Adaptive Test Case Discovery for LLM-Assisted Decision Making in High-Stakes Domains

Anjali Parashar[*], **Carson Sobolewski**, **Yingke Li**, **Fei Chen**, **Chuchu Fan**
Laboratory of Information & Decision Systems (LIDS)
MIT
Cambridge, MA
{anjalip,csobo,yingkeli,feic,chuchu}@mit.edu

Abstract

Large language models (LLMs) are increasingly used for assistive decision-making in high-stakes domains, yet their outputs can be unreliable and difficult to audit. This necessitates pre-deployment testing, which can be used to identify scenarios where an LLM-assisted pipeline fails to produce satisfactory decisions. Existing testing strategies are often not sample-efficient and struggle to adapt to diverse testing objectives, application domains, and rapidly evolving models. We introduce a sample-efficient scenario design strategy for testing LLM-based decision-making pipelines. Our approach formulates testing as adaptive experimental design, using a flow-based surrogate model to sequentially acquire scenarios that balance optimization to discover challenging test cases with diverse exploration. We evaluate the method across several LLM-assisted decision-making tasks: bank loan approval, stock prediction from news articles, disaster resource allocation from tweets and images and tumor detection in histopathology images. Our approach consistently discovers challenging scenarios that satisfy the testing objective across all tasks, providing high coverage over scenario space. Results and code available at this URL.

1 Introduction

Large language models (LLMs) are being rapidly deployed as assistants in real-world decision-making processes across a range of domains [Azime et al., 2025, Damle and Faltings, 2025, Wang et al., 2025a, 2024, Yu et al., 2023, Otal et al., 2024, Liu et al., 2024]. LLM-assisted decision-making allows incorporation of diverse real world input, such as user portfolios, satellite images, and news articles, making these systems flexible and broadly applicable [Yang et al., 2023, Rane, 2023, Eigner and Händler, 2024]. However, their black-box nature leads to dissatisfactory performance that is difficult to diagnose during deployment [Liu et al., 2024]. Therefore, rigorous pre-deployment testing across diverse scenarios is essential.

Sufficient assessment requires performance evaluation over a diverse range of test cases or scenarios. Conventional evaluation techniques rely on static datasets of handcrafted prompts or LLM-generated test cases [Jiao et al., 2025, Xie et al., 2024], but these approaches have key limitations concerning scalability. As models evolve rapidly, handcrafted static benchmarks require continual maintenance [Berrada et al.]. Moreover, complex decision-making tasks may make exhaustive evaluation over thousands of scenarios impractical due to query cost and decision time [Li et al., 2023, Ko et al., 2025]. These benchmarks are useful for average performance measurement of the models [Xie et al., 2024], however, they cannot stress test a model for a given testing objective in a sample efficient manner [Li et al., 2023].

[*]Corresponding author

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

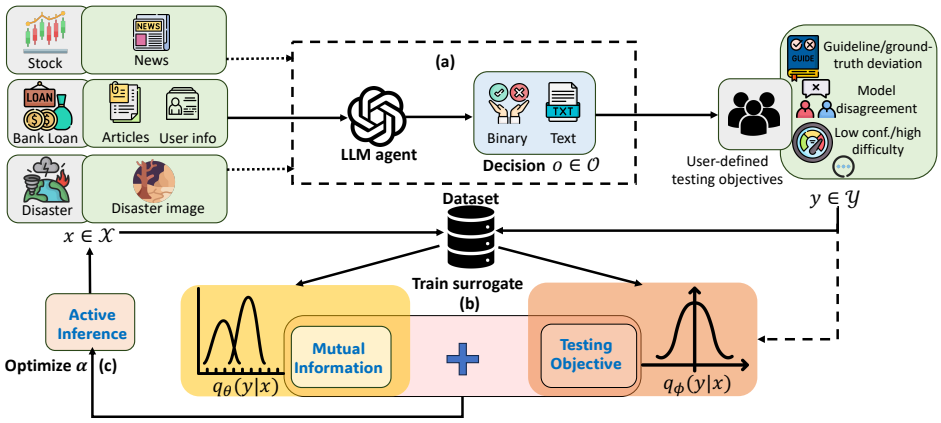


Figure 1: **Adaptive Experiment Design for generating diverse and high quality candidates.** Our pipeline has three main components, (a) data collection using sequential LLM queries, (b) training surrogate flow models q_θ, q_ϕ to estimate mutual information and learn the testing objective for data acquisition objective α (Equation (1)), (c) adaptive scenario generation for finding diverse and high utility candidates by optimizing α.

Prior works have noted that the amount of data needed for sufficient evaluation is significantly smaller than the size of conventional static benchmarks Polo et al. [2024]. This motivates us to construct a concise, and targeted pool of *diverse* and *high quality* scenarios that test LLM-assisted workflows in a principled manner for a user-defined objective. In our approach, we leverage adaptive experimental design to sequentially design a suite of scenarios that are challenging for the LLM for decision-making, where the definition of challenging scenarios can be user defined, in a sample-efficient manner Rainforth et al. [2024].

Recent works have proposed a combination of exploration and optimization for scenario design Li et al. [2026], Meera and Kouw [2026], Parashar et al., with exploration quantified through mutual information and motivated by active inference [Friston et al., 2017]. Leveraging this idea, we formulate a scenario design strategy that explicitly balances scenario diversity and quality, where diversity is quantified using mutual information, and quality is measured as alignment of a scenario to a given testing objective. For example, for a testing objective of measuring model performance with respect to a guideline or ground truth, scenarios that deviate from the guideline represent high quality candidates, and we aim to discover a diverse pool of such scenarios within a limited sampling budget. This yields a task and model agnostic framework for sample-efficient testing of LLM-assisted workflows under user-defined objectives. To ensure scalability to complex tasks involving real-world inputs, we further introduce a flow-based probabilistic generative surrogate model to learn which regions in scenario space align with the testing objective, and assist the scenario generation process in a scalable manner [Lipman et al., 2022].

Figure 1 provides an overview of our approach. We perform experimental validation across several assistive decision-making tasks using LLMs, [Azime et al., 2025, Wang et al., 2024, 2025a]. These tasks represent different scenario types (categorical, images, text), and forms of data acquisition (continuous scenario representation, dataset sampling). We show that our underlying data acquisition strategy, combined with flow based surrogate, discovers informative, diverse candidate scenarios and scales efficiently to high dimensional scenario spaces. We summarize our main contributions as:

- We propose an adaptive experimental design methodology that frames testing LLM-assisted decision-making workflows as the generation of *diverse* and *high-quality* scenarios.
- To accommodate a generalized representation of high dimensional scenarios commonly used as inputs to LLMs, we propose a hybrid flow-based surrogate model for scalable estimation of our dual objective and data acquisition, along with efficient training.
- We evaluate our approach on several LLM-assisted decision-making applications: bank loan approval, news-based stock prediction, and disaster resource allocation from images in the main paper, using cross-model disagreement and deviation from ground truth as testing objectives. We

also report results for an additional task and objective of disaster resource allocation using tweets in the Appendix.

Our strategy produces a targeted test suite of real-world scenarios where further intervention may be needed. We evaluate against sampling-based and LLM-based test-case generation baselines [Daulton et al., 2021, Lee et al., 2017, Raman et al., 2025]. Across tasks, our method outperforms all baselines, discovering up to twice as many challenging scenarios as canonical adaptive testing methods [Lee et al., 2017]. We also introduce a pairwise-distance coverage metric to measure diversity, under which our approach consistently achieves the highest coverage. Finally, we use interpretable scenario spaces to identify dominant failure modes where contrasting positive and negative decision factors induce high cross-model inconsistency [Wu et al., 2025, Grayeli et al., 2024].

2 Problem Formulation & Preliminaries

We consider a decision-making task in which a scenario $x \in \mathcal{X}$ is provided to an LLM agent via a prompt $\mathcal{P}_x = (\mathcal{C}, x)$, combining a task description $\mathcal{C}$ with scenario information x. The agent consists of a model m and prompt $\mathcal{P}_x$, and generates an output $o \in \mathcal{O}$ (Figure 1 (a)). For instance, in bank loan approval, x may encode an applicant's credit history, education, and financial background, while the output $o \in \mathcal{O}$ is an accept/deny decision [Azime et al., 2025]. Scenarios are obtained from a simulator $\mathcal{S}$ or a fixed dataset $\mathcal{D}_X$, inducing $x \sim p(x)$. To evaluate the LLM's behavior, we define a user-specified testing objective $h : \mathcal{X} \times \mathcal{O} \to \mathbb{R}$ that assigns a scalar score to each scenario-output pair, with larger values indicating stronger satisfaction of the objective. We discuss concrete choices of h in Section 2.1.

Sample-efficient scenario design. Given a testing objective h and budget K, we seek a set of *high-quality* and *diverse* scenarios $\mathcal{X}^* = \{x_i\}_{i=1}^K$. We formulate this as adaptive experimental design. At iteration k, given observed scenarios $\mathcal{X}_k = \{x_i\}_{i=1}^k$, we select the next query $x_{k+1} \in \mathcal{X}$ to maximize both testing objective and diversity. This process continues until the budget is exhausted. We define quality through a user-specified testing objective and diversity through mutual information, as described next.

2.1 Measuring quality using testing objective

For a given output o, we define a testing objective $h : (x, o) \to \mathbb{R}$ such that $x^* = \arg\max_{x \in \mathcal{X}} h(x, o)$ corresponds to the scenarios we aim to discover. The form of h depends on the evaluation goal. For instance, to identify scenarios that are difficult for decision-making, h can encode different analytical notions of difficulty. The testing objective is task and user-specific, therefore, we do not prescribe a single, 'correct' notion of a testing objective. Instead, we aim to accommodate different forms of testing objectives within our scenario generation paradigm. We instantiate this idea with three canonical testing objectives used in prior work.

Deviation from ground truth. When ground truth or domain guidelines are available, h can measure deviation from the reference output o_x^* for a given x, i.e., $h(x, o) := \|o_x - o_x^*\|_2^2$, which we use as the first example of testing objective. In the absence of ground truth, as is common in LLM-assisted decision-making, uncertainty is often used as a proxy for difficulty [Swayamdipta et al., 2020, Berrada et al., 2025].

Self-assessment. Recently, verbalized self-assessment has become a common approach for estimating LLM uncertainty and difficulty [Barkan et al., Wang et al., 2025b, Lin et al.]. For example, Yang et al. use verbalized confidence scores to identify low-confidence decisions, which is particularly useful for closed-source models [Wang et al., 2025b]. Natural Language Inference (NLI)-based confidence metrics have also been used to assess whether an LLM's reasoning entails its stated decision [Tarunesh et al., 2021, Swayamdipta et al., 2020]. We combine self-reported difficulty with NLI-based confidence to define our second testing objective.

Cross-model disagreement. Cross-model disagreement provides another useful signal of difficulty, which is not tied to a specific model, but can be relevant for scenario interpretation. Hamidieh et al. [2026] show that LLMs can be self-consistent while still exhibiting substantial disagreement with

other models, and Yang and Wang [2026] find that models with similar benchmark accuracy can disagree on 16–66% of items. We therefore use cross-model disagreement as an additional testing objective for identifying difficult test cases and use it as a primary testing objective in the experimental validation.

In Section 5, we demonstrate adaptive scenario design for several choices of h. Our formulation is open-ended, h need not be analytically tractable and is treated as a black-box objective learned alongside scenario design. This makes our framework compatible with tools such as pairwise preference elicitation [Parashar et al.] and symbolic regression for modeling objectives from user preferences and interpretable signals [Zhao et al., 2025]. We discuss these extensions and related works in Appendix A.

2.2 Adaptive experiment design for sample-efficient evaluation

Consider a distribution $p(y|x)$, where y represents the output for a given testing objective h, scenario x and LLM output o. Due to the fundamentally black-box nature of h as well as the LLM, we approximate the distribution $p(y|x)$ as $q_\theta(x)$, where θ denotes the model parameters of the learned model. We treat $q_\theta(\cdot)$ as a surrogate model to be learned, as well as used for suggesting scenarios for evaluation, by maximizing a testing objective in a sample-efficient manner.

Specifically, we use a dataset $\mathcal{D}_k = \{x, y\}_{i=1}^k$ of $k < K$ samples, to learn a surrogate model $q_\theta : \mathcal{X} \to [0, 1]$ Rainforth et al. [2024]. We subsequently suggest a potential scenario for evaluation, $x_{t+1} = x^*$ using an *acquisition function* $\alpha : \mathcal{X} \to \mathbb{R}$, such that $x^* = \arg\max_{x \in \mathcal{X}} \alpha(x)$, using the surrogate model q_θ. The training and scenario acquisition is conducted sequentially (Figure 1(b) and (c)).

Our goal is to sequentially acquire scenarios that balance exploration with optimality under a given testing objective. Recent work proposes Active Inference as an acquisition objective α for joint learning and optimization in adaptive experimental design [Li et al., 2026, Friston et al., 2017]. This objective combines the surrogate-based testing objective q_θ with *mutual information* (MI) over model parameters. MI is widely used in sample-efficient learning and optimization to quantify epistemic uncertainty [Parashar et al., Cao and Tsang, 2021, Rainforth et al., 2024, Chaloner and Verdinelli, 1995]. We combine these learning and optimization goals in the following acquisition function:

$$\alpha(x) = I(\theta, y|x, \mathcal{D}_k) + \mathbb{E}_{q_\theta(y|x,\mathcal{D}_k)}[y]. \tag{1}$$

Here, I denotes *Mutual Information* (MI) between model parameters and outcome y conditioned on x, and encourages exploration of scenarios where uncertainty in learned model parameters is high, thereby promoting diversity in scenario acquisition. Next we propose a scalable surrogate model design based on epistemic uncertainty quantification in flow based probabilistic models to represent complex scenario spaces used as inputs to LLMs in high-stakes decision making tasks [Berry and Meger, 2023, Berry and Meger].

3 Adaptive Experiment Design using Hybrid Flow Surrogate

Equation (1) is optimized sequentially to generate candidate scenarios. For sample efficiency, this optimization is typically performed over a surrogate model q_θ, often chosen as a Gaussian Process (GP) [Schulz et al., 2018]. However, GPs face scalability limitations, especially in high-dimensional or non-continuous input spaces [Zhang et al., 2025]. Their uncertainty estimates are also tightly coupled to the kernel choice, imposing restrictive smoothness assumptions on the scenario space.

These limitations are especially relevant for LLM-assisted decision-making, where inputs may include high-dimensional images [Wang et al., 2025a] or discrete and mixed attributes such as user biographical data [Azime et al., 2025]. As a result, standard GP surrogates are often ill-suited to the settings we consider. Berrada et al. address this challenge by using in-context prompted LLMs as surrogate models. However, because this surrogate is not updated online, it may not fully adapt to the user-defined testing objective during acquisition. Moreover, as shown in Section 4, LLM-based surrogates can introduce representational biases in the scenario space.

In this work, we explore flow-based surrogate models. Continuous Normalizing Flows (CNFs) trained with Flow Matching (FM) have shown strong empirical performance in generative modeling [Lipman

et al., 2022], and their flexibility makes them well suited for modeling diverse LLM-compatible input spaces. While we instantiate our approach with FM-based architectures, the scenario design strategy uses general generative modeling principles and can be extended to other generative surrogates.

3.1 Training surrogate model using flow matching

We aim to learn a surrogate q for the distribution $p_1(y|x) := p(y|x)$. Following Lipman et al. [2022], let p_t be a probability path such that $p_0 = p$ is a simple distribution, chosen as standard normal distribution, $p_0 = \mathcal{N}(y|0, I)$ in this work. The stochastic path between p_0 to p_t is defined using $\frac{dy}{dt} = y - y_0$. This gives the oracle velocity $v^*(y_t, t, x) = y - y_0$. Flow matching pertains to learning the oracle velocity and a parametric model defined using a Deep Neural Network (DNN) as $v_\theta : \mathcal{X} \times \mathcal{Y} \times [0, 1] \to \mathcal{Y}$, by minimizing the loss $\mathcal{L}_{\text{FM}}(\theta)$ using dataset $\mathcal{D}_k$ as:

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{t\sim\mathcal{U}[0,1],(x,y)\in\mathcal{D}_k}[\|v_\theta(y_t, x, t) - (y - y_0)\|_2^2]. \tag{2}$$

3.2 CNF mixture model for MI estimation

A key challenge in FM-based generative modeling is scalable mutual information (MI) estimation, since MI typically requires entropy estimation. Although normalizing flows provide tractable likelihoods and can be used to estimate entropy [Berry and Meger, 2023], such estimates often rely on Monte Carlo sampling, which can scale poorly in high dimensions. Recently, Berry and Meger proposed a mixture-based MI estimator for normalizing flows that avoids direct entropy estimation by using pairwise KL divergences between transformed latent distributions p_0^i, p_0^j for mixture components θ_i, θ_j, $i, j \in [1, \ldots, M]$. We adopt this approach to approximate MI for CNFs trained with flow matching. Specifically, we define $q_\theta(y; x)$ as a mixture over M ensemble components with a shared bijection and component-specific base distributions p_0^i as $q_\theta(y; x) = \sum_{i=1}^M \pi_i q_{\theta_i}(y; x)$, where $(\pi_i)_{i=1}^M$ are ensemble weights satisfying $\sum_{i=1}^M \pi_i = 1$.

Berry and Meger showed that for such flow ensembles, estimating epistemic uncertainty in the output distribution is equivalent to estimating it in the base distribution due to shared bijection. Therefore, epistemic uncertainty in $q_{\theta_i}(y; x), q_{\theta_j}(y; x)$ for $i, j \in [1, \ldots, M]$, reduces to uncertainty in base distributions p_0^i, p_0^j, which can be measured using Pairwise Distance Estimators (PaiDEs). Using this property, we can construct an estimator of MI, $\hat{I}_{KL}$ using PaiDEs as:

$$\hat{I}_{KL}(\theta, y) = -\sum_{i=1}^M \pi_i \log \sum_{j=1}^M \pi_j \exp\big(-D_{KL}(p_0^i||p_0^j)\big), \tag{3}$$

where $D_{KL}(p_0^i||p_0^j)$ denotes KL divergence and has a closed form expression for Gaussian distributions discussed in Appendix G.1. We choose $\pi_i = 1/M$ for uniform weights.

While mixture models are effective for MI estimation, it poses practical challenges in estimation of testing objective h for optimization due to high model bias between various mixture components [Berry and Meger]. This leads to a loss in accuracy of prediction, rendering it unsuitable for constructing our second term in Equation (1).

As a practical solution, we propose a hybrid surrogate approach, using an ensemble-based mixture model for estimating epistemic uncertainty trained using bootstrapping with parameters $\theta = (\theta_i)_{i=1}^M$, and a nominal surrogate model parameterized by ϕ, trained using a nominal FM approach for accurate predictions. This results in an acquisition function estimated using a hybrid surrogate architecture as:

$$\alpha(x) = \hat{I}(\theta, y|x, \mathcal{D}_k) + \mathbb{E}_{q_\phi(y|x,\mathcal{D}_k)}[y]. \tag{4}$$

Algorithm 1 outlines our complete sequential scenario design process with α in Equation (1) estimated using hybrid surrogate (Equation (4)). We now demonstrate applications of our approach, with hybrid surrogate model combined with acquisition function α towards efficient scenario generation in LLM decision-making paradigms.

4 Experimental Validation

Our central hypothesis is that our approach generates high quality dataset consisting of diverse candidates that efficiently test the LLM's decision-making capabilities within limited number of

Algorithm 1 Adaptive Experimental Design for Probing LLMs in Decision-Making Tasks

Input: Initial dataset $\mathcal{D}_0 = \{(x_i, y_i)\}_{i=1}^{K_0}$, total budget $B = K + K_0$
Train surrogate models q_ϕ and q_θ on $\mathcal{D}_0$ using the loss in Equation (2)
for $k = 0$ to $K - 1$ **do**
  Select query point $x_{k+1} \leftarrow \arg\max_{x \in \mathcal{X}} \left(\hat{I}(\theta, y \mid x, \mathcal{D}_k) + \mathbb{E}_{q_\phi(y|x,\mathcal{D}_k)}[y] \right)$
  Query LLM, observe decision output o_{k+1} and compute evaluation metrics y_{k+1}
  Update dataset $\mathcal{D}_{k+1} \leftarrow \mathcal{D}_k \cup \{(x_{k+1}, y_{k+1})\}$
  Update models q_θ and q_ϕ using $\mathcal{D}_{k+1}$ and the loss in Equation (2)
end for
Return $\mathcal{D}_K$

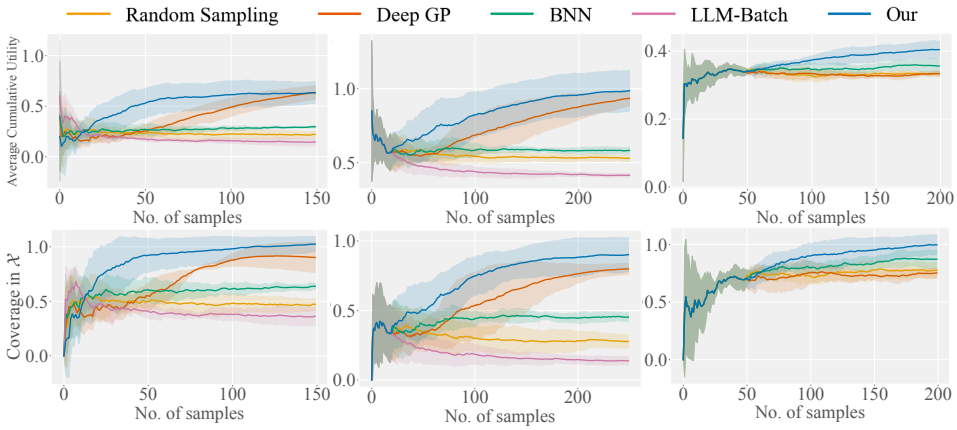


Figure 2: **Baseline comparison for h_3: Average Cumulative Utility (top) and Coverage (bottom).** Left to right: Bank Loan, News, Disaster Image (mean $\pm$ std. computed across five seeds).

samples. In the main paper, we show validations of this hypothesis across four tasks involving LLM-assisted decision-making, reported as Bank Loan, News, Disaster Image and Camelyon17. In Appendix, we also show results for another task, Disaster Tweet.

Tasks The tasks cover diverse scenario representations for real-world LLM-assisted decision making. Complete implementation details are provided in Appendix C.

Bank Loan uses an LLM to *accept* or *deny* a loan application from a user profile, represented as a 10-dimensional mix of biographic, education, and financial attributes (Table 10), with binary output $o \in \{0, 1\}$ [Azime et al., 2025].

News task consists of an LLM based stock prediction of next-month stock returns from prior returns and financial news [Wang et al., 2024]. Here, scenarios are defined by six categorical markers, prior trend, sentiment, news topic, prediction horizon, evidence strength, and conflicting trends, plus prior percentage return. Together, these are used to generate synthetic news summaries for stock prediction test cases.

Disaster Image task uses before and after satellite images from Disaster M3 dataset [Wang et al., 2025a] to select one of eight immediate recovery actions. Here, images are represented by 384-dimensional DINOv2 embeddings [Oquab et al., 2023] with a ViT-Small backbone [Touvron et al., 2021]. We include an additional Disaster Tweet task in the appendix, where an LLM recommends immediate recovery actions from tweets.

Camelyon17 is a medical domain task providing direct grounded validation [Koh et al., 2021], where the LLM agent detects tumor from histopathology images, for which ground truth labels are available.

Testing objectives. To validate generalizability of our approach for different testing objectives, we test the tasks with three testing objectives, (1) deviation from ground truth (h_1 Bank Loan, Camelyon17), (2) combination of self assessment and NLI entailment for measuring confidence of LLM generated decision, h_2 (Bank Loan, News, Disaster Tweet), and (3) disagreement between

Table 1: **Method ablations.** Mean ± standard deviation. Bold indicates the best result within each setting and metric.

Setting	Method	Avg. cum. utility	Coverage	Vendi	Pool clusters
Bank Loan (h_1)	Flow (ours)	0.829 ± 0.023	$\mathbf{1.00 \pm 0.032}$	2.80 ± 0.17	17.8 ± 1.5
	No-MI	$\mathbf{0.899 \pm 0.026}$	0.914 ± 0.144	2.07 ± 0.34	8.0 ± 4.6
	Only-MI	0.362 ± 0.019	0.883 ± 0.043	$\mathbf{3.45 \pm 0.14}$	$\mathbf{20.2 \pm 1.3}$
	No-ens-MI	0.460 ± 0.041	0.887 ± 0.047	2.48 ± 0.31	15.6 ± 2.2
Bank Loan (h_3)	Flow (ours)	0.541 ± 0.046	$\mathbf{1.00 \pm 0.029}$	2.29 ± 0.42	11.8 ± 1.8
	No-MI	$\mathbf{0.624 \pm 0.101}$	0.929 ± 0.138	1.77 ± 0.38	7.4 ± 5.1
	Only-MI	0.230 ± 0.045	0.622 ± 0.101	$\mathbf{3.23 \pm 0.24}$	$\mathbf{14.0 \pm 1.4}$
	No-ens-MI	0.279 ± 0.058	0.639 ± 0.110	2.56 ± 0.29	13.4 ± 0.9
News (h_3)	Flow (ours)	0.987 ± 0.138	$\mathbf{0.902 \pm 0.121}$	1.87 ± 0.24	9.0 ± 2.4
	No-MI	$\mathbf{1.151 \pm 0.050}$	0.733 ± 0.094	1.23 ± 0.22	3.6 ± 2.3
	Only-MI	0.791 ± 0.020	0.821 ± 0.041	$\mathbf{2.46 \pm 0.10}$	$\mathbf{12.6 \pm 1.3}$
	No-ens-MI	0.669 ± 0.044	0.590 ± 0.072	2.26 ± 0.26	9.4 ± 2.3
Disaster Image (h_3)	Flow (ours)	0.426 ± 0.029	1.00 ± 0.077	1.15 ± 0.02	$\mathbf{11.6 \pm 2.2}$
	No-MI	$\mathbf{0.480 \pm 0.014}$	$\mathbf{1.00 \pm 0.008}$	1.10 ± 0.02	8.2 ± 1.5
	Only-MI	0.311 ± 0.017	0.704 ± 0.061	$\mathbf{1.18 \pm 0.01}$	11.0 ± 1.7
	No-ens-MI	0.408 ± 0.027	0.953 ± 0.071	1.14 ± 0.02	11.0 ± 0.6
Camelyon17 ($h_1^\dagger$)	Flow (ours)	0.650 ± 0.080	$\mathbf{1.00 \pm 0.044}$	2.40 ± 0.35	$\mathbf{6.8 \pm 1.5}$
	No-MI	0.690 ± 0.093	0.956 ± 0.029	2.14 ± 0.26	5.2 ± 1.2
	Only-MI	0.220 ± 0.064	0.562 ± 0.147	$\mathbf{4.18 \pm 1.20}$	5.4 ± 1.9
	No-ens-MI	$\mathbf{0.697 \pm 0.056}$	0.997 ± 0.076	2.39 ± 0.47	6.6 ± 1.5

various models h_3 (Bank Loan, News, Disaster Image). Design of testing objectives is discussed in Appendix D.

Metrics. We evaluate methods along quality and diversity of scenarios evaluated using a variety of metrics: *(1) Avg. cumulative utility*: $c(\boldsymbol{y}_k) = \frac{1}{k}\sum_{i=1}^{k} y_i$ measures acquisition quality with respect to evaluation objective, for $\boldsymbol{y}_k = [y_1, \ldots, y_k]$ observed over $k \leq K$ samples. *(2) Coverage*: diversity of high utility region in scenario space using pairwise distance, defined as $g(\mathcal{X}_u^k) = \lambda \sum_{j=1}^{k}\sum_{i=1}^{k} \|x_i - x_j\|_2^2$ for $x_i \in \mathcal{X}_u^k$, where $X_u = \{x_i | (x_i, y_i) \in \mathcal{D}_k, h(y_i) \geq h_{\text{th}}\}$. Here λ is used to normalize g within $[0, 1]$ for consistent reporting, and the value of $h_{\text{th}} = 0.5 h_{\max}$, where $h_{\max}$ is the maximum value of objective observed in the experiment for each task. *(3) Positive samples (PS-τ)*: fraction of acquisitions with $h \geq h_{\text{th}}$, with $h_{\text{th}} = k h_{\max}$, adopted from [Malkomes et al., 2021]. *(4) Vendi score*: $V = \exp\big(H(\text{eig}(K_{\cos}))\big)$ where $K_{\cos}$ is the cosine-similarity kernel matrix of acquired failure embeddings [Friedman and Dieng, 2022] and measures effective number of distinct failure modes. *(5) Pool clusters*: number of distinct KMeans clusters (out of $K_c = 25$ fit on the full pool) covered by acquired failures. *(6) Hit-rate* (only for Camelyon17 h_1): fraction of acquisitions that are true failures ($Y = 1$). Experimental details and values of constants such as λ, K are discussed in Appendix E and Table 11.

Baselines. We compare our approach with the following baselines. Random Sampling samples data uniformly in the scenario space $\mathcal{X}$. We test alternate surrogate model choices using GPs with Deep Kernels [Lee et al., 2017], and Bayesian Neural Network (BNN), implemented as Deep-GP and BNN respectively. For Bank Loan, News task, we compare LLM-Batch that uses LLM to generate test cases by providing in-context learning examples Raman et al. [2025] (more results in Appendix B.9). Results are reported for GPT-5.2 Achiam et al. [2023], OpenAI [2026] for h_1, h_2 and different variants of GPT models for h_3 unless stated otherwise. Figure 2 and Table 3 shows baseline comparison for h_3, and Camelyon17 (h_1) respectively. Figure 5 show baseline comparison for h_2. In Figure 5, we also compare against a multiobjective optimization baseline q-NEHVI Daulton et al. [2021] for h_2, that optimizes testing objectives defined using multiple sub-objectives. A discussion on computational cost of all approaches is in Appendix F.2.

Ablations. We ablate three components of our acquisition function across *all tasks*, isolating the contribution of each: **No-MI** ($\beta = 0$): removes the exploration term, reducing to pure utility maximization; **Only-MI** (utility term omitted): pure epistemic-uncertainty-driven exploration; **No-ens-MI**: replaces the ensemble PaiDE estimator with single-model MC-dropout MI (BNN-style). Table 1 presents results uniformly across tasks and objectives. We perform additional ablations on Bank Loan task with testing objective h_2 reported in Appendix: (1) temperature ablations (Table 9), (2) LLM model ablation with Llama-3.1-8B-Instruct Touvron et al. [2023], Meta AI [2024] (Appendix B.8), (3) surrogate model ablations using nominal FM surrogate to test ensemble based MI estimation (Appendix B.5, Figure 6), (4) data acquisition ablations with and without MI (Table 7,Appendix B.6).

5 Results & Discussion

We present key results from experiments on Bank Loan (h_1, h_3) News (h_3), Disaster Image (h_3), and Camelyon17 (h_1). For each task, we apply a continuous, normalizing transformation to propose normalized scenario space for data acquisition and training the surrogate models (see Appendix C for details). For optimizing Equation (1), we use BoTorch library for performing sampling based acquisition for Bank Loan, News, since optimization in a continuous space is feasible. For Disaster Image, we select top candidate from the dataset. For each task, we recover dominant modes shown in Tables 2, 5 and 6, which, in addition to high scenario coverage, confirm that our method performs diverse exploration of search space in limited budget.

5.1 Bank Loan

Figure 8 shows the distribution of key scenario drivers, with high loan amount and positive credit history consistent across objectives. We present key insights drawn from h_1 and h_3 below.

Analysis for h_1. The objective h_1 uses Debt-To-Income (DTI) and credit history as loan qualifiers, with DTI below 0.36 and positive credit history. The scenarios collected by our approach have DTI 0.11 ± 0.62 and credit history 0.96 ± 0.19 (where 1 denotes positive credit history). Notably, 83% of discovered scenarios $(415/500)$ correspond to cases approved by the reference criterion but rejected by the LLM, indicating that our method converges to scenarios that optimize the testing objective. The loan amount distribution is unimodal and skewed high (628.4 ± 169.9K), while income is multimodal (Figure 8, Appendix H.1). For low-income applicants, rejections are often justified by a high loan-amount-to-income ratio, consistent with the low effective DTI cutoff used by GPT-5.2.

Analysis for h_3. We use GPT-5.2, GPT-5.4-mini, GPT-5.5 as candidate models for evaluating disagreement. On scenarios $x^* = \{x | h(x, o) = 1\}$, approval rate for the three LLMs is observed to be $21.2\%, 54.2\%, 49\%$ respectively. The dominant disagreement pattern observed over 34.7% disagreeing scenarios is REJECT, APPROVE, REJECT, showing that GPT-5.4-mini is the most lenient in loan approval (dominant modes shown in Table 6).

5.2 Analysis for h_3 in News

Analyzing the final 50 scenarios across all seeds, we find dominant modes characterized by mixed sentiment (largely negative with minor positive details-85.6%), large prior rise (74.8%), conflicting stock-valuation interpretations (55.2%), and steep reported prior return drops ($D10+$). Table 5 reports the top-5 modes, showing that disagreement concentrates in scenarios where strong prior trend signals conflict with negative news tone and partial positive evidence. Table 2 further shows that GPT-5.4-mini tends toward negative predictions, GPT-5.5 is more self-consistent and comparatively optimistic, while GPT-5.2 produces a wider and noisier range of numerical predictions. Although GPT-5.2 and GPT-5.5 rely on similar qualitative factors such as horizon and evidence strength, GPT-5.2 is noisier in mapping these judgments to numerical forecasts (examples in Appendix H.3).

5.3 Analysis for h_3 in Disaster Image

We use GPT-5.4-mini, GPT-4.1-mini and GPT-5.2 here. We observe that the most common form of disagreement appears over three action choices, 'evacuate high risk areas', 'close unsafe roads'

Table 2: News scenario analysis. LLM outputs for the five most frequent scenario modes, reported as mean ± standard deviation. Highest and lowest means are highlighted in each mode.

	Mode-1	Mode-2	Mode-3	Mode-4	Mode-5
GPT-5.2	0.2 ± 2.0	$\mathbf{-1.7} \pm 1.8$	2.2 ± 1.4	1.1 ± 1.4	1.1 ± 1.9
GPT-5.4-mini	$\mathbf{-0.9} \pm 1.3$	-1.4 ± 1.6	$\mathbf{-0.8} \pm 1.1$	$\mathbf{-1.3} \pm 1.7$	$\mathbf{0.0} \pm 1.4$
GPT-5.5	$\mathbf{2.6} \pm 0.6$	$\mathbf{2.0} \pm 1.5$	$\mathbf{2.4} \pm 0.6$	$\mathbf{2.4} \pm 0.5$	$\mathbf{2.7} \pm 0.3$



Figure 3: Disaster Image examples with high model disagreement. For Hurricane Michael (left) and Matthew (right), GPT-5.4-mini selects evacuation, GPT-4.1-mini selects infrastructure assessment, while GPT-5.2 alternates between boat surveillance and road closure. Images sampled from the Disaster-M3 Dataset [Wang et al., 2025a]

and 'infrastructure assessment'. (Figure 3). Figure 4 shows subset of scenarios sampled by our approach that have high variance for each model. GPT-5.2 shows high variance across larger number of scenarios, which also leads to a self-inconsistency driven disagreement for this task, despite 66.5% scenarios exhibiting high inter-model disagreement.

5.4 Analysis for h_1 in Camelyon17

Baseline results for this task (Table 3) validate our pipeline against a ground truth reference. With true failure labels, we show that our method achieves $2\times$ the hit-rate of Random Sampling (72.4% vs. 36.0%) and of the best-performing Bayesian baseline (BNN: 34.0%), while maintaining superior coverage and diversity (6.0 distinct patients, 7.8 slides vs. 6.2 and 6.6 for Random Sampling, Table 3). Deep-GP performs worst here (30.4% hit-rate), consistent with its failure on the other image based task (Figure 2). This grounded result confirms that the performance gains on proxy-objective tasks reflect genuine improvements in targeted failure discovery.

6 Discussion

We summarize key results below in the form of research questions supported by our method.

(Q1) How does our framework support the generation of high quality and diverse scenarios?

Figures 2 and 5 show superior performance of our method, achieving high cumulative utility and scenario coverage across all tasks over all baselines. This provides empirical evidence to support our theoretical formulation in Section 3. Deep-GP performs well on low-dimensional tasks (Bank Loan), but degrades on high-dimensional image embeddings (Disaster Image, Camelyon17) where GP kernel approximations lose fidelity; Figure 6 further illustrates these limitations.

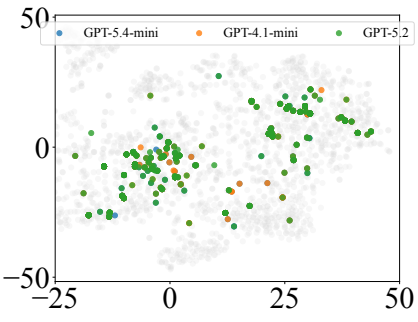


Figure 4: Disaster Image. t-SNE image embeddings with high self-variance cases highlighted.

While BNN is better suited for high dimensional surrogate modeling, it provides reasonable utility but limited coverage, owing to its noisy MC-dropout based stochasticity, leading to ineffective epistemic uncertainty quantification.

LLM-Batch achieves the lowest utility and coverage on every task (Figure 11), reflecting the representation bias of an LLM generating scenarios that mirror its own blind spots. This also exposes the risk of biased representation in using an LLM as a surrogate model in active testing [Berrada et al.].

Our flow-based surrogate scales to high dimensional scenarios, provides reliable estimation of epistemic uncertainty, and avoids biased scenario generation. This is a result of scalable flow surrogate combined with reliable MI estimation (Equation (3)) with the hybrid objective in Equation (4). Figure 5 also shows that when scenario quality is captured by a single utility, our scalar acquisition samples more efficiently than multi-objective baseline q-NEHVI.

(Q2) Is our method robust to uncertainties?

Table 8 reports cumulative objective and coverage under prompt ablations, with P_0 as the reference prompt and P_1, P_2 as ChatGPT-generated variants, showing robustness to prompt specification. Table 9, Figure 9, and Figure 10 further show temperature and model ablations using Llama-3.1-8B-Instruct for h_2, confirming robustness to model uncertainty. We attribute this to the flow-based probabilistic surrogate, which captures input-dependent noise as predictive variance and directs MI-based acquisition toward uncertain regions. While BNN also captures uncertainty using MC-dropouts, we find that epistemic uncertainty representation can be noisy and estimated inaccurately due to MC sampling. The pairwise-distance based MI estimation alleviates this problem by representing epistemic uncertainty analytically, therefore accurately representing unexplored areas.

Robustness to label noise. A concern for h_1-style objectives is sensitivity to label corruption. We ablate this on Camelyon17 by randomly flipping ground-truth labels at rates 0–50% (Table 4). The framework remains robust up to 30% corruption (true hit-rate: 43.3% at 30% vs. 86.7% at 0%), because random corruption increases aleatoric uncertainty without disrupting epistemic uncertainty-driven exploration.

(Q3) How does our acquisition function support diverse and targeted discovery?

Three consistent patterns emerge across all tasks and objectives. **(1) MI is critical for diversity.** NO-MI consistently exceeds the full method on utility but pays a steep cost in diversity. Vendi and pool-cluster counts drop sharply, confirming that the MI term is the primary driver of exploration into novel failure regions. **(2) MI alone over-explores.** ONLY-MI achieves the highest Vendi scores (Bank Loan h_1: 3.45) and pool-cluster counts (20.2) but with severely degraded utility (0.362 vs. 0.829), showing that without explicit objective guidance the acquisition visits diverse but low-value scenarios. **(3) Ensemble MI outperforms single-model MI.** NO-ENS-MI is consistently worse than the full method on utility across all tasks except Camelyon17, with the largest gaps on Bank Loan h_1 (0.460 vs. 0.829), confirming that the PaiDE-based ensemble diversity estimate provides a meaningfully more reliable epistemic signal than MC-dropout.

7 Limitations

Our convergence plots in Figure 2 suggest that testing budgets scale not only with dimensionality (Disaster Image) but also with task complexity (News), which is an open-ended problem in the realm of testing and evaluation. Here, feature-learning approaches such as Recursive Feature Machines [Radhakrishnan et al., 2025], may help reduce the effective search space. Finally, we acknowledge that the superior performance of our approach involves marginally higher computational cost compared to Deep GP due to ensemble training, however it shows scalability across dimensions, unlike Deep GP where performance in complex scenario space plummets (Appendix F).

8 Conclusion & Future Work

We presented a novel adaptive design paradigm for testing LLM based decision making applications, that uses flow matching based surrogate model for generalizable scenario representation. To our knowledge, this is the first work that investigates the use of flow matching surrogates for sample-efficient adaptive design. As a future work, we aim to combine concept learning via hypothesis testing in our framework to deduce key emergent concepts tied to open ended testing objectives, using decision modeling ideas [Zhao et al., 2025] and neuro-symbolic regression [Grayeli et al., 2024].

Acknowledgement

This work was partly supported by the Office of Naval Research (ONR) grant N000142512080, Air Force Office of Scientific Research (AFOSR) grant FA9550-23-1-0099, and a grant from Liberty Mutual. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the sponsors.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Firoj Alam, Umair Qazi, Muhammad Imran, and Ferda Ofli. Humaid: Human-annotated disaster incidents data from twitter. In *Proceedings of the 15th International AAAI Conference on Web and Social Media (ICWSM)*, 2021.

architsharma01. Loan approval prediction dataset. Kaggle Dataset. URL `https://www.kaggle.com/datasets/architsharma01/loan-approval-prediction-dataset`. `https://www.kaggle.com/datasets/architsharma01/loan-approval-prediction-dataset`.

Robert B. Avery, Raphael W. Bostic, Paul S. Calem, and Glenn B. Canner. Credit risk, credit scoring, and the performance of home mortgages. *Federal Reserve Bulletin*, 82:621–648, 1996. URL `https://www.federalreserve.gov/pubs/bulletin/1996/796lead.pdf`.

Israel Abebe Azime, Deborah D Kanubala, Tejumade Afonja, Mario Fritz, Isabel Valera, Dietrich Klakow, and Philipp Slusallek. Accept or deny? evaluating llm fairness and performance in loan approval across table-to-text serialization approaches. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 17478–17503, 2025.

Maximilian Balandat, Brian Karrer, Daniel Jiang, Samuel Daulton, Ben Letham, Andrew G Wilson, and Eytan Bakshy. Botorch: A framework for efficient monte-carlo bayesian optimization. *Advances in neural information processing systems*, 33:21524–21538, 2020.

Casey O Barkan, Sidney Black, and Oliver Sourbut. Do large language models know what they are capable of? In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*.

Gabrielle Berrada, Jannik Kossen, Freddie Bickford Smith, Muhammed Razzak, Yarin Gal, and Tom Rainforth. Scaling up active testing to large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Gabrielle Berrada, Jannik Kossen, Freddie Bickford Smith, Muhammed Razzak, Yarin Gal, and Tom Rainforth. Scaling up active testing to large language models. *arXiv preprint arXiv:2508.09093*, 2025.

Lucas Berry and David Meger. Epistemic uncertainty estimation in regression ensemble models with pairwise epistemic estimators. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Lucas Berry and David Meger. Normalizing flow ensembles for rich aleatoric and epistemic uncertainty modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6806–6814, 2023.

Xiaofeng Cao and Ivor W Tsang. Bayesian active learning by disagreements: A geometric perspective. *arXiv preprint arXiv:2105.02543*, 2021.

Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical science*, pages 273–304, 1995.

Consumer Financial Protection Bureau. 12 CFR Section 1041.5: Ability-to-Repay Determination Required. `https://www.consumerfinance.gov/rules-policy/regulations/1041/2018-01-16/5/`, January 2018. Accessed: 2026-04-26.

Consumer Financial Protection Bureau. What is the ability to repay rule? https://www.consumerfinance.gov/ask-cfpb/what-is-the-ability-to-repay-rule-en-1787/, April 2024. Accessed: 2026-04-26.

Longchao Da, Xiaoou Liu, Jiaxin Dai, Lu Cheng, Yaqing Wang, and Hua Wei. Understanding the uncertainty of llm explanations: A perspective based on reasoning topology. *arXiv preprint arXiv:2502.17026*, 2025.

Sankarshan Damle and Boi Faltings. Llms for resource allocation: A participatory budgeting approach to inferring preferences. *arXiv preprint arXiv:2508.06060*, 2025.

Samuel Daulton, Maximilian Balandat, and Eytan Bakshy. Parallel bayesian optimization of multiple noisy objectives with expected hypervolume improvement. *Advances in neural information processing systems*, 34:2187–2200, 2021.

Eva Eigner and Thorsten Händler. Determinants of llm-assisted decision-making. *arXiv preprint arXiv:2402.17385*, 2024.

Fannie Mae. Selling guide: B3-6-02, debt-to-income ratios. https://selling-guide.fanniemae.com/sel/b3-6-02/debt-income-ratios, 2026. Accessed: 2026-04-26.

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *arXiv preprint arXiv:2210.02410*, 2022.

Karl Friston, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, and Giovanni Pezzulo. Active Inference: A Process Theory. *Neural Computation*, 29(1):1–49, 1 2017. ISSN 0899-7667. doi: 10.1162/NECO{_}a{_}00912.

Jacob Gardner, Geoff Pleiss, Kilian Q Weinberger, David Bindel, and Andrew G Wilson. Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration. *Advances in neural information processing systems*, 31, 2018.

Arya Grayeli, Atharva Sehgal, Omar Costilla-Reyes, Miles Cranmer, and Swarat Chaudhuri. Symbolic regression with a learned concept library. *Advances in Neural Information Processing Systems*, 37: 44678–44709, 2024.

Kimia Hamidieh, Veronika Thost, Walter Gerych, Mikhail Yurochkin, and Marzyeh Ghassemi. Complementing self-consistency with cross-model disagreement for uncertainty quantification. *arXiv preprint arXiv:2604.17112*, 2026.

Matthew Henderson, Paweł Budzianowski, Iñigo Casanueva, Sam Coope, Daniela Gerz, Girish Kumar, Nikola Mrkšić, Georgios Spithourakis, Pei-Hao Su, Ivan Vulić, et al. A repository of conversational datasets. *arXiv preprint arXiv:1904.06472*, 2019.

Alex Hernandez-Garcia, Nikita Saxena, Moksh Jain, Cheng-Hao Liu, and Yoshua Bengio. Multi-fidelity active learning with gflownets. *arXiv preprint arXiv:2306.11715*, 2023.

Valentin Hofmann, David Heineman, Ian Magnusson, Kyle Lo, Jesse Dodge, Maarten Sap, Pang Wei Koh, Chun Wang, Hannaneh Hajishirzi, and Noah A Smith. Fluid language model benchmarking. *arXiv preprint arXiv:2509.11106*, 2025.

Junfeng Jiao, Saleh Afroogh, Abhejay Murali, Kevin Chen, David Atkinson, and Amit Dhurandhar. Llm ethics benchmark: a three-dimensional assessment system for evaluating moral reasoning in large language models. *Scientific Reports*, 15(1):34642, 2025.

Myeongseob Ko, Hoang Anh Just, Charles Fleming, Ming Jin, and Ruoxi Jia. Probing knowledge holes in unlearned llms. *arXiv preprint arXiv:2511.00030*, 2025.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pages 5637–5664. PMLR, 2021.

Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as gaussian processes. *arXiv preprint arXiv:1711.00165*, 2017.

Kalev Leetaru and Philip A Schrodt. Gdelt: Global data on events, location, and tone, 1979–2012. In *ISA annual convention*, volume 2, pages 1–49. Citeseer, 2013.

Jiatong Li, Rui Li, and Qi Liu. Beyond static datasets: A deep interaction approach to llm evaluation. *arXiv preprint arXiv:2309.04369*, 2023.

Yang Li, Jie Ma, Miguel Ballesteros, Yassine Benajiba, and Graham Horwood. Active evaluation acquisition for efficient llm benchmarking. *arXiv preprint arXiv:2410.05952*, 2024.

Yingke Li, Anjali Parashar, Enlu Zhou, and Chuchu Fan. Pragmatic curiosity: A hybrid learning-optimization paradigm via active inference. *arXiv preprint arXiv:2602.06104*, 2026.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Ollie Liu, Deqing Fu, Dani Yogatama, and Willie Neiswanger. Dellma: Decision making under uncertainty with large language models. *arXiv preprint arXiv:2402.02392*, 2024.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Gustavo Malkomes, Bolong Cheng, Eric H Lee, and Mike Mccourt. Beyond the pareto efficient frontier: Constraint active search for multiobjective experimental design. In *International Conference on Machine Learning*, pages 7423–7434. PMLR, 2021.

Ajith Anil Meera and Wouter Kouw. Curvature-aware expected free energy as an acquisition function for bayesian optimization. *arXiv preprint arXiv:2603.26339*, 2026.

Meta AI. Meta-llama-3.1-8b-instruct. `https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct`, 2024. Accessed: 2026-05-07.

Anil B Murthy, Jaron Mink, and Lindsay Sanneman. Metrics for holistic evaluation of llm reasoning about action, change, and planning. In *Bridging Planning and Reasoning in Natural Language with Foundational Models*, 2026.

OpenAI. Openai api model documentation. `https://developers.openai.com/api/docs/models`, 2026. Accessed: 2026-05-07.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Hakan T Otal, Eric Stern, and M Abdullah Canbaz. Llm-assisted crisis management: Building advanced llm platforms for effective emergency response and public collaboration. In *2024 IEEE Conference on Artificial Intelligence (CAI)*, pages 851–859. IEEE, 2024.

Anjali Parashar, Yingke Li, Eric Yang Yu, Fei Chen, James Neidhoefer, Devesh Upadhyay, and Chuchu Fan. Seed-set: Scalable evolving experimental design for system-level ethical testing. In *The Fourteenth International Conference on Learning Representations*.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinybenchmarks: evaluating llms with fewer examples. *arXiv preprint arXiv:2402.14992*, 2024.

Adityanarayanan Radhakrishnan, Mikhail Belkin, and Dmitriy Drusvyatskiy. Linear recursive feature machines provably recover low-rank matrices. *Proceedings of the National Academy of Sciences*, 122(13):e2411325122, 2025.

Tom Rainforth, Adam Foster, Desi R Ivanova, and Freddie Bickford Smith. Modern bayesian experimental design. *Statistical Science*, 39(1):100–114, 2024.

Narun Raman, Taylor Lundy, Thiago Amin, Jesse Perla, and Kevin Leyton-Brown. Steer-me: Assessing the microeconomic reasoning of large language models. *arXiv preprint arXiv:2502.13119*, 2025.

Nitin Rane. Chatgpt and similar generative artificial intelligence (ai) for smart industry: role, challenges and opportunities for industry 4.0, industry 5.0 and society 5.0. *Challenges and Opportunities for Industry*, 4, 2023.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.

Eric Schulz, Maarten Speekenbrink, and Andreas Krause. A tutorial on gaussian process regression: Modelling, exploring, and exploiting functions. *Journal of mathematical psychology*, 85:1–16, 2018.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.746. URL https://aclanthology.org/2020.emnlp-main.746/.

Ishan Tarunesh, Somak Aditya, and Monojit Choudhury. Trusting roberta over bert: Insights from checklisting the natural language inference task. *arXiv preprint arXiv:2107.07229*, 2021.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Junjue Wang, Weihao Xuan, Heli Qi, Zhihao Liu, Kunyi Liu, Yuhan Wu, Hongruixuan Chen, Jian Song, Junshi Xia, Zhuo Zheng, et al. Disasterm3: A remote sensing vision-language dataset for disaster damage assessment and response. *arXiv preprint arXiv:2505.21089*, 2025a.

Shengkun Wang, Taoran Ji, Linhan Wang, Yanshen Sun, Shang-Ching Liu, Amit Kumar, and Chang-Tien Lu. Stocktime: A time series specialized large language model architecture for stock price prediction. *arXiv preprint arXiv:2409.08281*, 2024.

Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. Minilmv2: Multi-head self-attention relation distillation for compressing pretrained transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021.

Xunzhi Wang, Zhuowei Zhang, Gaonan Chen, Qiongyu Li, Bitong Luo, Zhixin Han, Haotian Wang, Zhiyu Li, Hang Gao, and Mengting Hu. Ubench: Benchmarking uncertainty in large language models with multiple choice questions. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8076–8107, 2025b.

Yuchen Wu, Edward Sun, Kaijie Zhu, Jianxun Lian, Jose Hernandez-Orallo, Aylin Caliskan, and Jindong Wang. Personalized safety in llms: A benchmark and a planning-based agent approach. *arXiv preprint arXiv:2505.18882*, 2025.

Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, et al. Finben: A holistic financial benchmark for large language models. *Advances in Neural Information Processing Systems*, 37:95716–95743, 2024.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. In *The Twelfth International Conference on Learning Representations*, 2023.

Daniel Yang, Yao-Hung Hubert Tsai, and Makoto Yamada. On verbalized confidence scores for llms. In *ICLR Workshop: Quantify Uncertainty and Hallucination in Foundation Models: The Next Frontier in Reliable AI*.

Eddie Yang and Dashun Wang. Benchmark illusion: Disagreement among llms and its scientific consequences. *arXiv e-prints*, pages arXiv–2602, 2026.

Jihan Yao, Peter Jin, Ke Bao, Qiaolin Yu, Khushi Bhardwaj, Chang Su, Jialei Wang, Yikai Zhu, Sugam Devare, Damon Mosk-Aoyama, et al. The measure of all measures: Quantifying llm benchmark quality. In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*.

Jing Yao, Xiaoyuan Yi, and Xing Xie. Clave: An adaptive framework for evaluating values of llm generated responses. *Advances in Neural Information Processing Systems*, 37:58868–58900, 2024.

Ziyu Ye, Rishabh Agarwal, Tianqi Liu, Rishabh Joshi, Sarmishta Velury, Quoc V Le, Qijun Tan, and Yuan Liu. Reward-guided prompt evolving in reinforcement learning for llms. In *Forty-second International Conference on Machine Learning*.

Xinli Yu, Zheng Chen, Yuan Ling, Shujing Dong, Zongyi Liu, and Yanbin Lu. Temporal data meets llm–explainable financial time series forecasting. *arXiv preprint arXiv:2306.11025*, 2023.

Renfei Zhang, Mohit Pandey, Artem Cherkasov, and Martin Ester. Why pool when you can flow? active learning with gflownets. *arXiv preprint arXiv:2509.00704*, 2025.

Yibo Zhao, Yang Zhao, Hongru Du, and Hao Frank Yang. Personalized decision modeling: Utility optimization or textualized-symbolic reasoning. *arXiv preprint arXiv:2511.02194*, 2025.

A Related Works

Active sampling and data acquisition for LLM based applications. This work is closely related to approaches that consider sample efficient evaluation of LLMs. These techniques often leverage existing leader-boards and performance of other models to encourage active search or sub-sampling of datasets Li et al. [2024], Polo et al. [2024], Hofmann et al. [2025]. These approaches focus on creating a smaller dataset of a few hundred prompts for effective evaluation, typically sampled from an existing large scale benchmark. However, most of these works assume access to evaluation metrics from other models to speed-up the evaluation of a specific model. Similar to static datasets, these approaches cannot be applied to new tasks if a large collection of prompts is not available. Additionally, the testing objective in these works is defined specifically with respect to disagreement between models, as opposed to our work, that can handle any open-ended testing objective.

Another related line of works uses LLM for adaptive prompt modification Raman et al. [2025], Yao et al. [2024], which we use to construct the LLM-Batch. We see that in structured scenario spaces, this approach may fail and would need a lot more in-context learning examples to adequately uncover the underlying biases and challenges of LLM under test. Our results on model disagreement further show that usage of LLMs as surrogate models for testing other LLMs, as done in Berrada et al. (which also motivates our LLM-Batch baseline with ICL), can lead to biased representation of the target model. Further, the approach cited in Berrada et al. only works for classification tasks, and cannot be easily extended to open-ended tasks.

There are RL-based approaches and sampling based works that train policies for eliciting biases and specific model responses Ye et al., or use MCTS or another search approach for searching through large spaces Wu et al. [2025], but have larger data dependency, needing thousands of prompts to train.

Active learning architectures. Within the scope of large scale active learning that uses non GP based approaches, BNNs are a popular choice Cao and Tsang [2021]. Recently, Gflow-Net based active learning works have also been proposed Zhang et al. [2025], Hernandez-Garcia et al. [2023]. These works require a surrogate model but alleviate dependency on optimization of acquisition function by replacing it with a sampling of distribution that is proportional to the acquisition objective. This approach is more scalable in higher dimensional domains such as molecular discovery, but also needs a larger number of samples to warm start. Potential high dimensional, high budget extensions of our work can involve integration of Gflow-net based data acquisition with Flow matching surrogate.

Interpretable scenario representation. Complementary to our key contribution, efficient scenario representation has been explored in LLM based applications, to make scenario space interpretable and low dimensional using neuro-symbolic paradigm Wu et al. [2025], Liu et al. [2024], Grayeli et al. [2024]. Feature extraction approaches such as Recursive Feature Machines (RFMs) can also be used to extract low dimensional features from data Radhakrishnan et al. [2025]. A key challenge in direct implementation of these approaches is that they require existing labeled data. Our approach performs simultaneous learning of the testing objective, and its optimization online. We adopt the key underlying idea discussed in these works to use symbolic schema for efficient representation of scenario spaces, which enables interpretability.

Defining testing objectives for testing. A large body of works focuses on proposing open-ended and unifying definitions of difficulty, which is often formalized using uncertainty at token level, reasoning topologies, or decision disagreement [Yao et al., Da et al., 2025, Yang and Wang, 2026, Berrada et al.]. Application specific definitions have also been proposed in literature Murthy et al. [2026]. We aim for our approach to be agnostic to the specifics of testing objective, and treat it as a black-box model to be learnt during evaluation, allowing a large body of these definitions to be accommodated.

B Additional results

Below we present some additional results on Camelyon17 task, and other method, model, prompt, temperature ablations, primarily using h_2 testing objective and average cumulative utility and coverage as performance metrics.

B.1 Camelyon17 task results

Refer to Table 3 for baseline + ablation results and Table 4 for label corruption results.

Table 3: Camelyon17 (h_1, ablations + baselines with new metrics) — mean ± std, 5 seeds..

Method	Hit-rate	Avg cum. utility	Coverage (norm.)	Vendi	Pool clusters/25	Patients	Slides
RS	0.360 ± 0.052	0.370 ± 0.041	0.797 ± 0.056	2.34 ± 0.42	7.2 ± 1.7	6.2 ± 0.7	6.6 ± 0.5
IBNN (Deep-GP)	0.304 ± 0.083	0.297 ± 0.064	0.679 ± 0.115	2.30 ± 0.47	6.4 ± 2.3	6.8 ± 1.0	8.2 ± 0.4
BNN	0.340 ± 0.059	0.330 ± 0.056	0.741 ± 0.094	2.54 ± 0.28	8.4 ± 2.3	5.2 ± 0.7	6.4 ± 1.2
Flow (ours)	0.724 ± 0.097	0.650 ± 0.080	1.000 ± 0.044	2.40 ± 0.35	6.8 ± 1.5	6.0 ± 0.6	7.8 ± 0.7
No-MI ($\beta = 0$)	0.772 ± 0.111	0.690 ± 0.093	0.956 ± 0.029	2.14 ± 0.26	5.2 ± 1.2	5.2 ± 0.7	6.6 ± 1.4
Only-MI ($\beta = 1$, util=0)	0.208 ± 0.061	0.220 ± 0.064	0.562 ± 0.147	4.18 ± 1.20	5.4 ± 1.9	5.4 ± 1.0	6.2 ± 1.2
No-ensemble-MI (single)	0.788 ± 0.053	0.697 ± 0.056	0.997 ± 0.076	2.39 ± 0.47	6.6 ± 1.5	5.8 ± 0.7	7.2 ± 1.0

Table 4: Camelyon17 label-noise ablation (mean ± std over 3 seeds).

Flip (%)	Observed hit-rate (%)	True hit-rate (%)	Vendi (true)	Pool clusters (true)
0	86.7 ± 1.9	86.7 ± 1.9	2.30 ± 0.13	7.3 ± 0.9
10	71.3 ± 10.6	76.0 ± 10.7	2.59 ± 0.07	8.0 ± 0.8
20	62.0 ± 12.8	62.7 ± 14.3	2.19 ± 0.38	6.7 ± 2.4
30	41.3 ± 13.9	43.3 ± 22.5	2.45 ± 0.34	7.3 ± 0.9
50	43.3 ± 6.8	18.0 ± 2.8	2.45 ± 0.11	6.3 ± 0.9

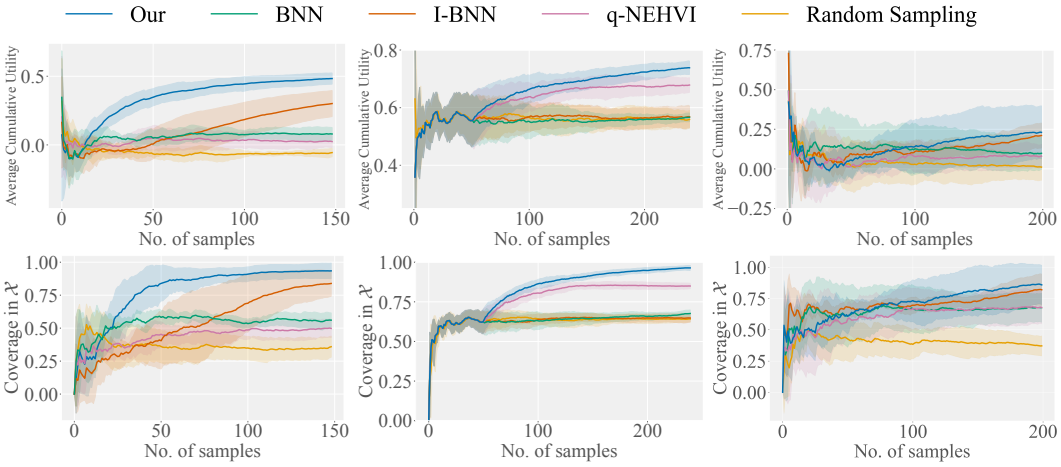


Figure 5: **Baseline comparison for Average Cumulative Utility (top) and Coverage (bottom).** Left to right: Bank Loan, News, Disaster Tweet. Figures report metric ± standard deviation for each baseline across five seeds.

B.2 Dominant scenario modes in News task

Table 5 shows dominant scenario modes discovered by our method.

B.3 Dominant modes for Bank Loan, h_3

Table 6 shows dominant scenarios for Bank Loan, h_3.

B.4 Baseline comparison for h_2

Figure 5 shows baseline comparison for h_2 testing objective, on Bank Loan, News, Disaster Tweet, comparing our approach against BNN, Deep-GP, q-NEHVI and Random Sampling. Figure 12 shows baseline comparison for h_2 on Disaster Image comparing Random Sampling and acquisition strategy ablations.

Table 5: Dominant scenario modes in News.

Mode	Trend	News Driver	Sentiment	Horizon	Evidence	Conflict	Cat.
1 (7.2%)	Large Rise	AI	Neg.+pos.	Near	Mod.	None	D10+
2 (6.4%)	Large Rise	Earnings	Neg.+pos.	Near	Strong	Mom./rev.	D10+
3 (3.2%)	Large Rise	Earnings	Neg.+pos.	Near	Weak	Mom./rev.	D10+
4 (3.2%)	Large Rise	AI	Neg.+pos.	Uncertain	Weak	None	D10+
5 (2.8%)	Large Rise	AI	Neg.+pos.	Near	Strong	Mom./rev.	D10+

Table 6: Bank Loan. Dominant Scenario features and decision analysis.

Mode	M	Dep.	Edu.	Self-emp.	Inc.	Loan	Term	Credit	Prop.	GPT-5.2	GPT-5.4-mini	GPT-5.5
Mode 1	yes (0.64)	1	no	no (0.48)	33,760	278	344	Pos.	Rural	0.08	0.76	0.54
Mode 2	no (0.49)	2	yes	yes (0.51)	89,680	655	178	Pos.	Rural	0.50	0.23	0.80

B.5 Empirical justification for mixture model based approach

We investigate the role of mixture model based MI estimation against MI constructed for a nominal flow using MC sampling approach using the Bank Loan task. The results in Figure 6 show clear superiority of mixture modeling based MI estimation compared to MC sampling based MI (Nominal $\beta = 1$). We theorize that this is because MC sampling approach is sensitive to the design of likelihood in regression tasks, and flow based generative models do not capture epistemic uncertainty explicitly in a direct manner. This is also the reason BNN does not perform well in h_3, Figure 2.

For regression tasks, the expression for MI works well for models where epistemic uncertainty is represented explicitly, such as in GPs, with a kernel choice. We attribute the good performance of Deep-GP in Bank Loan, News in h_3 Figure 2 to that. However, this setup is not scalable, leading to poor performance in Disaster Image in h_3, Figure 2, and also for News in h_2, Figure 5. Our approach scales across scenario dimension, and outperforms both these approaches.

We also note that the performance of our approach with MI is comparable to $\beta = 0$ (no MI) in both utility and coverage. This is due to the fact that the true high utility dataset for GPT-5.2 is concentrated in a small region and is fairly unimodal, leading to localized exploration even with MI (see Figure 7). We discuss the effect of MI in detail in Appendix B.6.

B.6 MI ablation over GPT models, h_2, Bank Loan

To investigate the role of MI, we compare scenario design from full acquisition function in Equation (4) against $\alpha_U(x)$ defined as:

$$\alpha_U(x) = \mathbb{E}_{q_\phi(y|x,\mathcal{D}_k)}[h(y)], \tag{5}$$

for three GPT models (GPT-4-o, GPT-4.1-mini, and GPT-5.2), with h_2 testing objective for Bank Loan task. Both α_U and α give comparable performance in utility, *showing that discovery of optimal candidates is not compromised by exploration*. For both GPT-5.2 and GPT-4o, adding MI improves

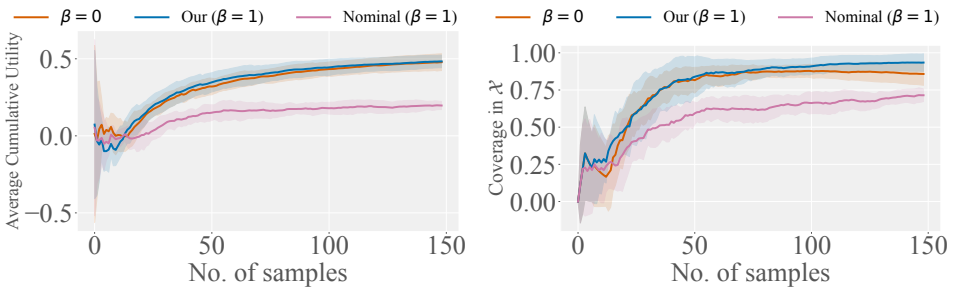


Figure 6: **MI estimation analysis using Bank Loan task.** for MC sampling based MI estimation using nominal FM only (Nominal $\beta = 1$), mixture model based MI estimation (Our $\beta = 1$). We also show $\beta = 0$ for which both approaches are equivalent since there is no MI estimation.

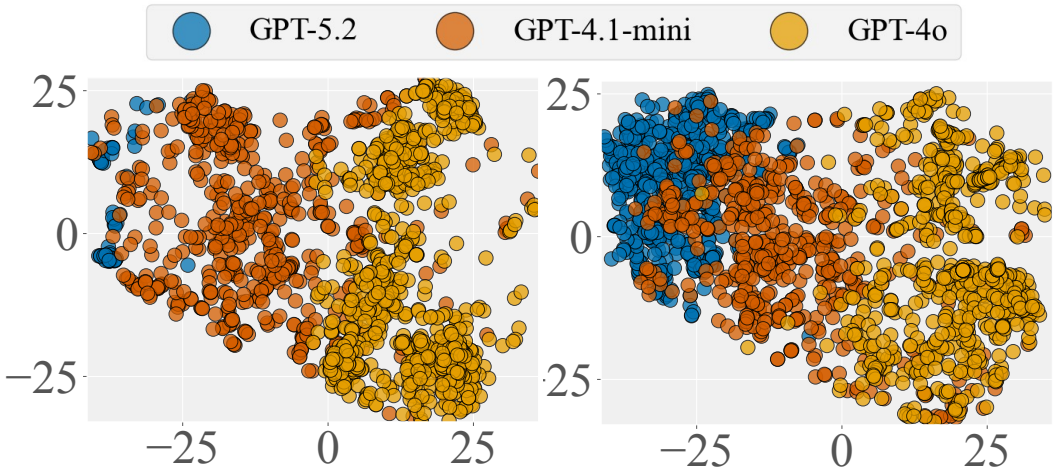


Figure 7: **Visualizing MI-driven exploration for Bank Loan task, h_2 testing objective.** T-SNE embeddings for high utility samples generated by our algorithm with no MI (left) and with MI (right) for three GPT models. Samples are concentrated in a dense region without MI, with limited exploration, whereas inclusion of MI encourages diversity.

Table 7: **GPT Model and MI Ablation for h_2, Bank Loan task.** Average cumulative utility and coverage for 150 samples over five seeds, comparing sampling with and without mutual information (MI).

	Without MI		With MI	
	Utility	Coverage	Utility	Coverage
GPT-5.2	0.4787	0.6857	**0.4837**	**0.7465**
GPT-4.1-mini	0.7536	**0.8437**	**0.7630**	0.8167
GPT-4o	**0.8563**	0.6613	0.8279	**0.7364**

average cumulative coverage by 12%. For GPT-4.1-mini we observe that the coverage decreases slightly, but the difference is marginal. This is also confirmed visually in Figure 7, where the spread of scenarios for GPT-4.1-mini changes marginally with $\beta = 1$. On the other hand, for GPT-5.2 we observe diversified local exploration, and for GPT-4o we observe discovery of new modes by including MI.

B.7 Testing objective and Prompt ablation

Table 8 shows testing objective and prompt ablations for Bank Loan and shows that our method is robust to prompt specifications, in discovering useful and diverse scenarios.

B.8 Model ablations for Bank Loan and Disaster Tweet tasks, h_2 testing objective

Figure 9 and Figure 10 show baseline comparison of our approach with Random Sampling, q-NEHVI, Deep-GP (Deep-GP) for coverage and utility for Llama-3.1-8B-Instruct model (GGUF quantized release), for Bank Loan and Disaster Tweet tasks for h_2 testing objective. We also compare against our method with no MI using a β ablation by defining the acquisition strategy as (Our $\beta = 0$) for Bank Loan task, showing no MI based exploration:

$$\alpha(x) = \beta I(\theta, y|x, \mathcal{D}_k) + \mathbb{E}_{q_\theta(y|x,\mathcal{D}_k)}[y]. \tag{6}$$

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

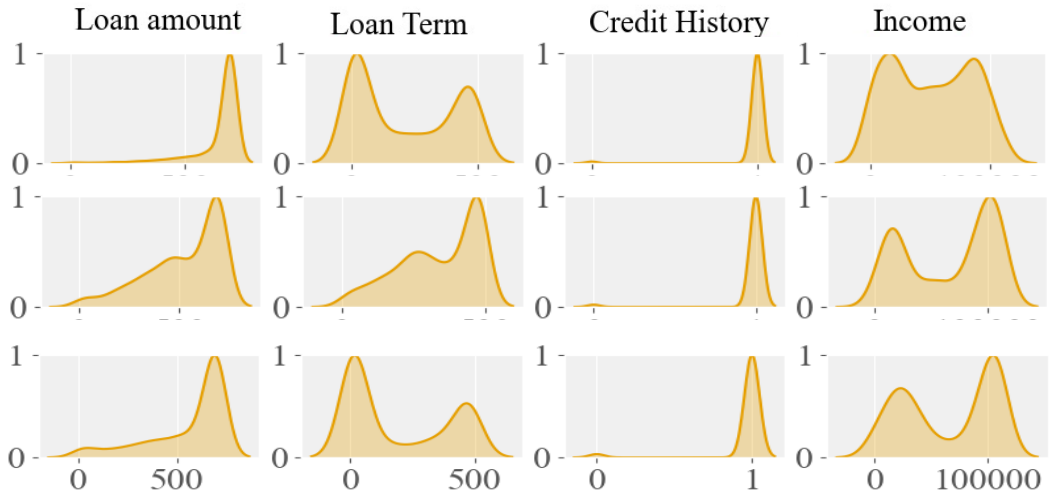


Figure 8: Bank Loan marginal distribution of part of the scenario space, for testing objectives $h_1 h_2, h_3$ (top-bottom). Credit history and loan amount always skew towards unimodal distributions across all objectives, representing two key clashing factors introducing uncertainty in decision-making.

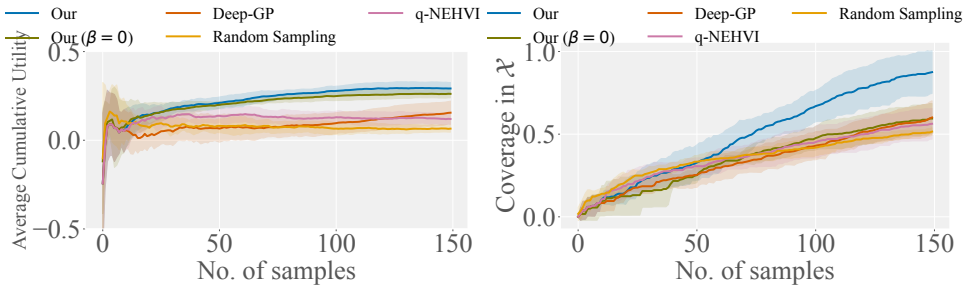


Figure 9: **Baseline comparison for Llama-3.1-8B-Instruct model.** Cumulative Utility and Coverage for Bank Loan task.

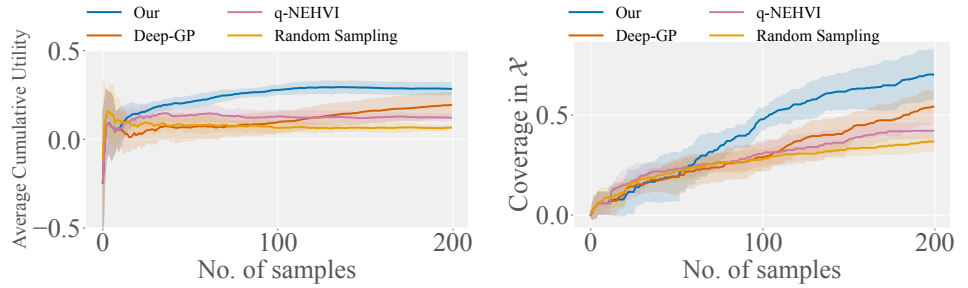


Figure 10: **Baseline comparison for Llama-3.1-8B-Instruct model.** Cumulative Utility and Coverage for Disaster task.

Table 8: **Testing Objective & Prompt Ablation, Bank Loan task.** Final values of cumulative objective $c(y_K)$ and coverage $g(\mathcal{X}_u^K)$ for $K = 150$ (mean $\pm$ std for five seeds).

	h_1		h_2		h_3	
	Obj.	Coverage	Obj.	Coverage	Obj.	Coverage
P_0	0.88 ± 0.03	0.93 ± 0.05	0.45 ± 0.03	0.95 ± 0.03	0.64 ± 0.01	0.96 ± 0.07
P_1	0.91 ± 0.02	0.93 ± 0.05	0.54 ± 0.03	0.96 ± 0.05	0.57 ± 0.06	0.89 ± 0.08
P_2	0.90 ± 0.02	0.96 ± 0.05	0.45 ± 0.05	0.85 ± 0.06	0.66 ± 0.07	0.95 ± 0.06

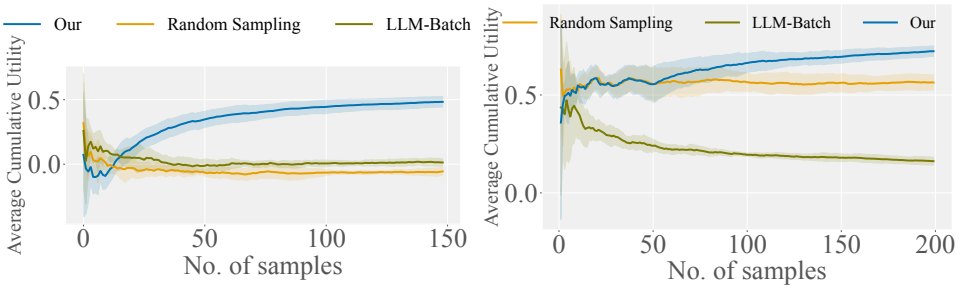


Figure 11: Average Cumulative Utility Comparison of LLM Batch with our approach for Bank Loan task (left) and News task (right)

B.9 Comparison against LLM-Batch for h_2, Bank Loan

We use LLM in this baseline with 10 in-context learning examples of high utility candidates to generate 5 scenarios in a batch, by reasoning over possible latent factors that can be challenging. Figure 11 shows the comparison. To mimic an adaptive experiment design setup, and in the absence of a larger pool of candidates, we provide generated candidates by the LLM in prior batches as samples for in-context learning in consecutive batches. We observe that the performance oscillates between batches.

With some batches the LLM correctly infers the reason for challenging scenarios and generates them, in others it fails to do so and generates candidates that do not represent difficulty. Appendix H.2.4 shows examples of such scenarios in second category.

We use GPT-4.1-mini for results reported in main paper for Bank Loan, News, for h_2 with Bank Loan, we use GPT-5.2 as the generator LLM for this baseline.

B.10 Temperature ablations for h_2 with GPT-5.2, Bank Loan

Table 9 shows temperature ablations with GPT-5.2, no reasoning mode, showing that our method is robust to aleatoric noise in output. This is due to the fact that noisy output for a given scenario is well represented as higher variance in our flow based surrogate model.

B.11 Optimization strategy ablation for h_2, with Disaster Image

We compare the standard optimization approach for data acquisition (ACF) with a dataset sampling with replacement (Dataset I) and without replacement (Dataset II). Optimizing directly with a high dimensional encoding generates scenario candidates that are not well represented in the sparse dataset. This motivates a sampling based approach for scenario selection in this task. At each iteration we estimate α for each candidate in the dataset and choose the most optimal one. Figure 12 shows the utility and coverage for these baselines. We observe that (Dataset I) performs most optimally. While (Dataset I) generates fewer unique candidates, it outperforms (Dataset II) substantially, because the assessment and confidence metrics are noisy, and (Dataset I) encourages repeat sampling of images that show higher uncertainty.

Table 9: **Temperature Ablation for h_2, Bank Loan task.** Final cumulative utility and coverage for $K = 150$ across different temperatures (mean $\pm$ std over five seeds). Percent changes are reported relative to temperature 0.2.

$T = 1.0$		$T = 0.2$		$T = 0.0$	
Utility	Coverage	Utility	Coverage	Utility	Coverage
0.47 ± 0.02	0.96 ± 0.05	0.48 ± 0.03	0.93 ± 0.05	0.47 ± 0.01	0.97 ± 0.05

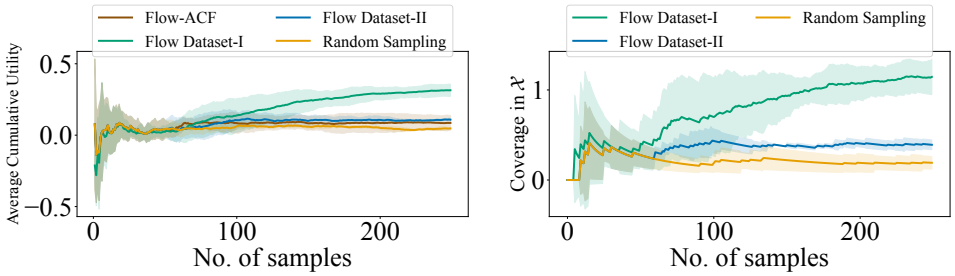


Figure 12: **Baseline comparison Disaster Image task.** Average cumulative utility (top) and coverage (bottom) for various ablations of our method, compared against Random Sampling for reference.

C Design of scenario $\mathcal{X}$ and output space $\mathcal{Y}$ across tasks

Across a broad range of applications discussed in the work, we leverage existing scenario definitions from literature to create the initial expression of the scenario space. For designing the scenario space $\mathcal{X}$ based on these inputs, we normalize the scenario to a continuous vectorized input using transformations and embeddings. This is done to facilitate ease of training surrogate model and data acquisition by optimization.

C.1 Bank Loan

We choose 10-dimensional scenario space of mixed variables shown in Table 10. The ranges and dimensions are chosen using US Bank Loan dataset as a reference from Azime et al. [2025], that is available on Kaggle [architsharma01]. Continuous piecewise transformation is applied to all categorical variables, and normalization is applied using upper and lower limits of the domain of each variable to achieve $\mathcal{X} = [0, 1]^{10}$. Figure 13 shows the scenario generation and LLM-assisted decision-making pipeline.

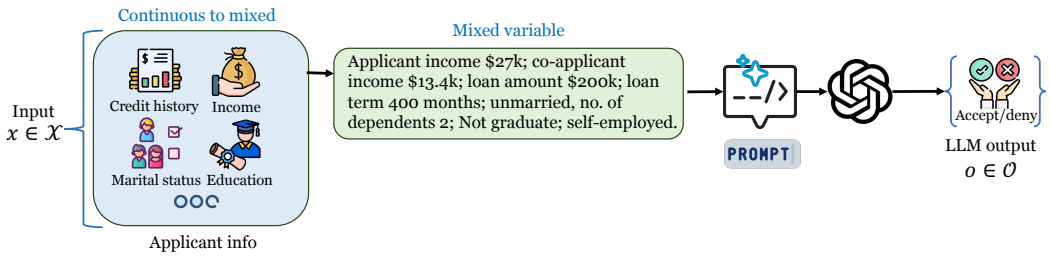


Figure 13: Visualization of scenario space and LLM-assisted decision-making for Bank Loan.

Feature	Description	Domain
x_1	Marital status	{yes, no}
x_2	Number of dependents	{0, 1, 2, 3+}
x_3	Graduate degree	{yes, no}
x_4	Self-employed	{yes, no}
x_5	Applicant income (USD)	[150, 90,000]
x_6	Co-applicant income (USD)	[0, 20,000]
x_7	Loan amount (thousand USD)	[17, 700]
x_8	Loan term (months)	[12, 480]
x_9	Credit history meets guidelines	{yes, no}
x_{10}	Property type	{rural, urban}

Table 10: Definition of scenario variables $x = [x_1, \ldots, x_{10}]$ used in the Bank loan approval task.

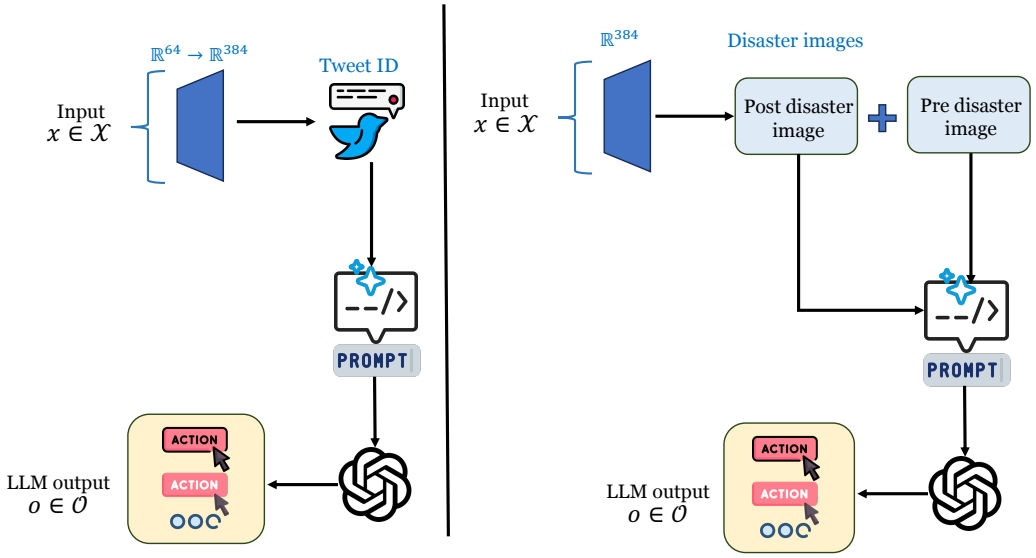


Figure 14: Visualization of scenario space and LLM-assisted decision-making for Disaster Tweet and Disaster Image respectively

Continuous transformation We use a piecewise continuous mapping, $T(c) = \begin{cases} \frac{c-1}{n-1}, & n > 1, \\ 0, & n = 1. \end{cases}$ and $x_i = T(c_i)$ for each categorical variable with n categories.

Output. Output is binary, and represents APPROVE or REJECT of the loan application.

C.2 Disaster Tweet

We use HumAID dataset [Alam et al., 2021] consisting of 1,570 tweets labeled across 8 categories: caution and advice, infrastructure and utility damage, requests or urgent needs, not humanitarian, displaced people and evacuation, sympathy and support, rescue volunteering or donation, other relevant information. We use all-MiniLM-L6-v2 sentence transformer to retrieve 384 dimensional encoding for each of the tweets in the dataset Henderson et al. [2019], Reimers and Gurevych [2019], Wang et al. [2021]. We further compress this to 64-dimensional encoding using PCA transformation for data efficiency. Figure 14 shows pipeline for implementation of Disaster Image, Disaster Tweet.

Output. Output is categorical, and represents one of the seven actions:

`evacuation_messaging` (1; evacuation or air-quality alerts), `road_access_clearance` (2; clear blocked emergency routes), `power_infrastructure_assessment` (2; assess utilities), `fire_damage_survey_drone` (2; assess fire extent and hotspots), `medical_response_team` (3; treat burns, injuries, and smoke inhalation), `evacuation_shelter_setup` (3; shelter displaced residents), and `aerial_fire_suppression` (4; deploy helicopters or air tankers).

C.3 Disaster Image

The Disaster Image task uses satellite images as inputs to the LLM assistant to provide recommendations on recovery strategy. We use Disaster M3 dataset Wang et al. [2025a] consisting of 2,303 before and after satellite images of disasters such as hurricanes and floods. The scenario space is represented by a 384 dimensional encoding generated using DINOv2 Oquab et al. [2023] with a ViT-Small Touvron et al. [2021] backbone and 14×14 patch size. For cost efficiency, we demonstrate our results using GPT-4.1-mini with three seeds, and mainly compare against Random Sampling and other variants of our approach better suited for dataset sampling.

C.3.1 Output for h_2

Output is categorical, as one of nine priority actions:

`search_for_stranded_people` (1), `evacuate_high_risk_areas` (2), `close_unsafe_roads` (3), `set_up_temporary_shelter` (4), `deploy_drone_or_boat_reconnaissance` (5), `supply_delivery` (6), `infrastructure_assessment` (7), `inspect_water_sanitation` (8), and `deliver_food_water_and_supplies` (9).

C.3.2 Output for h_3

We use eight priority actions, each associated to an index, and clustered in three groups. Indices (1,2,3,4) form cluster-1, (9,10) is cluster-2, and (14,15) is cluster-3. This is done to also encode semantic distance of choices made by various LLMs, instead of just binary agree/disagree for more interpretability.

`search_for_stranded_people` (1),
`evacuate_high_risk_areas` (2),
`set_up_temporary_shelter` (3),
`deliver_food_water_and_supplies` (4),
`deploy_drone_or_boat_reconnaissance` (9),
`close_unsafe_roads` (10),
`infrastructure_assessment` (14), and
`inspect_water_sanitation` (15).

C.4 News

This pipeline is adopted from news based stock prediction in Wang et al. [2024]. We perform a continuous linear transformation to map the categorical scenario space to normalized, continuous domain in $[0, 1]^7$ for optimization in h_3, and in $[0, 1]^{131}$ for h_2.

C.4.1 Scenario design in h_3

Our scenario space consists of 7 dimensional keywords that are used for generating synthetic news articles using GPT-4.1-mini in h_3 experiments in main paper. Final dimension is quantitative value of prior month's stock prediction. The keywords are shown below. **News scenario space.** We define news scenarios using seven categorical factors: *prior trend* $\in$ {large drop, small drop, flat, small rise, large rise}; *sentiment* $\in$ {mostly positive, mostly negative, balanced mixed, positive w/ negative caveat, negative w/ positive caveat}; *news driver* $\in$ {earnings/guidance, product demand, analyst/institutional, macro/sector, legal/regulatory, supply chain, competitive AI strategy}; *horizon* $\in$ {near-term, long-term, uncertain timing}; *evidence strength* $\in$ {weak, moderate, strong}; *conflict mode* $\in$ {none, sentiment vs. news type, short-term vs. long-term, company vs. macro, momentum vs. reversal}; and prior month stock return bin.

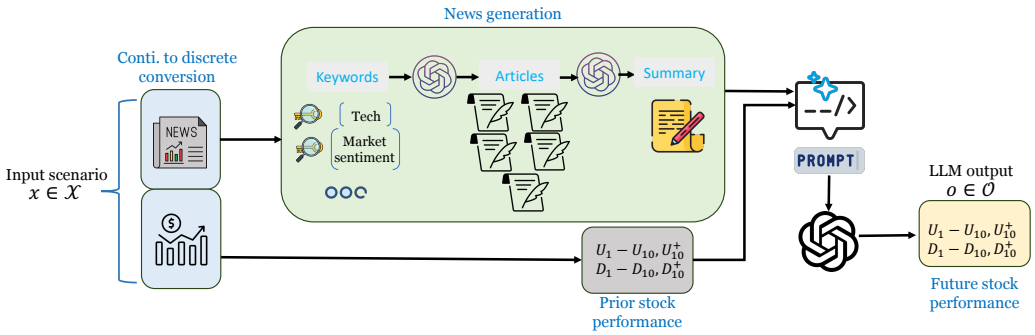


Figure 15: Visualization of scenario space and LLM-assisted decision-making for News

C.4.2 Scenario design in h_2

In h_2 experiments, the keyword space represents 5 articles, is 131 dimension, with each article represented by 26 keywords using one-hot encoding. Final dimension is quantitative value of prior month's stock prediction.

Figure 15 visualizes the complete pipeline. We collect news articles from online finance news dated between January 2024 to May 2024 using GDELT 2.1 Document API (Leetaru and Schrodt [2013]). These are used to extract 26 relevant keywords using ChatGPT-5.2. Each article a_i is represented as a binary vector $(b_j)_{j=1}^{26}$, where $b_j \in \{0, 1\}$, showing whether the keyword can be associated to that article or not. Our scenario space consists of keyword mapping for five articles, leading to a 26×5 dimensional search space. This is used alongwith a stock return from prior month k that can take one in 22 values as in $\{U_1, \ldots, U_{10}, U_{10}^+, D_1, \ldots, D_{10}, D_{10}^+\}$ to create the scenario space. This is normalized between [0,1] and the combined 131 dimensional vector is used for optimization using Botorch's optimization pipeline.

Article generation from keywords Data acquisition strategy returns a vectorized data-point, which we then convert to a collection of five articles and a stock return. To generate articles using keywords, we use GPT-5.2 by providing actual news articles from Yahoo finance as in-context learning examples.

Output. Output is one of 22 categories, showing predicted future stock increase or decrease, in the same format as input stock return for prior month.

C.5 Camelyon17

We include this task with 10 initial and 50 iterations of active learning across five seeds, across baselines and ablations, we observe that our approach balances optimization (hit rate, utility) and exploration (coverage, vendi, pool clusters) in the most efficient way without compromising significantly on one for another.

Goal: find histopathology patches where LLM (GPT-4.1-mini) confidently misclassifies (i.e., always misclassifies across 3 predictions) tumor status.

Dataset: 2,000 balanced 96×96 lymph-node histopathology patches from the Camelyon17-WILDS (`https://huggingface.co/datasets/wltjr1007/Camelyon17-WILDS`) breast-cancer metastasis dataset [Koh et al., 2021]. Each patch is embedded to a 384-d DINOv2 ViT-S/14 vector (L2-normalized). These embeddings define the search space X. Metadata for patches available: patient, slide, center, and gold label.

LLM task: GPT-4.1-mini is shown the patch image and asked to classify it as NORMAL vs TUMOR, returning JSON {answer, rationale}.

Testing objective (what we maximize): a *consistent-failure* signal. The LLM is queried three times; then we annotate the result using Y, where $Y = 1$: all 3 trials agree on the same wrong label (consistently incorrect) , $Y = 0$: they agree and are correct, or the trials disagree (inconsistent).

So a failure is a patch the model is reliably and confidently wrong about, not just occasionally wrong.

D Testing Objective Design

D.1 Design of testing objective h_1 (Bank Loan).

We define a separate, rule-based oracle o_x^* as a baseline for Bank Loan for a given user profile, using affordability (loan amount, loan term and income, x_7, x_5, x_6) and credit trustworthiness (x_9) to construct Debt-to-income (DTI) as a key criterion for approval:

$$\mathrm{DTI}(x) = \frac{x_7/x_8}{x_5 + x_6}. \tag{7}$$

This is consistent with regulatory guidelines on ability to repay loans and mortgages [Consumer Financial Protection Bureau, 2018, Avery et al., 1996, Consumer Financial Protection Bureau, 2024].

$$o_x^* = \begin{cases} 0, & \text{if } (x_9 = 0) \text{ or } (x_9 = 1 \text{ and } \mathrm{DTI}(x) > 0.45), \\ 1, & \text{if } x_9 = 1 \text{ and } \mathrm{DTI}(x) \leq 0.36, \\ 0.5, & \text{if } x_9 = 1 \text{ and } 0.36 < \mathrm{DTI}(x) \leq 0.45. \end{cases} \tag{8}$$

Under this criterion, loan is rejected if credit history criteria is not met, or if DTI is above 0.45 [Fannie Mae, 2026] ($o_x^* = 0$), and accepted for good credit history and low DTI. For the case of marginal DTI with strong credit history, we introduce a soft approve, $o_x^* = 0.5$. Based on this criteria, we define a testing objective as deviation from o_x^* as:

$$h(x, o) = \begin{cases} 1, & \text{if } (o = 1 \text{ and } o_x^* = 0), \text{ or } (o = 0 \text{ and } o_x^* = 1), \\ 0, & (o = 1 \text{ and } o_x^* = 1), \text{ or } (o = 0 \text{ and } o_x^* = 0), \\ 0.5, & \text{if } o_x^* = 0.5. \end{cases} \tag{9}$$

Note that this oracle is not intended to reproduce real lending decisions or regulatory compliance. It is a simplified reference policy for stress-testing whether an AI decision system respects basic affordability and creditworthiness constraints.

D.2 Design of testing objective h_1 (Camelyon17).

Here, h_1 is simply deviation from a known ground truth label. See Appendix C.5 for more details on the objective design.

D.3 Design of testing objective h_2.

D.3.1 Bank Loan

We collect three responses consisting of brief verbal reasoning for the decision from the LLM. This is used to estimate confidence using NLI entailment of reasoning with decision using the roberta-large-mnli model Liu et al. [2019]. We perform NLI based estimation for all three responses and take the average as confidence value y_1. The confidence values are automatically reported between $[0, 1]$.

For assessing difficulty by self-assessment, the LLM is asked to provide a score s between 1-7, with 1 being the easiest and 7 being the most difficult. We normalize these scores to get $y_2 \in [0, 1]$.

D.3.2 News

We estimate confidence by using NLI entailment from roberta-large-mnli for reasoning provided by the LLM. The entailment is done to test if the reasoning supports the decision of increasing or decreasing the price. Therefore, confidence is a coarse evaluation in this case.

For assessing difficulty by self-assessment, the LLM is asked to provide a score s between 1-7, with 1 being the easiest and 7 being the most difficult. We normalize these scores to get $y_2 \in [0, 1]$.

D.4 Disaster Image, Disaster Tweet

We estimate confidence by querying LLM 3 times, and computing Jaccard similarity s of the set of suggested actions. This is used to construct confidence measurement as $y_1 = \exp(6(s-1))$.

For self-assessment in Disaster Tweet, we request binary labels from the LLM in the following format:

```
Assessment:
- 0 = tweet provides clear, concrete evidence of wildfire impacts and urgency or is
    completely irrelevant to the disaster or for resource allocation
- 1 = this tweet is potentially sharing something very crucial that could be helpful
     for resource allocation but it needs more investigation
```

For self-assessment in Disaster Image, we repeat the process with score range from 1-3.

D.5 Design of testing objective h_3.

News, Disaster Image use scaling constants which come from the range of outputs, to normalize the output for the sake of consistency across objectives. We run repeat evaluations for h_3 to observe consistent results.

D.5.1 Bank Loan

$$h_3^r(x, o_1, o_2, o_3) = \begin{cases} 1, & \text{if } \exists\, i, j \in \{1,2,3\} \text{ such that } o_i \neq o_j, \\ 0, & \text{otherwise,} \end{cases}$$
$$h_3(x, o_1, o_2, o_3) = \frac{1}{3}\sum_{r=1}^{3} h_3^r(x, o_1, o_2, o_3). \tag{10}$$

D.5.2 News

$$h_3(x, o_1, o_2, o_3) = \frac{10}{22}\frac{1}{3}\sum_{i=1}^{3}(|o_1^i - o_2^i| + |o_3^i - o_2^i| + |o_1^i - o_3^i|) \tag{11}$$

D.5.3 Disaster Image

$$h_3(x, o_1, o_2, o_3) = \frac{1}{25}\frac{1}{2}\sum_{i=1}^{2}(|o_1^i - o_2^i| + |o_3^i - o_2^i| + |o_1^i - o_3^i|) \tag{12}$$

E Experiment Setting

E.1 Training, model architecture

We use an MLP as the backbone architecture for our approach, BNN, q-NEHVI. Our method and q-NEHVI also use the same flow model and flow matching loss for training, and differ only in acquisition strategies. All models for these baselines are kept the same, with same training parameters, that are task specific, and fine-tuned over the initial data and discussed below. The initial dataset for training surrogate model is randomly sampled and also shown in Table 11. The total budget B is fixed to 250 across all tasks, and early stopping is used for cases such as Bank Loan and Disaster Tweet where early convergence to high average cumulative utility is observed.

E.2 Training params for tasks.

See Table 11 for training hyperparam chosen for each task. The models are chosen based on training loss analysis on the initial, randomly sampled dataset. The size of initial dataset is chosen based on the amount of data needed to get good representation of mutual information in the ensemble based flow matching surrogate model for a held out test dataset that is far away from training distribution. We use the same amount of data and same samples for warm-starting all surrogate models.

Table 11: Training, optimization and reported parameter settings across tasks and testing objectives (Adam optimizer with lr 1e-3 was used for all training).

Setting	Bank Loan			News		Disaster Image		Disaster Tweet
	h_1	h_2	h_3	h_3	h_2	h_3	h_2	h_2
Training epochs (q_ϕ, nominal)	1000	1000	1000	2000	2000	5000	5000	5000
Training epochs (q_θ, ensemble model)	1000	1000	1000	100	100	5000	5000	400
MLP parameters (layers and largest hidden dim.)	3, 64	3, 64	3, 64	3, 32	4, 256	6, 128	6, 128	3, 64
LLM query/epoch	3	3	3	3	3	2	3	3
No. of initial samples	10	10	10	20	50	50	50	20
No. of total samples (B)	150	150	150	250	250	250	250	200
λ used in coverage def.	$\frac{1}{1.5}$	$\frac{1}{1.5}$	$\frac{1}{1.5}$	$\frac{1}{1.5}$	$\frac{1}{1.5}$	5	5	$\frac{1}{3.4}$

E.3 Optimization strategy for optimizing Equation (1).

Bank loan, News. On the transformed continuous space for BankLoan, News, Disaster Tweet, we use Monte Carlo (MC) acquisition strategy using BoTorch library in python for performing optimization of the acquisition function at each step [Balandat et al., 2020]. This optimization procedure has two hyperparameter; (1) no. of MC samples used in estimating acquisition function value at random points, and (2) number of restarts to repeat optimum search based on acquisition value obtained from MC samples.

Optimization hyperparameter Each hyperparameter presents a solution quality and computational memory trade-off. We observe from initial test runs that for MC samples greater than 128 and number of restarts greater or equal to 10, the solution quality is comparable in the cumulative utility achieved. We fix MC samples to 512, and number of restarts to 10 for both tasks across all testing objectives.

Disaster Tweet, Disaster Image These two tasks consist of scenarios represented as datasets. We create encodings corresponding to each tweet/image in dataset. For Disaster tweet, the encoding is 64 dimensional, and we perform optimization in continuous encoding space and find nearest neighbor solution in the dataset. The optimization procedure uses same hyperparameter and procedure as before. This approach does not scale well to the 384 dimensional encoding space for Disaster Image task. Hence, we attempted alternate optimization procedures summarized below.

Optimization in Disaster Image.

Continuous optimization (ACF baseline) Optimization in continuous encoding space led to poor performance on utility optimization in this task (see baseline ACF in Figure 12). This is due to the closest match in dataset being far from the discovered solution, as the encodings are high dimensional, and multiple solutions in continuous space correspond to the same data point, leading to inefficiency.

Dataset optimization (Dataset-I,II baseline) We performed optimization by directly selecting the top acquisition value candidate from the dataset at each pass (Dataset-I baseline, Figure 12), which worked better than continuous optimization. Since the dataset was small, we were able to estimate acquisition value of all candidates using GPU. For larger datasets, batched estimation of acquisition can be performed. Same procedure was followed for Camelyon17.

The idea behind mapping of mixed space to continuous transformation is used from Bayesian optimization literature, but does not scale well in high dimensional dataset sampling scenarios. For such cases, batched or full pass of top candidate selection is more efficient.

F Computational Resources & Reproducibility

F.1 Assets & Licenses

We use DisasterM3 for satellite-image disaster scenarios, its dataset card lists CC BY-NC-SA 4.0, and the project page restricts images and annotations to academic, non-commercial use. Image embeddings are computed with DINOv2 using a ViT-Small backbone; DINOv2 code and model weights are released under Apache 2.0. Tweet embeddings are computed with all-MiniLM-L6-v2,

which is also released under Apache 2.0. For tweet-based experiments, we use HumAID, a publicly released disaster-tweet dataset, and follow the dataset and Twitter/X-derived-data usage constraints. For news-based experiments, we use the GDELT DOC 2.0 API; GDELT describes its database as free and open, and we use it only to access news metadata and source articles for research. OpenAI GPT models are accessed through the OpenAI API under OpenAI's service terms, and Llama-3.1-8B-Instruct is used under the Llama 3.1 Community License. We do not redistribute third-party datasets, raw publisher text, or modified model weights.

F.2 Model initialization & computational cost.

We initialize all surrogate models using a collection of randomly sampled scenarios. Retraining of surrogate model at each step incurs increasing computational cost with data-points for Deep-GP with $\mathcal{O}(n^3)$. Computational cost is comparable across BNN, and q-NEHVI and our method ($\mathcal{O}(n)$) in number of datapoints) as training uses gradient based updates that can be done on GPU. Training has a computational cost for these three methods with number of epochs, k, but is substantially fast for real-time online update. Our approach has an additional online update cost with number of ensemble models, m and also uses gradient based updates over bootstrapped samples, sampled with replacement. However, we parallelize the training of ensembles across the GPU and cut down the online training cost.

Random Sampling and LLM-Batch do not require retraining of surrogate, but LLM-Batch incurs additional cost of LLM query at each batched step, which generates 5 scenarios at once.

The major runtime bottleneck of our approach is not in the training or retraining of surrogate models, rather in LLM querying, which has to be done 2-3 times per epoch for getting stable evaluations as the outputs can be very noisy.

All simulations were conducted on a Linux workstation with Ubuntu 22.04 LTS equipped with an Intel 13th Gen Core i7-13700KF CPU (16 cores, 24 threads, up to 5.4 GHz) and an NVIDIA GeForce RTX 4090 GPU (24 GB VRAM). The Bayesian Experimental Design (BED) loops were implemented using wrapper BoTorch [Balandat et al., 2020] and GPyTorch [Gardner et al., 2018] libraries. The compute requirements were consistent with standard usage of these libraries and did not require additional specialized hardware. OpenAI API keys were used to access GPT models, and Llama-3.1-8B-Instruct was used locally by downloading from Huggingface Touvron et al. [2023]. Examples of LLM prompts used in the generation of results reported in the paper are also provided in the Appendix.

G Mutual Information detail

The use of MI as a metric for uncertainty measurement is prevalent in active learning literature Chaloner and Verdinelli [1995].

$$I(\theta, y|x, \mathcal{D}_k) = H(y|x, \mathcal{D}_k) - \mathbb{E}_{p(\theta|\mathcal{D}_k)} H(y|x, \theta, \mathcal{D}_k), \tag{13}$$

where $H(\cdot)$ denotes entropy.

The first term in (13) indicates entropy of model prediction, which is high when the model prediction is uncertain. The right term is expectation of entropy of model prediction over θ posterior, which is low when each prediction over each posterior has low uncertainty. This difference leads to high MI, which is equivalent to the case of model predictions disagreeing with each other. Conceptually this can happen either due to high uncertainty of output y, or due to lack of sufficient exploration around x, leading to high epistemic uncertainty at x. Thus, optimizing MI promotes exploration in $\mathcal{X}$ and diversity of scenarios.

G.1 MI in pairwise estimation

With the PaiDE form of MI estimator, the closed form expression for KL divergence of Gaussian distributions can be used [Berry and Meger]. We use Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ as base distribution in our flow matching loss), resulting in:

$$D_{\mathrm{KL}}(p_0^i \,\|\, p_0^j) = \frac{1}{2}\left[\mathrm{tr}\left(\Sigma_{j|x}^{-1}\Sigma_{i|x}\right) - D + \ln\left(\frac{\det \Sigma_{j|x}}{\det \Sigma_{i|x}}\right) + (\mu_{j|x} - \mu_{i|x})^\top \Sigma_{j|x}^{-1}(\mu_{j|x} - \mu_{i|x})\right], \tag{14}$$

where tr denotes trace of a matrix. $D_{\mathrm{KL}}(p_0^i \,\|\, p_0^j)$ as defined here can be directly used in Equation (3). Along the lines of Berry and Meger [2023], we use the same velocity model v_θ across ensembles, with bootstrapping with sample replacement to train the different ensemble heads $\mathcal{N}(\mu_{i|x}, \Sigma_{i|x})$. In nominal FM training, these are typically kept fixed. Here, $\mu_{i|x}, \Sigma_{i|x}$ are predicted using an MLP for each x, and our implementation closely follows the Nflow-Base model in Berry and Meger [2023].

H Scenario examples

H.1 Scenario examples h_1, Bank Loan, GPT-5.2

Shown below are assessment reports for several examples for h_1 testing objective with $h = 1$, i.e., optimal objective values. All correspond to the decision of loan rejection.

- Strong credit history supports approval, but affordability/DTI concerns from a large loan size and long term create a significant trade-off.
- Strong credit history supports approval, but the large loan size and extended term create a significant affordability/risk trade-off.
- Credit history is positive, but affordability metrics (high loan size and long term vs income) make the outcome fairly clear
- The income-to-loan mismatch is severe and makes the decision straightforward even with a positive credit history.
- The income-to-loan mismatch is severe and makes the decision straightforward even with a positive credit history.
- Good credit and no dependents support approval, but the loan size-to-income mismatch and self-employment income stability concerns make the decision challenging.
- Credit history is positive, but the loan-to-income mismatch is a clear primary rejection driver, making the decision fairly straightforward.

H.2 Scenario examples h_2, Bank Loan

We show metric scores and reasoning of LLM for various scenarios across three models.

H.2.1 GPT-4-o

Confidence: 0.013, **Subj. Hardness**:0.91, **Reasoning:**The applicant does not meet the credit record guidelines, which is a critical factor for loan approval.

Confidence: 0.52, Subj. **Hardness**:0.99. **Reasoning:** The applicant does not meet credit record guidelines, and the loan amount requested is high relative to the combined income.

Confidence: 0.8, **Subj. Hardness**:0.12, **Reasoning:** The applicant has a high income, a college degree, a good credit history, and no dependents, which supports their ability to repay the loan.

Confidence: 0.64, **Subj. Hardness**:0.66, **Reasoning:** The applicant has a college degree, a satisfactory credit history, and a reasonable income to support the requested loan amount, despite being self-employed.

H.2.2 GPT-4.1-mini

Confidence: 0.19, **Subj. Hardness**:0.99, **Reasoning:** High loan amount relative to income, long loan term, self-employment without degree, and rural property location increase risk.

Confidence: 0.16, **Subj. Hardness**:0.99, **Reasoning:** High loan amount relative to income and self-employment status increase risk despite good credit history.

Confidence: 0.12, **Subj. Hardness**:0.99, **Reasoning:** Poor credit history and high loan amount relative to income with long term.

H.2.3 GPT-5.2

Confidence: 0.38, **Hardness**: 0.66

Reasoning Strong positives (credit history, education, no dependents) are outweighed by major affordability and structure red flags (very large principal and very long term), making the decision trade-off heavy. Requested loan amount is extremely high relative to stated incomes and the term is unusually long, creating high repayment and risk concerns despite good credit history.

Confidence: 0.35, **Subj. Hardness**:0.99

Reasoning Requested loan amount is very high relative to combined income and the term is extremely long, creating elevated repayment and risk concerns despite acceptable credit history.

H.2.4 Scenarios generated by LLM-Batch, h_2, Bank Loan

LLM generated example Scenario A

Married, 3 or more dependents, no education, not self-employed, very high income, no co-applicant income, highest loan amount, shortest loan term, good credit history, semi-urban or rural area.
Explanation: The combination of high income with a high loan amount and short term could lead to model uncertainty.

LLM generated example Scenario B

Married, 3 or more dependents, educated, self-employed, highest applicant and co-applicant income, lowest loan amount, shortest loan term, good credit history, semi-urban or rural area.
Explanation: This profile is complex due to the high income and low loan request, which might confuse the model.

H.3 Scenario examples h_3, News

Table 12 shows examples from dominant modes of disagreement in the News task for h_3 objective.

H.4 Commonly occurring keywords in top scenarios sampled for h_2, News

Using cut-offs $\sum_j x_i^j > 0.65$ for keywords, we find the following keywords recur often in challenging scenarios.

Article-1: (SelloffOversold, RetailSentimentConsumerConfidence, LeadershipStrategyShift).

Article-2: MarketSentimentShifts.

Article-3: EcosystemPlatformCompetition.

Article-4: LeadershipStrategyShifts.

Article-5: SoftwareServicesApps

H.5 Examples of evaluations, News, h_2

Figure 16 shows an example of scenario with high testing objective score for h_2, i.e., showing low confidence and high internal reported difficulty. Shown below are summaries, prior stock return, LLM output and reasoning for different types of scenarios.

H.5.1 Low confidence, High Difficulty

Prior Stock return: D10
XX stock has been experiencing notable fluctuations, with periods of both outperformance and underperformance. Despite recent short-term volatility, analysts remain optimistic about company's growth prospects, particularly driven by robust demand. The company's AI innovations and supply chain resilience are also seen as positive factors bolstering investor confidence.

Table 12: Representative model justifications for dominant News modes. Scores are ordered as GPT-5.4-mini, GPT-5.5, and GPT-5.2.

Mode	GPT-5.4-mini	GPT-5.5	GPT-5.2
Mode-1: large rise; competitive AI strategy; negative sentiment with positive caveat; near-term; moderate evidence; no conflict; D10+.			
Mode-1	**-3**: Mixed fundamentals and cautious sentiment suggest downside risk, with limited support from demand and services strength.	**2**: The prior selloff appears overextended; stable demand and services resilience support a small positive move.	**3**: Oversold conditions and stable demand support a small recovery despite AI-cost and valuation headwinds.
Mode-1	**1**: Durable ecosystem and cash generation support a modest rebound, but slower AI rollout and valuation cap upside.	**2**: A small recovery is likely after the D10+ selloff, though AI concerns should cap gains.	**3**: AAPL may bounce modestly from an extreme drawdown, with AI and valuation limiting magnitude.
Mode-2: large rise; earnings guidance; negative sentiment with positive caveat; near-term; strong evidence; momentum-vs-reversal; D10+.			
Mode-2	**-3**: Softer outlook and margin pressure make a pullback more likely than another leg higher.	**-3**: Negative guidance revisions and margin concerns suggest modest downside continuation.	**-3**: Cautious outlook and margin sensitivity after a momentum run-up increase downside risk.
Mode-2	**-2**: Soft guidance and margin pressure after an extended rally make a modest pullback more likely.	**2**: Oversold conditions and institutional tech momentum support a small recovery, but weak fundamentals cap gains.	**3**: Trend-following inflows and mean reversion after a $> 10\%$ drop should lift shares modestly.
Mode-3: large rise; earnings guidance; negative sentiment with positive caveat; near-term; weak evidence; momentum-vs-reversal; D10+.			
Mode-3	**-3**: Stretched valuation and softer guidance outweigh supportive sentiment, suggesting a modest decline.	**2**: Supportive positioning suggests a small recovery, but fundamental caution limits expected gains.	**2**: Oversold conditions and supportive flows favor a small bounce, though fundamentals limit upside.
Mode-3	**-3**: Softer-than-hoped guidance makes further decline or consolidation more likely than a quick rebound.	**2**: A small recovery is likely after the steep decline, but weak fundamentals and crowded positioning limit gains.	**3**: Oversold conditions and incremental AI/supply-chain positives support a small bounce.
Mode-4: large rise; competitive AI strategy; negative sentiment with positive caveat; uncertain timing; weak evidence; no conflict; D10+.			
Mode-4	**-2**: Cautious sentiment, tougher AI competition, and stretched valuation make a small decline more likely.	**-2**: Competitive AI concerns and less forgiving valuation suggest modest downside.	**-3**: Elevated expectations and uncertain AI revenue/margin impact increase downside risk.
Mode-4	**-3**: Slower AI monetization and valuation risk suggest downside, though ecosystem strength limits the drop.	**3**: A small positive bounce is likely, but AI strategy and earnings uncertainty limit upside.	**3**: The D10+ drop raises odds of a relief rally, with AI and margin uncertainty capping gains.
Mode-5: large rise; competitive AI strategy; negative sentiment with positive caveat; near-term; strong evidence; momentum-vs-reversal; D10+.			
Mode-5	**-1**: Elevated expectations and AI margin/cost concerns make a small pullback more likely.	**3**: A modest recovery is likely after the prior sharp drop, though AI cost concerns cap gains.	**-3**: Elevated valuation and AI cost/margin uncertainty increase downside risk.
Mode-5	**-3**: Crowded positioning and higher AI expectations make a modest decline more likely than a rebound.	**3**: The sharp decline creates room for limited recovery, but AI execution concerns cap upside.	**4**: Mean reversion plus supportive demand should lift shares despite valuation and execution headwinds.

Stock Prediction: U2

LLM provided reasoning: Despite recent volatility and a significant drop last month, Apple's strong fundamentals, high institutional interest, and robust iPhone demand suggest potential for recovery. Positive factors such as AI innovations and supply chain resilience bolster investor confidence, indicating a likely rebound in stock performance.

H.5.2 High confidence, High Difficulty

Prior Stock return: D1
Apple's stock has recently experienced both positive and negative movements. On the positive side, the stock has seen a 2% rise, outperforming the tech sector, driven by strong iPhone sales, innovation in wearables, and a strategic shift to India that enhances supply chain resilience. Analysts have raised price targets, reflecting strong consumer confidence and robust earnings growth. However, the stock has also faced sharp selloffs and short-term volatility, leading to oversold conditions.

Stock Prediction: U2

LLM provided reasoning: Despite recent volatility and selloffs, Apple's strong iPhone sales, innovation in wearables, and strategic expansion into India suggest a positive outlook. Raised price



Figure 16: **Example of a scenario generated by our method for the News task**. Our data acquisition strategy samples a collection of keywords, which are used by an agent to generate five news articles with mixed sentiments. Even though LLM predicts increase in stock prices, due to conflicting evidences coupled with prior month's stock return to be decreasing, the scenario is marked challenging and incurs low confidence from NLI entailment.

targets and positive retail sentiment indicate potential for a stock price increase, although institutional caution may temper gains.

H.5.3 High confidence, Low Difficulty

Prior Stock return: D2
Apple's stock performance has been mixed, with notable rises driven by strong iPhone demand, innovative security features, and advancements in AI strategies. Positive market sentiment and increased institutional interest have contributed to these gains. However, the stock has also faced declines due to concerns over iPhone demand fluctuations and supply chain disruptions. Despite these challenges, analysts see potential buying opportunities, supported by Apple's strong fundamentals and brand loyalty. The company continues to face competitive pressures in the semiconductor and hardware sectors, which may impact its long-term growth.

Stock Prediction: U3

LLM provided reasoning: Apple's stock is likely to see a moderate increase next month due to strong iPhone demand, positive market sentiment, and increased institutional interest. Despite challenges like supply chain disruptions and competitive pressures, Apple's strong fundamentals and brand loyalty provide a solid foundation for growth.

Prior Stock return: D8
Apple's stock has experienced fluctuations, initially underperforming compared to its tech peers and the broader market due to short-term volatility and shifting market sentiment. However, recent developments have seen a notable rise in its stock price, driven by strong demand for its latest iPhone release and increased investor confidence. Analysts have adjusted price targets upward, reflecting optimism in Apple's long-term growth prospects, bolstered by strong fundamentals, innovation, and AI-driven strategies. Institutional interest is growing, with hedge funds increasing their stakes, betting on Apple's value and stability. The company's robust services growth and brand loyalty further enhance its appeal to long-term investors.

Stock Prediction: U3

LLM provided reasoning: Apple's stock is likely to rebound due to strong iPhone demand, increased investor confidence, and upward price target adjustments. Institutional interest and robust services growth further support a positive outlook.

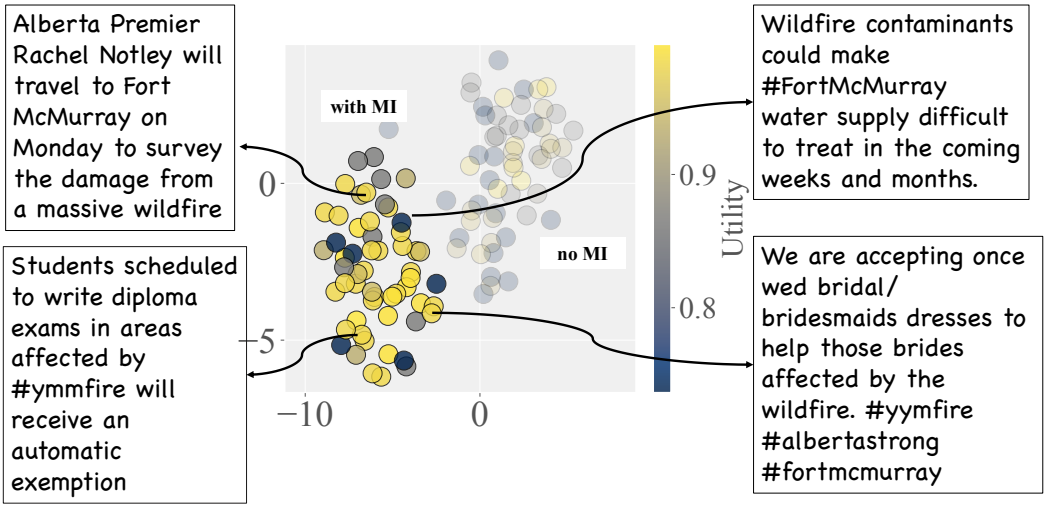


Figure 17: **Effect of MI on exploration, Tweet task.** Tweets sampled using our method with MI shown against tweets sampled with only optimization (no MI) in the background for the same seed. MI enables discovery of new optimal areas with our combined objective. Without exploration, sampling often gets stuck in localized regions. Examples of tweets that align with our testing objective (low confidence, high difficulty) show that GPT-5.2 categorizes tweets with presence of human activity with missing or partial information as challenging. The task allocation for such tweets displays high uncertainty between actions due to resource constraints.

H.6 Examples for Disaster Tweet, h_2

Below we show a range of examples for GPT-5.2. We first discuss low utility tweets and then show tweets sampled by our approach.

H.6.1 Examples of High confidence, Low difficulty

These tweets are generally not conveying any useful information pertaining to the wildfire, hence marked low difficulty and high confidence because the action space is null.

> Easiest way to help FortMacFire is text REDCROSS to 30333 - Donate 5$ on your next phone bill - Help save #Ymm.
>
> Great opportunity for the students. Hope #smokeybear talked about the difference between #goodfires and bad fires.
>
> Great opportunity for the students. Hope #smokeybear talked about the difference between #goodfires and bad fires. This is our opportunity to love our #neighbours in
>
> Fort #McMurray . #canadianredcross cross taking donations. Wildfire contaminants
>
> could make #FortMcMurray water supply difficult to treat in the coming weeks and months. Wonder if #normmacdonald would care to donate to Norm MacDonald?

The second category of tweets that are also low difficulty and high confidence are those that provide relevant information, but the LLM is confident in the set of actions to choose, and hence confidence of action is high. These tweets are also clear in tone, and do not suggest urgency, leading to low difficulty.

Wildfire official Chad Morrison says between 18 and 20 structures in Anzac were destroyed when fire hit late Thurs. night Yes, applauding destruction of personal property for your ends continues to be a smart way to go. #yymfire #Cdnpoli Thanks again to everyone who has raised awareness for #ymmfire and for your kind words. Donate redcrosscanada

H.6.2 Examples of High Confidence, High Difficulty tweets

These tweets are high difficulty due to suggestion of human activity, and ambiguity or lack of complete information, and high confidence due to known or no set of actions to be performed.

Strangers help bride who lost dress in Fort McMurray wildfire

H.6.3 Examples of Low Confidence, High Difficulty tweets

Shown below are tweets sampled by our approach, prioritized due to high utility.

These tweets are informative for rescue, donation needs but require more detail (high difficulty), and due to lack of complete information the LLM shows uncertainty in suggested actions (low confidence).

More information needed:

Devastating and Heartbreaking. Please help if you can. SHOCKING video of the destruction of the #ymmfire in #FortMcMurray. #FortMacFire Online food drive supports wildfire victims Thunder Bay women rally donations for Fort McMurray wildfire evacuees: Residents in Thunder Bay, Ont., have l

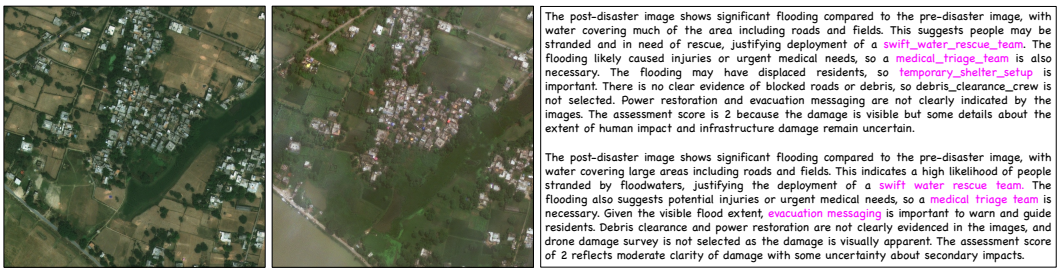
We also observe the tweet shown below to be evaluated repeatedly because it dwindles between low and high confidence. This suggests that our approach accommodates uncertainty of observation by suggesting repeat evaluation.

Tough day.for the evacuees who have to go through this #ymmfire #FortMacFire

H.7 Example images from Disaster Image, for h_2.

Figure 18 shows an exemplar scenario sampled using our approach. The before and after disaster images reveal some damage, but the extent of damage is unclear, leading to moderate difficulty and high uncertainty of actions, leading to low confidence.

We show examples of images that score low on testing objective (Figures 19 to 21) compared against images that score high (Figure 22).



The post-disaster image shows significant flooding compared to the pre-disaster image, with water covering much of the area including roads and fields. This suggests people may be stranded and in need of rescue, justifying deployment of a swift_water_rescue_team. The flooding likely caused injuries or urgent medical needs, so a medical_triage_team is also necessary. The flooding may have displaced residents, so temporary_shelter_setup is important. There is no clear evidence of blocked roads or debris, so debris_clearance_crew is not selected. Power restoration and evacuation messaging are not clearly indicated by the images. The assessment score is 2 because the damage is visible but some details about the extent of human impact and infrastructure damage remain uncertain.

The post-disaster image shows significant flooding compared to the pre-disaster image, with water covering large areas including roads and fields. This indicates a high likelihood of people stranded by floodwaters, justifying the deployment of a swift water rescue team. The flooding also suggests potential injuries or urgent medical needs, so a medical triage team is necessary. Given the visible flood extent, evacuation messaging is important to warn and guide residents. Debris clearance and power restoration are not clearly evidenced in the images, and drone damage survey is not selected as the damage is visually apparent. The assessment score of 2 reflects moderate clarity of damage with some uncertainty about secondary impacts.

Figure 18: **High utility scenario example Disaster-Image task.** Pre and post-disaster images (left, right) provided as an input to LLM for immediate task allocation. Text box shows GPT-4.1-mini reasoning over scenario difficulty and task allocation. The scenario is ranked moderately difficult (assessment score=2) due to unclear visible damage to human life, leading to high uncertainty in preferred immediate actions between the two outputs.



The post-disaster image shows a change in the color and condition of the large rectangular area adjacent to the buildings, suggesting possible flooding or contamination. However, there is no clear evidence of stranded people, injuries, blocked roads, or power failures visible. Given the ambiguity of the damage and the need for more detailed information, deploying a drone damage survey (cost 2) is prioritized to map the damage and identify access issues. Additionally, evacuation messaging (cost 1) is selected to ensure safety communication in case of unseen hazards. This selection uses only 3 points, leaving room for further actions once more information is available.

The post-disaster image shows a significant change in the color and appearance of the large rectangular area near the buildings, suggesting possible flooding or contamination. However, there is no clear evidence of stranded people, injuries, blocked roads, or power failures visible from the image alone. Due to this ambiguity, deploying a drone damage survey (cost 2) is prioritized to map damage and identify blocked access routes. Additionally, evacuation messaging (cost 1) is selected to warn and inform potentially affected people. Other actions cannot be justified without more visible evidence.

Figure 19: High Confidence, High Difficulty, Hurricane Florence



The post-disaster image shows significant flooding around the central structure and surrounding fields, which were dry in the pre-disaster image. This flooding likely stranded people, so deploying a swift_water_rescue_team is critical. The flooding also suggests displacement, so temporary_shelter_setup is necessary. Evacuation_messaging is important to warn and guide affected residents. Due to the extensive flooding and potential access issues, a drone_damage_survey will help map the extent of flooding and identify blocked routes for better response planning. The damage and flooding are clearly visible, so the assessment score is 1

The post-disaster image shows extensive flooding around the central structure and surrounding fields, which were dry in the pre-disaster image. This flooding likely stranded people and displaced residents, justifying the deployment of a swift water rescue team (action 1) and temporary shelter setup (action 4). The presence of floodwaters also supports sending evacuation messaging (action 7) to inform and guide affected populations. There is no clear evidence of blocked roads or debris, so debris clearance is not selected. Medical needs and power issues are not visually evident, so those actions are not prioritized. The damage and flooding are clearly visible, so the assessment score is 1

Figure 20: Low Confidence, Low Difficulty, Nepal Flooding

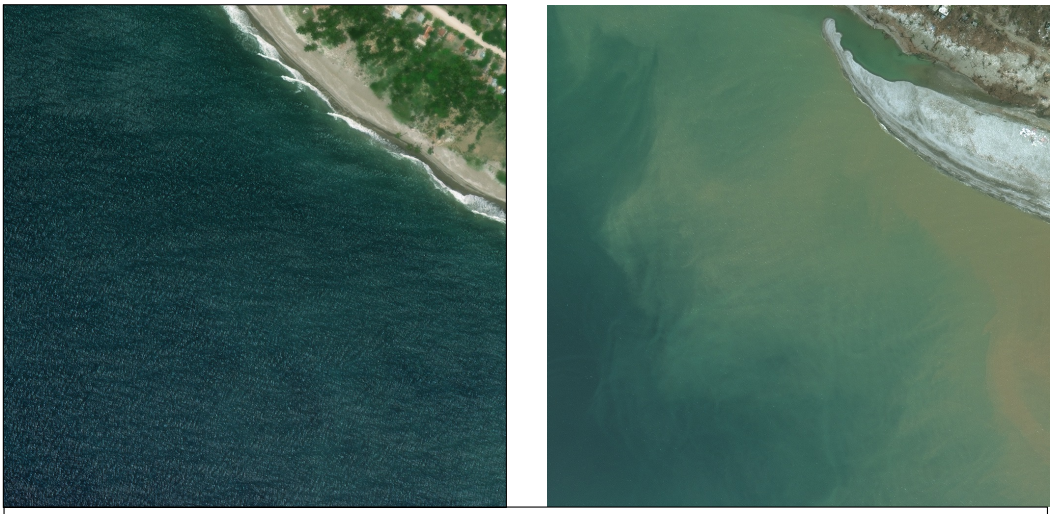


The post-disaster image shows extensive flooding with water covering large areas and surrounding many buildings, indicating people may be stranded and in need of rescue, justifying the deployment of a swift_water_rescue_team. The flooding likely displaced residents, so temporary_shelter_setup is necessary to provide shelter for those affected. Evacuation_messaging is important to inform residents about safety measures and ongoing rescue efforts. Due to the widespread flooding and potential blocked access routes, deploying a drone_damage_survey will help map the damage and identify safe routes for response teams. The damage and flooding are clearly visible when comparing pre- and post-disaster images, allowing confident action selection.

The post-disaster image shows extensive flooding with many buildings partially submerged, indicating people may be stranded and in need of rescue (action 1). The flooding likely displaced residents, so setting up temporary shelters (action 4) is critical. Evacuation messaging (action 7) is necessary to warn and guide affected populations. Due to the widespread flooding and potential blocked routes, deploying a drone damage survey (action 6) will help map damage and identify accessible routes for response teams. The damage and flooding are clearly visible compared to the pre-disaster image, justifying these targeted actions."

The post-disaster image shows extensive flooding with many buildings partially submerged in water, indicating a high risk of people stranded and displaced. Therefore, deploying a swift_water_rescue_team is critical to rescue stranded individuals. Temporary_shelter_setup is necessary to accommodate displaced residents. A drone_damage_survey will help map the flooded area and identify blocked access routes for effective response. Evacuation_messaging is important to inform residents about safety measures and evacuation routes. The damage is clearly visible and significant, so the assessment score is 1.

Figure 21: High Confidence, Low Difficulty, Hurricane Florence



The post-disaster image shows significant changes in water color and shoreline, indicating possible flooding, sediment displacement, or contamination. There is visible debris along the shore and some structural damage to the land area. However, the extent of human impact, stranded people, or medical needs is unclear. Therefore, a drone damage survey (action 6) is prioritized to better map damage and identify blocked access routes. Debris clearance (action 3) is selected due to visible debris along the shore, which may block access or pose hazards. Evacuation messaging (action 7) is included to warn or guide people in the affected area.

The post-disaster image shows significant changes in water color and shoreline condition compared to the pre-disaster image, indicating possible sediment displacement and damage to the coastal area. However, there is no clear visual evidence of stranded people, injuries, blocked roads, or power failures. Due to the ambiguity of visible damage and lack of clear evidence for specific response actions, deploying a drone damage survey (action 6) is prioritized to gather more detailed information quickly. Additionally, evacuation messaging (action 7) is selected to ensure public safety given the uncertain conditions. This combination uses minimal resources while addressing uncertainty and potential risk.

Figure 22: Hurricane Harvey: Low confidence, high difficulty

I Prompts

Prompts shown for h_2 for all tasks.



The post-disaster image shows significant flooding compared to the pre-disaster image, with water covering much of the area including roads and fields. This suggests people may be stranded and in need of rescue, justifying deployment of a swift_water_rescue_team. The flooding likely caused injuries or urgent medical needs, so a medical_triage_team is also necessary. The flooding may have displaced residents, so temporary_shelter_setup is important. There is no clear evidence of blocked roads or debris, so debris_clearance_crew is not selected. Power restoration and evacuation messaging are not clearly indicated by the images. The assessment score is 2 because the damage is visible but some details about the extent of human impact and infrastructure damage remain uncertain.

The post-disaster image shows significant flooding compared to the pre-disaster image, with water covering large areas including roads and fields. This indicates a high likelihood of people stranded by floodwaters, justifying the deployment of a swift water rescue team. The flooding also suggests potential injuries or urgent medical needs, so a medical triage team is necessary. Given the visible flood extent, evacuation messaging is important to warn and guide residents. Debris clearance and power restoration are not clearly evidenced in the images, and drone damage survey is not selected as the damage is visually apparent. The assessment score of 2 reflects moderate clarity of damage with some uncertainty about secondary impacts.

Figure 23: Nepal Flooding: Low confidence, high difficulty

I.1 Bank Loan Task

```
*Task: You are a loan approver, and you have to decide using the
presented information of a candidate if a loan must be approved or not.
Based on the applicant information, decide approve or reject. Return your
answer in a proper JSON format with the following entries:

'decision': 'APPROVE' or 'REJECT'.
'reasoning': A very short explanation supporting the decision including
primary factors, without explaining in detail.
'assessment': Report a number on a scale of 1 to 7 scoring the
difficulty of this case for loan approval, with 1 being the easiest/most
straightforward, and 7 being the least easy/challenging. You can use your
internal metrics for making this assessment.
'dimension': If the difficulty is higher than 5, provide dimensions
between 1-10 that are imposing a trade-off or causing difficulty in
decision. Report as a list, [dim1, dim2,...]. Otherwise report an empty
list.
```

```
*Candidate details:
1. Marital status: {marital_status}
2. No. of dependents: {x[1]}
3. Does the applicant have a college degree: {education}
4. Is the applicant self-employed: {self_employed}
5. Applicant income: {x[4]}
6. Co-applicant income: {x[5]}
7. Loan amount requested in thousand USD: {x[6]}
8. Term of loan in months: {x[7]}
9. Does the credit history of the applicant meet credit record
guidelines: {credit_history}
10. What type of area does the applicant own property in:
{property_area}

Use the provided information to give a response
```

I.2 News Task

```
*Task: You are an analyst forecasting the NEXT MONTH return category for
a stock.

Return categories: D10+ (drop >10%), D1, D2, D3, D4, D5, D6, D7, D8, D9,
D10 (drop between (k-1)% and k%), U1, U2, U3, U4, U5, U6, U7, U8, U9, U10
(rise between (k-1)% and k%), U10+ (rise >10%).

Stock: {ticker}

Last-month realized return:
- return_bin: {month_ret_bin}

News (same month) meta-summary & keywords:
{news_meta_summary}

Task:
1) Predict NEXT MONTH return_bin.
2) Provide a short 'Summary' explaining the key drivers behind your
forecast.
3) Provide 'Keywords'.
4) Provide output: 'INCREASE' if stock price increases or 'DECREASE' if
stock price decreases.
5) Provide 'assessment': Report a number on a scale of 1 to 7 scoring the
difficulty of this case, with 1 being the easiest/most straightforward,
and 7 being the least easy/challenging. You can use your internal metrics
for making this assessment.
```

```
*Output format follow exactly:
Summary: ...
Keywords: ...
Stock Return: <one of the bins above>
Output: INCREASE/DECREASE
Assessment: <1,2,3,4,5,6,7>

Only output one of bins for stock returns.
Be concise. Do not output anything else.
```

I.3 Disaster Tweet Task

```
*Task: You are a wildfire response operations planner. You must extract
actionable situational information from a single tweet and decide which
response actions to deploy in the next 6 hours to maximize harm reduction,
under resource constraints and uncertainty.

Goal: Choose a set of response actions (from the list below) with total
cost <= 10 points, based ONLY on information contained in the tweet. Do
NOT assume facts not stated or strongly implied by the tweet.

Budget: - 10 points total.
```

```
*Available actions (cost -> description):
1) 'aerial_fire_suppression' (4) -> deploy helicopters or air tankers to
slow fire spread
2) 'medical_response_team' (3) -> treat burns, injuries, smoke inhalation
3) 'evacuation_shelter_setup' (3) -> provide shelter for displaced
residents
4) 'evacuation_messaging' (1) -> issue evacuation or air-quality safety
alerts
5) 'road_access_clearance' (2) -> clear debris blocking evacuation or
emergency routes
6) 'power_infrastructure_assessment' (2) -> assess damage to power lines
and utilities
7) 'fire_damage_survey_drone' (2) -> assess fire extent, hotspots, and
access constraints
```

```
*Rules:
- You may select multiple actions, but total_cost must be <= 10.
- Only select an action if you can cite specific evidence from the tweet
supporting it.
- If evidence is ambiguous or incomplete, prefer selecting
'fire_damage_survey_drone' (7) and/or 'evacuation_messaging' (4).
- Prioritize actions that address immediate threats to life and safety.

Priority scale:
- 1 = low urgency
- 5 = immediate life-threatening risk

Return ONLY valid JSON (no markdown) with exactly:
{
''selected_actions'':[ { ''action'': int, ''cost'': int, ''priority'': int } ],
''total_cost'': int,
''assessment'': int
}

Assessment:
- 1 = tweet provides clear, concrete evidence of wildfire impacts and
urgency
- 2 = some relevant information but partial or indirect
- 3 = insufficient or highly ambiguous information; more assessment needed
```

I.4 Disaster Image Task

```
*Task: You are a disaster response operations planner. You have ONE pair
of images of a site before and after a disaster.

Your goal: choose a set of response actions to deploy in the next 6 hours
to maximize harm reduction, subject to limited resources and uncertainty.
Use ONLY what is visible in the image.

Budget: 10 points total.
```

```
*Available actions (cost -> description):
1) 'swift_water_rescue_team' (4) -> rescue people stranded by floodwater
2) 'medical_triage_team' (3) -> treat injuries / urgent medical needs
3) 'debris_clearance_crew' (3) -> clear blocked roads, remove fallen
trees/debris
4) 'temporary_shelter_setup' (3) -> set up shelter for displaced people
5) 'power_restoration_assessment' (2) -> assess downed lines / critical
power failures
6) 'drone_damage_survey' (2) -> map damage quickly, identify blocked
access routes
7) 'evacuation_messaging' (1) -> send targeted evacuation/safety messaging
```

```
*Rules:
- You may choose multiple actions but total cost must be <= 10.
- Only select an action if you can cite specific visual evidence
supporting it.
- If evidence is ambiguous, you may select 'drone_damage_survey' to reduce
uncertainty.

Return ONLY valid JSON (no markdown) with exactly:
{
''selected_actions'': [{''action'': int (indicating action number), ''cost'':
int, ''priority'': 1-5}],
''total_cost'': int,
''assessment'': 1-3 #Use this to score the images, by comparing evidence
from pre and post disaster images, lower score indicates higher clarity of
damage caused by disaster as a significant difference, that is clearly
visible in the post disaster image. Compare pre and post disaster
scenarios to get this score. Use 3 when you think more information has
to be provided to evaluate the scenario.
}
```

J Broader impacts

This work leverages active inference for adaptive test case discovery with LLM-assisted decision making, with the goal of maximizing societal productivity, and highlighting scenarios where the performance of LLMs is unpredictable, or rooted in fundamental biases, across diverse domains. This work has tremendous future scope for early debugging of LLMs before deploying them in real-world decision making tasks. In more open-ended settings, where testing objective is not well specified or theoretically grounded, we depend on user-defined definitions for constructing testing objective, in the form of existing user data, natural language specifications or symbolic concepts. Applied with the right testing objective, the work has tremendous advantages, however, if user data is being used to train a testing objective, care must be taken to ensure that data being used aligns with the high level testing goal. Misrepresentation of testing objective due to bias from user data must be factored in before real-world deployment.

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 paper proposes a new methodology for testing LLM based decision making systems in real world applications in Introduction and abstract. We provide rigorous empirical evaluation of our approach across several real world applications, alongwith baseline comparison, and model ablations considering a wide range of LLM models, in Section 4 and 5, and additional results in Appendix to support the main claims.

   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: We discuss key limitations of the work in Section 6. We also provide computational cost comparison of all approaches in Appendix F.2.

   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 is centered towards a novel application procedure, hence we provide rigorous empirical evidence.

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: The problem setup of each problem is discussed in the main paper, including number of samples. We report our results as mean and standard deviation across five seeds for each experiment and use the collected data from these experiments for further analysis. Appendices C to F and I includes all implementation details, concerning computational costs, LLM prompts used in the various tasks and baselines, design of objectives, as well as computational resources used in the experiments.

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: [No]

   Justification: All the main results reported in the paper use OpenAI API keys for accessing GPT models, which are non-open source. The code will be released with open source alternatives upon acceptance for general public access. We provide extensive implementation details in Appendices C to F for all results reported in the paper.

   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: All training details concerning size of initial datasets, training epochs, model architectures, optimization protocols are provided in Appendix E. All modeling details are covered in Appendices C and D

   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 mean $\pm$ 1 standard deviation for all training curves, calculated over five seeds, shown across number of samples in Figure 2.

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: Appendix F discusses all hardware details pertaining to the computations done for the results in the paper.

   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: The research conforms with the NeurIPS Code of Ethics in all aspects.

   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: [Yes]

Justification: The broader impacts have been discussed in Appendix J.

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 do not release models or data as a contribution of this work, the contribution is a standalone algorithm design that can be implemented on any LLM based decision making task, with any choice of LLM. To our knowledge, there is no risk of misuse for our algorithm.

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: All the assets (code, baseline methods) are properly credited in this manuscript. Besides inline citations, we provide a list of assets used with licenses and terms of use in Appendix F.1.

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 dataset/code/model. Code will be released upon acceptance.

    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: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    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: [N/A]

Justification: The paper does not involve crowdsourcing or research with human subjects.

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: [Yes]

Justification: The paper proposes an approach for testing LLM based decision making tasks. We have declared every LLM used to validate our approach in the experiment Section 4 and 5, along with prompts used for each task in Appendix I

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.