intermediate] Input:

- **CLAIMS:**
  1. This listing of claims will replace all prior versions, and listings, of claims in the application [...]
- **Candidate Pool:**
  – **Oracle Art: Ref US11117934**: (A myxoma virus (MYXV) having enhanced anti-cancer activity, wherein the MYXV is genetically engineered to attenuate an activity or expression level of its M153 proteinn[...] [... Other pertinent references omitted for brevity ...]
  – **BM25 Set: Ref US20180042847A1**: A method of treating and/or reducing the incidence of chronic obstructive pulmonary disease (COPD)n[...] [... Other pertinent references omitted for brevity ...]
- **rebuttal:**REMARKS Claims 1-20 are currently pending and under consideration[...]
- **Examination History:**revious Round Office Action: DETAILED ACTION Notice of Pre-AIA or AIA Status The present application[...]

  Task summary: Act as a USPTO examiner to conduct a formal re-examination by navigating a high-entropy candidate pool containing both BM25-retrieved noise and Oracle-level references. The process requires the model to autonomously evaluate and select the most pertinent prior art to address amended claims and applicant rebuttals, applying Active Statute Gating to determine whether to maintain or withdraw previous rejections across all applicable statutory grounds. Each argument must be assessed as Persuasive or Unpersuasive through evidence-first mapping and a strict Scope Lock on provided technical summaries, ensuring the final report acknowledges amendments, identifies potential new matter violations, and presents updated rejections within a structured, realistic USPTO-style Office Action.

[Intermediate] Output:

DETAILED ACTION
The present application is being examined under the first inventor to file provisions of the AIA.
RESPONSE TO ARGUMENTS
The applicant argues that the double patenting rejection has been overcome by filing a Terminal Disclaimer. This argument is Persuasive as the Terminal Disclaimer is properly filed and the reference patent, US11117934, is commonly owned.
No rejection under 35 U.S.C. 102 is made at this time. However, a rejection under 35 U.S.C. 103 is made as the claims are obvious over [Ref US11117934]. The reference teaches a MYXV with a genetic modification at its M153 locus [...] similar to the claims of the present application. [... Detailed reference selection and evidence-first mapping omitted for brevity ...] The nonstatutory double patenting rejection is based on a judicially created doctrine grounded in public policy so as to prevent the unjustified or improper timewise extension of the “right to exclude” granted by a patent and to prevent possible harassment by multiple assignees...
CONCLUSION
The nonstatutory double patenting rejection is withdrawn, and the application is still subject to examination under 35 U.S.C. 102 and 103. A rejection under 35 U.S.C. 103 is made.

Table 20: Example of generated Office Action under Direct-Prompting Generation (OA-DP) setting.

Task: *Direct Prompting Generation* (OA-DP)

[Initial] Input:

- **CLAIMS**:
  1: An extendable support leg comprising:a top leg tube (1), a middle leg tube (2) and an end leg tube (3) arranged axially telescopically and sequentially from outside to inside [...]
- **Task summary:** Act as a USPTO examiner. Given current claims without any prior art documents, evaluate statutory rejections (e.g., 101, 102, 103, 112, or DP). Rely solely on your internal technical knowledge and common sense available to a PHOSITA. State technical reasoning before deciding on allowance or rejection.

[Initial] Output:

DETAILED ACTION
OPENING STATEMENT
The present examination has been completed with respect to claims 1-10. The claims are examined for compliance with 35 U.S.C. § 112(a) and (b).
[CLAIM REJECTIONS - 35 U.S.C. 112(b)]
Claims 1-10 are rejected under 35 U.S.C. § 112(b) as being indefinite. The specification does not provide sufficient support to enable a person having ordinary skill in the art to make and use the claimed invention. Specifically, the description of the first clamping assembly in claim 10 lacks clarity and precision. The relationship between the various components (drive tube, upper locking piece, locking body, and expansion member) and their function in achieving clamping is not adequately explained. The description of the expansion member's behavior, transitioning between a tightened and loosen state based on the movement of the upper locking piece, is vague and lacks sufficient detail to allow for practical implementation. The specification does not provide a clear understanding of how the expansion member interacts with the inner wall of the top leg tube, nor does it provide sufficient detail regarding the forces and mechanisms involved...
CONCLUSION
Claims 1-10 are currently rejected under 35 U.S.C. § 112(b) for indefiniteness. The applicant is invited to amend the claims to provide a clear and concise description of the invention, enabling a person having ordinary skill in the art to make and use the claimed invention without undue experimentation.

[Intermediate] Input:

- **CLAIMS**:
  (AMENDMENTS TO THE CLAIMS: The following listing of claims will replace all prior versions and listings of claims in this application ...).
- **rebuttal:**REMARKS: Favorable reconsideration of this application is respectfully requested in view of the foregoing amendments and the following remarks...
- **Examination History:**Previous Round Office Action: DETAILED ACTION Notice of Pre-AIA or AIA Status The present application, filed on or after March 16, 2013...
- **Task summary:** Act as a USPTO examiner to conduct a formal re-examination by evaluating amended claims and applicant rebuttals relying solely on internal technical knowledge and PHOSITA common sense in the absence of external prior art. The process requires applying Active Statute Gating to maintain or withdraw existing rejections across all applicable statutory grounds, assessing each argument as Persuasive or Unpersuasive through an Evidence-First Logic derived from technical principles. Adhering to a strict Scope Lock, the model must acknowledge amendments, identify any new matter violations, and present updated rejections—specifically focusing on inherent obviousness or indefiniteness—within a structured, realistic USPTO-style Office Action.

[Intermediate] Output:

DETAILED ACTION
This is a first Office action in response to the applicant's amendment filed February 12, 2025. Currently, claims 1, 2, 4-10 are pending and examined below. Claim 3 has been canceled.
The applicant's arguments in response to the prior rejections under 35 U.S.C. § 112(b) and § 102 are addressed.
Regarding the objection to the specification (title), the applicant has amended the title. The objection is withdrawn.
Regarding the rejection of claims 1-9 under 35 U.S.C. § 102(a)(1) as anticipated by Baek, the applicant presents two primary arguments distinguishing amended claim 1.
First, the applicant argues that Baek does not disclose "a thickness of the pad from an upper surface of the first interconnect structure to the upper surface of the pad is greater than or equal to a thickness of the encapsulating resin layer from the upper surface of the first interconnect structure to an upper surface of the encapsulating resin layer." The applicant asserts that in Baek, the thickness of the redistribution layer (pad) 112c is smaller than the thickness of the encapsulant 130. The examiner's review of the annotated Fig. 14 of Baek, as cited in the prior Office action, supports the applicant's assertion. The pad 112c is shown embedded within the insulating layer 111a, and its exposed upper surface appears to be at or below the level of the upper surface of the insulating layer 111a. The encapsulant 130 is shown covering the side of the structure, with its upper surface significantly above the upper surface of the pad 112c. Therefore, the thickness of the pad from the upper surface of the first interconnect structure (which includes insulating layer 111a) to the pad's upper surface is less than the thickness of the encapsulant layer measured from that same reference plane. This does not meet the "greater than or equal to" limitation. The applicant's argument on this point is persuasive...
CONCLUSION
Claims 1, 2, 4-9 are allowable subject to the withdrawal of the prior rejections. Claim 10 is rejected under 35 U.S.C. § 112(b). A shortened statutory period for reply to this Office action is set to THREE MONTHS.

Office Action Generation Prompt

[System Message] You are a professional USPTO patent examiner. Your tone is formal, objective, and authoritative. You strictly adhere to 35 U.S.C. standards. You must not refuse or mention limitations about external databases or access; use ONLY the provided inputs and make best-effort determinations. If information is missing, state reasonable assumptions and proceed.
[User Message] Perform an initial examination of claims 1-{claim_count}.
Input: Current Claims: {current_claims}
Instructions:
1. ------------ Different OA settings ------------
[Directly Prompting (OA-DP)] **Evidence-First Logic (Intrinsic Knowledge):** No prior art documents are provided. Evaluate the claims based on your internal technical knowledge base and common sense available to a PHOSITA. State your technical reasoning **before** deciding allowance or rejection.
[Reference-Oracle Generation (OA-RO)] **Reference Selection & Evidence Logic (Critical):** The input contains references that were cited in the actual examination. You must **not** blindly use all of them. **First**, evaluate the provided references and select the most pertinent ones to support your specific rejection or allowance logic for each claim.
[Retrieval-Simulated Generation (OA-RS)] **Reference Selection Logic (Crucial):** The input contains a mixed pool of 10 candidates (BM25 5 + Oracle 5, deduplicated; Oracle fills if fewer). **First**, evaluate the candidate pool and select the most relevant references to address the claims. Briefly explain why the selected references are the most pertinent.
1. ------------ Different OA settings ------------
2. **Disposition (No Bias):** Decide whether the claims should be **Allowed** or **Rejected** based strictly on the claim language and your technical reasoning. Do not assume a rejection merely because you are an examiner. **Active Statute Gating (Strict):** Before drafting, determine the set of statutes you will actually apply (e.g., 101, 102, 103, 112, double patenting). You must not include headings for statutes you do not substantively apply. If no rejection applies, omit all rejection headings.
3. **Updated Rejections:** Provide analysis for any maintained or new rejections (primarily 112 or inherent obviousness) based on the claim limitations. If no rejection applies, explain why allowance is warranted. **Scope Lock:** All technical content must be grounded in the provided inputs. Do not introduce unrelated technical topics, inventions, or references.
4. **No-Citation Rule:** Do not cite any specific prior art publications or formal reference numbers.
5. **Structure (Real USPTO Style):** Use natural USPTO-style paragraphs with fixed headings in plain text. Include only applicable sections: DETAILED ACTION, OPENING STATEMENT, CLAIM REJECTIONS - 35 U.S.C. 101, CLAIM REJECTIONS - 35 U.S.C. 102, CLAIM REJECTIONS - 35 U.S.C. 103, CLAIM REJECTIONS - 35 U.S.C. 112(a)/(b), DOUBLE PATENTING and CONCLUSION.
6. **Realistic Phrasing:** Use standard Office Action phrasing (e.g., "Claims 1-10 are rejected under 35 U.S.C. 112(b)...").
7. **Decision Logic:** Initial Office Actions are generally Non-Final unless clearly warranted; do not label Final unless criteria are met.
8. **Conclusion:** Summarize claim status and set expectations for continued examination.
Output Format: Start with the heading **DETAILED ACTION** and present the Office Action in a realistic style. Do not append JSON.

Figure 8: Prompt for the Office Action generation across three settings.

Rebuttal Generation Prompt

[System Message] You are a senior patent attorney. Your goal is to draft a persuasive response to a USPTO Office Action while ensuring strict adherence to MPEP standards and formatting rules. You must not refuse or mention limitations about external databases or access; use ONLY the provided inputs and make best-effort determinations. If information is missing, state reasonable assumptions and proceed.
[User Message] Draft a formal response to the following Office Action.
Input: Current Claims: {current_claims}; Office Action: {office_action}; Reference Claims: {reference_summaries}; Prosecution History (prior rounds): {prosecution_history}.
Instructions:
1. **Realistic PTO Structure (mirror real replies):** Use the following structure: RESPONSE TO OFFICE ACTION; AMENDMENT AND REMARKS; I. Status of the Claims (identify canceled/amended/pending and independent claims); II. Objection to the Specification (only if raised); III. Objection to the Claims (only if raised); IV. Section 102 Rejections (only if raised); V. Section 103 Rejections (only if raised); VI. Section 112 Rejections (only if raised); Conclusion; REFINED CLAIMS.
2. **Strict Statute Gating:** Only address statutes explicitly raised in the Office Action. Do not include 101/102/103/112 sections unless the Office Action states them. If a statute is not raised, omit that section entirely.
3. **Date Handling:** If the Office Action includes a date, it must be used. If no date is available, the placeholder "[insert date]" may be retained. **Short-Form Response for Limited OA:** If the Office Action is a Notice of Allowance or raises only double patenting, keep the response brief and limited to those issues. Do not introduce additional statutory sections. **Use OA-Specific Strategy:** If the Office Action identifies allowable subject matter in dependent claims, incorporate those limitations into the relevant independent claims and explicitly argue that the amended independent claims are now allowable. If the Office Action raises an informality (e.g., typographical error), acknowledge and correct it.
4. **Point-by-Point Response:** For each rejection group, respond in the same order as presented in the Office Action and identify the claim numbers and cited references. Each argument must be tied to specific claim language and the Office Action's reasoning. Avoid boilerplate responses.
5. **Statute-Specific Arguments: 35 U.S.C. §102 (only if raised):** Traverse each rejection group and explain why the claims are not anticipated, with explicit reference to claim language and cited portions. **35 U.S.C. §103 (only if raised):** Address deficiencies such as lack of motivation to combine, teaching away, or technical incompatibility, tied to claim language. **35 U.S.C. §112 (only if raised):** Address indefiniteness or enablement issues with arguments grounded in the specification.
6. **Amendment Compliance:** If claims are amended, use standard USPTO markings (underlining for additions and strikethrough for deletions). Each claim must include a status identifier such as (Original), (Currently Amended), (Withdrawn), or (Canceled).
7. **No-New-Matter Rule:** Ensure all amendments are fully supported by the original specification. Do not introduce new matter.
Output Format: Start with the heading **RESPONSE TO OFFICE ACTION** and present the full attorney rebuttal in a realistic style. Do not append JSON.

Figure 9: Prompt for the rebuttal generation.

LLM-as-a-judge Evaluation Prompt for Generated Office Action

[System Message] You are a Senior Patent Attorney auditing an AI-generated Office Action against a gold-standard Ground Truth. Your role is not to re-examine the case, but to verify the legal and technical integrity of the generated text.
Audit Mandate:
1. **Fact-Check Against Ground Truth:** The provided Ground Truth (disposition, statutes, and citations) is the absolute reference. Any contradiction between the Generated Office Action and the Ground Truth constitutes a critical failure.
2. **Detect "Right Answer, Wrong Reason":** Even if the overall disposition aligns with the Ground Truth, strictly penalize the response if it relies on hallucinated features or incorrect prior art mappings to reach that conclusion.
3. **Evidence-Based Scrutiny:** Compare the claim-to-prior-art mapping in the Generated Office Action against the provided Reference Summaries. Ensure that every technical assertion is directly supported by the reference text.
Hard Constraints:
1. Use only the provided inputs; do not reference external databases or limitations of access.
2. Do not apologize, request additional data, or include disclaimers.
3. Maintain strict scrutiny for hallucinated technical details; if a feature is not present in the reference summary, it must be treated as nonexistent.
4. Always produce the required professional report and structured JSON output.
[User Message] Evaluate the generated Office Action based on the provided context.
Context: Ground Truth OA: {ground_truth}; {response_target_section}; Generated Office Action: {generated_text}.
Core Rules:
1. Use only the provided text. Do not assume facts not stated in Ground Truth or Reference Summaries.
2. If the disposition label (Non-Final / Final / Allowance) does not match Ground Truth, cap soundness at 3.
3. Figure/Drawing Exception: If a discrepancy is solely due to figures or drawings not provided in text, do not deduct points.
4. Do not apologize or refuse. Provide determinate scores with evidence.
Evaluation Criteria (1–10 unless noted):
1. **soundness:** Conclusion correctness and reasoning alignment. Must be 0–3 if the disposition is incorrect... (Omit)
2. **clarity:** Professionalism and readability.
3. **completeness:** Coverage of claims and rejection grounds.
4. **constructiveness:** Quality of actionable guidance.
Metric-Specific Scoring Guidelines:
soundness: 10: Full match with Ground Truth; 7–9: correct disposition with minor issues; 4–6: partial reasoning mismatch; 1–3: incorrect disposition and/or major reasoning flaws... (Omit)
clarity: 10: highly professional and clear; 7–9: generally clear; 4–6: some ambiguity; 1–3: unclear or incoherent.
completeness: 10: full coverage; 7–9: minor omissions; 4–6: partial coverage; 1–3: major omissions.
constructiveness: 10: highly actionable guidance; 7–9: mostly actionable; 4–6: generic suggestions; 1–3: no useful guidance.
Output Format:
Return a professional audit report followed by a JSON block (Omit):

Figure 10: Prompt for LLM-as-a-judge evaluation of generated Office Actions.

LLM-as-a-judge Evaluation Prompt for Generated Rebuttal

[System Message] You are a Patent Law Partner evaluating an associate's rebuttal. You must not refuse or mention limitations about external databases or access; use ONLY the provided inputs and make best-effort determinations. If information is missing, state reasonable assumptions and proceed.
[User Message] Evaluate the generated rebuttal based on the provided context.
Context: Office Action: {office_action}; Ground Truth Rebuttal: {ground_truth}; Generated Rebuttal: {generated_text}
Core Rules:
1. Use only the provided text; do not assume unstated facts.
2. Figure/Drawing Exception: discrepancies due solely to missing figures should not be penalized.
3. Avoid boilerplate; require point-specific reasoning.
Evaluation Criteria (1-10):
effectiveness: strength of counter-arguments against 102/103 rejections; 10: directly and convincingly rebuts all theories; 7-9: strong with minor gaps; 4-6: partially responsive; 1-3: weak or off-point.
soundness: specificity to OA points; 10: fully claim-specific and detailed; 7-9: mostly specific; 4-6: partially generic; 1-3: largely boilerplate or non-responsive.
reliability: factual grounding in specification/prior art; 10: fully supported; 7-9: minor uncertainty; 4-6: some unsupported points; 1-3: significant hallucination or new matter.
statute_compliance: correctness of legal strategy under 101/102/103/112; 10: precise and fully compliant; 7-9: minor issues; 4-6: some misapplication; 1-3: incorrect or irrelevant statute use.
feature_distinction: clarity and precision of distinguishing features; 10: precise and claim-grounded; 7-9: mostly clear; 4-6: vague; 1-3: absent or incorrect.
specification_support: whether amendments are supported by the original disclosure; 10: fully supported; 7-9: minor gaps; 4-6: partial support; 1-3: new matter introduced.
clarity: legal writing quality and structure; 10: highly professional; 7-9: clear; 4-6: uneven; 1-3: unclear or disorganized.
completeness: coverage of all OA rejections; 10: fully covered; 7-9: minor omissions; 4-6: partial coverage; 1-3: major gaps.
point_wise_coverage: extent of addressing atomic OA points; 10: nearly all addressed; 7-9: strong coverage; 4-6: partial; 1-3: minimal.
Output Format:

```
{
 "total_score": score,
 "scores": {
  ...
 },
 "justification": "...",
 "justification_details": {
  "point_coverage": "...",
  "amendment_compliance": "...",
  "legal_strategy": "..."
 }
}
```

Figure 11: Prompt for LLM-as-a-judge evaluation of generated rebuttal.