PILOT-Bench: A Benchmark for Legal Reasoning in the Patent Domain with IRAC-Aligned Classification Tasks

Yehoon Jang[1*] **Chaewon Lee**[1*] **Hyun-seok Min**[2] **Sungchul Choi**[1†]
[1]Major in Industrial Data Science & Engineering,
Department of Industrial and Data Engineering, Pukyong National University
[2]Tomocube Inc.
{jangyh0420, oochaewon}@pukyong.ac.kr, min6284@gmail.com, sc82.choi@pknu.ac.kr

Abstract

The Patent Trial and Appeal Board (PTAB) of the USPTO adjudicates thousands of *ex parte* appeals each year, requiring the integration of technical understanding and legal reasoning. While large language models (LLMs) are increasingly applied in patent and legal practice, their use has remained limited to lightweight tasks, with no established means of systematically evaluating their capacity for structured legal reasoning in the patent domain. To address this gap, we introduce **PILOT-Bench** (**P**atent **I**nva**L**idati**O**n **T**rial Benchmark), a dataset and benchmark that aligns PTAB decisions with USPTO patent data at the case-level and formalizes three IRAC-aligned classification tasks: Issue Type, Board Authorities, and Subdecision. We evaluate a diverse set of close-source(commercial) and open-source LLMs and conduct analyses across multiple perspectives, including input-variation settings, model families, and error tendencies. Notably, on the Issue Type task, closed-source(commercial) models consistently exceed 0.75 in Micro-F1 score, whereas the strongest open-source model (Qwen-8B) achieves performance around 0.56, highlighting the substantial gap in reasoning capabilities. PILOT-Bench establishes a foundation for the systematic evaluation of patent-domain legal reasoning and points toward future directions for improving LLMs through dataset design and model alignment. All data, code, and benchmark resources are available at https://github.com/TeamLab/pilot-bench.

1 Introduction

As the volume of patent applications and examinations continues to grow, the Patent Trial and Appeal Board (PTAB) of the US Patent and Trademark Office (USPTO) handles a substantial number of appeals and invalidation proceedings each year (USPTO, 2025). The *ex parte* appeal, which challenges the rejection of an examiner, requires a precise interpretation of patent—such as claims and prior art—and legal reasoning to identify and apply the relevant provisions of 35 U.S.C. and 37 C.F.R. to reach a conclusion.

Large language models (LLMs) are increasingly used in patent and legal practice to reduce repetitive reading tasks (USPTO, 2024; Simmons, 2024; Wang et al., 2024; Makover and Boynes, 2025). However, their adoption remains largely limited to such lightweight tasks, while *ex parte* appeals demand deep reasoning—issue identification, rule mapping, rule application, and conclusion determination—that go well beyond them. Furthermore, the lack of a systematic public dataset or benchmark hinders quantitative assessment of whether LLMs possess the technical understanding and legal reasoning required in PTAB invalidity review. As a result, using LLMs for these tasks remains challenging.

In this paper, we propose the **P**atent **I**nva**L**idati**O**n **T**rial Benchmark (PILOT-Bench), a dataset and benchmark for evaluating the legal reasoning abilities of LLMs in the patent domain. We combine PTAB decisions with USPTO data per case and construct classification tasks aligned with the Issue–Rule–Application–Conclusion (IRAC) framework commonly used in legal practice. Our contributions are threefold:

- **PILOT-Bench dataset & benchmark.** PILOT-Bench is, to our knowledge, the first *benchmark* that integrates 18K PTAB *ex parte* appeals with USPTO patent text at the case-level and provides 15K opinion-split instances explicitly engineered to prevent label leakage.
- **IRAC-aligned tasks.** We design three classification tasks; Issue Type(5 labels, multi-label),

* Equal contribution.
† Corresponding author.

Proceedings of the Natural Legal Language Processing Workshop 2025, pages 240–280
November 8, 2025 ©2025 Association for Computational Linguistics

Board Authorities(9 labels, multi-label), Subdecision(23 fine/6 coarse grained labels, multi-class), directly aligned with the IRAC framework to measure patent-domain legal reasoning.

- **Empirical evaluation.** We conduct input variation experiments to assess the respective contributions of role segmentation and claim-text augmentation across multiple LLMs.

PILOT-Bench establishes a benchmark for evaluating LLMs' legal reasoning in the patent domain—specifically, PTAB *ex parte* appeals where technical understanding and legal reasoning meet. Our objective is to open a durable, reusable point of comparison that can anchor subsequent model, data, and methodology work and, ultimately, support responsible use of LLMs in patent practice. Accordingly, we fix the evidence boundary via the Opinion Split: inputs contain only `appellant_arguments` and `examiner_findings`, with all `ptab_opinion` text excluded. We keep the label schema fixed across Issue Type, Board Authorities, and Subdecision (fine/coarse) and evaluate under a unified zero-shot protocol with task-appropriate metrics (Exact Match/Macro-F1/Micro-F1 for multi-label; Accuracy/Macro-F1/Weighted-F1 for multi-class). We also report results for both closed-source(commercial) and open-source model families and for the Split (Base), Merge, and Split+Claim input-variation settings, providing reference baselines for subsequent work.

2 Preliminaries

2.1 PTAB *ex parte* Appeal

The PTAB *ex parte* appeal process is initiated after a final rejection by a patent examiner. The appellant submits an Appeal Brief, followed by an Examiner's Answer and, optionally, a Reply Brief. The Board then issues a written decision. PTAB decisions are conventionally organized into sections such as the *Statement of the Case*, outlining the procedural and factual background, and the *Analysis*, presenting the legal reasoning. The concluding portion records the outcome at the claim or case-level and cites the statutory or regulatory authorities (e.g., 35 U.S.C., 37 C.F.R.) that ground the ruling. In this way, PTAB decisions closely reflect the flow of legal reasoning.

Dataset / Study	Patent	Legal	LLM
Patent			
WIPO-alpha	✓	✗	✗
CLEF-IP	✓	✗	✗
USPTO-2M	✓	✗	✗
BIGPATENT	✓	✗	✗
HUPD	✓	✗	✓
IMPACT	✓	✗	✓
Patent-CR	✓	✗	✓
Legal			
LegalBench	✗	✓	✓
LexGLUE	✗	✓	✗
CaseHOLD	✗	✓	✗
CUAD / LEDGAR[1]	✗	✗	✗
Pile of Law	✗	✓	✗
MultiLegalPile	✗	✓	✗
PTAB studies			
Winer (2017)	✓	✓	✗
Rajshekhar (2017)	✓	✗	✗
Love (2019)	✓	✓	✗
Garcia (2022)	✓	✓	✗
Sokhansanj & Rosen (2022)	✓	✓	✗
Fu (2021)	✓	✗	✗
PILOT-Bench	✓	✓	✓

Table 1: Comparison by three criteria: (1) patent tasks, (2) legal/adjudicatory tasks, (3) ability to evaluate LLM in the patent/legal domain. Legal/adjudicatory tasks denote tasks leveraging statutory/regulatory mappings and decision structure. PTAB entries are research studies (not reusable corpora).

2.2 IRAC Framework

In PTAB *ex parte* appeals, IRAC maps naturally onto the decision flow: Issue identifies the contested statutory grounds; Rule maps those issues to the governing legal provisions; Application weighs the parties' arguments and facts against those provisions; and Conclusion renders the Board's ruling. We operationalize Issue, Rule, and Conclusion as three classification tasks and leave Application to future, generation-based work.

Our benchmark translates three of these IRAC stages—Issue, Rule, and Conclusion—into three concrete classification tasks to evaluate LLMs' capacity for patent-domain legal reasoning.

3 Related Work

3.1 Patent Corpora/Benchmarks

Public patent corpora have largely been constructed around technical-text tasks such as summariza-

[1]CUAD/LEDGAR focus on contract clause extraction/-classification; they are not decision/holding–centric and do not map statutes/regulations, hence marked ✗ under Legal/adjudicatory.

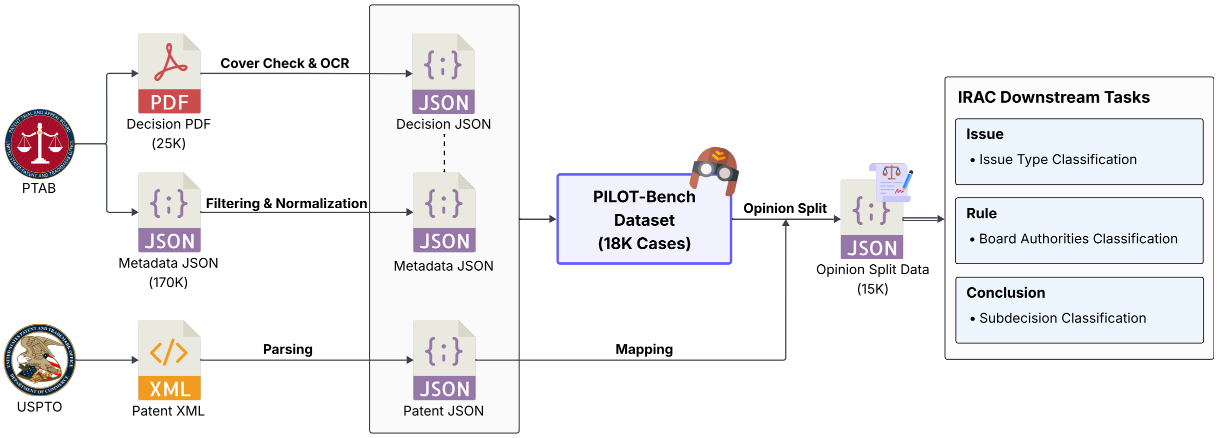


Figure 1: PILOT-Bench: Data sources, processing pipeline, and tasks. PTAB metadata JSONs and decision JSONs are aligned with USPTO patent JSONs to form PILOT-Bench (18K). From this base, we map each case to the appellant's patent and apply an LLM opinion split, yielding the 15K Opinion Split Data used for IRAC-aligned classification tasks.

tion and classification. WIPO-alpha (Fall et al., 2003), CLEF-IP (Piroi, 2010; Piroi et al., 2011), and USPTO-2M (Li et al., 2018) provide patent full text together with bibliographic metadata and introduce evaluation setups for IPC/CPC classification and prior-art retrieval research. BIGPATENT (Sharma et al., 2019) releases roughly 1.3 million description–abstract pairs and establishes a long-document summarization benchmark. HUPD (Suzgun et al., 2022) links patent documents filings from 2004–2018 with metadata, enabling multiple tasks including classification and binary decision prediction. More recently, IMPACT (Shomee et al., 2024) introduces a multimodal dataset by combining design images with patent information, while Patent-CR (Jiang et al., 2024) expands the scope of patent datasets by defining a claim-centric corpus for claim-revision tasks.

3.2 Legal Corpora/Benchmarks

LegalBench (Guha et al., 2023) covers legal reasoning broadly with 162 tasks and defines IRAC-stage tasks. LexGLUE (Chalkidis et al., 2022) is a multi-task legal NLU benchmark that offers evaluation setups for case classification, topic classification, and clause identification in contracts. CUAD (Hendrycks et al., 2021) and LEDGAR (Tuggener et al., 2020) construct clause extraction and classification tasks from contracts. CaseHOLD (Zheng et al., 2021) targets holding identification within judicial opinions. Pile of Law (Henderson et al., 2022) and MultiLegalPile (Niklaus et al., 2024) offer large-scale pretraining corpora aggregating diverse legal subdomains.

3.3 PTAB Studies

Prior PTAB prediction and analysis studies can be organized by procedure type and input modality. Winer (2017) targets Post-Grant Review (PGR) disputes and uses SVM and random forests to predict institution and invalidation outcomes. Rajshekhar et al. (2017) works in *Ex Parte* Reexamination (EPR), performing prior-art retrieval from the abstract, the first claim, and the title. Love et al. (2019) studies Inter Partes Review (IPR), predicting institution from metadata such as the number of unique words in the first independent claim and specification length. Garcia et al. (2022) combines claims with rejection grounds and classifies PTAB final decisions using BERT. Sokhansanj and Rosen (2022) uses the Patent Owner Preliminary Response (POPR) and decision text as inputs and applies XGBoost and a CNN-Attention model to predict IPR institution. Fu (2021) leverages IPR institution and final outcomes to estimate firm-level patent performance measures.

Limitations across Domains. Taken together, these studies reveal persistent gaps across patent, legal, and PTAB corpora. Patent benchmarks remain confined to technical-text problems such as summarization, classification, and retrieval, without capturing legal reasoning grounded in statutory authorities or decision structure. Legal corpora address reasoning tasks broadly, yet largely overlook the patent domain. PTAB studies have primarily examined procedures distinct from *ex parte* appeal, such as Post-Grant Review (PGR), Inter Partes Review (IPR), and *Ex Parte* Reexamination (EPR), or

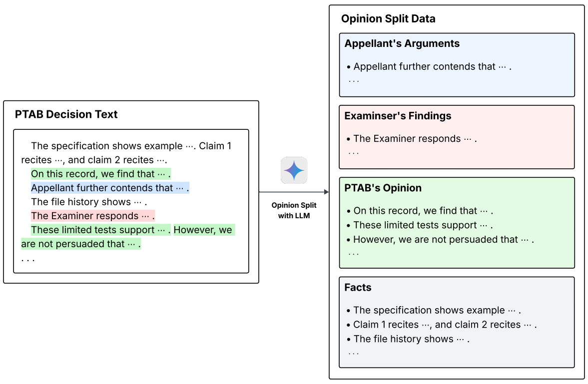


Figure 2: Opinion Split of PTAB Decisions. Given a PTAB decision, an LLM segments the text at the sentence-level and, using context, classifies each sentence into four roles; `appellant_arguments`, `examiner_findings`, `ptab_opinion`, and `facts`. The resulting Opinion Split Data serves as the base input for our IRAC-aligned classification tasks.

have focused on predicting outcomes from text and metadata, with little attention to integrated legal reasoning or LLM evaluation.

PILOT-Bench directly addresses these shortcomings by targeting *ex parte* appeals, aligning PTAB decisions with USPTO patent information at the case-level, and translating the IRAC framework into classification tasks that enable systematic assessment of LLMs' legal-reasoning performance in the patent domain.

4 Data Construction

This section describes the construction of the PILOT-Bench dataset, including source collection, case-level alignment, text normalization, opinion splitting, and label refinement. The goals are threefold: (i) to consistently align PTAB decisions with USPTO patent information; (ii) to prevent answer leakage by excluding the Board's opinion from inputs via the Opinion Split; and (iii) to provide input–label sets that reflect PTAB practice and are directly applicable to IRAC-aligned classification tasks.

4.1 Data Sources & Scope

- **PTAB Metadata (JSON, 170K)** Using USPTO's PTAB API v2[2], we collect metadata such as proceeding identifiers, application/publication numbers, proceeding type, panel judges, decision dates, and decision types.
- **PTAB Decisions (PDF, 25K)** We apply OCR to the original PDF decisions to extract the full opinion text and segment conventional sections such as *Decision on Appeal*, *Statement of the Case*, and *Analysis*.
- **USPTO Patent (XML)** From USPTO bulk XML[3], we extract only textual components—titles, claims, and specifications—and preprocess claims to preserve their dependency structures.

We set the PTAB window to 2009–2024 to ensure consistent document formatting and reliable OCR (standardized cover pages). For USPTO patent text, we use 2006–2024 to approximate a 20-year horizon relative to appeal filings and to cover applications linked to appeals decided after 2009.

4.2 Opinion Split

PTAB decisions intermix the appellant's arguments, the examiner's findings, and the PTAB's opinion. To prevent answer leakage, we exclude the Board's opinion from model inputs and retain only the appellant's and examiner's arguments. This design ensures that classification tasks such as Issue Type, Board Authorities, and Subdecision measure an LLM's ability to compare and synthesize conflicting arguments, rather than relying on the Board's conclusions.

The split dataset is primarily derived from the *Statement of the Case* and *Analysis* sections, which encompass the substantive exchanges between the appellant and the examiner. To construct the split dataset, each decision is processed by an LLM instructed to classify sentences into four categories: `appellant_arguments`, `examiner_findings`, `ptab_opinion`, and `facts`. After evaluating outputs across multiple models, we selected Gemini-2.5-pro as the final splitter for large-scale classification. The full prompt used in this task is provided in the Appendix D.3.

In addition, we further analyzed document-level statistics of the Opinion Split data to assess input scale and variability across decisions. On average, each split decision contains approximately 1.4K words and 8.7K characters, reduced by about 25% relative to the original sections (*Statement of the Case* + *Analysis*) due to the exclusion of PTAB opinion text. Among the original sections,

[2] https://developer.uspto.gov/api-catalog/ptab-api-v2

[3] https://data.uspto.gov/bulkdata/datasets

the *Statement of the Case* averages 430 words while the *Analysis* section averages 1.4K words, indicating that most of the reasoning content resides in the latter. Within the split data, the `appellant_arguments` and `examiner_findings` segments are similar in length (about 300 words each), whereas the `ptab_opinion` portion, retained only for reference, is substantially longer and more variable (820 words on average). These findings suggest that the input texts used for model evaluation maintain a balanced representation of opposing arguments while preserving realistic document scale. Full descriptive statistics, including word- and character-level summaries and role-wise distributions, are provided in Appendix E.4.

4.3 Labeling Sources & Regularization

We refine labels for three classification tasks, starting from the metadata in PTAB JSON and consolidating them into a schema restricted to merits determinations in *ex parte* appeals.

For the Issue Type task, the raw metadata contained six statutory sections under 35 U.S.C. (§100, 101, 102, 103, 112, and 120). To improve consistency and focus on the most frequent and practically relevant issues, we reduced these to five labels: *101*, *102*, *103*, *112*, and an *Others* category. Because a single appeal may raise multiple issues, this task is modeled as multi-label.

For the Board Authorities task, we identified the regulatory provisions cited in PTAB's opinions as the operative authorities for decisions. Although 35 U.S.C. sections appear in the raw data, the operative authority in *ex parte* appeals is generally 37 C.F.R.; accordingly, we select the most frequent provisions—§*1.131*, *1.132*, *41.50*, *41.50(a)*, *41.50(b)*, *41.50(c)*, *41.50(d)*, and *41.50(f)*—and group the remainder under *Others*, yielding a nine-label schema. Boilerplate references such as 35 U.S.C. §134 were excluded. Like Issue Type, this task is modeled as multi-label.

For the Subdecision task, we standardized the final outcomes of PTAB decisions. In the base dataset, we initially observed 34 distinct outcome labels. Since our corpus is restricted to appeal proceedings, we excluded reexamination appeals as well as AIA trial outcomes (e.g., IPR, PGR, CBM), removing AIA-specific categories such as Institution Granted. This reduction yielded 23 appeal-specific outcomes. We then applied normalization (case folding, whitespace and punctuation unification) and synonym merging to consolidate the labels. We provide these 23 outcomes as a set of fine-grained labels, which include an *Others* category grouping infrequent outcomes. In addition, we map them into six coarse-grained labels that dominate in *ex parte* appeals: *Affirmed*, *Affirmed with New Ground of Rejection*, *Affirmed-in-Part*, *Affirmed-in-Part with New Ground of Rejection*, *Reversed*, *Reversed with New Ground of Rejection*, and *Others*.

After defining these schemas, we examined their distributions. As shown in Figure 3, all tasks are highly imbalanced. Additional information on the labels is provided in the Appendix D.2.

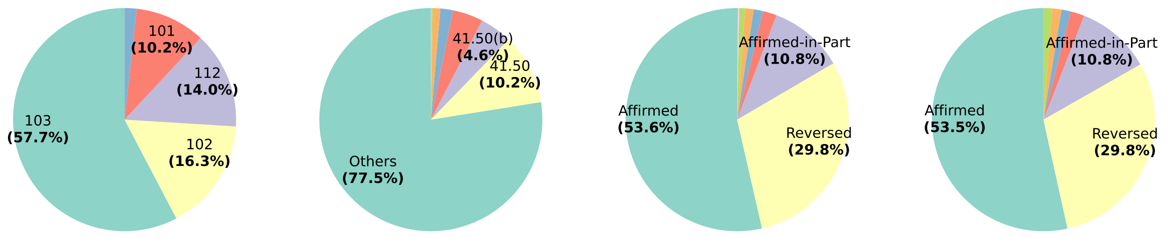
5 Tasks

In this section, we formalize the benchmark's three classification tasks in alignment with the IRAC framework. While we follow IRAC's logical order, the tasks are defined as independent evaluation units without dependencies across them. A uniform input and leakage-prevention policy applies: to avoid answer leakage, we exclude all PTAB's opinion text, and by default inputs consist only of the `appellant_arguments` and `examiner_findings` produced by the Opinion Split.

We note that the benchmark does not include a task corresponding to the Application stage of IRAC. Application requires multi-step reasoning that connects legal rules to case-specific facts, which goes beyond the scope of classification. In this work, we focus on classification tasks as a first step, and leave Application to future research, where it can be more appropriately modeled through generation tasks that capture complex legal reasoning.

5.1 Issue Type (IRAC–Issue)

This task identifies which statutory grounds are disputed in a case. The model must contrast and synthesize the competing arguments of the appellant and the examiner to determine the contested legal issues, corresponding directly to the Issue stage of IRAC. The task is formulated as multi-label classification at the case-level. For evaluation, we report three complementary metrics: Exact Match as an overall case-level measure, Macro-F1 to capture performance under label imbalance, and Micro-F1 to reflect overall distributional performance. Additional evaluation metrics are reported in Appendix 10.



(a) Issue type (b) Board Authorities (c) Subdecision (Fine, Top10) (d) Subdecision (Coarse)

Figure 3: Label distributions across tasks are imbalanced; for Subdecision (fine), only the top 10 labels are shown. Bold values under the labels are the proportion each label occupies in the dataset.

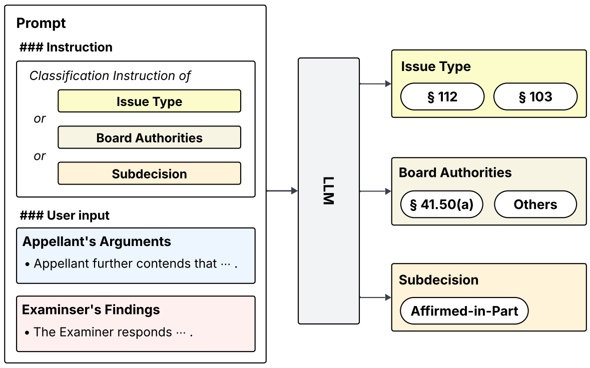


Figure 4: Task-specific prompting. A standardized prompt combines a task-specific instruction with the `appellant_arguments` and `examiner_findings` segments; the LLM then executes the chosen task–Issue, Board Authorities, or Subdecision–and outputs from the predefined label set.

5.2 Board Authorities (IRAC–Rule)

This task predicts which procedural provisions under 37 C.F.R. are cited as authority for the Board's decision, given the parties' arguments and evidence. This corresponds to the Rule stage of IRAC. Like the Issue Type task, this task is modeled as a case-level multi-label classification and evaluated using the same metrics: Exact Match, Macro-F1, Micro-F1. Other evaluation metrics are provided in the Appendix 11.

5.3 Subdecision (IRAC–Conclusion)

This task predicts the Board's final outcome for an appeal. The model must integrate conflicting claim-level arguments and facts from both sides and select a single conclusion for the case, corresponding to the Conclusion stage of IRAC. The task is framed as multi-class classification. For evaluation, we report Accuracy as the baseline overall measure, Macro-F1 to account for class imbalance, and Weighted-F1 to reflect performance across the empirical label distribution. Other evaluation metrics, such as micro-F1, are reported in the Appendix 12 and 13.

6 Experiments

We describe the experimental setup, model lineup, and evaluation protocol for the three classification tasks. Unless otherwise noted, inputs are restricted to the `appellant_arguments` and `examiner_findings` obtained from the Opinion Split, with all PTAB's opinion text excluded. For input-variation experiments, we compare three configurations under identical instructions: Split (Base), Merge, and Split+Claim. In the Split (Base) setting, appellant and examiner arguments are separated into distinct segments. Merge combines the two roles into a single role-neutral input, while Split+Claim augments the role-separated arguments with the patent's claim text. These variants allow us to analyze the relative contributions of role signals (the distinction between appellant and examiner) and technical signals (the claim text) to model performance.

The model lineup includes five closed-source(commercial) LLMs and four open-source LLMs. The closed-source(commercial) models are Claude-Sonnet-4 (Anthropic, 2025), Gemini-2.5-pro (Gemini Team, 2025), GPT-4o, GPT-o3 (OpenAI, 2024), and Solar-pro2 (Upstage, 2025). The open-source models are LLaMA-3.1 (Meta AI, 2024), Mistral (Jiang et al., 2023), Qwen (Qwen Team, 2025), and T5 (Google DeepMind, 2025). For closed-source(commercial) models, structured output features such as function calling were used to guarantee JSON-only responses. For open-source models, which lack native structured output capabilities, we enforced consistency by providing explicit format examples

Model	Exact Match	Macro-F1	Micro-F1
	Split (Base)		
Claude-Sonnet-4	0.5871	0.5457	0.7905
Gemini-2.5-pro	0.5874	0.6630	0.7923
GPT-4o	0.5751	0.6519	0.7860
GPT-o3	**0.5955**	**0.6639**	**0.7968**
Solar-pro2	0.5583	0.5240	0.7707
LLaMA-3.1(8B)	0.1826	0.1051	0.5793
Mistral(7B)	0.3405	0.2111	0.6080
Qwen(8B)	0.5561	0.5251	0.7741
T5(2B)	0.0772	0.3845	0.4469
	Merge		
Claude-Sonnet-4	0.5879	0.5468	0.7915
Gemini-2.5-pro	0.5810	0.6625	0.7889
GPT-4o	0.5516	0.6422	0.7758
GPT-o3	**0.5943**	**0.6645**	**0.7961**
Solar-pro2	0.5466	0.6249	0.7643
LLaMA-3.1(8B)	0.1334	0.4517	0.5801
Mistral(7B)	0.2639	0.1356	0.5760
Qwen(8B)	0.5322	0.6255	0.7634
T5(2B)	0.0057	0.3534	0.4050
	Split+Claim		
Claude-Sonnet-4	0.5869	0.5443	0.7915
Gemini-2.5-pro	0.5911	0.6632	0.7955
GPT-4o	0.5658	0.6492	0.7828
GPT-o3	**0.5946**	**0.6639**	**0.7967**
Solar-pro2	0.5355	0.6225	0.7596
LLaMA-3.1(8B)	0.1785	0.4360	0.5928
Mistral(7B)	0.4200	0.2662	0.6767
Qwen(8B)	0.5631	0.6353	0.7782
T5(2B)	0.0155	0.0024	0.4545

(a) Issue Type

Model	Exact Match	Macro-F1	Micro-F1
	Split (Base)		
Claude-Sonnet-4	0.4945	0.2397	0.5444
Gemini-2.5-pro	0.5906	**0.2665**	**0.6916**
GPT-4o	**0.6314**	0.2589	0.6522
GPT-o3	0.5302	0.1940	0.6236
Solar-pro2	0.4293	0.1014	0.6179
LLaMA-3.1(8B)	0.0000	0.0843	0.1230
Mistral(7B)	0.0028	0.0075	0.2762
Qwen(8B)	0.1542	0.1420	0.1966
T5(2B)	0.0064	0.0026	0.2116
	Merge		
Claude-Sonnet-4	**0.7761**	0.2128	**0.8033**
Gemini-2.5-pro	0.6323	0.3062	0.7387
GPT-4o	0.6032	**0.2486**	0.6179
GPT-o3	0.6459	0.2160	0.7344
Solar-pro2	0.2531	0.0620	0.5524
LLaMA-3.1(8B)	0.0000	0.0882	0.1629
Mistral(7B)	0.0028	0.0038	0.2729
Qwen(8B)	0.4266	0.1897	0.4531
T5(2B)	0.0026	0.0032	0.1757
	Split+Claim		
Claude-Sonnet-4	0.2026	0.1530	0.2636
Gemini-2.5-pro	**0.4913**	**0.2201**	**0.5795**
GPT-4o	0.0035	0.1425	0.1431
GPT-o3	0.2477	0.2109	0.4194
Solar-pro2	0.0041	0.0485	0.1780
LLaMA-3.1(8B)	0.0001	0.0923	0.1950
Mistral(7B)	0.0003	0.0044	0.1603
Qwen(8B)	0.0134	0.1136	0.0574
T5(2B)	0.0009	0.0037	0.1442

(b) Board Authorities

Table 2: Exact Match, Macro-F1 and Micro-F1 scores of Issue Type and Board Authorities classification

in the instruction and applying post-processing to convert outputs into valid JSON. This ensured that parsing errors were minimized across all runs.

All tasks are evaluated in a zero-shot setting under a unified protocol. Detailed instruction templates, and prompts are provided in Appendix D.3 and model specifications are provided in the Appendix F .

7 Results

We evaluate model performance across the three classification tasks, with task-level results reported in Tables 2a–3b; confusion heatmaps appear in the Appendix E.2. Overall, closed-source(commercial) models consistently outperform open-source models, although all models exhibit limitations under long-tailed label distributions. Macro-F1 remains low across tasks, reflecting persistent difficulty with rare labels.

7.1 Closed-Source(commercial) vs. Open-Source Models

As shown in the confusion heatmaps (Figures 16–27), closed-source(commercial) models (Claude-Sonnet-4, Gemini-2.5-pro, GPT-4o, GPT-o3, Solar-pro2) achieve consistently higher accuracy and exhibit a stronger diagonal concentration, indicating greater reliability in classification performance. In the Issue Type task under the Split (Base) setting, closed-source(commercial) models reach Exact Match scores around 55–60% with Micro-F1 scores close to 0.80, whereas open-source models are far less consistent: LLaMA-3.1 and Mistral remain below 35% Exact Match, T5 collapses to below 10%, and only Qwen approaches closed-source(commercial)-level performance. The Issue Type results thus provide the clearest illustration of the performance gap between closed-source(commercial) and open-source models.

7.2 Input-Setting Effects

Split (Base) provides the most reliable performance across tasks. Merge occasionally improves consistency for certain models, such as Claude-Sonnet-4 and GPT-o3, suggesting that role separation can sometimes introduce unnecessary variability. Split+Claim generally degrades performance: input length increases by roughly twice on average, and by a factor of three to four in terms of maximum token count, compared to Split (Base) (Table 8). This dilutes the salience of arguments and introduces irrelevant claim text as noise. The effect is

Model	Accuracy	Macro-F1	Weighted-F1
Split (Base)			
Claude-Sonnet-4	0.5658	0.1296	0.4854
Gemini-2.5-pro	0.5050	0.1635	0.4982
GPT-4o	0.4924	0.0997	0.4907
GPT-o3	**0.5918**	**0.1639**	**0.5541**
Solar-pro2	0.5369	0.0779	0.3923
LLaMA-3.1(8B)	0.4364	0.0767	0.4006
Mistral(7B)	0.1241	0.0251	0.1284
Qwen(8B)	0.4794	0.1024	0.4450
T5(2B)	0.0419	0.0142	0.0617
Merge			
Claude-Sonnet-4	0.5590	0.1129	0.4320
Gemini-2.5-pro	0.5114	0.1443	0.5036
GPT-4o	0.4592	0.0912	0.4353
GPT-o3	**0.6086**	**0.1683**	**0.5682**
Solar-pro2	0.5420	0.0804	0.3932
LLaMA-3.1(8B)	0.5036	0.0696	0.0676
Mistral(7B)	0.1265	0.0572	0.0407
Qwen(8B)	0.4266	0.0698	0.4264
T5(2B)	0.0191	0.0794	0.0437
Split+Claim			
Claude-Sonnet-4	0.5620	0.1272	0.4842
Gemini-2.5-pro	0.4908	0.4854	0.1433
GPT-4o	0.3804	0.0892	0.3581
GPT-o3	**0.5884**	**0.1692**	**0.5538**
Solar-pro2	0.5373	0.0608	0.3966
LLaMA-3.1(8B)	0.4125	0.0642	0.3938
Mistral(7B)	0.1209	0.0295	0.1205
Qwen(8B)	0.4368	0.0794	0.4364
T5(2B)	0.0225	0.0436	0.0168

(a) Subdecision (Fine-grained)

Model	Accuracy	Macro-F1	Weighted-F1
Split (Base)			
Claude-Sonnet-4	0.5625	0.2116	0.4900
Gemini-2.5-pro	0.5063	**0.2366**	0.4927
GPT-4o	0.5045	0.2037	0.4863
GPT-o3	**0.5863**	0.2126	**0.5511**
Solar-pro2	0.5389	0.1356	0.3929
LLaMA-3.1(8B)	0.4764	0.1551	0.4024
Mistral(7B)	0.0726	0.0758	0.0994
Qwen(8B)	0.4733	0.1692	0.4404
T5(2B)	0.0254	0.0499	0.0146
Merge			
Claude-Sonnet-4	0.5607	0.1788	0.4456
Gemini-2.5-pro	0.5119	**0.2381**	0.5001
GPT-4o	0.4972	0.1820	0.4638
GPT-o3	**0.6020**	0.2125	**0.5631**
Solar-pro2	0.5423	0.1390	0.3967
LLaMA-3.1(8B)	0.5229	0.1253	0.3922
Mistral(7B)	0.0823	0.0821	0.1168
Qwen(8B)	0.4163	0.1761	0.4223
T5(2B)	0.0234	0.0446	0.0092
Split+Claim			
Claude-Sonnet-4	0.5639	0.2018	0.4889
Gemini-2.5-pro	0.4915	0.4840	0.2111
GPT-4o	0.3046	0.1206	0.2027
GPT-o3	**0.5783**	**0.2068**	**0.5426**
Solar-pro2	0.5364	0.1210	0.3977
LLaMA-3.1(8B)	0.4741	0.1259	0.3909
Mistral(7B)	0.0587	0.0549	0.0721
Qwen(8B)	0.4605	0.1655	0.4439
T5(2B)	0.0136	0.0053	0.0142

(b) Subdecision (Coarse-grained)

Table 3: Accuracy, Macro-F1 and Weighted-F1 scores of Subdecision (Fine-grained) and Subdecision (Coarse-grained) classification

most pronounced in the Board Authorities task (Table 2b), where all models except Gemini-2.5-pro show a clear decline. Unlike Issue Type or Subdecision, which integrate technical facts with legal reasoning, Board Authorities is narrowly focused on mapping arguments to procedural rules. In this setting, claim text contributes little useful information and instead confuses the model, leading to a sharper performance drop. These results highlight that more input context is not uniformly beneficial: when tasks hinge primarily on legal rule alignment rather than technical content, excessive claim context may actively impair model reasoning.

7.3 Invalid Response Patterns

Another clear pattern, especially among open-source models, is the generation of labels outside the predefined set. For example, in Issue Type and Board Authorities tasks, models occasionally output arbitrary numbers or provisions not included in the label schema. This indicates both a failure to strictly follow instructions and a lack of domain alignment. Potential remedies include stronger prompt constraints (explicitly requiring outputs to be drawn only from the label set), post-filtering to reject out-of-label responses, and instruction tuning to reduce invalid or incomplete responses. Example cases of label deviations and invalid responses are presented in Appendix F.2.

7.4 Summary

Taken together, these results show that while closed-source(commercial) models can handle frequent labels and surface-level reasoning, all models struggle with long-tailed label distributions. The IRAC-based task design exposes these weaknesses across different stages, while the input-setting analysis underscores the importance of careful input design. Future work will build on these findings by exploring selective claim augmentation and instruction tuning as ways to improve alignment with PTAB-specific reasoning tasks.

8 Conclusion

We presented PILOT-Bench, a benchmark to evaluate legal reasoning in the patent domain by aligning PTAB *ex parte* appeals with USPTO patent data. By framing three IRAC-aligned classification tasks, we enable systematic assessment of LLMs' ability to identify issues, map rules, and predict conclusions in appeal proceedings. Our experiments show that while closed-source(commercial) LLMs outperform open-source models, all models face persistent challenges with label imbalance and procedural-rule mapping. Input-variation analysis further demonstrates that simply adding all claims can harm performance, underscoring the need for more targeted data design.

PILOT-Bench thus provides both a resource and an evaluation protocol to study how LLMs reason in a domain where technical detail and legal precision must be combined. We hope this benchmark will encourage further work at the intersection of NLP, law, and intellectual property.

9 Future Work

Beyond this study, we plan to pursue research-driven extensions of PILOT-Bench. A first direction is to expand beyond classification by introducing generation-based tasks that capture the IRAC Application stage, directly testing whether models can reason through the application of legal rules to facts. Second, we aim to explore selective claim augmentation and instruction tuning to mitigate noise and hallucination, thereby improving alignment with task constraints. Finally, we envision extending the benchmark to broader PTAB and USPTO contexts, enabling multi-procedure comparisons and richer evaluation of patent-domain legal reasoning.

Limitations

This study has several limitations related to data collection and task design. First, the scope is restricted to PTAB *ex parte* appeals, excluding AIA trial proceedings. While this aligns with source availability and our intended focus, it confines evaluation to appeal-centered cases. Second, although OCR quality is generally stable, no systematic, line-by-line correction against the source PDFs was performed; the converted text should not be regarded as a fully verified transcription. Similarly, the Opinion Split was generated solely via an LLM without human validation, so misclassifications may propagate into downstream tasks. Finally, the dataset exhibits substantial label imbalance. To address this, Subdecision outcomes were consolidated into six coarse labels via LLM-based normalization without additional rebalancing. Partnering with domain experts to vet and refine this schema may yield further gains in robustness and interpretability.

Ethical Considerations

This benchmark is released for research purposes only and must not be used to automate, replace, or appear to provide legal advice or adjudicative decisions. All documents originate from public USPTO/PTAB sources; we redistribute only derived annotations/splits/metadata and remove any incidental PII found during OCR. Users remain responsible for compliance with applicable laws and professional standards. Model outputs may contain errors and require qualified human review.

Acknowledgments

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under grants through the National Research Foundation of Korea (NRF) (No. RS-2024-00354675, 70%) and the ICAN (ICT Challenge and Advanced Network of HRD) support program supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation) (IITP-2023-RS-2023-00259806, 30%)

References

Anthropic. 2025. Claude sonnet 4.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, Dublin, Ireland. Association for Computational Linguistics.

Caspar J. Fall, Attila Törcsvári, Karim Benzineb, and Gábor Karetka. 2003. Automated categorization in the international patent classification. *SIGIR Forum*, 37(1):10–25.

Xiaoyong (Jack) Fu. 2021. Patents: Ability or choice? SSRN working paper.

Oscar A. Garcia, Naisargi Dave, Qie Tang, Josvin John, Anthony Topper, Kashyap Bhuva, Manasi Shrotri, Sayali Shelke, Xiaosong Wen, Reza Mollaaghababa, Fatemeh Emdad, Chun-Kit Ngan, Elke Rundensteiner, and Seyed A. Zekavat. 2022. A deep learning model for predicting patent applications outcomes. *The Journal of Robotics, Artificial Intelligence & Law (RAIL)*, 5(5):347–356.

Gemini Team. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities.

Google DeepMind. 2025. T5gemma: Encoder–decoder gemma models.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, and 21 others. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Computing Research Repository*, arXiv:2308.11462.

Peter Henderson, Mark S. Krass, Lucia Zheng, Neel Guha, Christopher D. Manning, Dan Jurafsky, and Daniel E. Ho. 2022. Pile of law: Learning responsible data filtering from the law and a 256GB open-source legal dataset. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022), Datasets and Benchmarks Track*.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. Cuad: An expert-annotated NLP dataset for legal contract review. *Computing Research Repository*, arXiv:2103.06268.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *arXiv*.

Lekang Jiang, Pascal A. Scherz, and Stephan Goetz. 2024. Patent-cr: A dataset for patent claim revision. *Computing Research Repository*, arXiv:2412.02549.

Shaobo Li, Jie Hu, Yuxin Cui, and Jianjun Hu. 2018. Deeppatent: patent classification with convolutional neural networks and word embedding. *Scientometrics*, 117:721–744.

Brian J. Love, Shawn P. Miller, and Shawn Ambwani. 2019. Determinants of patent quality: Evidence from Inter Partes review proceedings. *University of Colorado Law Review*, 90:67–165.

Matthew S. Makover and Lexi Boynes. 2025. Uspto introduces AI strategy to drive innovation and balance IP protections.

Meta AI. 2024. Llama-3.1-8b-instruct (checkpoint used in experiments).

Joel Niklaus, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, and Daniel Ho. 2024. Multilegalpile: A 689GB multilingual legal corpus. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15077–15094, Bangkok, Thailand. Association for Computational Linguistics.

OpenAI. 2024. Models.

Florina Piroi. 2010. Clef-ip 2010: Retrieval experiments in the intellectual property domain. In *CLEF 2010 LABs and Workshops, Notebook Papers*.

Florina Piroi, Mihai Lupu, Allan Hanbury, and Veronika Zenz. 2011. Clef-ip 2011: Retrieval in the intellectual property domain. In *CLEF 2011 Labs and Workshop, Notebook Papers*, volume 1177 of *CEUR Workshop Proceedings*, Amsterdam, The Netherlands. CEUR-WS.org.

Qwen Team. 2025. Qwen3 technical report. *arXiv*.

Kripa Rajshekhar, Wlodek Zadrozny, and Sri Sneha Garapati. 2017. Analytics of patent case rulings: Empirical evaluation of models for legal relevance. In *Proceedings of the 16th International Conference on Artificial Intelligence and Law (ICAIL 2017)*, London, United Kingdom.

Eva Sharma, Chen Li, and Lu Wang. 2019. Bigpatent: A large-scale dataset for abstractive and coherent summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2204–2213, Florence, Italy. Association for Computational Linguistics.

Homaira Huda Shomee, Zhu Wang, Sourav Medya, and Sathya N. Ravi. 2024. Impact: A large-scale integrated multimodal patent analysis and creation dataset for design patents. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Datasets and Benchmarks Track*.

Ryan K. Simmons. 2024. Artificial intelligence and the patent application process: A synopsis of the potential benefits and risks.

Bahrad A. Sokhansanj and Gail L. Rosen. 2022. Predicting institution outcomes for inter partes review (ipr) proceedings at the united states patent trial & appeal board by deep learning of patent owner preliminary response briefs. 12(7):3656.

Mirac Suzgun, Luke Melas-Kyriazi, Suproteem K. Sarkar, Scott Duke Kominers, and Stuart M. Shieber. 2022. The harvard uspto patent dataset: A large-scale, well-structured, and multi-purpose corpus of patent applications.

Don Tuggener, Pius von Däniken, Thomas Peetz, and Mark Cieliebak. 2020. Ledgar: A large-scale multi-label corpus for text classification of legal provisions in contracts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1235–1241, Marseille, France. European Language Resources Association.

Upstage. 2025. Solar pro 2: Fluent. reasoning. frontier.

USPTO. 2024. Guidance on use of artificial intelligence-based tools in practice before the united states patent and trademark office. 89:25609–25617.

USPTO. 2025. Patent trial and appeal board (ptab) statistics.

Qiyao Wang, Shiwen Ni, Huaren Liu, Shule Lu, Guhong Chen, Xi Feng, Chi Wei, Qiang Qu, Hamid Alinejad-Rokny, Yuan Lin, and Min Yang. 2024. Autopatent:

A multi-agent framework for automatic patent generation.

David Winer. 2017. Predicting bad patents: Employing machine learning to predict post-grant review outcomes for us patents. Technical Report UCB/EECS-2017-60, EECS Department, University of California, Berkeley.

Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. 2021. When does pretraining help? assessing self-supervised learning for law and the CaseHOLD dataset of 53,000+ legal holdings. In *Proceedings of the 18th International Conference on Artificial Intelligence and Law (ICAIL '21)*, pages 159–168, São Paulo, Brazil. Association for Computing Machinery.

Appendix

A Data Card

- **Licensing Information** The dataset is released under the Creative Commons Attribution 4.0 International License.
- **Data Domain** Patent Domain
- **Languages** The dataset contains English text only.
- **Dataset Composition** PTAB OCR, PTAB Opinion Split, PTAB Metadata, and USPTO Structured Data.
- **Computational Resources** Experiments were run on two RTX 4090(24GB) and two H100(80GB) GPUs

B Data Format and Structure

B.1 PTAB Decision

Each PTAB decision is distributed as a JSON file named after the official decision filename (e.g., `2018004769_DECISION.json`). We release two corpus variants: PTAB OCR and PTAB Opinion Split. PTAB OCR provides page-level Optical Character Recognition (OCR) text, providing extracted from each decision. PTAB Opinion Split segments the decision text into four categories: `appellant_arguments`, `examiner_findings`, `ptab_opinion`, and `facts`.

B.2 PTAB Metadata

we release a PTAB Metadata JSON aligned PTAB decision JSON files. PTAB Metadata contains 35 fields per decision, including the targets used in our classification tasks: `issueType`, `boardRulings`, and `subdecisionTypeCategory`. Table 4 shows the metadata JSON fields.

B.3 USPTO Structured Data

For each decision, we include the corresponding USPTO patent data as a single JSON file within the directory for that PTAB Decision filename, named by the patent's application or publication number (e.g., `2018004769_DECISION/US20140127537A1.json`).

C Dataset Creation

C.1 Source Data

We collected 25,829 PTAB decisions (1993–2024) and 176,627 metadata records (1997–2025) via the PTAB API v2[4]. We also retrieved patent full texts and bibliographic metadata from USPTO Bulk Data[5], covering 2006–2024.

C.2 Patent-Term Filtering

Considering the statutory patent term (typically 20 years from the filing date), we restrict our analysis to PTAB decisions dated 2006 or later, yielding 22,439 cases.

C.3 OCR Quality Filtering

We require page-level OCR for decision text analysis. Nonstandard layouts—often due to missing cover pages—disrupted caption normalization and section detection. To stabilize OCR, we retain only decisions with a cover page, resulting in 18,738 cases.

C.4 Case-Thread Normalization

We define the analysis scope for *ex parte* appeal case threads and apply metadata-driven preprocessing to normalize threads and remove duplicates. To ensure a reproducible one-to-one mapping between each case and its associated patent text, we adopt a single target per case and restrict the analysis to a subset of procedural variants. Records that could yield duplicate or ambiguous labels are excluded.

- **Exact duplicates** Decision records Decision records that are identical across all fields; a single canonical decision record is retained.
- **Application number / document name duplicates** When multiple decision records share `documentName` and `appellantApplicationNumberText`, we reconcile the PTAB Decision with PTAB Metadata and preserve one consistent decision record.
- **Subsequent proceedings (rehearing/reconsideration/reexamination)** Subsequent decisions within the same proceeding can produce multiple decision records for a single dispute. we retain one representative decision record per (`documentName`, `decisionDate`) pair.
- **Separate opinions (dissent/concurring)** Separately authored opinions are excluded because they may introduce competing rationales and thus ambiguous case-level labels.

[4]https://developer.uspto.gov/api-catalog/ptab-api-v2

[5]https://data.uspto.gov/bulkdata/datasets

Only the unified decision record is kept for downstream tasks.

C.5 OCR Parsing

From the OCR text, we removed cover-page bibliographic fields (e.g., `Application No.`, `Filing Date`, `First Named Inventor`) that duplicate metadata entries, thereby preventing redundancy. To maintain linguistic consistency and improve OCR robustness, we also removed non-English text.

C.6 Section Segmentation

To support a logical decomposition of each decision, we defined a header dictionary comprising `DECISION ON APPEAL`, `STATEMENT OF THE CASE`, `ANALYSIS`, `DECISION/ORDER`, and `FOOTNOTES`, and we then performed section-level segmentation using GPT-o3 (3-2025-04-16). Decisions in which `STATEMENT OF THE CASE` or `ANALYSIS` could not be extracted—e.g., dismissals following a Request for Continued Examination (RCE) or express abandonment—were excluded from the analysis.

C.7 PTAB Opinion Split

Using the primary reasoning sections `STATEMENT OF THE CASE` and `ANALYSIS` as input, we split each decision with gemini-2.5-pro into four categories: `appellant_arguments`, `examiner_findings`, `ptab_opinion`, and `facts`. Only `appellant_arguments` and `examiner_findings` are used as inputs to downstream tasks. Figure 6 presents the prompt for opinion splitting.

C.8 PTAB to USPTO Mapping

We align PTAB decision records with USPTO patent records via the application number, matching PTAB `appellantApplicationNumberText` to USPTO `application-reference/doc-number`. When a single application number is associated with multiple publications, we select one representative publication anchored to the PTAB `decisionDate`. Applications predating 2006 fall outside the coverage of our USPTO corpus and are omitted. This alignment yields 15,482 PTAB–USPTO links.

C.9 USPTO Structured Data

To preserve claim dependencies, each claim carries a `depend_on` pointer to its parent claim. We further factor claim text into component-level units and arrange them hierarchically to support granular analyses in subsequent work. Figure 7 depicts the schema.

D Classification Tasks

D.1 Prediction Targets

Our tasks comprise three targets: issue type, board authorities, and subdecision. For consistency in evaluation, instances with missing `Board Authorities` (empty) are systematically mapped to `Others` label.

D.2 Label Details

Table 14–19 enumerates the full labels used in our experiments and their definitions.

D.3 Prompt

Figure 8–10 are the prompts used for each task; Issue Type, Board Authorities, Subdeicision (Fine/-Coarse).

E Statistics and Analysis

E.1 Input Tokens per Variants

Table 8 reports the average and maximum input token counts per input variant for the Board Authorities task, measured with the Gemini tokenizer.

E.2 Experiment Results

Tables 10–13 present results for all evaluation metrics. Table 10 shows that T5 attains unusually high recall despite weaker Exact Match, Micro-F1, and Macro-F1. Inspection of Figure 13-15 reveals a systematic tendency to emit the full five-label set (`[101,102,103,112,Others]`), which mechanically inflates recall in the multi-label setting by covering most labels while simultaneously depressing precision and exact match. All models' confusion heatmaps can be found in Figures 16–27

E.3 PTAB Subproceeding Types by Year

To illustrate the oral distribution and procedural composition of the PTAB corpus, we analyzed the number of decisions per year and subproceeding type (*REEXAM*, *REGULAR*, and *REISSUE*) based on the PTAB Document JSON metadata. Figure 5 and Table 5 show a steady increase in *REGULAR* appeal decisions from 2010 to 2017, followed by a gradual decline consistent with overall PTAB appeal volume trends. *REEXAM* and *REISSUE* proceedings account for less than 5% of total decisions, confirming that the dataset is dominated by regular *ex parte* appeals—the intended focus of PILOT-Bench.

E.4 Document Length Statistics of Opinion Split Data

We provide document and role aspect descriptive statistics to quantify the scale and variability of the Opinion Split data. Table 6 summarizes the word-level statistics, and Table 7 presents the corresponding character-level statistics. These results show that PTAB *ex parte* decisions vary widely in length, with the *Analysis* section dominating the total word count and the split inputs maintaining a balanced representation of opposing arguments.

E.5 Linked Patents per PTAB Case

To quantify the connectivity between PTAB decisions and their associated patents, we examined the number of linked patents per case after PTAB–USPTO alignment. Each PTAB case contains one *base patent* (the appellant's patent) and zero or more *prior patents* cited as prior art or reference patents in the appeal record. Figure 11 and Figure 12 visualize the distribution of linked patents across cases and its yearly trend.

On average, each PTAB case is connected to approximately **2.05 patents**, consisting of one base patent and roughly one additional prior patent. The average base-to-prior ratio is about **0.64**, indicating that while most cases are linked to a single prior reference, a small number of cases involve more complex prior-art networks (up to 14 linked patents). Table 9 reports detailed summary statistics.

F Model

This study evaluates both closed-source(commercial) and open-source models. For the open-source group, we primarily used small models in the 2B–8B parameter range due to computational constraints. We expect larger variants of the same architectures (>8B parameters) and models with dedicated reasoning modes to achieve higher performance. Details on model sizes are provided below.

- **Closed-source(commercial) Models** gpt-4o-2024-08-06, gpt-o3-2025-04-16, claude-sonnet-4-20250514, gemini-2.5-pro, solar-pro2-250710
- **Open-source Models** Llama-3.1-8B-Instruct, Qwen3-8B, Mistral-7B-Instruct-v0.3, t5gemma-2b-2b-ul2-it

F.1 Post-Processing of Model Outputs

For open-source models, we instructed JSON only output at the prompt stage. In practice, some responses exhibited formatting errors, so we applied content-preserving normalization. Specifically, (i) we corrected parsing errors caused by missing or superfluous brackets or quotation marks with minimal edits, (ii) we restored character-level fragmented outputs (e.g., "", "i", "s", "s", "u", . . .) to valid contiguous strings, and (iii) we removed duplicated labels such as "103", "103", "103". This pipeline was designed to enforce schema consistency without altering the meaning of the original responses.

F.2 Response Tendencies

F.2.1 Closed-Source(commercial) Models

- **Issue Type** Claude intermittently returned `<UNKNOWN>`.

- **Board Authorities** According to the labels, citations such as `37 CFR 1.104`, `37 CFR 1.111`, `37 CFR 41.37(c)(iv)` should be assigned to `Others`; nevertheless, the model occasionally emitted them as distinct labels.

F.2.2 Open-Source Models

- **Issue Type** We observed frequent deviations from the label set, bare numerals (e.g., `51`, `22`); subsection-annotated variants (e.g., `102(b)`, `103(a)`, `102(e)` instead of base labels `102`, `103`); and unstructured natural language text (e.g., "The Examiner found that claims . . . ").

- **Board Authorities** Category confusions and hallucinated citations were common. Statutory grounds intended for the Issue Type task (e.g., `35 U.S.C. § 103(a)`, `35 U.S.C. § 102(b)`) were misassigned as Board Authorities. Provisions outside our label set (e.g., `37 C.F.R. § 41.37(c)(1)(ii)`)—which should map to `Others`—were emitted as labels. We also observed nonexistent citations in our dataset (e.g., `37 C.F.R. § 41.132`, `§ 101`, `§ 102(e)`).

- **Subdecision** Mistral tended to produce natural language text rather than schema labels (e.g., "Claims 1–3, 17–23, 25, and 28–30 stand rejected.").

F.3 Evaluation Protocol and Response Rates

F.3.1 Evaluation Protocol

By default, we evaluated 15,482 cases. For each model–task pair, we allowed up to ten retries. A case was marked as a `non-answer` if (i) no output was produced, (ii) the model provided a rationale without a final label, or (iii) the input text was echoed verbatim or the response consisted of repetitive content.

F.3.2 Response Rates

- **Solar-pro2** Owing to maximum context-length limits, evaluation under Split+Claim covered 15,481 samples. See Table 8 for average input length.

- **T5** Under the Base and Merged, evaluations of Subdecision-Fine and Subdecision-Coarse yielded on average 15,470 valid responses. Despite up to ten retries, we frequently observed outputs consisting only of explanatory text without a label or terminating in repetitive content. Under Split+Claim, response rates declined across all tasks, with non-answers increasing via partial claim echoes or verbatim reproductions of the input; accordingly, metrics for Split+Claim were computed on approximately 15,040 samples.

- **Mistral.** Under Split+Claim for Board Authorities, the model frequently returned the input verbatim. Evaluation proceeded with 15,481 samples.

Name	**Definition**	**Example**
proceedingNumber	PTAB proceeding ID	2018004769
decisionTypeCategory	Decision type	"Decision"
subdecisionTypeCategory	Final outcome of decision	"Affirmed"
documentName	Decision PDF filename	"2018004769_DECISION.pdf"
proceedingTypeCategory	Proceeding type	"Appeal"
subproceedingTypeCategory	Sub-type of proceeding	"REGULAR"
documentIdentifier	Document ID	"201800476914127348Appeal ..."
objectUuId	Internal repository ID	"workspace: ..."
respondentTechnologyCenterNumber	Respondent USPTO Technology Center(TC)	"1700"
respondentPartyName	Respondent party name	"Samsung SDI Co., Ltd. et al"
respondentGroupArtUnitNumber	Respondent Group Art Unit(GAU) number	"1727"
respondentPatentNumber	Respondent patent number	"10028104"
respondentApplicationNumberText	Respondent application number	14127348
appellantTechnologyCenterNumber	Appellant USPTO Technology Center(TC)	"1700"
appellantPatentOwnerName	Appellant name	"Samsung SDI Co., Ltd. et al"
appellantPartyName	Appellant party name	"Samsung SDI Co., Ltd. et al"
appellantGroupArtUnitNumber	Appellant Group Art Unit(GAU) number	"1727"
appellantInventorName	Appellant inventor(s) name	"Claus Gerald Pflueger et al"
appellantCounselName	Appellant Counsel/firm	"Maginot, Moore & Beck LLP"
appellantGrantDate	Appellant patent grant date	"03-27-2018"
appellantPatentNumber	Appellant patent number	"9925542"
appellantApplicationNumberText	Appellant application number.	14127348
appellantPublicationDate	Appellant publication date	"05-08-2014"
appellantPublicationNumber	Appellant publication number	"20140127537A1"
ocrSearchText	OCR text by USPTO	"14127348,Patent_Board ..."
issueType	Statutory sections under 35 U.S.C.	["103"]
boardRulings	Regulatory provisions cited	["35 USC 134"]
decisionDate	Decision date	"03-21-2019"
documentFilingDate	Filing date of the decision doc	"03-21-2019"
thirdPartyName	Third party name	"SMITH & NEPHEW, INC."
file_name	Basename without extension.	"2018004769_DECISION"
issueType_label	Label of Issue Type task	["103"]
boardAuthorities_label	Label of Board Authorities task	[Others]
subdecisionType_label	Fine-grained label of Subdeicision task	"Affirmed"
subdecisionTypeCoarse_label	Coarse-grained label of Subdeicision task	"Affirmed"

Table 4: PTAB metadata fields

Year	**REEXAM**	**REGULAR**	**REISSUE**
2007	1	0	0
2008	0	1	0
2009	0	9	0
2010	19	410	7
2011	25	949	11
2012	36	1314	6
2013	35	1498	4
2014	44	1256	4
2015	34	1758	5
2016	25	2192	1
2017	14	1734	2
2018	8	1452	0
2019	5	1205	0
2020	6	1078	7
2021	4	1038	6
2022	7	830	6
2023	5	469	1
2024	6	518	3

Table 5: Number of PTAB decisions by subproceeding type from 2007 to 2024.

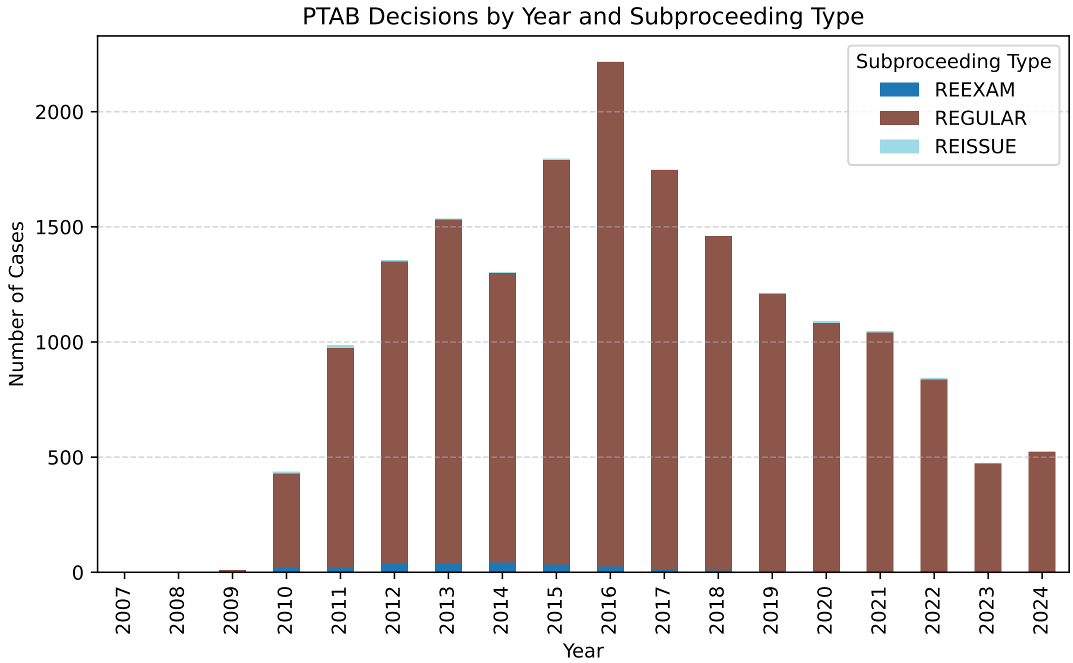


Figure 5: PTAB decisions by year and subproceeding type (2007–2024).

Section / Role	**Count**	**Mean (Words)**	**Median**	**Std**	**Min**	**Max**
Overall (Pre-Split)	18,049	1,864.3	1,551	1,143.6	0	10,261
Statement of the Case	17,919	433.4	366	276.5	19	4,685
Analysis	18,042	1,434.5	1,130	1,064.9	9	9,764
Overall (Post-Split)	18,049	1,409.1	1,173	935.7	0	10,039
appellant_arguments	17,445	296.5	235	242.6	3	2,613
examiner_findings	17,766	306.7	248	239.4	10	2,827
ptab_opinion	18,041	821.0	634	674.2	5	8,532

Table 6: Descriptive statistics of document and role-level word counts in the PTAB Opinion Split dataset.

Section / Role	**Count**	**Mean (Chars)**	**Median**	**Std**	**Min**	**Max**
Overall (Pre-Split)	18,049	11,565.6	9,563	7,202.5	1	64,872
Statement of the Case	17,919	2,690.3	2,241	1,749.8	120	28,950
Analysis	18,042	8,875.3	7,126	6,730.4	85	62,180
Overall (Post-Split)	18,049	8,748.5	7,245	5,883.9	2	64,594
appellant_arguments	17,445	1,856.2	1,468	1,525.4	14	17,163
examiner_findings	17,766	1,876.9	1,511	1,475.3	53	17,486
ptab_opinion	18,041	5,107.2	3,926	4,250.6	30	54,854

Table 7: Descriptive statistics of document and role-level character counts in the PTAB Opinion Split dataset.

Statistic	**Split (Base)**	**Merge**	**Split+Claim**
Average	2026.14	1730.00	4876.58
Maximum	6109.00	5193.00	20924.00

Table 8: Average and Maximum input tokens by variant (Board Authorities; Gemini tokenizer)

You are given two sections from a PTAB (Patent Trial and Appeal Board) decision:
Instruction

Classification Criteria
1. Appellant Arguments
Description of criteria classified as Appellant Arguments

2. Examiner Findings
Description of criteria classified as Examiner Findings

3. PTAB Opinion
Description of criteria classified as PTAB Opinion

4. Facts
Description of criteria classified as Facts

Rules
Description the rules the model must follow when responding

Few-Shot Example
Input
Few-Shot Input Example

Output
Response Examples with Output Format

Data to Classify
<STATEMENT OF THE CASE>*{statement_of_the_case}***</STATEMENT OF THE CASE>**
<ANALYSIS>*{analysis}***</ANALYSIS>**

Figure 6: Opinion Split prompt construction

US20140127537A1

BATTERY CELL MODULE, METHOD FOR OPERATING A BATTERY CELL MODULE AND BATTERY AND MOTOR VEHICLE

Claims

1. A battery cell module, comprising:

a plurality of battery cells each having a degassing orifice; and

a gas receiving chamber allocated to several battery cells of the plurality of battery cells, the gas receiving chamber configured to at least temporarily receive gases escaping from the several battery cells,

wherein a volume of the gas receiving chamber is directly connected to the degassing orifices of the several battery cells.

2. The battery cell module as claimed in claim 1 , wherein the gas receiving chamber is open in a direction towards the several battery cells and an opening region of the gas receiving chamber is configured to extend ...

Description

PRIOR ART

There is a considerably high demand for batteries for use in a wide range of applications for example for vehicles, stationary installations, for example wind power installations, and mobile electronic devices, for example laptops and communication ...


```
{
    "filename" : "US20140127537A1-20140508",
    "title" : "BATTERY CELL MODULE, METHOD ...",
    "claims" : [
        {
            "claim_id" : "CLM-00001",
            "claim_num" : "1",
            "depends_on" : "null",
            "claim_text" : {
                "text" : "1 . A battery cell module, comprising:",
                "components" : [
                    {"text" : "a plurality of battery cells each ..."},
    {"text" : "a gas receiving chamber ..."}
                ]
            }
        },
        {
            "claim_id" : "CLM-00002",
            "claim_num" : "2",
            "depends_on" : "CLM-00001",
            "claim_text" : { "text" : "..." }
        },
    ]
    "description" : [
        {
            "heading" : "PRIOR ART",
            "paragraphs" : [ "...", "..." ]
        }
    ]
}
```


Figure 7: USPTO Structured Data structure

Statistic	Base Count	Prior Count	Total
Count	78,480	78,480	78,480
Mean	0.99	1.06	2.05
Std. Dev.	0.10	1.47	1.47
Min	0	0	1
Max	1	13	14

Table 9: Summary statistics of linked patents per PTAB case. Each case contains one base patent and zero or more prior patents.

[Role & Mission]
Persona setting and Instruction

[Evidence Scope]
Description of the input setting

[Task]
Description of the Issue Type classification task

[Rules]
Description the rules the model must follow when responding

<Issue Type Set>
["101","102","103","112","Others"]
</Issue Type Set>

<Issue Type Definitions>
Issue Type label Dictionary
</Issue Type Definitions>

[Output Format]
Response Examples with Output Format

---- INPUT ----
<Appellant Arguments>*{appellant}***</Appellant Arguments>**
<Examiner Findings>*{examiner}***</Examiner Findings>**

Figure 8: Issue Type classification prompt construction

[Role & Mission]
Persona setting and Instruction

[Evidence Scope]
Description of the input setting

[Task]
Description of the Board Authorities classification task

[Rules]
Description the rules the model must follow when responding

<Board Ruling Dictionary>
[
"37 CFR 1.131",
"37 CFR 1.132",
"37 CFR 41.50",
"37 CFR 41.50(a)",
"37 CFR 41.50(b)",
"37 CFR 41.50(c)",
"37 CFR 41.50(d)",
"37 CFR 41.50(f)",
"Others"
]
</Board Ruling Dictionary>

<Board Ruling Definitions>
Board Authorities label Dictionary
</Board Ruling Definitions>

[Output Format]
Response Examples with Output Format

---- INPUT ----
<Appellant Arguments>*{appellant}***</Appellant Arguments>**
<Examiner Findings>*{examiner}***</Examiner Findings>**

Figure 9: Board Authorities classification prompt construction

[Role & Mission]
Persona setting and Instruction

[Evidence Scope]
Description of the input setting

[Task]
Description of the Subdecision classification task

[Rules]
Description the rules the model must follow when responding

<Decision Type Dictionary>
fine/coarse subdecision dictionary in the for of {index: label}
</Decision Type Dictionary>

[Output Format]
Response Examples with Output Format

---- INPUT ----
<Appellant Arguments>*{appellant}***</Appellant Arguments>**
<Examiner Findings>*{examiner}***</Examiner Findings>**

Figure 10: Subdecision (Fine/Coarse) classification prompt construction

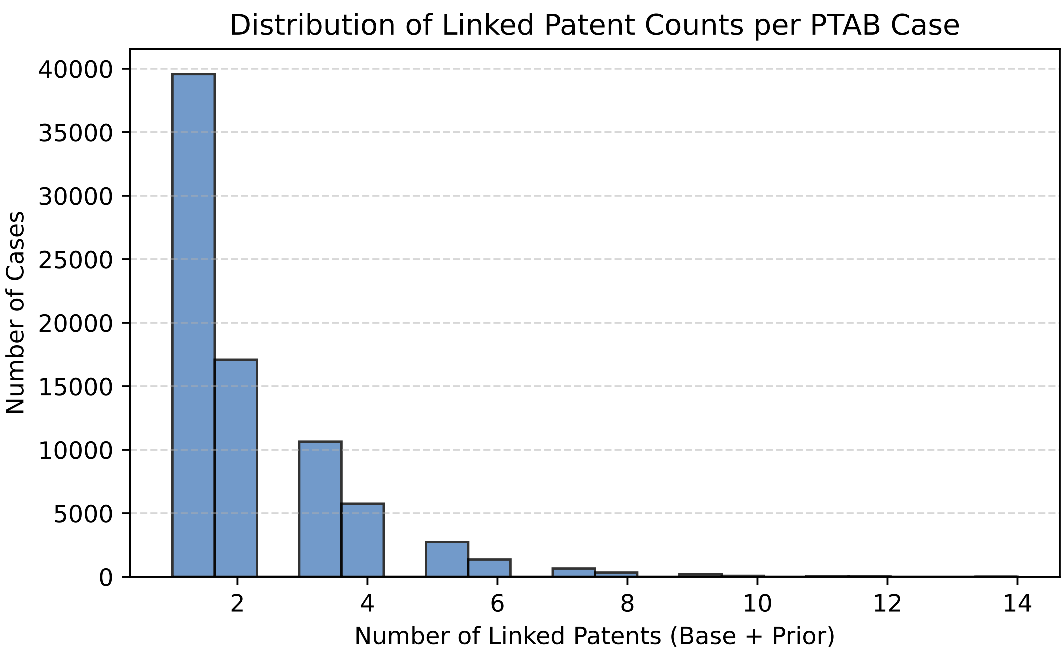


Figure 11: Distribution of the number of linked patents (base + prior) per PTAB case.

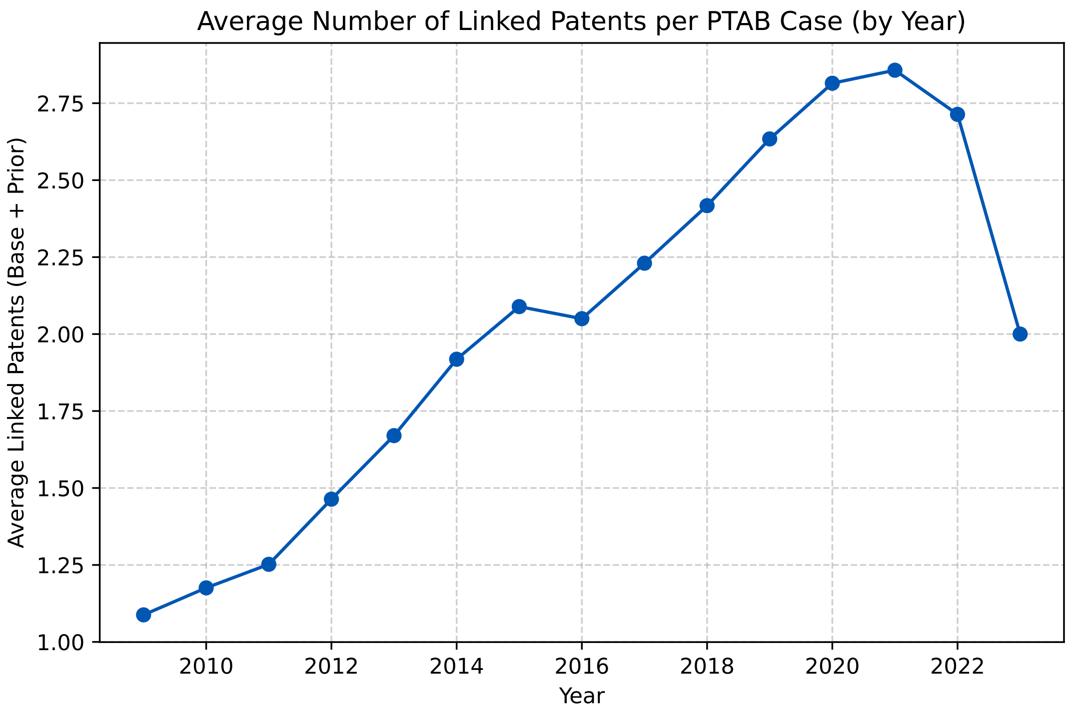


Figure 12: Average number of linked patents per PTAB case by year.

Model	Exact Match	Micro-P	Micro-R	Micro-F1	Macro-P	Macro-R	Macro-F1	HL
Split (Base)								
Claude-Sonnet-4	0.5871	0.7322	0.8589	0.7905	0.5340	0.5735	0.5457	0.0893
Gemini-2.5-pro	0.5874	0.7285	0.8683	0.7923	0.6427	0.7137	0.6630	0.1072
GPT-4o	0.5751	0.7215	0.8633	0.7860	0.6284	0.6997	0.6519	0.1107
GPT-o3	0.5955	0.7404	0.8624	0.7968	0.6567	0.6969	0.6639	0.1036
Solar-pro2	0.5583	0.7072	0.8467	0.7707	0.4988	0.5653	0.5240	0.0989
LLaMA-3.1(8B)	0.1826	0.4512	0.8092	0.5793	0.0920	0.1530	0.1051	0.0659
Mistral(7B)	0.3405	0.5302	0.7126	0.6080	0.1936	0.2650	0.2111	0.0902
Qwen(8B)	0.5561	0.7114	0.8489	0.7741	0.5006	0.5598	0.5251	0.0972
T5(2B)	0.0772	0.2945	0.9265	0.4469	0.2812	0.9118	0.3845	0.5401
Merge								
Claude-Sonnet-4	0.5879	0.7330	0.8602	0.7915	0.5348	0.5745	0.5468	0.0889
Gemini-2.5-pro	0.5810	0.7220	0.8694	0.7889	0.6351	0.7241	0.6625	0.1096
GPT-4o	0.5516	0.6984	0.8726	0.7758	0.6039	0.7129	0.6422	0.1188
GPT-o3	0.5943	0.7375	0.8648	0.7961	0.6535	0.7025	0.6645	0.1043
Solar-pro2	0.5466	0.6919	0.8535	0.7643	0.5817	0.6975	0.6249	0.1240
LLaMA-3.1(8B)	0.1334	0.4408	0.8482	0.5801	0.3689	0.7003	0.4517	0.2892
Mistral(7B)	0.2639	0.4631	0.7617	0.5760	0.1117	0.2013	0.1356	0.0777
Qwen(8B)	0.5322	0.6825	0.8660	0.7634	0.5732	0.6973	0.6255	0.1264
T5(2B)	0.0057	0.2563	0.9643	0.4050	0.2535	0.9624	0.3534	0.6674
Split+Claim								
Claude-Sonnet-4	0.5869	0.7339	0.8589	0.7915	0.5342	0.5707	0.5443	0.0888
Gemini-2.5-pro	0.5911	0.7334	0.8690	0.7955	0.6475	0.7062	0.6632	0.1052
GPT-4o	0.5658	0.7077	0.8759	0.7828	0.6155	0.7127	0.6492	0.1144
GPT-o3	0.5946	0.7393	0.8639	0.7967	0.6550	0.6991	0.6639	0.1038
Solar-pro2	0.5355	0.6808	0.8589	0.7596	0.5736	0.7066	0.6225	0.1281
LLaMA-3.1(8B)	0.1785	0.4587	0.8377	0.5928	0.3477	0.6530	0.4360	0.2710
Mistral(7B)	0.4200	0.5964	0.7820	0.6767	0.2439	0.3113	0.2662	0.0880
Qwen(8B)	0.5631	0.7229	0.8426	0.7782	0.6204	0.6599	0.6353	0.1131
T5(2B)	0.0155	0.3048	0.8931	0.4545	0.0018	0.0052	0.0024	0.0030

Table 10: Results for the Issue Type classification task with 8 evaluation metrics. Exact Match, Micro-P (Micro-Precision), Micro-R (Macro-Recall), Micro-F1 (Micro-F1), Macro-P (Macro-Precision), Macro-R (Macro-Recall), Macro-F1 (Macro-F1) and HL (Hamming Loss) are reported.

Model	Exact Match	Micro-P	Micro-R	Micro-F1	Macro-P	Macro-R	Macro-F1	HL
Split (Base)								
Claude-Sonnet-4	0.4945	0.6038	0.4956	0.5444	0.2499	0.3503	0.2397	0.1012
Gemini-2.5-pro	0.5906	0.8158	0.6003	0.6916	0.2549	0.4277	0.2665	0.0725
GPT-4o	0.6314	0.7004	0.6102	0.6522	0.3177	0.3509	0.2589	0.0882
GPT-o3	0.5302	0.6831	0.5736	0.6236	0.2787	0.2504	0.1940	0.0603
Solar-pro2	0.4293	0.5825	0.6279	0.6179	0.1054	0.2274	0.1014	0.0584
LLaMA-3.1(8B)	0.0000	0.0934	0.1801	0.1230	0.1359	0.3945	0.0843	0.3132
Mistral(7B)	0.0028	0.2043	0.4263	0.2762	0.0100	0.0300	0.0075	0.0211
Qwen(8B)	0.1542	0.1899	0.2039	0.1966	0.1860	0.4106	0.1420	0.2258
T5(2B)	0.0064	0.1508	0.3548	0.2116	0.0030	0.0079	0.0026	0.0064
Merge								
Claude-Sonnet-4	0.7761	0.8924	0.7304	0.8033	0.2105	0.2919	0.2128	0.0364
Gemini-2.5-pro	0.6323	0.9148	0.6194	0.7387	0.3551	0.4168	0.3062	0.0594
GPT-4o	0.6032	0.6525	0.5868	0.6179	0.2419	0.4041	0.2486	0.0984
GPT-o3	0.6459	0.8436	0.6503	0.7344	0.2732	0.2705	0.2160	0.0441
Solar-pro2	0.2531	0.4928	0.6284	0.5524	0.0628	0.1502	0.0620	0.0460
LLaMA-3.1(8B)	0.0000	0.1169	0.2685	0.1629	0.1218	0.3772	0.0882	0.3061
Mistral(7B)	0.0028	0.1984	0.4372	0.2729	0.0050	0.0146	0.0038	0.0112
Qwen(8B)	0.4266	0.4641	0.4427	0.4531	0.1960	0.3699	0.1897	0.1448
T5(2B)	0.0026	0.1105	0.4283	0.1757	0.0035	0.0117	0.0032	0.0099
Split+Claim								
Claude-Sonnet-4	0.2026	0.2920	0.2402	0.2636	0.1838	0.2837	0.1530	0.1364
Gemini-2.5-pro	0.4913	0.6261	0.5394	0.5795	0.2122	0.4493	0.2201	0.1061
GPT-4o	0.0035	0.1206	0.1760	0.1431	0.1806	0.4817	0.1425	0.2856
GPT-o3	0.2477	0.4011	0.4396	0.4194	0.2444	0.2991	0.2109	0.1060
Solar-pro2	0.0041	0.1596	0.2011	0.1780	0.0732	0.2122	0.0485	0.1133
LLaMA-3.1(8B)	0.0001	0.1408	0.3171	0.1950	0.1296	0.3130	0.0923	0.2904
Mistral(7B)	0.0003	0.1154	0.2627	0.1603	0.0070	0.0197	0.0044	0.0185
Qwen(8B)	0.0134	0.0544	0.0606	0.0574	0.1917	0.3804	0.1136	0.2700
T5(2B)	0.0009	0.0912	0.3431	0.1442	0.0051	0.0248	0.0037	0.0206

Table 11: Results for the Board Authorities classification task with 8 evaluation metrics. Exact Match, Micro-P (Micro-Precision), Micro-R (Macro-Recall), Micro-F1 (Micro-F1), Macro-P (Macro-Precision), Macro-R (Macro-Recall), Macro-F1 (Macro-F1) and HL (Hamming Loss) are reported.

Model	Acc	Balanced Acc	Macro-P	Macro-R	Macro-F1	Micro-F1	Weighted-F1
Split (Base)							
Claude-Sonnet-4	0.5658	0.1681	0.1767	0.1569	0.1296	0.5658	0.4854
Gemini-2.5-pro	0.5050	0.1765	0.2473	0.1647	0.1635	0.5050	0.4982
GPT-4o	0.4924	0.1327	0.0944	0.1283	0.0997	0.4924	0.4709
GPT-o3	0.5918	0.1519	0.3295	0.1519	0.1639	0.5918	0.5541
Solar-pro2	0.5369	0.1225	0.1509	0.1143	0.0779	0.5369	0.3923
LLaMA-3.1(8B)	0.4364	0.0927	0.0841	0.0927	0.0767	0.4364	0.4006
Mistral(7B)	0.1241	0.0603	0.0461	0.0422	0.0251	0.1241	0.1284
Qwen(8B)	0.4793	0.1106	0.1057	0.1032	0.0977	0.4793	0.4457
T5(2B)	0.0419	0.0917	0.0501	0.0583	0.0142	0.0419	0.0617
Merge							
Claude-Sonnet-4	0.5590	0.1614	0.1872	0.1509	0.1129	0.5590	0.4320
Gemini-2.5-pro	0.5114	0.1925	0.1661	0.1685	0.1443	0.5114	0.5036
GPT-4o	0.4592	0.1257	0.1381	0.1173	0.0912	0.4592	0.4353
GPT-o3	0.6086	0.1580	0.3244	0.1580	0.1683	0.6086	0.5682
Solar-pro2	0.5420	0.1248	0.1790	0.1164	0.0804	0.5420	0.3932
LLaMA-3.1(8B)	0.5036	0.0650	0.0536	0.5036	0.0696	0.3971	0.0676
Mistral(7B)	0.1265	0.0364	0.0229	0.1265	0.0572	0.1249	0.0407
Qwen(8B)	0.4266	0.1096	0.0707	0.0768	0.0698	0.4266	0.4264
T5(2B)	0.0191	0.0463	0.0092	0.0191	0.0794	0.0270	0.0437
Split+Claim							
Claude-Sonnet-4	0.5620	0.1616	0.1725	0.1509	0.1272	0.5620	0.4842
Gemini-2.5-pro	0.4908	0.1518	0.1832	0.1417	0.1433	0.4908	0.4854
GPT-4o	0.3804	0.1275	0.0944	0.1190	0.0892	0.3804	0.3581
GPT-o3	0.5884	0.1610	0.3241	0.1610	0.1692	0.5884	0.5538
Solar-pro2	0.5373	0.0762	0.0993	0.0762	0.0608	0.5373	0.3966
LLaMA-3.1(8B)	0.4125	0.0664	0.0830	0.0664	0.0642	0.4125	0.3938
Mistral(7B)	0.1209	0.0536	0.0533	0.0417	0.0295	0.1209	0.1205
Qwen(8B)	0.4368	0.0872	0.0831	0.0814	0.0794	0.4368	0.4364
T5(2B)	0.0225	0.1699	0.1655	0.1322	0.0436	0.0225	0.0168

Table 12: Results for the Subdecision (Fine-grained) classification task with 7 evaluation metrics. Acc (Accuracy), Balanced Acc (Balanced Accuracy), Macro-P (Macro-Precision), Macro-R (Macro-Recall), Macro-F1 (Macro-F1), Micro-F1 (Micro-F1), and Weighted-F1 are reported. In single-label multiclass classification, Accuracy and Micro-F1 coincide because both measure the proportion of correctly classified samples.

Model	Acc	Balanced Acc	Macro-P	Macro-R	Macro-F1	Micro-F1	Weighted-F1
Split (Base)							
Claude-Sonnet-4	0.5652	0.2108	0.2865	0.2105	0.2116	0.5625	0.4900
Gemini-2.5-pro	0.5063	0.2270	0.3351	0.2270	0.2366	0.5063	0.4927
GPT-4o	0.5045	0.1988	0.2350	0.1988	0.2037	0.5045	0.4863
GPT-o3	0.5863	0.2099	0.3802	0.2099	0.2126	0.5863	0.5511
Solar-pro2	0.5389	0.1621	0.2303	0.1621	0.1356	0.5389	0.3929
LLaMA-3.1(8B)	0.4764	0.1635	0.1770	0.1635	0.1551	0.4764	0.4024
Mistral(7B)	0.0726	0.1590	0.1725	0.1590	0.0758	0.0726	0.0994
Qwen(8B)	0.4733	0.1739	0.2298	0.1739	0.1692	0.4733	0.4404
T5(2B)	0.0254	0.2177	0.1446	0.2177	0.0499	0.0254	0.0146
Merge							
Claude-Sonnet-4	0.5607	0.1952	0.2872	0.1952	0.1788	0.5607	0.4456
Gemini-2.5-pro	0.5119	0.2390	0.2771	0.2390	0.2381	0.5119	0.5001
GPT-4o	0.4972	0.1794	0.2635	0.1794	0.1820	0.4972	0.4638
GPT-o3	0.6020	0.2101	0.3814	0.2101	0.2125	0.6020	0.5631
Solar-pro2	0.5423	0.1631	0.2598	0.1631	0.1390	0.5423	0.3967
LLaMA-3.1(8B)	0.5229	0.1515	0.1908	0.1515	0.1253	0.5229	0.3922
Mistral(7B)	0.0823	0.1552	0.1685	0.1552	0.0821	0.0823	0.1168
Qwen(8B)	0.4163	0.1760	0.2219	0.1760	0.1761	0.4163	0.4223
T5(2B)	0.0234	0.2238	0.1593	0.2238	0.0446	0.0234	0.0092
Split+Claim							
Claude-Sonnet-4	0.5639	0.2011	0.2646	0.2011	0.2018	0.5637	0.4889
Gemini-2.5-pro	0.4915	0.2142	0.3409	0.2142	0.2111	0.4915	0.4840
GPT-4o	0.3046	0.1633	0.1982	0.1633	0.1206	0.3046	0.2027
GPT-o3	0.5783	0.2099	0.5012	0.2099	0.2068	0.5783	0.5426
Solar-pro2	0.5364	0.1514	0.1819	0.1514	0.1210	0.5364	0.3977
LLaMA-3.1(8B)	0.4741	0.1447	0.1505	0.1447	0.1259	0.4741	0.3909
Mistral(7B)	0.0587	0.1568	0.2767	0.1568	0.0549	0.0587	0.0721
Qwen(8B)	0.4605	0.1660	0.2083	0.1660	0.1655	0.4605	0.4439
T5(2B)	0.0136	0.0440	0.0376	0.0246	0.0053	0.0136	0.0142

Table 13: Results for the Subdecision (Coarse-grained) classification task with 7 evaluation metrics. Acc (Accuracy), Balanced Acc (Balanced Accuracy), Macro-P (Macro-Precision), Macro-R (Macro-Recall), Macro-F1 (Macro-F1), Micro-F1 (Micro-F1), and Weighted-F1 are reported. In single-label multiclass classification, Accuracy and Micro-F1 coincide because both measure the proportion of correctly classified samples.

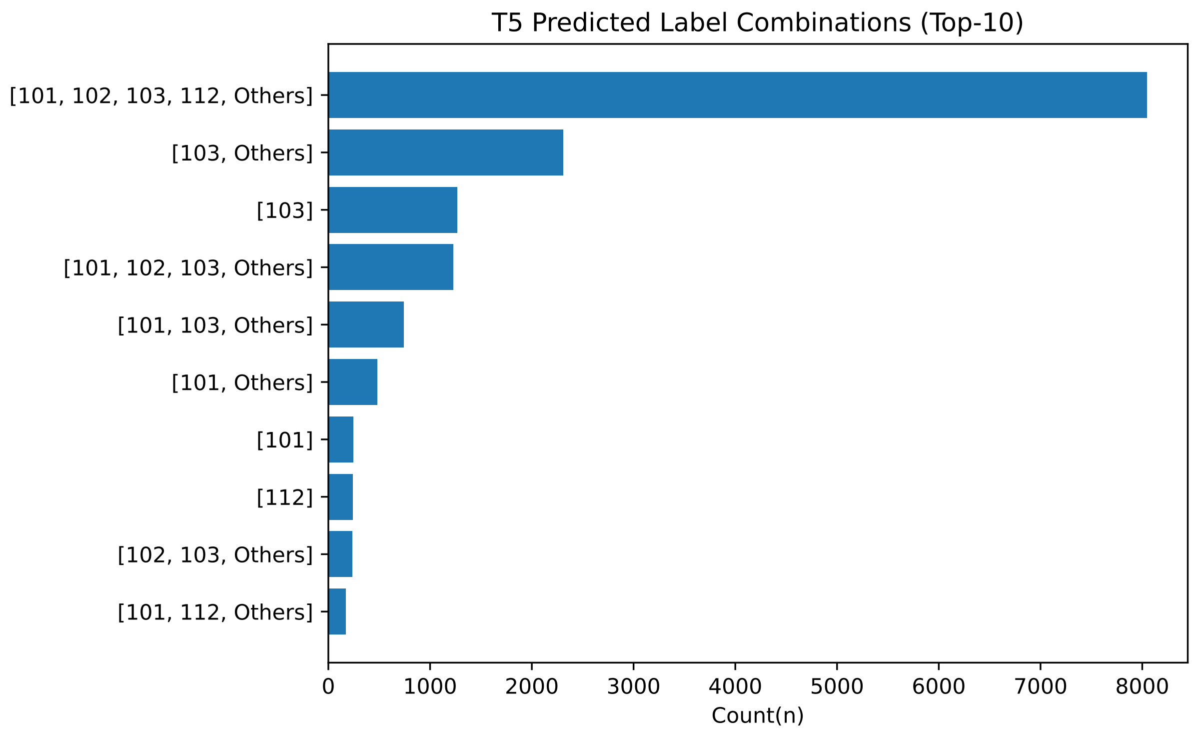


Figure 13: Top-10 predicted IssueType label combinations by T5 under Split (Base).

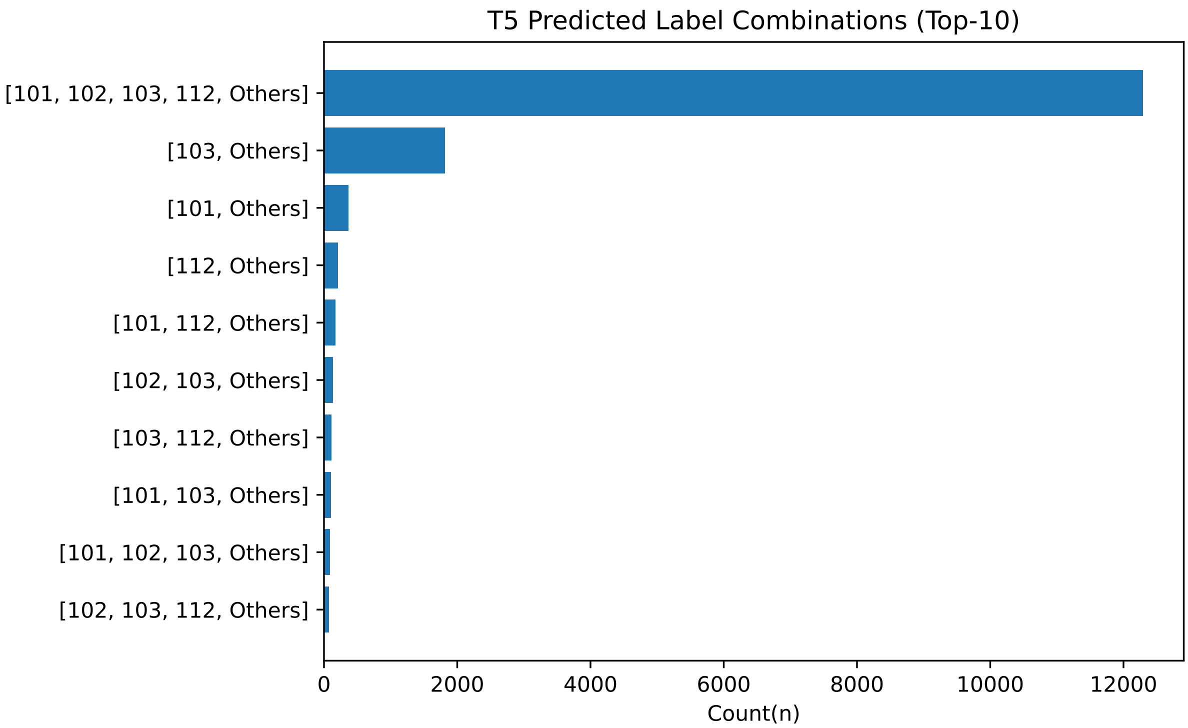


Figure 14: Top-10 predicted IssueType label combinations by T5 under Merge.

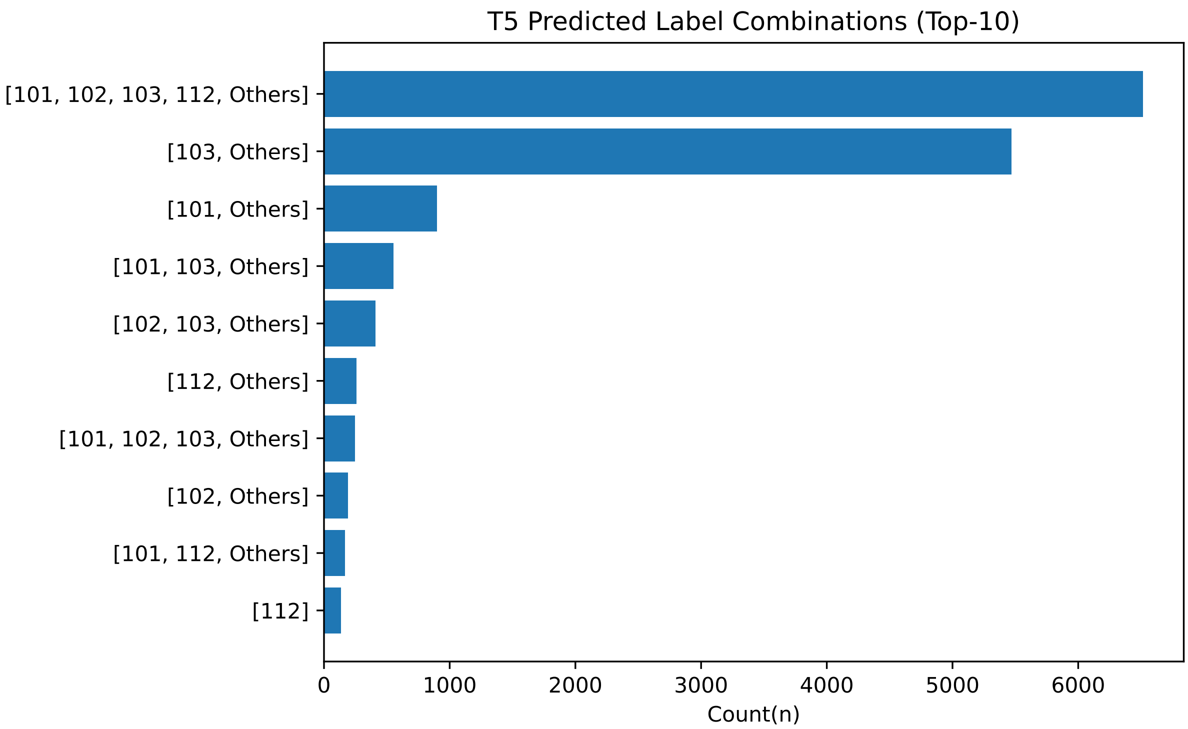


Figure 15: Top-10 predicted IssueType label combinations by T5 under Split+Claim.

Label	Definition
101	Patent eligibility (Subject-matter eligibility)
102	Novelty
103	Non-obviousness
112	Specification requirements (Written description / Enablement / Definiteness)
Others	All other issues (e.g., OTDP, priority, new matter, reissue, design)

Table 14: Labels used in the Issue Type classification task and their definitions. The dictionary was also provided within the classification prompt so that the LLM could reference these descriptions while reasoning about applicable statutory issues.

Label	Definition
37 CFR 41.50	General framework for PTAB decisions/actions in ex parte appeals (affirm/reverse/remand, new ground, additional briefing, time extensions).
37 CFR 41.50(a)	Merits decision on appeal (affirm/reverse/remand) and post-decision options.
37 CFR 41.50(b)	Board-designated New Ground of Rejection (non-final for judicial review); appellant may request rehearing or reopen prosecution.
37 CFR 41.50(c)	Procedure to address an undesignated new ground via rehearing request.
37 CFR 41.50(d)	Authority to order additional briefing/information; non-compliance may lead to dismissal.
37 CFR 41.50(f)	Rules for extensions of time for replies in ex parte appeals.
37 CFR 1.131	Pre-AIA affidavit/declaration of prior invention (swear behind) to overcome prior art.
37 CFR 1.132	Affidavits/declarations traversing rejections or objections (e.g., objective evidence, secondary considerations).
35 USC 251	Reissue of defective patents (broadening/narrowing; correction of error).
35 USC 161	Plant patent requirements (asexual reproduction, cultivar/variety).

Table 15: Labels used in the Board Authorities classification task and their definitions. This dictionary was also embedded in the classification prompt, so that the LLM could reference these descriptions while reasoning and assigning labels.

ID	Label	Variants / Mappings
1	Affirmed	affirmed
2	Affirmed with New Ground of Rejection	affirmed with new ground of rejection affirmed with new ground(s) of rejection affirmed w/ new ground(s) of rejection
3	Affirmed-in-Part	affirmed-in-part affirmed in part affirmed-in part affirmed/reversed in part reversed/affirmed in part reversed in-part reversed in part reversed-in part
4	Affirmed-in-Part and Remanded	affirmed-in-part and remanded affirmed-in-part and remanded with new ground of rejection
5	Affirmed-in-Part with New Ground of Rejection	affirmed-in-part with new ground of rejection affirmed-in-part with new ground(s) of rejection affirmed-in-part w/ new ground(s) of rejection
6	Reversed	reversed
7	Reversed with New Ground of Rejection	reversed with new ground of rejection reversed with new ground(s) of rejection reversed w/ new ground(s) of rejection
8	Reexam affirmed	reexam affirmed
9	Reexam Affirmed-in-part	reexam affirmed-in-part
10	Reexam Affirmed-in-part with New Ground of Rejection	reexam affirmed-in-part with new ground of rejection
11	Reexam reversed	reexam reversed
12	Inter Partes Reexam Affirmed	inter partes reexam affirmed
13	Inter Partes Reexam Affirmed-in-part	inter partes reexam affirmed-in-part
14	Inter Partes Reexam Reversed	inter partes reexam reversed
15	Inter Partes Reexam New Ground of Rejection	inter partes reexam new ground of rejection
16	Inter partes reexam rehearing decision is a new decision	inter partes reexam rehearing decision is a new decision
17	Affirmed-in-Part and Remanded with New Ground of Rejection	affirmed-in-part and remanded with new ground of rejection
18	Reversed and Remanded	reversed and remanded
19	Vacated	vacated vacated with new ground of rejection vacated-in-part with new ground of rejection vacated/remanded vacated and remanded vacatur vacated in part vacate and remand
20	Granted	granted granted (petitioner) granted (patent owner) granted-in-part granted-in-part (petitioner) granted-in-part (patent owner)
21	Denied	denied denied (petitioner) denied (patent owner)
22	Rehearing Decision - Granted	rehearing decision - granted Rehearing Decision Ãć Grante rehearing decision - granted rehearing decision-granted
23	Reexam rehearing decision final and appealable	reexam rehearing decision final and appealable

Table 16: Normalized subdecision fine categories (excluding **Others**) and their variants. Each variant was normalized by converting raw labels to lowercase and stripping leading/trailing whitespace before mapping them to a canonical label. The canonical labels are further incorporated into the classification prompt, enabling the LLM to consult these standardized categories during subdecision reasoning.

Label	Variants / Mappings
	dismissed
	dismissal
	voluntarily dismissed
	dismissed before institution
	dismissed after institution
	decision on rehearing
	decision on petition
	rehearing decision
	Rehearing Decision Ãć Granted w/ New Ground of Rejection
	rehearing decision - granted with new ground of rejection
	Rehearing Decision Ãć Denied
	rehearing decision - denied
	Rehearing Decision Ãć Denied w/ New Ground of Rejection
	rehearing decision - denied with new ground of rejection
	Rehearing Decision Ãć Granted-in-Part
	rehearing decision - granted-in-part
	remand
	administrative remand
	affirmed and remanded
	reverse and remanded with new ground of rejection
	panel remand
	panel remand with new ground of rejection
Others	remanded-in part
	institution granted
	institution granted (joined)
	institution denied
	decision on petition - denied
	settlement
	settlement before institution
	settlement after institution
	settled before institution
	settled after institution
	termination
	terminated
	termination before institution
	termination after institution
	request for adverse judgment before institution
	request for adverse judgment after institution
	institution-rehearing hybrid
	po rehearing request granted on institution decision granted (trial denied)
	petitioner's rehearing request granted on institution decision denied (reinstituted)
	final decision
	final written decision
	final written decision on cafc remand
	subsequent final written decision after rehearing
	subsequent decision
	judgment
	adverse judgment
	decision on motion
	order
	order on rehearing

Table 17: Variants mapped to Others. The Others category serves as a residual class, collecting normalized raw labels that did not align with any of the explicit subdecision fine categories.

ID	Label	Variants / Mappings
1	Affirmed	affirmed
2	Affirmed with New Ground of Rejection	affirmed with new ground of rejection affirmed with new ground(s) of rejection affirmed w/ new ground(s) of rejection
3	Affirmed-in-Part	affirmed-in-part affirmed in part affirmed-in part affirmed/reversed in part reversed/affirmed in part reversed in-part reversed in part reversed-in part
4	Affirmed-in-Part with New Ground of Rejection	affirmed-in-part with new ground of rejection affirmed-in-part with new ground(s) of rejection affirmed-in-part w/ new ground(s) of rejection
5	Reversed	reversed
6	Reversed with New Ground of Rejection	reversed with new ground of rejection reversed with new ground(s) of rejection reversed w/ new ground(s) of rejection

Table 18: Normalized subdecision coarse categories (excluding **Others**) and their variants. Each variant was normalized by converting raw labels to lowercase and stripping leading/trailing whitespace before mapping them to a canonical category. The canonical labels are further incorporated into the classification prompt, enabling the LLM to consult these standardized categories during subdecision reasoning.

Label	Variants / Mappings
Others	reexam affirmed
	inter partes reexam affirmed
	reexam affirmed-in-part
	inter partes reexam affirmed-in-part
	reexam affirmed-in-part with new ground of rejection
	reexam reversed
	inter partes reexam reversed
	inter partes reexam new ground of rejection
	reexam rehearing decision final and appealable
	inter partes reexam rehearing decision is a new decision
	granted
	granted (petitioner)
	granted (patent owner)
	granted-in-part
	granted-in-part (petitioner)
	granted-in-part (patent owner)
	denied
	denied (petitioner)
	denied (patent owner)
	dismissed
	dismissal
	voluntarily dismissed
	dismissed before institution
	dismissed after institution
	decision on rehearing
	decision on petition
	rehearing decision
	Rehearing Decision Ãć Granted
	rehearing decision - granted
	rehearing decision-granted
	Rehearing Decision Ãć Granted w/ New Ground of Rejection
	rehearing decision - granted with new ground of rejection
	Rehearing Decision Ãć Denied
	rehearing decision - denied
	Rehearing Decision Ãć Denied w/ New Ground of Rejection
	rehearing decision - denied with new ground of rejection
	Rehearing Decision Ãć Granted-in-Part
	rehearing decision - granted-in-part
	remand
	administrative remand
	affirmed-in-part and remanded
	affirmed-in-part and remanded with new ground of rejection
	affirmed and remanded
	reversed and remanded
	reverse and remanded with new ground of rejection
	panel remand
	panel remand with new ground of rejection
	remanded-in part
	vacated
	vacated with new ground of rejection
	vacated-in-part with new ground of rejection
	vacated/remanded
	vacated and remanded
	vacatur
	vacated in part
	vacate and remand
	institution granted
	institution granted (joined)
	institution denied
	decision on petition - denied
	settlement
	settlement before institution
	settlement after institution
	settled before institution
	settled after institution
	termination
	terminated
	termination before institution
	termination after institution
	request for adverse judgment before institution
	request for adverse judgment after institution
	institution-rehearing hybrid
	po rehearing request granted on institution decision granted (trial denied)
	petitioner’s rehearing request granted on institution decision denied (reinstituted)
	final decision
	final written decision
	final written decision on cafc remand
	subsequent final written decision after rehearing
	subsequent decision
	judgment
	adverse judgment
	decision on motion
	order
	order on rehearing

Table 19: Variants mapped to Others. The Others category serves as a residual class, collecting normalized raw labels that did not align with any of the explicit subdecision coarse categories.

Figure 16: Heatmaps of model performance on the Issue Type classification task under the Split (Base) input setting. Each subplot visualizes the distribution of predicted versus true labels across models.

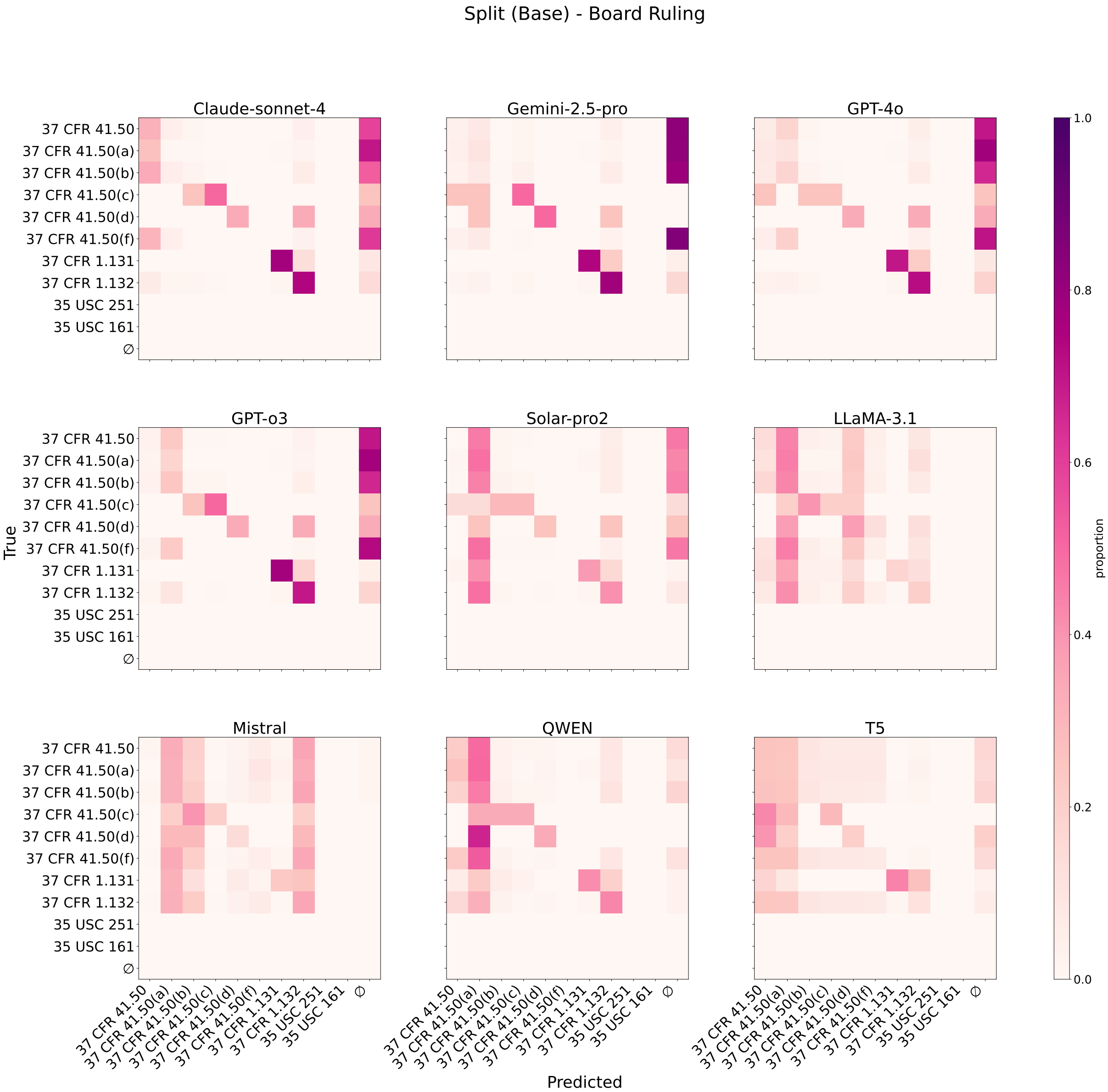


Figure 17: Heatmaps of model performance on the Board Authorities classification task under the Split (Base) input setting. Each subplot visualizes the distribution of predicted versus true labels across models.

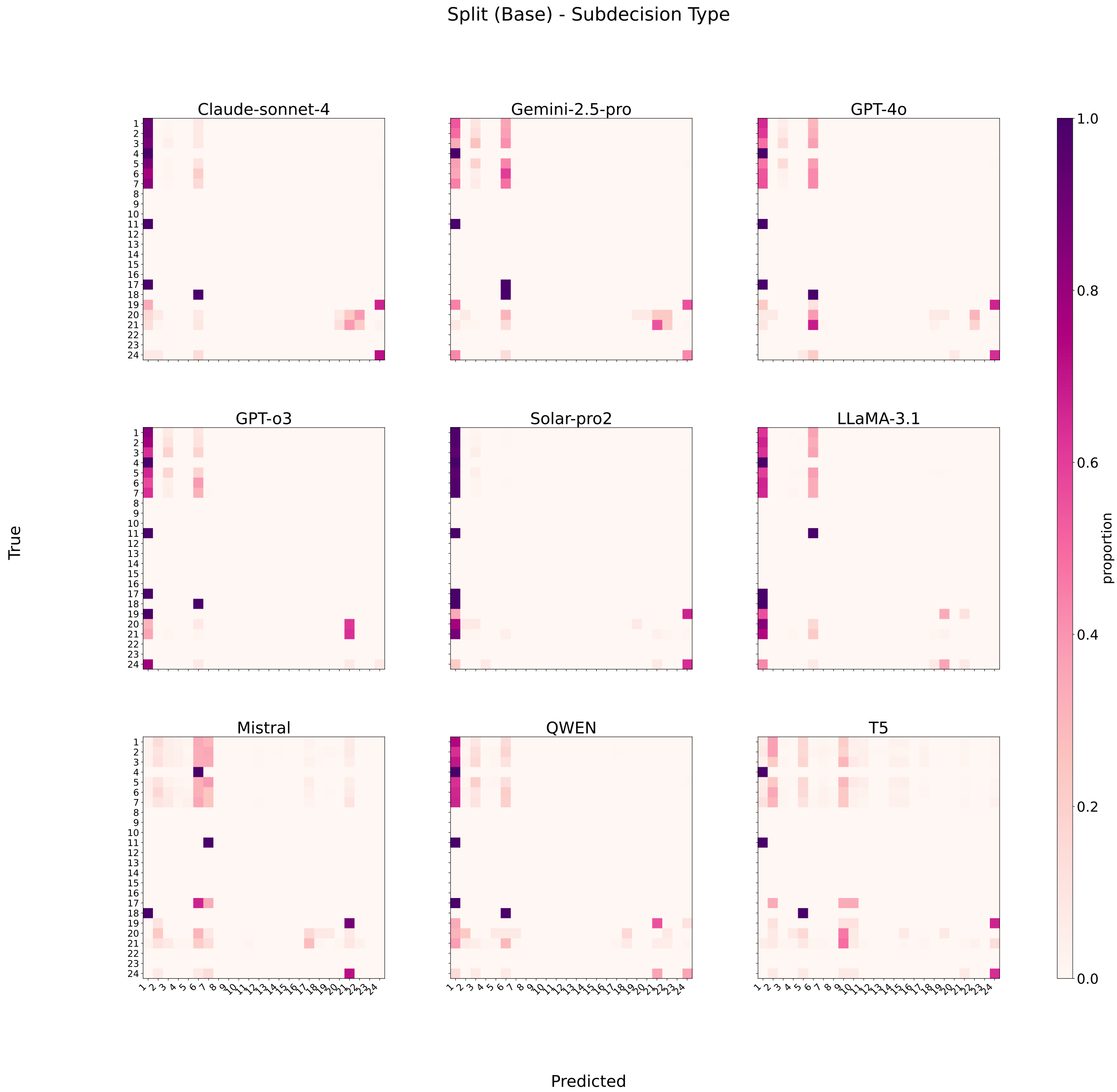


Figure 18: Heatmaps of model performance on the Subdecision (Fine-grained) classification task under the Split (Base) input setting. Each subplot visualizes the distribution of predicted versus true labels across models. The numerical indices on the axes correspond to the canonical labels defined in Table 16, where each index maps to a specific subdecision category.

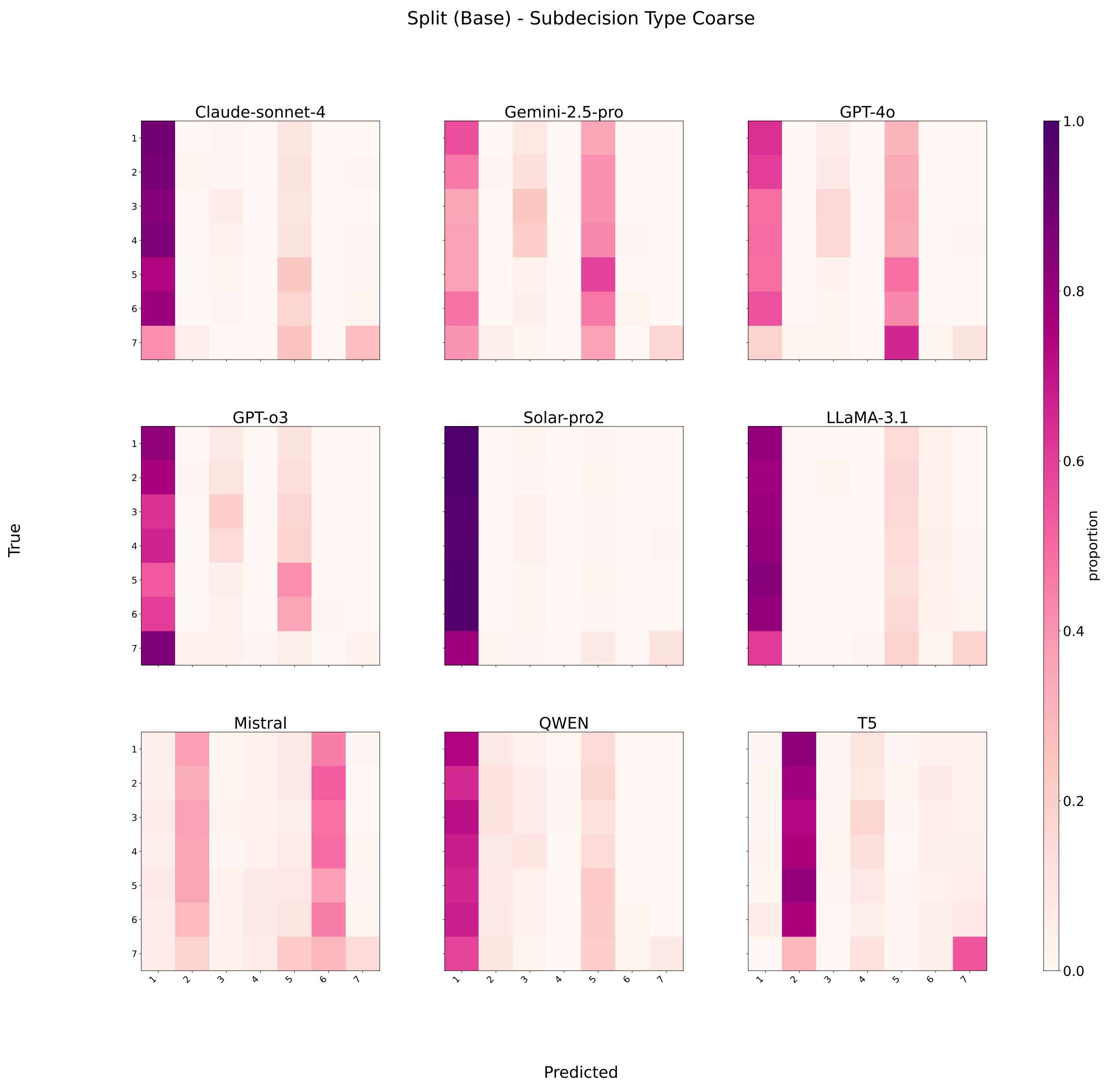


Figure 19: Heatmaps of model performance on the Subdecision (Coarse-grained) classification task under the Split (Base) input setting. Each subplot visualizes the distribution of predicted versus true labels across models. The numerical indices on the axes correspond to the canonical labels defined in Table 18, where each index maps to a specific subdecision category.

Figure 20: Heatmaps of model performance on the Issue Type classification task under the Merge input setting. Each subplot visualizes the distribution of predicted versus true labels across models.

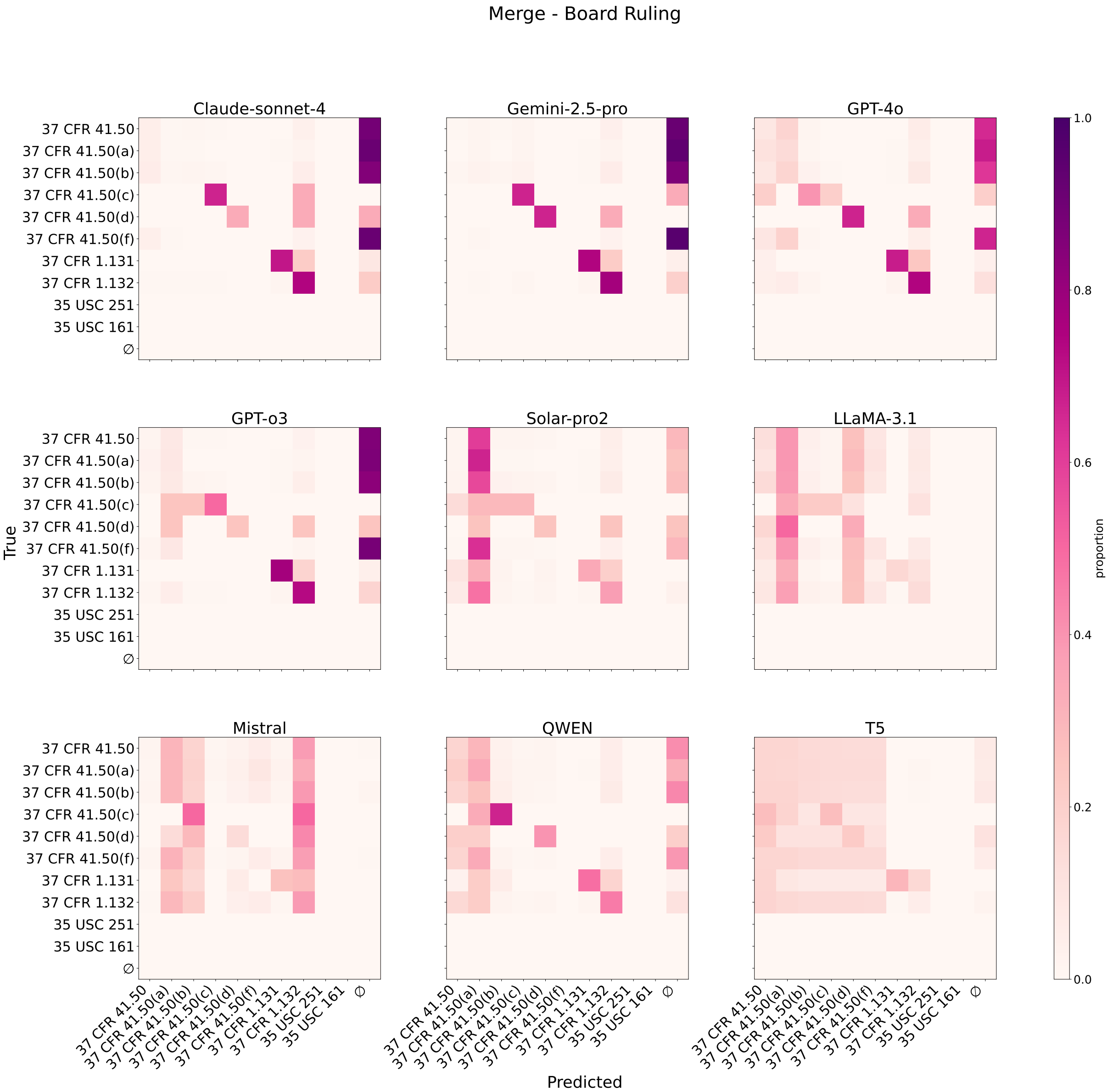


Figure 21: Heatmaps of model performance on the Board Authorities classification task under the Merge input setting. Each subplot visualizes the distribution of predicted versus true labels across models.

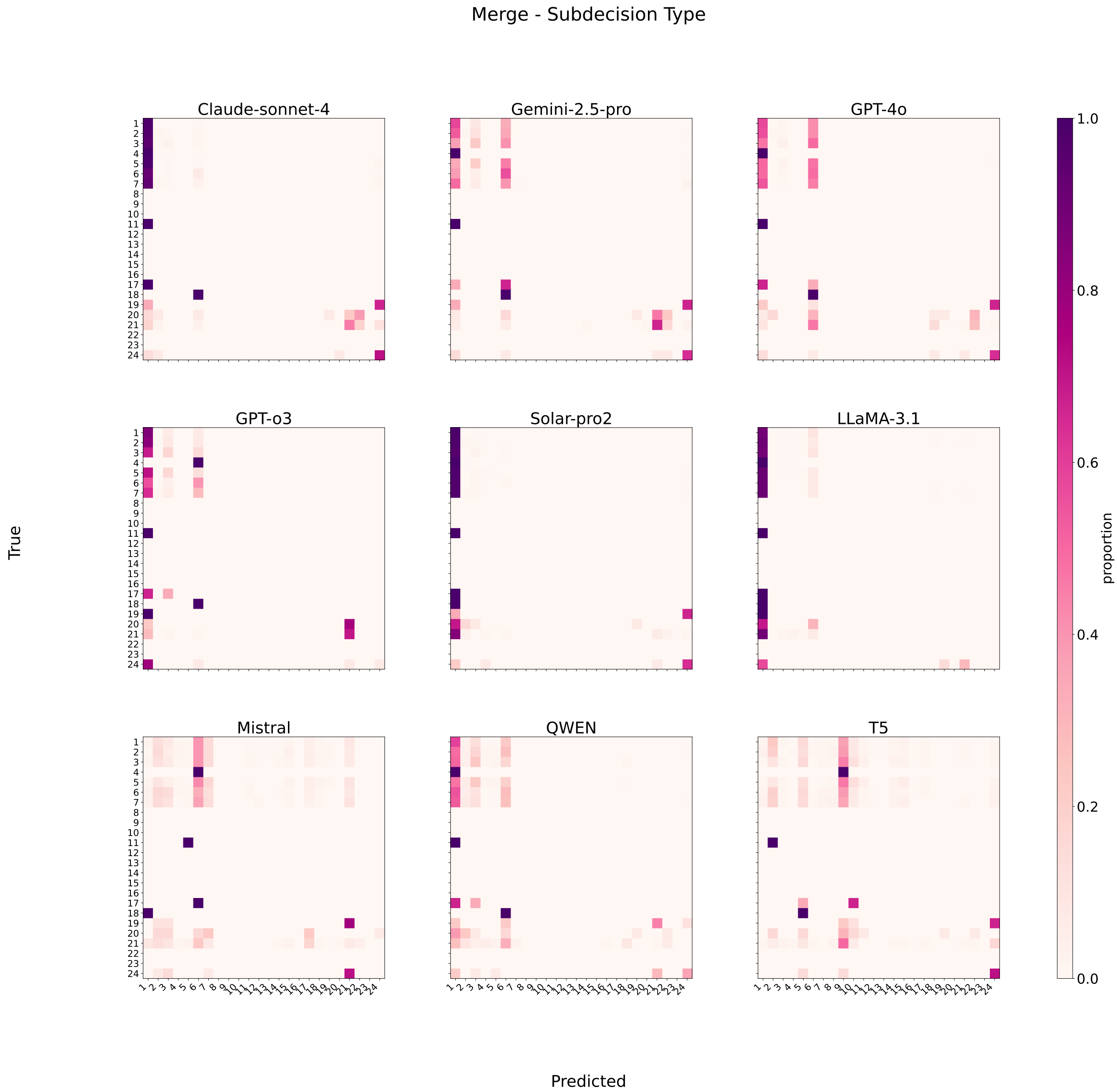


Figure 22: Heatmaps of model performance on the Subdecision (Fine-grained) classification task under the Merge input setting. Each subplot visualizes the distribution of predicted versus true labels across models. The numerical indices on the axes correspond to the canonical labels defined in Table 16, where each index maps to a specific subdecision category.

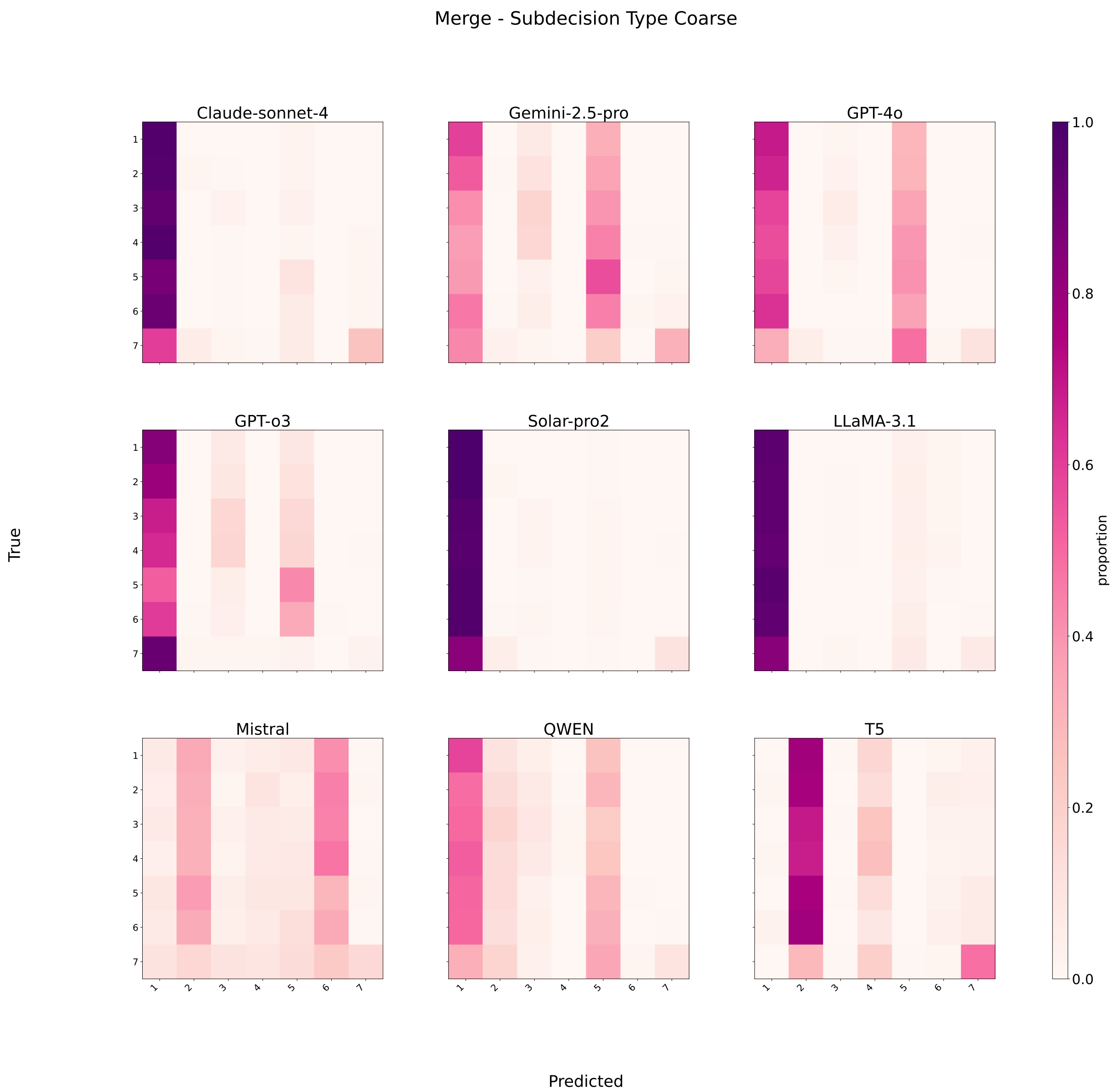


Figure 23: Heatmaps of model performance on the Subdecision (Coarse-grained) classification task under the Merge input setting. Each subplot visualizes the distribution of predicted versus true labels across models. The numerical indices on the axes correspond to the canonical labels defined in Table 18, where each index maps to a specific subdecision category.

Figure 24: Heatmaps of model performance on the Issue Type classification task under the Split+Claim input setting. Each subplot visualizes the distribution of predicted versus true labels across models.

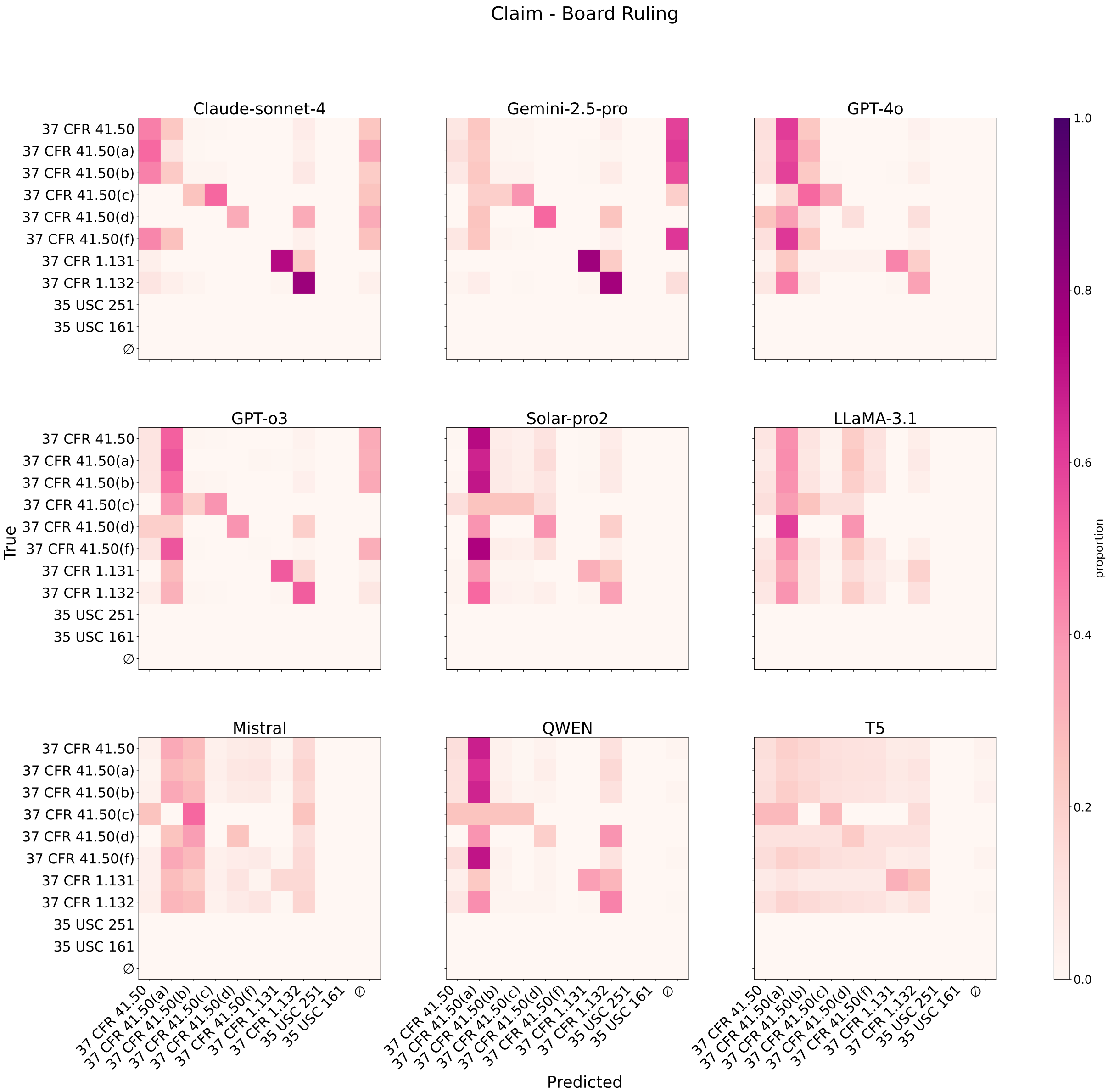


Figure 25: Heatmaps of model performance on the Board Authorities classification task under the Split+Claim input setting. Each subplot visualizes the distribution of predicted versus true labels across models.

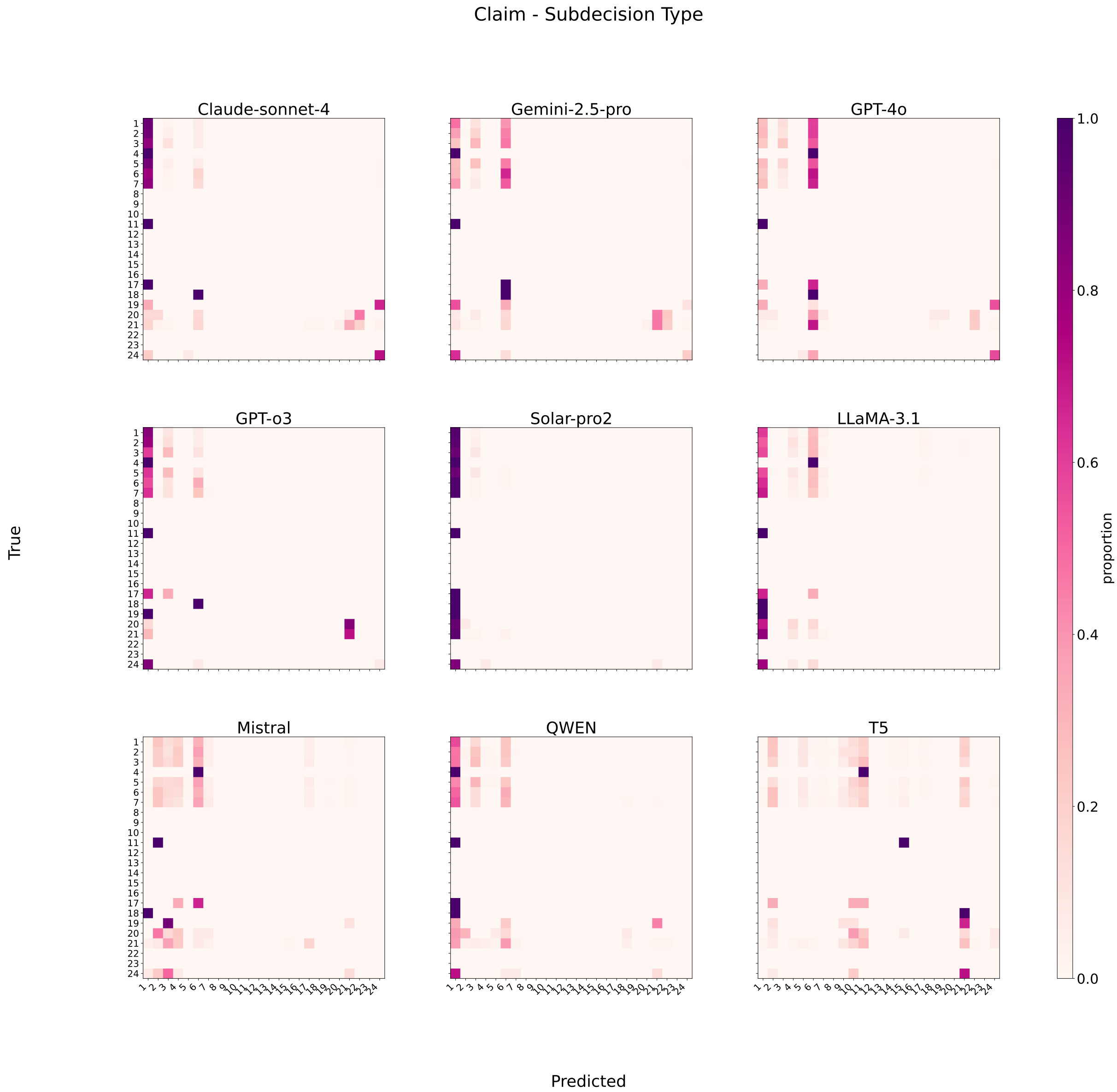


Figure 26: Heatmaps of model performance on the Subdecision (Fine-grained) classification task under the Split+Claim input setting. Each subplot visualizes the distribution of predicted versus true labels across models. The numerical indices on the axes correspond to the canonical labels defined in Table 16, where each index maps to a specific subdecision category.

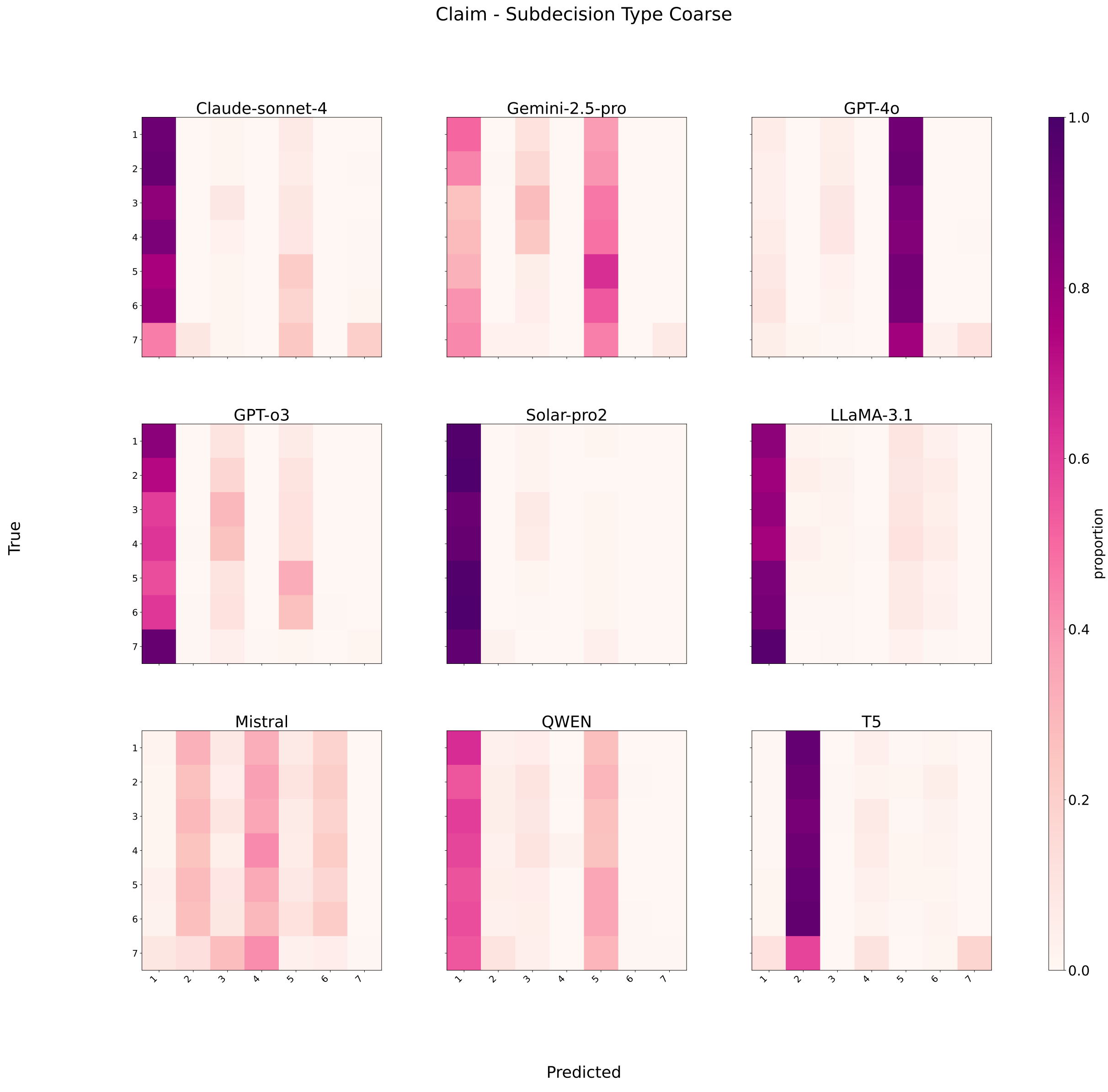


Figure 27: Heatmaps of model performance on the Subdecision (Coarse-grained) classification task under the Split+Claim input setting. Each subplot visualizes the distribution of predicted versus true labels across models. The numerical indices on the axes correspond to the canonical labels defined in Table 18, where each index maps to a specific subdecision category.