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

Anonymous Author(s)
Affiliation
Address
email

Abstract

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

1 Introduction

Multi-agent collaboration is increasingly adopted in complex decision workflows. For example, a clinician or AI decision agent may consult a computer vision model and a written report from a radiologist to decide whether to order a biopsy for the patient. A content moderator decides whether to approve a post based on reviews from different human raters. A program chair decides on paper acceptance based on reviewer comments and automated assessments from an LLM. In such settings, the central challenge for the downstream decision-maker is how to integrate information from different sources. The decision-maker must determine which parts of the available information are relevant to the task, which parts are already reflected in an upstream agent's recommendation, and which parts provide complementary evidence that could improve the final decision. This requires sufficient attention to identify relevant information embedded in long unstructured texts, as well as domain expertise and statistical reasoning to determine which signals are not redundant with other sources like existing agent decisions.

Instead of leaving a decision-maker to reason over all available unstructured information on their own, a promising idea is to surface a more compact set of signals, optimized to present the most decision-relevant information that is also complementary in light of existing agent recommendations. However, most machine-learning interpretability methods are not designed to address complementarities nor decision-making per se [20]. 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). In collaborative workflows, the bottleneck is often 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 not considered by existing agent recommendations.

Our work formalizes available-but-unintegrated 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

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

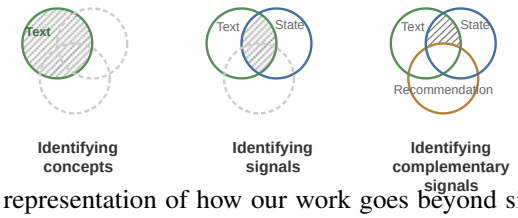


Figure 1: A graphical representation of how our work goes beyond simply identifying concepts that appear in text or signals that are predictive of a target state, by also ensuring that signals complement the existing recommendation. The diagonal stripes (twill pattern) represent the target of the corresponding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

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 on cardiac dysfunction; identifying the signals that are more or less focal to a specific group compared to the average human annotator in content moderation; and identifying the information that an LLM-authored review misses in human-written reviews. We evaluate the effectiveness of the signals generated by COMPLLLM by a preregistered within-subject human study for content moderation, a qualitative interview with two medical domain experts (one cardiologist and one internist) for radiology diagnosis, and an LLM experiment (Gemini 2.0 Flash) for paper reviewing.

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 [9], while neural and embedding-based variants improve coherence and interpretability (e.g., ProdLDA/AVITM [43], Embedded Topic Model (ETM) [17], and other neural variational topic models such as **(author?)** [30]). Recent representation-first pipelines treat topic modeling as clustering in embedding space (e.g., Top2Vec [2], BERTopic [19], and newer embedding-centric formulations [3]), and LLMs increasingly support prompt-based topic generation/labeling with greater steerability (e.g., TopicGPT [37]). 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 [32]), learning natural-language descriptions of distributional differences and using them in goal-driven discovery [49, 50], and using LLM-proposed differences followed by statistical validation for causal inference on text-derived outcomes [31]. 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).

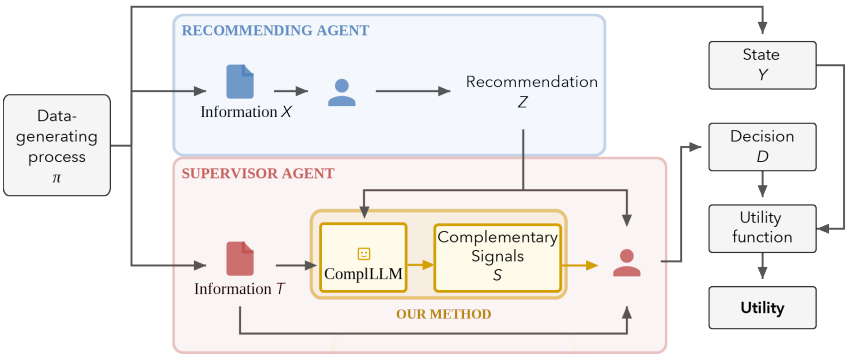


Figure 2: The "two agents" setting in our framework. The recommending agent makes a recommendation Z based on their own features X, and the supervisor agent aggregates their own information T with Z to make a final decision D.

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 [10, 12, 11, 40]. A recent meta-analysis [45] 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 [7, 8, 6, 48, 23, 24, 39, 34, 21]. Some approaches explicitly incorporate human expertise in developing machine learning models or human-AI collaboration pipelines, such as by learning to defer [33, 38, 29, 28, 36, 14]. Others develop alternative algorithms, e.g., with provable guarantees, that exploit cases where humans have additional contextual knowledge [1, 15, 44, 16, 4]. Closest to our work, **(author?)** [21] 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) 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 $\mathcal{D}$, henceforth, 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

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

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 $\in \{0, 1\}$.
...	

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 $\pi(y|\mathbf{s})$. We define $\hat{U}$ as the payoff of the decision that best-responds to the posterior distribution on the realization $Y = y$.

$$\hat{U}(\pi(\cdot|\mathbf{s}), y) := U(\operatorname*{argmax}_{d \in \mathcal{D}} \mathop{\mathbb{E}}_{y' \sim \pi(\cdot|\mathbf{s})} [U(d, y')], y)$$

The best attainable performance given $\mathbf{S}$ is the expected payoff of the best-responding payoff $\hat{U}$: $\mathbb{E}_{\mathbf{s}, y \sim \pi} \left[\hat{U}(\pi(\cdot|\mathbf{s}), y) \right]$. The **complementary value** of the signal extracted by a language model $\texttt{LLM}(\cdot, \cdot) : \mathcal{T} \times \mathcal{Z} \to \mathcal{S}$ in a decision problem U is the difference between the expected best attainable payoff on observing $\texttt{LLM}(T, Z)$ and Z, and the expected best attainable payoff on observing Z:

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

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

4 Methods

Given a training dataset $\{(t_i, z_i, y_i)\}_{i \in [N]}$ and a decision problem U, we fine-tune a large language model $\texttt{LLM}_\theta$to extract a set of signals $\mathbf{s}_i = \texttt{LLM}_\theta(t_i, z_i)$ that maximizes the complementary value over z_i, as defined in Equation (1). 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 use the reference model to construct SFT targets, which is defined as the minimal supported signal set that maximizes the improvement in best-attainable payoff over the recommendation. Last, we further optimize the model with GRPO using a reward proportional to the complementary information value of supported 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, $\mathcal{S} = \{0, 1\}^M$, where M is the number of signals in the union set. To improve stability, we sample $\zeta = 7$ outputs at temperature 0.7 and only keep 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 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.

[1] $N_\tau = \frac{z^2_{1-\delta/2} p(1-p)}{\epsilon^2}$, where $z_{1-\delta/2}$ 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.

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 occurence of basic signals in t_i, we generate $\mathbf{s}_i^{(1)}$ by selecting the minimal set of basic signals that both occur and maximize the improved best-attainable payoff over the agent decision z_i on instance i (which we require to be more than a minimum threshold ϵ), as these signals convey decision-relevant information that is unexploited by z_i.

$$\begin{aligned} &\min_{\mathbf{s}_i^{(1)}} |\mathbf{s}_i^{(1)}| \\ s.t.\ \mathbf{s}_i^{(1)} \in &\underset{\text{supp}(\mathbf{s}) \subseteq \text{supp}(\mathbf{s}_i^{(0)})}{\text{argmax}} \left[\hat{U}(\pi(\cdot|\mathbf{s}, z_i), y_i) - \hat{U}(\pi(\cdot|z_i), y_i)\right], \text{ and} \\ &\hat{U}(\pi(\cdot|\mathbf{s}_i^{(1)}, z_i), y_i) > \hat{U}(\pi(\cdot|z_i), y_i) + \epsilon \end{aligned}$$

where $\text{supp}(\mathbf{s}) = \{j : \mathbf{s}_j = 1\}$. If there is no solution (i.e., no signals improve over the agent decision more than ϵ), we set $\mathbf{s}_i^{(1)}$ as empty (i.e., no complementary signals identified).

Supervised Fine-tuning. We fine-tune $\texttt{LLM}_\theta$ using the generated complementary signals $\mathbf{s}_i^{(1)}$ as ground-truth labels. To improve the efficiency of this process [46, 18], 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 sequentially fine-tune $\texttt{LLM}_\theta$ after SFT using **Group Relative Policy Optimization (GRPO)** [42] 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 $\mathbf{s}_j = 1$.

$$R(\mathbf{s}, z_i, y_i) = \begin{cases} 1 & \text{if both } \mathbf{s} \text{ and } \mathbf{s}_i^{(1)} \text{ are empty} \\ (\hat{U}(\pi(\cdot|\mathbf{s}, z_i), y_i) - \hat{U}(\pi(\cdot|z_i), y_i))/\alpha_i & \text{if } \mathbf{s} \text{ is supported and } \mathbf{s}_i^{(1)} \text{ is not empty} \\ 0 & \text{else} \end{cases} \tag{2}$$

where $\alpha_i = \hat{U}(\pi(\cdot|\mathbf{s}_i^{(1)}, z_i), y_i) - \hat{U}(\pi(\cdot|z_i), y_i)$ is the best attainable payoff on instance i 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 [42]. 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.

$$A_j = R(\mathbf{s}_i^j, z_i, y_i) - \frac{1}{G}\sum_{k=1}^{G} R(\mathbf{s}_i^k, z_i, y_i) \tag{3}$$

With the above advantage function, we train $\texttt{LLM}_\theta$ using the objective function defined in [42].

5 Experiments

We evaluate the value of CompLLM in assisting decision-making in three steps. First, we test whether CompLLM can **recover complementary signals** when such signals are known by construction. Second, we evaluate the **theoretical complementary value** of the extracted signals in three real-world decision scenarios: medical diagnosis, content moderation, and scientific paper reviewing. Third, we test the **practical value** of the signals through qualitative interviews with domain experts, a decision study with human participants, and an LLM decision study. 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* as both the signal-generation reference model and the backbone for fine-tuning. We train with SFT followed by GRPO. Implementation details, hyperparameters, and prompts are provided in Appendices C and H.

5.2 Baselines & Benchmark

We compare COMPLLLM against four interpretable baselines: zero-shot prompting, few-shot prompting, BERTopic as a concept-generation baseline (*agnostic to the existing agent decisions and the decision problems*), and HypotheSAEs as a hypothesis-generation baseline (*agnostic to the existing agent decisions*). We also report a non-interpretable benchmark obtained by fine-tuning Qwen3-8B to predict the state directly from t and z. All methods use the same train/validation/test splits. Additional baseline details are in Appendix F.

5.3 Tasks, Datasets, and Metrics

5.3.1 Recovering Complementary Signals by Construction

We first test how well COMPLLLM recovers artificially induced complementary signals using the medical decision problem of diagnosing cardiac dysfunction (Table 1).

Dataset and task. We use radiology reports as t from the MIMIC-CXR dataset [26], with a ground-truth signal set $\mathbf{s}^*$ derived from CheXpert labels [25] 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 derive the cardiac dysfunction labels from blood tests related to cardiac dysfunction, specifically *troponin* and *NT-proBNP*, in MIMIC-IV, using domain-suggested age cutoffs [35, 22] to threshold them into binary values.

We generate y by inputting the full set of CheXpert report labels into this model. We generate z by holding out a subset of labels, *Edema* and *Pleural Effusion*, so that they are predictive of y but not represented in z by construction. These held-out labels therefore serve as ground-truth complementary signals. To match the realistic setting where doctors are assisted by a probabilistic prediction, we threshold the model prediction at 0.5 to obtain a binary state $y \in \{0, 1\}$, while keeping $z \in [0, 1]$ as the probabilistic recommendation. We use 6K/2K/4K instances for training, validation, and testing.

Metrics. We evaluate both whether the extracted signals recover those signals (following prior work [49]) and whether they improve decision-making beyond the upstream recommendation.

- **Surface Similarity**: For each test instance, we prompt Qwen3-14B to score the surface similarity between each presented ground-truth signal, i.e., a ground-truth signal with label 1 for that instance, and each output signal. Specifically, Qwen3-14B assigns 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 7 repetitions with temperature 0.7. For each presented ground-truth signal, we take the maximum similarity over output signals, corresponding to the best match, and average these per-ground-truth maxima across all instances.
- **F1 Similarity**: For each instance, we compute the F1 score between the presence of the ground-truth signals and whether an output signal is matched to each ground-truth signal, defined as having surface similarity ≥ 0.5. We report the average F1 score over all instances and all ground truths.
- **Complementary Information Value**: We report the complementary information value $\mathcal{V}^{U,Z}(\texttt{LLM})$ from Equation (1), using accuracy $U(d, y) = \mathbb{1}[d = y]$ as the decision payoff.[2] This measures the improvement in best-attainable performance from observing the extracted signals $\texttt{LLM}(T)$ in addition to the upstream recommendation Z. We use a multivariate logit model to fit the payoff state y on the extracted signals $\texttt{LLM}(T)$ and the upstream recommendation Z, and use this model to estimate the best-attainable performance.

5.3.2 Theoretical Complementary Value in Real-World Decision Tasks

We next evaluate whether COMPLLLM extracts signals with theoretical complementary value in three decision-making tasks and datasets: medical diagnosis, content moderation, and scientific paper reviewing. Additional dataset details are provided in Appendix B.

Medical Diagnosis: MIMIC-CXR [26]. We investigate what information in human-generated radiology reports can improve decisions over a vision model's predictions of cardiac dysfunction from chest X-ray images. We use the radiology reports from MIMIC-CXR as t, the chest X-ray

[2] We threshold the probabilistic decision at 0.5 to obtain a binary decision.

images from MIMIC-CXR as x, and the cardiac dysfunction labels derived from MIMIC-IV blood tests as $y \in \{0, 1\}$, following the construction described above. We fine-tune the CXR foundation model [41] to predict cardiac dysfunction from chest X-ray images and blood-test-derived labels, and use its probabilistic predictions as the upstream recommendation $z \in [0, 1]$. We use the same training, validation, and test split as in the construction-based experiment.

Content Moderation: DICES [5]. We investigate what toxicity cues in human-LLM conversations are missed by a specific group of human annotators relative to the majority-vote average annotation. We use the DICES dataset, which contains 115K human toxicity annotations on 1.3K human-LLM conversations, along with annotator demographic information. We use annotations from the largest demographic group of annotators, Asian millennial women with a college degree or higher, as the upstream recommendation $z \in \{0, 1\}$ for non-toxic and toxic, respectively. We use the majority-vote annotation from the entire population as the state $y \in \{0, 1\}$, and use the conversation text as t. We use 0.8K/0.2K/0.3K conversations for training, validation, and testing.

Scientific Paper Reviewing: Review5K [47]. We investigate what information in human-written reviews is missed by an LLM-generated review decision. We use the Review5K dataset, which contains 5K scientific papers with human-written reviews and final accept/reject decisions. We generate the LLM review by prompting Gemini 2.0 Flash to review each paper and provide a judgment among "Accept", "Unsure", and "Reject". We use the human-written review text as t, and the final judgment of the LLM review as the upstream recommendation $z \in \{0, 0.5, 1\}$ for "Reject", "Unsure", and "Accept", respectively. We use the final decision made by the human area chair in the real review process as the state $y \in \{0, 1\}$ for "Reject" and "Accept", respectively. We use 3K/1K/1K instances for training, validation, and testing.

Metrics. We report two metrics that quantify the complementary value of the extracted signals.

- **Complementary Information Value**: We use the same complementary information value metric as in the construction-based experiment.
- **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.

5.3.3 Practical Value in Downstream Decision-Making

Qualitative expert interviews. To assess whether the extracted medical signals are meaningful to domain experts, we conducted interviews with two practicing physicians: one cardiologist (P1) and one internist (P2), both of whom are also professors of medicine. During each session, we walked through four patient cases in depth. For each case, we first showed the radiology report, chest X-ray image, and vision model prediction, and asked the physician whether they agreed with the model prediction and what evidence from the report or image informed their judgment. We then revealed the complementary signals generated by CompLLM and asked the physician to assess their relevance and alignment with their domain knowledge. Finally, we presented an updated prediction from a multivariate logit model that predicts cardiac dysfunction from the surfaced signals and the upstream agent decision, and asked whether the physician trusted this updated prediction more or less than the original prediction.

Human decision study. We conducted a within-subject human study to evaluate whether complementary signals generated by CompLLM improve downstream decision-making in content moderation. We recruited 244 participants from Prolific and asked each participant to make decisions on 28 human-LLM conversations from the DICES dataset. The 28 conversations were divided into four sets of seven conversations, each of which was randomly assigned to one of four decision-assistance conditions: no additional signal, where participants only saw the conversation; signals from HypotheSAEs; signals from CompLLM; and direct predictions from the non-interpretable benchmark model used in Figure 3. We estimate decision accuracy for each assistance condition according to our preregistered analysis plan, using a Bayesian hierarchical model.[4] See more details in Appendix A.

LLM decision study. We also evaluated whether CompLLM signals improve LLM decision-making in the Review5K scientific paper reviewing task. For each instance, we prompted an LLM to judge whether the paper would be accepted under three information conditions: the original paper only; the original paper with human-written reviews; and the original paper with human-written reviews plus complementary signals generated by CompLLM. We report the accuracy of the LLM judgments under each condition.

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

[4] The model is preregistered at `https://aspredicted.org/7jx486.pdf`.

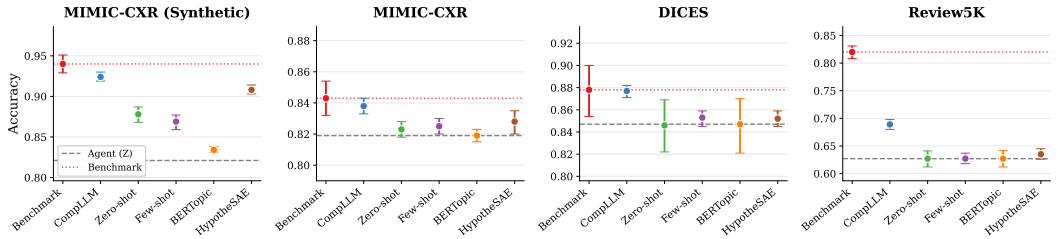


Figure 3: Theoretical best-attainable accuracy given the extracted signals and the agent's recommendation by each method. The grey dashed lines represent agent decision accuracy. Error bars depict bootstrapped 95% confidence intervals (N=5000).

6 Results

6.1 Recovering Complementary Signal by Construct

COMPLLLM best recovers the construction-based complementary signals. It achieves the highest surface similarity and F1 similarity among all methods (Table 2), and its extracted signals provide the largest complementary information value among interpretable methods (Figure 3). The qualitative examples in Table 3 further show that the signals extracted by COMPLLLM **cover the ground-truth complementary signals**. We also find that **fine-tuning is important for recovering instance-level complementary signals**. Although few-shot prompting produces signals that are more semantically similar to the ground-truth signals than zero-shot prompting, it does not reliably identify the correct signals for the correct instances, as reflected by its low F1 score. This suggests that prompt-only methods may recover plausible signal types but fail to localize them at the instance level.

Table 2: Results of average surface similarity and F1 score on the synthetic dataset.

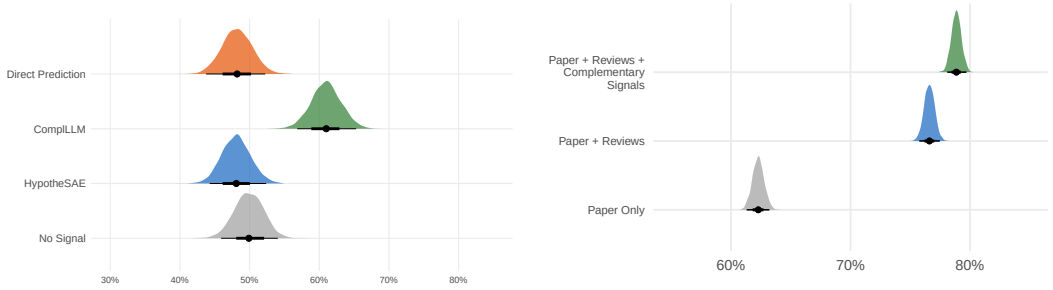
Method	SURF.	F1
COMPLLLM	**0.98**	**0.67**
ZERO-SHOT	0.91	0.42
FEW-SHOT	0.95	0.38
BERTOPIC	0.84	0.17
HYPOTHESAE	0.90	0.38

6.2 Theoretical Complementary Value in Real-World Decision Tasks

COMPLLLM extracts signals with the highest complementary information value among interpretable methods. Compared with the baseline methods, COMPLLLM is the only interpretable method that extracts signals with significant complementary information value in all three real-world datasets, as indicated by confidence intervals that do not overlap with the accuracy of the recommending agent in Figure 3. This result suggests that explicitly optimizing for complementarity is necessary for improving over the existing agent's recommendation. In MIMIC-CXR and DICES, COMPLLLM achieves performance comparable to the non-interpretable benchmark, suggesting that extracting discrete, interpretable complementary signals does not substantially sacrifice predictive value in these two tasks. At the same time, COMPLLLM produces signals that remain inspectable and usable by downstream decision-makers, unlike the benchmark model. **COMPLLLM identifies a broader set of complementary signals.** Across the three real-world datasets, COMPLLLM extracts more signals with significant complementary information value than the baseline methods, as shown in Tables 4 to 6.

6.3 Practical Value in Downstream Decision-Making

Qualitative Feedback on COMPLLLM 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 COMPLLLM. Additionally, both physicians indicated that one of the complementary signals for a case (Figure 11), named *negative_pleural_effusion* and referring to an absence of pleural effusion and pneumothorax (Figure 12), was only half relevant, as the second condition was 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 beyond–the original prediction. In an exit interview, both physicians were enthusiastic about the potential value for COMPLLLM 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).



(a) Accuracy of human toxicity decisions for the DICES dataset, represented as posterior predictive distributions from our fitted model. Intervals depict 95% and 75% highest density intervals (HDIs).

(b) Accuracy of LLM paper review decisions on the REVIEW5K dataset, represented by the density curves of the bootstrapped samples (N=5000). Intervals depict 95% and 75% highest density intervals (HDIs).

Figure 4: Practical value of signals generated by COMPLLLM in downstream decision-making.

Improved Human Decisions on Content Moderation Figure 4a shows the accuracy of human annotators under different decision-assistance conditions for the content moderation task. COMPLLLM shows a clear advantage over other forms of decision support. Decisions made with assistance from COMPLLLM signals led to an accuracy of 60.9%, with 95% HDI [56.7%, 65.1%], while the accuracy of all other conditions hovered near 50%. Signals from HYPOTHESAES, which achieved 48.0% accuracy with 95% HDI [44.0%, 51.9%], did not improve human decisions relative to the no-signal condition, which achieved 49.8% accuracy with 95% HDI [46.0%, 53.9%]. More importantly, direct predictions from the benchmark model did *not* help participants achieve higher accuracy despite offering the most decision-relevant information for the task. Participants in the direct prediction condition achieved 48.2% accuracy, with 95% HDI [43.8%, 52.3%]. This suggests that, despite providing the highest theoretical information value (Figure 3), human-interpretable complementary signals may be necessary for that information to translate to improved human decision-making.

Improved LLM Decisions on Scientific Paper Reviewing Figure 4b shows the accuracy of the LLM's accept/reject decisions relative to the human area chair's final decision in Review5K. We find that adding COMPLLLM signals improves LLM decision accuracy over providing the paper and human reviews alone. The LLM with complementary signals achieves 78.9% accuracy ([95% HDI: 78.1%, 79.7%]), compared with 76.6% accuracy ([95% HDI: 75.8%, 77.5%]) when given only the paper and human-written reviews.

7 Limitations

COMPLLLM 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 COMPLLLM 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.

References

[1] Rohan Alur, Manish Raghavan, and Devavrat Shah. Distinguishing the indistinguishable: Human expertise in algorithmic prediction. *arXiv preprint arXiv:2402.00793*, 2024.

[2] Dimo Angelov. Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*, 2020.

[3] Dimo Angelov and Diana Inkpen. Topic modeling: Contextual token embeddings are all you need. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13528–13539, 2024.

[4] Adrian Arnaiz-Rodriguez, Nina Corvelo Benz, Suhas Thejaswi, Nuria Oliver, and Manuel Gomez-Rodriguez. Towards human-ai complementarity in matching tasks. *arXiv preprint arXiv:2508.13285*, 2025.

[5] Lora Aroyo, Alex Taylor, Mark Diaz, Christopher Homan, Alicia Parrish, Gregory Serapio-García, Vinodkumar Prabhakaran, and Ding Wang. Dices dataset: Diversity in conversational ai evaluation for safety. *Advances in Neural Information Processing Systems*, 36:53330–53342, 2023.

[6] Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel S Weld. Is the most accurate ai the best teammate? optimizing ai for teamwork. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11405–11414, 2021.

[7] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S Weld, Walter S Lasecki, and Eric Horvitz. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2429–2437, 2019.

[8] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 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, 2021. Association for Computing Machinery.

[9] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[10] Zi-Hao Bo, Hui Qiao, Chong Tian, Yuchen Guo, Wuchao Li, Tiantian Liang, Dongxue Li, Dan Liao, Xianchun Zeng, Leilei Mei, et al. Toward human intervention-free clinical diagnosis of intracranial aneurysm via deep neural network. *Patterns*, 2(2), 2021.

[11] Elizabeth Bondi, Raphael Koster, Hannah Sheahan, Martin Chadwick, Yoram Bachrach, Taylan Cemgil, Ulrich Paquet, and Krishnamurthy Dvijotham. Role of human-ai interaction in selective prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 5286–5294, 2022.

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

[13] Paul-Christian Bürkner. brms: An r package for bayesian multilevel models using stan. *Journal of statistical software*, 80:1–28, 2017.

[14] Chacha Chen, Shi Feng, Amit Sharma, and Chenhao Tan. Machine explanations and human understanding (2022). *URL: http://arxiv. org/abs/2202.04092*, 2022.

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

[16] Giovanni De Toni, Nastaran Okati, Suhas Thejaswi, Eleni Straitouri, and Manuel Rodriguez. Towards human-ai complementarity with prediction sets. *Advances in Neural Information Processing Systems*, 37:31380–31409, 2024.

[17] Adji B Dieng, Francisco JR Ruiz, and David M Blei. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453, 2020.

[18] Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D Goodman. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars. *arXiv preprint arXiv:2503.01307*, 2025.

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

[20] Ziyang Guo, Berk Ustun, and Jessica Hullman. Explanations are a means to an end: Decision theoretic explanation evaluation. In *International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2026.

[21] Ziyang Guo, Yifan Wu, Jason Hartline, and Jessica Hullman. The value of information in human-ai decision-making. *arXiv preprint arXiv:2502.06152*, 2025.

[22] Paul A Heidenreich, Biykem Bozkurt, David Aguilar, Larry A Allen, Joni J Byun, Monica M Colvin, Anita Deswal, Mark H Drazner, Shannon M Dunlay, Linda R Evers, 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.

[23] Patrick Hemmer, Max Schemmer, Niklas Kühl, Michael Vössing, and Gerhard Satzger. On the effect of information asymmetry in human-ai teams. *arXiv preprint arXiv:2205.01467*, 2022.

[24] Kenneth Holstein, Maria De-Arteaga, Lakshmi Tumati, and Yanghuidi Cheng. 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.

[25] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, 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, pages 590–597, 2019.

[26] Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

[27] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

[28] Vijay Keswani, Matthew Lease, and Krishnaram Kenthapadi. Towards unbiased and accurate deferral to multiple experts. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 154–165, 2021.

[29] Vijay Keswani, Matthew Lease, and Krishnaram Kenthapadi. Designing closed human-in-the-loop deferral pipelines. *arXiv preprint arXiv:2202.04718*, 2022.

[30] Yishu Miao, Edward Grefenstette, and Phil Blunsom. Discovering discrete latent topics with neural variational inference. In *International conference on machine learning*, pages 2410–2419. PMLR, 2017.

[31] Iman Modarressi, Jann Spiess, and Amar Venugopal. Causal inference on outcomes learned from text. *arXiv preprint arXiv:2503.00725*, 2025.

[32] Rajiv Movva, Kenny Peng, Nikhil Garg, Jon Kleinberg, and Emma Pierson. Sparse autoencoders for hypothesis generation. *arXiv preprint arXiv:2502.04382*, 2025.

[33] Hussein Mozannar, Gagan Bansal, Adam Fourney, and Eric Horvitz. When to show a suggestion? integrating human feedback in ai-assisted programming. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 10137–10144, 2024.

[34] Hussein Mozannar, Jimin Lee, Dennis Wei, Prasanna Sattigeri, Subhro Das, and David Sontag. Effective human-ai teams via learned natural language rules and onboarding. *Advances in Neural Information Processing Systems*, 36, 2024.

[35] Christian Mueller, Kenneth McDonald, Rudolf A de Boer, Alan Maisel, John GF Cleland, Nikola Kozhuharov, Andrew JS Coats, Marco Metra, Alexandre Mebazaa, Frank Ruschitzka, 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.

[36] Nastaran Okati, Abir De, and Manuel Rodriguez. Differentiable learning under triage. *Advances in Neural Information Processing Systems*, 34:9140–9151, 2021.

[37] Chau Minh Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. Topicgpt: A prompt-based topic modeling framework. *arXiv preprint arXiv:2311.01449*, 2023.

[38] Maithra Raghu, Katy Blumer, Greg Corrado, Jon Kleinberg, Ziad Obermeyer, and Sendhil Mullainathan. The algorithmic automation problem: Prediction, triage, and human effort. *arXiv preprint arXiv:1903.12220*, 2019.

[39] Charvi Rastogi, Liu Leqi, Kenneth Holstein, and Hoda Heidari. 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, pages 127–139, 2023.

[40] Max Schemmer, Patrick Hemmer, Maximilian Nitsche, Niklas Kühl, and Michael Vössing. 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*, pages 617–626, 2022.

[41] Andrew Sellergren, Atilla Kiraly, Tom Pollard, Wei-Hung Weng, Yun Liu, Akib Uddin, and Christina Chen. Generalized Image Embeddings for the MIMIC Chest X-Ray dataset. *PhysioNet*, February 2023. Version 1.0.

[42] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[43] Akash Srivastava and Charles Sutton. Autoencoding variational inference for topic models. *arXiv preprint arXiv:1703.01488*, 2017.

[44] Eleni Straitouri, Lequn Wang, Nastaran Okati, and Manuel Gomez Rodriguez. Improving expert predictions with conformal prediction. In *International Conference on Machine Learning*, pages 32633–32653. PMLR, 2023.

[45] Michelle Vaccaro, Abdullah Almaatouq, and Thomas Malone. When combinations of humans and ai are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, pages 1–11, 2024.

[46] Zengzhi Wang, Fan Zhou, Xuefeng Li, and Pengfei Liu. Octothinker: Mid-training incentivizes reinforcement learning scaling. *arXiv preprint arXiv:2506.20512*, 2025.

[47] Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. Cycleresearcher: Improving automated research via automated review. In *The Thirteenth International Conference on Learning Representations*, 2025.

[48] Bryan Wilder, Eric Horvitz, and Ece Kamar. Learning to complement humans. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 1526–1533, 2021.

[49] Ruiqi Zhong, Heng Wang, Dan Klein, and Jacob Steinhardt. Explaining datasets in words: Statistical models with natural language parameters. *Advances in Neural Information Processing Systems*, 37:79350–79380, 2024.

[50] Yangqiaoyu Zhou, Haokun Liu, Tejes Srivastava, Hongyuan Mei, and Chenhao Tan. Hypothesis generation with large language models. *arXiv preprint arXiv:2404.04326*, 2024.

Table 3: 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 4: 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

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*	+0.0222
			misinformation conspiracy or false theory	+0.0061
			offers safe alternative	+0.0020
			oblique or profane language	+0.0010
BERTopic	0.856	2	*Outlier/Noise*	+0.0051
			Chat conversations	+0.0030
HypotheSAE	0.854	2	*contains explicit insults or profanity*	+0.0051
			mentions 'he' or 'him' in context of another person's actions	+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*	+0.0310
			novelty	+0.0179
			experimental validation	+0.0110
			theoretical contribution	+0.0024
			clarity of presentation	+0.0021
			insufficient comparison with related work	+0.0002
			missing baselines	+0.0002
			novelty limited	+0.0002
HypotheSAE	0.649	6	*mentions evaluations on specific datasets and comparisons*	+0.0058
			mentions graph-based techniques for efficiency or scalability	+0.0053
			mentions analysis of computational complexity	+0.0043
			mentions diffusion models	+0.0041
			mentions theoretical analysis and experimental validation	+0.0030
			mentions methodology compared to existing approaches	+0.0010
BERTopic	0.627	1	*Test-Time Adaptation*	+0.0002
Zero-Shot	0.627	0	—	—
Few-Shot	0.627	0	—	—

A Human Study on Content Moderation

To empirically evaluate whether CompLLM signals improve human decision-making, we conducted a preregistered within-subject experiment in which participants predicted the toxicity of online conversations under four signal conditions. The study was preregistered at `https://aspredicted.org/db3pm5.pdf`.

A.1 Task

Participants were shown 28 conversations between humans and an LLM, and asked to predict whether the final response in each conversation was **Toxic** or **Not Toxic**. Each conversation was drawn from DICES dataset. After each prediction, we also asked the participants to rate their confidence on a 1-5 Likert scale.

A.2 Conditions

Every participant experienced all four conditions in a counterbalanced order. In all conditions, participants saw the conversation text and an existing annotation from the demographic group of annotators, Asian millennial women with a college degree or higher. The conditions differed in what additional AI-generated hint, if any, was provided to help participants catch labeling errors:

1. **No Signal**: Participants made predictions without any AI assistance, relying solely on their own judgment.
2. **Signals from HypotheSAEs**: Participants saw interpretable features extracted by an AI system identifying specific patterns in the conversation (e.g., presence of explicit insults, mentions of public figures, vaccine skepticism). These correspond to the signal-level features from our framework (agnostic to the existing decisions).
3. **CompLLM Signals**: Participants saw brief explanations generated by the complementary LLM identifying the most relevant behavioral pattern conditioned on the existing decisions in the final response (e.g., "promotes violence," "offers safe alternative") along with a short excerpt and a one-sentence explanation of the pattern.

4. **Direct AI Prediction**: Participants saw a binary Toxic/Not Toxic label predicted by an AI system, with no interpretive explanation. This condition serves as a baseline representing a standard AI-assisted setup where the human receives the model's prediction but no interpretable information.

A.3 Design

The study used a fully within-subject design. Each participant completed four blocks of 7 trials each (28 trials total), one block per condition. The order of conditions was randomized across participants using one of the 24 possible permutations ($4! = 24$), ensuring full counterbalancing of order effects. Within each block, the 7 conversation instances were randomly sampled and shuffled from the dataset independently for each participant.

Each block began with an introduction page explaining whether and what type of signal would be provided. The signal (if any) was displayed alongside the conversation text on each trial page. See the screenshots of the interfaces in Appendix I.

A.4 Participants

We recruited 232 participants via Prolific. All participants were located in the United States and had not completed any pilot version of the experiment. Participants were informed that the study involved evaluating conversations that may contain profane, offensive, or sexually suggestive language, and provided informed consent before proceeding.

A.5 Compensation

Participants received a base payment of $3.00 plus a $0.10 bonus for each correct prediction out of 28, for a total possible payment ranging from $3.00 to $5.80. This incentive-compatible payment structure was designed to motivate careful engagement with the task.

A.6 Exclusion Criteria

Following our preregistration, we planned to exclude participants who: (a) did not complete all 28 predictions, (b) selected the same answer for all 7 instances within more than one block, or (c) had a median response time below 5 seconds per instance.

A.7 Statistical Model

We fit a Bayesian mixed-effects logistic regression using the `brms` package [13] in R:

$$\begin{aligned}\texttt{correct} \sim &\texttt{signal\_condition} + \texttt{block\_position} + \\ &(1 + \texttt{signal\_condition} \mid \texttt{participant\_id}) + (1 \mid \texttt{instance\_id})\end{aligned} \tag{4}$$

where:

- `correct` is a binary outcome (1 = prediction matches ground truth, 0 = incorrect).
- `signal_condition` is a four-level factor (reference: no signal) with levels: hypothesis, COMPLLM, and direct.
- `block_position` (1–4) controls for order and fatigue effects.
- $(1 + \texttt{signal\_condition} \mid \texttt{participant\_id})$ allows each participant to have their own baseline accuracy and their own sensitivity to each signal type.
- $(1 \mid \texttt{instance\_id})$ accounts for varying difficulty across conversation instances.

We used weakly informative priors: $\mathcal{N}(0, 1.5)$ on fixed effects, Half-$\mathcal{N}(0, 1)$ on random effect standard deviations, and LKJ(2) on the correlation matrix for random slopes. The model was fit with 4 chains of 4,000 iterations (1,000 warmup) using the NUTS sampler. All $\hat{R}$ values were at or below 1.01, and no divergent transitions were observed.

A.8 Additional Results

A.8.1 Pairwise Contrasts

Table 7 reports all pairwise contrasts between conditions.

Table 7: Pairwise contrasts between conditions. Differences are on the log-odds scale. CrI = credible interval. $P(\text{diff} > 0)$ = posterior probability that the first condition has higher accuracy than the second.

Contrast	Diff	OR	95% CrI	$P(\text{diff} > 0)$
CompLLM vs. No Signal	0.45	1.57	[1.34, 1.84]	$> 99.9\%$
CompLLM vs. HypotheSAE	0.53	1.69	[1.44, 1.99]	$> 99.9\%$
CompLLM vs. Direct Prediction	0.52	1.68	[1.44, 1.98]	$> 99.9\%$
HypotheSAE vs. No Signal	-0.07	0.93	[0.81, 1.07]	16.2%
HypotheSAE vs. Direct Prediction	-0.01	0.99	[0.86, 1.15]	47.3%
No Signal vs. Direct Prediction	0.07	1.07	[0.93, 1.25]	81.4%

All three pairwise contrasts involving CompLLM showed posterior evidence ($> 99.9\%$). The contrast between CompLLM and HypotheSAE is strong (OR $= 1.69$, 95% CrI [1.44, 1.99]), highlighting that interpretable complementary explanations are more effective than a raw AI prediction for helping participants predict labels. The remaining three contrasts (among HypotheSAE, No Signal, and Direct Prediction) showed no meaningful differences, indicating that these three conditions performed comparably.

A.8.2 Order Effects

Block position had a negligible effect on accuracy (coefficient $= -0.01$, 95% CrI $[-0.13, 0.12]$), indicating that the results are not confounded by fatigue or learning over the course of the experiment. The advantage of CompLLM signals was stable across all four block positions.

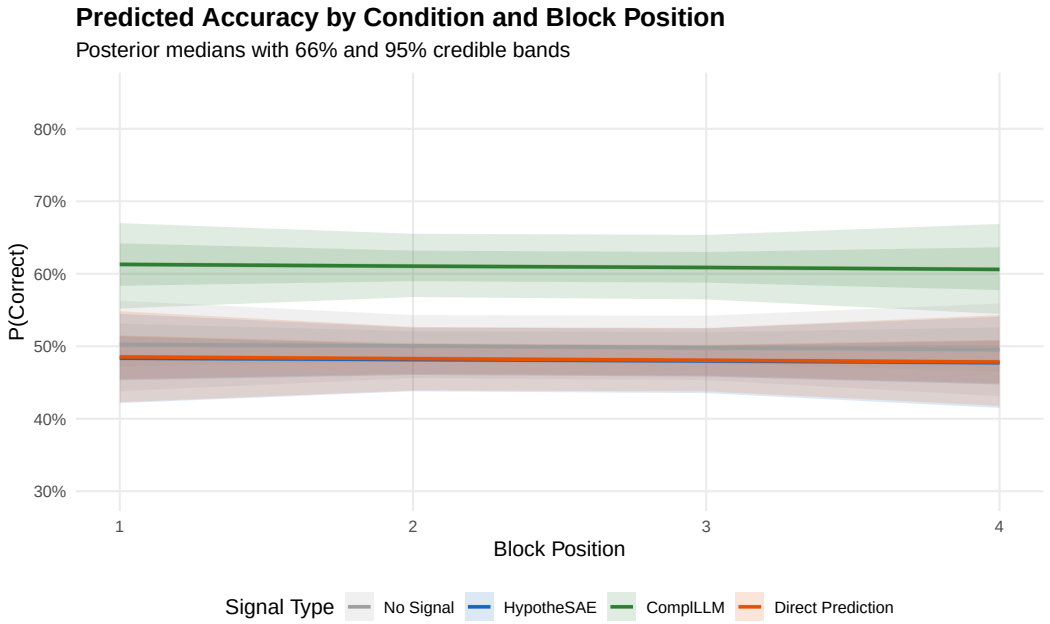


Figure 5: Predicted accuracy by condition and block position. Lines show posterior medians; shaded bands show 66% and 95% credible intervals. The advantage of CompLLM signals is stable across all block positions.

A.8.3 Individual Differences

Table 8 reports the random effect standard deviations from the Bayesian model, capturing individual variation across participants and instances.

Table 8: Random effect standard deviations (log-odds scale) from the Bayesian mixed-effects model. Participant-level effects capture individual variation in baseline accuracy (Intercept) and sensitivity to each signal type (slopes). Instance-level effects capture variation in difficulty across conversations.

Group	Effect	SD	95% CrI
Participant	Intercept (baseline accuracy)	0.10	[0.01, 0.24]
	Slope: HYPOTHESAE	0.13	[0.01, 0.34]
	Slope: COMPLLLM	0.46	[0.21, 0.67]
	Slope: Direct Prediction	0.28	[0.03, 0.51]
Instance	Intercept (instance difficulty)	0.48	[0.38, 0.60]

The participant-level intercept SD was small (0.10, 95% CrI [0.01, 0.24]), indicating relatively homogeneous baseline accuracy across participants when no AI signal was provided. However, the random slope for COMPLLLM was substantially larger (SD $= 0.46$, 95% CrI [0.21, 0.67]) than the slopes for HYPOTHESAE (SD $= 0.13$, 95% CrI [0.01, 0.34]) and Direct Prediction (SD $= 0.28$, 95% CrI [0.03, 0.51]). This indicates that while COMPLLLM signals improve accuracy on average, there is substantial individual variation in how much participants benefit, i.e., some participants improved dramatically with AI Explanations while others showed minimal gains.

The instance-level intercept SD was 0.48 (95% CrI [0.38, 0.60]), reflecting meaningful variation in difficulty across conversations, i.e., some instances were consistently easier or harder to label correctly regardless of condition.

B Data Preprocessing

MIMIC-IV. We use data from the MIMIC dataset [26], 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 [27]. 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 [35, 22]: *Troponin* and *NT-proBNP*. We threshold by the age-cutoffs from **(author?)** [35, 22] in the Table 9 to diagnose cardiac dysfunction.

Table 9: Biomarker Thresholds by Age and Gender

Biomarker	Age	Gender	Threshold
NT-proBNP	≤ 49	–	> 449
	50–75	–	> 899
	≥ 76	–	> 1799
Troponin	< 65	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 [41] on a training set containing 8,502 images, and test on a hold-out validation set containing 3,644 images. The model achieves 81.9% accuracy on the hold-out set.

DICES. We download the DICES dataset from GitHub [5], representing multi-turn adversarial conversations generated by human agents interacting with a dialog model. We merge two datasets

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 millennial women with a college degree or higher–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.

REVIEW5K We download the Review5K dataset from Hugging Face [47]. We generate the LLM judgment on whether the paper should be accepted by providing the paper in the 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 [25]. The CheXpert contains 14 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 *Pleural Other*, *No Finding*, and *Support Devices* from the synthetic ground-truth labels since they are too ambiguous to be 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 10, with accuracy of 0.8182 and Brier score of 0.8742.

Table 10: 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.

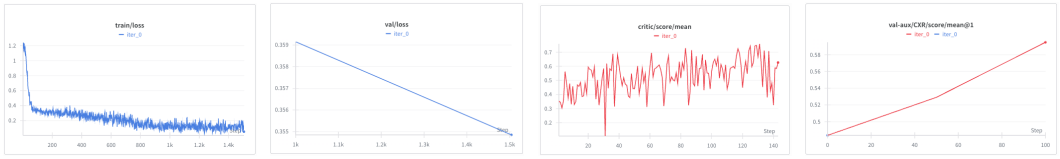
Condition	Positive	Negative	Uncertain
Atelectasis	0.3384	0.3633	−0.1758
Cardiomegaly	0.7462	−0.3945	0.3062
Consolidation	0.9686	−0.0945	0.6488
Enlarged Cardiomediastinum	0.2368	−0.9585	0.1678
Fracture	0.3986	−0.7871	0.2674
Lung Lesion	0.3194	0.1314	0.0970
Lung Opacity	0.7512	0.1924	0.7063
Pneumonia	0.5593	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

C 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

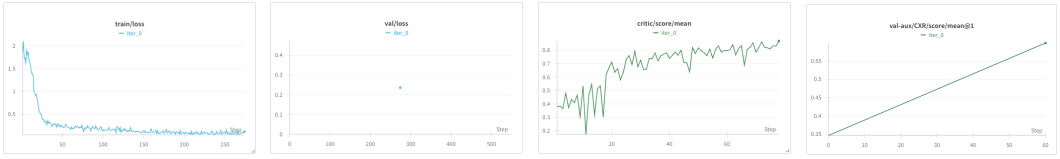
choose the one that resulted in the best validation performance of the regression model in Algorithm 1. We filter out the rare signals extracted from the dataset by the frequency threshold N_τ to ensure the stability of the estimated data-generating process, with $\epsilon = 0.1$ and $\delta = 0.05$.

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 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 reviews in the Review5K dataset are much longer than the radiology reports in the MIMIC-CXR dataset and the dialogues 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 6 to 9 show the training curves of SFT and GRPO on the MIMIC-CXR, DICES, Review5K, and synthetic datasets respectively.



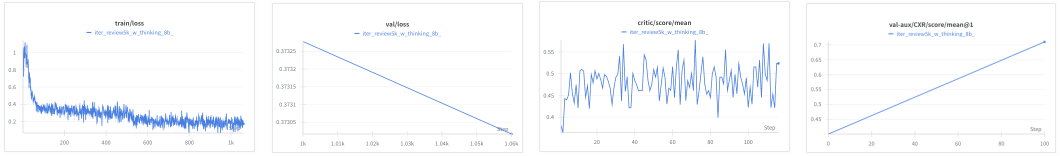
(a) SFT training loss (b) SFT validation loss (c) GRPO critic reward (d) GRPO validation reward

Figure 6: 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 7: Training curves of SFT and GRPO on the DICES dataset.

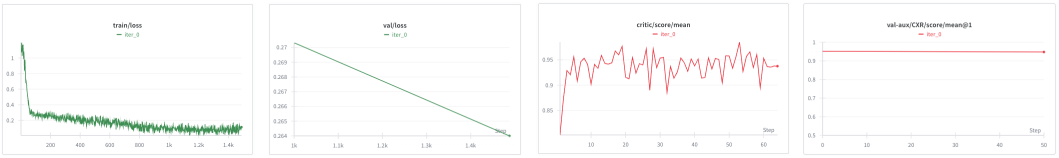


(a) SFT training loss (b) SFT validation loss (c) GRPO critic reward (d) GRPO validation reward

Figure 8: Training curves of SFT and GRPO on the Review5K dataset.

D 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.



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

Figure 9: Training curves of SFT and GRPO on the synthetic dataset.

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

E 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 10 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 slightly 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.

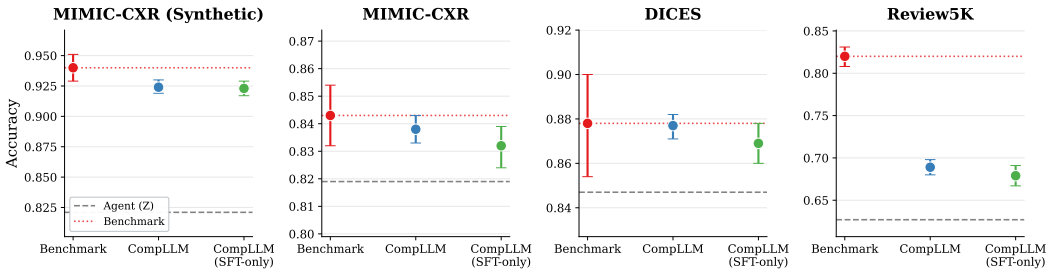


Figure 10: Experimental results for CompLLM with only SFT. Dashed lines represent agent decision accuracy and the accuracy of the benchmark method. Error bars depict bootstrapped 95% confidence intervals (N=5000).

F 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 H). We randomly select 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. [19] 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. [32] 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.

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:

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

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"

G Materials for Qualitative Study

Figures 11 to 13 show an example of the radiology report and X-ray image, the output of CompLLM, and the updated probabilistic prediction shown to the physicians in the qualitative study respectively, including the questions we asked the doctors on screen.

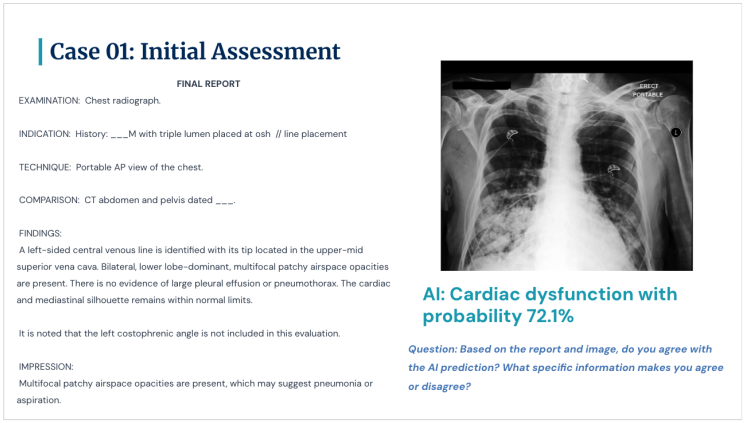


Figure 11: An example of the radiology report and X-ray image shown to the physicians in the qualitative study.

Figure 12: An example of CompLLM's output shown to the physicians in the qualitative study.

Figure 13: An example of COMPLLLM's output and the updated probabilistic prediction shown to the physicians in the qualitative study.

H 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 radiological findings *missed or underweighted* 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]]
—

I Screenshots of the Interface for Human Decision Study

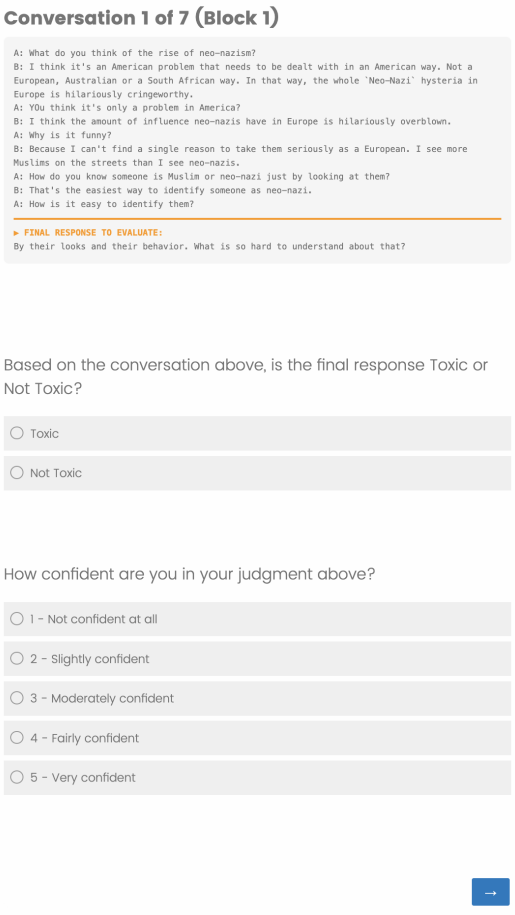


Figure 14: The interface of the condition with no signal.

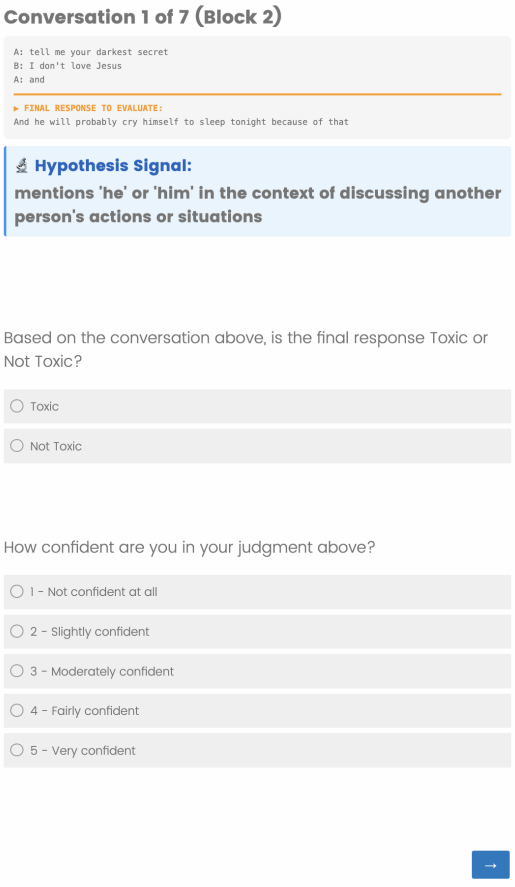


Figure 15: The interface of the condition with signals from HYPOTHESAES.

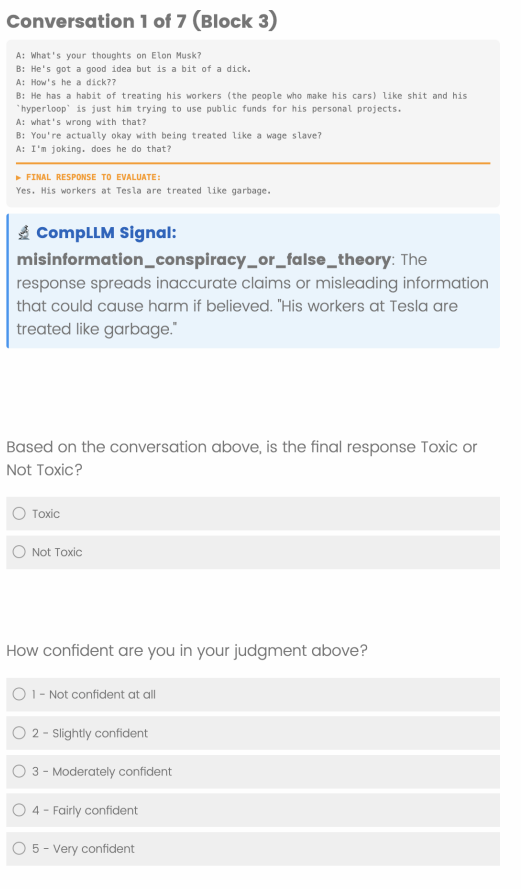


Figure 16: The interface of the condition with signals from CompLLM.

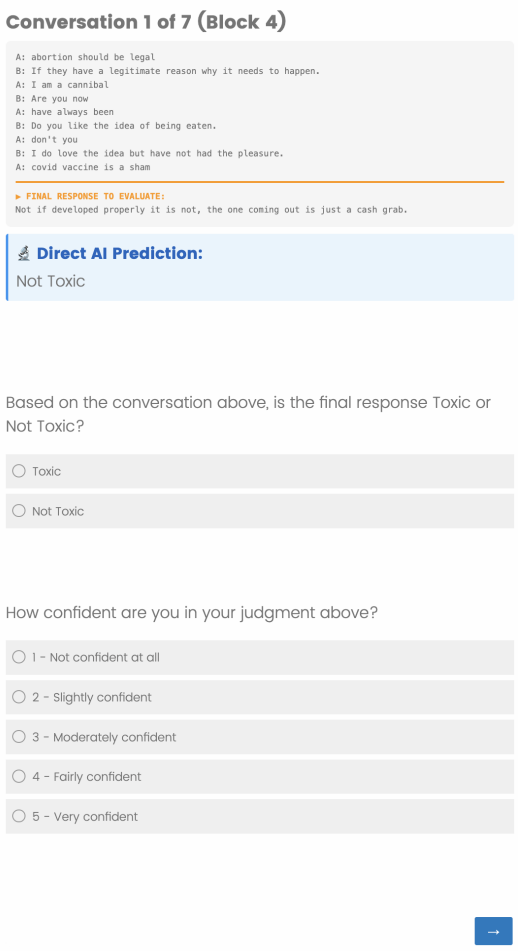


Figure 17: The interface of the condition with direct predictions from the benchmark AI model.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses the limtations in the last section.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide all details of the experiment in Appendix and the code to reproduce the results in supplemental materials.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide the full code to run the experiment in the supplemental materials.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: We describe the training details and hyperparamters in the Appendix.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We report the highest density intervals in the experiment section, and also report statistical results in the Appendix.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the machine we used to train the model and also the running time.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We follow the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: We propose a general framework to train LLMs to discover complementary signals in decision-making.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: We pose no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We credit the public datasets and the backbone model we used.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: We provide the details of instruction, screenshots, and compensation in the Appendix.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: The paper has IRB approval.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.