BEYOND STRUCTURE RECOGNITION: BENCHMARKING PATENTABILITY ASSESSMENT IN DRUG DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Assessing patentability in drug discovery requires more than recognizing chemical structures: models must interpret the scope jointly defined by structural drawings and patent claims, and Markush claims make this task particularly challenging, as a shared scaffold, variable substituents, and textual restrictions simultaneously define a family of compounds. In this paper, we introduce **`MarkPatentBench`**, a multimodal benchmark for assessing the patentability of small-molecule drug claims, curated from patent examination records of the European and U.S. patent office websites. The benchmark contains claim text and Markush chemical structure images with source-linked examination reference documents. It supports overall patentability prediction and assessment generation across five core dimensions: novelty, inventive step, support by the description, claim clarity, and patent-eligible subject matter. Human domain experts review the curated datasets for data quality and annotation reliability. By evaluating frontier large language and vision-language models under Naive context, evidence-augmented, and agentic configurations, we found that current models suffer from inherent limitations, including overlooking claim limitations, inferring obviousness from scaffold familiarity, and treating missing information as evidence of a substantive defect. **`MarkPatentBench`** provides a testbed for investigating whether models can move beyond structure recognition to preserve the claimed chemical scope and support their patentability assessments with applicable evidence.

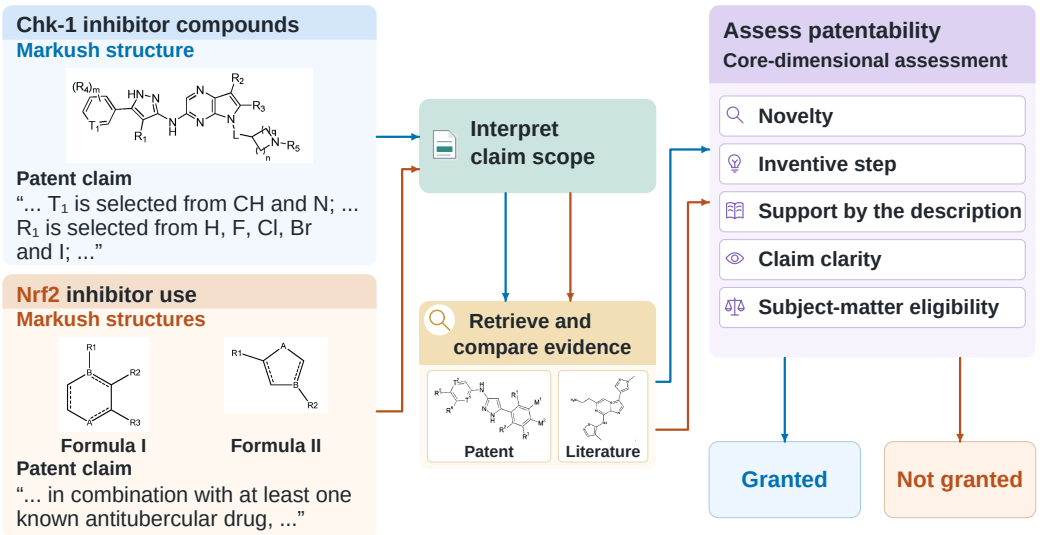


Figure 1: Illustration of the patentability assessment in drug discovery. The process requires joint reasoning over textual claims and Markush chemical structure images, alongside comparison with prior art (e.g., patents and literature). We focus on two primary objectives: core-dimensional assessment and grant-outcome prediction.

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.

1 INTRODUCTION

Patentability assessment connects molecular design with the claimed scope of chemical inventions. In drug discovery, assessing patent records requires more than recognizing a chemical structure: models must interpret the scope jointly defined by structural drawings and patent claims (Papadatos et al., 2016). This task is particularly challenging for Markush claims, where a shared scaffold, variable substituents, and textual restrictions simultaneously define a vast family of compounds. Evaluating these boundaries requires assessing five core dimensions: novelty, inventive step (non-obviousness), support by the description, claim clarity, and patent-eligible subject matter (European Patent Office, a; 2026). These dimensions connect chemical evidence with assessment questions that biological activity or scaffold recognition alone cannot resolve.

Reliable assessment therefore needs faithful multimodal inputs and traceable examination evidence. While optical structure recognition, chemical entity extraction, and recent Markush systems provide access to basic structural inputs (Filippov & Nicklaus, 2009; Zhai et al., 2019; Morin et al., 2025), evaluating a patent requires moving beyond structure recognition to connect structural resemblance with claim scope and examination grounds (Lim et al., 2025). Existing molecular-coverage and examination-reasoning benchmarks address parts of this problem (Shi et al., 2025; Lim et al., 2025), but they do not jointly cover graphical Markush claims and the five core dimensions considered here.

To bridge this gap, we introduce **`MarkPatentBench`**, a novel benchmark for assessing the patentability of small-molecule drug claims. Curated from publicly accessible European and U.S. examination records, the benchmark is designed for overall grant-outcome prediction and core-dimensional assessment generation. Models must process the claim text alongside one or more Markush chemical structure images and source-linked examination reference documents (Figure 1). To ensure data quality and annotation reliability, human drug-discovery experts rigorously review the curated datasets through a source-linked interface.

Our contributions are threefold. (1) We introduce a novel testbed that connects graphical chemical claims, core-dimensional assessment, and grant outcomes, validated by expert quality review. (2) We systematically evaluate frontier large language and vision-language models under Naive, evidence-augmented, and agentic configurations. (3) We uncover inherent limitations in current models when reasoning about complex chemical claims: they frequently overlook textual claim limitations, erroneously infer obviousness from scaffold familiarity, and treat missing information as evidence of a substantive defect. MarkPatentBench provides a foundation for investigating how models can preserve claimed chemical scope and support their patentability assessments with applicable evidence.

2 RELATED WORK

Chemical Patents and Assessment Benchmarks. OSRA extracts chemical drawings (Filippov & Nicklaus, 2009), SureChEMBL connects compounds with patents (Papadatos et al., 2016), and chemical named-entity recognition addresses patent language (Zhai et al., 2019). Recent systems further interpret graphical Markush representations (Morin et al., 2025; Strohmeyer et al., 2026). Complementary patent benchmarks study application outcomes (HUPD (Suzgun et al., 2022)), claim–prior-art correspondence (PatentMatch (Risch et al., 2020)), examination reasoning (PANORAMA (Lim et al., 2025)), and claim revision (Patent-CR (Jiang et al., 2025)). SciAssess includes Markush subtasks (Cai et al., 2025), while MolPatent-240 tests molecular coverage through PatentFinder (Shi et al., 2025). Table 1 contrasts these scopes with our core-dimensional assessment of graphical chemical claims. Human review denotes task-relevant annotation or validation; our experts assess curated material quality rather than adjudicate model outputs.

Evidence Retrieval and Assessment. Retrieval-augmented generation supplies external evidence (Lewis et al., 2020). ColPali represents document images for retrieval (Faysse et al., 2025); ViDoRAG and VLD-RAG use iterative retrieval and verification (Wang et al., 2025; Kim, 2026); EvoCtx considers whether to retrieve or refine reasoning within the current context (Chen et al., 2026). Verbal feedback and instruction refinement can guide subsequent inference (Shinn et al., 2023; Fernando et al., 2024). We adapt these principles to chemical patent assessment, preserving source-page identities and claim restrictions, without claiming full reproductions of the cited systems.

Table 1: Comparison of different patent-related benchmarks. N: novelty; I: inventive step; S: support by the description; C: claim clarity; E: subject-matter eligibility.

Benchmark	Task	Size	Structure	Evidence	Assessment	Drug discovery	Human review
PatentMatch (Risch et al., 2020)	Prior-art matching	6,259,703 pairs	✗	✓	N proxy	✗	✓
PANORAMA (Lim et al., 2025)	Examination reasoning	8,143 records	✗	✓	N, I	✗	✓
Patent-CR (Jiang et al., 2025)	Claim revision	22,606 pairs	✗	✗	–	✗	✓
SciAssess (Cai et al., 2025)	Chemical subtasks	87 examples	✓	✗	–	✓	✓
IP5-M (Strohmeyer et al., 2026)	Structure recognition	1,000 images	✓	✗	–	✓	✓
MolPatent-240 (Shi et al., 2025)	Molecular coverage	240 pairs	✓	✗	–	✓	✓
`MarkPatentBench` (Ours)	Patentability assessment	851 EP applications 308 US applications	✓	✓	N, I, S, C, E	✓	✓

3 MarkPatentBench

3.1 Task Definition and Assessment Targets

Assessing a chemical patent requires interpreting the scope jointly defined by claim text and Markush drawings, together with relevant cited evidence. To evaluate this process, `MarkPatentBench` pairs a review across the core patentability dimensions with a prediction of the recorded grant outcome.

Formally, for case i, let c_i be the selected claim, g_i its frozen chemical images, and E_i pages selected from optional cited documents. The model receives

$$x_i = (c_i, g_i, E_i), \qquad f_\theta(x_i) = (\hat{y}_i, \{a_{i,d}\}_{d\in\mathcal{K}}, u_i), \tag{1}$$

where $\mathcal{K} = \{\text{novelty}, \text{inventive step}, \text{support}, \text{clarity}, \text{eligibility}\}$ and u_i denotes uncertainty expressed within the assessments, not a separate output field. Naive inference uses $E_i = \varnothing$. Eligibility abbreviates patent-eligible subject matter, also termed subject-matter eligibility in the figures.

With this in mind, we define two key tasks: **(1) Core-dimensional Assessment Generation.** This task requires generating a comment $a_{i,d}$ for each core dimension in patentability assessment. Source-linked reference passages are withheld from evaluation inputs. Missing references are excluded. Evidence must identify a supplied page or claim span; the target Markush image is not external prior art. Missing evidence and invalid outputs are recorded separately from patent outcomes. Support and eligibility are not interchangeable with sufficiency and unity, respectively. **(2) Patentability-outcome Prediction.** This task aims to predict whether a patent claim alongside the Markush structure is worthy of being granted or not. Agreement with the recorded application outcome y_i^{hist} is historical outcome agreement, not expert adjudication of the supplied claim.

3.2 Benchmark Construction

Design Principles. Construction prioritizes source traceability, preservation of chemical scope, and separation of inputs from references. Claims retain document and version identities; drawings preserve the atoms, bonds, substituents, and labels needed to interpret their text. We preserve original examination wording and available source locations. Substituent omissions and claim-version mismatches can change the assessment target; source records make both inspectable.

Dataset Curation. As shown in Figure 2, we collect publicly accessible claims, descriptions, amendments, examination communications, and available cited patents or literature from the European Patent Office and the United States Patent and Trademark Office. Claim reconstruction retains document identity, available source pages, cross-page definitions, and exclusions. Associated structural drawings remain separately traceable even when displayed as a composite. Examination passages are organized by the five dimensions, recording their document, location, and identifiable claim scope. Absent discussion is distinguished from an explicit finding. Unresolved claim-version applicability remains recorded rather than being treated as a verified reference for the selected claim. Inputs and reference answers are stored separately. Human construction-stage feedback identifies omitted text, mixed claim boundaries, inappropriate versions, and mismatched references; corrections and rendered crops are checked against source material. This extraction review precedes the separate domain-expert quality study below.

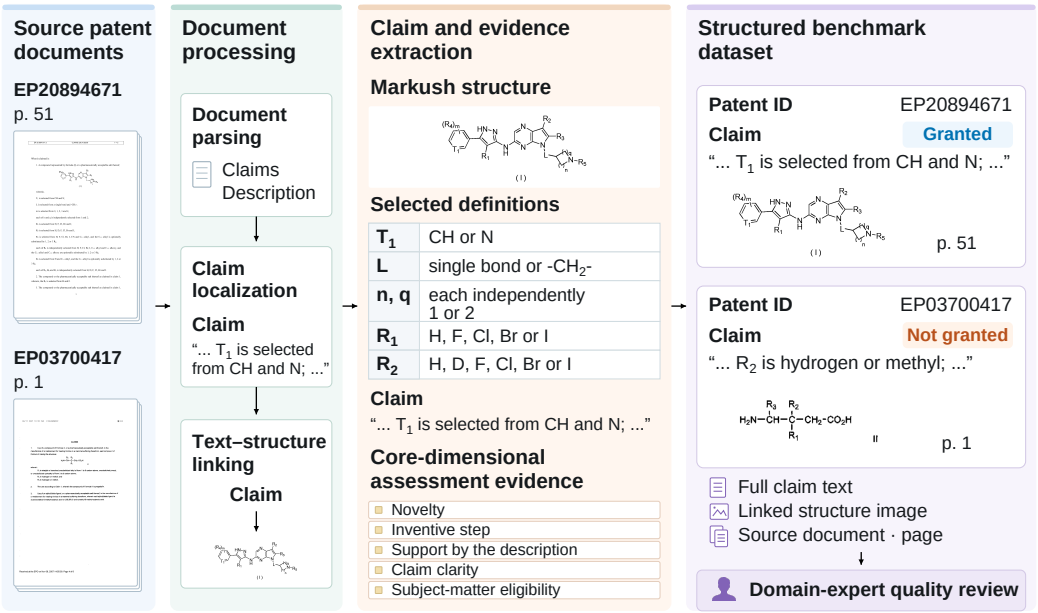


Figure 2: Overview of the dataset curation process of **MarkPatentBench**.

3.3 Dataset Statistics and Human-Expert Validation

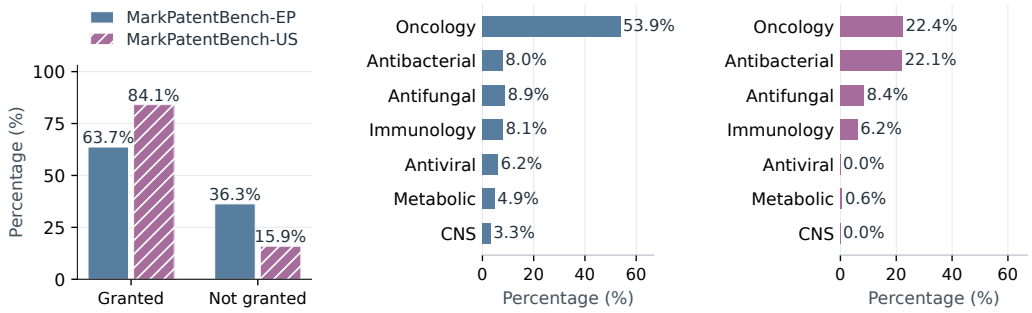
Table 2: Statistics of **MarkPatentBench**.

	MarkPatentBench-EP	MarkPatentBench-US
# Patents	851	308
# Avg. docs per patent	29.64	51.76
# Markush images	1,654	419
# Avg. Markush images per patent	1.94	1.36
Core-dimensional assessment evidence: average tokens (n)		
Novelty	258.4 (815)	225.1 (112)
Inventive step	363.0 (835)	428.1 (70)
Support	64.5 (186)	300.7 (74)
Clarity	203.9 (688)	126.4 (101)
Eligibility	80.1 (539)	287.4 (7)

Table 2 summarizes the full EP and US source collections, their PDF and image assets, and the lengths of core-dimensional assessment evidence. Not granted includes withdrawal and does not imply substantive refusal. Review samples overlap their source collections and are not random estimates of collection-wide quality.

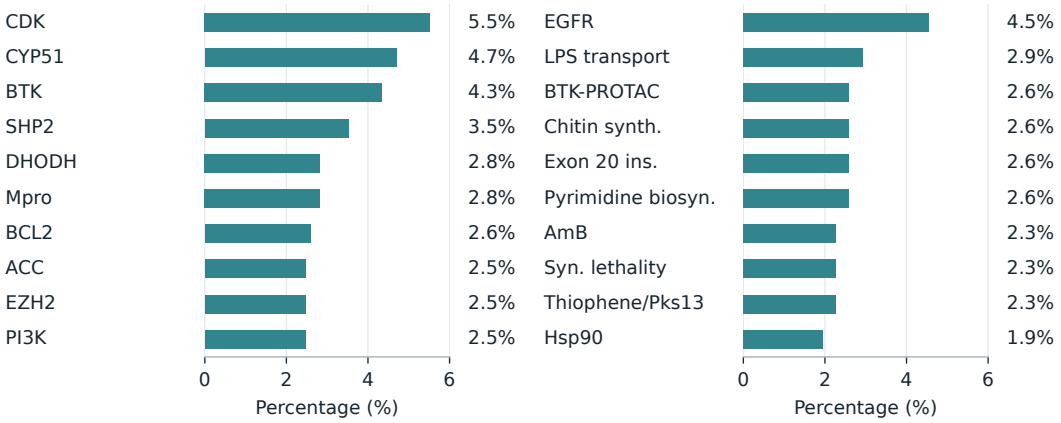
On the selected review samples, drug-discovery experts inspect claims, structural images, source documents, and five reference descriptions through a source-linked interface. Three reviewers independently score seven criteria: Markush fidelity, claim transcription, and each reference dimension, using low (0), medium (1), or high (2) quality. All assigned reviews are complete; Appendix C reports coverage, mean ratings, and agreement. Figure 5 summarizes the quality ratings across the seven criteria for the reviewed sample. These ratings assess material quality, not model performance or a patent's merits.

Figures 3 and 4 summarize the EP and US collections. Domain and target groups are acquisition metadata, not verified biological annotations; missing labels remain unrecorded. US collection themes are assigned from acquisition keywords; broad or ambiguous keywords remain unclassified. The complete target distribution appears in Appendix J.



(a) Grant-outcome statistics (b) MarkPatentBench-EP (c) MarkPatentBench-US

Figure 3: Distribution statistics of **MarkPatentBench**, where (a) shows overall grant-outcome statistics, (b) and (c) report distributions of top target diseases for patents in MarkPatentBench-EP and MarkPatentBench-US, respectively.



(a) MarkPatentBench-EP (b) MarkPatentBench-US

Figure 4: Distribution of the top-10 molecular compound types for patents in MarkPatentBench-EP (a) and MarkPatentBench-US (b). Details of abbreviations are expanded in Appendix J.

3.4 Our Method: Experience-Guided Assessment

To address patentability assessment in drug discovery, we propose a simple, effective method called *Experience-Guided Patent Assessment* (EGPA). The key idea is to leverage generated core-dimensional assessment reviews as reusable experience to guide grant-outcome prediction.

Following verbal-feedback and instruction-refinement approaches (Shinn et al., 2023; Fernando et al., 2024), we compare training reviews with available examiner passages, outcome labels, and claim-scope metadata. Reflection addresses three groups of reasoning challenges also illustrated in Figure 6: prior-art matching, scope and clarity, and evidence gaps. The resulting rules require comparisons of the full claimed combination, attention to Markush restrictions and variable definitions, and explicit uncertainty when evidence is missing. Feedback is consolidated into reusable instructions without case identifiers or copied examination text.

For core-dimensional assessment generation, candidates are ranked on a separate selection split using review similarity and balanced accuracy. The selected instructions guide claim–evidence comparisons to generate reviews of novelty, inventive step, support, clarity, and eligibility. The assessment uses claim text, original Markush images, and available cited evidence; evaluation-case

examiner references are withheld. Instructions remain fixed during evaluation. Appendix G.2 reports the search and the selected rules.

For grant-outcome prediction, we use a trained predictor that combines the refined and Naive reviews to predict patentability outcome. Fitting uses only training reviews and labels, with thresholds selected from out-of-fold training predictions. The predictor leaves the Task 1 assessment text unchanged. More implementation details are presented in Appendix G.2.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Backbone Models. We evaluate GPT-5.6 Sol (OpenAI, 2026), Gemini 3.1 Pro (Google, 2026), DeepSeek V4.1 Flash (DeepSeek, 2026), MiniMax M3 (MiniMax, 2026), and Qwen3.8-Max (Qwen Team, 2026) configurations. In the subsequent sections, GPT-5.6, Gemini 3.1, and Qwen3.8 abbreviate the above model names; DS-Flash denotes DeepSeek V4.1 Flash and DS-Official denotes official `deepseek-flash`. Iter.-Vis. and Src.-Agent abbreviate Iterative-Visual and Source-Agent.

Baseline Settings. We employ the following baseline settings for experiments: (1) *Naive* uses claim text and Markush images. (2) *Evidence-augmented* settings include Hybrid-DOS, which combines hybrid retrieval with DOS RAG's presentation of retrieved passages in original document order (Laitenberger et al., 2025), and Direct-LC, our direct-context baseline motivated by RAG–long-context comparisons (Li et al., 2024). Direct-LC supplies pages in fixed document order up to the input budget, without relevance ranking. (3) *Agentic* settings include Iterative-Visual, inspired by ViDoRAG's iterative visual evidence gathering (Wang et al., 2025), the Fixed RAG executor, and Source-Agent. Retrieval combines BM25 (Robertson & Zaragoza, 2009), BGE-M3 (Chen et al., 2024), and SigLIP2 (Tschannen et al., 2025); Appendix D presents full details of these task-specific implementations.

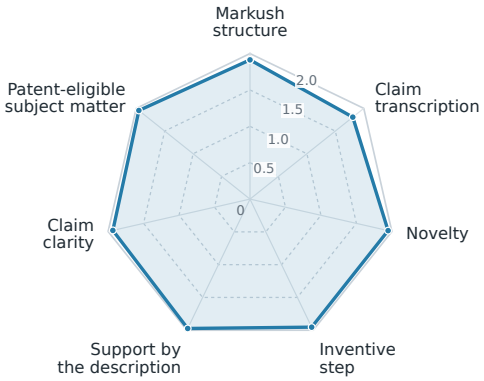


Figure 5: Human-expert review ratings across seven criteria on a 0–2 scale.

Evaluation Metrics. For the assessment generation task, we measure cosine similarity between generated assessments and examiner references using `text-embedding-3-large` (OpenAI, 2024). Scores are averaged within each dimension and then equally across dimensions. We adopt ROUGE-L (Lin, 2004) as a complementary measure of lexical overlap.

For the patentability-outcome prediction task, the primary metric is balanced accuracy (BA), the mean recall of the granted and not-granted classes. We also report Macro-F1, Matthews correlation coefficient (MCC), accuracy with Wilson confidence intervals, and granted-class precision, recall, and F1. Missing outputs and missing references are excluded separately. Appendix G specifies detailed aggregation and coverage.

4.2 MAIN RESULTS

We analyze the tasks separately: Task 1 evaluates generated reviews across novelty, inventive step, support, clarity, and eligibility (Table 3); Task 2 evaluates granted versus not-granted predictions (Table 4). Cross-model rankings describe the benchmark, while the official DeepSeek comparisons and component controls examine our method.

Task 1: Core-dimensional assessment. On EP, EGPA improves similarity to examiner references in all five dimensions over official DeepSeek Naive and Fixed RAG. The largest gain over Naive

Table 3: Evaluation results on core-dimensional assessment generation across MarkPatentBench-EP and MarkPatentBench-US datasets.

Model	Method	Nov.	Inv.	Supp.	Clar.	Elig.	Avg.	ROUGE-L
MarkPatentBench-EP								
GPT-5.6	Naive	0.6190	0.6436	0.5330	0.5320	0.5767	0.5809	0.0968
Gemini 3.1	Naive	0.5970	0.5954	0.5329	0.5675	0.5715	0.5728	0.1115
Qwen3.8	Naive	0.6003	0.6334	0.5754	0.5653	**0.5905**	0.5930	0.1129
MiniMax M3	Naive	0.5884	0.6231	0.5516	0.5362	0.5850	0.5769	0.1018
DS-Flash	Naive	0.6421	0.6511	0.5655	0.5862	0.5787	0.6047	0.1074
GPT-5.6	Hybrid-DOS	0.6421	0.6650	0.5557	0.5413	0.5784	0.5965	0.0963
GPT-5.6	Direct-LC	0.6459	0.6686	0.5558	0.5433	0.5785	0.5984	0.0971
GPT-5.6	Iter.-Vis.	0.6399	0.6620	0.5631	0.5423	0.5799	0.5974	0.0961
DS-Flash	Fixed-RAG	0.6507	0.6732	0.5820	0.5768	0.5876	0.6141	0.1034
DS-Flash	Src.-Agent	0.6489	0.6659	0.5805	0.5742	0.5859	0.6111	0.1037
DS-Official	Naive	0.6330	0.6302	0.5690	0.5747	0.5153	0.5844	0.1033
DS-Official	Fixed-RAG	0.6613	0.6693	0.5715	0.5824	0.5155	0.6000	0.1098
DS-Official	**EGPA (Ours)**	**0.6835**	**0.7077**	**0.5825**	**0.6148**	0.5357	**0.6249**	**0.1185**
MarkPatentBench-US								
GPT-5.6	Naive	0.5703	0.6304	0.6197	0.5338	0.4780	0.5665	0.0902
Gemini 3.1	Naive	0.5597	0.5538	0.5756	0.5332	0.5068	0.5458	0.1018
Qwen3.8	Naive	0.5684	0.6110	0.6439	0.5458	0.5161	0.5770	0.1029
MiniMax M3	Naive	0.5449	0.6034	0.6105	0.5187	**0.5274**	0.5610	0.0940
DS-Flash	Naive	0.5871	0.6210	0.6328	0.5652	0.5106	0.5833	0.1000
GPT-5.6	Hybrid-DOS	0.5924	0.6361	0.6280	0.5306	0.4895	0.5753	0.0885
GPT-5.6	Direct-LC	0.5981	0.6289	0.6339	0.5347	0.4964	0.5784	0.0914
GPT-5.6	Iter.-Vis.	0.5958	0.6323	0.6126	0.5213	0.5003	0.5724	0.0873
DS-Flash	Fixed-RAG	0.6052	0.6315	**0.6480**	0.5662	0.5194	0.5940	0.0996
DS-Flash	Src.-Agent	0.5998	0.6327	0.6423	0.5585	0.5130	0.5893	0.0995
DS-Official	Naive	0.5689	0.6132	0.6164	0.5631	0.4845	0.5692	0.0960
DS-Official	**EGPA (Ours)**	**0.6362**	**0.6605**	0.6347	**0.5910**	0.4661	**0.5977**	**0.1136**

occurs in inventive step, followed by novelty and clarity; support improves least. EGPA leads the displayed models in these four dimensions, while Qwen Naive has the highest eligibility score. The equal-dimension mean rises from 0.5844 for official Naive to 0.6249 for EGPA, and ROUGE-L also increases. The result is therefore broad within-backbone improvement, with substantial variation across dimensions rather than uniform gains.

On US, EGPA again has the highest mean similarity and ROUGE-L, leading novelty, inventive step, and clarity. Support improves over official Naive but remains below Fixed RAG, while eligibility decreases. Eligibility has only six nonempty references, so its ranking and contribution to the equally weighted mean require particular caution. EP support also has much sparser coverage than the other dimensions. These results measure alignment on available examination passages; they neither rank the intrinsic difficulty of the five dimensions nor verify that a review correctly assesses the supplied claim version.

Task 2: Grant-outcome prediction. On EP, EGPA achieves the highest BA, Macro-F1, and MCC among the displayed configurations. Relative to official DeepSeek Naive, BA increases from 55.80% to 65.14%, although accuracy and granted-class F1 decrease. The class recalls explain this difference: EGPA recognizes more not-granted applications while missing more granted applications. BA exposes this trade-off in a panel where granted applications are the majority; accuracy alone favors the more grant-heavy Naive predictions.

On US, EGPA has the highest accuracy, recall, and granted-class F1, but trails Sol Naive and Source-Agent on BA. Source-Agent has the highest Macro-F1, while Sol Naive leads BA and MCC. Thus,

Table 4: Evaluation results on grant-outcome prediction across MarkPatentBench-EP and MarkPatentBench-US datasets.

Model	Method	BA	Macro-F1	MCC	Acc. (95% CI)	Prec.	Rec.	$F1_G$
MarkPatentBench-EP								
GPT-5.6	Naive	54.40	0.423	0.102	43.11 (38.8–47.5)	0.775	0.232	0.357
Gemini 3.1	Naive	59.47	0.539	0.182	54.09 (49.7–58.4)	**0.788**	0.446	0.569
Qwen3.8	Naive	51.65	0.372	0.043	38.92 (34.8–43.3)	0.727	0.164	0.268
MiniMax M3	Naive	53.78	0.441	0.081	44.31 (40.0–48.7)	0.746	0.276	0.403
DS-Flash	Naive	58.58	0.580	0.165	61.68 (57.3–65.8)	0.741	0.672	0.705
GPT-5.6	Hybrid-DOS	53.24	0.402	0.078	41.32 (37.1–45.7)	0.758	0.202	0.319
GPT-5.6	Direct-LC	52.71	0.410	0.062	41.72 (37.5–46.1)	0.738	0.223	0.342
GPT-5.6	Iter.-Vis.	51.46	0.376	0.037	39.12 (34.9–43.5)	0.720	0.173	0.279
DS-Flash	Fixed-RAG	52.67	0.487	0.098	67.86 (63.7–71.8)	0.693	0.947	0.800
DS-Flash	Src.-Agent	61.98	0.622	0.244	67.66 (63.4–71.6)	0.755	0.777	0.766
DS-Official	Naive	55.80	0.537	0.189	**69.86** (65.7–73.7)	0.708	0.947	**0.811**
DS-Official	Fixed-RAG	50.19	0.416	0.017	67.86 (63.7–71.8)	0.681	**0.991**	0.808
DS-Official	EGPA (Ours)	**65.14**	**0.650**	**0.300**	69.26 (65.1–73.1)	0.779	0.765	0.772
MarkPatentBench-US								
GPT-5.6	Naive	**65.52**	0.580	**0.293**	58.28 (50.6–65.6)	**0.871**	0.474	0.614
Gemini 3.1	Naive	64.08	0.603	0.258	61.96 (54.3–69.1)	0.817	0.588	0.684
Qwen3.8	Naive	50.35	0.456	0.007	46.01 (38.5–53.7)	0.703	0.395	0.506
MiniMax M3	Naive	53.14	0.513	0.058	53.99 (46.3–61.5)	0.724	0.553	0.627
DS-Flash	Naive	55.93	0.558	0.116	61.96 (54.3–69.1)	0.736	0.711	0.723
GPT-5.6	Hybrid-DOS	60.54	0.503	0.209	50.31 (42.7–57.9)	0.854	0.360	0.506
GPT-5.6	Direct-LC	57.73	0.481	0.155	48.15 (40.6–55.8)	0.812	0.342	0.481
GPT-5.6	Iter.-Vis.	57.57	0.496	0.150	49.63 (41.3–58.0)	0.791	0.366	0.500
DS-Flash	Fixed-RAG	55.82	0.550	0.163	69.94 (62.5–76.5)	0.727	0.912	0.809
DS-Flash	Src.-Agent	64.83	**0.642**	0.286	68.71 (61.2–75.3)	0.794	0.746	0.769
DS-Official	EGPA (Ours)	53.20	0.486	0.156	**71.17** (63.8–77.6)	0.713	**0.982**	**0.827**

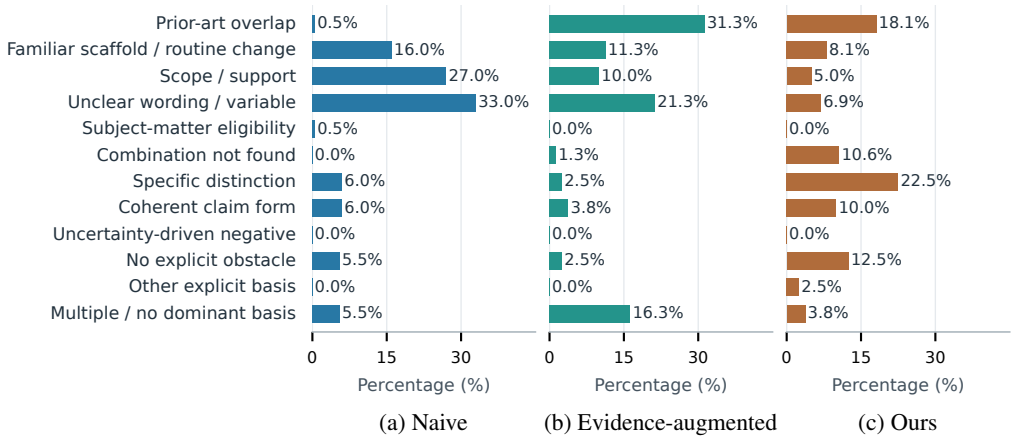
the strongest mean Task 1 alignment does not yield the strongest Task 2 BA. The US panel is selected for evidence availability, and some competing settings have incomplete outputs (Appendix G.4); these rankings describe the reported panel rather than population-level performance.

4.3 Discussions on Different Settings

Naive assessment. Model strengths differ across tasks even without external evidence. On EP, DeepSeek V4.1 Flash has the highest mean Task 1 similarity among Naive configurations, whereas Gemini has the highest Naive BA. Qwen leads eligibility similarity but has low outcome BA. Grant-prediction tendencies also differ: Sol and Qwen recover relatively few granted cases, whereas official DeepSeek Naive predicts granted much more often. Class-specific recall is therefore needed to interpret differences in outcome performance.

Evidence-augmented assessment. For Sol, Hybrid-DOS and Direct-LC increase mean Task 1 similarity on both jurisdictions but reduce Task 2 BA relative to Naive. On EP, their gains concentrate in novelty, inventive step, and support, with smaller changes in clarity and eligibility. US dimension-level changes are less consistent. Evidence access can therefore improve reference alignment without improving the recorded outcome decision. Differences in prompts, evidence budgets, and recovery procedures limit attribution of these changes to retrieval alone.

Agentic assessment. Iterative-Visual improves mean Task 1 similarity over Sol Naive but falls below Direct-LC and lowers Task 2 BA on both jurisdictions. For DeepSeek V4.1 Flash, Source-Agent has slightly lower mean similarity than Fixed RAG but higher BA in both panels. These



(a) Naive (b) Evidence-augmented (c) Ours

Figure 6: Statistics of error types across (a) Naive assessment, (b) evidence-augmented assessment, and (c) our method.

contrasting trends show why additional evidence-selection steps should be assessed separately for review generation and outcome prediction; a more elaborate workflow does not consistently improve both tasks.

Overall, our EGPA outperforms previous methods across two tasks for patentability assessment in drug discovery. For Task 1, paired similarity changes have positive medians in every EP dimension against both Naive and Fixed RAG, extending the gains beyond a few examples. For Task 2, applying the same predictor-fitting procedure to Naive reviews already improves class balance; EGPA does not significantly outperform this control in BA. On US, the transferred predictor slightly reduces BA relative to the refined review's own grant decision. The evidence supports improved text alignment from refinement and an EP outcome benefit from supervised prediction, without establishing an independent outcome benefit from refinement. Further feedback and refinement also failed to improve the recorded joint selection score, so the initial candidate was retained (Appendix I).

4.4 Error Analysis

Figure 6 examines the stated reasoning behind Task 2 errors in an earlier EP sample. Naive mismatches emphasize clarity and scope, evidence-augmented mismatches emphasize prior-art overlap, and agentic mismatches emphasize specific distinctions. These reasons connect the outcome errors to the evidence comparisons required by Task 1: prior-art matching for novelty and inventive step, and claim-scope interpretation for support and clarity. They also motivate the reflection themes in Section 3.4, including explicit uncertainty when evidence is missing. An outcome mismatch does not establish that the cited dimension was assessed incorrectly, and these descriptive categories are not verified causes or per-dimension error rates. Appendix H gives sampling and coding details.

5 Conclusion

We introduced `MarkPatentBench`, a multimodal benchmark that connects chemical claims and complex Markush structures with core-dimensional assessment and grant-outcome prediction, grounded in source-linked examination records and rigorously validated by domain-expert review. Our evaluations demonstrate that providing frontier models with access to examination evidence does not inherently guarantee reliable assessments. Models still exhibit fundamental limitations, such as overlooking precise textual constraints or misjudging obviousness based on scaffold familiarity. Future progress in this domain will necessitate models capable of independently verified claim-level reasoning to accurately preserve and evaluate the claimed chemical scope alongside these metrics.

AI USE STATEMENT

We used ChatGPT and Codex to assist with experimental implementation and debugging, data processing, and manuscript refinement. AI-assisted checks are not counted as domain-expert review. We affirm that all core content and findings in this paper are the original work of the authors. The authors are responsible for the reported evidence, interpretations, and final manuscript.

ETHICS STATEMENT

Public access does not imply unrestricted redistribution: patent-office terms distinguish official and third-party material (European Patent Office, b; United States Patent and Trademark Office). Sources retain their rights and provenance; cited literature does not acquire a new license through inclusion. Licenses for author-created code and annotations, document-level distribution permissions, and appropriate redaction remain to be finalized before release. Source identifiers and retrieval instructions support reproducibility without assuming permission to redistribute every document. Outputs are research predictions, not professional examinations.

REFERENCES

Hengxing Cai, Xiaochen Cai, Junhan Chang, Sihang Li, Lin Yao, Wang Changxin, Zhifeng Gao, Hongshuai Wang, Li Yongge, Mujie Lin, Shuwen Yang, Jiankun Wang, Mingjun Xu, Jin Huang, Xi Fang, Jiaxi Zhuang, Yuqi Yin, Yaqi Li, Changhong Chen, Zheng Cheng, Zifeng Zhao, Linfeng Zhang, and Guolin Ke. SciAssess: Benchmarking LLM proficiency in scientific literature analysis. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 2335–2357, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.125. URL https://aclanthology.org/2025.findings-naacl.125/.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. arXiv preprint arXiv:2402.03216, 2024. URL https://arxiv.org/abs/2402.03216v4. Version 4, 2024.

Rubing Chen, Jian Wang, Wenjie Li, Xiao-Yong Wei, and Qing Li. To retrieve or to think? cross-boundary context evolution for multi-hop complex reasoning. arXiv preprint arXiv:2601.08747, 2026. URL https://arxiv.org/abs/2601.08747.

DeepSeek. DeepSeek V4.1 Flash. Official release announcement, September 10, 2026. URL https://api-docs.deepseek.com/zh-cn/news/news260910/.

European Patent Office. European patent convention, article 52: Patentable inventions. European Patent Convention, a. URL https://www.epo.org/en/legal/epc/2020/a52.html. Accessed September 22, 2026.

European Patent Office. Terms and conditions of use for the website of the european patent office, b. URL https://www.epo.org/en/terms-of-use/terms-and-conditions-use-website-european-patent-office. Accessed September 22, 2026.

European Patent Office. Guidelines for examination, part f-iv, 4.1: Clarity, 2026. URL https://www.epo.org/en/legal/guidelines-epc/2026/f_iv_4_1.html. Accessed September 22, 2026.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/99e9e141aafc314f76b0ca3dd66898b3-Abstract-Conference.html.

Chrisantha Fernando, Dylan Sunil Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. Promptbreeder: Self-referential self-improvement via prompt evolution. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 13481–13544. PMLR, 2024. URL https://proceedings.mlr.press/v235/fernando24a.html.

Igor V. Filippov and Marc C. Nicklaus. Optical structure recognition software to recover chemical information: OSRA—an open source solution. *Journal of Chemical Information and Modeling*, 49(3):740–743, 2009. doi: 10.1021/ci800067r. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC2889020/.

Google. Gemini 3.1 Pro: A smarter model for your most complex tasks. Official model announcement, 2026. URL https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/.

Lekang Jiang, Pascal A. Scherz, and Stefan Goetz. Patent-CR: A dataset for patent claim revision. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *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, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.116. URL https://aclanthology.org/2025.naacl-long.116/.

Seonok Kim. Vld-rag: Agentic vision-language retrieval-augmented generation for long, visually-rich multi-page documents. arXiv preprint arXiv:2607.24748, 2026. URL https://arxiv.org/abs/2607.24748.

Alex Laitenberger, Christopher D Manning, and Nelson F. Liu. Stronger baselines for retrieval-augmented generation with long-context language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 32559–32569, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1656. URL https://aclanthology.org/2025.emnlp-main.1656/.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks, 2020. URL https://arxiv.org/abs/2005.11401.

Zhuowan Li, Cheng Li, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. Retrieval augmented generation or long-context LLMs? a comprehensive study and hybrid approach. In Franck Dernoncourt, Daniel Preoţiuc-Pietro, and Anastasia Shimorina (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 881–893, Miami, Florida, US, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-industry.66. URL https://aclanthology.org/2024.emnlp-industry.66/.

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 *Advances in Neural Information Processing Systems 38: Datasets and Benchmarks Track*, 2025. URL https://papers.neurips.cc/paper_files/paper/2025/hash/4aab82c8d6b77c0b6b010145c1bfcdd3-Abstract-Datasets_and_Benchmarks_Track.html.

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/.

MiniMax. MiniMax M3: Frontier coding, 1m context, native multimodality—all in one model. Official model announcement, 2026. URL https://www.minimax.io/blog/minimax-m3.

Lucas Morin, Valéry Weber, Ahmed Nassar, Gerhard Ingmar Meijer, Luc Van Gool, Yawei Li, and Peter Staar. Markushgrapher: Joint visual and textual recognition of markush structures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14505–14515, 2025. doi: 10.1109/CVPR52734.2025.01352. URL https://openaccess.thecvf.com/content/CVPR2025/html/Morin_MarkushGrapher_Joint_Visual_and_Textual_Recognition_of_Markush_Structures_CVPR_2025_paper.html.

OpenAI. New embedding models and api updates, January 2024. URL https://openai.com/index/new-embedding-models-and-api-updates/.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition. Official model announcement, 2026. URL https://openai.com/index/gpt-5-6/.

George Papadatos, Mark Davies, Nathan Dedman, Jon Chambers, Anna Gaulton, James Siddle, Richard Koks, Sean A. Irvine, Joe Pettersson, Nicko Goncharoff, Anne Hersey, and John P. Overington. SureChEMBL: a large-scale, chemically annotated patent document database. *Nucleic Acids Research*, 44(D1):D1220–D1228, 2016. doi: 10.1093/nar/gkv1253. URL https://pubmed.ncbi.nlm.nih.gov/26582922/.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026. URL https://qwen.ai/blog?id=qwen3.8.

Julian Risch, Nicolas Alder, Christoph Hewel, and Ralf Krestel. Patentmatch: A dataset for matching patent claims & prior art. arXiv preprint arXiv:2012.13919, 2020. URL https://arxiv.org/abs/2012.13919.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009. doi: 10.1561/1500000019. URL https://www.nowpublishers.com/article/Details/INR-019.

Yaorui Shi, Sihang Li, Taiyan Zhang, Xi Fang, Jiankun Wang, Zhiyuan Liu, Guojiang Zhao, Zhengdan Zhu, Zhifeng Gao, Renxin Zhong, Linfeng Zhang, Guolin Ke, Weinan E, Hengxing Cai, and Xiang Wang. Intelligent system for automated molecular patent infringement assessment. arXiv preprint arXiv:2412.07819, 2025. URL https://arxiv.org/abs/2412.07819v2. Version 2, 2025; first posted in 2024.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. arXiv preprint arXiv:2303.11366, 2023. URL https://arxiv.org/abs/2303.11366.

Tim Strohmeyer, Lucas Morin, Gerhard Ingmar Meijer, Valéry Weber, Ahmed Nassar, and Peter Staar. Markushgrapher-2: End-to-end multimodal recognition of chemical structures. arXiv preprint arXiv:2603.28550, 2026. URL https://arxiv.org/abs/2603.28550.

Mirac Suzgun, Luke Melas-Kyriazi, Suproteem K. Sarkar, Scott Duke Kominers, and Stuart M. Shieber. The harvard uspto patent dataset: A large-scale, well-structured, and multi-purpose corpus of patent applications. arXiv preprint arXiv:2207.04043, 2022. URL https://arxiv.org/abs/2207.04043.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. URL https://arxiv.org/abs/2502.14786.

United States Patent and Trademark Office. Terms of use for uspto websites. URL https://www.uspto.gov/terms-use-uspto-websites. Accessed September 22, 2026.

Qiuchen Wang, Ruixue Ding, Zehui Chen, Weiqi Wu, Shihang Wang, Pengjun Xie, and Feng Zhao. ViDoRAG: Visual document retrieval-augmented generation via dynamic iterative reasoning agents. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 9113–9134, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN

979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.464. URL https://aclanthology.org/2025.emnlp-main.464/.

Zenan Zhai, Dat Quoc Nguyen, Saber Akhondi, Camilo Thorne, Christian Druckenbrodt, Trevor Cohn, Michelle Gregory, and Karin Verspoor. Improving chemical named entity recognition in patents with contextualized word embeddings. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pp. 328–338, 2019. doi: 10.18653/v1/W19-5035. URL https://aclanthology.org/W19-5035/.

A Example Benchmark Input

EP20894671 (MKB-0015) illustrates a benchmark input: a compound claim paired with its original Markush drawing. The panel reproduces the complete stored claim, including its formula placeholder and paragraph breaks.

Claim 1 (EP20894671)

1. A compound represented by formula (I) or a pharmaceutically acceptable salt thereof,

[FORMULA_1]

wherein,

T_1 is selected from CH and N;

L is selected from a single bond and -CH_2-;

m is selected from 0, 1, 2, 3 and 4;

each of n and q is independently selected from 1 and 2;

R_1 is selected from H, F, Cl, Br and I;

R_2 is selected from H, D, F, Cl, Br and I;

R_3 is selected from H, F, Cl, Br, I, CN and C_{1-3} alkyl, and the C_{1-3} alkyl is optionally substituted by 1, 2 or 3 R_a;

each of R_4 is independently selected from H, F, Cl, Br, I, C_{1-3} alkyl and C_{1-3} alkoxy, and the C_{1-3} alkyl and C_{1-3} alkoxy are optionally substituted by 1, 2 or 3 R_b;

R_5 is selected from H and C_{1-3} alkyl, and the C_{1-3} alkyl is optionally substituted by 1, 2 or 3 R_c;

each of R_a, R_b and R_c is independently selected from H, D, F, Cl, Br and I.

Input and evidence. Naive inference receives the complete claim and full original structural image; the pixel crop in Figure 1 is for display only. Evidence-augmented settings also receive selected document pages and source identifiers. The target structure remains distinct from external prior art.

Assessment. Models interpret the scaffold together with substituent definitions and textual exceptions. Generated assessments are compared with reference passages from the corresponding dimension, excluding pairs without a reference. Figure 1 names the assessment targets without assigning case-specific answers.

B Protocols and Reproducibility

Model-visible claim inputs, source-page identities, request variants, and responses are retained for reproducibility. Missing documents do not, by themselves, support either grant outcome.

Input and encoder identifiers. The original inputs in Appendix A and the earlier BGE-M3 text encoder are identified below. Fingerprints use SHA-256; the model revision is listed separately.

Artifact	Fingerprint or revision
Original image	`1881e955ed019c4af77cfb131e9ceddeba428de181083fa3d02501d434cdc949`
Claim text	`78e016f00d0b2d7a2d7cdd39cb437a82e9425d17a88585c93d3dfc50e4efa325`
BGE-M3 revision	`a46a13810c9d7f876fccd8d7017512ebc265d2c1`
BGE-M3 weights	`b5e0ce3470abf5ef3831aa1bd5553b486803e83251590ab7ff35a117cf6aad38`

Earlier EP text scoring used these BGE-M3 weights, CLS pooling, FP32 inference, 8,190-token text windows, token-weighted chunk centroids, and L2 normalization. The final text evaluation uses

the API encoder and chunking protocol in Appendix G.2; the retrieval index retains its own shorter overlapping chunks. The reference audit checks whether passages occur continuously in the source text; it does not establish their applicability to the evaluated claim version.

C Domain-Expert Quality Review

The final review contains 100 applications: 50 EP and 50 US. Three reviewer accounts each completed all seven criteria for every assigned application, yielding 300 complete reviews and 2,100 ordinal ratings. The authors identify the reviewers as drug-discovery domain experts. The retained records do not specify individual credentials or years of experience, so no more specific qualification is claimed. All stored claim and image fingerprints match the frozen review manifest. Scores represent low (0), medium (1), or high (2) quality; they are not percentages of legally correct judgments.

The EP sample extends the earlier SHOWCASE 40 with ten additional cases selected by a fixed hash ordering from unused inputs that passed image–text screening, balancing recorded outcomes within the addition. The US sample comes from a separately supplied reviewed package and contains granted applications only. The two samples are not matched for document availability, outcome balance, or sampling procedure. Means are descriptive summaries of an ordinal scale, and the study is not a blinded assessment of model-generated answers. Historical outcome and examination material are available in the review interface.

Table 5 reports all criteria separately by jurisdiction, with 150 ratings per criterion in each jurisdiction. Exact pairwise agreement averages the three reviewer pairs across the 50 applications in each jurisdiction. This statistic describes reproducibility of the ratings, not their correctness. The US material shows greater reviewer disagreement and lower mean quality, supporting caution in interpreting the US model results. The study reports independent ratings without converting them into consensus labels.

Table 5: Domain-expert material-quality ratings and pairwise agreement.

	EP		US	
Criterion	Mean	Agree. (%)	Mean	Agree. (%)
Markush fidelity	1.913	88.0	1.113	43.3
Transcription	1.807	69.3	1.367	27.3
Novelty	1.947	92.7	1.400	35.3
Inventive step	1.953	93.3	1.340	26.7
Support	1.973	94.7	1.253	17.3
Clarity	1.933	92.0	1.273	20.0
Eligibility	1.953	90.7	1.027	2.7

Earlier construction-stage review. A retained field-review inventory contains 33 reviews from five accounts covering 20 distinct applications, with 201 fields marked corrected and verified: 25 benchmark-input fields, 130 claim fields, and 46 assessment fields. The associated report records six completed reruns and explicitly retains unresolved extraction failures. This inventory supports the construction-feedback description; it does not establish an independent completed 50-case validation or provide final quality scores. The earlier SHOWCASE 40 single-criterion ratings are likewise historical and are not added to the final seven-criterion denominator.

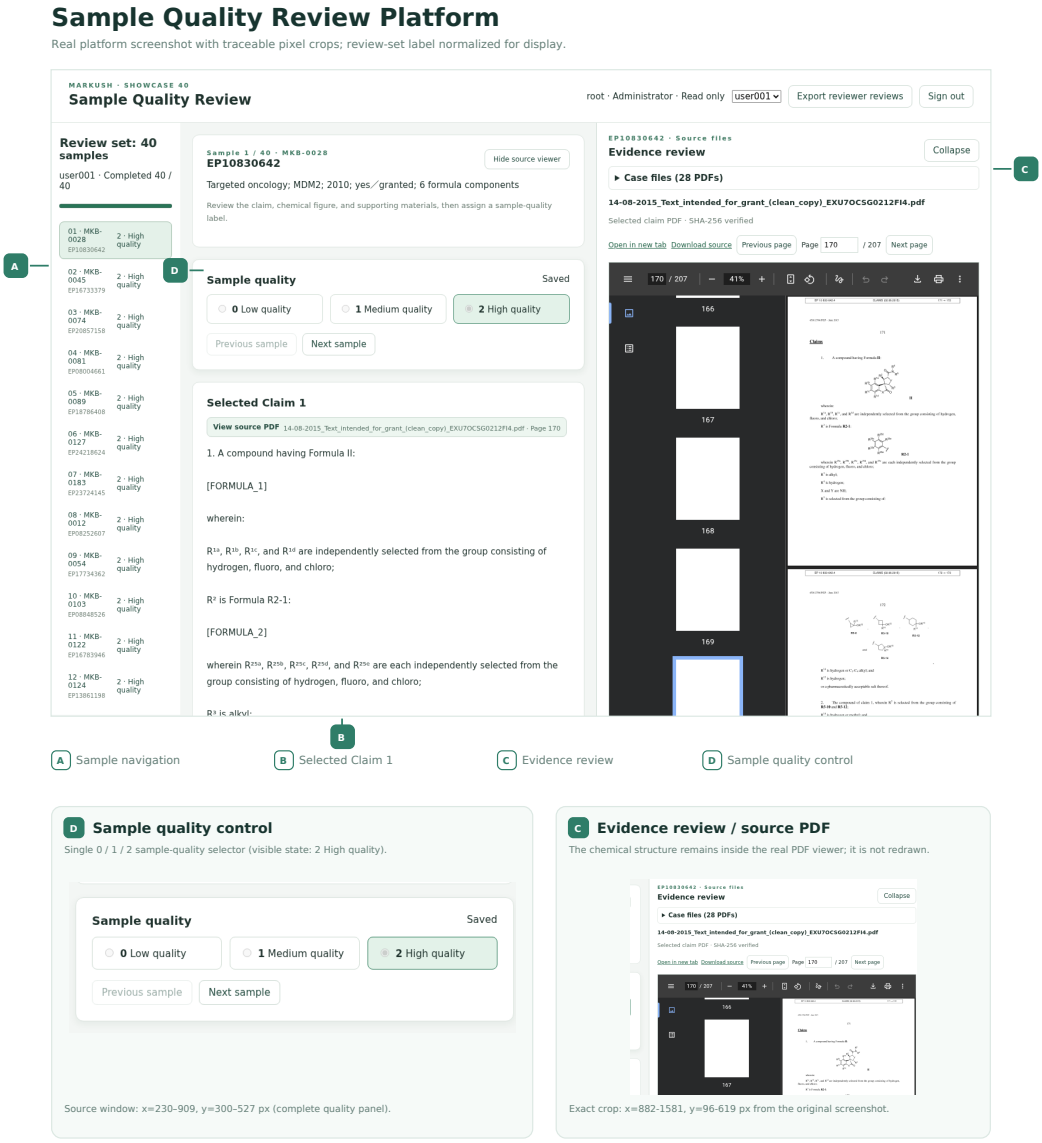


Figure 7: Historical source-linked review interface. Claim text and original documents can be inspected together. This screenshot shows the earlier single-criterion interface, not the final seven-criterion form.

D EVIDENCE-AUGMENTATION DETAILS

The three evidence configurations use an application's documented cited corpus. This provides information unavailable to an open-world retrieval system: the candidate set itself can originate from later examination records. Document identity and publication-date qualification are not fully resolved, so we call these *cited-document* experiments rather than unrestricted, chronologically certified prior-art search.

Hybrid-DOS. The retriever combines BM25, BGE-M3 text similarity, and SigLIP2 visual similarity. The implementation retrieves candidates from each channel and combines ranks using reciprocal rank fusion,

$$\mathrm{RRF}(p) = \sum_{m \in \mathcal{M}} \frac{1}{60 + \mathrm{rank}_m(p)}. \tag{2}$$

The baseline uses equal channel weights, up to 60 candidates per channel, and restores document and page order after selection. Neighboring pages can preserve local definitions. At most 24 cited pages enter the bounded evidence set. This is the project's implementation, not a claim of reproducing another system's entire algorithm or trained retriever.

Direct-LC and Iterative-Visual. Direct-LC selects a prefix in frozen document and page order, without relevance ranking. It stops at the verified input budget and records the omitted suffix; it is not a full-corpus oracle. Iterative-Visual starts from retrieval, inspects evidence, and formulates additional queries around unresolved comparisons. It permits at most three rounds, at most 24 selected pages per round, and at most 49 distinct cited pages overall. The final request includes the retained original pages. All three configurations preserve one target image and operate within the experiment's 50-image request bound as well as text and output budgets. Their unequal evidence and navigation budgets are reported as configuration differences.

E Earlier Ours-C: Experience Memory

Executor and evidence budget. Earlier Ours-C development uses official `deepseek-flash`, with `gpt-5.6-sol` for reflection and update review (DeepSeek, 2026; OpenAI, 2026). The executor performs at most two searches and inspects at most eight new pages per search. For the 651-case EP comparison, the frozen C memory guides OpenLux `deepseek-v4.1-flash`; a separate call generates core-dimensional reviews from the claim and inspected pages. Source-Agent additionally produces source-verification commentary. These settings are separate from the full-framework generation protocol.

Frozen configurations and selection. Experience refinement uses 50 development cases and 20 selection cases over three rounds. The memory is frozen before the 651-case evaluation; it is not updated using evaluation labels. Development starts from an empty memory; the selected C configuration contains 19 experience cards. The frozen memory and original development templates are retained with the source assets.

The selection criterion is a count of correct historical outcome predictions, with mean squared probability error as the secondary criterion:

$$\mathrm{Brier}(\pi) = \frac{1}{|S|} \sum_{i \in S} \big(p_\pi(x_i) - y_i^{\mathrm{hist}}\big)^2. \tag{3}$$

A candidate with more correct predictions is preferred even if its Brier score worsens. At equal correct count, at least 0.005 Brier improvement is required. Invalid or incomplete candidates cannot replace a complete incumbent. This repeated use of 20 selection cases introduces selection-overfitting risk; selection performance is not reported as test performance.

Experience retrieval and prompts. A card has four fields: trigger, action, evidence boundary, and counter-risk. The query is the supplied claim concatenated with accumulated page observations; it contains no evaluation labels. BGE-M3 text-window vectors are averaged and L2-normalized, then ranked against similarly encoded cards by cosine similarity. Up to four cards are inserted into navigation and final-decision context. For example, the frozen C configuration asks for a topology fingerprint and warns that a minimal feature conjunction can omit a decisive limitation; these instructions are not documentary evidence. Retrieved cards must never be cited as prior art.

The original reflection and update templates remain with the source assets. These development settings apply only to earlier Ours-C, not to the full framework's training, selection, or predictor. Current full-framework prompts appear in Appendix G.3.

F Paired Drug-Discovery Case Illustration

Figure 1 uses original chemical images and abbreviated claim text. The complete supplied claims, displayed text, editable figure, and source notes are retained with the manuscript assets. The shared workflow is illustrative and is executed conceptually for each case separately; it does not document a new database search or model run. Ellipses indicate omitted source text, so the displayed excerpts do not replace full claims.

Granted Chk-1 example. MKB-0015 (EP20894671) supplies the compound claim and original Markush image. The displayed claim is from 2024, while the cited European search opinion is from 2023: novelty is on physical page 2, inventive step on page 3, original disclosure under Article 123(2) on page 1, and the clarity objection under Article 84 and Rule 43(2) on page 3. No new applicability check aligns those earlier findings with the displayed later claim. In particular, the clarity objection is historical, not a demonstrated unresolved defect of the granted claim. The retained historical excerpt concerning Article 123(2) is not a comprehensive favorable finding on sufficiency or every support question. The supplied opinion contains no explicit independent eligibility finding. The historical grant is documented by the decision dated 7 January 2025.

Not-granted Nrf2 example. EP14826727 supplies independent use claim 5, spanning physical pages 24–25 of the published application, and its two original formula images. The EPO opinion dated 4 August 2017 supplies the novelty objection (page 3, claims 1–14), clarity discussion (pages 1–2), and treatment-method exclusion under Article 53(c) (page 4, claims 5–13). The inventive-step objection instead comes from the international written opinion marked 10 December 2014, on physical page 4 of the report entered in the file in 2016: claims 5–14 received an unfavorable assessment, whereas claims 1–4 received a favorable assessment. That earlier opinion also assessed novelty favorably for claims 1–14. The EPO Article 83 passage (page 2) addresses the application and explicitly claim 1; it is not an independently established defect of displayed claim 5. The 28 February 2018 loss-of-rights notice attributes deemed withdrawal to unpaid fourth-year renewal and additional fees. These recorded objections do not establish that the application was formally refused because of five substantive defects. The illustration is not an additional model-evaluation result.

G Experimental Scope and Metrics

The accepted EP benchmark contains 851 applications with 542 granted and 309 not-granted labels. Tables 3 and 4 use the same 501-case EP subset: the intersection of saved multi-model outputs and the full framework's evaluation partition, excluding its training and selection applications. US results use the separate panel. Collection statistics, quality review, and the earlier decision-basis analysis retain their stated scopes. Frozen outcomes are not expert judgments of the supplied claim. BA averages the two class recalls. Supplementary Macro-F1 equally averages granted and not-granted F1; MCC uses all four confusion-matrix counts and ranges from -1 to 1. These supplementary metrics are computed from the same frozen predictions and were not used to select instructions, predictors, or thresholds. ROUGE-L averages F1 over available case–dimension pairs; mean cosine similarity equally weights dimension means.

In the main result tables, Nov., Inv., Supp., Clar., and Elig. abbreviate novelty, inventive step, support, clarity, and eligibility; R-L denotes ROUGE-L and Avg. the equal-dimension mean similarity. Acc., Prec., and Rec. denote accuracy, precision, and recall. BA and accuracy are percentages; accuracy includes Wilson 95% confidence intervals. Precision, recall, and $\mathrm{F1_G}$ refer to granted. Bold marks the best displayed value within each jurisdiction, including ties at the reported precision. Text similarity does not certify assessment correctness.

Final assessment scores use 3,072-dimensional `text-embedding-3-large` vectors and the common chunking protocol in Appendix G.2. Multiple reference passages form a character-length-weighted normalized centroid. For generated assessment $a_{i,d}$ and available reference passages $r_{i,d,j}$,

$$s_{i,d} = e(a_{i,d})^{\top} \frac{\sum_j |r_{i,d,j}| e(r_{i,d,j})}{\left\| \sum_j |r_{i,d,j}| e(r_{i,d,j}) \right\|_2}. \tag{4}$$

Only assessment text and the corresponding source passages enter this comparison.

G.1 EP Evaluation Coverage

The common EP evaluation subset contains 501 applications: 341 granted and 160 not granted. The historical multi-model run generated outputs for 651 accepted applications; the full-framework run evaluated 701 after setting aside 120 training and 30 selection applications. Their intersection determines this comparison without selecting cases by performance. Each displayed EP configuration has 501 valid predictions. Saved responses and original failures remain retained; rescoring adds no generation.

Recovery was not identical-request repetition throughout. The recorded citation-safe variant instructed the model to leave optional evidence-memory and source-evidence arrays empty, omit citation objects and quotes, and continue reasoning from the supplied pages in its prose. Model-visible instructions therefore changed for some recovered outputs. These results include recovery variants and must not be presented as a uniform first-attempt protocol or as uniformly populated source-citation output. Completion establishes technical coverage, not substantive correctness.

The historical model comparison uses OpenLux identifiers `gpt-5.6-sol`, `gemini-3.1-pro-preview`, `qwen3.8-max`, `MiniMax-M3`, and `deepseek-v4.1-flash`. EGPA and its controls use official `deepseek-flash`. Naive, Sol RAG methods, Iterative-Visual, and Source-Agent assess the exact unamended claim; Fixed RAG predicts the historical outcome. Using the same cases and scoring does not erase these target, prompt, evidence, or provider differences. The official DeepSeek rows in Tables 3 and 4 provide comparisons within that backbone; Appendix I reports the component controls.

All displayed EP configurations use the same expanded V20 references and valid-pair mask: 479 novelty, 488 inventive step, 105 support, 400 clarity, and 324 eligibility cases. The resulting 1,796 pairs per configuration exclude missing and placeholder text rather than replacing it with labels or summaries. Support has the least coverage. Application linkage and continuous source wording do not establish applicability to the selected claim version.

Canonical input, claim-text, and original-image identities were checked before reusing outputs. The full-framework generation also retains its recorded corrections to claim-number headers and incorporated-claim text; using the same source input does not imply identical model-visible prompts. No predictor or threshold is refitted for the common subset.

G.2 Full Framework: Training and Text Scoring

The original full-framework split assigns the accepted 851 applications to 120 training (86 granted, 34 not granted), 30 selection (15 each), and 701 evaluation cases (441 granted, 260 not granted). The reported common subset retains 501 of those evaluation cases, with no training or selection overlap. A48 uses 48 cases within training, not additional cases. Instructions, predictors, and thresholds remain frozen. Earlier experiments included some evaluation applications, so disjoint identities do not establish a fresh blind test.

Candidate learning and selection. The search considers nested feedback sets of 48, 96, and 120 training applications. Candidate families use evidence-chain instructions, draft-and-verifier assessment, or bounded selection of at most 16 available evidence-page identifiers followed by evidence-chain assessment. Up to two additional training-feedback refinements follow the strongest initial candidate. On the 30 selection cases, each candidate with its fitted predictor is compared with direct Naive inference using

$$J = \min\left(\frac{\Delta \mathrm{BA}}{0.05}, \frac{\Delta S_{\mathrm{BGE}}}{0.01}\right), \tag{5}$$

where BA is on $[0, 1]$ and S_{BGE} is the equal-dimension BGE-M3 mean. Ties favor the sum of normalized improvements, fewer feedback cases, fewer inference stages, and finally the configuration identifier. Selection uses BGE-M3, not the final table's `text-embedding-3-large` metric.

The selected A48 candidate reuses verified rules learned from 48 training comparison cards, processed in three batches of 16 and consolidated. Cards contain available baseline and examiner passages, outcome feedback, and claim-scope metadata; historical-version passages may teach writing patterns but do not establish a defect of the selected claim. Learning requests prohibit case identifiers, copied

official phrases, and case-specific chemical structures in reusable rules. The initial A48 candidate is retained after selection. Appendix G.3 reproduces its learning templates and all nine selected rules.

Generation and input reuse. Official `deepseek-flash` calls use temperature zero, a 9,000-token output limit, JSON output, and reasoning effort set to `none`. Each request permits two actual attempts, with original attempts retained. Inference uses selected claims and original formula images, together with fixed evidence packets where available; empty packets are explicit. The final input-correction pass preserves frozen rules, predictors, and thresholds. It regenerates three configurations for 36 evaluation applications with corrected claim headers and reuses verified outputs for the remaining 665. The development artifacts retain their original input provenance and are not retrained during this correction.

Predictor controls. Each model represents candidate and Naive core-dimensional reviews in two tagged text slots. The Naive control repeats its own review in both slots; Fixed RAG pairs its review with Naive. Word unigram and bigram TF-IDF uses at most 4,000 features, minimum document frequency 2, maximum document frequency 0.98, sublinear term frequency, and Unicode accent stripping. Class-balanced logistic regression uses $C = 0.03$, the `liblinear` solver, and at most 3,000 iterations. Five stratified shuffled folds (seed 20260924) fit the vocabulary and predictor within each training fold. Threshold candidates include 0, 0.5, 1, and midpoints between distinct out-of-fold probabilities. Selection maximizes training out-of-fold BA, breaking ties by proximity to 0.5 and then lower threshold. The final vocabulary and predictor are fitted on all 120 training cases. Writer features can be in-sample to instruction learning; out-of-fold predictor diagnostics are not an unbiased writer evaluation. Evaluation labels are not used for fitting or threshold selection. The predictor changes only the grant output.

Text scoring and uncertainty. After excluding empty and placeholder passages, the expanded examiner-reference set (V20) covers 479 novelty, 488 inventive-step, 105 support, 400 clarity, and 324 eligibility cases in the common EP subset. Long texts are split into continuous chunks of at most 1,800 UTF-8 bytes without discarding characters. The OpenLux `text-embedding-3-large` endpoint returns 3,072-dimensional vectors; normalized chunk embeddings are combined by character-length weighting and normalized again. References form a length-weighted normalized centroid. Cosine similarity is averaged within each dimension and then equally across dimensions. Paired uncertainty estimates resample applications within grant-label strata using 10,000 bootstrap replicates and 97.5% intervals. V20 is application-linked and does not guarantee that every passage assesses the supplied claim version.

G.3 Prompts

These panels reproduce the selected A48 candidate's stored prompt components. Bracketed uppercase fields are presentation placeholders for serialized training data, not literal request text. The inference system concatenates the base instruction, refinement instruction, and nine learned rules in order (space-separated). User messages contain the selected claim, evidence packet with page identifiers, and output-format instruction; original formula and evidence-page images accompany the text. Naive omits external evidence; Fixed RAG uses the base instruction without the refinement or learned rules. The output-format example is a schema instruction, not a measured probability; its JSON is indented for display.

Learning: system instruction

You are analyzing TRAINING-ONLY chemical-patent review examples. Derive transferable rules, independent verifier checks and de-identified error-memory lessons. A historical or unverified claim-version passage is only same-EP writing context and must never be treated as a contradiction of the selected claim. Do not copy EP numbers, source document IDs, exact official phrases or case-specific chemical structures. Missing evidence is not proof of grant or a technical defect. Respond as valid JSON only.

Learning: batch instruction

These are training-only comparison cards. For each, note that only selected_claim scope can support an error conclusion about the selected claim. Historical scope may teach general caution or writing patterns only. Return JSON with arrays A_rules (3-6 concise rules), B_checks (3-6 verifier checks), C_lessons (3-8 objects with trigger_terms array and lesson string). All content must be generic and de-identified. Cards: [TRAINING_CARDS_JSON]

Learning: consolidation

Consolidate these training-only generic notes. Return exactly JSON with A_rules (5-12 short strings), B_checks (5-12 short strings), C_lessons (8-30 objects, each trigger_terms string array and lesson string). Deduplicate, make claim-version and evidence-absence checks explicit. Do not include EP numbers, verbatim official text, document identifiers or case-specific structures. Notes: [BATCH_NOTES_JSON]

Inference: base instruction

You assess the historical grant outcome of this EP application using the supplied claim, original Markush image and selected source pages. Separately write five assessments of the supplied claim: novelty, inventive step, support by description, clarity, and patent-eligible subject matter. Grant outcome does not imply that every dimension is favorable. Do not assume missing evidence proves a defect or compliance. Distinguish direct source observation, inference and uncertainty. Treat page dates and document identities as unresolved unless supplied. The target Markush image is not prior art. Avoid copying examiner wording. Return one JSON object matching the schema.

Inference: refinement instruction

Evidence-chain task: for every dimension first identify the selected claim number and full Markush feature combination, then compare it with a specifically visible document/page, and finally state a qualified conclusion. A shared fragment does not establish anticipation. Do not state that inventive step fails merely because novelty fails. Distinguish absent evidence from positive proof. Do not transfer an examiner objection from an earlier or unverified claim version. Check that the conclusion direction follows the stated comparison. Training-only transferable rules:

Inference: selected learned rules

1. Only the selected claim version can support a conclusion that the selected claim is defective; historical or unverified claim versions are context for caution and drafting patterns only.

2. A missing-evidence warning (missing pages, unresolved document identity or date) prevents a definitive prior-art or eligibility conclusion; absence of evidence is not proof of grant or defect.

3. For novelty, require a direct disclosure of the full claimed combination; a broad genus or category-A search listing is not a novelty-destroying disclosure without a specific pointer.

4. For inventive step, identify the closest prior art, the distinguishing feature, the technical problem and any comparative effect; a mere arbitrary selection among known possibilities supports an obviousness objection.

5. For clarity, flag functional terms, undefined structural boundaries, self-referential or inconsistent variable definitions, and claims dependent on non-existent claims as potential clarity defects.

6. For eligibility, distinguish product/process claims from excluded methods of treatment; a method-of-treatment claim may need reformatting as a product-for-use claim.

7. For support, compare the breadth of a Markush genus with the exemplified disclosure; a large discrepancy between claim scope and description can indicate a support/enablement issue.

8. A proviso or disclaimer in the selected claim must be considered when comparing against prior art; a prior-art disclosure that falls within the proviso may not anticipate the selected claim.

9. A chemical product claim is generally eligible subject matter; eligibility issues arise when the claim is directed to an abstract idea, natural phenomenon, or an excluded method of treatment as claimed.

Inference: output format

Return JSON with exactly these top-level fields:

```
{
  "predicted_granted": true,
  "grant_probability": 0.5,
  "dimensions": {
    "novelty": "At least 30 characters of review text",
    "inventive_step": "At least 30 characters of review text",
    "support": "At least 30 characters of review text",
    "clarity": "At least 30 characters of review text",
    "eligibility": "At least 30 characters of review text"
  }
}
```

Each dimension value must be a plain string, not an object. No grant_outcome, assessment, reasoning, or type fields.

The rules are empirical model-generated artifacts reproduced without editorial correction, not authoritative legal guidance. Their treatment of support and eligibility can conflate concepts or jurisdictions; the benchmark retains separate dimension labels and evaluates the resulting text rather than endorsing every learned instruction.

G.4 US Evaluation Coverage and Controls

The US result tables use 163 applications: all 49 not-granted cases and 114 granted cases retained from the 308-application source collection. Positive-case selection follows cited-material availability, the number of dimensions with examination evidence, and reference-quote counts; the saved membership list is reused for every configuration. The panel is therefore evidence-selected, not a representative sample of US applications. The dataset distribution figures retain the complete source collection.

The parent collection contains 7,107 distinct cited documents (7,035 PDFs and 72 HTML files). These are retrieval materials, not complete dossiers or the quality-review PDF counts. Retrieval fuses BM25, BGE-M3, and SigLIP2 rankings. Hybrid-DOS, Direct-LC, and Iterative-Visual use `gpt-5.6-sol`; Fixed RAG and Source-Agent use `deepseek-v4.1-flash`. Their historical calls use temperature zero, high image detail and reasoning effort, and a 900-second timeout. Recovery retains original failures and includes instruction variants, so completion does not establish equal cumulative budgets or identical decision targets.

Frozen US transfer. Official `deepseek-flash` generates matched Naive and refined reviews using the original US claims and images. Refined reviews reuse the existing US fixed-RAG evidence selections. The EP-learned instructions, TF-IDF vocabulary, predictor coefficients, and threshold are frozen; no US labels are used for fitting or selection. A shared US-jurisdiction instruction clarifies nonobviousness, description support, definiteness, and eligibility, and prevents EP-specific treatment exclusions from establishing a US defect. Missing claim numbers and formula images remain explicit. Generation retains the settings in Appendix G.2. The table also exposes the refined LLM prediction before applying the predictor: the frozen EP predictor increases not-granted recall but loses enough granted recall to lower BA relative to that component.

Table 6: US grant-outcome prediction coverage and confusion counts.

Method	Valid	BA (%)	TP	FP	FN	TN
Sol Naive	163	65.5	54	8	60	41
Gemini Naive	163	64.1	67	15	47	34
Qwen Naive	163	50.3	45	19	69	30
MiniMax Naive	163	53.1	63	24	51	25
Flash Naive	163	55.9	81	29	33	20
Hybrid-DOS	161	60.5	41	7	73	40
Direct-LC	162	57.7	39	9	75	39
Iterative-Visual	135	57.6	34	9	59	33
Fixed RAG	163	55.8	104	39	10	10
Source-Agent	163	64.8	85	22	29	27
Official Naive	163	53.2	112	45	2	4
EGPA (Ours)	163	57.2	56	17	58	32
EGPA before predictor	163	58.0	74	24	40	25
Always granted	163	50.0	114	49	0	0

The confusion counts in Table 6 treat granted as positive; the pre-predictor row is a component diagnostic. The always-granted baseline has 69.94% accuracy and 0.823 granted-class F1. Hybrid-DOS supplies 161 outputs, Direct-LC 162, and Iterative-Visual 135; the remaining displayed configurations cover all 163 cases. Unavailable outputs are excluded rather than treated as not granted. These denominators describe the retained panel; original outputs and failure histories from the parent collection remain preserved.

Nonempty references cover 308 of the panel's 815 possible case–dimension pairs: 93 novelty, 61 inventive step, 63 support, 85 clarity, and six eligibility. Empty and placeholder text is excluded. Missing outputs can further reduce a configuration's scored pairs. Some retained passages describe general allowance rather than dimension-specific findings, so coverage does not certify applicability. US uses the same `text-embedding-3-large` chunking, normalization, and character-length weighting as EP. Table 3 equally weights the dimension means; Table 7 reports pair-weighted means and claim-text controls on the same scored pairs. Neither weighting establishes performance on absent references. Similarity to examiner references and to claim text measures different forms of alignment; neither certifies correctness. ROUGE-L uses fixed-order concatenated references without stemming.

Table 7: US assessment similarity to examiner references and claim text.

		`text-embedding-3-large`		ROUGE-L F1	
Method	Pairs	Reference	Claim text	Reference	Claim text
Sol Naive	308	0.5805	0.5443	0.0902	0.0997
Gemini Naive	308	0.5534	0.5288	0.1018	0.0989
Qwen Naive	308	0.5850	0.5154	0.1029	0.1021
MiniMax Naive	308	0.5623	0.5214	0.0940	0.1037
Flash Naive	308	0.5956	0.5293	0.1000	0.1112
Hybrid-DOS	306	0.5895	0.5458	0.0885	0.0996
Direct-LC	308	0.5920	0.5362	0.0914	0.0977
Iterative-Visual	258	0.5852	0.5412	0.0873	0.1015
Fixed RAG	308	0.6067	0.5641	0.0996	0.1276
Source-Agent	308	0.6019	0.5482	0.0995	0.1219
Official Naive	308	0.5841	0.5378	0.0960	0.1228
EGPA (Ours)	308	0.6249	0.5606	0.1136	0.1690

H Sampled Decision-Basis Analysis

Figure 6 retains a separate analysis of the earlier 651-case EP evaluation set and the final saved grant prediction for each configuration–application pair. The initial sample was expanded without replacement from each configuration's unselected mismatches to 40 outputs per configuration. Cu-

mulative false-positive and false-negative quotas at 40 follow their within-configuration population percentages, rounded by the largest-remainder method; ties favor false negatives. The combined 440 outputs span 326 distinct applications, which may recur under different configurations. The group denominators are 200 Naive, 80 evidence-augmented, and 160 agentic sampled outputs. Iterative-Visual and Source-Agent appear in the agentic group alongside Fixed RAG and earlier Ours-C. This historical sample is not a new error analysis of EGPA or of the common 501-case tables.

One dominant basis was coded from each saved final grant prediction using the earlier twelve-category decision-basis taxonomy. Multiple equally decisive reasons, or an unstated dominant reason, retain their own category. Figure 6 shows all twelve categories and labels each count with its within-setting percentage. A zero means that no sampled output received that code; it does not establish that the category is absent from all mismatches. The class assignments were checked against exact response quotations and source hashes, but they were not independently double-coded or expert validated. The expanded sample was labeled by a single coder from the saved final-response rationale; these are descriptive labels, not expert judgments. In particular, this analysis does not establish the legal effect of cited art, the correctness of chemical readings, or the cause of outcome disagreement. Equal sampling per configuration and rounded direction quotas preclude extrapolating the plotted proportions to all mismatches.

I Full-Framework Analysis

The quantitative analysis uses the saved outputs and rescored similarities underlying Tables 3 and 4; it introduces no new generation, predictor fitting, or threshold search. Evaluation applications are the same across methods. Development comparisons use only the recorded selection split and the original BGE-M3 selection objective.

Distribution of text improvements. Table 8 reports paired changes in `text-embedding-3-large` cosine similarity. Each dimension uses its shared valid-reference mask. The table gives median changes and percentages of scored cases with positive changes; mean scores are in Table 3. Positive medians and improvement on a majority of cases occur against both controls in every dimension. Support has the fewest references and the smallest mean change. This is descriptive evidence about text alignment, not independent verification of reasoning or a population estimate over missing references.

Table 8: Paired assessment-similarity changes for EGPA against official DeepSeek controls.

		Versus Naive		Versus Fixed RAG	
Dimension	Cases	Median Δ	Improved (%)	Median Δ	Improved (%)
Novelty	479	+0.0491	84.1	+0.0222	69.5
Inventive step	488	+0.0735	90.0	+0.0353	79.5
Support	105	+0.0132	60.0	+0.0134	59.0
Clarity	400	+0.0396	84.2	+0.0296	79.5
Eligibility	324	+0.0216	74.1	+0.0217	75.9

Outcome trade-offs. Table 9 separates the two class recalls on the same 501 evaluation applications. BA and recalls are percentages, and granted is the positive class. Predictor controls use the same fitting and threshold-selection procedure. Each predictor reduces false-positive grant predictions but also increases false negatives. The gain is therefore a trade-off between class recalls, rather than uniform correction of both error types. EGPA versus Naive with the same predictor procedure has a 97.5% paired bootstrap BA interval of $[-1.10, 7.07]$ percentage points; an independent grant-prediction benefit from refinement is not established.

Table 9: Official DeepSeek component comparisons on the same EP evaluation applications.

Method	BA	Granted	Not granted	TP	FN	FP	TN
Naive	55.80	94.72	16.88	323	18	133	27
Fixed RAG	50.19	99.12	1.25	338	3	158	2
EGPA without predictor	52.72	97.95	7.50	334	7	148	12
Naive + predictor	62.15	77.42	46.88	264	77	85	75
Fixed RAG + predictor	60.64	80.65	40.62	275	66	95	65
EGPA (Ours)	65.14	76.54	53.75	261	80	74	86

Feedback scale and refinement. Table 10 shows all initial candidates and subsequent refinements evaluated on the same 30 selection applications. Round zero denotes the initial candidate; similarity uses the original BGE-M3 selection metric. Feedback sets are nested within training; every trained predictor still uses the full 120-case training partition. A is evidence-chain writing, B adds draft verification, and C selects available page identifiers before writing. A48 is retained by the joint ranking and tie-breaking procedure in Appendix G.2. Increasing feedback size does not monotonically improve selection performance within these families. Neither subsequent A48 refinement exceeds the initial candidate's joint score. All initial candidates fall below direct Naive in selection BA; the selected candidate is therefore the best under the recorded search criterion, not a demonstrated improvement on both selection metrics. These development observations neither rank candidates on the final evaluation set nor establish a generally optimal training size.

Table 10: Development comparisons on the selection set.

Candidate	Feedback cases	Round	BA (%)	BGE-M3	Joint score
Naive	–	–	63.33	0.6082	–
A48_r0	48	0	56.67	0.6297	-1.333
A48_r1	48	1	50.00	0.6164	-2.667
A48_r2	48	2	56.67	0.6200	-1.333
A96_r0	96	0	50.00	0.6102	-2.667
A120_r0	120	0	46.67	0.5881	-3.333
B48_r0	48	0	56.67	0.6028	-1.333
B96_r0	96	0	53.33	0.6011	-2.000
B120_r0	120	0	56.67	0.6078	-1.333
C48_r0	48	0	46.67	0.6306	-3.333
C96_r0	96	0	53.33	0.6133	-2.000
C120_r0	120	0	46.67	0.5631	-4.510

Illustrative change in a saved review. EP24195442, a retained example outside the common evaluation subset, illustrates how wording changes without changing the raw grant decision. It was chosen near the median inventive-step similarity increase in the original full-framework evaluation; it is an illustration, not a blinded correctness audit. Both requests use the same selected-claim input and an empty evidence packet. Naive describes the structural change as "modest" and suggests it "may be an obvious further option," while acknowledging that this is an inference. The refined review instead identifies the missing closest prior art and comparative effect, leaving inventive step unresolved. Both raw outputs predict granted, disagreeing with the recorded not-granted label. The example is consistent with the learned instructions on evidence sufficiency, but it neither verifies the model's chemical interpretation nor proves that a particular rule caused the change.

J Complete Target Collection Distribution

The labels in Figure 4 use the following abbreviations. EP: CDK, Cyclin-dependent kinases; CYP51, Sterol 14α-demethylase; BTK, Bruton tyrosine kinase; SHP2, Tyrosine-protein phosphatase non-receptor type 11; DHODH, Dihydroorotate dehydrogenase; Mpro, Main protease; BCL2, B-cell lymphoma 2; ACC, Acetyl-coenzyme A carboxylase; EZH2, Enhancer of zeste homolog 2; PI3K, Phosphoinositide 3-kinase. US: EGFR, Epidermal growth factor receptor; LPS transport, Lipopolysaccharide transport complex; BTK-PROTAC, Bruton tyrosine kinase–targeting proteolysis-targeting chimeras; Chitin synth., Chitin synthase; Exon 20 ins., Exon 20 insertion; Pyrimidine biosyn., Pyrimidine biosynthesis; AmB, Amphotericin B; Syn. lethality, Synthetic lethality; Thiophene/Pks13, Thiophene / polyketide synthase 13; Hsp90, Heat shock protein 90.

The distribution uses the accepted 851-case EP membership and its frozen outcome labels. Historical acquisition records provide traceable collection metadata for 795 applications across 67 displayed target groups after pooling CDK4, CDK6, CDK7, and CDK9 as CDK; 56 applications remain unrecorded. Assignments describe collection queries rather than verified biological targets. Groups are ranked by decreasing count with alphabetical tie-breaking: ACC, EZH2, and PI3K each contain 21 cases and all enter the Top 10 after the CDK merge. Figure 4 shows the Top 10; Figure 8 includes all groups and unrecorded cases. The latter all carry the frozen not-granted label; missing metadata should not be interpreted as biological evidence or a reason for non-grant. Historical and current review-tier distinctions are preserved, and these counts do not establish expert claim-level gold.

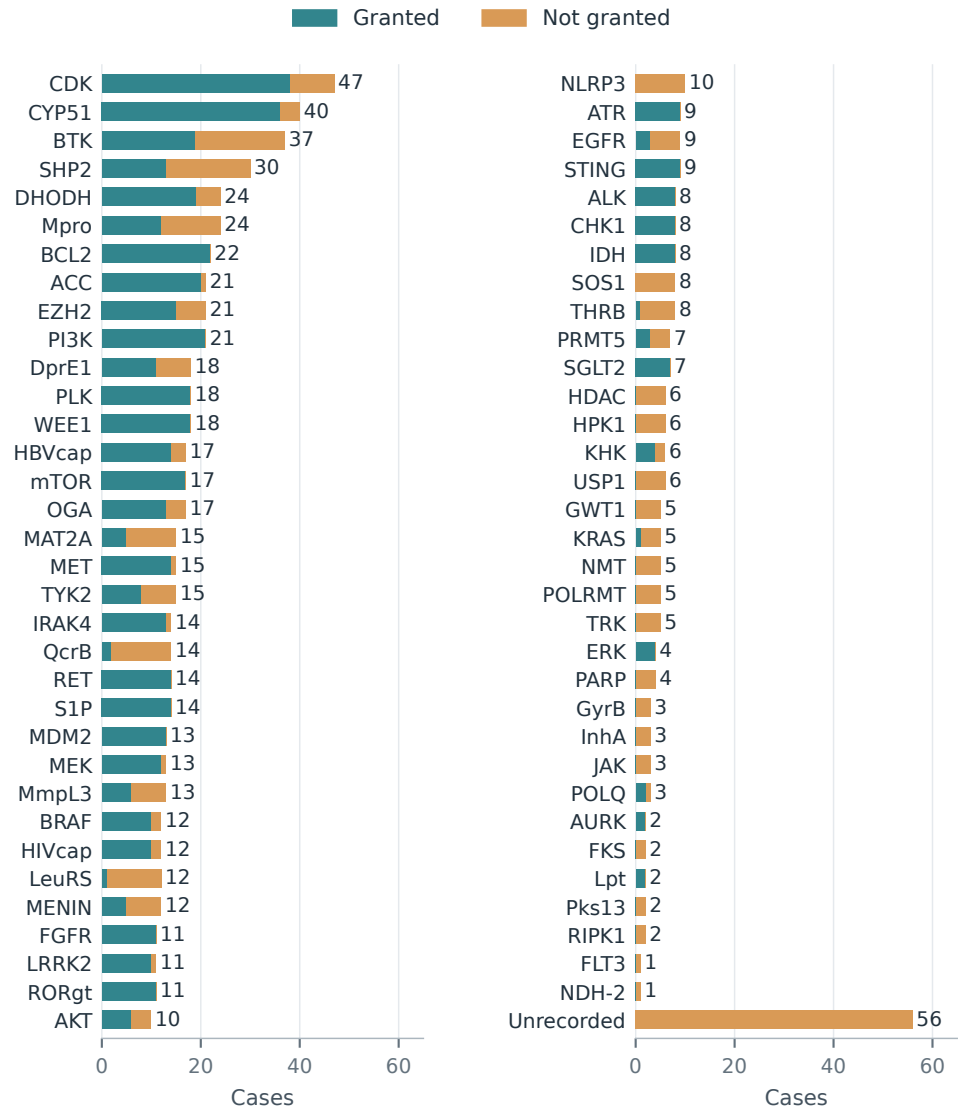


Figure 8: Complete distribution of target collection groups in the 851-case EP benchmark, partitioned by frozen grant-outcome label, with CDK4, CDK6, CDK7, and CDK9 pooled as CDK. Unrecorded cases are shown explicitly.

The parent US collection contains 308 applications with 259 granted and 49 not-granted labels; the collection figures retain all of these cases, whereas the result tables use the selected 163-case panel. Its source-directory records assign one collection keyword to every application, yielding 96 distinct groups. Figures 9 and 10 retain these original query groups, with long names abbreviated for display. Figure 4 pools the EGFR and epidermal growth factor receptor aliases into one group of distinct applications, yielding 95 display groups; its US Top 10 contains 82 cases, with 226 in other groups. No other US groups are merged, so the display is not a harmonized cross-jurisdiction target taxonomy. Ties use alphabetical ordering of the source labels after this alias merge. Figure 3 uses a post-hoc mapping of acquisition keywords to broad collection themes. The mapping assigns 184 applications to a theme and retains 124 as unclassified because their keywords are broad or cross-domain. All 308 applications remain in the denominator. These are collection-theme inferences, not verified patent indications; the original keyword groups and outcome labels are unchanged.

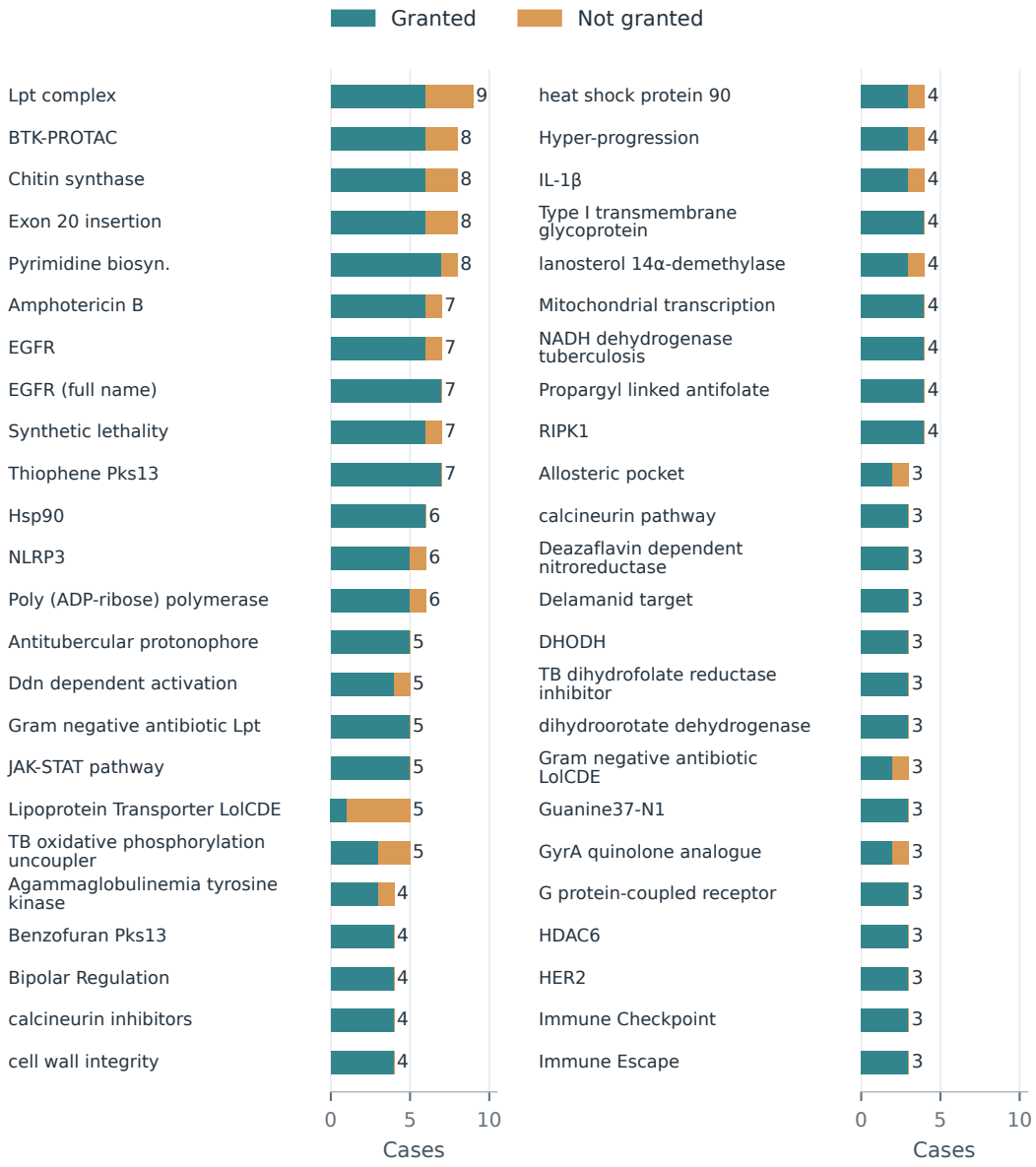


Figure 9: US collection keyword groups ranked 1–48, partitioned by recorded grant outcome. Groups describe acquisition queries, not verified biological targets.

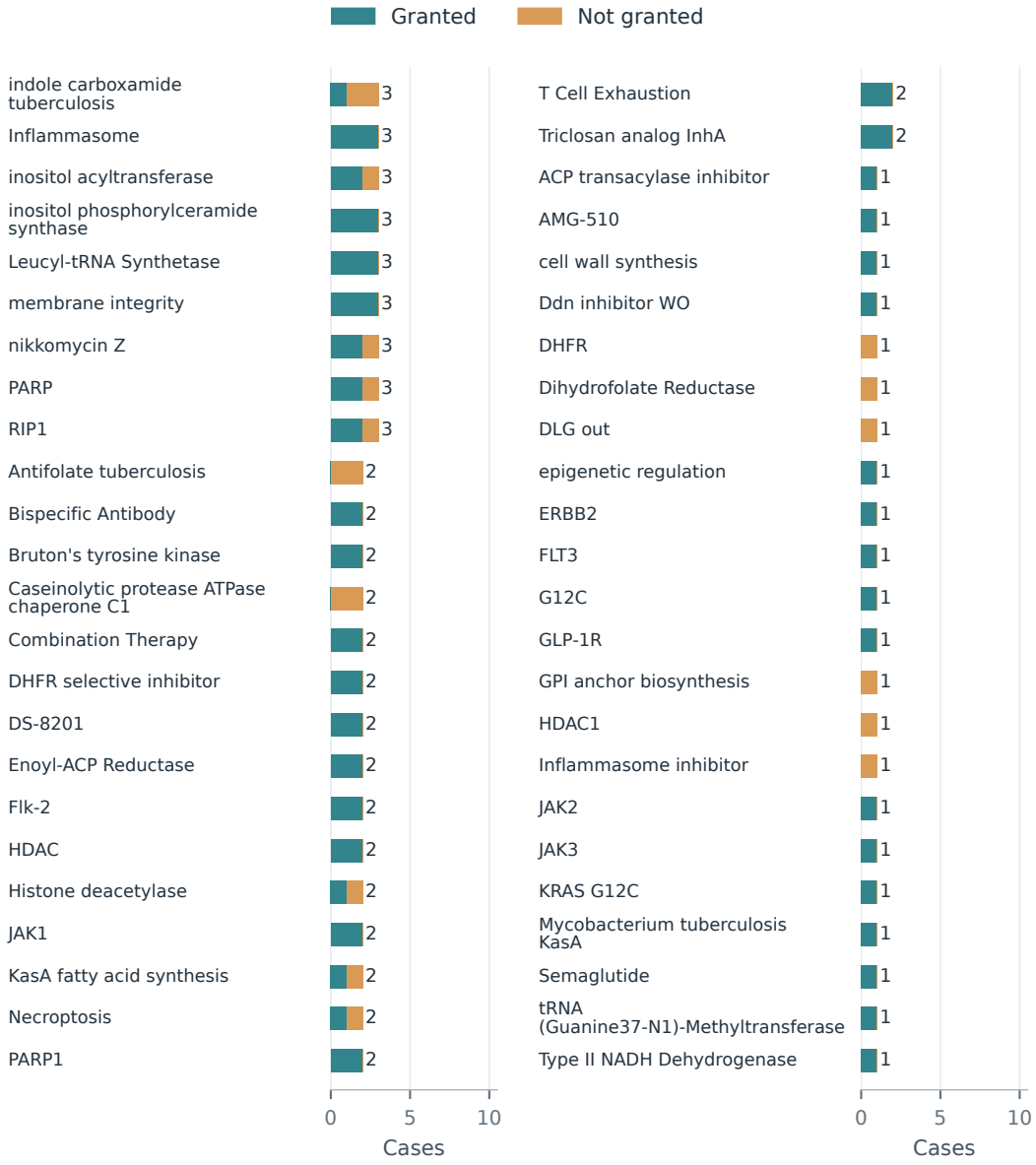


Figure 10: US collection keyword groups ranked 49–96, partitioned by recorded grant outcome. Groups describe acquisition queries, not verified biological targets.