# CancerGUIDE: Cancer Guideline Understanding via Internal Disagreement Estimation

**Alyssa Unell**[1,2] **Noel C. F. Codella**[3*] **Sam Preston**[3*] **Peniel Argaw**[1*]
**Wen-wai Yim**[3] **Zelalem Gero**[1] **Cliff Wong**[1] **Eric Horvitz**[4] **Amanda K. Hall**[1]
**Ruican Rachel Zhong**[1] **Jiachen Li**[1] **Shrey Jain**[3] **Mu Wei**[3]
**Matthew Lungren**[3**] **Hoifung Poon**[1**]

[1]Microsoft Research [2]Stanford University [3]Microsoft Health and Life Sciences [4]Microsoft
[*] Project Mentors [**] Equal Leadership

## Abstract

The National Comprehensive Cancer Network (NCCN) provides evidence-based guidelines for cancer treatment. Translating complex patient presentations into guideline-compliant treatment recommendations is time-intensive, requires specialized expertise, and is prone to error. Advances in large language model (LLM) capabilities promise to reduce the time required to generate treatment recommendations and improve accuracy. We present an LLM agent-based approach to automatically generate guideline-concordant treatment trajectories for patients with non-small cell lung cancer (NSCLC). Our contributions are threefold. First, we construct a novel longitudinal dataset of 121 cases of NSCLC patients that includes clinical encounters, diagnostic results, and medical histories, each expertly annotated with the corresponding NCCN guideline trajectories by board-certified oncologists. Second, we demonstrate that existing LLMs possess domain-specific knowledge that enables high-quality proxy benchmark generation for both model development and evaluation, achieving strong correlation (Spearman coefficient $r = 0.88$, RMSE = 0.08) with expert-annotated benchmarks. Third, we develop a hybrid approach combining expensive human annotations with model consistency information to create both the agent framework that predicts the relevant guidelines for a patient, as well as a meta-classifier that verifies prediction accuracy with calibrated confidence scores for treatment recommendations (AUROC=0.804). Calibrated confidence scoring is a critical capability for communicating the accuracy of outputs, custom-tailoring tradeoffs in performance, and supporting regulatory compliance. This work establishes a framework for clinically viable LLM-based guideline adherence systems that balance accuracy, interpretability, and regulatory requirements while reducing annotation costs, providing a scalable pathway toward automated clinical decision support. Code and synthetic patient data are made available here: CancerGUIDE repository.

## 1 Introduction

Cancer treatment decisions require oncologists to synthesize complex patient histories with evolving clinical guidelines to recommend the appropriate next treatments. The National Comprehensive Cancer Network (NCCN) guidelines provide evidence- and consensus-based recommendations for cancer diagnosis, treatment, and management [1], with adherence promoting higher quality and consistent care between providers [2; 3; 4]. However, guideline navigation presents significant challenges: the guidelines are extensive, frequently updated as new research emerges, and require

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance.

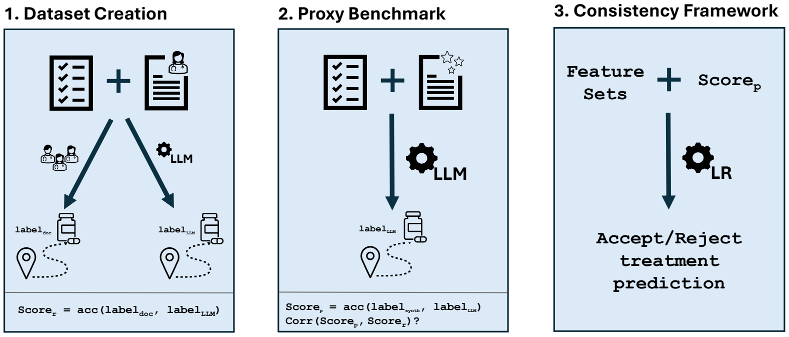

1. Dataset Creation
2. Proxy Benchmark
3. Consistency Framework
LLM
$label_{doc}$
$label_{LLM}$
$Score_r$ = acc($label_{doc}$, $label_{LLM}$)
LLM
$label_{LLM}$
$Score_p$ = acc($label_{synth}$, $label_{LLM}$)
Corr($Score_p$, $Score_r$)?
Feature Sets
+
$Score_p$
LR
Accept/Reject treatment prediction

Figure 1: **CancerGUIDE framework.** (1) Clinicians derive patient pathways ($label_{doc}$) from NCCN guidelines and real notes. These serve as gold-standard references to compare with LLM-derived pathways ($label_{LLM}$), producing a reference accuracy score ($Score_r$). (2) NCCN guidelines and both synthetic and real clinical notes are used to generate weak labels ($label_{synth}$). LLM predictions ($label_{LLM}$) are compared to these proxy labels to compute proxy performance scores ($Score_p$), enabling evaluation of how well synthetic supervision approximates expert annotations. (3) Model consistency features (Feature Sets) and $Score_p$ are used to train a logistic regression meta-classifier that predicts whether a treatment recommendation is likely correct. This classifier is fit on labelled data from step 1 and unlabelled features from step 2. The classifier is applied at test time to accept or reject LLM-generated recommendations, supporting confidence estimation and threshold selection for clinical deployment.

time-intensive review of complex patient documentation [4]. These factors contribute to variability in guideline adherence and treatment recommendations, particularly in resource-constrained settings where specialist expertise is limited [5; 6].

Large language models (LLMs) offer promising potential to address these challenges by automatically processing clinical notes and recommending guideline-concordant treatment plans [7; 8; 9; 10]. However, deploying LLMs for clinical decision support requires rigorous evaluation to ensure accuracy and safety. The complexity of clinical reasoning, combined with the high stakes of treatment decisions, demands robust validation methods that can assess model performance at scale while maintaining clinical safety standards. In addition, current FDA guidelines for evaluation of AI systems recommend ROC curve measurement as a part of comprehensive clinical performance assessment endpoints [11], which is challenging to produce from generative outputs that are not associated with semantically aligned confidence scores.

Despite growing interest in clinical LLMs, rigorous evaluation remains a fundamental bottleneck due to the scarcity of expert-annotated datasets. High-quality ground truth labels for complex clinical reasoning tasks require substantial investment in specialist time and expertise, limiting the scale at which models can be validated [12; 13]. Common evaluation approaches face significant limitations: synthetic data generation often fails to capture clinical complexity and is vulnerable to distributional shift [14; 15; 16; 17], while using actual patient treatments as ground truth is problematic, since real-world decisions frequently incorporate factors exogenous to guideline recommendations, such as patient preferences, drug availability, institutional protocols, and physician experience [18; 19].

This evaluation challenge is particularly acute for guideline adherence tasks, where the gold standard requires expert oncologists to determine whether complex, multi-step clinical reasoning aligns with evidence-based recommendations. The resulting annotation bottleneck creates a critical gap: while LLMs show promise for clinical decision support, practitioners lack scalable methods to assess model reliability before deployment in high-stakes healthcare settings.

We address this evaluation bottleneck through two complementary approaches that enable scalable assessment of LLM performance on guideline adherence without extensive expert annotation. First, we evaluate models across six different proxy-benchmark generation methods: two synthetically generated datasets and four based on real clinical notes with consistency-derived labels. This analysis enables model selection and preliminary capability assessment without ground-truth labels. Second, we demonstrate that self- and cross-model consistency (the degree to which a model agrees with itself and other models) serve as reliable predictors of accuracy on expert-annotated cases.

To validate these approaches, we construct the first benchmark for NCCN guideline adherence on non-small cell lung cancer (NSCLC) by eliciting 13 oncologists to annotate 121 complete patient pathways, representing 130+ hours of specialist expertise. This expert-validated dataset enables us to systematically evaluate six different proxy benchmarking methods and quantify the relationship between model consistency and clinical accuracy across multiple frontier LLMs. We then develop a meta-classifier framework that combines these weak supervision signals to classify individual prediction correctness, achieving robust performance while requiring minimal expert validation (Figure 1).

Our evaluation highlights the role of synthetic data and consistency-based data in proxy benchmarking, with these approaches achieving high Spearman correlation coefficients (r=0.88) and low RMSE values (RMSE=0.08). We demonstrate that model consistency can also serve as a reliable accuracy predictor, enabling our meta-learning framework to achieve an average of 0.804 AUROC in classifying individual prediction accuracy across all models, while unsupervised clustering using consistency signals alone achieves 0.666 F1 at separating correct from incorrect predictions.

Our primary contributions are as follows:

1. **NCCN guideline-adherent dataset and task formalization**: The first rigorous ML problem formulation for clinical guideline adherence, with an expert-annotated dataset of 121 patient pathways and benchmarking across eight frontier LLMs.
2. **Proxy benchmark to validate performance in zero-label settings**: A systematic evaluation that identifies which proxy methods best predict clinical performance without expert labels.
3. **Consistency framework for reliable treatment prediction**: A hybrid agent and meta-classifier framework that generates relevant guideline paths and produces confidence values correlated with accuracy, enabling ROC curve calculation to clearly convey performance to clinicians and ensure compliance with regulatory standards [11].

## 2 Related Work

### 2.1 Guideline Adherent Treatment Recommendations

Guideline-adherent treatments have consistently been associated with improved clinical outcomes, such as overall survival [20; 21; 22]. Clinical decision-support systems have the ability to translate clinical guidelines into point-of-care recommendations. Several studies have demonstrated the use of AI-driven systems in predicting guideline-adherent treatment recommendations in oncology [23; 24]. Emerging work has also explored the utility of LLMs; for example, preliminary investigations suggest that ChatGPT can effectively summarize guideline content [9] and exhibits partial concordance in identifying guideline-adherent treatments [10]. Furthermore, recent benchmarking efforts have systematically evaluated the alignment of LLMs with established medical guidelines, underscoring both their potential utility and the current limitations imposed by the relatively small number of verified clinical cases available [7].

### 2.2 Synthetic Generation of Clinical Data

Synthetic data generation is a promising strategy to address the challenges of privacy and data scarcity in healthcare. Past works have focused on generating high-fidelity electronic health records (EHRs) using generative models, such as generative adversarial networks, variational autoencoders and autoregressive models [25; 26; 27; 28; 29]. More recent frameworks have used LLMs to generate synthetic data and have shown improvements in privacy guarantees and scalability [30; 31; 32; 33]. In parallel, standardized evaluation metrics and scorecards have been proposed to assess the fidelity, privacy, and clinical utility of synthetic health data, providing a foundation for more rigorous benchmarking and deployment [14; 15; 16; 34]. In addition, synthetic cohorts have been developed to model guideline-adherent treatment pathways, such as synthetic stroke registries for adherence benchmarking [35] and synthetic EHR modules that embed clinical practice guidelines and protocols [36]. Previous work has explored guideline-following capability across frontier models on synthetic clinical data[37]. These efforts demonstrate the potential of synthetic data for adherence-focused tasks. However, significant challenges remain in generating high-fidelity, guideline-adherent cases while maintaining performance on downstream evaluation tasks.

### 2.3 Weak Supervision and Consistency

Weak supervision and consistency-driven learning have emerged as key strategies for training models when labeled data is limited or noisy. Weak supervision techniques, such as programmatic labeling and distant supervision, integrate heterogeneous, partially labeled sources to produce probabilistic labels for downstream predictive tasks [38; 39; 40; 41; 42]. Consistency-based methods encourage models to produce stable predictions under input perturbations, data augmentations, or repeated evaluations, helping to regularize training and improve generalization [43; 44; 45; 46]. *Pseudo-labeling*, a related strategy, generates labels for unlabeled or partially labeled data by treating confident model predictions as additional supervised training data, enhancing model performance by expanding the training set with its own high-confidence predictions [47; 48; 45]. Within healthcare, pseudo-labeling has shown to improve the reliability of EHR phenotyping and imaging tasks under label scarcity [49; 50; 51]. Together, these techniques provide a framework for leveraging limited labeled data while maintaining prediction stability and adherence to domain-specific constraints, which is particularly relevant for guideline-adherent treatment modeling.

### 2.4 Non-Verifiable Task Evaluation

Evaluating LLMs on open-ended or non-verifiable tasks, such as treatment recommendation prediction, is challenging due to the scarcity of ground truth labels. Traditional metrics like accuracy or BLEU are insufficient in zero-label settings, prompting the use of human-aligned and/or proxy evaluation frameworks [52; 53; 54]. Further, studies have shown that consistency-derived benchmarks provide improved predictive fidelity relative to ground-truth outcomes [55; 56]. Proxy frameworks and consistency-derived benchmarks offer partial solutions, but their alignment with expert judgment remains imperfect, motivating the continued development of hybrid evaluation strategies.

## 3 Methods

### 3.1 Guideline-Compliance Task Formalization

We formalize guideline-compliant treatment prediction as a structured prediction problem. Let $x \in \mathcal{X}$ denote patient notes and $y \in \mathcal{Y}$ the corresponding guideline-compliant pathway, where $\mathcal{Y}$ is the space of decision graphs with nodes as clinical decisions and terminal nodes as treatments. An LLM $f : \mathcal{X} \rightarrow \mathcal{Y}$ produces predictions $\hat{y} = f(x)$.

To ground this formalization in clinical practice, we leveraged o3's vision capabilities [57] to extract the NCCN NSCLC guideline decision tree [1] and curated an expert-annotated dataset. Oncologists traced patient notes through the decision tree, recording node sequences and assessing guideline adherence for each case. Further details on annotation procedures, participant recruitment, and quality control measures are provided in Appendix A.1.

Since gold-standard pairs $(x, y)$ are difficult to obtain, direct supervision is not scalable as $X$ grows. To address this, we first introduce *proxy benchmarking* methods that act as substitutes for direct supervision. We define synthetic inputs $\hat{x}$ and corresponding predicted pathways $\hat{y}$, forming two primary classes of proxy evaluations:

1. **Proxy using synthetic inputs:** $(\hat{x}, \hat{y})$, where $\hat{x}$ represents synthetic patient notes generated conditionally on guideline paths. This allows assessment of model reliability under controlled perturbations or alternative representations of patient data.
2. **Proxy using real inputs:** $(x, \hat{y})$, where the model prediction $\hat{y}$ is generated multiple times by $f$ and the $(x, \hat{y})$ pair is kept only if minimum self-consistency is reached.

These proxy-based evaluations provide measurable signals of model performance, enabling ranking of models or detection of likely errors in zero-label settings. We learn a surrogate evaluator $g$ that predicts whether $\hat{y}$ is guideline-compliant. Concretely, we define a feature mapping

$$\phi : (\mathcal{X}, \mathcal{Y}, f) \rightarrow \mathbb{R}^d$$

that extracts signals such as (i) model self-consistency across rollouts, (ii) agreement across models, and (iii) alignment with proxy benchmarks. The evaluator is then trained as a binary classifier

$$g(\phi(x, \hat{y}, f)) \approx \mathbf{1}\{\hat{y} = y\}.$$

Our objective is to minimize the expected classification loss

$$\min_{g} \; \mathbb{E}_{(x,y)\sim\mathcal{D}}\big[\mathcal{L}(g(\phi(x, \hat{y}, f)), \mathbf{1}\{\hat{y} = y\})\big],$$

where $\mathcal{D}$ denotes the distribution over patients and $\mathcal{L}$ is standard loss. This formulation enables evaluation of $f$ in settings with limited or no access to human-labeled $(x, y)$ pairs, by leveraging model agreement, self-consistency, and benchmark proxy-derived features as signals for meta-classification.

### 3.1.1 Evaluation Preliminaries

To evaluate model performance, we employ two complementary metrics:

1. **Path Overlap**: Measures the proportion of nodes in predicted paths that are repeated, relative to the total nodes in all compared paths. This captures consistency of decision sequences with the full derivation in Appendix A.2
2. **Treatment Match**: A binary score indicating whether the final predicted treatment matches the ground truth when available. In settings without ground truth, it is computed as the proportion of repeated final treatments across multiple predictions, with full derivation in Appendix A.3.

## 3.2 Zero-Label Benchmark Generation

To approximate $(x, y)$ pairs for evaluation, we introduce two complementary approaches (**synthetic supervision** and **consistency-based pseudo-labeling**), with a total of six proxy benchmarking methods for zero-label performance estimation contributed.

### 3.2.1 Synthetic Supervision

We generate high-fidelity synthetic patient notes $\hat{x}$ paired with generated guideline paths $\hat{y}$ to simulate real clinical cases. The goal is to maximize fidelity to realistic patient notes while ensuring the target guideline path is accurate. Our multi-step pipeline separates the generation of $\hat{x}$ and the selection of $\hat{y}$ to filter incorrect labels from the benchmark, while still maintaining meaningfully realistic patient cases.

**Generating $\hat{x}$** Two complementary strategies are used: **Structured Generation** fills empty structured fields conditioned on the generated path and full clinical guidelines, performs consistency checks by reconstructing implied paths (discarding mismatched cases), then generates unstructured notes from structured data, target path, guidelines, and real clinical note examples. **Unstructured Generation** bypasses structured fields, generating synthetic notes directly from target paths, guidelines, and clinical note examples. Synthetic datasets were generated with GPT-4.1, except for those used to evaluate GPT-4.1, which were produced by GPT-5 under minimal-reasoning conditions.

**Selecting $\hat{y}$ via LLM Preference** Once $\hat{x}$ is generated, we obtain $\hat{y}$ by having the LLM generate the path from $\hat{x}$. If the prediction matches the target path, we accept the pair directly. Otherwise, the LLM chooses between predicted and target paths in position-agnostic format, and we retain only cases where the target path is selected.

The final dataset is then composed of $(\hat{x}, \hat{y})$ pairs which the generation model either correctly regenerated $\hat{y}$ from $\hat{x}$ or was able to select $\hat{y}$ from a pair of available $(\hat{y}^*, \hat{y})$, with $\hat{y}^*$ being the path prediction generated for $\hat{x}$. This captures examples where direct generation fails but verification remains feasible (e.g., generating the correct path is difficult, but identifying it is easier) [58; 42].

### 3.2.2 Consistency-Based Pseudo-Labeling

We propose a consistency-based pseudo-labeling strategy to construct proxy benchmarks that approximate model performance on treatment prediction and guideline path generation. Consistency-based pseudo-labels are derived from two sources: (i) *Self-consistency*, which leverages agreement within repeated predictions of a single model, and (ii) *Cross-Model Consistency*, which relies on agreement across different models. For each source, we define two benchmarks based on whether consistency is measured with respect to the *path overlap* metric or the *treatment match* metric, yielding four benchmarks in total.

**Self-Consistency** For real clinical notes, pseudo-labels are generated using model self-consistency. Specifically, we sample $k$ independent predictions $f(x) = \hat{y}_1, \ldots, \hat{y}_k$ from the same model $m$ across $X$ questions. Agreement among these predictions is assessed along two axes: *path overlap* (structural alignment) and *treatment match* (final treatment concordance). Notes with agreement above a threshold $\delta = 0.9$ on the target metric are retained, while inconsistent cases are assigned a score of 0. The retained $(x, \hat{y})$ pairs are then used to evaluate model $m$, where $\hat{y}$ corresponds to the most frequent prediction among $f(x)$. Performance is computed as accuracy over the retained subset of $X$, counting inconsistent answers as incorrect.

**Cross-Model Consistency** We further assign pseudo-labels by comparing predictions across the whole set of models, $M$. For a note $x$, if two or more models converge on the same pseudo-label after $k$ independent samples, we define this as the aggregated label. Convergence in this setting is exact path match. Inconsistent cases are simply excluded rather than penalized in contrast to the Self-Consistency method. We obtain the subset of (x, $\hat{y}$) pairs with the aggregated labels and evaluate all models in $M$ on this proxy benchmark.

### 3.3 Final Treatment Accuracy Prediction

**Supervised Accuracy Classification** We use consistency signals and model performance on proxy benchmarks to predict accuracy of a generated final treatment prediction. We train a meta-classifier using features derived from self-consistency metrics ($k$-rollout path overlap and treatment match), cross-model consistency (fraction of models with the same generated path), and proxy benchmark scores (synthetic and consistency-derived).

**Unsupervised Performance for Accuracy Classification and Error Identification** To demonstrate label-free evaluation capabilities, we apply unsupervised methods to both accuracy classification and error identification. Clustering self- and cross-model consistency features naturally separates high-confidence correct from low-confidence incorrect predictions. For error identification, we tabulate inconsistencies across $k$ rollouts for each model $m$ on patient $p$. This approach leverages consistency signals to detect potential errors without relying on human labels.

## 4 Results

### 4.1 Expert-Annotated Dataset

To evaluate LLM performance on guideline-concordant treatment prediction, we constructed a high-quality expert-annotated dataset.

**Annotation Details** A total of 13 oncologists annotated 121 patient notes, averaging 46.5 minutes per patient note with a cost of $500 per US Board certified clinician hour. To assess inter-annotator reliability, 11 examples were dually annotated, showing an average of 0.636 treatment match and 0.692 path overlap score, with further disagreement analysis present in Appendix A.8. Notes were on average 54755 characters long and contained 82 distinct paths through the NCCN decision tree and 48 distinct final treatment recommendations. We use this annotation as ground truth for both evaluating model performance and validating proxy benchmarking approaches. By establishing a high-quality reference standard, we can interpret LLM performance, detect systematic errors, and calibrate models in high-stakes medical settings without relying exclusively on expensive ongoing human annotation.

**Model Performance on Human-Annotated Benchmark** Table 1 illustrates the results of eight frontier models on the expert-annotated dataset. We report performance on GPT-5 [59], GPT-4.1 [60], o3 [57], o4-mini [61], DeepSeek-R1 [62], and LLaMA-3.3-70B-Instruct [63]. We evaluate GPT-5

| Model | Path Overlap | Treatment Match |
|---|---|---|
| GPT-5-High | 0.455 ± 0.035 | 0.339 ± 0.043 |
| GPT-5-Medium | 0.483 ± 0.036 | 0.364 ± 0.044 |
| GPT-5-Minimal | 0.441 ± 0.032 | 0.322 ± 0.043 |
| GPT-4.1 | 0.388 ± 0.035 | 0.298 ± 0.042 |
| o3 | 0.477 ± 0.038 | 0.364 ± 0.044 |
| o4-mini | 0.433 ± 0.037 | 0.339 ± 0.043 |
| Deepseek-R1 | 0.419 ± 0.038 | 0.355 ± 0.044 |
| LLaMA-3.3-70B-Instr. | 0.174 ± 0.022 | 0.112 ± 0.029 |

Table 1: Model performance on Expert-Annotated Data.

with varying reasoning efforts (minimal, medium, and high) to directly assess the impact of reasoning on performance. All models are evaluated with default temperature 1.0. We measured two clinically relevant metrics: (i) *path overlap*, measuring structural agreement with the annotated guideline path, and (ii) *treatment match*, indicating whether the recommended treatment node matches expert annotation.

### 4.2 Proxy Benchmarks Enable Model Evaluation in Zero-Label Settings

We evaluated six proxy benchmarks for assessing LLM performance without human labels (Figure 2). Among synthetic approaches, adding structure to the generation pipeline substantially increased error (RMSE rising from 0.11 in Synthetic Unstructured to 0.43 in Synthetic Structured) while yielding only a marginal correlation gain (Spearman $r$: 0.86 → 0.90). For consistency-based methods, thresholding by treatment match outperformed path overlap, likely because treatment match reflects a more generalizable prediction goal, whereas path overlap penalizes minor deviations that have little effect on downstream treatment accuracy. Aggregating outputs across models in cross-model consistency failed to improve correlations and instead increased RMSE. Overall, synthetic benchmarks are appealing for their robustness to variation in model consistency, while consistency-based benchmarks are strong in domains where accuracy and consistency are correlated. In this setting, self-consistency is aligned with model accuracy (Figure 3), and per-model Pearson correlations reported in Appendix A.5 further confirm that consistency reliably indicates instance-level accuracy.

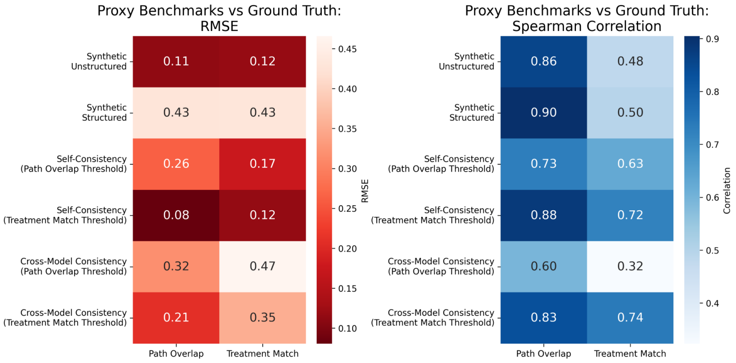


Figure 2: **Self-consistency pseudo-labels provide a robust proxy for benchmarking.** Six approaches are evaluated: synthetic data (structured and unstructured), self-consistency pseudo-labeling (with varying acceptance criteria), and cross-model consistency pseudo-labeling. Correlation is measured using Spearman coefficients as well as root mean-squared error with color intensity indicating magnitude.

### 4.3 Predicting Treatment Accuracy Using Model Self-Consistency

We develop a meta-classifier to predict when treatment recommendations are correct, using features from model consistency patterns and benchmark performance. Our approach achieves 0.804 AUROC by leveraging inference-time signals to predict accuracy without requiring ground truth labels at test time. We evaluate five feature sets to identify which signals contribute most to performance (Table 2).

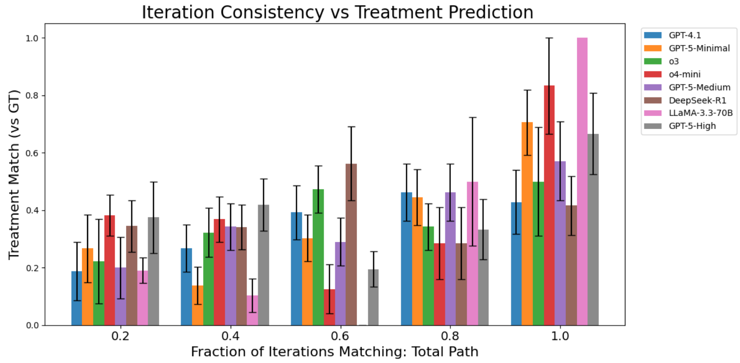


Figure 3: **Model accuracy increases with self-consistency across prediction runs**. Higher consistency (fraction of runs producing identical paths) correlates with improved performance for treatment matching.

We stratify train and test sets of the human annotated benchmark by models and patient IDs with a 70/30 split. Performance generalizes across models and patients, enabling downstream classification with minimal human supervision.

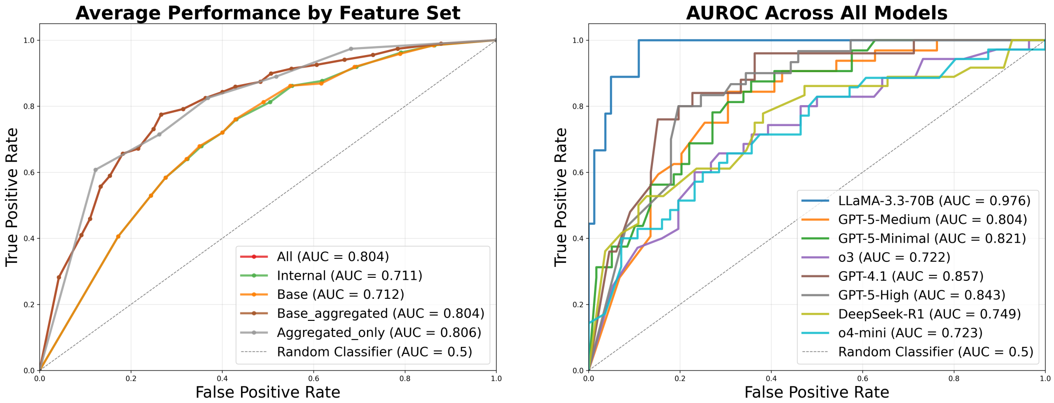
| **Feature Set** | **Self-Consistency**$^*$ | **Synthetic Benchmarks** | **Consistency Benchmarks** | **Cross-Consistency**$^*$ |
|---|---|---|---|---|
| Base | ✓ | | | |
| Internal | ✓ | ✓ | ✓ | |
| Agg. | | | | ✓ |
| Base+Agg. | ✓ | | | ✓ |
| All | ✓ | ✓ | ✓ | ✓ |

Agg. = Aggregated.
$^*$Self-consistency and Cross-Consistency features are computed per sample.

Table 2: Feature sets for predicting treatment recommendation accuracy

We find in Figure 4a that **Base_aggregated**, **Aggregated_only**, and **All** outperform other feature sets, indicating cross-model consistency provides the strongest signal for accuracy classification. We see minimal gain from proxy benchmark performance, suggesting consistency-based approaches are better for real-time analysis of model outputs than synthetic data. Even without cross-model information (**Base** and **Internal**), we see high AUROC scores, showing model self-consistency is a strong signal for classification in isolation. Figure 4b shows intra-model variability of the classifier trained using **Base_aggregated** features, with **Internal** features reported in Appendix A.7 to highlight that even where consistency and accuracy are not strongly correlated, e.g., DeepSeek-R1, these features still provide signal for accuracy classification with an AUROC of 0.582. The strong classification performance of LLaMA-3.3-70B-Instruct outputs highlights a key issue with including lower-quality models: its NCCN task performance is lower than other models, arbitrarily inflating AUROC as the meta-classifier's confidence task becomes easier. Besides LLaMA-3.3-70B-Instruct, GPT-4.1 has the highest AUROC and Deepseek-R1 the lowest, indicating GPT-4.1's consistency better correlates with accuracy.

Furthermore, in a fully unsupervised setting, we can separate accurate from inaccurate predictions with an F1 score of 0.666, showing consistency signals carry meaningful information for error detection without labels. Notably, 40.42% of model errors on human-labeled data can be detected without any human labels (detailed analysis in Appendix A.4). These results indicate consistency signals alone provide structure to distinguish accurate from inaccurate predictions and identify failure modes, enabling model developers to detect potential issues before deployment and support iterative refinement of clinical decision support systems.



(a) ROC curves by feature set, averaged over all models.

(b) ROC curves across models for the `Base_aggregated` feature set.

Figure 4: **Using signals from self- and cross-model consistency provides high AUC for accuracy classification.** Proxy benchmark results do not provide significant prediction signal compared to consistency. Model AUCs range from 0.703 to 0.981, indicating strong prediction capability. LLaMA-3.3-70B-Instruct's high AUC can be attributed to its low performance on the given task, creating an arbitrary classification problem and highlighting limitations of including lower-performing models in analyses.

## 5 Discussion

As LLMs become increasingly capable across tasks, understanding their limitations and quantifying accuracy for specific applications is critical for deployment [64]. This work makes three main contributions in the high-stakes medical domain. First, we formalize guideline-concordant treatment recommendation generation and introduce a dataset of 121 patient cases annotated by board-certified oncologists. Second, we compare synthetic supervision and consistency-based supervision for benchmarking models on tasks lacking human annotations, finding that consistency-based benchmarks best approximate human judgments (Spearman $r = 0.88$, RMSE=0.08). Third, we develop a hybrid agent and meta-classifier trained using unsupervised features, achieving an average of 0.804 AUC on held-out models and producing confidence scores that support ROC analysis, error trade-off optimization, and regulatory compliance [11]. ROC curves let clinicians or developers adjust thresholds to balance sensitivity and specificity, aligning with FDA guidance. Calibrated confidence scores from our meta-classifier provide interpretable reliability indicators, supporting informed decisions in high-stakes settings. Consistency signals reveal model failure modes—like guideline non-compliance or TNM staging errors—even without extensive human annotation, highlighting opportunities for targeted improvement. Feature analysis shows self- and cross-model consistency are the strongest predictors of recommendation accuracy, enabling error detection and confidence calibration in a label-free setting.

Several promising directions emerge from this work. Expanding the dataset to other cancer types and guidelines would help characterize the generalizability of our consistency-based benchmarking approach, potentially enabling cross-domain accuracy prediction [65; 66]. Further evaluation of self- and cross-model consistency as a signal for accuracy classification across broader model sizes and architectures could deepen understanding of this method's robustness. Increasing both the size of the human-annotated dataset and the number of dually annotated samples would strengthen confidence in ground-truth labels. Since clinicians exhibit variability in path selection, future work should also explore ways to explicitly model and incorporate human uncertainty into evaluation. Adaptive learning approaches that account for uncertainty across clinical scenarios may further improve sensitivity of proxy metrics [67; 68]. Finally, investigating how proxy benchmark data can be leveraged for alignment with downstream human preferences could help mitigate data bottlenecks that limit current methods [69; 70].

## References

[1] National Comprehensive Cancer Network. NCCN clinical practice guidelines in oncology (NCCN guidelines®). 2025. Version 1.2025.

[2] Lenora A. Pluchino and Thomas A. D'Amico. National comprehensive cancer network guidelines: Who makes them? what are they? why are they important? *The Annals of Thoracic Surgery*, 110(6):1789–1795, December 2020. doi: 10.1016/j.athoracsur.2020.03.022.

[3] Rodger J. Winn and Joan McClure. The NCCN clinical practice guidelines in oncology: A primer for users. *Journal of the National Comprehensive Cancer Network*, 1(1):5–13, March 2003. doi: 10.6004/jnccn.2003.0003.

[4] Al B Benson III and Elizabeth Brown. Role of NCCN in integrating cancer clinical practice guidelines into the healthcare debate. *American Health & Drug Benefits*, 1(1):28, 2008.

[5] Wui-Jin Koh, Benjamin O. Anderson, and Robert W. Carlson. NCCN resource-stratified and harmonized guidelines: A paradigm for optimizing global cancer care. *Cancer*, 126(S10): 2416–2423, April 2020. doi: 10.1002/cncr.32880.

[6] Priyanka Kumar, Michael Del Rosario, Jenny Chang, Argyrios Ziogas, Mehraneh D. Jafari, Robert E. Bristow, Sora Park Tanjasiri, and Jason A. Zell. Population-based analysis of national comprehensive cancer network (NCCN) guideline adherence for patients with anal squamous cell carcinoma in California. *Cancers*, 15(5):1465, February 2023. doi: 10.3390/cancers15051465.

[7] Dennis Fast, Lisa C. Adams, Felix Busch, Conor Fallon, Marc Huppertz, Robert Siepmann, Philipp Prucker, Nadine Bayerl, Daniel Truhn, Marcus Makowski, Alexander Löser, and Keno K. Bressem. Autonomous medical evaluation for guideline adherence of large language models. *npj Digital Medicine*, 7(1), December 2024. doi: 10.1038/s41746-024-01356-6.

[8] Leyao Wang, Zhiyu Wan, Congning Ni, Qingyuan Song, Yang Li, Ellen Clayton, Bradley Malin, and Zhijun Yin. Applications and concerns of ChatGPT and other conversational large language models in health care: Systematic review. *Journal of Medical Internet Research*, 26:e22769, November 2024. doi: 10.2196/22769.

[9] Ehab Hamed, Ahmad Eid, and Medhat Alberry. Exploring ChatGPT's potential in facilitating adaptation of clinical guidelines: A case study of diabetic ketoacidosis guidelines. *Cureus*, May 2023. doi: 10.7759/cureus.38784.

[10] Brian Schulte. Capacity of ChatGPT to identify guideline-based treatments for advanced solid tumors. *Cureus*, April 2023. doi: 10.7759/cureus.37938.

[11] U.S. Food and Drug Administration. Clinical performance assessment: Considerations for computer-assisted detection devices applied to radiology images and radiology device data in premarket notification (510(k)) submissions. Guidance for industry and fda staff, U.S. Food and Drug Administration, September 2022. Docket Number: FDA-2009-D-0503.

[12] Hubert Dariusz Zajac, Natalia Rozalia Avlona, Finn Kensing, Tariq Osman Andersen, and Irina Shklovski. Ground truth or dare: Factors affecting the creation of medical datasets for training AI. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 351–362, August 2023. doi: 10.1145/3600211.3604766.

[13] Jennifer J Liang, Ching-Huei Tsou, and Murthy V Devarakonda. Ground truth creation for complex clinical nlp tasks–an iterative vetting approach and lessons learned. *AMIA Summits on Translational Science Proceedings*, 2017:203, 2017.

[14] Gabriele Santangelo, Giovanna Nicora, Riccardo Bellazzi, and Arianna Dagliati. How good is your synthetic data? SynthRO, a dashboard to evaluate and benchmark synthetic tabular data. *BMC Medical Informatics and Decision Making*, 25(1), February 2025. doi: 10.1186/s12911-024-02731-9.

[15] Chao Yan, Yao Yan, Zhiyu Wan, Ziqi Zhang, Larsson Omberg, Justin Guinney, Sean D. Mooney, and Bradley A. Malin. A multifaceted benchmarking of synthetic electronic health record generation models. *Nature Communications*, 13(1), December 2022. doi: 10.1038/s41467-022-35295-1.

[16] Xingran Chen, Zhenke Wu, Xu Shi, Hyunghoon Cho, and Bhramar Mukherjee. Generating synthetic electronic health record data: a methodological scoping review with benchmarking on phenotype data and open-source software. *Journal of the American Medical Informatics Association*, 32(7):1227–1240, June 2025. doi: 10.1093/jamia/ocaf082.

[17] Hejie Cui, Alyssa Unell, Bowen Chen, Jason Alan Fries, Emily Alsentzer, Sanmi Koyejo, and Nigam Shah. TIMER: Temporal instruction modeling and evaluation for longitudinal clinical records, 2025. arXiv preprint arXiv:2503.04176.

[18] Derk L. Arts, Albertine G. Voncken, Stephanie Medlock, Ameen Abu-Hanna, and Henk C. P. M. van Weert. Reasons for intentional guideline non-adherence: A systematic review. *International Journal of Medical Informatics*, 89:55–62, May 2016. doi: 10.1016/j.ijmedinf.2016.02.009.

[19] Silvana Quaglini, Paolo Ciccarese, Giuseppe Micieli, and Anna Cavallini. Non-compliance with guidelines: Motivations and consequences in a case study. In *Computer-based Support for Clinical Guidelines and Protocols*. IOS Press, 2004. doi: 10.3233/978-1-60750-944-8-75.

[20] Jhalak Dholakia, Elyse Llamocca, Allison Quick, Ritu Salani, and Ashley S. Felix. Guideline-concordant treatment is associated with improved survival among women with non-endometrioid endometrial cancer. *Gynecologic Oncology*, 157(3):716–722, June 2020. doi: 10.1016/j.ygyno.2020.03.016.

[21] Jonatan Lindqvist, Antti Jekunen, Eero Sihvo, Mikael Johansson, and Heidi Andersén. Effect of adherence to treatment guidelines on overall survival in elderly non-small-cell lung cancer patients. *Lung Cancer*, 171:9–17, September 2022. doi: 10.1016/j.lungcan.2022.07.006.

[22] Achim Wöckel, Christian Kurzeder, Verena Geyer, Igor Novasphenny, Regine Wolters, Manfred Wischnewsky, Rolf Kreienberg, and Dominic Varga. Effects of guideline adherence in primary breast cancer–a 5-year multi-center cohort study of 3976 patients. *The Breast*, 19(2):120–127, April 2010. doi: 10.1016/j.breast.2009.12.006.

[23] Abdulwadud Nafees, Maha Khan, Ronald Chow, Rouhi Fazelzad, Andrew Hope, Geoffrey Liu, Daniel Letourneau, and Srinivas Raman. Evaluation of clinical decision support systems in oncology: An updated systematic review. *Critical Reviews in Oncology/Hematology*, 192:104143, December 2023. doi: 10.1016/j.critrevonc.2023.104143.

[24] Suthida Suwanvecho, Harit Suwanrusme, Tanawat Jirakulaporn, Surasit Issarachai, Nimit Taechakraichana, Palita Lungchukiet, Wimolrat Decha, Wisanu Boonpakdee, Nittaya Thanakarn, Pattanawadee Wongrattananon, Anita M Preininger, Metasebya Solomon, Suwei Wang, Rezzan Hekmat, Irene Dankwa-Mullan, Edward Shortliffe, Vimla L Patel, Yull Arriaga, Gretchen Purcell Jackson, and Narongsak Kiatikajornthada. Comparison of an oncology clinical decision-support system's recommendations with actual treatment decisions. *Journal of the American Medical Informatics Association*, 28(4):832–838, January 2021. doi: 10.1093/jamia/ocaa334.

[25] Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F. Stewart, and Jimeng Sun. Generating multi-label discrete patient records using generative adversarial networks. In *Proceedings of the 2nd Machine Learning for Healthcare Conference (MLHC)*, volume 68, pages 286–305. PMLR, 2017.

[26] Chang Lu, Chandan K. Reddy, Ping Wang, Dong Nie, and Yue Ning. Multi-label clinical time-series generation via conditional GAN. *IEEE Trans. on Knowl. and Data Eng.*, 36(4):1728–1740, April 2024. doi: 10.1109/TKDE.2023.3310909.

[27] Jin Li, Benjamin J. Cairns, Jingsong Li, and Tingting Zhu. Generating synthetic mixed-type longitudinal electronic health records for artificial intelligent applications. *npj Digital Medicine*, 6(1), May 2023. doi: 10.1038/s41746-023-00834-7.

[28] Jinsung Yoon, Michel Mizrahi, Nahid Farhady Ghalaty, Thomas Jarvinen, Ashwin S. Ravi, Peter Brune, Fanyu Kong, Dave Anderson, George Lee, Arie Meir, Farhana Bandukwala, Elli Kanal, Sercan Ö. Arık, and Tomas Pfister. EHR-Safe: generating high-fidelity and privacy-preserving synthetic electronic health records. *npj Digital Medicine*, 6(1), August 2023. doi: 10.1038/s41746-023-00888-7.

[29] Daniel Jarrett, Ioana Bica, and Mihaela van der Schaar. Time-series generation by contrastive imitation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 28968–28982. Curran Associates, Inc., 2021.

[30] Bayan Altalla', Sameera Abdalla, Ahmad Altamimi, Layla Bitar, Amal Al Omari, Ramiz Kardan, and Iyad Sultan. Evaluating GPT models for clinical note de-identification. *Scientific Reports*, 15(1), January 2025. doi: 10.1038/s41598-025-86890-3.

[31] Austin A. Barr, Joshua Quan, Eddie Guo, and Emre Sezgin. Large language models generating synthetic clinical datasets: a feasibility and comparative analysis with real-world perioperative data. *Frontiers in Artificial Intelligence*, 8, February 2025. doi: 10.3389/frai.2025.1533508.

[32] Andrea Taloni, Giulia Coco, Marco Pellegrini, Matthias Wjst, Niccolò Salgari, Giovanna Carnovale-Scalzo, Vincenzo Scorcia, Massimo Busin, and Giuseppe Giannaccare. Exploring detection methods for synthetic medical datasets created with a large language model. *JAMA Ophthalmology*, 143(6):517, June 2025. doi: 10.1001/jamaophthalmol.2025.0834.

[33] Onkar Litake, Brian H Park, Jeffrey L Tully, and Rodney A Gabriel. Constructing synthetic datasets with generative artificial intelligence to train large language models to classify acute renal failure from clinical notes. *Journal of the American Medical Informatics Association*, 31 (6):1404–1410, April 2024. doi: 10.1093/jamia/ocae081.

[34] Suhana Bedi, Hejie Cui, Miguel Fuentes, Alyssa Unell, Michael Wornow, Juan M. Banda, Nikesh Kotecha, Timothy Keyes, Yifan Mai, Mert Oez, Hao Qiu, Shrey Jain, Leonardo Schettini, Mehr Kashyap, Jason Alan Fries, Akshay Swaminathan, Philip Chung, Fateme Nateghi, Asad Aali, Ashwin Nayak, Shivam Vedak, Sneha S. Jain, Birju Patel, Oluseyi Fayanju, Shreya Shah, Ethan Goh, Dong-han Yao, Brian Soetikno, Eduardo Reis, Sergios Gatidis, Vasu Divi, Robson Capasso, Rachna Saralkar, Chia-Chun Chiang, Jenelle Jindal, Tho Pham, Faraz Ghoddusi, Steven Lin, Albert S. Chiou, Christy Hong, Mohana Roy, Michael F. Gensheimer, Hinesh Patel, Kevin Schulman, Dev Dash, Danton Char, Lance Downing, Francois Grolleau, Kameron Black, Bethel Mieso, Aydin Zahedivash, Wen-wai Yim, Harshita Sharma, Tony Lee, Hannah Kirsch, Jennifer Lee, Nerissa Ambers, Carlene Lugtu, Aditya Sharma, Bilal Mawji, Alex Alekseyev, Vicky Zhou, Vikas Kakkar, Jarrod Helzer, Anurang Revri, Yair Bannett, Roxana Daneshjou, Jonathan Chen, Emily Alsentzer, Keith Morse, Nirmal Ravi, Nima Aghaeepour, Vanessa Kennedy, Akshay Chaudhari, Thomas Wang, Sanmi Koyejo, Matthew P. Lungren, Eric Horvitz, Percy Liang, Mike Pfeffer, and Nigam H. Shah. MedHELM: Holistic evaluation of large language models for medical tasks, 2025. arXiv preprint arXiv:2505.23802.

[35] Lanjing Wang, Vihaan Manchanda, Holly Picotte, Chandler Beon, Jennifer L. Hall, Juan Zhao, and Xue Feng. Synthetic data for the get with the guidelines–stroke registry. *Journal of the American Heart Association*, 14(5), March 2025. doi: 10.1161/jaha.124.039667.

[36] Jason Walonoski, Mark Kramer, Joseph Nichols, Andre Quina, Chris Moesel, Dylan Hall, Carlton Duffett, Kudakwashe Dube, Thomas Gallagher, and Scott McLachlan. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*, 25(3): 230–238, August 2017. doi: 10.1093/jamia/ocx079.

[37] X. Li, M. Gao, Y. Hao, T. Li, G. Wan, Z. Wang, Y. Wang, and X. Chen. Medguide: Benchmarking clinical decision-making in large language models. `https://arxiv.org/abs/2505.11613`, 2025. arXiv preprint arXiv:2505.11613.

[38] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, 2009.

[39] Alexander Ratner, Christopher De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. Data programming: creating large training sets, quickly. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 3574–3582, 2016.

[40] Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment*, 11(3):269–282, November 2017. doi: 10.14778/3157794.3157797.

[41] Enshuo Hsu and Kirk Roberts. Leveraging large language models for knowledge-free weak supervision in clinical natural language processing. *Scientific Reports*, 15(1), March 2025. doi: 10.1038/s41598-024-68168-2.

[42] Jon Saad-Falcon, E. Kelly Buchanan, Mayee F. Chen, Tzu-Heng Huang, Brendan McLaughlin, Tanvir Bhathal, Shang Zhu, Ben Athiwaratkun, Frederic Sala, Scott Linderman, Azalia Mirhoseini, and Christopher Ré. Shrinking the generation-verification gap with weak verifiers, 2025. arXiv preprint arXiv:2506.18203.

[43] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning, 2016. arXiv preprint arXiv:1610.02242.

[44] Takeru Miyato, Shin-Ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1979–1993, August 2019. doi: 10.1109/tpami.2018.2858821.

[45] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with noisy student improves ImageNet classification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2020. doi: 10.1109/cvpr42600.2020.01070.

[46] B. Brown, J. Juravsky, R. Ehrlich, R. Clark, Q. V. Le, C. Ré, and A. Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. `https://arxiv.org/abs/2407.21787`, 2024. arXiv preprint arXiv:2407.21787.

[47] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)*, July 2013.

[48] Yuliang Zou, Zizhao Zhang, Han Zhang, Chun-Liang Li, Xiao Bian, Jia-Bin Huang, and Tomas Pfister. PseudoSeg: Designing pseudo labels for semantic segmentation. In *International Conference on Learning Representations (ICLR)*, 2021.

[49] Hang Dong, Víctor Suárez-Paniagua, Huayu Zhang, Minhong Wang, Arlene Casey, Emma Davidson, Jiaoyan Chen, Beatrice Alex, William Whiteley, and Honghan Wu. Ontology-driven and weakly supervised rare disease identification from clinical notes. *BMC Medical Informatics and Decision Making*, 23(1), May 2023. doi: 10.1186/s12911-023-02181-9.

[50] Wenqiao Zhang, Lei Zhu, James Hallinan, Shengyu Zhang, Andrew Makmur, Qingpeng Cai, and Beng Chin Ooi. BoostMIS: Boosting medical image semi-supervised learning with adaptive pseudo labeling and informative active annotation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20666–20676, June 2022.

[51] Isabelle-Emmanuella Nogues, Jun Wen, Yucong Lin, Molei Liu, Sara K. Tedeschi, Alon Geva, Tianxi Cai, and Chuan Hong. Weakly semi-supervised phenotyping using electronic health records. *Journal of Biomedical Informatics*, 134:104175, October 2022. doi: 10.1016/j.jbi.2022.104175.

[52] Xinxi Lyu, Yizhong Wang, Hannaneh Hajishirzi, and Pradeep Dasigi. HREF: Human response-guided evaluation of instruction following in language models, 2024. arXiv preprint arXiv:2412.15524.

[53] Yusuke Yamauchi, Taro Yano, and Masafumi Oyamada. An empirical study of LLM-as-a-judge: How design choices impact evaluation reliability, 2025. arXiv preprint arXiv:2506.13639.

[54] Haochen Tan, Zhijiang Guo, Zhan Shi, Lu Xu, Zhili Liu, Yunlong Feng, Xiaoguang Li, Yasheng Wang, Lifeng Shang, Qun Liu, and Linqi Song. ProxyQA: An alternative framework for evaluating long-form text generation with large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 6806–6827, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.368.

[55] Noah Lee, Jiwoo Hong, and James Thorne. Evaluating the consistency of LLM evaluators, 2024. arXiv preprint arXiv:2412.00543.

[56] Yubo Li, Yidi Miao, Xueying Ding, Ramayya Krishnan, and Rema Padman. Firm or fickle? evaluating large language models consistency in sequential interactions. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6679–6700, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.347.

[57] OpenAI. OpenAI o3 system card. `https://openai.com/index/o3-o4-mini-system-card/`, 2025. Accessed: 2025-09-05.

[58] Zihan Ma, Taolin Zhang, Maosong Cao, Junnan Liu, Wenwei Zhang, Minnan Luo, Songyang Zhang, and Kai Chen. Rethinking verification for LLM code generation: From generation to testing, 2025. arXiv preprint arXiv:2507.06920.

[59] OpenAI. GPT-5 system card. `https://cdn.openai.com/gpt-5-system-card.pdf`, 2025. Accessed: 2025-09-05.

[60] OpenAI. GPT-4.1 system card. `https://openai.com/index/gpt-4-1/`, 2025. Accessed: 2025-09-05.

[61] OpenAI. OpenAI o4-mini system card. `https://openai.com/index/o3-o4-mini-system-card/`, 2025. Accessed: 2025-09-05.

[62] DeepSeek-AI. DeepSeek-R1 system card. `https://huggingface.co/deepseek-ai/DeepSeek-R1`, 2025. Accessed: 2025-09-05.

[63] Meta. LLaMA 3.3-70B system card. `https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct`, 2024. Accessed: 2025-09-05.

[64] Alyssa Unell, Mehr Kashyap, Michael Pfeffer, and Nigam Shah. Real-world usage patterns of large language models in healthcare. *medRxiv*, 2025. doi: 10.1101/2025.05.02.25326781.

[65] V Sorin, BS Glicksberg, Y Artsi, Y Barash, E Konen, GN Nadkarni, and E Klang. Utilizing large language models in breast cancer management: systematic review. *Journal of Cancer Research and Clinical Oncology*, 150(3):140, March 2024. doi: 10.1007/s00432-024-05678-6.

[66] Tristen Pool and Dennis Trujillo. A large language model pipeline for breast cancer oncology, 2024. arXiv preprint arXiv:2406.06455.

[67] Benjamin Kompa, Jasper Snoek, and Andrew L. Beam. Second opinion needed: communicating uncertainty in medical machine learning. *npj Digital Medicine*, 4(1):4, 2021. doi: 10.1038/s41746-020-00367-3.

[68] X Jing, R Goli, K Komatineni, S Alluri, N Hubig, H Min, Y Gong, DF Sittig, P Biondich, D Robinson, C Nøhr, A Faxvaag, A Wright, T Law, L Rennert, and R Gimbel. Active learning pipeline to identify candidate terms for a CDSS ontology. *Studies in Health Technology and Informatics*, 316:1338–1342, August 2024. doi: 10.3233/SHTI240660.

[69] Sungdong Kim and Minjoon Seo. Rethinking the role of proxy rewards in language model alignment, 2024. arXiv preprint arXiv:2402.03469.

[70] Seohyeong Lee, Eunwon Kim, Hwaran Lee, and Buru Chang. Dataset cartography for large language model alignment: Mapping and diagnosing preference data. 2025.

[71] Insightful. Insightful: Workforce intelligence & productivity insights. `https://www.insightful.io/`, 2025. Accessed: 2025-08-28.

# A Appendix

## A.1 Annotation Details

We begin by extracting the decision tree pages from the NCCN NSCLC guidelines [1] and creating a JSON-based representation where each decision point is labeled as a “node." Terminal nodes with “recommendation" labels contain final treatment recommendations for clinicians, while intermediate nodes represent decision points requiring additional clinical information.

To evaluate model capability in mapping these guidelines onto real clinical data, we created a human-annotated dataset for guideline-adherent reasoning in NSCLC. Expert annotators traced patient notes through the NCCN guideline decision tree, recording the ordered sequence of page–node identifiers (e.g., NSCL-1-1 $\rightarrow$ NSCL-2-1 $\rightarrow \ldots \rightarrow$ terminal). Each annotation aimed to reach either the appropriate “recommendation” node when sufficient clinical information was available, or the furthest node that could be accurately determined given the available patient data. Annotators additionally marked whether the observed care was guideline-compliant and provided brief rationales for any identified as non-compliance. Non-compliance is treated as a valid “path” and evaluates model’s capacity to identify patient treatment trajectories as non-compliant with given guidelines.

Thirteen physicians performed the annotations: twelve board-certified oncologists and one hematology–oncology fellow, averaging 13 years of clinical experience. Participants were recruited through data vendors, screened for oncology knowledge and rubric literacy, and verified via official certification registries. All physicians were compensated, and annotation instructions are provided.

A subset of cases was double-annotated, and disagreements were resolved through adjudication meetings, with all decisions documented for consistency. A strict no-AI-use policy was enforced via Insightful time-tracking with periodic screenshot audits [71]. Annotation time averaged 46.5 minutes per case. All cases were fully de-identified and contained no protected health information, so the study did not constitute human-subjects research and no IRB was required.

**Clinical Annotation Instructions**

You are helping us understand how a synthetic patient progressed through the NCCN guidelines for non-small cell lung cancer (NSCLC), based on information in their clinical note. Your task is to carefully read the clinical note and determine which steps in the NCCN guideline the patient appears to have gone through. Using your clinical judgment and the structure of the guideline, trace the path the patient has followed so far.

**Path Format:** You will identify the treatment path in the following format:

NSCL-x-y → NSCL-a-b → … → Final Treatment Node

Each step should include the full node ID from the guideline (e.g., NSCL-1-1). Do not skip any nodes, even if some are broader or less specific than others. The final step should be a terminal node—a node that includes a clear treatment recommendation.

If the patient's treatment has not reached a terminal node, you should provide the recommended pathway to the next terminal node, as permitted by the information in the note.

**Compliance Assessment:** There is a box to check if the patient's path complies with the NCCN guidelines. If you believe the patient's path does not align with the NCCN guideline, please do not check the box. In this case, you will also include a brief explanation of why the treatment path doesn't match the guidelines. Only include a "reason" if the path is not compliant. If the path is guideline-compliant, just provide the full treatment path.

**Important Notes:**

- You do not need to speculate beyond what is described in the note
- You do not need to assume the patient has already received the final treatment—just determine the most appropriate next step or current position in the guideline
- In some cases, the NSCLC Guidelines may not apply. For these cases, please note that the guidelines do not apply and proceed to the next task

**Materials Provided:**

- A clinical note (the patient's case)
- The full NCCN guideline (for reference)

**Tips and Tricks:** We recommend utilizing "Command + f" (Mac) or "Ctrl+f" (Windows) to find key terms and parts of the note/guideline tree that are relevant. The guidelines consist of many "pages" (i.e. NSCLC-1, NSCLC-2) which contain further questions that will bring you to other pages.

**Example Workflow:**

*Step 1:* **REDACTED DUE TO NCCN LICENSING REGULATIONS**

*Step 2:* Using Ctrl+F (Command+F on Mac), proceed to the next step based on the patient's clinical notes.

Per the example above, search "NSCL-2" to skip to the section beginning: `‘‘NSCL-2’’: { ‘‘page_id’’: ‘‘NSCL-2’’`

Under NSCL-2, you see "nodes" and then a numbered list of different criteria. Begin with node 1 and follow the branching guidance from there.

At each step, confirm that the treatment outlined in the patient's clinical note is in accordance with the NCCN Guidelines. If the treatment diverges from the path outlined by the guidelines, it would be considered noncompliant. The treatment path you provide should be the treatment path that was followed in the patient's care. If this differs from the path recommended by the NCCN Guidelines, indicate that the path was noncompliant.

## A.2 Path Overlap Score

Let $\mathcal{P} = \{P_1, P_2, \ldots, P_k\}$ be a collection of $k$ predicted paths, where

$$P_i = (p_{i,1}, p_{i,2}, \ldots, p_{i,n_i}), \quad i = 1, \ldots, k,$$

and $p_{i,j}$ denotes the $j$-th node in path $P_i$.

Define the node set of path $P_i$ as

$$V_i = \{p_{i,1}, p_{i,2}, \ldots, p_{i,n_i}\},$$

and the union and intersection of all paths as

$$U = \bigcup_{i=1}^{k} V_i, \quad I = \bigcap_{i=1}^{k} V_i.$$

The **Path Overlap Score** (corresponding to `path_match_fraction` in code) is defined as the Jaccard similarity:

$$\text{Overlap}(\mathcal{P}) = \begin{cases} \frac{|I|}{|U|}, & |U| > 0, \\ 1, & |U| = 0. \end{cases}$$

### Properties

- $\text{Overlap}(\mathcal{P}) \in [0, 1]$.
- $\text{Overlap}(\mathcal{P}) = 1$ if and only if all paths share exactly the same set of nodes.
- Higher scores indicate greater similarity in node coverage across paths.
- Symmetric with respect to path ordering.
- Sensitive only to node membership, not ordering or repetition.

## A.3 Treatment Match

### Final Treatment Consistency Score

Let $\mathcal{P} = \{P_1, P_2, \ldots, P_k\}$ be a set of $k$ predicted treatment paths for a given patient, where each path $P_i = (p_{i,1}, p_{i,2}, \ldots, p_{i,n_i})$ terminates with a final treatment recommendation $f_i = p_{i,n_i}$.

### Case 1: Ground Truth Available

When ground truth final treatment $f^*$ is available, the score is computed as the indicator function:

$$\mathbf{1}_{f_i = f^*} = \begin{cases} 1 & \text{if } f_i = f^* \\ 0 & \text{otherwise} \end{cases}$$

This returns whether there was an exact match between the ground truth and the prediction final treatment.

### Case 2: No Ground Truth Available

In the absence of ground truth, we measure internal consistency by computing the proportion of repeated final treatments:

$$S_{\text{final}}(\mathcal{P}) \;=\; \frac{\max_{t \in T} c(t)}{k}$$

where:

$$k = \text{total rollouts}$$

$$c(t) = \sum_{i=1}^{k} \mathbf{1}_{\{f_i = t\}}$$

The numerator $\max_{t \in T} c(t)$ counts the total number of “repeated" final treatment occurrences (i.e., how many times each treatment appears beyond its first occurrence) and returns the frequency of the mode treatment selection. For all models $k = 10$.

### Properties

- $S_{\text{final}}(\mathcal{P}) \in [0, 1]$ for both cases
- Higher scores indicate better accuracy (Case 1) or higher consensus (Case 2)
- **Case 2**: Score of 0 indicates complete disagreement; score approaching 1 indicates strong consensus

### A.4 Error Analysis

We find that 40.42% of all errors made by models on human-annotated clinical text are flagged within the top 5 points of confusion by the model itself during its internal rollouts. This enables practitioners to refine instructions, perform fine-tuning, or apply other interventions to address these frequent errors—without relying on human annotations to detect them. Figure 5 shows the overlap between errors identified by model consistency vs. the true errors made by the model against human annotated data. Discrepancies mostly arise around guideline compliance and tumor staging, indicating that models struggle to differentiate between non-compliant and compliant cases with high accuracy as well as are unable to leverage parametric knowledge to perform tumor staging. Potential interventions could be training a model specifically for either task or providing additional context with clear instructions regarding clinical expectations.

### A.5 Consistency-Accuracy Correlation

We show that for both Path Overlap metric and Treatment Match metrics, consistency of a prediction is correlated with accuracy (Table 3). We do see, however, that there are different levels of correlation, indicating that consistency-based approaches inherently are more accurate for some model families than others. We note that DeepSeek-R1 has an outlier correlation coefficient value, being negative for Path Overlap score and extremely low for Treatment Match score's correlation with accuracy. This highlights the robustness of the meta-classification pipeline to varying model internal consistencies. Despite the lack of direct correlation between self-consistency and accuracy for this model, we are still able to classify accuracy with high fidelity.

| **Model** | **Path Overlap** | **Treatment Match** |
|---|---|---|
| GPT-5-High | 0.477 | 0.472 |
| GPT-4.1 | 0.812 | 0.923 |
| GPT-5-Minimal | 0.925 | 0.866 |
| o3 | 0.700 | 0.795 |
| o4-mini | 0.566 | 0.491 |
| GPT-5-Medium | 0.794 | 0.935 |
| DeepSeek-R1 | -0.647 | 0.129 |
| LLaMA-3.3-70B-Instr. | 0.688 | 0.789 |
| **Mean ± SEM** | 0.540 ± 0.179 | 0.675 ± 0.103 |

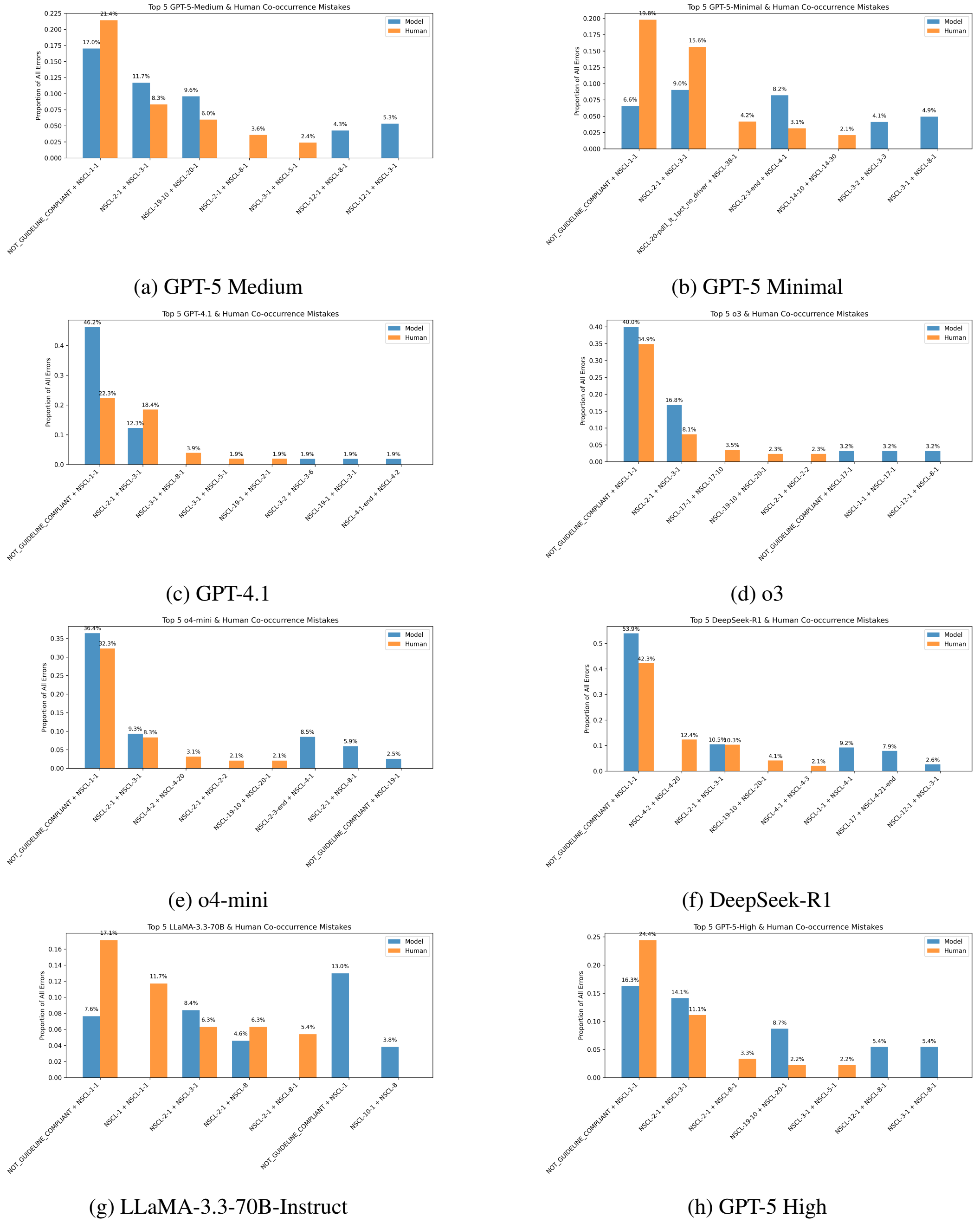
Table 3: Iteration consistency vs. accuracy correlation ($r$) for path overlap score and treatment match score. Mean correlation values across all models ± SEM are reported at the bottom.

### A.6 Unsupervised clustering performance

We cluster on features related to self- and cross-model consistency (Figure 6). We are able to distinguish between True Negatives and False Negatives with high accuracy, and are able to derive some signal regarding the True Positives. This indicates that consistency alone can be used as a baseline in an unsupervised method to identify accuracy of prediction. The performance is below supervised approaches, as to be expected, but is promising for further applications in which supervision is not viable or possible. F1 score is .666, indicating potential of unsupervised methods in conjunction with consistency approaches to evaluate model performance in zero-label settings.

### A.7 Meta-Classifier Performance on Internal Features

We report the performance of our meta-classifier using internal features only for training, as shown in Figure 7. This enables us to disaggregate classification performance between cross-model and self consistency. We see here that the AUROC scores for each model is lower than when we include cross-model information, but it is still significantly above random, indicating that while cross-model consistency is a useful feature, it is not the only feature that allows for insight into model performance.



(a) GPT-5 Medium

(b) GPT-5 Minimal

(c) GPT-4.1

(d) o3

(e) o4-mini

(f) DeepSeek-R1

(g) LLaMA-3.3-70B-Instruct

(h) GPT-5 High

Figure 5: Identification of most common discrepancies between human annotations and model annotations compared to most common discrepancies between $k$ model rollouts of path prediction.

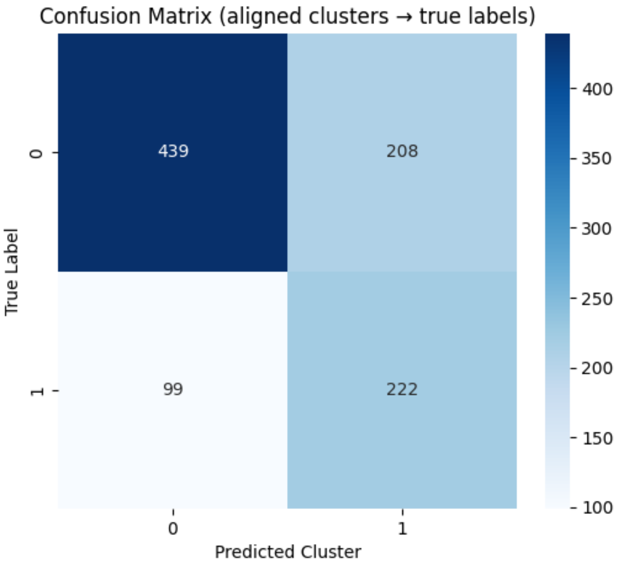


Figure 6: Confusion Matrix for K-means clustering over consistency-derived features.

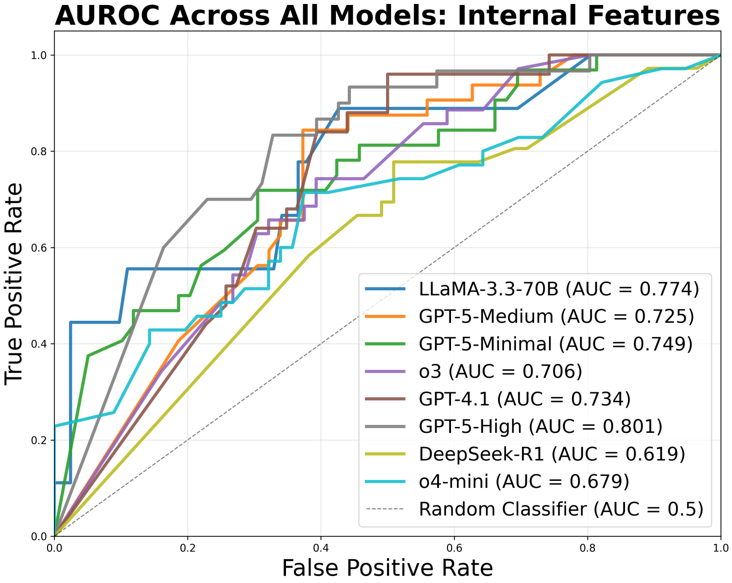


Figure 7: ROCs disaggregated by model for classification of treatment prediction accuracy using exclusively internal features during training.

### A.8 Model Information

For supervised classification, we trained a logistic regression classifier with L2 regularization (C=0.1), optimized using LBFGS with up to 10,000 iterations, and applied class-balanced weighting to account for label imbalance. Table 4 reports further API details.

| Model | Provider | API Version |
|---|---|---|
| DeepSeek-R1 | Azure AI | 2024-05-01-preview |
| LLaMA-3.3-70B-Instr. | Azure OpenAI | 5 (Model release: 2024-12-06) |
| GPT-4.1 | Azure OpenAI | 2024-12-01-preview |
| GPT-5 | Azure OpenAI | 2024-12-01-preview |
| o4-mini | Azure OpenAI | 2024-12-01-preview |
| o3 | Azure OpenAI | 2024-12-01-preview |

Table 4: Model release dates and API versions.

### A.9 Human Annotation Disagreement Analysis

Of the 11 dually annotated records:

1. 7 were agreed upon by all physicians.
2. 4 required adjudication
3. 2 involved disagreements about whether the patient's history complied with guidelines (these cases are excluded from our benchmark).
4. 1 involved assumptions regarding the patient's history; after clarification to rely strictly on the provided record, both clinicians agreed.
5. 1 involved a TNM staging error by one clinician, resolved upon discussion.

After removing non-compliant cases, agreement across the benchmark was 7/9. This reflects the inherent difficulty and nuance of the task, setting a realistic upper bound for LLM performance in clinical decision support scenarios.

### A.10 NCCN Tree Generation

To construct the clinical decision tree, we first manually structured a subset of sample guideline pages into a JSON-based hierarchical format representing decision nodes and treatment branches. These manually curated examples served as few-shot demonstrations for prompting the language model to generate additional decision trees for the remaining guideline sections. The generated JSON outputs were automatically validated to ensure structural integrity and adherence to the required schema. While this process provided an initial level of validation, the resulting trees were not independently verified by clinical experts, and thus may contain minor inconsistencies or omissions relative to expert-curated guideline representations. The generated NCCN tree for NSCLC has 3,302 nodes, 3,301 edges, and a max depth of 8.

### A.11 Clinical Data Details

The patient data used for analysis comprised approximately 2,900 patients diagnosed with breast and lung cancer at various stages, drawn from two U.S. health systems: a community-based system and a large nonprofit health system, each spanning multiple hospital locations. The dataset included electronic medical records, radiology reports, pathology reports, and other clinical documents such as next-generation sequencing (NGS) reports. Tumor registry data were obtained from three complementary sources: (1) the hospitals' internal registries, (2) registry submissions to the state (for one system), and (3) manually labeled registry data extracted by over 30 registered oncology nurses. All data were de-identified by the vendor prior to delivery, following HIPAA-compliant procedures that included redaction of all PHI, date shifting relative to the date of birth, and removal of DICOM headers and pathology metadata containing identifiers.