WHICH LLM TO PICK? ONLINE ACTIVE MODEL SELECTION FOR LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Language Models (LLMs) are increasingly applied to process streaming data, with practitioners relying on benchmarks to select the best model even though these signals only approximate real performance. While oracle annotations can provide reliable feedback, they are often costly and difficult to obtain at scale. To address this challenge, we propose ONLINE LLM PICKER, the first framework for active model selection for LLMs in online settings. Given an arbitrary stream of queries and a limited annotation budget, ONLINE LLM PICKER selects the most informative prompts for annotation to identify the best LLM among candidate models. Across multiple tasks including 10 datasets, for over 130 language models, we show that ONLINE LLM PICKER saves annotation cost by up to 71.67% while reliably identifying the best or near-best model for the stream. We also show that using the returned model for sequential generation on unannotated prompts across the stream reduces regret by up to a factor of $2.51\times$, indicating that ONLINE LLM PICKER can identify the best or near-best model well before processing all streaming prompts.

1 INTRODUCTION

Large Language Models (LLMs) are widely applied across various domains (Fan et al., 2023; Tan, 2023), from medicine (Boll et al., 2025; Veen et al., 2024; Biswas & Talukdar, 2024) and industry (Angelopoulos et al., 2025; Kok et al., 2024; Li et al., 2024b), to education (Dan et al., 2023; Caines et al., 2023) and law (Pont et al., 2023; Lee, 2023), (Guha et al., 2023). Alongside this growing adoption, hundreds of publicly available LLMs have emerged (Google-t5, 2024; Meta-Llama, 2024; Falcon, 2023; DeepSeek, 2025; MistralAI; Qwen, 2025; Nvidia, 2025), many demonstrating strong performance without the need for fine-tuning and the ability to follow instructions and adapt to new tasks with little or no task-specific supervision (Brown et al., 2020; Kojima et al., 2023; Dong et al., 2024; Liu et al., 2021).

Given the abundance of off-the-shelf models, selecting the most suitable LLM for a particular task or data stream is a non-trivial problem. Existing evaluation pipelines and benchmarks assess performance across diverse datasets and tasks using multiple metrics (Guo et al., 2023), but no single model consistently excels across domains or datasets (Chang et al., 2023; Liang et al., 2023). Because LLM effectiveness is highly context-dependent and varies substantially across scenarios, standard evaluation metrics alone cannot fully capture model quality or practical utility (Ouyang et al., 2022; Kocoń et al., 2023). A common approach to model selection is therefore to rely on randomly or heuristically chosen small subsets of annotated data (Polo et al., 2024; Vivek et al., 2024). However, such strategies often lead to inefficient use of resources and fail to reliably capture differences across models (Kossen et al., 2021). Active Model Selection (Madani et al., 2012; Karimi et al., 2021; Ashury-Tahan et al., 2024; Okanovic et al., 2024; Kay et al., 2025; Liu et al., 2022; Li et al., 2024a; Hara et al., 2024; Gardner et al., 2015) addresses this limitation by selectively annotating a small set of queries to identify the best model for arbitrary data examples. Yet, prior work has largely focused on classification tasks rather than generation (Karimi et al., 2021; Liu et al., 2022; Okanovic et al., 2024; Kay et al., 2025; Li et al., 2024a; Hara et al., 2024), with the exception of Ashury-Tahan et al. (2024), which studies language model selection under the assumption that all

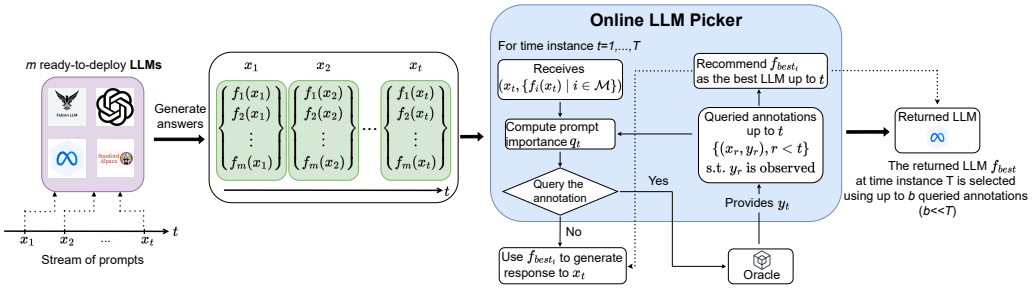

Figure 1: An overview of ONLINE LLM PICKER workflow.

prompts are available from the outset. To the best of our knowledge, no prior work has explored active model selection for LLMs in online settings to date.

Contributions: To fill this gap, in this work we propose ONLINE LLM PICKER, an active model selection strategy designed to efficiently identify the best candidate LLM in online settings. Given an incoming stream of prompts and a small annotation budget, ONLINE LLM PICKER selects the most informative prompts to annotate in order to reliably identify the best LLM for the stream. Our framework is based on the combined informativeness criterion that maximizes variance across model responses to unannotated prompts while also accounting for posterior uncertainty. ONLINE LLM PICKER makes no assumptions about the underlying LLMs, allowing it to be applied to any generative task and any model collection. ONLINE LLM PICKER is depicted in Figure 1.

We validate ONLINE LLM PICKER through a comprehensive set of experiments spanning multiple LLM generation tasks, including open-ended question answering (Rajpurkar et al., 2016), open-ended question answering with no correct response (Rajpurkar et al., 2018), grammar correction of natural language text (Loem et al., 2023), and both arithmetic (Gambardella et al., 2024) and calculus problems expressed in code-like notation (Gao et al., 2023). Our model collections across these tasks contain over 130 language models. We rank the models using both ROUGE (Lin, 2004) and embedding-based similarity measures (e.g. BERTScore (Zhang et al., 2020)), so that we can capture lexical and semantic agreement with the reference outputs. Our results demonstrate that ONLINE LLM PICKER identifies the best or near-best LLM in online environments while requiring up to 71.67% fewer annotations than competing baselines, and reduces regret by up to a factor of $2.51\times$ when applied to sequential generation on unannotated prompts across the stream.

2 RELATED WORK

A wide range of metrics are proposed for the **evaluation of language models**, from early automatic metrics such as ROUGE (Lin, 2004), BLEU (Papineni et al., 2002), and METEOR (Lavie & Agarwal, 2007) to benchmark-driven progress with GLUE (Wang et al., 2019), SuperGLUE (Wang et al., 2020), MMLU (Hendrycks et al., 2021), and BIG-bench (Srivastava et al., 2023). More recently, evaluation has shifted toward human feedback (Stiennon et al., 2022); (Ouyang et al., 2022), but reliance on annotators is costly and limits scalability in large-scale or streaming settings. LLM-as-a-judge methods (Gu et al. (2025); Bai et al. (2024); Chiang et al. (2024); Dubois et al. (2025)) offer more scalable pipelines, though they introduce bias. Efficiency is also identified as a central concern with Active Testing (Berrada et al. (2025)) which reduces annotation cost of evaluation through adaptive sampling. However, such methods focus on evaluating a fixed model or small set of models, whereas our setting requires identifying the best LLM for an arbitrary data stream under limited annotations.

The most relevant line of work for our setting is **active model selection**, which seeks to identify the best model from a set of candidates by querying informative examples (Madani et al., 2012; Karimi et al., 2021; Ashury-Tahan et al., 2024; Okanovic et al., 2024; Kay et al., 2025; Liu et al., 2022; Li et al., 2024a; Hara et al., 2024; Gardner et al., 2015) serving as the primary motivation for our work. However, most of these approaches assume a pool-based setup where all queries are available upfront(Okanovic et al., 2024; Ashury-Tahan et al., 2024; Kay et al., 2025) , or focus on

classification tasks(Karimi et al., 2021; Liu et al., 2022; Okanovic et al., 2024; Kay et al., 2025; Li et al., 2024a; Hara et al., 2024) , and therefore not directly applicable to our setting.

In this work, we introduce the novel problem of online active model selection for large language models and present several strategies as baselines (introduced in Section 4) along with ours. We relegate discussions of related but less directly relevant frameworks to Appendix A.

3 Online LLM Picker

3.1 Problem Statement and Background

Consider the inference phase with a stream of T prompts $\{x_t \in \mathcal{X} \mid t \in [T]\}$ drawn from an unknown distribution. Each prompt x_t is associated with a reference annotation $y_t \in \mathcal{Y}$, which remains hidden unless queried from an oracle.

We consider m pretrained language models $\mathcal{M} = \{f_i : \mathcal{X} \to \mathcal{Y} \mid i \in [m]\}$. For each prompt x_t, model i produces a response $f_i(x_t)$. At time step t, given the set of model responses $\{f_i(x_t) \mid i \in \mathcal{M}\}$, we decide whether to query the oracle for the reference annotation y_t or not. If y_t is not queried, it remains hidden; otherwise, we compute the loss vector as

$$\boldsymbol{\ell}_t^y = [\ell_{t,i}^y]_{i\in\mathcal{M}} \quad \text{where } \ell_{t,i}^y = 1 - d\big(y_t, f_i(x_t)\big) \tag{1}$$

for some similarity score $d : \mathcal{Y} \mapsto \mathcal{Y}$. In our framework, we choose $d(\cdot,\cdot)$ to be ROUGE-L (Lin, 2004) or BERTScore (Zhang et al., 2020) in our case (Rehman et al., 2025; Sul & Choi, 2023).

Given an annotation budget $b \ll T$ for the entire stream, our objective is to select the best language model for the remaining $T - b$ prompts using only the annotated prompts. At each time step t, our method returns the language model f_{best_t} based on the annotations observed up to time t such that $f_{\text{best}_t} \sim \mathbf{p}_t$. Our objective is to annotate the most informative examples across the stream so that f_{best_t} closely approximates the *true* best language model, defined as the model that incurs the minimum cumulative loss if all annotations in the stream were available: $f_{\text{best}} \coloneqq \arg\min_{i\in\mathcal{M}} \sum_{t=1}^{T} \ell_{t,i}^y$. Formally, we aim to minimize the cumulative loss of f_{best_t} relative to f_{best}, which is captured by the standard regret:

$$R_T = \sum_{t=1}^{T} \ell_{t,f_{\text{best}_t}}^y - \sum_{t=1}^{T} \ell_{t,f_{\text{best}}}^y. \tag{2}$$

Since our ultimate goal is to identify the best model under the annotation budget b, we also evaluate the quality of the returned model using two complementary metrics. The first is *identification probability*, which measures the probability of successfully recovering the true best model in a stochastic data stream. The second is *annotation efficiency*, which quantifies the annotation cost saved by identifying the best or near-best models under budget b. We provide their formal definitions in Section 4.

3.2 The Algorithm

In this section, we introduce Online LLM Picker, which aims to identify the best language model for a given stream of prompts with limited reference annotations from the oracle.

At a high level, our algorithm operates as follows. At each time step t, we decide whether to query the annotation of the prompt x_t via a random experiment: we draw a Bernoulli random variable $Q_t \sim \text{Bernoulli}(q_t)$, with *query probability* q_t, where the choice of q_t depends on factors introduced later. If a query is made ($Q_t = 1$), we update our posterior belief over the models using the newly observed annotation. We denote this posterior by $\mathbf{p}_t = [p_{t,i}]_{i\in\mathcal{M}}$ where $p_{t,i}$ represents the probability each model i at time t being the true best language model f_{best}. If no query is made ($Q_t = 0$), the reference annotation for x_t remains hidden and the posterior remains unchanged.

To perform the posterior update, we first construct an importance-weighted loss estimate using $\ell_{t,i}^y$ in Equation 1:

$$\hat{\ell}_{t,i}^y \;=\; \frac{\ell_{t,i}^y}{q_t}\, Q_t \quad \forall i \in \mathcal{M}$$

where $1/q_t$ serves as an importance weight, correcting sampling bias from the querying process by assigning greater weight to under-sampled instances.

We then obtain the cumulative loss estimate up to time t as

$$\hat{L}_{t,i} \;=\; \hat{L}_{t-1,i} \;+\; \hat{\ell}^{y}_{t,i} \quad \forall i \in \mathcal{M}$$

and update the posterior distribution $\mathbf{p}_t = [p_{t,i}]_{i\in\mathcal{M}}$ using the Exponential Weights (EW) algorithm by Littlestone & Warmuth (1994) with an adaptive learning rate η_t:

$$p_{t,i} \;\propto\; \exp\{-\eta_t \hat{L}_{t-1,i}\}. \tag{3}$$

Appendix E details the derivation of $\boldsymbol{p}_t$. We expand on our learning rate η_t as follows.

3.2.1 Adaptive Learning Rate

While time-based decay schedules are common (Karimi et al., 2021), inspired by AdaHedge (Rooij et al., 2013), we propose an adaptive learning rate η_t that updates based on observed annotations. In AdaHedge, the dynamic learning rate is defined in terms of the cumulative mixability gap, which measures the cumulative approximation error incurred when approximating the Hedge loss with the mix loss (Rooij et al., 2013). By means of Bernstein's bound, the mixability gap at a certain round t can be expressed in terms of variance of the losses. In addition, Rooij et al. (2013) shows that the cumulative mixability gap grows at most as the square root of the cumulative loss variance up to t. Motivated by this, we introduce an adaptive learning rate that depends on the variance of the losses over the queried examples. The variance of the loss can be computed as:

$$\operatorname*{Var}_{i\sim\boldsymbol{p}_t} \ell^{y}_{t,i} = \langle \boldsymbol{p}_t, \boldsymbol{\ell}^{y}_{t} \odot \boldsymbol{\ell}^{y}_{t} \rangle - (\langle \boldsymbol{p}_t, \boldsymbol{\ell}^{y}_{t} \rangle)^2.^{1} \tag{4}$$

Although variance-based strategies exist for adapting the learning rate (Rooij et al., 2013), they are less effective with limited annotations, where sparse queries and a few noisy examples can disproportionately distort the variance estimate. Following the idea of exponential moving averages in optimization (Kingma & Ba, 2017), which keeps exponential moving averages of the gradient and its square (first and second moments), applies bias correction, and uses them to form a per-parameter adaptive step, we instead introduce an exponential moving average of the variance of the observed examples as a proxy. This proxy does not track the exact variance of the underlying losses, but provides a smoothed, adaptive estimate that emphasizes recent observations while discounting stale ones, thus makes the learning rate more robust under sparse and noisy feedback.

We compute the exponential moving average of the variance as:

$$\mathcal{V}_t = \begin{cases} \beta \cdot \mathcal{V}_{t-1} + (1-\beta) \cdot \operatorname*{Var}_{i\sim\boldsymbol{p}_t} \ell^{y}_{t,i}, & \text{if } y_t \text{ is observed,} \\ \mathcal{V}_{t-1}, & \text{otherwise,} \end{cases} \tag{5}$$

where $\beta \in [0,1]$ is the decay rate of the exponential moving average. The effective window size of this exponential moving average is approximately $(1-\beta)^{-1}$, meaning that values of $\beta \approx 1$ place more weight on past observations and correspond to a longer memory.

At the beginning of the stream, we have $\mathcal{V}_0 = 0$ and no annotations are available yet. This initialization biases the variance proxy toward zero, particularly in the early time instances. As discussed in Kingma & Ba (2017), this bias can be corrected by a bias-adjusted proxy of the average variance:

$$\hat{\mathcal{V}}_t = \frac{\mathcal{V}_t}{1-\beta^t}. \tag{6}$$

To connect back to the framework of Rooij et al. (2013), which relates the cumulative mixability gap to the cumulative variance, we estimate the cumulative variance up to time t by scaling the number of time steps with the bias-adjusted proxy of the average variance with $t \cdot \hat{\mathcal{V}}_t$ (6). Eventually, at any time instance t, we compute the adaptive learning rate η_t as follows:

$$\eta_t = \sqrt{\frac{\log m}{t \cdot (\hat{\mathcal{V}}_t + \epsilon)}} \tag{7}$$

[1] Here, $\odot$ denotes the element-wise product and $\langle \cdot, \cdot \rangle$ denotes the inner product.

where $\epsilon > 0$ is a small constant added for numerical stability. This makes sure that η_t remains finite and well-defined for all t, including the case $\hat{\mathcal{V}}_t \to 0$, which corresponds to (near-) identical pairwise losses among the answers generated by language models. With a uniform prior over the m experts, the initial information cost of not knowing the best expert is $\log m$, which appears as $(\log m)/\eta$ in the regret bound (Rooij et al., 2013).

Plugging Equation 7 into Equation 3 completes the model posterior update. We now discuss the query probability q_t (10), which determines the outcome of the query decision at each time step t.

3.2.2 Query Probability

The query probability q_t at each time step is designed to capture our uncertainty regarding the identity of the best model on the given prompt x_t, and subsequently guide our decision of whether to request the annotation of x_t or not. Importantly, this probability should not only reflect predictive uncertainty but also integrate the current posterior belief over the language models, and balance the competing objectives of exploration and exploitation by doing so. A natural way to operationalize uncertainty in streaming settings is through variance-based measures(Karimi et al., 2021), which serve as a principled proxy for the potential information gain associated with querying an annotation. Intuitively, the goal is to estimate the expected value of acquiring the annotation for x_t, that is, the degree to which it would refine our estimate of the best model. More concretely, this quantity can be estimated from the losses incurred by the model responses $\{f_i(x_t) \mid i \in \mathcal{M}\}$, together with the posterior distribution over models $\boldsymbol{p}_t$ at time t. Since the reference annotation y_t is not available, this variance estimation is hypothetical. In the classification setting, such a hypothetical variance is well-defined because the true label at time t is assumed to lie within a finite set of predefined classes. In contrast, in the generative setting there is no fixed label space: the set of valid responses is open-ended and potentially unbounded, which makes evaluation and comparison substantially more challenging. To address this, we approximate the response space at each time t by considering the complete set of language model outputs $\{f_i(x_t) \mid i \in \mathcal{M}\}$.

Towards that, we denote the (pairwise) loss between the response generated by the language models and that generated by the model k by $\boldsymbol{\ell}_t^k = [\ell_{t,i}^k]_{i\in\mathcal{M}}$ where $\ell_{t,i}^k = 1 - d(f_i(x_t), f_k(x_t))$. At each time instance t, we treat the response of each language model $k \in \mathcal{M}$ as the hypothetical reference annotation. We then compute the maximum hypothetical variance among the losses with:

$$\max_{k\in\mathcal{M}} \operatorname*{Var}_{i\sim\boldsymbol{p}_t} \ell_{t,i}^k = \langle \boldsymbol{p}_t, \boldsymbol{\ell}_t^k \odot \boldsymbol{\ell}_t^k \rangle - (\langle \boldsymbol{p}_t, \boldsymbol{\ell}_t^k \rangle)^2 \tag{8}$$

which represents our uncertainty about the annotation of x_i given the model posterior $\boldsymbol{p}_t$. If the language model responses are nearly identical where $\boldsymbol{\ell}_t^k \approx 0$, or if they differ to the same degree such that $\boldsymbol{\ell}_t^k$ is similar for all $k \in \mathcal{M}$, then the dispersion under $\boldsymbol{p}_t$ is negligible. In this case, the annotation y_t provides little information, contributes minimally to model selection, and has only a negligible effect on regret. Conversely, if the maximum hypothetical variance is large, for instance, when $\boldsymbol{\ell}_t^k$ exhibits substantial heterogeneity across model responses, then the annotation y_t is informative.

Revisiting the model posterior, which our algorithm updates based on the observed annotations, we find that in some cases it concentrates around a single best model, whereas in others it remains diffuse across multiple candidates. When the posterior distribution is already concentrated after some annotations, additional annotations have very little impact on the posterior update. Otherwise, a few strategically chosen annotations can shift the balance and determine the best model. This perspective is rooted in the Bayesian experimental design, where the value of new information measured by its expected reduction in posterior uncertainty. Shannon entropy of the posterior provides a natural quantification of 'how much remains to be learned' about which model or parameter is best (Lindley, 1956; MacKay, 1992; Sebastiani & Wynn, 2025). A parallel also exists in online decision making: exploration is most valuable when the uncertainty expressed in terms of posterior entropy about the best choice is high, since this is when information can most effectively reduce future regret (Russo & Roy, 2016; 2017).

Combining multiple signals such as variance and entropy is a common design pattern in active learning and streaming settings, such as scaling committee disagreement by input density (McCallum & Nigam, 1998), weighting predictive entropy by local density (Zhu et al., 2008), querying when both uncertainty and density are high (Ienco et al., 2014), and multiplying predictive entropy by

a coverage-based representativeness factor (Katragadda et al., 2023). Following this principle, we incorporate posterior entropy into our query probability. Specifically, we compute the normalized entropy of the posterior distribution over models:

$$\bar{\mathbb{H}}(\boldsymbol{p}_t) \;=\; \frac{\mathbb{H}(\boldsymbol{p}_t)}{\log m} \in [0,1] \qquad \text{where } \mathbb{H}(\boldsymbol{p}_t) \;=\; -\sum_{i=1}^{m} p_{t,i}\,\log p_{t,i}. \tag{9}$$

We normalize entropy by $\log m$ to place uncertainty on a fixed $[0,1]$ scale, independent of the number of models m. This makes thresholds and query schedules comparable across settings and interpretable: $\bar{\mathbb{H}} = 0$ when one model dominates (high confidence) and $\bar{\mathbb{H}} = 1$ under a uniform posterior (maximal uncertainty). We then incorporate this normalized entropy by scaling the variance in 8 such that $\underset{i\sim\boldsymbol{p}_t}{\mathrm{Var}}\, \ell_{t,i}^k \cdot \bar{\mathbb{H}}(\boldsymbol{p}_t)$.

Finally, we define the query probability as

$$q_t = \begin{cases} \max\left\{\max_{k\in\mathcal{M}} \underset{i\sim\boldsymbol{p}_t}{\mathrm{Var}}\, \ell_{t,i}^k \cdot \bar{\mathbb{H}}(\boldsymbol{p}_t),\, \eta_t\right\}, & \text{if } \max_{k\in\mathcal{M}} \underset{i\sim\boldsymbol{p}_t}{\mathrm{Var}}\, \ell_{t,i}^k \neq 0 \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

where η_t is a time-decaying lower bound based on the adaptive learning rate. This bound prevents two problems: when variance is extremely small, importance-weighted loss estimates can become unstable and inflate regret; and when predictions are overly confident, the algorithm may skip prompts that are actually informative.

In essence, the query rule is based on the hypothetical loss variance across models under the current posterior as well as the direct posterior entropy. High variance and entropy trigger more queries, while a concentrated posterior reduces them. Early on, queries are mainly variance-driven, but over time q_t balances both factors and reflects model disagreement and overall uncertainty about the annotation of the incoming queries.

The pseudocode of ONLINE LLM PICKER is depicted in Algorithm 1.

Algorithm 1 Online LLM Picker

Require: Language models $\mathcal{M}$
  Set $\hat{L}_{0,i} = 0 \quad \forall\, i \in \mathcal{M},\ \mathcal{V}_0 = 0$
  for $t = 1, 2, ..., T$ **do**
    $\eta_t = \sqrt{\frac{\log m}{t\cdot(\hat{\mathcal{V}}_t + \epsilon)}}$
    $p_{t,i} \propto \exp\{-\eta_t \hat{L}_{t-1,i}\} \quad \forall\, i \in \mathcal{M}$ ▷ update posterior over experts
    Get models predictions $\{f_i(x_t) \mid i \in \mathcal{M}\}$ for x_t
    Recommend $f_{\text{best}_t} := \arg\max_{i\in\mathcal{M}} p_{t,i}$ as the best model up to time instance t
    Compute q_t as in *(10)* and sample $Q_t \sim \text{Bernoulli}(q_t)$
    if Q=1 **then** ▷ query the annotation y_t
      Update $\mathcal{V}_t$ as in *(5)*
      $\hat{L}_{t,i} = \hat{L}_{t-1,i} + \frac{\ell_{t,i}^y}{q_t} \quad \forall\, i \in \mathcal{M}$
    else
      Annotate x_t with $f_{\text{best}_t}(x_i)$
      $\hat{L}_{t,i} = \hat{L}_{t-1,i}$
    end if
  end for

Return: $f_{\text{best}} = \arg\max_{i\in\mathcal{M}} p_{i,T}$ ▷ return the best model for the stream

4 EXPERIMENTS

We evaluate ONLINE LLM PICKER for active LLM selection in streaming settings using several public generative datasets and model collections containing more than 130 language models. As this is the first study of its kind, we introduce baselines and compare performance across multiple

metrics: *identification probability*, *regret*, and *annotation efficiency* for identifying best or near-best models across the stream.

4.1 Datasets and Models

We benchmark Online LLM Picker against competing baselines across datasets from different generative tasks. First, we test the algorithms on LLMs collections for the SQuAD (Rajpurkar et al., 2016) and SQuAD v2 (Rajpurkar et al., 2018) datasets, which are designed for question answering tasks, with the latter including questions that may not have an answer in the given context. Question answering is often framed as a non-generative task, especially in the extractive setting where models directly copy answer spans from the input text. In contrast, our models generate answers in their own words, producing open-ended text instead of directly extracting it. We also use Comprehensive Arithmetic Problems dataset (Lee, 2024) that features a range of algebraic expressions, the Calculus Datasets (FDU) (Zhang, 2025) that consist of more advanced numerical and symbolic problems represented in LaTeX code, and the Grammar Correction dataset (Agentlans, 2024), which challenges models to fix grammar mistakes in English sentences. In addition, we also consider MT-Bench (Bai et al., 2024), a benchmark for evaluating multi-turn conversational ability of LLMs, and multiple datasets from the HELM Benchmark(Liang et al., 2023). In particular, we include two Longform Question Answering (LF-QA) from classic HELM, LF-QA Canonical and LF-QA Prompt, respectively. From MedHELM (Bedi et al., 2025) we add MedCalc (Khandekar et al., 2024) to our experiments. We also include FinQA(Chen et al., 2022) from HELM Finance.

As for the language models, we include several open-source and proprietary LLMs (Almazrouei et al. (2023), Radford et al. (2019), Brown et al. (2020), Raffel et al. (2023), OpenAI et al. (2024), DeepSeek-AI et al. (2025), Touvron et al. (2023), Grattafiori et al. (2024)), ranging from those with a few million parameters to those with several billion. Some of these models are fine-tuned on the specific datasets of interest, while others are general-purpose models suitable for a wide range of generative tasks. We also expand our model collection to DeepSeek (DeepSeek-AI et al., 2025) and various GPT-based models (Brown et al., 2020; Kocoń et al., 2023; Ouyang et al., 2022). An overview of our datasets and models are in Table 2 as well as LLM scores on the test datasets in Figure 4 in Appendix B. To simulate multiple LLMs using the same underlying model, we also apply prompt engineering techniques to instruction-tuned LLMs, encouraging them to behave in agent-specific ways. Prompt engineering is also used to improve performance on our benchmark datasets by reducing unnecessary verbosity and avoiding irrelevant explanations in the output. Our work is model-agnostic and makes no assumptions about LLM architectures or performance. To approximate real-world use, we evaluate across a diverse suite of datasets. More details on LLM collections and datasets can be found in Appendix B.

4.2 Baselines

We introduce several strategies and baselines and evaluate the performance of Online LLM Picker against them. These methods typically follow a coin-flipping strategy. At each time instance t, when a new prompt x_t is received, the decision to query the reference annotation y_t is made by sampling a Bernoulli random variable Q_t with bias q_t, which is usually adaptive. The reference annotation y_t is queried only if $Q_t = 1$.

For **Random** (passive learning) baseline, we query annotation of each time instance with a fixed probability $q_t = b/T$, having an expected number of b annotations queried over a stream of length T. For **Disagreement** baseline, we only consider prompts where the language models strongly disagree and use a fixed $q_t = b/T$ on those instances. Disagreement is quantified as the variance across experts in their average pairwise discrepancies, and a query is triggered when this measure exceeds a small threshold we introduced. We adapt the **Kullback–Leibler** baseline where KL queries with probability proportional to the Kullback–Leibler divergence (Shlens, 2014) between distribution of pairwise losses at time instance t and the posterior belief, to encourage annotations when the loss distribution deviates strongly from the learned belief. We adapt **Uncertainty**(Dagan & Engelson, 1995) to our setting to query annotations with probability similar to the Shannon entropy of the pairwise losses over candidate models. It estimates how evenly the models support competing hypotheses and regards higher entropy as greater uncertainty.

4.3 Evaluation Protocol

We evaluate Online LLM Picker separately on each dataset. For a given dataset, we generate a stream by drawing T i.i.d. instances uniformly at random and feeding them to each method, referring to each such stream as a *realization*. For each realization, we evaluate performance based on the language model returned by each method. We repeat this process over multiple independent realizations with fresh streams drawn from the test set, and average the results to estimate the expected performance for each metric.

We define the budget b as the maximum number of annotations that a method can query within a given realization. We evaluate the methods across different budget levels. For a fair comparison under the same budget, we tune the hyperparameter(s) of each method to query the same number of annotations in average at the end of the stream and compare their performance. Hyperparameter tuning adjusts the query probability of a method by multiplying it with different up- or down-scaling factors to control how many annotations each method ends up requesting in average across all realizations.

4.4 Performance Metrics

We evaluate each algorithm using the following metrics. First, for a fixed annotation budget, we compute the *regret* to measure how well the models returned by each method perform in sequential generation on the unannotated prompts across the stream. Second, we measure *identification probability*, defined as the fraction of realizations in which the best model is correctly identified by the end of the stream. Finally, we assess *annotation efficiency*, which captures how efficiently a method identifies the best or near-best language models relative to the number of annotations used.

4.5 Experimental Results

4.5.1 Regret

For a given budget where Online LLM Picker returns a model with high confidence, we report the expected regret averaged over all realizations. Our experiments are robust on the choice of budgets and we evaluate Online LLM Picker and baselines at very different budget levels. In all cases, Online LLM Picker consistently achieves a reduction in regret, in some cases reaching up to a factor of $2.51\times$ with respect to the best competing baseline, which shows its ability for sequential generation even well before exhausting its annotation budget. Details on annotation budget levels and further results are reported in Table 3 of Appendix D.

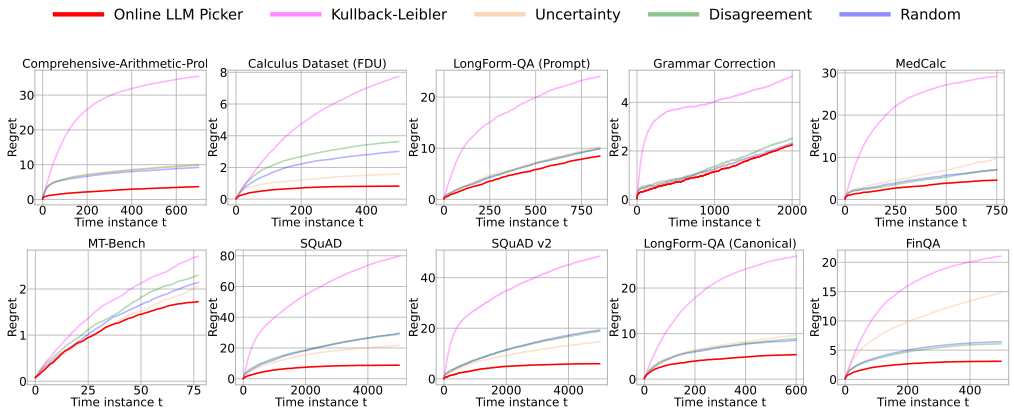

Figure 2: Regret for Online LLM Picker and baselines across 10 datasets.

4.5.2 Identification Probability

Figure 3 shows the identification probabilities for Online LLM Picker and the baselines. For each dataset, we extend the annotation budget until the best competing baseline achieves 100%

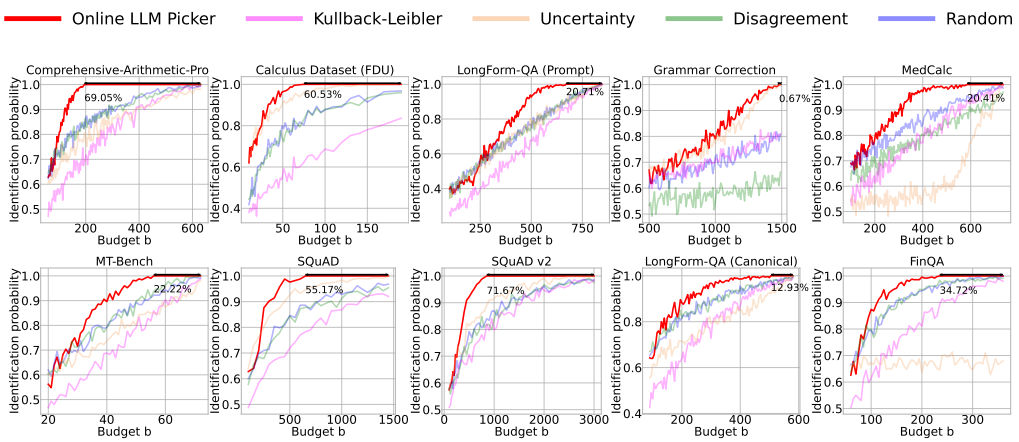

Figure 3: Identification probabilities of ONLINE LLM PICKER and baselines across different annotation budgets.

identification probability, and then report the percentage of annotation cost saved by ONLINE LLM PICKER for confidently identifying the best model. Across all datasets and annotation budgets, ONLINE LLM PICKER consistently requires the fewest annotations to reach maximum identification probability, showing its ability in correctly identifying the most informative examples to annotate, saving up to 71.67% with respect to the budget needed by the best competing baseline.

4.5.3 ANNOTATION EFFICIENCY

We examine the annotation cost of selecting a best or near-best model with accuracy within the δ-vicinity of that of the true best model across the entire stream. In particular, we focus on the required number of annotations where, in all realizations, the selected models are within 2.5%, 1%, 0.5%, and 0.1% of the true best model score. We evaluate the percentage reduction in annotation cost relative to the best competing baseline, reported separately for each dataset and each δ-vicinity. Our results in Table 1 show that ONLINE LLM PICKER significantly reduces up to 70.69% the number of annotations required to identify a near-best model whose accuracy is close to that of the best one. ONLINE LLM PICKER is consistently more annotation-efficient in identifying both the best and a near-best model across different tasks and datasets.

Dataset	δ=2.5%	δ=1.0%	δ=0.5%	δ=0.1%
Comprehensive–Arithmetic–Problems	↓ **24.39%**	↓ **64.89%**	↓ **68.00%**	↓ **69.05%**
Calculus Dataset (FDU)	↑ 28.00%	↓ **53.75%**	↓ **46.43%**	↓ **60.53%**
LongForm–QA (Prompt)	0.00%	↓ **27.70%**	↓ **25.79%**	↓ **20.96%**
Grammar Correction	0.00%	0.00%	0.00%	↓ **0.67%**
MedCalc	↓ **14.66%**	↓ **38.46%**	↓ **35.51%**	↓ **21.09%**
MT–Bench	↓ **22.22%**	↓ **22.22%**	↓ **22.22%**	↓ **22.22%**
SQuAD	↑ 20.00%	↓ **52.00%**	↓ **55.17%**	↓ **55.17%**
SQuAD v2	0.00%	↓ **33.33%**	↓ **63.04%**	↓ **70.69%**
LongForm-QA (Canonical)	↓ **30.30%**	↓ **27.00%**	↓ **33.93%**	↓ **23.28%**
FinQA	↓ **34.69%**	↓ **31.88%**	↓ **34.72%**	↓ **34.72%**

Table 1: Annotation efficiency for the near-best model.

5 DISCUSSIONS

We introduce the novel problem of online model selection for language models with limited annotation evidence, which is a technically challenging setting given the open-ended nature of generative outputs. We propose ONLINE LLM PICKER, a method tailored to this task that confidently identifies the best language model for the task in an annotation-efficient manner. As LLMs are increasingly deployed in domains where annotation budgets are limited and data distributions shift, ONLINE LLM PICKER enables adaptive model selection that reduces annotation costs, maintains performance under evolving conditions, and improves the robustness of real-world LLM deployment.

Ethics statement. We do not foresee ethical concerns arising from this work. We introduce a novel active model selection algorithm that performs well under budget constraints and compare it with established baselines in the literature. The study uses public benchmarks and does not involve human subjects or sensitive data.

Reproducibility statement. Our results are completely reproducible. We present the rigorous workflow in the paper and report experiments in the main text and the appendix. We provide the complete source code and scripts as supplementary material to reproduce all experiments.

REFERENCES

Agentlans. agentlans/grammar-correction · Datasets at Hugging Face, December 2024. URL https://huggingface.co/datasets/agentlans/grammar-correction.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. The Falcon Series of Open Language Models, November 2023. URL http://arxiv.org/abs/2311.16867. arXiv:2311.16867 [cs].

John Angelopoulos, Christos Manettas, and Kosmas Alexopoulos. Industrial Maintenance Optimization Based on the Integration of Large Language Models (LLM) and Augmented Reality (AR). In Kosmas Alexopoulos, Sotiris Makris, and Panagiotis Stavropoulos (eds.), *Advances in Artificial Intelligence in Manufacturing II*, pp. 197–205, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-86489-6. doi: 10.1007/978-3-031-86489-6_20.

Shir Ashury-Tahan, Ariel Gera, Benjamin Sznajder, Leshem Choshen, Liat Ein-Dor, and Eyal Shnarch. Label-Efficient Model Selection for Text Generation, June 2024. URL http://arxiv.org/abs/2402.07891. arXiv:2402.07891 [cs].

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and Wanli Ouyang. MT-Bench-101: A Fine-Grained Benchmark for Evaluating Large Language Models in Multi-Turn Dialogues. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7421–7454, 2024. doi: 10.18653/v1/2024.acl-long.401. URL http://arxiv.org/abs/2402.14762. arXiv:2402.14762 [cs].

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

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 41–48, Montreal Quebec Canada, June 2009. ACM. ISBN 978-1-60558-516-1. doi: 10.1145/1553374.1553380. URL https://dl.acm.org/doi/10.1145/1553374.1553380.

Gabrielle Berrada, Jannik Kossen, Muhammed Razzak, Freddie Bickford Smith, Yarin Gal, and Tom Rainforth. Scaling Up Active Testing to Large Language Models, August 2025. URL http://arxiv.org/abs/2508.09093. arXiv:2508.09093 [cs].

Anjanava Biswas and Wrick Talukdar. Intelligent Clinical Documentation: Harnessing Generative AI for Patient-Centric Clinical Note Generation. *International Journal of Innovative Science and Research Technology (IJISRT)*, pp. 994–1008, May 2024. ISSN 2456-2165. doi: 10.38124/ijisrt/IJISRT24MAY1483. URL http://arxiv.org/abs/2405.18346. arXiv:2405.18346 [cs].

Heloisa Oss Boll, Antonio Oss Boll, Leticia Puttlitz Boll, Ameen Abu Hanna, and Iacer Calixto. DistillNote: LLM-based clinical note summaries improve heart failure diagnosis, June 2025. URL http://arxiv.org/abs/2506.16777. arXiv:2506.16777 [cs].

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners, July 2020. URL http://arxiv.org/abs/2005.14165. arXiv:2005.14165 [cs].

Andrew Caines, Luca Benedetto, Shiva Taslimipoor, Christopher Davis, Yuan Gao, Oeistein Andersen, Zheng Yuan, Mark Elliott, Russell Moore, Christopher Bryant, Marek Rei, Helen Yannakoudakis, Andrew Mullooly, Diane Nicholls, and Paula Buttery. On the application of Large Language Models for language teaching and assessment technology, July 2023. URL http://arxiv.org/abs/2307.08393. arXiv:2307.08393 [cs].

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A Survey on Evaluation of Large Language Models, December 2023. URL http://arxiv.org/abs/2307.03109. arXiv:2307.03109 [cs].

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. FinQA: A Dataset of Numerical Reasoning over Financial Data, May 2022. URL http://arxiv.org/abs/2109.00122. arXiv:2109.00122 [cs].

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference, March 2024. URL http://arxiv.org/abs/2403.04132. arXiv:2403.04132 [cs].

Ido Dagan and Sean P. Engelson. Committee-based sampling for training probabilistic classifiers. In *Proceedings of the Twelfth International Conference on International Conference on Machine Learning*, ICML'95, pp. 150–157, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc. ISBN 978-1-55860-377-6.

Yuhao Dan, Zhikai Lei, Yiyang Gu, Yong Li, Jianghao Yin, Jiaju Lin, Linhao Ye, Zhiyan Tie, Yougen Zhou, Yilei Wang, Aimin Zhou, Ze Zhou, Qin Chen, Jie Zhou, Liang He, and Xipeng Qiu. EduChat: A Large-Scale Language Model-based Chatbot System for Intelligent Education, August 2023. URL http://arxiv.org/abs/2308.02773. arXiv:2308.02773 [cs].

DeepSeek. deepseek-ai/DeepSeek-V3.1 · Hugging Face, August 2025. URL https://huggingface.co/deepseek-ai/DeepSeek-V3.1.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J.

Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. DeepSeek-V3 Technical Report, February 2025. URL http://arxiv.org/abs/2412.19437. arXiv:2412.19437 [cs].

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A Survey on In-context Learning, October 2024. URL http://arxiv.org/abs/2301.00234. arXiv:2301.00234 [cs].

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-Controlled AlpacaEval: A Simple Way to Debias Automatic Evaluators, March 2025. URL http://arxiv.org/abs/2404.04475. arXiv:2404.04475 [cs].

Falcon. tiiuae/falcon-7b-instruct · Hugging Face, June 2023. URL https://huggingface.co/tiiuae/falcon-7b-instruct.

Lizhou Fan, Lingyao Li, Zihui Ma, Sanggyu Lee, Huizi Yu, and Libby Hemphill. A Bibliometric Review of Large Language Models Research from 2017 to 2023, April 2023. URL http://arxiv.org/abs/2304.02020. arXiv:2304.02020 [cs].

Simin Fan and Martin Jaggi. Irreducible Curriculum for Language Model Pretraining, October 2023. URL http://arxiv.org/abs/2310.15389. arXiv:2310.15389 [cs].

Andrew Gambardella, Yusuke Iwasawa, and Yutaka Matsuo. Language Models Do Hard Arithmetic Tasks Easily and Hardly Do Easy Arithmetic Tasks, June 2024. URL http://arxiv.org/abs/2406.02356. arXiv:2406.02356 [cs].

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. PAL: Program-aided Language Models, January 2023. URL http://arxiv.org/abs/2211.10435. arXiv:2211.10435 [cs].

Jacob Gardner, Gustavo Malkomes, Roman Garnett, Kilian Q Weinberger, Dennis Barbour, and John P Cunningham. Bayesian Active Model Selection with an Application to Automated Audiometry. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://papers.nips.cc/paper_files/paper/2015/hash/d9731321ef4e063ebbee79298fa36f56-Abstract.html.

Google-t5. google-t5/t5-small · Hugging Face, March 2024. URL https://huggingface.co/google-t5/t5-small.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret,

Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang,

Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The Llama 3 Herd of Models, November 2024. URL http://arxiv.org/abs/2407.21783. arXiv:2407.21783 [cs].

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A Survey on LLM-as-a-Judge, March 2025. URL http://arxiv.org/abs/2411.15594. arXiv:2411.15594 [cs].

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan H. Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models, August 2023. URL http://arxiv.org/abs/2308.11462. arXiv:2308.11462 [cs].

Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Supryadi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, and Deyi Xiong. Evaluating Large Language Models: A Comprehensive Survey, November 2023. URL http://arxiv.org/abs/2310.19736. arXiv:2310.19736 [cs].

Satoshi Hara, Mitsuru Matsuura, Junya Honda, and Shinji Ito. Active model selection: A variance minimization approach. *Machine Learning*, 113(11):8327–8345, December 2024. ISSN 1573-0565. doi: 10.1007/s10994-024-06603-1. URL https://doi.org/10.1007/s10994-024-06603-1.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring Massive Multitask Language Understanding, January 2021. URL http://arxiv.org/abs/2009.03300. arXiv:2009.03300 [cs].

Dino Ienco, Indrė Žliobaitė, and Bernhard Pfahringer. High density-focused uncertainty sampling for active learning over evolving stream data. In *Proceedings of the 3rd International Workshop*

on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications, pp. 133–148. PMLR, August 2014. URL https://proceedings.mlr.press/v36/ienco14.html. ISSN: 1938-7228.

Mohammad Reza Karimi, Nezihe Merve Gürel, Bojan Karlaš, Johannes Rausch, Ce Zhang, and Andreas Krause. Online Active Model Selection for Pre-trained Classifiers, April 2021. URL http://arxiv.org/abs/2010.09818. arXiv:2010.09818 [cs].

Sai Prathyush Katragadda, Tyler Cody, Peter Beling, and Laura Freeman. Active Learning with Combinatorial Coverage, February 2023. URL http://arxiv.org/abs/2302.14567. arXiv:2302.14567 [cs].

Justin Kay, Grant Van Horn, Subhransu Maji, Daniel Sheldon, and Sara Beery. Consensus-Driven Active Model Selection, July 2025. URL http://arxiv.org/abs/2507.23771. arXiv:2507.23771 [cs].

Nikhil Khandekar, Qiao Jin, Guangzhi Xiong, Soren Dunn, Serina S. Applebaum, Zain Anwar, Maame Sarfo-Gyamfi, Conrad W. Safranek, Abid A. Anwar, Andrew Zhang, Aidan Gilson, Maxwell B. Singer, Amisha Dave, Andrew Taylor, Aidong Zhang, Qingyu Chen, and Zhiyong Lu. MedCalc-Bench: Evaluating Large Language Models for Medical Calculations, June 2024. URL http://arxiv.org/abs/2406.12036. arXiv:2406.12036 [cs].

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization, January 2017. URL http://arxiv.org/abs/1412.6980. arXiv:1412.6980 [cs].

Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, Anna Kocoń, Bartłomiej Koptyra, Wiktoria Mieleszczenko-Kowszewicz, Piotr Miłkowski, Marcin Oleksy, Maciej Piasecki, Łukasz Radliński, Konrad Wojtasik, Stanisław Woźniak, and Przemysław Kazienko. ChatGPT: Jack of all trades, master of none. *Information Fusion*, 99:101861, November 2023. ISSN 15662535. doi: 10.1016/j.inffus.2023.101861. URL http://arxiv.org/abs/2302.10724. arXiv:2302.10724 [cs].

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large Language Models are Zero-Shot Reasoners, January 2023. URL http://arxiv.org/abs/2205.11916. arXiv:2205.11916 [cs].

Ibrahim Kok, Orhan Demirci, and Suat Ozdemir. When IoT Meet LLMs: Applications and Challenges, November 2024. URL http://arxiv.org/abs/2411.17722. arXiv:2411.17722 [cs].

Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Tom Rainforth. Active Testing: Sample-Efficient Model Evaluation. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 5753–5763. PMLR, July 2021. URL https://proceedings.mlr.press/v139/kossen21a.html. ISSN: 2640-3498.

Alon Lavie and Abhaya Agarwal. Meteor: an automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pp. 228–231, USA, 2007. Association for Computational Linguistics.

Garreth Lee. garrethlee/comprehensive-arithmetic-problems · Datasets at Hugging Face, December 2024. URL https://huggingface.co/datasets/garrethlee/comprehensive-arithmetic-problems.

Jieh-Sheng Lee. LexGPT 0.1: pre-trained GPT-J models with Pile of Law, June 2023. URL http://arxiv.org/abs/2306.05431. arXiv:2306.05431 [cs].

Po-han Li, Oyku Selin Toprak, Aditya Narayanan, Ufuk Topcu, and Sandeep Chinchali. Online Foundation Model Selection in Robotics, February 2024a. URL http://arxiv.org/abs/2402.08570. arXiv:2402.08570 [cs].

Yiwei Li, Huaqin Zhao, Hanqi Jiang, Yi Pan, Zhengliang Liu, Zihao Wu, Peng Shu, Jie Tian, Tianze Yang, Shaochen Xu, Yanjun Lyu, Parker Blenk, Jacob Pence, Jason Rupram, Eliza Banu, Ninghao Liu, Linbing Wang, Wenzhan Song, Xiaoming Zhai, Kenan Song, Dajiang Zhu, Beiwen Li, Xianqiao Wang, and Tianming Liu. Large Language Models for Manufacturing, October 2024b. URL http://arxiv.org/abs/2410.21418. arXiv:2410.21418 [cs] version: 1.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic Evaluation of Language Models, October 2023. URL http://arxiv.org/abs/2211.09110. arXiv:2211.09110 [cs].

Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

D. V. Lindley. On a Measure of the Information Provided by an Experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, December 1956. ISSN 0003-4851. doi: 10.1214/aoms/1177728069. URL http://projecteuclid.org/euclid.aoms/1177728069.

N. Littlestone and M. K. Warmuth. The Weighted Majority Algorithm. *Information and Computation*, 108(2):212–261, February 1994. ISSN 0890-5401. doi: 10.1006/inco.1994.1009. URL https://www.sciencedirect.com/science/article/pii/S0890540184710091.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing, July 2021. URL http://arxiv.org/abs/2107.13586. arXiv:2107.13586 [cs].

Xuefeng Liu, Fangfang Xia, Rick L. Stevens, and Yuxin Chen. Contextual Active Model Selection, July 2022. URL http://arxiv.org/abs/2207.06030. arXiv:2207.06030 [cs].

Mengsay Loem, Masahiro Kaneko, Sho Takase, and Naoaki Okazaki. Exploring Effectiveness of GPT-3 in Grammatical Error Correction: A Study on Performance and Controllability in Prompt-Based Methods, May 2023. URL http://arxiv.org/abs/2305.18156. arXiv:2305.18156 [cs].

David J. C. MacKay. Information-Based Objective Functions for Active Data Selection. *Neural Computation*, 4(4):590–604, July 1992. ISSN 0899-7667. doi: 10.1162/neco.1992.4.4.590. URL https://doi.org/10.1162/neco.1992.4.4.590.

Omid Madani, Daniel J. Lizotte, and Russell Greiner. Active Model Selection, July 2012. URL http://arxiv.org/abs/1207.4138. arXiv:1207.4138 [cs].

Andrew McCallum and Kamal Nigam. Employing EM and Pool-Based Active Learning for Text Classification. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pp. 350–358, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc. ISBN 978-1-55860-556-5.

Meta-Llama. meta-llama/Llama-3.1-8B · Hugging Face, December 2024. URL https://huggingface.co/meta-llama/Llama-3.1-8B.

MistralAI. mistralai/Mistral-7B-Instruct-v0.3 · Hugging Face. URL https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3.

Nvidia. nvidia/NVIDIA-Nemotron-Nano-9B-v2 · Hugging Face, September 2025. URL https://huggingface.co/nvidia/NVIDIA-Nemotron-Nano-9B-v2.

Patrik Okanovic, Andreas Kirsch, Jannes Kasper, Torsten Hoefler, Andreas Krause, and Nezihe Merve Gürel. All models are wrong, some are useful: Model Selection with Limited Labels, October 2024. URL http://arxiv.org/abs/2410.13609. arXiv:2410.13609 [cs].

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, C. J. Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. GPT-4 Technical Report, March 2024. URL http://arxiv.org/abs/2303.08774. arXiv:2303.08774 [cs].

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, March 2022. URL http://arxiv.org/abs/2203.02155. arXiv:2203.02155 [cs].

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a Method for Automatic Evaluation of Machine Translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin (eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinyBenchmarks: evaluating LLMs with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *ICML'24*, pp. 34303–34326, Vienna, Austria, 2024. JMLR.org.

Thiago Dal Pont, Federico Galli, Andrea Loreggia, Giuseppe Pisano, Riccardo Rovatti, and Giovanni Sartor. Legal Summarisation through LLMs: The PRODIGIT Project, August 2023. URL http://arxiv.org/abs/2308.04416. arXiv:2308.04416 [cs].

Hadi Pouransari, Chun-Liang Li, Jen-Hao Rick Chang, Pavan Kumar Anasosalu Vasu, Cem Koc, Vaishaal Shankar, and Oncel Tuzel. Dataset Decomposition: Faster LLM Training with Variable Sequence Length Curriculum, January 2025. URL http://arxiv.org/abs/2405.13226. arXiv:2405.13226 [cs].

Qwen. Qwen/Qwen3-Next-80B-A3B-Instruct · Hugging Face, September 2025. URL https://huggingface.co/Qwen/Qwen3-Next-80B-A3B-Instruct.

Alec Radford, Jeff Wu, R. Child, D. Luan, Dario Amodei, and I. Sutskever. Language Models are Unsupervised Multitask Learners. 2019. URL https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14dfe.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, September 2023. URL http://arxiv.org/abs/1910.10683. arXiv:1910.10683 [cs].

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ Questions for Machine Comprehension of Text, October 2016. URL http://arxiv.org/abs/1606.05250. arXiv:1606.05250 [cs].

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know What You Don't Know: Unanswerable Questions for SQuAD, June 2018. URL http://arxiv.org/abs/1806.03822. arXiv:1806.03822 [cs].

Tohida Rehman, Soumabha Ghosh, Kuntal Das, Souvik Bhattacharjee, Debarshi Kumar Sanyal, and Samiran Chattopadhyay. Evaluating LLMs and Pre-trained Models for Text Summarization Across Diverse Datasets, March 2025. URL http://arxiv.org/abs/2502.19339. arXiv:2502.19339 [cs].

Steven de Rooij, Tim van Erven, Peter D. Grünwald, and Wouter M. Koolen. Follow the Leader If You Can, Hedge If You Must, January 2013. URL http://arxiv.org/abs/1301.0534. arXiv:1301.0534 [cs].

Daniel Russo and Benjamin Van Roy. An Information-Theoretic Analysis of Thompson Sampling. *Journal of Machine Learning Research*, 17(68):1–30, 2016. ISSN 1533-7928. URL http://jmlr.org/papers/v17/14-087.html.

Daniel Russo and Benjamin Van Roy. Learning to Optimize via Information-Directed Sampling, July 2017. URL http://arxiv.org/abs/1403.5556. arXiv:1403.5556 [cs].

Paola Sebastiani and Henry P Wynn. Maximum Entropy Sampling and Optimal Bayesian Experimental Design | Request PDF. *ResearchGate*, August 2025. ISSN 1467-9868. URL https://www.researchgate.net/publication/4771982_Maximum_Entropy_Sampling_and_Optimal_Bayesian_Experimental_Design.

Yunfan Shao, Linyang Li, Zhaoye Fei, Hang Yan, Dahua Lin, and Xipeng Qiu. Balanced Data Sampling for Language Model Training with Clustering, June 2024. URL http://arxiv.org/abs/2402.14526. arXiv:2402.14526 [cs].

Jonathon Shlens. Notes on Kullback-Leibler Divergence and Likelihood, April 2014. URL http://arxiv.org/abs/1404.2000. arXiv:1404.2000 [cs].

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan A. Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar,

Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models, June 2023. URL http://arxiv.org/abs/2206.04615. arXiv:2206.04615 [cs].

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback, February 2022. URL http://arxiv.org/abs/2009.01325. arXiv:2009.01325 [cs].

Jeewoo Sul and Yong Suk Choi. Balancing Lexical and Semantic Quality in Abstractive Summarization, May 2023. URL http://arxiv.org/abs/2305.09898. arXiv:2305.09898 [cs].

Wang-Chiew Tan. Unstructured and structured data: Can we have the best of both worlds with large language models?, July 2023. URL http://arxiv.org/abs/2304.13010. arXiv:2304.13010 [cs].

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and Efficient Foundation Language Models, February 2023. URL http://arxiv.org/abs/2302.13971. arXiv:2302.13971 [cs].

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerova, Nidhi Rohatgi, Poonam Hosamani, William Collins, Neera Ahuja, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, John Pauly, and Akshay S. Chaudhari. Adapted Large Language Models Can Outperform Medical Experts in Clinical Text Summarization. *Nature Medicine*, 30(4): 1134–1142, April 2024. ISSN 1078-8956, 1546-170X. doi: 10.1038/s41591-024-02855-5. URL http://arxiv.org/abs/2309.07430. arXiv:2309.07430 [cs].

Rajan Vivek, Kawin Ethayarajh, Diyi Yang, and Douwe Kiela. Anchor Points: Benchmarking Models with Much Fewer Examples. In Yvette Graham and Matthew Purver (eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1576–1601, St. Julian's, Malta, March 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.eacl-long.95. URL https://aclanthology.org/2024.eacl-long.95/.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding, February 2019. URL http://arxiv.org/abs/1804.07461. arXiv:1804.07461 [cs].

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems, February 2020. URL http://arxiv.org/abs/1905.00537. arXiv:1905.00537 [cs].

Zhenting Wang, Guofeng Cui, Yu-Jhe Li, Kun Wan, and Wentian Zhao. DUMP: Automated Distribution-Level Curriculum Learning for RL-based LLM Post-training, June 2025. URL http://arxiv.org/abs/2504.09710. arXiv:2504.09710 [cs].

Di Zhang. di-zhang-fdu/calculus-dataset · Datasets at Hugging Face, January 2025. URL https://huggingface.co/datasets/di-zhang-fdu/calculus-dataset.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating Text Generation with BERT, February 2020. URL http://arxiv.org/abs/1904.09675. arXiv:1904.09675 [cs].

Yang Zhang, Amr Mohamed, Hadi Abdine, Guokan Shang, and Michalis Vazirgiannis. Beyond Random Sampling: Efficient Language Model Pretraining via Curriculum Learning, June 2025. URL http://arxiv.org/abs/2506.11300. arXiv:2506.11300 [cs].

Jingbo Zhu, Huizhen Wang, Tianshun Yao, and Benjamin K. Tsou. Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pp. 1137–1144, USA, 2008. Association for Computational Linguistics. ISBN 978-1-905593-44-6.

A Related Work

Data sampling methods for LLMs Data sampling has long been an important topic in machine learning, and with the rise of LLMs it has gained renewed attention. Recent research explores diverse strategies for selecting and ordering training data to improve efficiency, scalability, and final model quality. A first branch of approaches follows *curriculum learning*(Bengio et al., 2009), where examples are ordered from easy to hard to smooth optimization. Zhang et al. (2025) ranks data incrementally by means of different metrics for Language Models pretraining, Fan & Jaggi (2023) quantifies example learnability to prioritize those that may be most beneficial and Pouransari et al. (2025) uses variable sequence length and batch-size to improve long-context modeling. Wang et al. (2025) applies a similar approach for in-context learning. Other works are draw on *importance or loss-based sampling*, where training examples are weighted by estimated contribution to learning progress. At large scale, related ideas appear in corpus-level mixture design, such as the fixed data-mixture weights used to balance sources in GPT-3 (Brown et al., 2020), which adjust the proportions of different domains to achieve a desired training distribution. Some methods address corpus in a heterogeneous way through balanced or cluster-based selection, which preserves diversity by balancing common and rare examples to improve model training(). Recently, Shao et al. (2024) introduces *distribution-level curriculum learning* for LLM post-training, dynamically balancing exploration and exploitation across reward-conditioned datasets. While these approaches significantly reduce training cost or improve data efficiency, they generally assume offline access to the full corpus or slowly changing data distributions. They thus differ from our setting, which requires adaptive decisions on streaming data with a limited annotation budget.

B Datasets and LLM collections

More than 130 LLMs over 10 different datasets have been employed including different generative tasks. Except for LLMs of MT-Bench, which were evaluated using BERTScore (Zhang et al., 2020), the performance of the LLMs on the test sets of the other datasets was assessed using ROUGE-L (Lin, 2004).

Dataset	No. of instances	No. of LLMs	LLM scores
Comprehensive–Arithmetic–Problems	700	13	0.17 - 0.63
Calculus Dataset (FDU)	500	5	0.84 - 0.95
LongForm–QA (Prompt)	850	11	0.02 - 0.54
Grammar Correction	2000	5	0.76 - 0.84
MedCalc	750	9	0.12 - 0.41
MT–Bench	78	6	0.76 - 0.79
SQuAD	5000	7	0.47 - 0.93
SQuAD v2	5000	5	0.39 - 0.77
LongForm–QA (Canonical)	600	54	0.16 - 0.66
FinQA	500	22	0.39 - 0.81

Table 2: Overview of datasets and LLM collections, including number of models and the scores achieved on the test set. The number of instances represents the stream length.

Figure 4 shows the distribution of LLM scores across datasets. Depending on the task, scores are computed with ROUGE-L (Lin, 2004) or BERTScore (Zhang et al., 2020). We evaluate a diverse collection of LLMs, yielding varied score distributions and enabling a comprehensive assessment. As reported in Table 2, the observed scores span 0.02–0.97 across models and datasets. Where available, we use benchmarks that release both LLM outputs and oracle annotations. In contrast, for the Calculus Dataset (FDU) (Zhang, 2025), Grammar Correction (Agentlans, 2024), Comprehensive Arithmetic Problems (Lee, 2024), SQuAD (Rajpurkar et al., 2016), and SQuAD v2 (Rajpurkar et al., 2018), only oracle annotations are provided; therefore we ran LLMs ourselves to generate the model outputs. As detailed in Section 4, this includes instruction-tuned LLMs (to reduce output noise and elicit agent-specific behavior), models fine-tuned on the corresponding datasets, and off-the-shelf

general LLMs. Our study is model-agnostic: we make no assumptions about LLM architectures or task-specific use cases and assess models in a manner consistent with realistic practitioner workflows.

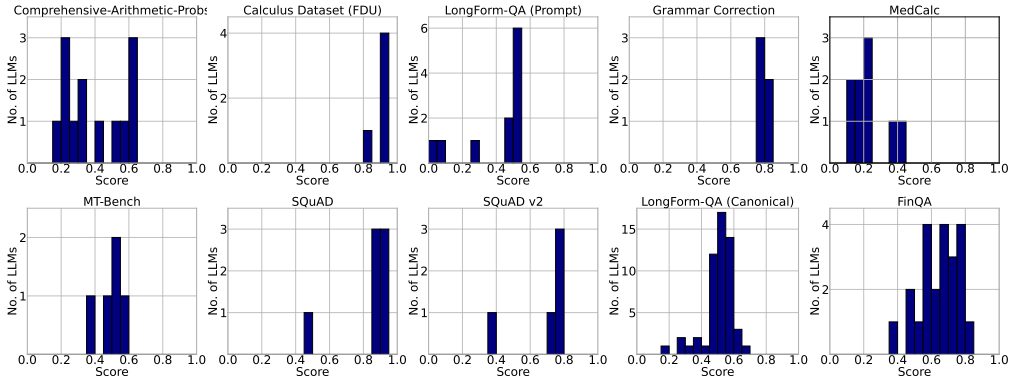

Figure 4: Score of LLM collections across the datasets.

C Baselines

Disagreement Similar to the Random algorithm, the probability of querying the annotation at each time instance is fixed to $q_t = b/T$. However, we restrict our attention to instances where there is significant disagreement among experts. Specifically, cases in which the answers generated by the LLMs differ substantially from one another. We quantify disagreement at time instance t as the variance—across experts—of each expert's average pairwise discrepancy from the others in the set of generated answers. Denote this statistic by Ξ_t. An instance is deemed to exhibit disagreement whenever Ξ_t exceeds a small near-zero threshold δ.

Kullback–Leibler (KL) We maintain a belief $\boldsymbol{p}_t$ over the m models from past annotations (Hedge-style posterior). At each time instance t, we turn the mean of the pairwise losses among the answers generated by the LLMs into a normalized predictive loss distribution $\tilde{p}_t$ (via a softmax over the losses) and measure its mismatch from our belief with $D_t = \mathrm{KL}(\tilde{p}_t \,\|\, p_t)$ (Shlens, 2014). The query probability is set proportional to D_t, so we are more likely to annotate when the present loss pattern looks uncertain under our belief and less likely when LLMs agree. We sample a Bernoulli query with bias D_t and spend from a fixed budget b until exhaustion.

Uncertainty At each time instance t, we form a per–model average loss profile p_t by averaging the current (unlabeled) pairwise prediction losses across models. We then compute an entropy–like uncertainty score $H_t = -\frac{1}{m}\sum_i p_{t,i}\log(p_{t,i})$ and set the query probability q_t proportional to H_t. Thus, instances that bring higher uncertainty are queried more often. We sample a Bernoulli query with bias H_t and spend from a fixed budget b until exhaustion.

D Extended Results

This section presents additional results and comprehensive data from the experiments described in Section 4. Separately for each dataset and for each budget level, we conduct 500 independent realizations, ensuring robust and reliable results.

D.1 Regret

We evaluate the regret of Online LLM Picker and baselines at different budget levels where our method identifies the best LLM with high confidence. Table 3 reports the budget levels used for each dataset. Across datasets, Online LLM Picker attains consistently lower regret than the baselines. For each dataset, we compare Online LLM Picker against the strongest competing baseline at the same budget and report the reduction factor (best-baseline regret divided by Online LLM Picker regret). As shown in Table 3, Online LLM Picker achieves regret reductions of

$2.51\times$, $2.47\times$, $2.44\times$, and $1.97\times$ on Comprehensive–Arithmetic–Problems, SQuAD, SQuAD v2, and FinQA, respectively.

Dataset	Budget level	Regret
Comprehensive–Arithmetic–Problems	120	$\downarrow \mathbf{2.51}\times$
Calculus Dataset (FDU)	65	$\downarrow \mathbf{1.93}\times$
LongForm–QA (Prompt)	450	$\downarrow \mathbf{1.17}\times$
Grammar Correction	1200	$\downarrow \mathbf{1.01}\times$
MedCalc	400	$\downarrow \mathbf{1.54}\times$
MT–Bench	50	$\downarrow \mathbf{1.19}\times$
SQuAD	450	$\downarrow \mathbf{2.47}\times$
SQuAD v2	800	$\downarrow \mathbf{2.44}\times$
LongForm–QA (Canonical)	300	$\downarrow \mathbf{1.58}\times$
FinQA	200	$\downarrow \mathbf{1.97}\times$

Table 3: Budget levels and regret analysis.

Dataset	Online LLM Picker	Kullback–Leibler	Uncertainty	Disagreement	Random
Comprehensive-Arithmetic-Problems	**3.67**	35.39	10.09	9.89	<u>9.22</u>
Calculus Dataset	**0.82**	7.75	<u>1.58</u>	3.62	3.01
LongForm-QA (Prompt)	**8.45**	24.06	10.31	<u>9.88</u>	9.97
Grammar Correction	**2.25**	5.07	<u>2.27</u>	2.52	2.31
MedCalc	**4.57**	29.18	9.67	<u>7.03</u>	7.07
MT-Bench	**1.72**	2.71	<u>2.06</u>	2.29	2.14
SQuAD	**8.78**	79.70	<u>21.67</u>	28.93	29.58
SQuAD v2	**5.98**	48.45	<u>14.58</u>	18.83	19.31
LongForm–QA (Canonical)	**5.39**	26.98	9.63	8.88	<u>8.53</u>
FinQA	**3.10**	21.04	14.83	<u>6.10</u>	6.44

Table 4: Regret of ONLINE LLM PICKER and baselines. Bold denotes the best value, underlining denotes the second-best. The regret of ONLINE LLM PICKER is consistently lower with respect to baselines.

D.2 IDENTIFICATION PROBABILITY

An overview of the percentage of budget needed for each method to reach the 100% identification probability for the first time can be found in Table 5. Note that the percentage of budget for each method is now computed with respect to the total length of the stream. ONLINE LLM PICKER consistently requires less annotations compared to baselines to reach the maximum identification probability, showing its effectiveness in properly selecting the most informative annotations to improve the selection strategy.

Dataset	Online LLM Picker	Kullback–Leibler	Uncertainty	Disagreement	Random
Comprehensive-Arithmetic-Problems	**28.26**%	97.10%	100.00%	<u>91.30</u>%	95.65%
Calculus Dataset	**15.00**%	<u>80.00</u>%	38.00%	<u>80.00</u>%	90.00%
LF-QA-Prompt	**78.82**%	100.00%	<u>99.41</u>%	100.00%	<u>99.41</u>%
Grammar Correction	**74.00**%	98.00%	<u>74.50</u>%	100.00%	98.00%
MedCalc	**78.00**%	100.00%	<u>98.00</u>%	100.00%	100.00%
MT-Bench	**71.79**%	96.15%	98.71%	<u>92.30</u>%	93.58%
SQuAD	**13.00**%	90.00%	<u>29.00</u>%	80.00%	58.00%
SQuAD v2	**17.00**%	100.00%	<u>60.00</u>%	90.00%	80.00%
LongForm–QA (Canonical)	**84.16**%	100.00%	98.33%	<u>96.67</u>%	99.17%
FinQA	**47.00**%	88.00%	100.00%	84.00%	<u>72.00</u>%

Table 5: Percentage of annotations (w.r.t. total stream length) required by ONLINE LLM PICKER and baselines to reach 100% identification probability. Bold denotes the best value, underlining denotes the second-best.

D.3 Robustness Analysis

We compute the 95-th percentile accuracy gap at budget needed by Online LLM Picker to reach certain identification probability levels. Specifically, we focus on 70%, 80%, 90% and 100% identification probabilities. If the exact desired identification probability value is unavailable, the next higher closest value is used. Table 6 shows the results. Best values are in bold, second-best values are underlined. In 36 cases out of 40, Online LLM Picker achieves the lowest 95-th percentile accuracy gap. However, even for the four remaining cases, Online LLM Picker performs competitively, achieving the second lowest accuracy gap.

Dataset Identification probability	Online LLM Picker (70% / 80% / 90% / 100%)	Kullback–Leibler (70% / 80% / 90% / 100%)	Uncertainty (70% / 80% / 90% / 100%)	Disagreement (70% / 80% / 90% / 100%)	Random (70% / 80% / 90% / 100%)
Comprehensive-Arithmetic-Problems	1.81 / **1.33** / **0.86** / **0.00**	3.71 / 2.67 / 2.57 / 2.19	2.10 / 1.71 / 1.71 / 1.43	**1.62** / 1.62 / 1.33 / 1.43	1.81 / 1.71 / 1.52 / 1.24
Calculus Dataset	1.12 / **0.91** / **0.66** / **0.00**	3.11 / 3.03 / 2.85 / 2.76	**1.02** / 0.95 / 0.87 / **0.00**	2.92 / 2.65 / 1.22 / 0.88	2.81 / 2.46 / 1.10 / 0.91
LongForm–QA (Prompt)	**1.45** / **1.23** / **0.63** / **0.00**	2.89 / 1.82 / 1.68 / 0.98	1.72 / 1.43 / 1.35 / 0.85	1.53 / 1.54 / 1.26 / 0.69	1.59 / 1.37 / 1.30 / 0.65
Grammar Correction	**0.11** / **0.11** / **0.11** / **0.00**	**0.11** / **0.11** / **0.11** / 0.11	**0.11** / **0.11** / **0.11** / **0.00**	**0.11** / **0.11** / **0.11** / 0.11	**0.11** / **0.11** / **0.11** / 0.11
MedCalc	2.13 / **1.84** / **0.99** / **0.00**	4.40 / 2.62 / 2.24 / 0.48	2.55 / 2.45 / 2.44 / 1.59	2.37 / 2.19 / 1.99 / 0.89	**2.10** / 1.99 / 1.79 / **0.00**
MT-Bench	**6.61** / **3.86** / **3.28** / **0.00**	7.59 / 7.50 / 6.73 / 3.79	7.03 / 4.78 / 3.92 / 3.71	6.93 / 6.42 / 3.83 / 2.82	6.99 / 4.61 / 3.86 / 3.01
SQuAD	1.69 / **1.33** / **1.12** / **0.00**	2.05 / 1.95 / 2.15 / 1.72	**1.54** / 1.48 / 1.37 / 0.57	1.74 / 1.68 / 1.61 / 1.24	1.76 / 1.66 / 1.61 / 1.24
SQuAD v2	**1.30** / **0.92** / **0.65** / **0.00**	1.77 / 1.58 / 1.44 / 1.06	1.31 / 1.21 / 0.94 / 0.78	1.49 / 1.28 / 1.18 / 0.92	1.39 / 1.29 / 1.10 / 0.85
LongForm–QA (Canonical)	**2.91** / **2.47** / **1.41** / **0.00**	7.78 / 7.46 / 4.22 / **0.00**	4.11 / 3.42 / 2.59 / 0.43	3.21 / 2.79 / 2.01 / **0.00**	3.51 / **3.46** / 1.96 / **0.00**
FinQA	**3.56** / **2.92** / **1.99** / **0.00**	6.72 / 4.26 / 5.21 / 2.74	3.67 / 3.67 / 3.83 / 3.96	3.67 / 3.33 / 2.74 / **0.00**	**3.56** / 3.29 / 2.98 / **0.00**

Table 6: Robustness analysis: 95-th percentile accuracy gap at the budget needed for Online LLM Picker to reach identification probabilities of 70%, 80%, 90%, and 100%. Bold denotes the best value, underlining denotes the second-best.

D.4 Annotation Efficiency

For completeness, we show the extended plots related to the results of annotation efficiency described in Table 1.

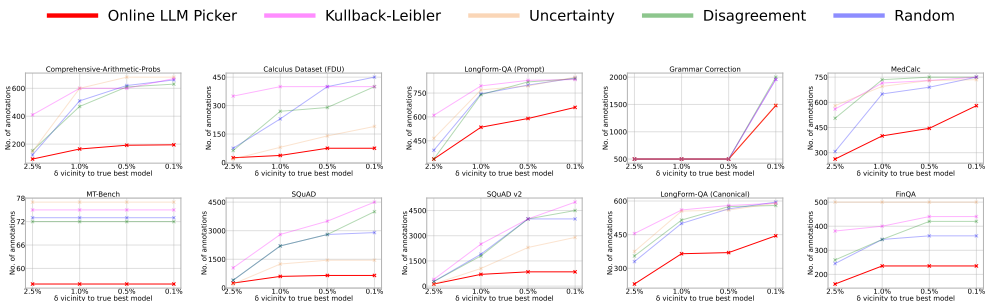

Figure 5: Identification probabilities of Online LLM Picker and baselines across different annotation budgets.

E Additional Details

Let $\hat{L}_{t-1,i}$ be the cumulative loss estimate and η_t the adaptive learning rate. The posterior distribution $\mathbf{p}_t = [p_{t,i}]_{i\in\mathcal{M}}$ is updated as:

$$p_{t,i} = \frac{\exp\left\{-\eta_t \hat{L}_{t-1,i}\right\}}{\sum_{j\in\mathcal{M}} \exp\left\{-\eta_t \hat{L}_{t-1,j}\right\}} \tag{11}$$