OpenMTB-Audit: Exposing Over-Refusal and Clinical Expert Perspectives in LLM-Based Molecular Tumor Board Safety Evaluation

Negin Ashrafi[1], Jia Luo[2], Stacey M. Frumm[2], and Roxana Daneshjou[1,†]

[1]*Department of Biomedical Data Science, Stanford University, Stanford, CA, USA*
[2]*Lowe Center for Thoracic Oncology, Dana-Farber Cancer Institute, and Harvard Medical School, Boston, MA, USA*

[†]*Corresponding author: roxanad@stanford.edu*

Molecular tumor boards integrate genomic findings, clinical context, and therapeutic evidence to support precision oncology. As AI enters this workflow, a key safety challenge is distinguishing truly unsupported recommendations from evidence-supported options that still require oncologist review because of incomplete information, poor ECOG performance status, or other clinical caveats. We introduce OpenMTB-Audit, an open-source benchmark of 500 synthetic non-small cell lung cancer cases spanning five adversarial error categories and four safety labels: *Supported*, *Partially Supported*, *Unsupported*, and *Insufficient Information*. Across eight large language model configurations, we identify pervasive *over-refusal*: all LLM configurations failed to retain the *Partially Supported* label in 83.3–100% of true *Partially Supported* cases, achieving high aggregate safety scores through label collapse rather than clinically calibrated reasoning. To address this limitation, we developed MTB-AuditAgent, a deterministic seven-module framework separating evidence verification, missing-information detection, safety classification, and abstention. It reduces over-refusal to 6.7% and achieves 91.2% accuracy (95% CI: 88.6–93.6%). A two-oncologist annotation study found disagreement concentrated at the boundary between information sufficiency and treatment optimization, underscoring the need to preserve clinically meaningful distinctions.

Keywords: Molecular Tumor Board; LLM Safety; Over-Refusal; NSCLC; Clinical AI.

1. Introduction

Molecular tumor boards (MTBs) are an important component of precision oncology. The availability of more than fifteen biomarker-directed treatment options for non-small cell lung cancer (NSCLC) has increased the need for systematic, evidence-based case review.[1,2] Prospective studies have also associated greater molecular matching with improved outcomes in advanced cancers.[3] MTBs support this process by bringing together oncologists, molecular pathologists, pharmacists, radiologists, and other specialists to interpret genomic findings, evaluate therapeutic evidence, and formulate individualized recommendations.

Complex MTB workflows have driven interest in AI-assisted support. Large language models (LLMs) perform strongly on medical knowledge benchmarks and licensing-style examinations,[4–6] and have been explored for treatment recommendation, evidence retrieval, and automated review of MTB cases.[7–9] However, prior work has largely emphasized recommendation correctness, evidence retrieval, or detection of clearly unsafe outputs, with less attention to intermediate cases in which a therapy is evidence-supported but requires clinical review.

This distinction is clinically consequential. A therapy may be molecularly appropriate yet depend on performance status, incomplete documentation, prior treatment, or treatment

© 2026 The Authors. Open Access chapter published by World Scientific Publishing Company and distributed under the terms of the Creative Commons Attribution Non-Commercial (CC BY-NC) 4.0 License.

sequencing. A system that cannot represent this intermediate state may either endorse a recommendation without sufficient caution or reject an appropriate option without justification. Moreover, aggregate metrics may not capture safety in clinically consequential settings,[10,11] motivating task-specific evaluation.

Over-refusal is a failure mode of safety-aligned LLMs, in which models reject benign inputs near an overly conservative safety boundary.[12–14] In molecular tumor board auditing, it takes a distinct form: the model does not decline to respond but fails to preserve the *Partially Supported* label, most consequentially classifying evidence-supported therapies requiring review as *Unsupported* or *Supported*.

To our knowledge, over-refusal has not previously been quantified as a label-calibration failure in molecular tumor board auditing. Across eight LLM configurations evaluated on OpenMTB-Audit, structured approaches consistently fail to preserve the *Partially Supported* category, achieving high aggregate safety scores despite poor calibration at this clinical decision boundary. Our contributions are as follows:

(1) We introduce OpenMTB-Audit, an open-source benchmark of 500 synthetic NSCLC tumor board audit cases spanning five adversarial error categories and a clinically informed four-label safety framework derived from public evidence sources.
(2) We identify a task-specific form of *over-refusal* in molecular tumor board auditing, whereby aggregate safety metrics obscure systematic misclassification of evidence-supported recommendations requiring clinical review.
(3) We introduce MTB-AuditAgent, a deterministic seven-module audit pipeline that integrates evidence verification, missing-information detection, four-label classification, and human-referral abstention.
(4) We present expert concordance and qualitative disagreement analyses that identify information sufficiency and treatment optimization as key clinical decision boundaries.

2. Related Work

2.1. *LLMs for Molecular Tumor Board Support*

Recent work has explored large language models for precision-oncology and molecular tumor board workflows. Schmutz et al.[7] compared ChatGPT-4.0 recommendations with expert decisions on real-world oncology cases, identifying clinically relevant suggestions alongside variation in evidence prioritization. MEREDITH uses retrieval-augmented generation and expert-guided refinement to generate treatment options from biomedical literature, trial registries, drug approvals, and guidelines, but was evaluated on only ten fictional cases and focused on recommendation relevance.[8] MTBBench evaluates multimodal and longitudinal oncology decision-making and reports persistent challenges in hallucination, evidence reconciliation, and temporal reasoning.[9] In contrast, our work focuses on downstream safety auditing: given an existing recommendation, the system classifies it as supported, requiring clinical review, lacking sufficient information, or unsupported. This separates auditing from treatment generation and enables direct evaluation of label calibration.

2.2. *Clinical LLM Evaluation and Safety*

LLMs perform strongly on medical knowledge questions and licensing-style examinations,[4–6] but these evaluations do not capture clinical workflow safety requirements.[10] MedHELM shows substantial task-level variation, supporting application-specific evaluation.[10] Strong factual performance does not ensure clinical reliability: outputs may include unsupported claims,

harmful omissions, or inappropriate options,[11,15,16] and errors persist in specialist-consultation cases.[11]

2.3. *Over-Refusal and Safety Calibration*

Safety alignment can reduce harmful model behavior but may also produce over-refusal, in which models reject benign or permissible requests near a conservative safety boundary.[12–14] This has been studied in healthcare: CARES shows that models may both comply with harmful requests and over-refuse safe medical prompts,[17] and Health-ORSC-Bench extends this analysis to dual-use health prompts.[18]

The failure studied here differs from conventional textual refusal: in molecular tumor board auditing, a model may respond yet fail to assign the *Partially Supported* label to a recommendation that warrants qualified clinical review. We define this as *clinical over-refusal*—the systematic failure to correctly classify *Partially Supported* cases, whether misclassified as *Unsupported* (conservative escalation) or as *Supported* (without required caveats). To our knowledge, prior work has not quantified over-refusal as a label-calibration error in molecular tumor board auditing.

3. The OpenMTB-Audit Benchmark

3.1. *Design Rationale and Data Sources*

OpenMTB-Audit focuses on non-small cell lung cancer (NSCLC), a clinically important setting with 15 molecular profiles (EGFR, ALK, ROS1, BRAF, MET, KRAS, RET, NTRK, HER2, PD-L1, and others) and 20 FDA-labeled therapies, supported by regulatory indications, clinical guidelines, and public evidence resources.[1,2] This makes NSCLC well suited for auditable benchmark construction.

All benchmark cases were generated from publicly accessible sources containing no protected health information. CIViC[19] provided 464 accepted predictive biomarker–treatment evidence records, OpenFDA identified 20 NSCLC therapies with relevant FDA labeling, and ClinicalTrials.gov provided 146 active or completed NSCLC studies. We sampled 170 molecular profiles from publicly available NSCLC cohorts in cBioPortal,[20] including TCGA lung adenocarcinoma, TCGA lung squamous cell carcinoma, and the MSK-2015 cohort (data accessed January–March 2026). Evidence records were harmonized by normalizing gene and therapy names to a controlled vocabulary and linking records by molecular profile key; cBioPortal profiles were mapped to the same 15-profile vocabulary. All 500 patient summaries are fully synthetic; no electronic health records or protected health information were used.

3.2. *Case Construction*

OpenMTB-Audit contains 500 synthetic cases: 200 clean and 300 adversarial. Clean cases pair a therapeutically relevant molecular alteration with an evidence-supported treatment recommendation, varied in disease stage, histology, ECOG performance status, prior treatment, and brain metastasis status to produce clinically plausible combinations. The adversarial subset contains 60 cases per error category, each targeting a distinct failure mode (Table 1); representative examples are shown in Figure 1.

3.3. *Four-Label Safety Framework*

Each case is assigned a reference Support Label that captures both evidentiary support and the sufficiency of the available clinical information. The four labels are defined as follows.

Table 1. Adversarial error categories. Each category contains 60 cases.

Error Type	Description
Wrong Drug	The therapy lacks support for the reported molecular alteration.
Wrong Context	The therapy has evidence or regulatory support in a different cancer type, but not in the presented NSCLC context.
Missing Critical Information	One or more fields required to determine treatment appropriateness are absent, such as disease stage, ECOG performance status, prior treatment, or biomarker validation.
Hallucinated Variant	A variant of uncertain significance, non-sensitizing alteration, or otherwise unsupported biomarker is presented as clinically actionable.
Unsafe Overconfidence	The molecularly matched therapy may be appropriate, but the recommendation omits important caveats, such as ECOG performance status ≥ 3 or extensive prior treatment.

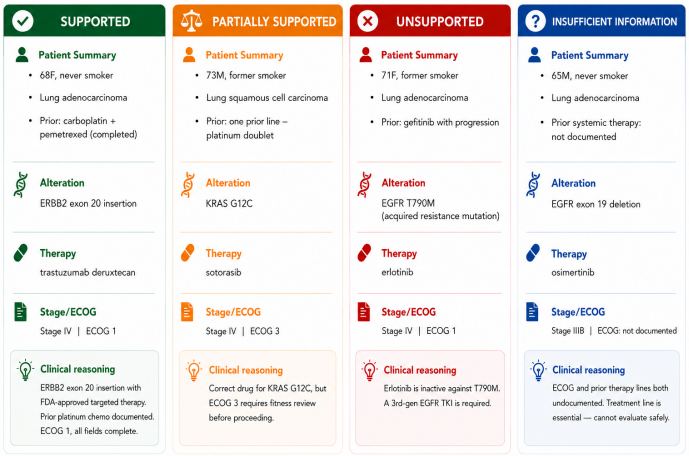


Fig. 1. Worked examples illustrating the four reference labels in OpenMTB-Audit. All cases are synthetic and contain no real patient data.

- **Supported:** The proposed therapy is supported by the reported molecular and disease context, and the information required to assess its appropriateness is available.
- **Partially Supported:** The proposed therapy has an appropriate molecular or clinical rationale, but patient-specific considerations require oncologist review before implementation. Examples include poor ECOG performance status or extensive prior treatment history.
- **Unsupported:** The therapy lacks adequate support for the molecular alteration or disease context, including mismatched drugs, therapies supported only in another cancer type, and recommendations based on non-actionable or hallucinated biomarkers.
- **Insufficient Information:** One or more critical fields are missing, preventing a reliable determination of whether the proposed therapy is appropriate.

Each case additionally includes a binary abstention target, missing-information indicators, and a structured audit rationale. The distinction between *Partially Supported* and *Unsupported* is central: *Partially Supported* indicates a valid therapeutic rationale requiring patient-specific review, whereas *Unsupported* indicates inadequate evidentiary support. Collapsing these labels

conflates uncertainty about implementation with lack of evidence. *Insufficient Information* warrants abstention rather than a definitive label. Expert assessment of these boundaries is reported in Section 8.

4. MTB-AuditAgent

MTB-AuditAgent is a deterministic seven-module framework for auditing proposed molecular tumor board recommendations rather than generating them. Given a patient summary and proposed therapy, it verifies molecular and clinical support, identifies missing information, determines whether human review is required, and produces a structured audit report. Because decision logic is rule-based after evidence retrieval, all predictions are fully transparent and reproducible (Figure 2).

M1. Case Parser. Extracts structured clinical variables from free-text case summaries, including cancer type, histology, stage, genomic alteration, proposed therapy, ECOG performance status, prior treatment, PD-L1 expression, and brain metastasis status. Potential variants of uncertain significance and non-actionable alterations are identified during parsing.

M2. Evidence Retriever. Retrieves supporting evidence from the benchmark knowledge base by matching the reported biomarker, disease context, and therapy to FDA approvals and CIViC records. This evidence supports subsequent verification.

M3. Evidence Verifier. Determines whether the proposed therapy is supported for the reported molecular alteration and clinical context. This module performs explicit biomarker-therapy concordance checking and provides the principal evidence signal used by the downstream safety classifier.

M4. Missing Information Detector. Evaluates the completeness of clinically relevant fields. Disease stage, ECOG performance status, prior treatment history, and biomarker validation are treated as critical variables because they directly affect treatment appropriateness, whereas features like PD-L1 expression and brain metastasis status are treated as contextual variables that refine, but do not necessarily prevent, safety assessment.

M5. Safety Classifier. Integrates evidence-verification and information-completeness signals to assign one of the four reference labels defined in the four-label safety framework. Classification is based on evidentiary support, the availability of critical clinical information, and whether additional clinical review is required.

M6. Abstention Module. Identifies cases that should be referred for oncologist review, including recommendations with missing critical information, uncertain biomarker actionability, poor ECOG performance status, or unsupported therapies.

M7. Audit Report Generator. Produces a structured audit report containing the predicted safety label, abstention recommendation, missing-information analysis, supporting evidence, and a concise clinical rationale.

5. Experimental Setup

We evaluated MTB-AuditAgent and eight LLM-based baselines on all 500 benchmark cases. The baselines comprise progressively stronger prompting strategies designed to isolate the contribution of retrieval, structured reasoning, and explicit verification. All LLM experiments were conducted via the OpenAI API from June to July 2026, using GPT-4o (`gpt-4o-2024-08-06`) and GPT-4o-mini (`gpt-4o-mini-2024-07-18`) with temperature=0 and max_tokens=512. MTB-AuditAgent makes no LLM calls at inference time; all seven modules are deterministic and rule-based following evidence retrieval. Full prompts, evaluation code, and benchmark data are available at `https://github.com/Negiiinx/OpenMTB-Audit`.

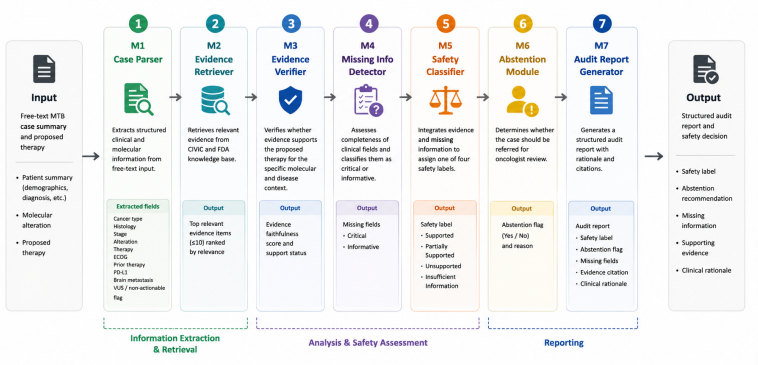


Fig. 2. MTB-AUDITAGENT workflow from free-text case input through seven deterministic modules to a structured safety audit.

- **Base LLM:** Zero-shot prompting with GPT-4o-mini and GPT-4o.
- **Simple RAG:** Base prompting augmented with the top five CIViC evidence records.
- **RAG+EV:** Simple RAG with an explicit seven-step evidence verification prompt.
- **Prompt-Checklist:** A structured six-step auditing prompt that explicitly reasons over the four benchmark safety labels, evaluated with both GPT-4o-mini and GPT-4o.
- **MTB-AuditAgent:** A deterministic seven-module audit pipeline without LLM inference.

6. Evaluation Metrics

Clinical errors differ substantially in their potential impact on patient care. Incorrectly endorsing an unsupported therapy is considerably more serious than failing to identify missing information or misclassifying a partially supported recommendation. To reflect these differences, we define a weighted Safety Score ranging from 0 to 100:

$$\mathrm{SS} = 100 \times \Big(0.30(1 - R_{\mathrm{miss}}) + 0.20F1_{\mathrm{abs}} + 0.20R_{\mathrm{MI}} + 0.15C_{\mathrm{WD}} + 0.10C_{\mathrm{Hall}} + 0.05C_{\mathrm{OC}}\Big) \qquad (1)$$

where R_{miss} denotes the fraction of unsupported therapies incorrectly classified as Supported, $F1_{\mathrm{abs}}$ is the abstention F1 score, R_{MI} is the fraction of Missing Critical Information cases assigned any non-*Supported* label (i.e., correctly flagged by the system), and C_{WD}, C_{Hall}, and C_{OC} denote catch rates (fraction of cases not labeled *Supported*) for Wrong Drug, Hallucinated Variant, and Unsafe Overconfidence errors, respectively. Wrong Context cases are labeled *Unsupported* in the ground truth; endorsing them as *Supported* constitutes a false approval and is penalized through R_{miss}. Weights reflect clinical risk priority: false approval (R_{miss}, weight 0.30) carries the highest penalty because it risks harm through active endorsement of an unsupported therapy; abstention and missing-information detection are weighted equally (0.20 each) as both affect whether clinically uncertain cases reach appropriate review; specific adversarial catch rates are weighted by clinical severity. A sensitivity analysis varying each weight by ± 0.05 with proportional redistribution across remaining components confirmed that the relative ranking of all evaluated systems remained stable across all perturbations. In addition to the Safety Score, we report overall accuracy, macro F1, and class-specific precision and recall (Table 3; per-error-type recall for all nine systems in Figure 4). MTB-AUDITAGENT's full predictions and all baseline prompts are in the public repository. Confidence intervals for accuracy were estimated using 10,000 bootstrap resamples, and paired model comparisons were performed using McNemar's test.

Table 2. Clinical over-refusal and false-approval rates. Over-refusal is $1 - \text{recall}_{\text{PS}}$: the proportion of ground-truth Partially Supported cases not predicted as Partially Supported (misclassified as Supported or Unsupported). False approval denotes Unsupported cases classified as Supported.

System	Over-Refusal	False Approval
GPT-4o-mini		
Base LLM	83.3%	2.2%
Simple RAG	100.0%	0.0%
RAG+EV	100.0%	0.0%
Prompt-Checklist	100.0%	2.2%
GPT-4o		
Base LLM	98.3%	1.7%
Simple RAG	100.0%	1.1%
RAG+EV	98.3%	0.0%
Prompt-Checklist	100.0%	1.7%
MTB-AuditAgent (ours)	**6.7%**	**0.0%**

7. Over-Refusal: A Failure of Clinical Safety Calibration

7.1. *Definition and Quantification*

The principal failure mode identified in this study is *clinical over-refusal*: failing to preserve the *Partially Supported* label for an evidence-supported therapy that requires clinical review. Unlike conventional refusal, the model does not decline to respond; instead, it assigns a therapeutically valid recommendation to an incorrect category.

We operationalize over-refusal as the proportion of ground-truth *Partially Supported* cases not correctly predicted as *Partially Supported* (equivalently, $1 - \text{recall}_{\text{PS}}$), including cases misclassified as either *Unsupported* or *Supported*. In OpenMTB-Audit, these 60 cases correspond to the Unsafe Overconfidence category: molecularly appropriate therapies that require oncologist review because of factors such as ECOG performance status ≥ 3 or extensive prior treatment. They therefore distinguish uncertainty about treatment implementation from lack of therapeutic evidence.

All LLM configurations were explicitly instructed that *Partially Supported* was a valid label and were provided its operational definition, including the ECOG threshold and examples of patient-specific caveats. Despite this, on the 60 ground-truth *Partially Supported* cases, all configurations almost never assigned the label correctly (Table 2), misclassifying cases as *Unsupported* or *Supported* depending on the configuration. Among GPT-4o configurations, the 98.3% and 100% rates were not statistically distinguishable (two-sided Fisher's exact $p = 1.00$, 59/60 vs. 60/60), indicating near-complete failure across all LLM settings.

7.2. *Clinical Consequences*

Clinical over-refusal is not apparent from aggregate Safety Scores alone. Table 3 reports class-specific precision and recall for the highest-accuracy LLM baseline and MTB-AuditAgent. Although their Safety Scores of 90.1 and 92.4 are close, their clinical error profiles differ substantially.

Prompt-Checklist GPT-4o correctly identifies 98% of ground-truth *Unsupported* cases yet fails to assign the *Partially Supported* label to any of the 60 ground-truth PS cases, 31 are misclassified as *Supported* and 29 as *Unsupported*, reducing *Unsupported* precision to 0.75 (one-quarter of its *Unsupported* calls are false positives, of which 29 are misclassified PS cases). MTB-AuditAgent achieves 0.91 *Unsupported* precision and 0.93 *Partially Supported* recall,

Table 3. Per-class precision and recall for the highest-accuracy LLM baseline and MTB-AUDITAGENT. Similar aggregate Safety Scores conceal differences in class-specific performance, particularly for Partially Supported cases.

Label	Prompt-Checklist GPT-4o		MTB-AuditAgent	
	Prec	Rec	Prec	Rec
Supported	0.97	0.75	0.98	0.84
Partially Supported	0.00	0.00	**0.71**	**0.93**
Unsupported	0.75	0.98	**0.91**	**1.00**
Insufficient Info	0.97	0.58	**1.00**	**0.87**

preserving the clinically meaningful distinction. Figure 3 summarizes this trade-off.

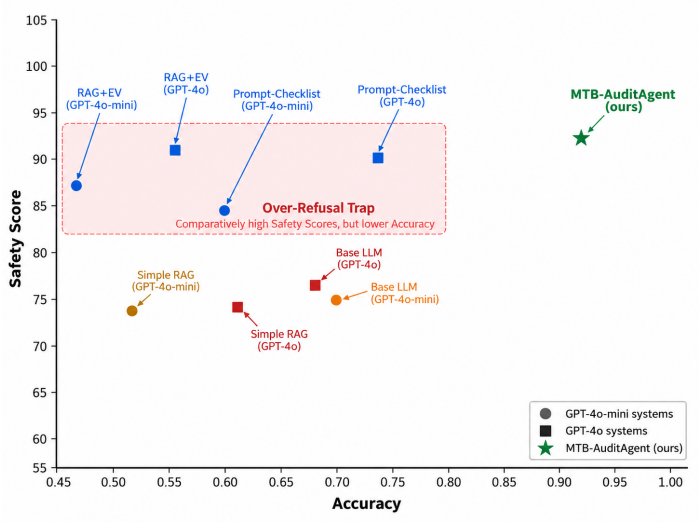


Fig. 3. Safety Score versus accuracy across systems. Structured LLMs show high safety but lower accuracy because of over-refusal, whereas MTB-AUDITAGENT achieves both.

8. Expert Annotation Study

8.1. *Study Design*

To assess the clinical plausibility of the benchmark reference labels and to characterize areas of expert disagreement, two oncologists with molecular tumor board experience who routinely care for patients with NSCLC independently annotated a stratified subset of 50 benchmark cases. Annotators were blinded to AI predictions, benchmark labels, and each other's assessments. The subset contained 20 Unsupported cases and 10 cases from each of the other three reference labels. It was intentionally enriched for difficult examples, with 30 of the 50 cases selected from cases exhibiting disagreement among the evaluated AI systems.

Each oncologist assigned one of the four benchmark labels (*Supported*, *Partially Supported*, *Unsupported*, or *Insufficient Information*) and provided free-text clinical reasoning for every case. Throughout this paper they are referred to as Oncologist A and Oncologist B. Because this annotation subset was reviewed during development of MTB-AUDITAGENT, agreement between MTB-AUDITAGENT and the expert annotations is reported as post-refinement con-

cordance rather than held-out validation. The primary purpose of the study is therefore to assess the clinical plausibility of the benchmark labels and to characterize the clinical reasoning boundaries that generate disagreement. Disagreements between the two oncologists were not adjudicated to consensus; they were preserved as study data to characterize naturally occurring expert disagreement at clinical decision boundaries, which is itself a study finding.

8.2. *Agreement and Sources of Disagreement*

Oncologist A demonstrated complete agreement with the benchmark reference labels (50/50, $\kappa = 1.000$), showing that the study-defined labeling framework can be applied consistently by an independent expert. Oncologist B agreed with the benchmark on 34 of 50 cases ($\kappa = 0.553$); because Oncologist A agreed perfectly with the benchmark, inter-oncologist agreement (A-vs-B) is identical to B-vs-benchmark by construction (Table 4). According to the interpretation of Landis and Koch,[21] this corresponds to moderate agreement. The difference may reflect both variation in clinical judgment and differences in whether the annotators applied the operational label definitions strictly or interpreted the cases according to preferred clinical practice.

Table 4. Agreement with benchmark labels and per-label recall on the 50-case expert review subset.[21]

Comparison	**Agreement**	**κ (Level)**	**Oncologist B Label**	**Recall**
Oncologist A vs. Benchmark	50/50	1.000 (Perfect)	Unsupported	19/20(95%)
Oncologist B vs. Benchmark	34/50	0.553 (Moderate)	Partially Supported	8/10(80%)
Oncologist A vs. Oncologist B	34/50	0.553 (Moderate)	Supported	6/10(60%)
			Insufficient Info	1/10(10%)

Importantly, disagreement was structured rather than random. Fifteen of the 16 discordant cases concerned information sufficiency or treatment optimization, while one involved an Unsupported reference-label case. Three recurring patterns accounted for 12 of the 16 disagreements: (1) missing clinical information (7 cases), where Oncologist A applied strict operational definitions requiring ECOG and prior treatment, while Oncologist B accepted available molecular evidence; (2) treatment optimization in ROS1 fusion-positive NSCLC (3 cases), where crizotinib was labeled *Supported* vs. *Partially Supported* depending on whether optimality vs. approval status was primary; and (3) ECOG 3 cases (2 cases) resolved via drug-specific tolerability judgment vs. required clinical review. These findings identify a challenging labeling boundary reflecting differences in clinical judgment and annotation interpretation.

8.3. *Implications for AI Safety*

The disagreement patterns contextualize LLM over-refusal: structured prompting strategies failed to preserve *Partially Supported*, collapsing these cases into *Unsupported* or *Supported* depending on the configuration, rather than distinguishing those requiring clinical review from those lacking evidentiary support. MTB-AuditAgent routes missing information to *Insufficient Information* and evidence-supported therapies requiring review to *Partially Supported* with an abstention recommendation. On the 50 expert-reviewed cases, MTB-AuditAgent achieved post-refinement $\kappa = 0.833$ versus $\kappa = 0.451$ for Prompt-Checklist (both vs. benchmark), correctly classifying all ten *Partially Supported* cases (Prompt-Checklist GPT-4o classified all ten as *Unsupported* within this subset). The primary challenge is not identifying

Table 5. Performance on OpenMTB-Audit ($N = 500$) under the four-label framework. Bold indicates the best value in each column.

System	Accuracy	Macro F1	Abstain F1	MI Catch	Safety Score
GPT-4o-mini					
Base LLM	0.694	0.499	0.348	0.567	74.9
Simple RAG	0.524	0.318	0.346	0.467	73.8
RAG+EV	0.462	0.317	0.832	0.650	87.0
Prompt-Checklist	0.600	0.422	0.851	0.567	83.9
GPT-4o					
Base LLM	0.676	0.441	0.392	0.600	76.4
Simple RAG	0.608	0.402	0.451	0.433	74.1
RAG+EV	0.554	0.402	0.871	0.800	**90.9**
Prompt-Checklist	0.740	0.601	0.884	0.783	90.1
MTB-AuditAgent	**0.912**	**0.898**	0.688	**0.933**	**92.4**

clearly incorrect recommendations—on which oncologists agreed almost unanimously—but distinguishing therapies requiring clinical review from those genuinely lacking evidentiary support. When AI and expert assessments diverge, clinicians should consult the M7 audit report; no flagged case should be acted upon without attending oncologist review.

9. Full Evaluation Results

9.1. *Main Results*

Table 5 summarizes the performance of all evaluated systems under the four-label framework. MTB-AuditAgent achieved an accuracy of 91.2% (95% CI: 88.6–93.6%) and a Safety Score of 92.4. The baseline with the highest Safety Score, RAG+EV with GPT-4o, achieved 55.4% accuracy and a Safety Score of 90.9. MTB-AuditAgent improved accuracy by 35.8 percentage points while also improving the aggregate Safety Score. The bootstrap confidence intervals did not overlap, and a paired McNemar test confirmed that the difference in case-level accuracy was statistically significant ($b = 197$, $c = 18$, $\chi^2 = 147.4$, $p < 0.001$).

Model scale alone produced modest safety gains relative to task structure. GPT-4o outperformed GPT-4o-mini across all matched strategies, but margins were small for unstructured configurations (Base LLM: 76.4 vs. 74.9; Simple RAG: 74.1 vs. 73.8). Structured evidence verification improved safety substantially: RAG+EV reached 87.0 and 90.9 for GPT-4o-mini and GPT-4o respectively, and Prompt-Checklist reached 83.9 and 90.1. However, the four structured LLM configurations (Safety Scores 83.9–90.9) had accuracies of only 46.2–74.0%, reflecting label collapse (Figure 3). MTB-AuditAgent combined a Safety Score of 92.4 with 91.2% accuracy, demonstrating that explicit evidence verification, information-sufficiency assessment, and clinical-review routing can preserve safety without collapsing the *Partially Supported* category.

9.2. *Module Ablation*

Table 6 summarizes the contribution of each major MTB-AuditAgent module. The ablation results demonstrate that the modules support complementary functions rather than providing interchangeable improvements.

Removing the Evidence Verifier (M3) caused the largest drop (−55.0 points), confirming that evidence concordance is the principal determinant of safety. Removing M4 degraded Macro F1 from 0.898 to 0.638 (−17.3 SS points): without it, cases with missing critical fields

Table 6. Ablation analysis of MTB-AUDITAGENT on OPENMTB-AUDIT. Values in parentheses indicate the change in Safety Score relative to the full system.

Configuration	Accuracy	Macro F1	Safety Score
Full MTB-AUDITAGENT	**0.912**	**0.898**	**92.4**
w/o Evidence Verifier (M3)	0.552	0.586	37.4 (−55.0)
w/o Missing Information Detector (M4)	0.808	0.638	75.1 (−17.3)
w/o ECOG Rule (M5)	0.800	0.667	87.8 (−4.6)
w/o Abstention Module (M6)	0.912	0.898	78.7 (−13.7)

cannot be distinguished from evaluable recommendations, collapsing the *Insufficient Information* class. Removing the ECOG rule from M5 reduced the Safety Score by 4.6 points, with disproportionate accuracy and F1 losses, illustrating how aggregate metrics can understate clinically important misclassification. Removing M6 left accuracy and Macro F1 unchanged (M6 does not alter four-class labels) but reduced the Safety Score by 13.7 points, equal to $100 \times 0.20 \times F1_{\text{abs}}$. The R_{MI} catch rate remains unchanged because M4 assigns the *Insufficient Information* label directly; abstention provides a redundant safety signal for borderline cases but is not the primary vehicle for MCI detection. Together, the modules address complementary aspects of safe auditing (Figure 4).

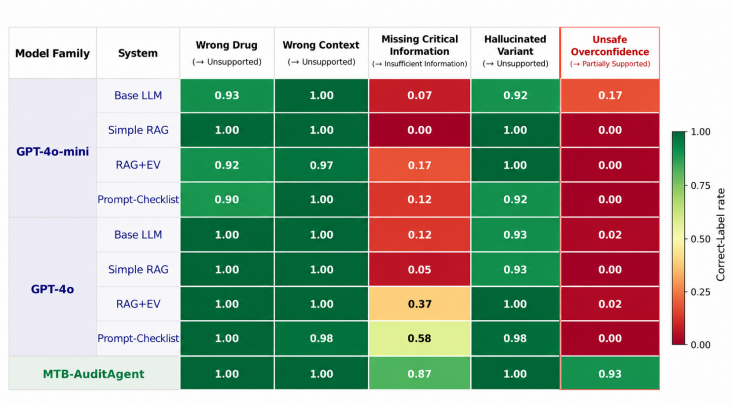


Fig. 4. Correct-label rates by error type ($N = 60$ each), with the largest gap on *Unsafe Overconfidence*.

9.3. *Benchmark Validity*

A potential concern is that MTB-AUDITAGENT uses the same evidence knowledge base for benchmark construction and inference. We examined this through two controlled experiments. First, in a text-only setting, the structured molecular profile identifier was removed, requiring MTB-AUDITAGENT to infer the alteration from the free-text case description. Performance remained nearly unchanged, with 92.2% accuracy and a Safety Score of 94.5, indicating that the system does not rely on oracle access to the benchmark profile. Second, we withheld five molecular profiles from the inference knowledge base. Across the 165 affected cases, MTB-AUDITAGENT achieved a Safety Score of 91.7 (95% CI: 90.1–93.1) and detected all predefined safety violations. Accuracy decreased to 47.9% because previously supported recommendations

were conservatively reclassified as *Insufficient Information*. This is the intended failure mode of a clinical auditing system: when evidence is unavailable, the system abstains rather than issuing unsupported judgments.

10. Discussion

Aggregate safety metrics are insufficient for evaluating AI systems that audit molecular tumor board recommendations: structured LLM configurations achieved comparatively high Safety Scores (84–91) while systematically failing to preserve the *Partially Supported* category, overestimating clinical reliability by obscuring performance on ambiguous cases. Expert annotation confirms that disagreement concentrates at the boundary between evidence sufficiency and clinical decision-making, on clearly incorrect recommendations, oncologists agreed almost unanimously, while disagreements concerned missing information, treatment optimization, and performance status. MTB-AuditAgent addresses this by explicitly modeling evidence validity, information sufficiency, and review need as separate decision states, enabling high safety with greater accuracy than prompting-based approaches.

Several limitations remain. First, OpenMTB-Audit covers only NSCLC and 15 molecular profiles; extension requires additional evidence knowledge bases. Second, rare alterations and complex co-mutation scenarios remain underrepresented. Third, only GPT-4o and GPT-4o-mini were evaluated; open-source and medical-specialized LLMs should be included to determine whether over-refusal generalizes across model families. Fourth, because the 50-case expert subset informed MTB-AuditAgent refinement, reported agreement reflects post-refinement concordance rather than prospective external validation, and independent multi-institutional validation is needed. Fifth, benchmark labels, evidence knowledge base, and decision rules share common design provenance, so performance on OpenMTB-Audit should not be interpreted as independent clinical validation. Sixth, the four-label taxonomy is one operationalization; multi-axis representations encoding evidentiary support, information sufficiency, and review need warrant future study. Seventh, the Safety Score weights embody judgments about relative harm that may require recalibration for other clinical settings. Finally, LLM baseline runtime and API costs were not characterized.

11. Conclusion

We introduce OpenMTB-Audit and MTB-AuditAgent, an open-source benchmark and deterministic audit framework for evaluating AI-assisted molecular tumor board safety in NSCLC. All LLM configurations failed to retain the *Partially Supported* label in 83.3–100% of true *Partially Supported* cases, achieving high aggregate Safety Scores through label collapse. Expert annotation showed moderate inter-annotator agreement ($\kappa = 0.553$), with disagreements concentrated at the boundary between information sufficiency and treatment optimization; Oncologist A fully agreed with the benchmark labels ($\kappa = 1.000$), supporting their plausibility.

By separating evidence verification, missing-information assessment, clinical-risk routing, and abstention, MTB-AuditAgent preserved these clinically meaningful distinctions, achieving 91.2% accuracy (95% CI: 88.6–93.6%), a Safety Score of 92.4, 6.7% over-refusal, and 91% precision on the *Unsupported* class. Its improvement over the best-safety LLM baseline was statistically significant (McNemar $p < 0.001$), with non-overlapping bootstrap confidence intervals. We release OpenMTB-Audit, the benchmark cases, expert annotations, evaluation code, and the MTB-AuditAgent framework as an open-source resource for evaluating whether systems avoid unsafe approvals while preserving clinically meaningful distinctions.

Acknowledgments

The benchmark incorporates CIViC[19] and cBioPortal[20] data under the Open Data Commons Open Database License, together with publicly available openFDA and ClinicalTrials.gov data.

References

1. S. Guo *et al.*, Biomarker-targeted therapies in non-small cell lung cancer: current status and perspectives, *Cells* **11**, p. 3200 (2022).
2. L. E. L. Hendriks *et al.*, Oncogene-addicted metastatic non-small-cell lung cancer: ESMO clinical practice guideline for diagnosis, treatment and follow-up, *Annals of Oncology* **34**, 339 (2023).
3. J. K. Sicklick *et al.*, Molecular profiling of cancer patients enables personalized combination therapy: the I-PREDICT study, *Nature Medicine* **25**, 744 (2019).
4. K. Singhal *et al.*, Large language models encode clinical knowledge, *Nature* **620**, 172 (2023).
5. H. Nori *et al.*, Capabilities of GPT-4 on medical challenge problems, *arXiv preprint arXiv:2303.13375* (2023).
6. A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan and D. S. W. Ting, Large language models in medicine, *Nature Medicine* **29**, 1930 (2023).
7. M. Schmutz, S. Sommer, C. Sander, D. Graumann, J. Raffler, I. Soto-Rey, S. Sheikhalishahi, L. Schmidt, L. P. Unkelbach, L. Ortak *et al.*, Large language model processing capabilities of ChatGPT 4.0 to generate molecular tumor board recommendations—a critical evaluation on real world data, *The Oncologist* **30**, p. oyaf293 (2025).
8. J. Lammert, T. Dreyer, S. Mathes, L. Kuligin, K. J. Borm, U. A. Schatz, M. Kiechle, A. M. Lörsch, J. Jung, S. Lange *et al.*, Expert-guided large language models for clinical decision support in precision oncology, *JCO Precision Oncology* **8**, p. e2400478 (2024).
9. K. Vasilev *et al.*, MTBBench: A multimodal sequential clinical decision-making benchmark in oncology, in *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2511.20490.
10. S. Bedi *et al.*, Holistic evaluation of large language models for medical tasks with MedHELM, *Nature Medicine* **32**, p. 943 (2026).
11. D. Wu *et al.*, First, do NOHARM: Towards clinically safe large language models, *arXiv preprint arXiv:2512.01241* (2025).
12. P. Röttger, H. R. Kirk, B. Vidgen, G. Attanasio, F. Bianchi and D. Hovy, XSTest: A test suite for identifying exaggerated safety behaviours in large language models, in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*, 2024.
13. J. Cui, W.-L. Chiang, I. Stoica and C.-J. Hsieh, OR-Bench: An over-refusal benchmark for large language models, in *Proceedings of the 42nd International Conference on Machine Learning*, 2025. arXiv:2405.20947.
14. L. Pan, Y. Tong, X. Zhang, X. Zhang, J. Zhou and Z. Chu, Understanding and mitigating overrefusal in LLMs from an unveiling perspective of safety decision boundary, in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, (Association for Computational Linguistics, Suzhou, China, 2025).
15. H. Alkaissi and S. I. McFarlane, Artificial hallucinations in ChatGPT: implications in scientific writing, *Cureus* **15**, p. e35179 (2023).
16. M. Sallam, ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns, *Healthcare* **11**, p. 887 (2023).
17. S. Chen *et al.*, CARES: Comprehensive evaluation of safety and adversarial robustness in medical large language models, in *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2505.11413.
18. Z. Zhang *et al.*, Health-ORSC-Bench: A benchmark for measuring over-refusal and safe completion in health contexts, in *Findings of the Association for Computational Linguistics: ACL 2026*, 2026. arXiv:2601.17642.
19. A. M. Danos *et al.*, Standard operating procedure for curation and clinical interpretation of variants in cancer, *Genome Medicine* **11**, p. 76 (2019).
20. E. Cerami *et al.*, The cBio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data, *Cancer Discovery* **2**, 401 (2012).

21. J. R. Landis and G. G. Koch, The measurement of observer agreement for categorical data, *Biometrics* **33**, 159 (1977).