COMPLLLM: Fine-tuning LLMs to Discover Complementary Signals for Decision-making

Anonymous Authors[1]

Abstract

Multi-agent decision pipelines can outperform single agent workflows when *complementarity* holds, i.e., different agents bring unique information to the table to inform a final decision. We propose COMPLLLM, a post-training framework based on decision theory that fine-tunes a decision-assistant LLM using complementary information as reward to output signals that complement existing agent decisions. We validate COMPLLLM on synthetic and real-world tasks involving domain experts, demonstrating how the approach recovers known complementary information and produces plausible explanations of complementary signals to support downstream decision-makers.

1. Introduction

Multi-agent collaboration (including with humans) is increasingly adopted in complex decision workflows. For example, a clinician may consult a computer vision model and a written report from a radiologist to decide whether to order a biopsy for the patient; a moderator decides whether approve a post based on reviews from different human raters; a program chair decides paper acceptance based on reviewer comments and automated assessments from an LLM. A central challenge to the downstream decision-maker who must integrate inputs from upstream agents is determining complementary information: when and how does each source of information contribute information that could improve the final decision over and above existing agents' decisions?

Most machine-learning interpretability methods are not designed to address complementarities. Explanations typically characterize why a single model produced its output, often with respect to that model's own inputs (e.g., feature attributions, saliency, rationales). But in collaborative workflows, the bottleneck is different. For a clinician using a vision model in diagnosis, for example, what is often needed is not simply a restatement of the model's reasoning, but a list of features of patient information that are inconsistent with existing agent recommendations.

Our work formalizes available-but-overlooked evidence as **complementary signals**. We consider settings with two "agents": an upstream agent that produces a recommendation Z (e.g., a vision model's risk score on an X-ray image), and a supervisor agent that has access to potentially complementary unstructured information T (e.g., a text such as a radiology report). The goal is to identify a set of discrete, interpretable *signals* (i.e., findings) $\mathbf{S}$ in T that capture complementary decision-relevant information not already conveyed by Z, i.e., $\mathbf{S}$ that can improve the best-attainable decision conditioned on Z.

We propose COMPLLLM, a framework that (1) defines complementary value using the best-attainable performance on the decision problem, and (2) trains a language model to act as a complementary signal extractor from unstructured text. Formally, COMPLLLM learns a mapping from unstructured textual information (T) and recommendations (Z) to a set of structured signals, $\texttt{LLM} : \mathcal{T} \times \mathcal{Z} \to \mathcal{S}$, which prioritizes signals that meaningfully improve best-attainable decision quality relative to using the recommendation alone, rather than signals that are merely frequent or important. This shifts the role of "explanation" from justifying the agent recommendation to surfacing actionable complementary information that a supervisor should consider precisely because it is not already reflected in Z. COMPLLLM can be used for cases where the two agents are distinct, or where they represent the same human or model, i.e., an agent makes an initial decision using its available information, and then uses an LLM to do a second pass to surface overlooked cues in that same information.

We validate COMPLLLM across multiple domains. We first use a synthetic setting where complementary signals are controlled by construction to test the recoverability of the complementary signals. We then demonstrate use of the framework on three real-world decision-making tasks: identifying complementary signals in radiology reports that improve upon a vision model's recommendation regarding cardiac dysfunction; identifying the signals that are more or less focal to a specific group compared to the average

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

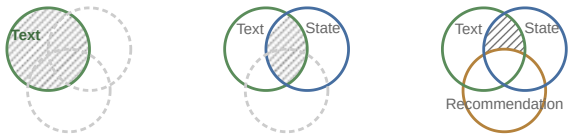


Figure 1. A graphical representation of how our work goes beyond simply identifying concepts that appear in text or signals that are predictive of a target state, by also ensuring that signals complement the existing recommendation. The diagonal stripes (twill pattern) represent the target of the corresponding methods.

human annotator in content moderation; and identifying the information that an LLM-authored review misses in human-written reviews. We further evaluate the relevance of the signals in the radiology diagnosis task by eliciting qualitiative feedback from two medical domain experts (one cardiologist and one internist). For the paper reviewing task, we validate that offering the complementary signals improves the accuracy of the LLM (Gemini 2.0 Flash) judgment on the paper acceptance.

2. Related Work

2.1. Topic & Hypothesis Generation

A large literature studies topic (or *concept*) generation as a way to summarize or organize corpora: classical probabilistic models such as LDA infer latent topics as word distributions (Blei et al., 2003), while neural and embedding-based variants improve coherence and interpretability (e.g., ProdLDA/AVITM (Srivastava & Sutton, 2017), Embedded Topic Model (ETM) (Dieng et al., 2020), and other neural variational topic models such as Miao et al. (2017)). Recent representation-first pipelines treat topic modeling as clustering in embedding space (e.g., Top2Vec (Angelov, 2020), BERTopic (Grootendorst, 2022), and newer embedding-centric formulations (Angelov & Inkpen, 2024)), and LLMs increasingly support prompt-based topic generation/labeling with greater steerability (e.g., TopicGPT (Pham et al., 2023)). These methods primarily target concepts or topics (i.e., patterns grounded in text) without necessarily requiring a link to a target state.

A parallel line of work targets hypothesis (or *signal*) generation, which produces interpretable candidate factors that are grounded in text and also correlated with a state of interest (Figure 1). Examples include using SAEs to generate hypotheses from latent features with LLM interpretations (e.g., HypotheSAEs (Movva et al., 2025)), learning natural-language descriptions of distributional differences and using them in goal-driven discovery (Zhong et al., 2024; Zhou et al., 2024), and using LLM-proposed differences followed by statistical validation for causal inference on text-derived outcomes (Modarressi et al., 2025). As shown in Figure 1, our work shifts focus to identify complementary signals, i.e., signals grounded in the supervisor information that add incremental decision value relative to an existing recommendation (not merely frequent, salient, or globally predictive signals).

2.2. Human-AI Decision-making

A growing body of work studies AI-assisted human decision-making based on its importance for legal and ethical accountability (Bo et al., 2021; Boskemper et al., 2022; Bondi et al., 2022; Schemmer et al., 2022). A recent meta-analysis (Vaccaro et al., 2024) finds that, on average, human–AI teams perform worse than the better of the two agents alone. A growing body of work seeks to evaluate and enhance complementarity in human–AI systems (Bansal et al., 2019; 2021b;a; Wilder et al., 2021; Hemmer et al., 2022; Holstein et al., 2023; Rastogi et al., 2023; Mozannar et al., 2024b; Guo et al., 2025). Some approaches explicitly incorporate human expertise in developing machine learning models or human-AI collaboration pipelines, such as by learning to defer (Mozannar et al., 2024a; Raghu et al., 2019; Keswani et al., 2022; 2021; Okati et al., 2021; Chen et al., 2022). Others develop alternative algorithms, e.g., with provable guarantees, that exploit cases where humans have additional contextual knowledge (Alur et al., 2024; Corvelo Benz & Rodriguez, 2023; Straitouri et al., 2023; De Toni et al., 2024; Arnaiz-Rodriguez et al., 2025). Closest to our work, Guo et al. (2025) provide a framework for assessing the complementary information value of arbitrary signals in a decision context. We demonstrate using complementary information as a fine-tuning objective, enabling LLM-based explanations of unexploited signals in unstructured text available at decision time.

3. Theoretical Framework

We consider a decision-making task with an upstream, recommending agent and a downstream, supervisor agent where, for each realization of the process, the recommending agent makes a recommendation $Z \in \mathcal{Z}$ based on their own features $X \in \mathcal{X}$, and the supervisor aggregates their information $T \in \mathcal{T}$ with Z to make a final decision $D \in \mathcal{D}$. The goal of the supervisor is to determine whether there is more information in T–in the form of a set of signals $\mathbf{S} \in \mathcal{S}$–over Z that can improve their utility.

We do not assume any relationship between the recommending agent's feature representation X and the supervisor's information T. For example, in the radiology diagnosis task, the features of the recommending agent (vision model) X are the X-ray image, and the features of the supervisor (clinician) T are the radiology report and the patient's medical history. In the content moderation task, the features of the recommending agent (a specific group of human annotators)

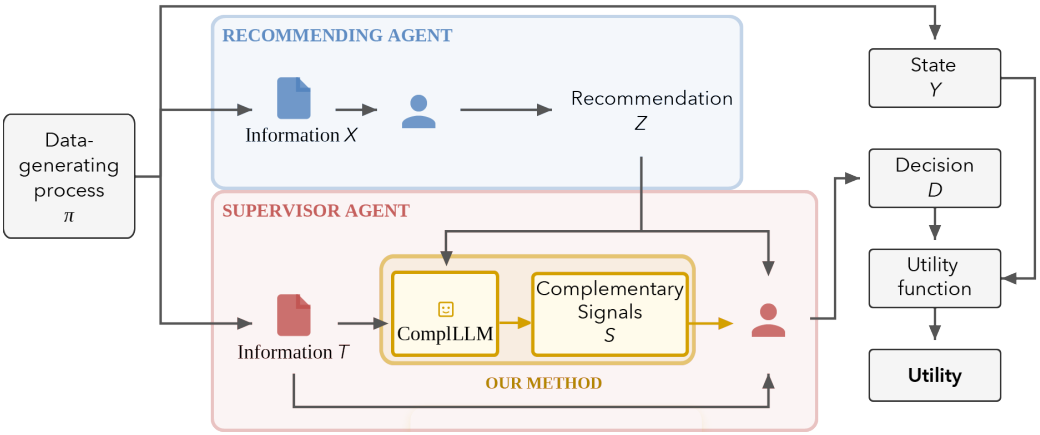


Figure 2. The "two agents" setting in our framework. The recommending agent makes a recommendation Z based on their own features X, and the supervisor agent aggregates their own information T with Z to make a final decision D.

X and the features of the supervisor (moderator) T are both the text content.

We use decision theory to characterize the decision-making process as a *decision problem* and an *information structure*. The decision problem consists of three components: the state space $\mathcal{Y}$, the decision space $\mathcal{D}$, and the utility function $U : \mathcal{D} \times \mathcal{Y} \to \mathbb{R}$. The state $y \in \mathcal{Y}$ represents the underlying state of the world that is relevant to the supervisor's utility. Since the utility function U implicitly identifies the other two components $\mathcal{Y}$ and $\mathcal{D}$, henceforce, we will use U to denote a decision problem.

The information structure is a joint distribution over the space of features of the recommending agent $\mathcal{X}$, the features of the supervisor $\mathcal{T}$, and the state $\mathcal{Y}$, denoted as $\pi \in \Delta(\mathcal{X} \times \mathcal{T} \times \mathcal{Y})$. Given a set of observations $\{(x_i, t_i, y_i)\}_{i\in[N]}$, we can estimate π on recommendation Z and derived discrete signals $\mathbf{S}$, assuming that the state is discrete, e.g., $Y \in \{0, 1\}$. We denote the probability of observing $Z = z$, $\mathbf{S} = \mathbf{s}$, and $Y = y$ as $\pi(z, \mathbf{s}, y)$. We denote the set of all possible signals as a combination of M basic signals, i.e., $\mathcal{S} = \{0, 1\}^M$. Each of the basic signals is a binary random variable, indicating the presence or absence of a certain signal, e.g., Table 1. We use upper-case letters to represent the random variables, e.g., $\mathbf{S}$, Z, Y, and lower-case letters to represent a value, i.e., $\mathbf{s}$, z, y. We denote the j-th basic signal as $\mathbf{S}_j \in \{0, 1\}$, and its absence or presence as $\mathbf{s}_j \in \{0, 1\}$.

The value of a signal is defined as the improvement that can be attained in the best attainable performance when that signal is provided. Concretely, a signal realization $\mathbf{s}$ induces posterior distribution $\pi(y|\mathbf{s})$. The best attainable performance ($\hat{U}$) with the signal is the expected payoff of the decision that best-responds to the posterior distribution:

$$\hat{U}(\pi(y|S), y) = \max_{d\in\mathcal{D}} \mathop{\mathbb{E}}_{y\sim\pi(y|S)} [U(d, y)] .$$

We denote the language model that extracts the signals as $\texttt{LLM}(\cdot) : \mathcal{T} \to \mathcal{S}$. The **complementary value** of $\texttt{LLM}(T)$ over the agent's decision Z in a decision problem U is the difference between the expected best attainable payoff on observing $\texttt{LLM}(T)$ and Z, and the expected best attainable payoff on observing Z:

$$\mathcal{V}^{U,Z}(\texttt{LLM}) := \mathop{\mathbb{E}}_{t,z,y\sim\pi} \left[\hat{U}(\pi(y|\texttt{LLM}(t), z), y)\right] - \mathop{\mathbb{E}}_{z,y\sim\pi} \left[\hat{U}(\pi(y|z), y)\right] \quad (1)$$

We demonstrate our framework in Table 1, using a radiology diagnosis task.

Notation	Radiology Diagnosis instantiation
Y	Payoff state $\in \{0, 1\}$ for the presence of cardiac dysfunction.
Z	Prediction from the vision model on the probability of the presence of cardiac dysfunction $\in [0, 1]$.
X	Features of the vision model (X-ray image).
T	Features of the clinician (radiology report and patient's medical history).
D	Final decision by the clinician $\in \{0, 1\}$ (e.g., whether to order a biopsy).
Signals in T	
$\mathbf{S}_1$	presence of enlarged cardiac silhouette $\in \{0, 1\}$
$\mathbf{S}_2$	presence of pleural abnormalities $\in \{0, 1\}$.
...	

Table 1. Notation summary (instantiated with the example of radiology-diagnosis).

4. Methods

Given a training dataset $\{(t_i, z_i, y_i)\}_{i\in[N]}$, where t_i is the supervisor's information, z_i is the agent's recommendation, and y_i is the state, and a decision problem U, we fine-tune a large language model $\texttt{LLM}$ parameterized by θ to extract a set of signals $\mathbf{s}_i = \texttt{LLM}_\theta(t_i)$ that maximizes the complementary value over the existing agent's recommendation z_i, as defined in Equation (1).

This approach involves three steps. First, we estimate the

data-generating process as the joint distribution between the signals, agent decision, and the state. We use this estimated distribution as the reference model to define the complementary value. Second, we identify a set of complementary signals on each instance with the reference model and do supervised fine-tuning with the identified complementary signals. Last, we use reinforcement learning to further enhance the complementary value extracted by the LLM without using the labeled complementary signals.

4.1. Estimating the Data-Generating Process

We estimate the data-generating process by prompting an LLM model in two rounds. In the first round, we identify the space of possible signals, denoted as $\mathcal{S} = \{0,1\}^M$, and in the second round, we identify whether a signal occurs in each instance (t_i, z_i, y_i), denoted as $\{\mathbf{s}_i^{(0)} : \mathbf{s}_i^{(0)} \in \mathcal{S}\}_{i\in[N]}$.

We identify the space of signals by initializing a set of signals on each instance. We use an LLM reference model $\texttt{LLM}_{\text{ref}}$ to output the signals that occur in the supervisor's information t_i. The union of the signals across all instances forms the space of all possible signals in our estimate, $\mathcal{S} = \{0,1\}^M$, wher M is the number of signals in the union set. To improve stability, we sample $\zeta = 7$ outputs at temperature 0.7 and only keeps the signals with frequency larger than $N_\tau * \zeta$ [1].

In the second round, for each signal $j \in [M]$ and each instance $i \in [N]$, we prompt $\texttt{LLM}_{\text{ref}}$ again to only output whether signal j occurs in t_i. We sample $\zeta = 7$ outputs at temperature 0.7 and use a majority vote, i.e., we mark the signal as occurred, $\mathbf{s}_{i,j}^{(0)} = 1$, if more than half of the samples indicate the signal occurs. We estimate the posterior distribution $\pi(y|\cdot)$ by a regression model (Algorithm 1) using $\{\mathbf{s}_i^{(0)}, y_i\}_{i\in[N]}$, which we subsequently use to generate training data for SFT and the reward function for RL.

4.2. Supervised Fine-tuning (SFT) with Generated Complementary Signals

Generating Complementary Signals. We generate the training data for SFT as the complementary signals $\mathbf{s}_i^{(1)} \in \{0,1\}^M$ for each instance (t_i, z_i, y_i). Given $\mathbf{s}_i^{(0)}$ denoting the occurance of basic signals in t_i, we generate $\mathbf{s}_i^{(1)}$ by selecting those basic signals that both occur and have larger best-attainable payoff than the agent decision z_i on instance i (more than a minimum threshold ϵ), as these signals convey decision-relevant information that is unexploited by z_i.

$$\begin{aligned} &\mathbf{s}_i^{(1)} = (\mathbf{s}_{i,1}^{(1)}, \ldots, \mathbf{s}_{i,M}^{(1)}) \\ &\text{s.t. } \mathbf{s}_{i,j}^{(0)} = 1 \text{ for any } \mathbf{s}_{i,j}^{(1)} = 1, \text{ and} \\ &\qquad \hat{U}(\pi(y|\mathbf{s}_i^{(1)}, z_i), y_i) > \hat{U}(\pi(y|z_i), y_i) + \epsilon \end{aligned} \tag{2}$$

Supervised Fine-tuning. We fine-tune the LLM $\texttt{LLM}_\theta$ using the generated complementary signals $\mathbf{s}_i^{(1)}$ as ground-truth labels. To improve the efficiency of this process ((Wang et al., 2025; Gandhi et al., 2025)), we use $\texttt{LLM}_{\text{ref}}$ to generate a Chain-of-Thought (CoT) given the supervisor's information t_i, the agent's decision z_i, and the ground truth of $\mathbf{s}_i^{(1)}$. We require the format of the CoT to identify the following: evidence for signals, the relevance to the state, and the complementary value relevant to the agent decision.

4.3. Reinforcement Learning

We squentially fine-tune $\texttt{LLM}_\theta$ after SFT using **Group Relative Policy Optimization (GRPO)** (Shao et al., 2024) to maximize the objective given in Equation (1).

Reward function. We design the reward function using the best-attainable payoff. The reward function $R : \mathcal{S} \times \mathcal{Z} \times \mathcal{Y} \to \mathbb{R}$ rewards signals only when they increase best-attainable payoff beyond the agent's recommendation, i.e., when they provide complementary information value. We only score signals that are *supported* by the supervisor's information to prevent rewarding hallucinations, where $\mathbf{s}$ is supported if every basic signal in $\mathbf{s}$ occurs in the supervisor's information, i.e., $\mathbf{s}_{i,j}^{(0)} = 1$ for all $j \in [M]$ such that $\mathbf{s}_j = 1$.

$$R(\mathbf{s}, z_i, y_i) = \begin{cases} 1 & \text{if both } \mathbf{s} \text{ and } \mathbf{s}_i^{(1)} \text{ are empty} \\ (\hat{U}(\pi(y|\mathbf{s}, z_i), y_i) - \hat{U}(\pi(y|z_i), y_i))/\alpha_i & \\ \qquad \text{else if } \mathbf{s} \text{ are supported} & \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where $\alpha_i = \hat{U}(\pi(y|\mathbf{s}_i^{(1)}, z_i), y_i) - \hat{U}(\pi(y|z_i), y_i)$ is the best attainable payoff with the identified complementary signals in SFT training data, which we use as a normalizer.

Group Relative Policy Optimization (GRPO). We fine-tune $\texttt{LLM}_\theta$ using the GRPO algorithm (Shao et al., 2024). At each training step, the model generates G candidate signals $\mathbf{s}_i^1, \cdots, \mathbf{s}_i^G$ for each instance i. The advantage function of GRPO is defined as the reward difference between the candidate signal and the average reward of the candidates.

$$A_j = R(\mathbf{s}_i^j, z_i, y_i) - \frac{1}{G}\sum_{k=1}^{G} R(\mathbf{s}_i^k, z_i, y_i) \tag{4}$$

With the above advantage function, we train $\texttt{LLM}_\theta$ using the objective function defined in (Shao et al., 2024).

[1] $N_\tau = \frac{z_{1-\delta/2}^2 p(1-p)}{\epsilon^2}$. $z_{1-\delta/2}$ is the $1-\delta/2$ quantile of the standard normal distribution and p is the prior $\pi(y)$. This is to ensure ample sample size for each signal, such that the estimation error of the posterior distribution of $y|\mathbf{s}_j$ is bounded by ϵ with confidence $1-\delta$ for the binary state space.

5. Experiments

We evaluate COMPLLLM by first testing its ability to **recover complementary signals** given by construction, then apply the approach to **evaluate practical utility** in three real-world scenarios: medical diagnosis, context moderation, and scientific paper reviewing. Across all experiments, we instantiate the "two agents" setting with the supervisor's information t, recommending agent's features x, and payoff state y.

5.1. Model and Training Details

We use *Qwen3-8B* to generate the signals for estimating the data-generating process and as the backbone language model for fine-tuning. For SFT, we train with 2 epochs, with a learning rate of 5×10^{-6} and a cosine learning rate scheduler. For GRPO, we initialize the training with 10 epochs but early stop when the improvement of the reward on validation set is less than 0.01, with a learning rate of 1×10^{-6} and a cosine learning rate scheduler. We sample 12 candidate completions for each instance in GRPO. For the hyperparameters that determine the threshold of the sample size N_τ for the data-generating process estimation, we use $\epsilon = 0.1$ and $\delta = 0.05$. We use the output token length of 1,500 tokens.

5.2. Baselines & Benchmark

We compare COMPLLLM with the following methods:

ZERO-SHOT and FEW-SHOT learning. We provide LLMs with task-specific instructions (zero-shot) and optionally with three demonstration examples (few-shot). We choose the demonstration examples from the generated SFT training dataset described in Section 4.2.

Topic-generation: BERTOPIC (Grootendorst, 2022) We compare with topic-generation methods that are agnostic to the existing agent decisions and the decision problems. BERTOPIC is a neural topic modeling method that produces topics by clustering text embeddings. We fit a multivariate logit model to predict the state y from the topic ID and the recommendation z. We assign each topic a natural language description by prompting Qwen3-8B with a sample of documents.

Hypothesis-generation: HYPOTHESAE (Movva et al., 2025) We compare with recent hypothesis-generation methods that take into account the decision problem and payoff-relevant state, but are agnostic to the existing agent decisions. HYPOTHESAE clusters the documents by the predictive power of selected neuron activations on the state y, and then produces hypotheses by prompting an LLM with a sample of documents in each cluster.

Benchmark To benchmark the value that text t and recommendation z have for predicting the ground-truth label y, we fine-tune Qwen3-8B with GRPO to predict the state, serving as a non-interpretable (i.e., without discrete signals) benchmark for achievable performance.

To ensure comparability, we use the same training, validation, and test splits for all the methods. We use the same model as used in COMPLLLM (Qwen3-8B), to summarize and annotate the topics in the BERTopic method and the hypotheses of the HypotheSAE method. See the hyperparameters of baseline & methods in Appendix E.

5.3. Tasks, Datasets, and Metrics

5.3.1. COMPLEMENTARITY BY CONSTRUCTION

We test how well COMPLLLM recovers artificially induced complementary signals, using the medical decision problem of diagnosing cardiac dysfunction defined in Table 1.

Dataset and task. We use radiology reports as t from the MIMIC-CXR dataset (Johnson et al., 2023) with a ground-truth signal set $\mathbf{s}^*$ derived from the CheXpert labels (Irvin et al., 2019) on these reports. To generate the state y and upstream recommendation z for each instance, we fit a logistic regression model that regresses real-world cardiac dysfunction labels on the CheXpert labels. We derived the cardiac dysfunction labels from blood tests related to cardiac dysfunction (*troponin* and *NT-proBNP*) in MIMIC-IV, using domain-suggested age-cutoffs (Mueller et al., 2019; Heidenreich et al., 2022) to threshold into binary values. We generate y by inputting the full set of report labels from CheXpert to this model, and generate z by holding out a subset of labels (*Edema* and *Pleural Effusion*) so they are, by construction, predictive of y but not represented in z. To be closer to the realistic setting where doctors are assisted by a probabilistic prediciton, we threshold the prediction by 0.5 to get a binary $y \in \{0, 1\}$ and keep $z \in [0, 1]$ as the probablisitic predictions. We use 6K/2K/4K instances for training/validation/test.

Metrics. We evaluate how well the output signals recover the ground-truth complementary signals, and the complementary information value of the output signals over the existing agent's decisions.

- **Surface Similarity**: For each test instance, we prompt Qwen3-14B to score surface similarity for every pair of presented ground-truth signal (i.e., a ground-truth signal with label 1 for that instance) and output signal. Specifically, we prompt Qwen3-14B to assign scores of 1, 0.5, and 0 for signals that are the same, related, or distinct, respectively. For stability, we report the average similarity score over all matched pairs given 7 repetitions and temperature 0.7. For each presented ground-truth signal, we take the maximum similarity over output signals (i.e., the best match), and report these per-ground-truth maxima averaged across all instances.
- **F1 Similarity**: For each instance, we compute the F1 score between the presence of the ground-truth signals

and whether a output signal is matched with it (i.e., has surface similarity ≥ 0.5). We report the average F1 score over all instances and all ground-truth signals.

- **Complementary Information Value (Improvement on accuracy by signals)**: We report the complementary information value $\mathcal{V}^{U,Z}(\texttt{LLM})$ from Equation (1) with accuracy $U(d, y) = \mathbb{1}[d = y]$[2] as the decision problem. This represents the improvement in best-attainable performance from observing extracted signals $\texttt{LLM}(T)$ in addition to Z. We use a multivariate logit model to fit the payoff state y on the extracted signals $\texttt{LLM}(T)$ and the existing agent's decision Z to compute the best-attainable performance.

The first and second metrics are derived from Zhong et al. (2024), who similarly evaluate recovery of ground-truth signals.

5.3.2. Complementarity in Decision-making

We test CompLLM on three realistic decision-making tasks: medical diagnosis, content moderation, and scientific paper reviewing. More details in Appendix A.

Medical Diagnosis: MIMIC-CXR (Johnson et al., 2023). We investigate what information in human-generated radiology reports can be used to improve a vision model's predictions of cardiac dysfunction on X-ray images. We use the radiology reports from MIMIC-CXR as t, the chest X-ray images from MIMIC-CXR as x, and the cardiac dysfunction labels derived from blood tests in MIMIC-IV as $y \in \{0, 1\}$, as described in the above experiment. We finetune the CXR foundation model (Sellergren et al., 2023) on the cardiac dysfunction labels based on the X-ray images and the blood test results and use its probablistic predictions as the agent recommendation $z \in [0, 1]$. We use use the same training/validation/test as above.

Content Moderation: DICES (Aroyo et al., 2023). We investigate what toxicity cues in human-LLM conversations a specific group of human annotators use differently from the majority-vote average annotations. We use the DICES dataset, which contains 115K human annotations of toxicity on 1.3K human-LLM conversations with demographic information about annotators. We use annoatations from one demographic group (i.e., 7K annotations from the largest demographic group of annotators: Asian millennial women with college degree or higher) as the recommending agent $z \in \{0, 1\}$ for non-toxicity and toxicity, and the average human annotation as the state $y \in \{0, 1\}$. We use the conversation text as t. We use 4K/1K/2K instances for training/validation/test respectively.

Scientific Paper Reviewing: Review5K (Weng et al., 2025). We investigate what information in human-written reviews is missed by an LLM review's decision. We use the Review5K dataset, which contains 5K scientific papers with human-written reviews and the final decision for each paper. We generate the LLM review by prompting Gemini 2.0 Flash to review the paper and provide a judgement on the acceptance of the paper between "Accept", "Unsure", and "Reject". We use the human-written review text as t and the final judgement of the LLM review as the recommending agent $z \in \{0, 0.5, 1\}$ for "Reject", "Unsure", and "Accept" respectively. We use the final decision (made by human AC in the real review process) on the paper as the state $y \in \{0, 1\}$ for "Reject" and "Accept" respectively. We use 3K/1K/1K instances for training/validation/test respectively.

Metrics.

- **Complementary Information Value**: Same as the metric for synthetic data.
- **Breadth**: We fit a multivariate logit model of y on z and $\texttt{LLM}(t)$, and report the number of extracted signals whose coefficients are significantly non-zero[3] when controlling for z (with multiple-comparisons correction). This captures how many extracted signals add unique information beyond the recommendation.
- **Qualitative assessment**: We conducted interviews with two practicing physicians–one cardiologist (P1) and one internist (P2)–both of whom are also professors of medicine. During the sessions, we walked through four patient cases. For each case, we first showed the radiology report, image, and vision model prediction, and asked for their agreement and reasons from the report or image on why they agreed or disagreed. We then revealed the complementary signals, and asked them assess their relevance and whether they aligned or not with their domain knowledge. Finally, we present an updated prediction based on a multivariate logit model predicting cardiac dysfunction from the surfaced signals and agent decisions, and asked if they trusted this prediction more or less than the original.

Method	Surf.	F1
CompLLM	**0.98**	**0.67**
Zero-shot	0.91	0.42
Few-shot	0.95	0.38
BERTopic	0.84	0.17
HypotheSAE	0.90	0.38

Table 2. Results comparing average surface similarity and F1 score of the extracted signals by each method on the synthetic dataset.

6. Results

6.1. Complementarity by Construction

CompLLM recovers the complementary signals on the synthetic dataset. Table 2 shows that **CompLLM beats**

[2]We threshold the probilistic decision by 0.5 into binary.

[3]We used a Bonferroni-correct p-value threshold of 5×10^{-3}.

Table 3. Signals on MIMIC-CXR dataset that are significant with $p < 0.05$ after Bonferroni correction. 'Δ Acc.' represents the marginal accuracy improvement each signal $\mathbf{s}_j$ marginally provides over the other signals and the agent recommendation. 'Total Acc.' shows the combined accuracy when all significant signals are included, i.e., sum of the signals' marginal gain. '# Sig' indicates the number of statistically significant signals discovered by each method.

Source	Total Acc.	# Sig	Significant Signals	Δ Acc.
Agent Decision	0.819	-	-	-
CompLLM	0.839	12	*negative pneumothorax*	+0.0048
			positive pleural effusion	+0.0040
			negative pleural effusion	+0.0032
			positive pulmonary congestion	+0.0024
			positive cardiac silhouette enlargement	+0.0016
			positive cardiac enlargement	+0.0008
			uncertain pleural effusion	+0.0008
			uncertain cardiomegaly	+0.0006
			positive pulmonary vascular congestion	+0.0006
			negative pulmonary edema	+0.0002
			negative cardiac enlargement	+0.0002
			positive cardiomegaly	+0.0002
HypotheSAE	0.831	6	*mentions no pleural effusion*	+0.0032
			mentions absence of pleural effusion or pneumothorax	+0.0024
			mentions pulmonary edema	+0.0020
			mentions presence of pleural effusion	+0.0020
			mentions presence of pulmonary edema	+0.0018
			mentions pulmonary edema or interstitial opacities	+0.0008
Zero-Shot	0.825	5	*cardiac enlargement*	+0.0014
			pleural effusion	+0.0012
			bilateral pleural effusion	+0.0012
			interstitial edema	+0.0010
			pulmonary vascular engorgement	+0.0010
Few-Shot	0.823	3	*no pleural effusion*	+0.0014
			positive interstitial edema	+0.0012
			positive cardiac silhouette enlargement	+0.0010
BERTopic	0.818	1	*endotracheal tube position*	+0.0005

all baseline methods in surface similarity and F1 score. As shown in Table 4, the signals extracted by CompLLM **cover the ground-truth complementary signals**. We also find that **fine-tuning helps recover ground-truth complementary signals**, i.e., though few-shot learning outputs signals that are more similar to the ground-truth signals than zero-shot learning, it does not predict those signals at the correct instance (i.e., low F1 score). The signals extracted by CompLLM also provide the **highest complementary information value** over the existing agent's decision recommendation in the synthetic dataset across all the methods (except for the non-interpretable benchmark) in Figure 3.

6.2. Complementarity in Decision-making

Figure 3 shows the complementary information value of the signals for the real-world datasets. Tables 3, 5 and 6 shows all significant signals on the MIMIC-CXR, Review5k, and DICES datasets.

6.2.1. Identifying complementary signals

CompLLM identifies the signals with the highest complementary information value among the methods in our experiments. Comparing to the baseline methods, CompLLM is the only method that extracts signals with significant complementary information value (i.e., with a non-overlapping confidence interval with the accuracy of the recommending agent) in all three datasets (as shown in Figure 3). This implies that designing for complementarity is necessary for improving over the existing agent's decision recommendation. We also find that CompLLM achieves a comparable performance with the non-interpretable benchmark in MIMIC-CXR dataset and DICES dataset (Figure 3), suggesting that extracting complementary signals does not harm the predictive power of the same LLM model in the decision problem in these two tasks. We also find that CompLLM extracts more signals with complementary information value than the baseline methods in all three datasets (Table 3), suggesting that the signals extracted by CompLLM are more diverse and comprehensive.

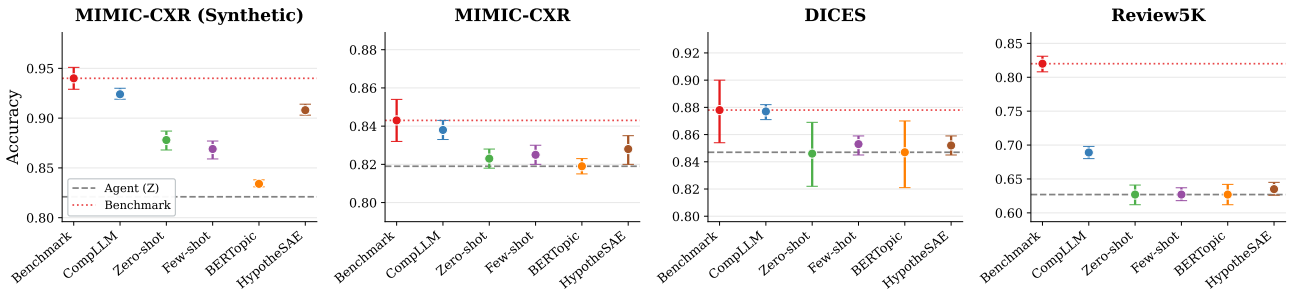


Figure 3. Expected accuracy given the extracted signals and the agent's recommendation by each method. Dashed lines represent agent decision accuracy and the accuracy of the benchmark method. Error bars depict bootstrapped 95% confidence intervals (N=5000).

6.2.2. Qualitative Feedback on CompLLM for Medical Diagnosis

The physicians viewed multiple overlapping cases (3 and 4 respectively, for a total of 6). Both **appraised the complementary signals as aligning with their domain knowledge**, with two exceptions. The first was a case where P1 noticed a signal that they believed provided relevant complementary information (the patient's history of cardiac problems), but was missed by CompLLM. Additionally, both physicians indicated that one of the complementary signals for a case (Figure 10) both saw, which, though named as *negative_pleural_effusion*, referred to absence of pleural effusion and pneumothorax (Figure 11), was only half relevant, as the second condition neither associated strongly with cardiac dysfunction or the most likely alternative condition. In one other case, P2 did not dispute the complementary signals, but felt they might be consistent with–rather than adding new information over–the original prediction.

In an exit interview, both physicians described seeing value for CompLLM in practice, including to assist in creating lists of supporting versus contradictory evidence that clinicians already make (P1, P2), and to support ER doctors and general internists who have less specific training in cardiology (P1).

6.2.3. Improved LLM Decisions on Paper Reviewing

We validate the usefulness of the complementary signals by giving the signals to a proxy, LLM decision-maker (Gemini 2.0 Flash) and seeing if they improve its performance in the scientific paper reviewing task. Figure 4 shows the accuracy of the LLM's paper review decision relative to the human AC's ground truth decision on the Review5K dataset. We test two settings over the baseline of giving the LLM only the paper text: also giving the LLM the human review text, and also giving the LLM the human review text and the complementary signals. We find that the LLM with the complementary signals as input achieves higher accuracy than the LLM only taking the paper text and the human review text as input, (79.7% [95% CI: 78.9%, 80.5%] vs. 78.8% [95% CI: 75.9%, 77.6%]).

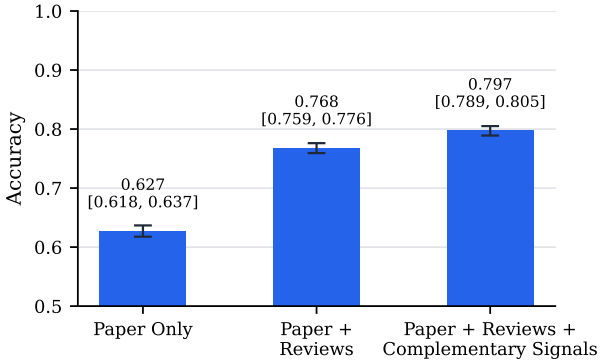


Figure 4. Accuracy of LLM paper review decisions on the Review5K dataset. Error bars depict bootstrapped 95% confidence intervals (N=5000).

7. Limitations

CompLLM uses the estimated data-generating process as the reference model to fine-tune and reinforce the LLM. We expect this approach to identify all important signals. However, there is a risk that CompLLM drops rare but important signals, i.e., signals that have frequency lower than N_τ in the dataset but would be considered important to domain experts.

We identify the signals and their decision-relevant value by the posterior distribution of the state and best-attainable performance given the occurrence of the signal as determined by the reference model. However, when the recommending agent and supervisor are the same, as in AI-assisted human decision workflows where a human first arrives at an independent judgment, then consults an AI, providing complementary signals may induce learning such that the predicted complementary value of signals no longer holds in hindsight. Future work could extend this by formalizing a continual learning problem that accounts for changes to human beliefs about the state after viewing the complementary signals and updates the LLM model correspondingly.[4]

[4]The data and code to reproduce our experimental results are available at https://osf.io/z5gyj/overview?view_

Impact Statement

Our work advances the field of human-AI or multi-agent collaboration and decision-making, which stands to contribute to a number of public-facing and scientific domains. To the best of our knowledge, there are no particular negative social consequences imposed by our work compared to machine learning research in general.

We use the DICES dataset to study differences in labeling patterns. Any deployment should include domain-appropriate oversight and evaluation to avoid over-reliance on automatically surfaced signals.

References

Alur, R., Raghavan, M., and Shah, D. Distinguishing the indistinguishable: Human expertise in algorithmic prediction. *arXiv preprint arXiv:2402.00793*, 2024.

Angelov, D. Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*, 2020.

Angelov, D. and Inkpen, D. Topic modeling: Contextual token embeddings are all you need. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 13528–13539, 2024.

Arnaiz-Rodriguez, A., Benz, N. C., Thejaswi, S., Oliver, N., and Gomez-Rodriguez, M. Towards human-ai complementarity in matching tasks. *arXiv preprint arXiv:2508.13285*, 2025.

Aroyo, L., Taylor, A., Diaz, M., Homan, C., Parrish, A., Serapio-García, G., Prabhakaran, V., and Wang, D. Dices dataset: Diversity in conversational ai evaluation for safety. *Advances in Neural Information Processing Systems*, 36:53330–53342, 2023.

Bansal, G., Nushi, B., Kamar, E., Weld, D. S., Lasecki, W. S., and Horvitz, E. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 2429–2437, 2019.

Bansal, G., Nushi, B., Kamar, E., Horvitz, E., and Weld, D. S. Is the most accurate ai the best teammate? optimizing ai for teamwork. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11405–11414, 2021a.

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., and Weld, D. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021b. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445717. URL https://doi.org/10.1145/3411764.3445717.

Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan): 993–1022, 2003.

Bo, Z.-H., Qiao, H., Tian, C., Guo, Y., Li, W., Liang, T., Li, D., Liao, D., Zeng, X., Mei, L., et al. Toward human intervention-free clinical diagnosis of intracranial aneurysm via deep neural network. *Patterns*, 2(2), 2021.

Bondi, E., Koster, R., Sheahan, H., Chadwick, M., Bachrach, Y., Cemgil, T., Paquet, U., and Dvijotham, K. Role of human-ai interaction in selective prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 5286–5294, 2022.

Boskemper, M. M., Bartlett, M. L., and McCarley, J. S. Measuring the efficiency of automation-aided performance in a simulated baggage screening task. *Human factors*, 64 (6):945–961, 2022.

Chen, C., Feng, S., Sharma, A., and Tan, C. Machine explanations and human understanding (2022). *URL: http://arxiv. org/abs/2202.04092*, 2022.

Corvelo Benz, N. and Rodriguez, M. Human-aligned calibration for ai-assisted decision making. *Advances in Neural Information Processing Systems*, 36:14609–14636, 2023.

De Toni, G., Okati, N., Thejaswi, S., Straitouri, E., and Rodriguez, M. Towards human-ai complementarity with prediction sets. *Advances in Neural Information Processing Systems*, 37:31380–31409, 2024.

Dieng, A. B., Ruiz, F. J., and Blei, D. M. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453, 2020.

Gandhi, K., Chakravarthy, A., Singh, A., Lile, N., and Goodman, N. D. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars. *arXiv preprint arXiv:2503.01307*, 2025.

Grootendorst, M. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

Guo, Z., Wu, Y., Hartline, J., and Hullman, J. The value of information in human-ai decision-making. *arXiv preprint arXiv:2502.06152*, 2025.

only=c32fcf0a14b54b9fa6ed22ddb8d1f774.

Heidenreich, P. A., Bozkurt, B., Aguilar, D., Allen, L. A., Byun, J. J., Colvin, M. M., Deswal, A., Drazner, M. H., Dunlay, S. M., Evers, L. R., et al. 2022 aha/acc/hfsa guideline for the management of heart failure: a report of the american college of cardiology/american heart association joint committee on clinical practice guidelines. *Journal of the American College of Cardiology*, 79(17): e263–e421, 2022.

Hemmer, P., Schemmer, M., Kühl, N., Vössing, M., and Satzger, G. On the effect of information asymmetry in human-ai teams. *arXiv preprint arXiv:2205.01467*, 2022.

Holstein, K., De-Arteaga, M., Tumati, L., and Cheng, Y. Toward supporting perceptual complementarity in human-ai collaboration via reflection on unobservables. *Proceedings of the ACM on Human-Computer Interaction*, 7 (CSCW1):1–20, 2023.

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 590–597, 2019.

Johnson, A. E., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-y., Mark, R. G., and Horng, S. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

Johnson, A. E., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Hao, S., Moody, B., Gow, B., et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

Keswani, V., Lease, M., and Kenthapadi, K. Towards unbiased and accurate deferral to multiple experts. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 154–165, 2021.

Keswani, V., Lease, M., and Kenthapadi, K. Designing closed human-in-the-loop deferral pipelines. *arXiv preprint arXiv:2202.04718*, 2022.

Miao, Y., Grefenstette, E., and Blunsom, P. Discovering discrete latent topics with neural variational inference. In *International conference on machine learning*, pp. 2410–2419. PMLR, 2017.

Modarressi, I., Spiess, J., and Venugopal, A. Causal inference on outcomes learned from text. *arXiv preprint arXiv:2503.00725*, 2025.

Movva, R., Peng, K., Garg, N., Kleinberg, J., and Pierson, E. Sparse autoencoders for hypothesis generation. *arXiv preprint arXiv:2502.04382*, 2025.

Mozannar, H., Bansal, G., Fourney, A., and Horvitz, E. When to show a suggestion? integrating human feedback in ai-assisted programming. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 10137–10144, 2024a.

Mozannar, H., Lee, J., Wei, D., Sattigeri, P., Das, S., and Sontag, D. Effective human-ai teams via learned natural language rules and onboarding. *Advances in Neural Information Processing Systems*, 36, 2024b.

Mueller, C., McDonald, K., de Boer, R. A., Maisel, A., Cleland, J. G., Kozhuharov, N., Coats, A. J., Metra, M., Mebazaa, A., Ruschitzka, F., et al. Heart failure association of the european society of cardiology practical guidance on the use of natriuretic peptide concentrations. *European journal of heart failure*, 21(6):715–731, 2019.

Okati, N., De, A., and Rodriguez, M. Differentiable learning under triage. *Advances in Neural Information Processing Systems*, 34:9140–9151, 2021.

Pham, C. M., Hoyle, A., Sun, S., Resnik, P., and Iyyer, M. Topicgpt: A prompt-based topic modeling framework. *arXiv preprint arXiv:2311.01449*, 2023.

Raghu, M., Blumer, K., Corrado, G., Kleinberg, J., Obermeyer, Z., and Mullainathan, S. The algorithmic automation problem: Prediction, triage, and human effort. *arXiv preprint arXiv:1903.12220*, 2019.

Rastogi, C., Leqi, L., Holstein, K., and Heidari, H. A taxonomy of human and ml strengths in decision-making to investigate human-ml complementarity. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 11, pp. 127–139, 2023.

Schemmer, M., Hemmer, P., Nitsche, M., Kühl, N., and Vössing, M. A meta-analysis of the utility of explainable artificial intelligence in human-ai decision-making. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 617–626, 2022.

Sellergren, A., Kiraly, A., Pollard, T., Weng, W.-H., Liu, Y., Uddin, A., and Chen, C. Generalized Image Embeddings for the MIMIC Chest X-Ray dataset. *PhysioNet*, February 2023. doi: 10.13026/pxc2-vx69. URL https://doi.org/10.13026/pxc2-vx69. Version 1.0.

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Srivastava, A. and Sutton, C. Autoencoding variational inference for topic models. *arXiv preprint arXiv:1703.01488*, 2017.

Straitouri, E., Wang, L., Okati, N., and Rodriguez, M. G. Improving expert predictions with conformal prediction. In *International Conference on Machine Learning*, pp. 32633–32653. PMLR, 2023.

Vaccaro, M., Almaatouq, A., and Malone, T. When combinations of humans and ai are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, pp. 1–11, 2024.

Wang, Z., Zhou, F., Li, X., and Liu, P. Octothinker: Mid-training incentivizes reinforcement learning scaling. *arXiv preprint arXiv:2506.20512*, 2025.

Weng, Y., Zhu, M., Bao, G., Zhang, H., Wang, J., Zhang, Y., and Yang, L. Cycleresearcher: Improving automated research via automated review. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=bjcsVLoHYs.

Wilder, B., Horvitz, E., and Kamar, E. Learning to complement humans. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pp. 1526–1533, 2021.

Zhong, R., Wang, H., Klein, D., and Steinhardt, J. Explaining datasets in words: Statistical models with natural language parameters. *Advances in Neural Information Processing Systems*, 37:79350–79380, 2024.

Zhou, Y., Liu, H., Srivastava, T., Mei, H., and Tan, C. Hypothesis generation with large language models. *arXiv preprint arXiv:2404.04326*, 2024.

Table 4. Summary of Extracted Signals by Method on Synthetic Dataset.

Method	Count	Signals
Ground Truth	2	Edema Pleural Effusion
COMPLLLM	16	positive_pleural_effusion positive_edema positive_cardiomegaly positive_pleural_effusion_left positive_pleural_effusion_right positive_pneumonia positive_edema_resolution positive_troponins positive_atelectasis no_pulmonary_edema ...
HYPOTHESAE	8	mentions presence of pulmonary edema mentions presence of interstitial edema mentions presence of interstitial lung disease or interstitial findings mentions no acute cardiopulmonary process mentions presence of pleural effusion and atelectasis mentions presence of pleural effusion mentions presence of bilateral pleural effusion mentions presence of interstitial pulmonary edema
BERTOPIC	15	Outlier/Noise Chest X-ray Findings Cardiac Enlargement Pancreatic Cancer and Chest Imaging Chest X-ray for TIA evaluation Chest X-ray for Upper GI Bleed Pulmonary Edema Monitoring Pulmonary Edema Evaluation Chest X-ray findings in edema Chest X-ray findings ...
FEW-SHOT	26	enlarged_cardiac_silhouette pulmonary_vascular_congestion pleural_effusion vascular_congestion moderate_cardiomegaly bilateral_pleural_effusions interstitial_edema cardiomegaly mild_pulmonary_edema pulmonary_congestion ...
ZERO-SHOT	55	pulmonary_vascular_congestion cardiac_enlargement pleural_effusions cardiac_silhouette_enlarged bilateral_pleural_effusions small_left_pleural_effusion no_cardiomegaly no_pleural_effusion moderate_cardiomegaly pulmonary_edema ...

Table 5. Signals on DICES dataset. Significant signals with $p < 0.05$ after Bonferroni correction. 'Δ Acc.' represents the incremental accuracy improvement each signal provides over the baseline agent decision (0.847). 'Total Acc.' shows the combined accuracy when all significant signals are included. '# Sig' indicates the number of statistically significant signals discovered by each method.

Source	Total Acc.	# Sig	Significant Signals	Δ Acc.
Baseline Agent Decision Accuracy: 0.847				
CompLLM	0.878	4	*promotes or condones violence* *misinformation conspiracy or false theory* *offers safe alternative* *oblique or profane language*	+0.0222 +0.0061 +0.0020 +0.0010
BERTopic	0.856	2	*Outlier/Noise* *Chat conversations*	+0.0051 +0.0030
HypotheSAE	0.854	2	*contains explicit insults or profanity* *mentions 'he' or 'him' in context of another person's actions*	+0.0051 +0.0010
Zero-Shot	0.847	0	—	—
Few-Shot	0.847	0	—	—

Table 6. Signals on Review5k dataset. Significant signals with $p < 0.05$ after Bonferroni correction. 'Δ Acc.' represents the incremental accuracy improvement each signal provides over the baseline agent decision (0.627). 'Total Acc.' shows the combined accuracy when all significant signals are included. '# Sig' indicates the number of statistically significant signals discovered by each method.

Source	Total Acc.	# Sig	Significant Signals	Δ Acc.
Agent Decision	0.627	-	-	-
CompLLM	0.692	8	*clarity* *novelty* *experimental validation* *theoretical contribution* *clarity of presentation* *insufficient comparison with related work* *missing baselines* *novelty limited*	+0.0310 +0.0179 +0.0110 +0.0024 +0.0021 +0.0002 +0.0002 +0.0002
HypotheSAE	0.649	6	*mentions evaluations on specific datasets and comparisons* *mentions graph-based techniques for efficiency or scalability* *mentions analysis of computational complexity* *mentions diffusion models* *mentions theoretical analysis and experimental validation* *mentions methodology compared to existing approaches*	+0.0058 +0.0053 +0.0043 +0.0041 +0.0030 +0.0010
BERTopic	0.627	1	*Test-Time Adaptation*	+0.0002
Zero-Shot	0.627	0	—	—
Few-Shot	0.627	0	—	—

A. Data Preprocessing

MIMIC-IV. We use data from the MIMIC dataset (Johnson et al., 2023), which contains anonymized electronic health records from Beth Israel Deaconess Medical Center (BIDMC), a large teaching hospital in Boston, Massachusetts affiliated with Harvard Medical School. We download the X-ray images and corresponding reports of radiologists from the MIMIC-CXR dataset (Johnson et al., 2019). We join the MIMIC-CXR dataset with MIMIC-IV by matching on patient and visit ID, and filter the radiology images and reports to those for which there is at least one follow-up blood test related to cardiac dysfunction for the patient at the same visit. We consider two types of blood tests following domain expert

suggestions (Mueller et al., 2019; Heidenreich et al., 2022): *Troponin* and *NT-proBNP*. We threshold by the age-cutoffs from Mueller et al. (2019); Heidenreich et al. (2022) in the Table 7 to diagnose cardiac dysfunction.

Table 7. Biomarker Thresholds by Age and Gender

Biomarker	Age	Gender	Threshold
NT-proBNP	≤ 49	–	> 449
	50–75	–	> 899
	≥ 76	–	> 1799
Troponin	< 64	Female	≥ 0.014
	≥ 65	Female	≥ 0.018
	< 50	Male	≥ 0.019
	50–64	Male	≥ 0.028
	≥ 65	Male	≥ 0.035

After filtering and processing, we got 12,146 records representing radiology reports with the corresponding X-ray images. We fine-tune the CXR foundation model (Sellergren et al., 2023) on a training set containing 8,502 images, and test on a hold-out validation set containing 3,644 images. The model acheives 81.9% accuracy on the hold-out set.

DICES. We download the DICES dataset from github (Aroyo et al., 2023), representing multi-turn adversarial conversations generated by human agents interacting with a dialog model. We merge two dataset from DICES, dataset 350 and dataset 990. We use the demographics (i.e., race, gender, age, and education) of the human annotators to group them. We pick the largest group to represent the recommending agent–Asian women with college degree or higher and were born in Millennials–and calculate the group's average annotation. We use the majority-vote annotation on each conversation to represent the state. We get 7,843 annotations from the group of Asian Millennials women with college degree or higher with the corresponding majority-vote annotations.

REVIW5K We download the Review5K dataset from huggingface (Weng et al., 2025). We generate the LLM judgment on whether the paper should be accepted by providing the paper in prompt. We use the final decisions on the papers from the dataset as the state, "Reject" or "Accept" (including the "Accept(poster)", "Accept(spotlight)", and "Accept(oral)"). We use the reviews written by humans as the supervisor's information. We include all 4,991 papers from Review5K in our experiment. We generate the LLM review judgment using the following prompt:

> You are an expert academic reviewer. Based on the following paper content, judge whether this paper will be accepted for publication at a top-tier conference (like ICLR).
> **Paper Content:**
> {paper_text_limited}
> Based on the paper's quality, novelty, technical soundness, clarity, and contribution, determine if this paper will be accepted.
> **IMPORTANT:** Respond with ONLY one word: "yes", "unsure", or "no". Do not include any other text or explanation.
> **Your judgment:**

Synthetic Dataset (MIMIC-CXR). We use the same X-ray images and reports as the MIMIC-CXR dataset for the synthetic dataset. Instead of the real cardiac dysfunction labels and predictions from the CXR-foundation model, we generate the synthetic ground-truth labels by a logistic regression on the annotated labels from CheXpert (Irvin et al., 2019). The CheXpert contains 13 labels of chest X-ray findings: *Atelectasis*, *Cardiomegaly*, *Consolidation*, *Edema*, *Enlarged Cardiomediastinum*, *Fracture*, *Lung Lesion*, *Lung Opacity*, *Pleural Effusion*, *Pneumonia*, *Pneumothorax*, *Pleural Other*, *Support Devices*, *No Finding*. We exclude the labels of *No Finding* and *Support Devices* from the synthetic ground-truth labels since they are not related to cardiac dysfunction. Since the CheXpert labels the findings into three categories–positively mentioned, negatively mentioned, and uncertain–we translate them into one-hot encoded binary labels to train the logistic regression model. The coefficients of the logistic regression model are shown in Table 8, with accuracy of 0.8182 and Brier score of 0.8742.

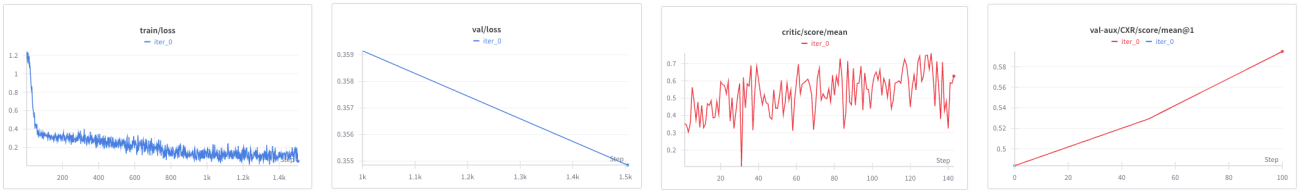
Table 8. Coefficients of the logistic regression model on the synthetic ground-truth labels. * indicates the labels are withheld from the model to generate the recommending agent's decision.

Condition	Positive	Negative	Uncertain
Atelectasis	0.3384	0.3633	−0.1758
Cardiomegaly	0.7462	−0.3945	0.3062
Consolidation	0.9686	−0.0945	0.6488
Enlarged Cardiomediastinum	0.2368	−0.9585	0.1678
Fracture	0.3986	−0.7871	0.2674
Lung Lesion	0.3194	0.1314	0.0970
Lung Opacity	0.7512	0.1924	0.7063
Pneumonia	0.5593	0.0380	0.2329
Pneumothorax	1.1280	0.2994	0.1718
Pleural Effusion*	1.6320	0.0933	0.9735
Edema*	1.8391	0.7875	1.3925

B. Hyperparameters & Training Details

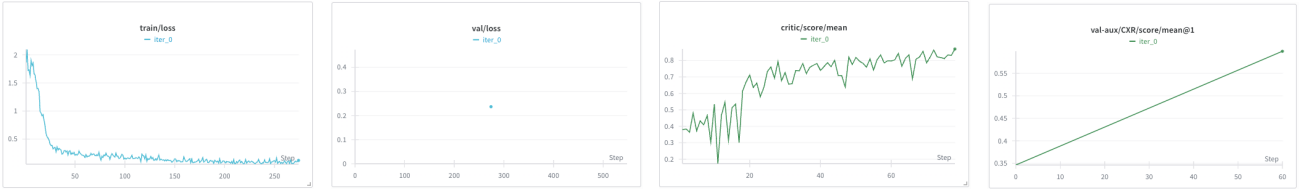
Estimating the Data-Generating Process. We set temperature to 0.7 when prompting the reference LLM model to extract the signals. We tested out different sample numbers $\zeta \in \{2, 3, 7, 14\}$ and choose the one that resulted in the best validation performance of the regression model in Algorithm 1. We filter out the rare signals extracted from the dataset by the frequency threshold N_τ to ensure the stability of the estimated data-generating process, with $\epsilon = 0.1$ and $\delta = 0.05$.

Fine-tuning the LLM. For SFT, we train with 2 epochs, with a learning rate of 5×10^{-6} and a cosine learning rate scheduler. For GRPO, we initialize the training with 10 epochs but early stop when the improvement of the reward on the validation set is less than 0.01, with a learning rate of 3×10^{-6} and a cosine learning rate scheduler. We sample 12 candidate completions for each instance in GRPO. We use limit the `max_prompt_length` to 1,024 tokens for the MIMIC-CXR dataset and DICES dataset and 4,096 tokens for the Review5K dataset, given that much longer reviews in the Review5K dataset than the radiology reports in the MIMIC-CXR dataset and the dialogs in the DICES dataset. We use the same output token length of 1,500 tokens for all the datasets. We trained on $4\times$H100 GPUs, resulting in a total training time of 1.5 hours for SFT and 30 hours for GRPO for each dataset. Figures 5 to 8 show the training curves of SFT and GRPO on the MIMIC-CXR, DICES, Review5K, and synthetic datasets respectively.



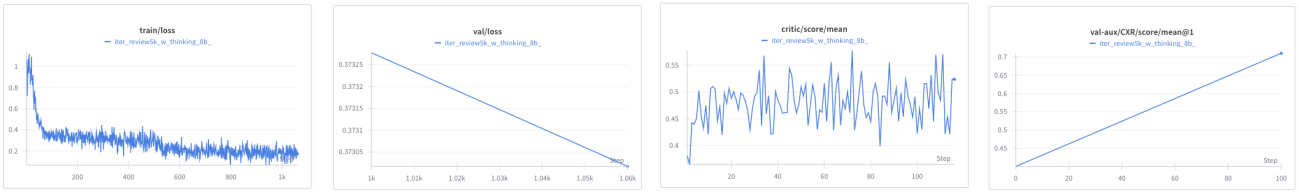
(a) SFT training loss (b) SFT validation loss (c) GRPO critic reward (d) GRPO validation reward

Figure 5. Training curves of SFT and GRPO on the MIMIC-CXR dataset.



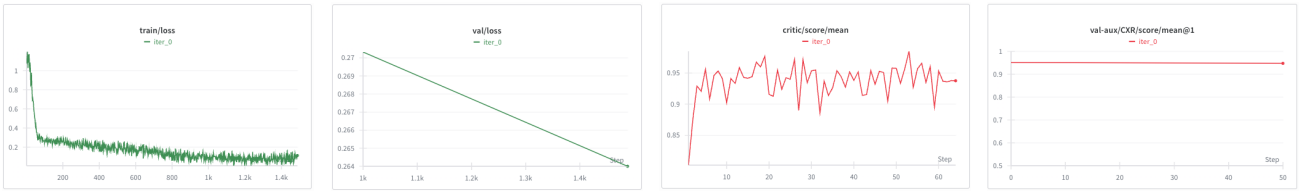
(a) SFT training loss (b) SFT validation loss (c) GRPO critic reward (d) GRPO validation reward

Figure 6. Training curves of SFT and GRPO on the DICES dataset.



(a) SFT training loss (b) SFT validation loss (c) GRPO critic reward (d) GRPO validation reward

Figure 7. Training curves of SFT and GRPO on the Review5K dataset.



(a) SFT training loss (b) SFT validation loss (c) GRPO critic reward (d) GRPO validation reward

Figure 8. Training curves of SFT and GRPO on the synthetic dataset.

C. Algorithm for estimating posterior distribution

We estimate the posterior distribution of the state given the signals and the agent's decision by a regression model to estimate the data-generating process. We design Algorithm 1 to select the signals and their interactions to be included in the regression model. This algorithm greedily selects the signals and their interactions with the largest marginal improvement to the prediction of the payoff state. Thus it approximates the minimal subset selection of the signals and their interactions to be included in the regression model to achieve the best prediction of the state.

D. Experimental Results for CompLLM with Only SFT

We run our experiments using SFT only to evaluate the performance of CompLLM without the GRPO component. Figure 9 shows the experimental results for CompLLM with only SFT. We observe that CompLLM with only SFT achieves a comparable performance with the CompLLM method in the synthetic dataset, but sightly lower than the CompLLM method in the real-world datasets. We also observed that there is more variance in the performance of CompLLM with only SFT than the CompLLM method in the real-world datasets.

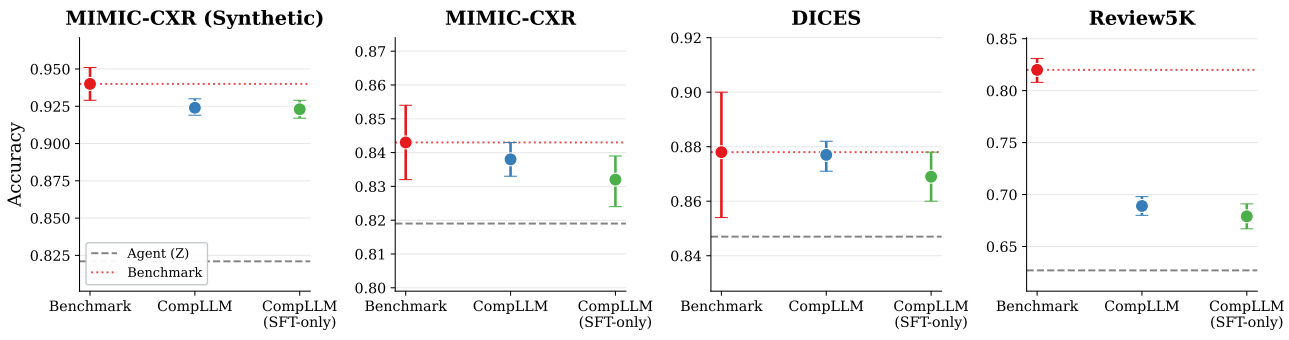


Figure 9. Experimental results for CompLLM with only SFT. Dashed lines represent agent decision accuracy and the accuracy of the benchmark method. Error bars depict bootstrapped 95% confidence intervals (N=5000).

E. Hyperparameters in Baseline & Benchmark Metrics

Zero-shot and Few-shot learning. We use the same LLM model choices as the backbone language model in CompLLM: Qwen3-8B. We use the same prompts for training and testing as the CompLLM (Appendix G). We randomly select

Algorithm 1 Greedy Feature and Interaction Selection

Require: Signals X, outcome Y, initial features $S_0 = \{Z\}$
Choose regression model and scoring function based on whether Y is binary or continuous
Initialize selected features $S \leftarrow S_0$

Phase 1: Select main effects
repeat
    Find the unused signal that most improves prediction
    if improvement exceeds threshold $\varepsilon_{\text{main}}$ **then**
        Add signal to S
    else
        stop
    end if
until no improvement

Phase 2: Select pairwise interactions
repeat
    Find the pair of selected signals whose interaction most improves prediction
    if improvement exceeds threshold ε_{int} **then**
        Add interaction to model
    else
        stop
    end if
until no improvement

Output: Final model using selected signals and their interactions

three examples from the SFT training dataset to generate the demonstration examples for few-shot learning. We insert the examples into the prompt with the following format:

—
Examples:
Example {`i`}:
Document: {`document`}
Agent Decision: {`agent_decision`}
Signals: {`signals`}

BERTopic. (Grootendorst, 2022) We use BERTopic's publicly available implementation to generate the topics[5]. We use the default model choices: all-MiniLM-L6-v2 for the embedding model, UMAP for dimensionality reduction, HDBSCAN for clustering, and c-TF-IDF to compute the top words associated with each topic. We tune the cluster size hyperparameter in BERTopic. We test the value in {10, 20, 50, 100, 200}. We choose the parameter values by maximizing the validation performance of the multivariate logit model on the state y. We use default values for the other hyperparameters.

HypotheSAE. (Movva et al., 2025) We use HypotheSAE's publicly available implementation[6]. We use the default OpenAI embedding model chosen by the HypotheSAE authors. We tune the number of hidden neurons (M) and the maximum number of active neurons (K) in the sparse autoencoder with a grid search in $(M, K) \in \{(64, 4), (256, 8), (1024, 8), (1024, 32), (2048, 32)\}$. We choose the parameter values by maximizing the validation performance of the multivariate logit model on the state y. We use the following task-specific instructions to generate the hypotheses.

[5] https://github.com/MaartenGr/BERTopic
[6] https://github.com/rmovva/HypotheSAEs

For MIMIC-CXR dataset:

> All of the texts are chest X-ray radiology reports.
> Features should describe specific findings or patterns in the reports. For example:
> - ”mentions presence of pleural effusion”
> - ”describes cardiomegaly or enlarged heart”
> - ”notes clear lung fields without abnormalities”
> - ”mentions presence of atelectasis or lung collapse”

For DICES dataset:

> All of the texts are conversational responses.
> Features should describe specific aspects of the response. For example:
> - ”contains harmful or inappropriate content”
> - ”demonstrates bias or stereotyping”
> - ”is a safe and appropriate response”

For Review5K dataset:

> All of the texts are paper review documents.
> Features should describe specific aspects of the review. For example:
> - ”mentions positive aspects of the paper”
> - ”identifies methodological concerns”
> - ”notes issues with presentation or clarity”
> - ”describes soundness or technical quality”

F. Materials for Qualitative Study

Figures 10 to 12 shows an example of the radiology report and X-ray image, the output of CompLLM, and the updated probablistic prediction shown to the physicians in the qualitative study respectively, including the questions we asked the doctors on screen.

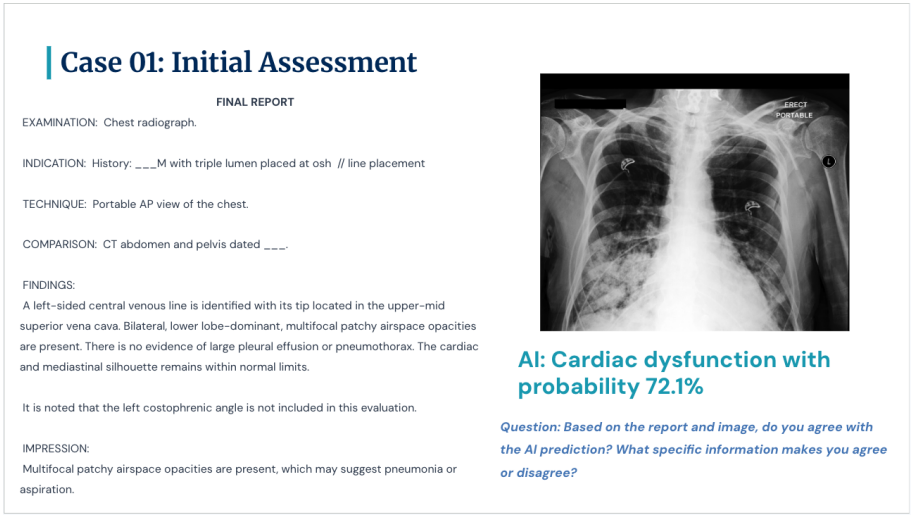


Figure 10. An example of the radiology report and X-ray image shown to the physicians in the qualitative study.

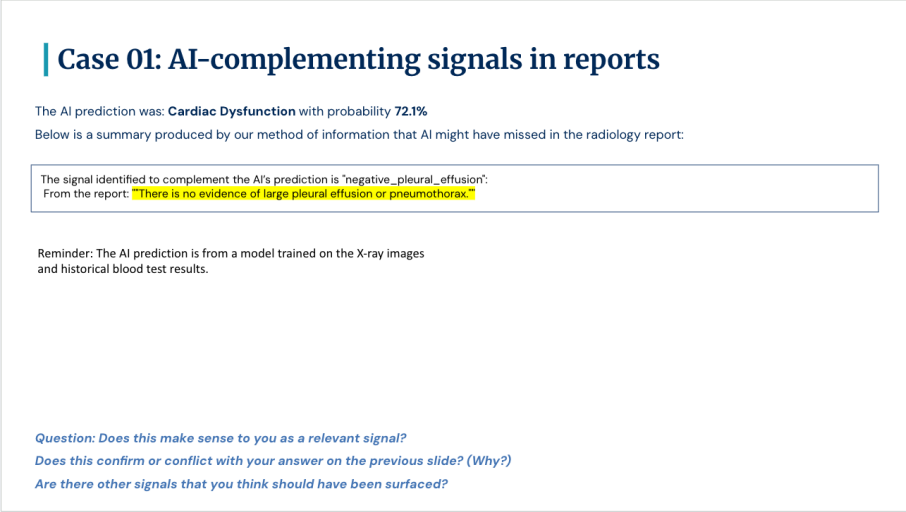


Figure 11. An example of COMPLLLM's output shown to the physicians in the qualitative study.

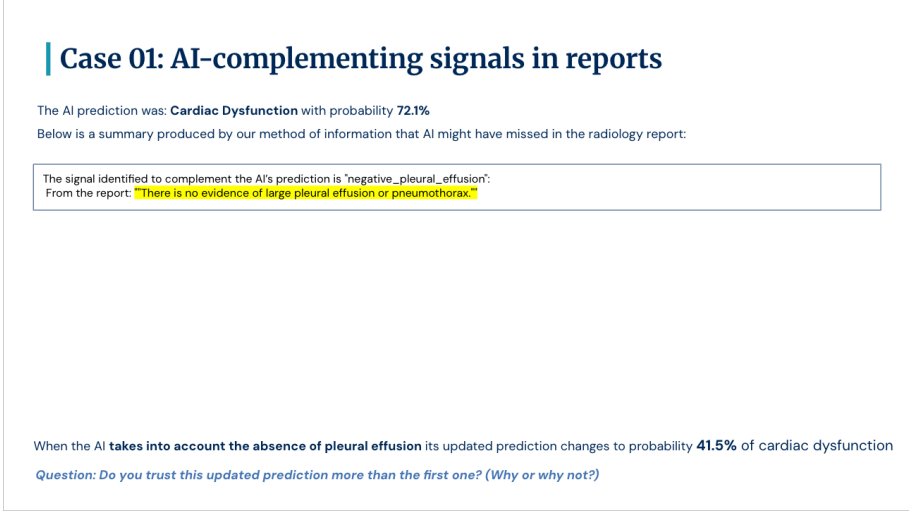


Figure 12. An example of COMPLLLM's output and the updated probablistic prediction shown to the physicians in the qualitative study.

G. Prompts

The prompt for extracting the signals to identify the signal space when estimating the data-generating process:

You are a clinical reasoning assistant that extracts radiological findings about cardiac dysfunction from chest X-ray reports.
Your input fields are:
'radiology_report' (str): a detailed radiology report for a chest X-ray.
Your task is to identify **clinical signals** in the report.
These signals must be:
- *Clinically relevant* to cardiac dysfunction (e.g., pleural effusion, pulmonary edema, cardiomegaly);
- *Explicitly mentioned* in the report (as present, absent, or uncertain);
If **no such signal can be confidently found**, you must output an *empty list* '[]'.
—
Suppression rules

- Do NOT output a signal if the finding is only implied or indirectly suggested.
- Do NOT output a signal if polarity cannot be clearly determined as **present**, **absent**, or **uncertain**.
- Do NOT output more than one polarity for the same base signal.

—

Output constraints
Your output must:
- Follow **valid JSON** syntax parsable by 'json.loads'.
- Contain **no commentary, explanations, or reasoning traces**.
- Include **no more than {k}** signals.
- Each signal must contain **exactly one field**:
- 'name' (str): lowercase, underscore-separated, polarity-encoded identifier.

—

Decision rule
Output a signal *only if*:
(a) it is explicitly stated as present, absent, or uncertain in the report, and
(b) it provides *information* about cardiac dysfunction.
If evidence is weak or ambiguous beyond explicit uncertainty, **output an empty list**.

—

Expected output format
[[radiology_report]]
`{radiology_report}`
[[signals]]
`[ {"name": "example_signal"} ]`
[[completed]]

The prompt for generating the reasoning traces for SFT:

You generate *structured clinical reasoning traces* explaining why a given set of complementary radiological signals (S) provide additional information beyond a model prediction (p). You MUST ground your reasoning strictly in the radiology report.
Your input fields:
- 'radiology_report' (str): the full chest X-ray report
- 'model_prediction' (float): probability of cardiac dysfunction assigned by an external model
- 'signals' (list[dict]): a list of ground-truth complementary signals, each with a 'name' and optional 'description'
Your output:
Produce a detailed thinking trace inside ¡thinking¿...¡/thinking¿ with EXACTLY the following sections:
IF 'signals' IS NON-EMPTY:
1. **EVIDENCE FROM REPORT (EXTRACTIVE)**
- For each signal in 'signals', quote the exact report span(s) that support it.
- If the report mentions a finding using different phrasing, point that out.
- If the report does not explicitly mention the signal, state: "No explicit mention; inferred from wording X".
2. **CLINICAL RELEVANCE**
- For each signal, explain why it is clinically relevant for cardiac dysfunction.
3. **COMPLEMENTARY VALUE RELATIVE TO MODEL PREDICTION p**
- Explain why this signal adds information *not already encoded* in p.
- Consider under-detection, subtle findings, uncertainty, or clinical decision thresholds.
- **If the report contains other cardiac-related findings that are *not* included in 'signals', explicitly state that these findings are assumed to be already correlated with or captured by the model prediction p, and therefore do not provide complementary information. Do NOT argue that they should have been included as complementary signals.**
IF 'signals' IS EMPTY:
1. **WHY NO COMPLEMENTARY SIGNALS WERE FOUND**

- Identify report content that suggests normal findings or absence of pathology.
- Explain whether the report lacks any findings strongly associated with cardiac dysfunction.
- Note if all cardiac-related findings are either explicitly normal, clinically insignificant, or already well captured by the model.
2. **MODEL PREDICTION p CONTEXT**
- Explain why the absence of complementary signals is consistent or inconsistent with p.
- For example: p already captures the risk, or the report does not contain findings that would shift the prediction.
Do NOT generate the final 'signals' list. That is provided as ground truth.
[[radiology_report]]
{document}
[[model_prediction]]
{agent_decision}
[[signals]]
{signals}

The prompt used for training and testing with SFT and GRPO:

You are a clinical reasoning assistant that extracts *missed or underweighted* radiological findings by a model prediction about cardiac dysfunction from chest X-ray reports.
Your input fields are:
1. 'radiology_report' (str): a detailed radiology report for a chest X-ray.
2. 'model_prediction' (float): the predicted probability of cardiac dysfunction from an external model.
Your task is to identify complementary clinical signals from the report.
A complementary clinical signal is a radiological finding that:
1. Is explicitly stated in the radiology report
2. Is clinically relevant to cardiac dysfunction
3. Provides additional or corrective information about cardiac dysfunction risk beyond what the model prediction already captures
Decision rules:
- Output a signal only if all three conditions are clearly satisfied
- If evidence is weak, ambiguous, or uncertain, output nothing
- If no complementary signals exist, output an empty list
- Output no more than k signals
Output requirements:
- Output only a list of signal names
- Each signal name must be lowercase and use underscores
- Do not include explanations, reasoning, or any extra text
—
Expected output format
[[radiology_report]]
{radiology_report}
[[model_prediction]]
{model_prediction}
[[signals]]
[{"name": "example_signal"}]
[[completed]]
—