OVERCOMING PRIOR BARRIERS: SUPERVISED FINE-TUNING UNDER LONG-TAIL DISTRIBUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Supervised fine-tuning (SFT) adapts pretrained large language models (LLMs) to downstream tasks, but the required concepts can receive substantially different levels of pretrained support. Frequent concepts are more likely to be well learned, whereas rare concepts may remain weakly represented. We introduce a novel notion named *prior barrier* to quantify how strongly the pretrained model supports competing concepts over the target concept. We observe that prior barriers follow a long-tail distribution, placing head and tail concepts at different starting points for SFT: head concepts face lower prior barriers, whereas tail concepts require additional instructions to overcome their higher prior barriers. Our theoretical analysis further derives a predictive risk bound for SFT under long-tail prior barriers, explicitly characterizing how the prior barrier and accumulated SFT evidence jointly determine predictive performance. Motivated by this prior barrier-dependent demand, we propose PASS, an adaptive SFT instruction selection method that constructs reference-derived concepts and estimates the distinguishing evidence provided by each instruction, and adaptively allocates the selection budget toward concepts that remain insufficiently covered under the current selection. In this way, PASS jointly considers which instructions can provide useful evidence and where additional supervision is needed under a limited budget. Experiments show that our method consistently outperforms seven state-of-the-art instruction selection methods on four backbone-budget settings. An ablation study further shows that PASS's adaptive allocation consistently improves over uniform allocation.

1 INTRODUCTION

Supervised fine-tuning (SFT) is widely used to adapt pretrained large language models (LLMs) to downstream requirements such as mathematical reasoning, instruction following, factuality, and domain-specific behaviors (Ouyang et al., 2022; Zhou et al., 2023; Yue et al., 2023). In practice, the available instruction pool can be much larger than the amount of data that can be used for fine-tuning, while candidate instructions can also vary substantially in quality, making data selection an important part of post-training (Xia et al., 2024). Given a downstream task and a large pool of candidate instructions, which examples should be selected under a limited SFT budget? Recent studies show that carefully selected subsets can match or outperform substantially larger SFT datasets (Zhou et al., 2023; Chen et al., 2024; Xia et al., 2024; Liu et al., 2024b), underscoring the importance of selecting valuable instructions under a limited SFT budget.

A central difficulty for effective SFT instruction selection is that the concepts required by the downstream task can receive substantially different levels of support from the pretrained model. Each task can involve a collection of concepts, ranging from common reasoning patterns and broadly observed instructions to rare knowledge and specialized behaviors. Frequently occurring concepts receive broader exposure during pretraining and are therefore more likely to receive strong support from the pretrained model, whereas rare concepts may remain weakly supported (Kandpal et al., 2023; Razeghi et al., 2022; Gekhman et al., 2024). We introduce a novel notion named *prior barrier* to quantify how strongly the pretrained model supports competing concepts over the target concept. Different prior barriers place target concepts at different starting points for SFT: a larger prior barrier indicates that more instructions is needed to achieve comparable performance. As illustrated in Figure 1, we observe a long-tail distribution in prior barriers, where head concepts face low prior barriers while tail concepts face high barriers, thus requiring more instructions. Under a limited SFT

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

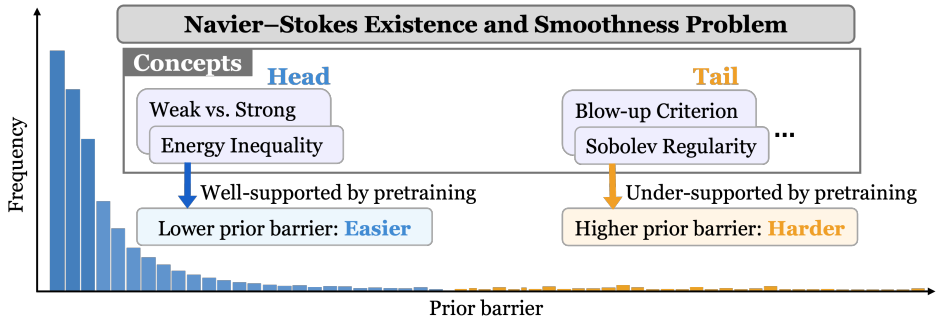


Figure 1: Illustration of the long-tail distribution of prior barriers using the Navier–Stokes existence and smoothness problem as an example.

budget, effective selection should determine how to allocate instructions across concepts and which candidate instructions provide useful evidence for each concept (Chen et al., 2023; Dai et al., 2026).

These observations motivate two fundamental challenges: **C1. Theoretical foundation.** How does the SFT evidence overcome the prior barrier, and how does this determine the amount of instructions required to learn a target concept? **C2. Computational framework.** How can we develop a method that identifies which instructions in the candidate pool provide useful evidence for each concept while adaptively allocating a limited selection budget according to concept-level demand?

To address **C1**, we first characterize the role of the prior barrier in SFT. In the *head case*, the target concept already receives sufficient pretrained support. SFT can move toward the target with relatively limited evidence. In the *tail case*, the target concept faces a positive prior barrier, so the accumulated evidence must additionally overcome this disadvantage from pretraining. This interaction yields different instruction demands across concepts. We further derive a predictive risk bound that connects the demands to model performance by quantifying predictive risk in terms of the prior barrier and the distinguishing evidence provided by the selected instructions.

To address **C2**, we propose Prior-barrier Aware SFT Instruction Selection (PASS), an adaptive SFT instruction selection framework motivated by the prior barrier-dependent demand. It first constructs a reference-derived concept space and estimates concept-specific distinguishing evidence for each candidate instruction. PASS further estimates the prior barrier and existing pretrained support of each concept to characterize how strongly that concept calls for additional SFT instruction. It then adaptively allocates the selection budget toward concepts that remain under-represented. As a result, PASS jointly determines which candidate instructions provide useful evidence and where the limited SFT budget should be allocated. Our empirical results show that PASS outperforms seven state-of-the-art instruction selection methods in all four backbone-budget settings. Ablation results further demonstrate the benefit of adaptively allocating the SFT budget according to prior barrier. The source code will be released upon publication.

2 Preliminary

In this section, we introduce the background, including a Bayesian concept-space view of SFT, and give the formal problem definition.

Notations. We consider a concept space Θ, where each concept $\theta \in \Theta$ describes a possible input–response relation through a conditional distribution $P_\theta(y \mid x)$. Here, x denotes an instruction input or query and y denotes the corresponding response. For example, for a mathematical reasoning concept, x may be a mathematical problem and y its correct solution. We denote the target concept required by the downstream task as $\theta^\star$. Pretraining does not provide equal support for all concepts. We represent this pretrained support by a distribution $\pi(\theta)$ over Θ, where a larger $\pi(\theta)$ indicates stronger pretrained support for concept θ. We interpret π as a prior distribution over concepts induced by pretraining. As a result, different target concepts can begin SFT from substantially different pretrained states.

Supervised fine-tuning. SFT adapts a pretrained language model to the downstream task using a set of supervised instruction input–response pairs. Given an SFT dataset $\mathcal{D}_n = \{(x_i, y_i)\}_{i=1}^n$, where x_i is the i-th instruction input and y_i is its target response, SFT updates the pretrained model to increase the likelihood of the target responses conditioned on their inputs. For an autoregressive

language model, this conditional probability factorizes as $P_\theta(y_i \mid x_i) = \prod_{t=1}^{T_i} P_\theta(y_{i,t} \mid x_i, y_{i,<t})$, where T_i is the number of tokens in response y_i, $y_{i,t}$ is the t-th token of y_i and $y_{i,<t}$ denotes its preceding response tokens. Standard SFT therefore trains the model by minimizing the token-level negative log-likelihood over the SFT dataset. We summarize the evidence provided by the SFT dataset for concept θ through its log-likelihood $S_n(\theta) = \sum_{i=1}^{n} \log P_\theta(y_i \mid x_i)$. We adopt a Bayesian concept-space view of SFT, where the pretrained distribution π serves as the prior and the supervised data update this prior through their likelihood. Let q denote a candidate distribution over the concept space Θ after observing the SFT dataset $\mathcal{D}_n$. We define $\mathcal{J}(q) = -\mathbb{E}_{\theta\sim q}[S_n(\theta)] + D_{\mathrm{KL}}(q\|\pi)$, The posterior distribution is defined as $q_n^\star = \arg\min_q \mathcal{J}(q)$, which takes the Gibbs-posterior form $q_n^\star(\theta) = \frac{e^{S_n(\theta)}\pi(\theta)}{\int_\Theta e^{S_n(u)}\pi(u)\,\lambda(du)}$. Here, $q_n^\star(\theta)$ represents the updated support for concept θ, obtained by updating the pretrained prior $\pi(\theta)$ with the likelihood evidence from the SFT data through $S_n(\theta)$.

Problem definition. Given a candidate SFT pool $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, a selection budget n, and a reference set $\mathcal{T}$ representing the desired behaviors in the downstream task with a long-tail distribution of prior barriers across concepts, our goal is to find a subset of instructions $\mathcal{D}_n \subseteq \mathcal{D}$ with $|\mathcal{D}_n| = n$ such that fine-tuning the pretrained model on $\mathcal{D}_n$ achieves strong overall performance on the downstream task.

3 Theoretical Analysis

In this section, we develop a theoretical analysis of how pretrained support and SFT evidence jointly shape predictive performance. First, we study posterior concentration around the target concept, showing how the prior barrier and distinguishing evidence together determine whether SFT can move the concept distribution toward the target. Second, we derive a predictive risk bound that connects posterior concentration to predictive performance and shows how the head–tail gap translates into different generalization guarantees.

Prior barrier. A pretrained model does not enter SFT from a neutral state: different concepts can receive substantially different levels of support from pretraining. This asymmetry is important because SFT adapts a pretrained model, where the target concept may compete with alternatives that are more strongly supported before fine-tuning begins. A key challenge is therefore to characterize this pretrained support relative to the target concept $\theta^\star$ and its competing concepts $\theta \in \Theta \setminus \{\theta^\star\}$. To this end, we consider concepts outside an r-neighborhood of the target concept $A_r = \{\theta \in \Theta : \|\theta - \theta^\star\| \geq r\}$. If $\theta^\star$ already receives stronger support than all concepts in A_r, the target concept does not face a pretrained disadvantage. Otherwise, some competing concept receives stronger pretrained support than $\theta^\star$, placing the target at an initial disadvantage before SFT begins. We characterize this relative pretrained disadvantage through the *prior barrier*.

Definition 1 (Prior barrier). *The prior barrier at resolution r is $B(r) = \sup_{\theta\in A_r} \log \frac{\pi(\theta)}{\pi(\theta^\star)}$.*

The prior barrier $B(r)$ compares the pretrained support of the target concept with its strongest competing concept. We observe that prior barriers exhibit a long-tail distribution across concepts, with many concepts facing relatively small barriers and few concepts facing substantially larger ones. For the analysis below, we refer to $B(r) \leq 0$ as the *head case*, where the target concept is at least as well supported as all concepts in A_r, and $B(r) > 0$ as the *tail case*, where at least one competing concept receives stronger pretrained support than the target concept.

Overcoming a positive prior barrier requires the SFT instructions to contain information that distinguishes the target concept from its competing concepts. The amount of useful information depends on the instruction distribution P_X: some instructions can clearly expose the difference between the target concept and an alternative concept, whereas others provide little distinguishing evidence. For a competing concept θ, define the expected evidence under P_X as

$$K(\theta) = \mathbb{E}_{x\sim P_X}\left[\mathrm{KL}\left(P_{\theta^\star}(\cdot \mid x)\|P_\theta(\cdot \mid x)\right)\right] = \mathbb{E}_{\substack{x\sim P_X \\ y\sim P_{\theta^\star}(\cdot|x)}}\left[\log P_{\theta^\star}(y \mid x) - \log P_\theta(y \mid x)\right]. \quad (1)$$

Thus, a larger $K(\theta)$ means that the observed instruction input–response pair drawn from the instruction distribution provides stronger expected evidence for separating the target concept from θ. Because SFT requires distinguishing the target from all concepts outside its neighborhood, we focus on the hardest competitor.

Definition 2 (Distinguishing Evidence). *The distinguishing evidence for the target concept at resolution r is $\epsilon(r) = \inf_{\theta \in A_r} K(\theta)$.*

The target concept $\theta^\star$ is identifiable under the SFT instruction distribution P_X when $\epsilon(r) > 0$.

Learning the target concept requires the posterior distribution after SFT to place little mass on the competing concepts in A_r, that is, to make $q_n^\star(A_r)$ small. The two central quantities governing it are $B(r)$, which measures how much pretrained disadvantage must be overcome, and $n\epsilon(r)$, which measures how much distinguishing evidence SFT accumulates. As the number of SFT instructions increases, the accumulated evidence grows with $n\epsilon(r)$. Therefore, once this evidence is sufficient to overcome the pretrained disadvantage, the posterior can concentrate around the target concept.

Theorem 1 (Posterior Concentration). *Fix $r > 0$ such that $A_r \neq \varnothing$ and $\epsilon(r) > 0$, and let $m(r) > 0$ denote the normalized prior mass in a sufficiently small neighborhood of the target concept $\theta^\star$. Under the assumptions given in Appendix A.1, for any $\eta \in (0, 1)$, if $n \geq n_{\mathrm{sep}}(r, \eta)$ where $n_{\mathrm{sep}}(r, \eta)$ denotes the sample-size threshold, with probability at least $1 - \eta$ over the SFT dataset $\mathcal{D}_n$,*

$$q_n^\star(A_r) \leq \min\left\{1, \frac{1}{m(r)} \exp\left(B(r) - \frac{n\epsilon(r)}{8}\right)\right\}. \tag{2}$$

Proof. The proof is provided in Appendix A.2. □

Theorem 1 shows that, once the target concept is identifiable under the SFT distribution, sufficiently many SFT examples drive posterior mass away from competing concepts and toward the target concept. Since the prior barrier is fixed while the accumulated distinguishing evidence grows with n, the concentration error decreases exponentially as supervision increases. The amount of supervision required, however, differs between the head and tail cases.

Corollary 1 (Head–Tail Supervision Gap). *Under the conditions of Theorem 1, let $\delta \in (0, 1)$ be a desired posterior error. The head case $B(r) \leq 0$ satisfies $q_n^\star(A_r) \leq \delta$ whenever*

$$n \geq \max\left\{\frac{8}{\epsilon(r)}\left(\log\frac{1}{m(r)\delta}\right)_+, n_{\mathrm{sep}}(r, \eta)\right\}, \tag{3}$$

where $(a)_+ = \max\{a, 0\}$. The tail case $B(r) > 0$ satisfies $q_n^\star(A_r) \leq \delta$ whenever

$$n \geq \max\left\{\frac{8}{\epsilon(r)}\left(B(r) + \log\frac{1}{m(r)\delta}\right)_+, n_{\mathrm{sep}}(r, \eta)\right\}. \tag{4}$$

Proof. The proof is provided in Appendix A.3. □

The posterior concentration result explains when SFT places most of its concept-level support near the target concept. We now ask whether this concept-level concentration also leads to predictive performance. Define the posterior predictive distribution as $P_{q_n^\star}(y \mid x) = \int_\Theta P_\theta(y \mid x) q_n^\star(d\theta)$, and measure its excess predictive log-loss relative to the target concept by $\mathcal{E}_n = \mathbb{E}_{x \sim P_X}\left[D_{\mathrm{KL}}\left(P_{\theta^\star}(\cdot \mid x) \| P_{q_n^\star}(\cdot \mid x)\right)\right]$. We now establish how posterior concentration around the target concept translates into predictive performance for SFT.

Theorem 2 (Predictive Risk Bound). *Let $\alpha(r) = \mathbb{E}_{x \sim P_X}\left[\sup_{\|\theta - \theta^\star\| < r} D_{\mathrm{KL}}\left(P_{\theta^\star}(\cdot \mid x) \| P_\theta(\cdot \mid x)\right)\right]$ denote the worst local predictive discrepancy within distance r of the target concept, and let $M_r < \infty$ be a uniform upper bound on the predictive KL discrepancy for competing concepts in A_r. Under the assumptions stated in Appendix A.1, SFT dataset $\mathcal{D}_n$ satisfies*

$$\mathcal{E}_n \leq \alpha(r) + M_r q_n^\star(A_r). \tag{5}$$

Consequently, under the conditions of Theorem 1, with probability at least $1 - \eta$,

$$\mathcal{E}_n \leq \alpha(r) + M_r \min\left\{\frac{1}{m(r)} \exp\left(B(r) - \frac{n\epsilon(r)}{8}\right), 1\right\}. \tag{6}$$

Proof. The proof is provided in Appendix A.4. □

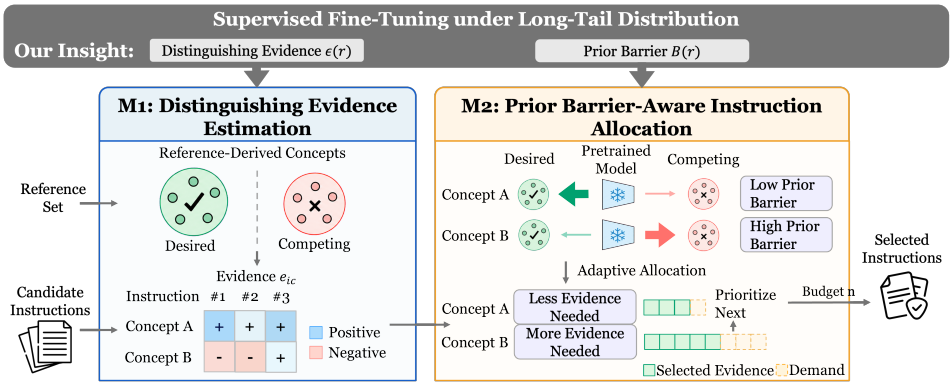


Figure 2: The proposed prior-barrier aware SFT instruction selection (PASS) framework.

Theorem 2 shows that predictive risk depends on two terms: how well concepts near the target concept approximate its predictions, and how much posterior mass remains on competing concepts. As SFT concentrates the posterior around the target concept, the contribution from competing concepts decreases, leading to a tighter predictive risk bound. In particular, if SFT achieves $q_n^\star(A_r) \leq \delta$, then $\mathcal{E}_n \leq \alpha(r) + M_r\delta$. Thus, a target concept in the head case can reach a given predictive risk with less supervision, while a target concept in the tail case must first accumulate additional evidence to overcome its positive prior barrier. As a result, the tail case generally requires more SFT instructions to reach the same level of predictive risk.

4 Method

Our theoretical analysis shows that achieving a small predictive risk requires SFT instructions to provide sufficient evidence for distinguishing the target concept from competing concepts. It further reveals different instruction requirements in the head and tail cases: a tail target concept requires additional instructions to overcome its positive prior barrier. These results motivate us to consider both how much distinguishing evidence each candidate instruction provides and how a fixed instruction budget should be allocated across target concepts with asymmetric pretrained support. In this section, we introduce PASS (shown in Figure 2), which consists of two major modules: *M1. Distinguishing evidence estimation* and *M2. Prior barrier-aware instruction allocation*. Accordingly, M1 constructs reference-derived concepts and estimates evidence for each candidate instruction by contrasting its support for the desired concept against relevant competing concepts. M2 uses the pretrained model to estimate the barrier and existing support of each concept, and incrementally allocates instructions according to both concept-level demand.

4.1 M1: Distinguishing Evidence Estimation

Given a reference set $\mathcal{T}$, we first construct a finite empirical discretization of the concept space Θ introduced in Section 2. We encode each reference instruction input x together with its target response y using the pretrained model M_0 to extract a response representation $z(x, y) \in \mathbb{R}^d$. We hierarchically cluster these reference representations, recursively splitting clusters that do not satisfy a coherence criterion. We measure coherence by the concentration of cosine similarity between the representations in a cluster and their centroid, and stop splitting once each cluster is sufficiently coherent. The resulting clusters instantiate a finite set of target concepts $\Theta = \theta_1, \ldots, \theta_C$, where C is the number of concepts. Each concept θ therefore corresponds to a subset $\mathcal{T}_\theta \subseteq \mathcal{T}$ of reference instructions that exhibit a coherent representation.

We represent the desired representation of concept θ by the normalized centroid of the representations in $\mathcal{T}_\theta$:

$$p_\theta = \mathrm{Norm}\left(\frac{1}{|\mathcal{T}_\theta|}\sum_{(x,y)\in\mathcal{T}_\theta} z(x, y)\right), \tag{7}$$

where $\mathrm{Norm}(v) = \frac{v}{\|v\|_2}$. To construct competing representations, we keep the same instruction inputs in $\mathcal{T}_\theta$ but replace their target responses with corresponding wrong responses, using available incorrect alternatives together with incorrect responses produced by M_0. These wrong responses

may reflect multiple competing concepts for the same concept. For concept θ, let $\mathcal{W}_{\theta k}$ denote the set of wrong response instructions corresponding to its k-th competing concept, with $k = 1, \ldots, K_\theta$, where K_θ denotes the number of competing concepts associated with target concept θ. We represent each competing representation by the normalized centroid:

$$n_{\theta k} = \mathrm{Norm}\left(\frac{1}{|\mathcal{W}_{\theta k}|}\sum_{(x,\tilde{y})\in\mathcal{W}_{\theta k}} z(x,\tilde{y})\right), \qquad k = 1, \ldots, K_\theta, \tag{8}$$

where $\tilde{y}$ denotes a wrong response. The resulting set $\mathcal{N}_\theta = \{n_{\theta 1}, \ldots, n_{\theta K_\theta}\}$ defines the competing representations of concept θ.

Our theoretical analysis suggests that a instruction may provide different distinguishing evidence for different concepts. We therefore evaluate candidate instruction $(x_i, y_i) \in \mathcal{D}$ separately against every concept. Let $z_i = z(x_i, y_i)$ be the representation of the i-th instruction. For concept θ, we measure its similarity to the desired representation and to the closest competing representation as

$$s_\theta^+(z_i) = \mathrm{Sim}(z_i, p_\theta), \qquad s_\theta^-(z_i) = \max_{v\in\mathcal{N}_\theta} \mathrm{Sim}(z_i, v), \tag{9}$$

where Sim denotes cosine similarity. The difference $s_\theta^+(z_i) - s_\theta^-(z_i)$ indicates whether the candidate favors the desired representation of concept θ over its strongest competitor. For directly comparable across concepts, we calibrate this difference using statistics estimated from the reference responses of each concept. We define the empirical distinguishing evidence as

$$\epsilon_{i\theta} = \mathrm{Calib}_\theta\left(s_\theta^+(z_i), s_\theta^-(z_i)\right), \tag{10}$$

with $\mathrm{Calib}_\theta(s^+, s^-) = r_\theta(s^+, s^-)\frac{s^+ - s^-}{\sigma_\theta}$. Here, $r_\theta(s^+, s^-) \in [0, 1]$ is a relevance factor determined by the candidate's affinity $\max\{s^+, s^-\}$ relative to the affinity observed for the reference responses of concept θ. It downweights candidates that are far from both the desired and competing representations while preserving candidates that are strongly aligned with either side. The scale $\sigma_\theta > 0$ normalizes the similarity difference between s_θ^+ and s_θ^- for concept θ, making evidence magnitudes comparable across concepts. Directly estimating the theoretical distinguishing evidence requires conditional response distributions, which are generally unavailable in practice. We therefore use the calibrated similarity margin as an empirical surrogate. A positive $\epsilon_{i\theta}$ indicates that candidate i provides evidence in favor of the desired representation of concept θ, whereas a negative value indicates stronger support for a competing concept. M1 outputs the evidence matrix $[\epsilon_{i\theta}] \in \mathbb{R}^{N\times C}$ for the reference-derived concepts rather than collapsing each candidate into a single score.

4.2 M2: Prior Barrier-Aware Instruction Allocation

M1 identifies how much distinguishing evidence each candidate can provide for different concepts, while the remaining question is how to allocate a fixed selection budget across concepts with different demands. Our theoretical analysis shows that a tail concept requires additional instructions to overcome its pretrained disadvantage. M2 therefore estimates concept-level demand from the prior barrier and incrementally allocates instructions as useful evidence is accumulated.

Since the theoretical prior barrier is not directly observable, we estimate it empirically from the pretrained model's responses. We let the pretrained model M_0 generate a response $\hat{y} \sim M_0(\cdot \mid x)$ and obtain its representation $\hat{z} = z(x, \hat{y})$ for each instruction input in $\mathcal{T}_\theta$. The prior barrier of concept θ is estimated from the normalized difference between competing and desired representations,

$$B_\theta = \mathrm{Agg}_{(x,y)\in\mathcal{T}_\theta}\left[\frac{s_\theta^-(\hat{z}) - s_\theta^+(\hat{z})}{\sigma_\theta}\right], \tag{11}$$

where Agg denotes the aggregation over the reference instructions of concept θ.

Complementary to this relative barrier, we separately estimate the absolute pretrained support for the target concept by measuring whether the generated response $\hat{y}$ lies sufficiently close to the desired representation p_θ:

$$\gamma_\theta = \frac{1}{|\mathcal{T}_\theta|}\sum_{(x,y)\in\mathcal{T}_\theta} \mathrm{Support}_\theta\left(\mathrm{Sim}(\hat{z}, p_\theta)\right), \tag{12}$$

where $\mathrm{Support}_\theta(\cdot) \in [0, 1]$ calibrates the similarity to p_θ using the similarities observed from the target responses of concept θ. A larger γ_θ indicates stronger pretrained support for the target concept.

A concept should receive more instruction when the pretrained model exhibits a higher prior barrier and weaker absolute target support. We therefore define its allocation weight as

$$A_\theta = \frac{\exp(B_\theta)/\gamma_\theta}{\sum_{\theta' \in \Theta} \exp(B_{\theta'})/\gamma_{\theta'}}, \tag{13}$$

so that $\sum_{\theta \in \Theta} A_\theta = 1$. While A_θ characterizes the relative instruction demand of each concept, the final selected subset should also account for how selected instructions contribute positive and negative distinguishing evidence, as well as the redundancy among their representations. We therefore define the selection objective by jointly modeling these three factors:

$$F(\mathcal{S}) = \sum_{\theta \in \Theta} A_\theta R_\theta^+(\mathcal{S}) - \lambda_h \sum_{\theta \in \Theta} A_\theta R_\theta^-(\mathcal{S}) - \lambda_r D(\mathcal{S}), \tag{14}$$

where R_θ^+ rewards positive distinguishing evidence for concept θ, R_θ^- penalizes negative distinguishing evidence for concept θ, D penalizes redundant selections, and λ_h and λ_r denote the weights of harm penalty and redundancy penalty.

For positive evidence, repeatedly selecting instructions for the same concept should yield diminishing returns. We therefore define $R_\theta^+(\mathcal{S}) = 1 - \exp\left\{-\beta_\theta^+ \sum_{i \in \mathcal{S}} \epsilon_{i\theta}^+\right\}$, where $\epsilon_{i\theta}^+ = (\epsilon_{i\theta})_+$. The coefficient $\beta_\theta^+ > 0$ controls how quickly the positive reward for concept θ saturates. Concepts with richer available evidence therefore saturate more slowly, while the marginal reward decreases as more positive evidence has already been selected.

For negative evidence, we define $R_\theta^-(\mathcal{S}) = \sum_{i \in \mathcal{S}} \left[\exp\left\{\beta^- \epsilon_{i\theta}^-\right\} - 1\right]$, where $\epsilon_{i\theta}^- = (-\epsilon_{i\theta})_+$, and $\beta^- > 0$ is an evidence scale shared across concepts for negative contributions. This term assigns a larger penalty to instructions that more strongly favor competing concepts.

Finally, we discourage selecting instructions that provide highly redundant information. We define $D(\mathcal{S}) = \frac{2}{|\mathcal{S}|(|\mathcal{S}|-1)} \sum_{i,j \in \mathcal{S}} \mathrm{Sim}(z_i, z_j),$. This term encourages the selected subset to cover distinct regions of the response representation space.

We initialize $\mathcal{S} = \varnothing$ and greedily select the instruction with the largest marginal gain,

$$i^\star \in \arg \max_{i:(x_i,y_i) \in \mathcal{D} \setminus \mathcal{S}} \left[F\big(\mathcal{S} \cup \{(x_i, y_i)\}\big) - F(\mathcal{S})\right]. \tag{15}$$

After each selection, the selection objective is updated. The procedure stops when $|\mathcal{S}| = n$, yielding the subset used for SFT. Algorithm 1 in Appendix B summarizes the complete procedure.

5 Experiments

In this section, we evaluate the performance of our method on four backbone-budget settings. PASS exhibits superior performance compared to various state-of-the-art baselines (Section 5.2). Further, through the ablation study (Section 5.3), we demonstrate the necessity of each component of PASS. We also report parameter analysis (Section 5.4), complexity sensitivity (Appendix C.2), and the composition of instructions selected by different instruction selection methods (Appendix C.3).

5.1 Experiment Setup

Candidate pool and target tasks. The candidate pool contains 300,932 instruction–response examples drawn from five open instruction datasets: FLAN V2 (Longpre et al., 2023), OASST1 (Köpf et al., 2023), WizardLM (Xu et al., 2024), Dolly (Conover et al., 2023) and Stanford Alpaca (Taori et al., 2023). To evaluate SFT instruction selection across concepts with long-tail prior barrier, we construct a downstream task spanning five capabilities: broad knowledge, multi-step reasoning, mathematics, truthfulness, and multilingual extractive QA. We form the reference set from examples covering these capabilities so that instruction selection is evaluated against

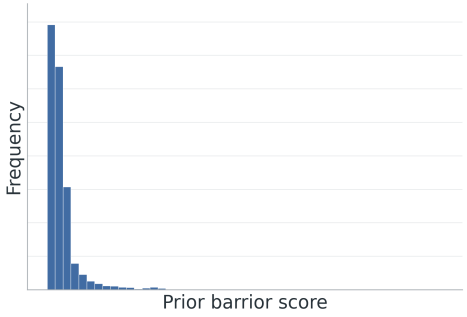


Figure 3: Long-tail distribution of prior barrier over the reference-derived concepts.

Table 1: Overall downstream performance on Mistral-7B and Qwen2.5-7B at two selection budgets. The best and second-best results within each model and budget are shown in **bold** and underlined.

	Mistral-7B		Qwen2.5-7B	
Method	$n = 5{,}000$	$n = 10{,}000$	$n = 5{,}000$	$n = 10{,}000$
BASE (no SFT)	26.9	26.9	52.8	52.8
RANDOM	49.3	50.4	61.7	**65.6**
RDS+	49.9	50.3	64.3	63.4
LESS	47.7	50.3	60.3	59.9
TSDS	49.4	49.8	63.1	63.0
TADS	48.7	<u>50.9</u>	<u>64.6</u>	63.2
ADG	50.0	<u>50.9</u>	62.4	64.1
DEITA	<u>50.6</u>	48.9	60.7	62.8
PASS	**51.5**	**51.8**	**64.8**	**65.6**

Table 2: Budget-aware instruction allocation versus uniform instruction allocation. $\Delta_{\text{uni}} = \text{PASS} - \text{UNIFORM}$. Bold identifies the better results.

	$n = 5{,}000$			$n = 10{,}000$		
Backbone	PASS	UNIFORM	Δ_{uni}	PASS	UNIFORM	Δ_{uni}
Mistral-7B	**51.46**	51.14	$+0.32$	**51.81**	51.54	$+0.26$
Qwen2.5-7B	**64.77**	64.48	$+0.29$	**65.60**	65.03	$+0.56$

Table 3: Parameter sensitivity of PASS. Δ_{def} is the difference relative to the default setting.

Setting	λ_h	λ_r	Mean	Δ_{def}
Default	1.0	0.10	51.46	0.00
Weak harm penalty	0.5	0.10	51.56	$+0.10$
Strong harm penalty	2.0	0.10	50.54	-0.92
Weak redundancy penalty	1.0	0.05	50.96	-0.50
Strong redundancy penalty	1.0	0.20	52.21	$+0.75$

the full range of target concepts. As shown in Figure 3, the prior barrier scores of the concepts in the reference set exhibit a long-tail distribution, providing a suitable setting for evaluating SFT under the long-tail distribution.

Backbones, budgets, and metrics. The main comparisons use Mistral-7B (Jiang et al., 2023) and Qwen2.5-7B (Qwen et al., 2024). We select $n \in \{5{,}000, 10{,}000\}$ records, corresponding to approximately 1.66% and 3.32% of the original candidate pool. Within each backbone and budget, all selectors share the fine-tuning settings, the training recipe uses three epochs of LoRA (Hu et al., 2021). We report the overall performance on the downstream task (the macro-average of the five capability benchmark scores).

Baselines. We compare against *Base* (backbone without fine-tuning) and seven instruction selection methods, including RANDOM that samples uniformly; representation-based RDS+ (Ivison et al., 2025); gradient-based LESS (Xia et al., 2024); TSDS that uses optimal transport (Liu et al., 2024c); TADS that combines divisity selections by voting (Cheng et al., 2026); answer-divergence-based ADG (Li et al., 2026); and instance-quality filtering by DEITA (Liu et al., 2024b).

5.2 PERFORMANCE ANALYSIS

Table 1 shows that PASS consistently achieves the best overall performance across all four settings, indicating consistent gains across backbone choice and selection budget. In particular, on Mistral-7B, PASS reaches a score of 51.5 with $5{,}000$ selected instructions, higher than every competing selector using twice the budget ($n = 10{,}000$), where the strongest overall performance is 50.9. This result indicates that a better composition of the selected subset can compensate for a substantially smaller SFT budget. In contrast, other selection methods may identify individually relevant or high-quality instructions, but they fail to adaptively cover all capabilities required for downstream tasks,

resulting in poor overall performance. This is particularly evident on Qwen2.5-7B with $n = 10{,}000$, where RANDOM achieves an average score of 65.6, outperforming all other baselines. Taken together, these results support the motivation behind PASS: the key challenge is not only to identify useful individual instructions, but also to adaptively allocate budget based on the prior barrier.

5.3 ABLATION STUDY

We compare our budget-aware instruction allocation strategy with *uniform instruction allocation*. Under uniform variant, we allocate an equal budget of $n/5$ separately to each of the five capabilities in the downstream task. Table 2 shows that budget-aware allocation consistently outperforms uniform allocation across both backbones and both selection budgets. On Mistral-7B, PASS improves the average score by 0.32 and 0.26 points at $n = 5{,}000$ and $n = 10{,}000$, respectively. A similar trend is observed on Qwen2.5-7B, where the gains are 0.29 and 0.56 points. These results indicate that an equal per-capability assignment is suboptimal. In contrast, PASS adaptively allocates instructions according to the prior barrier, thereby making better use of the limited SFT budget.

5.4 PARAMETER ANALYSIS

We study two hyperparameters of our method: (1) the harm coefficient λ_h weights the penalty on negative distinguishing evidence, and (2) the redundancy coefficient λ_r controls the penalty on redundant selections. Table 3 reports the overall performance under varying λ_h and λ_r using Mistral-7B with $n = 5{,}000$ selected instructions. The fluctuation of the performance is less than 2%, showing PASS is reliable and not sensitive to the hyperparameters over a wide range.

6 RELATED WORK

Instruction selection for supervised fine-tuning (SFT) has been studied from several perspectives, including instance quality, model-dependent utility, target-aware matching. Early work shows that small but carefully selected instruction sets can be highly effective (Zhou et al., 2023; Chen et al., 2024). Methods such as DEITA further combine response quality, instruction complexity, and embedding diversity to identify informative and non-redundant examples (Liu et al., 2024b). Other approaches make selection dependent on the model itself. For example, ADG measures the divergence among multiple generated answers to estimate instruction utility (Li et al., 2026), while related methods use model difficulty, uncertainty, or response likelihood as selection signals (Li et al., 2024; Liu et al., 2024a; Zhang et al., 2025).

Recent work uses downstream references to identify instructions that are relevant to the target task. LESS estimates candidate influence through low-rank gradient similarity to reference examples (Xia et al., 2024); RDS+ retrieves candidates using pretrained representations (Ivison et al., 2025); TSDS aligns the selected subset with the target distribution through optimal transport (Liu et al., 2024c); and TADS uses LLM-inferred proxy labels to improve task-aware distribution matching (Cheng et al., 2026). These methods provide strong mechanisms for estimating whether an instruction is relevant to downstream requirements, but generally summarize candidate utility through similarity, influence, or distribution matching. Beyond candidate scoring, several methods consider how supervision should be distributed across skills or tasks. Skill-It adapts training mixtures according to skill dependencies (Chen et al., 2023), while MIG and BIDS dynamically prioritize informative or underrepresented targets (Chen et al., 2025; Dai et al., 2026). Our work complements these approaches by connecting instruction allocation to unequal pretrained support through prior barrier. It combines prior barrier with distinguishing evidence to adaptively allocate the SFT budget.

7 CONCLUSION

This work studies supervised fine-tuning under long-tail distributions, where different target concepts can receive substantially different levels of pretrained support. We introduce the *prior barrier* to characterize this difference, and observe that prior barriers exhibit a long-tail distribution. Our theoretical analysis shows that the prior barrier and instruction evidence jointly determine the number of instructions required during SFT: head concepts require relatively limited additional supervision, whereas tail concepts require more evidence to overcome their prior barriers. Guided by this analysis, we propose PASS, which estimates concept-specific distinguishing evidence and adaptively allocates a limited SFT budget according to the prior barrier. PASS achieves the best overall performance compared with seven instruction selection baselines in all four backbone-budget settings, while its adaptive allocation consistently improves over uniform allocation. These results show that effective SFT instruction selection should account for both the evidence provided by candidate instructions and the adaptive allocation based on the prior barrier.

AI USE STATEMENT

Generative AI was used solely for writing refinement and source-code comments. Specifically, it was used to improve the grammar, clarity, wording, and sentence-level phrasing of text written by the authors, and to assist with writing or refining comments in the source code. Generative AI was not used to develop research ideas, theoretical results, methodology, experimental design, research code, analyses, or conclusions. All AI-assisted text and code comments were reviewed and approved by the authors.

ETHICS STATEMENT

This work does not involve human-subject experiments or the collection of new personal or sensitive data. The experiments use publicly available instruction datasets, evaluation benchmarks, and pretrained language models, as described in Section 5. These resources may contain biases, inaccuracies, or undesirable content inherited from their original sources, which may in turn affect the instructions selected by PASS and the behavior of the fine-tuned models. Our empirical results focus on instruction selection and downstream performance, and should not be interpreted as establishing the safety, fairness, or suitability of the resulting models for high-stakes deployment.

REPRODUCIBILITY STATEMENT

We provide detailed materials to support the reproducibility of both our theoretical and empirical results. For the theoretical analysis, Appendix A.1 presents the assumptions underlying our theoretical results together with detailed explanations. Complete proofs of the theoretical results in Section 3 are provided in Appendices A.2, A.3, and A.4. For our proposed method, Section 4 provides a detailed description of PASS, and Appendix B provides the corresponding pseudocode to specify the complete selection procedure. For the empirical evaluation, Section 5.1 and Appendix C.1 provide detailed experimental settings and implementation details, including the construction of the instruction pool, evaluation benchmarks, model backbones, selection budgets, evaluation metrics, and fine-tuning protocol. Together, these materials specify the theoretical assumptions, algorithmic procedure, and experimental protocol needed to reproduce the main results reported in this work. Upon acceptance, we will release the code for instruction selection, fine-tuning, and evaluation to further facilitate reproduction of our experiments.

REFERENCES

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, and Hongxia Jin. Alpagasus: Training a better alpaca with fewer data, 2024. URL https://arxiv.org/abs/2307.08701.

Mayee F. Chen, Nicholas Roberts, Kush Bhatia, Jue Wang, Ce Zhang, Frederic Sala, and Christopher Ré. Skill-it! a data-driven skills framework for understanding and training language models, 2023. URL https://arxiv.org/abs/2307.14430.

Yicheng Chen, Yining Li, Kai Hu, Zerun Ma, Haochen Ye, and Kai Chen. MIG: Automatic data selection for instruction tuning by maximizing information gain in semantic space. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 9902–9915. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.515. URL https://aclanthology.org/2025.findings-acl.515/.

Hao Cheng, Rui Zhang, Ling Li, Na Di, Jiaheng Wei, Zhaowei Zhu, and Bo Han. Task-aware data selection via proxy-label enhanced distribution matching for LLM finetuning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=R40WoYbYab.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470, 2020. doi: 10.1162/tacl_a_00317. URL https://aclanthology.org/2020.tacl-1.30/.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free dolly: Introducing the world's first truly open instruction-tuned llm, 2023. URL https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm.

Qirun Dai, Dylan Zhang, Jiaqi W. Ma, and Hao Peng. Improving influence-based instruction tuning data selection for balanced learning of diverse capabilities, 2026. URL https://arxiv.org/abs/2501.12147.

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. Does fine-tuning LLMs on new knowledge encourage hallucinations? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7765–7784, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.444. URL https://aclanthology.org/2024.emnlp-main.444/.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021. URL https://arxiv.org/abs/2009.03300.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Hamish Ivison, Muru Zhang, Faeze Brahman, Pang Wei Koh, and Pradeep Dasigi. Large-scale data selection for instruction tuning, 2025. URL https://arxiv.org/abs/2503.01807. arXiv preprint arXiv:2503.01807.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril,

Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 15696–15707. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/kandpal23a.html.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. Openassistant conversations - democratizing large language model alignment. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 47669–47681. Curran Associates, Inc., 2023. doi: 10.52202/075280-2064. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/949f0f8f32267d297c2d4e3ee10a2e7e-Paper-Datasets_and_Benchmarks.pdf.

Bo Li, Mingda Wang, Shikun Zhang, and Wei Ye. Instruction data selection via answer divergence. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4684–4702, 2026.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. From quantity to quality: Boosting LLM performance with self-guided data selection for instruction tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*, 2024.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022.

Liangxin Liu, Xuebo Liu, Derek F. Wong, Dongfang Li, Ziyi Wang, Baotian Hu, and Min Zhang. SelectIT: Selective instruction tuning for LLMs via uncertainty-aware self-reflection. In *Advances in Neural Information Processing Systems*, volume 37, 2024a. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/b130a5691815f550977e331f8bec08ae-Abstract-Conference.html.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning, 2024b. URL https://arxiv.org/abs/2312.15685.

Zifan Liu, Amin Karbasi, and Theodoros Rekatsinas. TSDS: Data selection for task-specific model finetuning. In *Advances in Neural Information Processing Systems*, volume 37, 2024c. URL https://openreview.net/forum?id=wjbTHLUSzU.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, and Adam Roberts. The flan collection: Designing data and methods for effective instruction tuning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 22631–22648. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/longpre23a.html.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, et al. Qwen2.5 technical report, 2024. URL https://arxiv.org/abs/2412.15115.

Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. Impact of pretraining term frequencies on few-shot numerical reasoning. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 840–854, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.59. URL https://aclanthology.org/2022.findings-emnlp.59/.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, and Jason Wei. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, 2023.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. LESS: Selecting influential data for targeted instruction tuning. In *International Conference on Machine Learning*, 2024.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. Wizardlm: Empowering large pre-trained language models to follow complex instructions. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 30745–30766, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/82eec786fdfbbfa53450c5feb7d1ac92-Paper-Conference.pdf.

Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mammoth: Building math generalist models through hybrid instruction tuning, 2023. URL https://arxiv.org/abs/2309.05653.

Dylan Zhang, Qirun Dai, and Hao Peng. The best instruction-tuning data are those that fit. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4717. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/cea04322465ad2f261f08e5b47ba9e7a-Abstract-Conference.html.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. LIMA: Less is more for alignment. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

A Details of Theoretical Analysis

A.1 Assumptions

We assume that the concept space $\Theta \subset \mathbb{R}^d$ is compact and is equipped with a Borel probability reference measure λ with full support. The prior has a strictly positive continuous density $\pi(\theta)$ with respect to λ. The input and output spaces are standard Borel spaces. There exists a common σ-finite measure ν on the output space such that each $P_\theta(\cdot \mid x)$ admits a density, denoted by $P_\theta(y \mid x)$, with $(\theta, x, y) \mapsto P_\theta(y \mid x)$ jointly measurable. Define $P^\star(dx, dy) = P_X(dx)P_{\theta^\star}(dy \mid x)$, and assume that the SFT examples are i.i.d. from $P^\star$. Let $\ell_\theta(x, y) = \log P_\theta(y \mid x)$ and $S_n(\theta) = \sum_{i=1}^n \ell_\theta(x_i, y_i)$.

Assumption 1. *There exists a measurable set $\mathcal{Z}_0$ of $P^\star$-probability one such that, for every $(x, y) \in \mathcal{Z}_0$,*

$$\ell_\theta(x, y) \in \mathbb{R} \quad \text{for every } \theta \in \Theta, \quad \sup_{\theta \in \Theta} \ell_\theta(x, y) < \infty. \tag{16}$$

Assumption 2 (Bounded log-likelihood ratios)**.** *For a fixed $r > 0$, there exists $M_r \in (0, \infty)$ such that, for every $(x, y) \in \mathcal{Z}_0$ and every $\theta \in A_r$,*

$$|\ell_\theta(x, y) - \ell_{\theta^\star}(x, y)| \le M_r. \tag{17}$$

Assumption 3. *For a fixed $r > 0$, there exists $L_r \in (0, \infty)$ such that, for every $(x, y) \in \mathcal{Z}_0$,*

$$|\ell_\theta(x, y) - \ell_{\theta'}(x, y)| \le L_r \|\theta - \theta'\|, \qquad \theta, \theta' \in A_r, \tag{18}$$

$$|\ell_\theta(x, y) - \ell_{\theta^\star}(x, y)| \le L_r \|\theta - \theta^\star\|, \qquad \|\theta - \theta^\star\| < r/2. \tag{19}$$

Assumption 4. *The local approximation term $\alpha(r)$ is finite, and the pointwise supremum used to define it is measurable.*

For posterior concentration, define

$$\rho_r = \min\left\{\frac{r}{2}, \frac{\epsilon(r)}{8L_r}\right\}, \quad U_r = \{\theta \in \Theta : \|\theta - \theta^\star\| < \rho_r\}, \tag{20}$$

and let

$$m(r) = \int_{U_r} \frac{\pi(\theta)}{\pi(\theta^\star)}\, \lambda(d\theta). \tag{21}$$

Because $\rho_r > 0$, λ has full support, and π is strictly positive and continuous, $m(r) > 0$.

Let

$$\mathcal{N}_r = N\left(A_r, \frac{\epsilon(r)}{8L_r}, \|\cdot\|\right) \tag{22}$$

be the covering number of A_r. Compactness of A_r implies $\mathcal{N}_r < \infty$. For $\eta \in (0, 1)$, define

$$n_{\mathrm{sep}}(r, \eta) = \frac{8M_r^2}{\epsilon(r)^2} \log\left(\frac{\mathcal{N}_r}{\eta}\right). \tag{23}$$

Finally, let

$$\Delta_n(\theta) = S_n(\theta) - S_n(\theta^\star). \tag{24}$$

A.2 Proof of Theorem 1

Proof. We first establish a uniform empirical separation over A_r. For any fixed $\theta \in A_r$, define

$$Z_i(\theta) = \ell_\theta(x_i, y_i) - \ell_{\theta^\star}(x_i, y_i).$$

By Assumption 2, $Z_i(\theta) \in [-M_r, M_r]$, while the definition of $\epsilon(r)$ gives

$$\mathbb{E}Z_i(\theta) = -K(\theta) \le -\epsilon(r).$$

Hence,

$$\left\{\Delta_n(\theta) \ge -\frac{n\epsilon(r)}{2}\right\} \subseteq \left\{\Delta_n(\theta) - \mathbb{E}\Delta_n(\theta) \ge \frac{n\epsilon(r)}{2}\right\}.$$

Hoeffding's inequality yields

$$\Pr\left\{\Delta_n(\theta) \geq -\frac{n\epsilon(r)}{2}\right\} \leq \exp\left(-\frac{n\epsilon(r)^2}{8M_r^2}\right). \tag{25}$$

Choose a cover $\mathcal{C}_r \subset A_r$ with $|\mathcal{C}_r| = \mathcal{N}_r$ and radius $\epsilon(r)/(8L_r)$. By a union bound,

$$\Pr\left\{\exists\theta_c \in \mathcal{C}_r : \Delta_n(\theta_c) \geq -\frac{n\epsilon(r)}{2}\right\} \leq \mathcal{N}_r \exp\left(-\frac{n\epsilon(r)^2}{8M_r^2}\right).$$

When $n \geq n_{\mathrm{sep}}(r,\eta)$, the right-hand side is at most η. On the complementary event, every $\theta \in A_r$ has a center $\theta_c \in \mathcal{C}_r$ such that, by Assumption 3,

$$|\Delta_n(\theta) - \Delta_n(\theta_c)| \leq \sum_{i=1}^{n} L_r\|\theta - \theta_c\| \leq \frac{n\epsilon(r)}{8}.$$

Therefore, with probability at least $1-\eta$,

$$\sup_{\theta\in A_r} \Delta_n(\theta) \leq -\frac{3n\epsilon(r)}{8} \leq -\frac{n\epsilon(r)}{4}. \tag{26}$$

On this event, the posterior mass of A_r can be written as

$$q_n^\star(A_r) = \frac{\displaystyle\int_{A_r} \exp\big(\Delta_n(\theta)\big)\frac{\pi(\theta)}{\pi(\theta^\star)}\lambda(d\theta)}{\displaystyle\int_{\Theta} \exp\big(\Delta_n(\theta)\big)\frac{\pi(\theta)}{\pi(\theta^\star)}\lambda(d\theta)}. \tag{27}$$

For every $\theta \in A_r$,

$$\log\frac{\pi(\theta)}{\pi(\theta^\star)} + \Delta_n(\theta) \leq B(r) - \frac{n\epsilon(r)}{4}.$$

Since $\lambda(A_r) \leq 1$, the numerator of 27 satisfies

$$\int_{A_r} \exp\big(\Delta_n(\theta)\big)\frac{\pi(\theta)}{\pi(\theta^\star)}\lambda(d\theta) \leq \exp\left(B(r) - \frac{n\epsilon(r)}{4}\right). \tag{28}$$

For every $\theta \in U_r$, Assumption 3 and the definition of ρ_r imply

$$\Delta_n(\theta) \geq -\sum_{i=1}^{n} |\ell_\theta(x_i,y_i) - \ell_{\theta^\star}(x_i,y_i)| \geq -\frac{n\epsilon(r)}{8}.$$

Restricting the denominator of 27 to U_r gives

$$\int_{\Theta} \exp\big(\Delta_n(\theta)\big)\frac{\pi(\theta)}{\pi(\theta^\star)}\lambda(d\theta) \geq m(r)\exp\left(-\frac{n\epsilon(r)}{8}\right). \tag{29}$$

Combining 28 and 29,

$$q_n^\star(A_r) \leq \frac{1}{m(r)}\exp\left(B(r) - \frac{n\epsilon(r)}{8}\right).$$

Together with $q_n^\star(A_r) \leq 1$, we obtain

$$q_n^\star(A_r) \leq \min\left\{1, \frac{1}{m(r)}\exp\left(B(r) - \frac{n\epsilon(r)}{8}\right)\right\}.$$

This completes the proof. □

A.3 Proof of Corollary 1

Proof. In the head case, $B(r) \leq 0$, so the posterior concentration bound becomes

$$q_n^\star(A_r) \leq \min\left\{1, \frac{1}{m(r)} \exp\left(-\frac{n\epsilon(r)}{8}\right)\right\}.$$

To guarantee $q_n^\star(A_r) \leq \delta$, it is sufficient that

$$\frac{1}{m(r)} \exp\left(-\frac{n\epsilon(r)}{8}\right) \leq \delta,$$

which gives

$$n \geq \frac{8}{\epsilon(r)} \left(\log \frac{1}{m(r)\delta}\right)_+.$$

In the tail case, $B(r) > 0$, and it is sufficient that

$$\frac{1}{m(r)} \exp\left(B(r) - \frac{n\epsilon(r)}{8}\right) \leq \delta.$$

Rearranging gives

$$n \geq \frac{8}{\epsilon(r)} \left(B(r) + \log \frac{1}{m(r)\delta}\right)_+.$$

This completes the proof. □

A.4 Proof of Theorem 2

Proof. Fix an SFT dataset $\mathcal{D}_n$. For P_X-almost every x, convexity of KL divergence in its second argument gives

$$D_{\mathrm{KL}}\left(P_{\theta^\star}(\cdot \mid x) \| P_{q_n^\star}(\cdot \mid x)\right) \tag{30}$$

$$\leq \int_\Theta D_{\mathrm{KL}}\left(P_{\theta^\star}(\cdot \mid x) \| P_\theta(\cdot \mid x)\right) q_n^\star(d\theta). \tag{31}$$

Splitting the integral over $\Theta \setminus A_r$ and A_r, the contribution from $\Theta \setminus A_r$ is bounded by the local approximation term $\alpha(r)$. For $\theta \in A_r$, Assumption 2 implies

$$D_{\mathrm{KL}}\left(P_{\theta^\star}(\cdot \mid x) \| P_\theta(\cdot \mid x)\right) = \mathbb{E}_{y \sim P_{\theta^\star}(\cdot|x)}\left[\ell_{\theta^\star}(x,y) - \ell_\theta(x,y)\right] \leq M_r.$$

Taking expectation over x therefore gives

$$\mathcal{E}_n \leq \alpha(r) + M_r q_n^\star(A_r).$$

Under the conditions of Theorem 1, with probability at least $1 - \eta$,

$$q_n^\star(A_r) \leq \min\left\{1, \frac{1}{m(r)} \exp\left(B(r) - \frac{n\epsilon(r)}{8}\right)\right\}.$$

Substituting this bound yields

$$\mathcal{E}_n \leq \alpha(r) + M_r \min\left\{1, \frac{1}{m(r)} \exp\left(B(r) - \frac{n\epsilon(r)}{8}\right)\right\}.$$

This completes the proof. □

B PSEUDO CODE

Algorithm 1 PASS: Prior-barrier Aware SFT Instruction Selection

Require: Candidate pool $\mathcal{D}$, reference set $\mathcal{T}$, pretrained model M_0, selection budget n
Ensure: Selected subset $\mathcal{D}_n$ with $|\mathcal{D}_n| = n$

M1: Distinguishing Evidence Estimation

for all $(x, y) \in \mathcal{T}$ **do**
  Encode (x, y) with M_0 to obtain the response representation $z(x, y)$
end for
Hierarchically cluster the reference representations using the coherence criterion
Obtain target concepts $\Theta = \{\theta_1, \ldots, \theta_C\}$ and their reference subsets $\{\mathcal{T}_\theta\}_{\theta \in \Theta}$
for all $\theta \in \Theta$ **do**
  Compute the desired representation p_θ from $\mathcal{T}_\theta$
  Collect wrong-response instructions associated with θ
  Construct the competing representations $\mathcal{N}_\theta = \{n_{\theta 1}, \ldots, n_{\theta K_\theta}\}$
end for
for all $(x_i, y_i) \in \mathcal{D}$ **do**
  Encode (x_i, y_i) to obtain $z_i = z(x_i, y_i)$
  for all $\theta \in \Theta$ **do**
    Compute $s_\theta^+(z_i)$ and $s_\theta^-(z_i)$ using Eq. 9
    Compute distinguishing evidence $\epsilon_{i\theta}$ using Eq. 10
  end for
end for
Form the evidence matrix $[\epsilon_{i\theta}] \in \mathbb{R}^{N \times C}$

M2: Barrier-Aware Instruction Allocation

for all $\theta \in \Theta$ **do**
  for all $(x, y) \in \mathcal{T}_\theta$ **do**
    Generate $\hat{y} \sim M_0(\cdot \mid x)$ and obtain $\hat{z} = z(x, \hat{y})$
  end for
  Compute the prior barrier B_θ using Eq. 11
  Compute the existing target-concept support γ_θ using Eq. 12
end for
Compute the allocation weights $\{A_\theta\}_{\theta \in \Theta}$ using Eq. 13
Initialize $\mathcal{S} \leftarrow \varnothing$
while $|\mathcal{S}| < n$ **do**
  for all $\theta \in \Theta$ **do**
    Compute $R_\theta^+(\mathcal{S})$ and $R_\theta^-(\mathcal{S})$
  end for
  Compute the redundancy term $D(\mathcal{S})$
  Evaluate the selection objective $F(\mathcal{S})$ using Eq. 14
  Select $i^\star$ by the marginal-gain rule in Eq. 15
  $\mathcal{S} \leftarrow \mathcal{S} \cup \{(x_{i^\star}, y_{i^\star})\}$
end while
return $\mathcal{D}_n \leftarrow \mathcal{S}$

C DETAILS OF EXPERIMENTS

C.1 IMPLEMENTATION DETAILS

We evaluate the five capabilities in the downstream task using MMLU (Hendrycks et al., 2021), BBH (Suzgun et al., 2023), GSM8K (Cobbe et al., 2021), TRUTHFULQA (Lin et al., 2022), and TYDIQA (Clark et al., 2020), respectively. Following prior work, each capability benchmark is partitioned into 20% reference data and 80% evaluation data. The reference partitions from all five benchmarks are combined into a single reference set $\mathcal{T}$ representing the desired behaviors of the downstream task, while the remaining examples are used for evaluation. The resulting evaluation sets contain 11,250, 1,080, 200, 632, and 885 examples for MMLU, BBH, GSM8K, TRUTHFULQA, and TYDIQA, respectively. All experiments are conducted on NVIDIA A100 GPUs with 80GB memory.

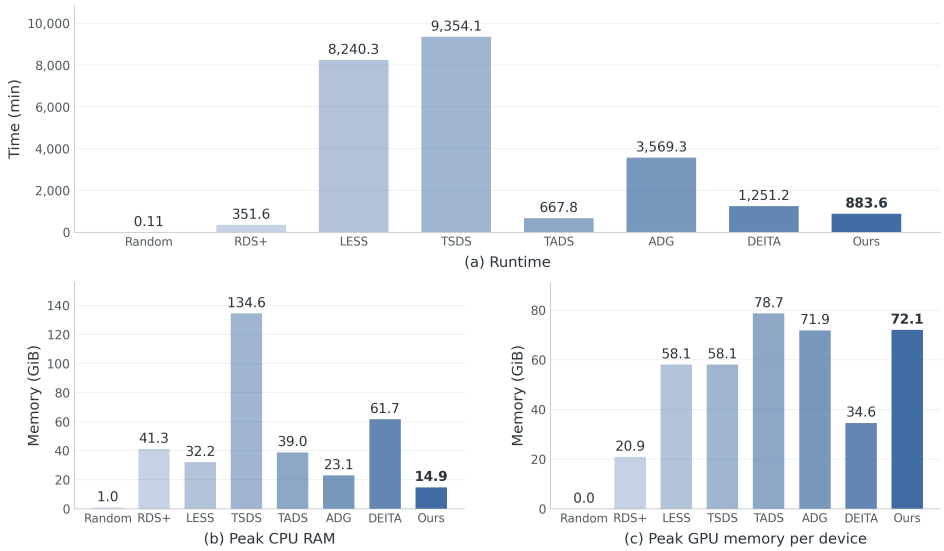


Figure 4: Runtime and memory consumption of all selection methods. (a) Total runtime in minutes. (b) Peak CPU memory. (c) Peak GPU memory per device.

C.2 Runtime and Memory Consumption

Figure 4 summarizes the total runtime and peak memory usage during instruction selection. As shown, LESS, TSDS, and ADG incur substantially higher running time, while our method requires considerably less time. Our method has the lowest CPU memory usage among the non-random selectors and comparable GPU memory usage.

C.3 Selection Composition

Figures 5 and 6 compare the source distributions of the instructions selected under the 5,000 budget. Random selection largely preserves the original pool distribution, whereas PASS substantially reweights the source mixture. On both Mistral-7B and Qwen2.5-7B, PASS increases the proportions of FLAN V2, OASST1, and Dolly while reducing the proportion of WizardLM relative to the candidate pool. The selected distributions also vary across the two backbones, reflecting the model-dependent nature of the prior barrier and instruction evidence used by PASS. In contrast, several baselines exhibit much stronger concentration on a small number of sources, such as TADS and TSDS. These results suggest that PASS adapts the composition of the selected subset rather than simply preserving the source distribution or concentrating on a single source.

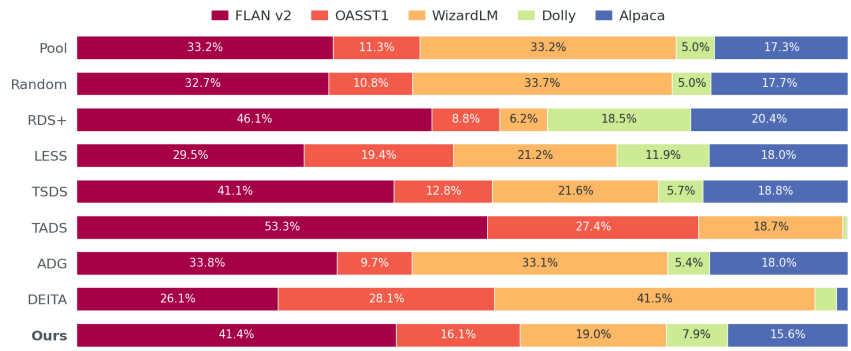


Figure 5: Distribution of instructions selected by Mistral-7B under the 5K selection budget.

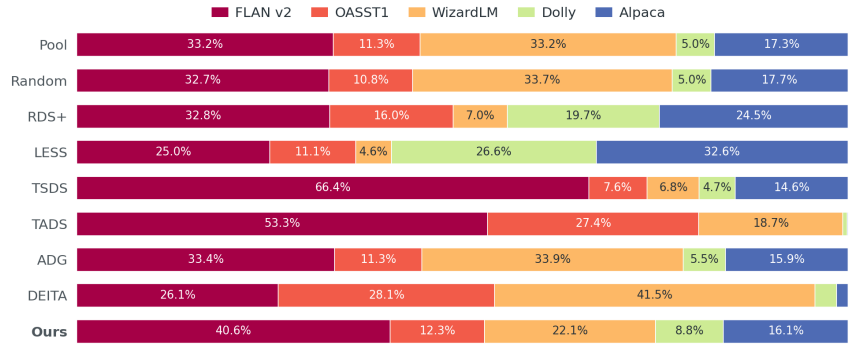


Figure 6: Distribution of instructions selected by Qwen2.5-7B under the 5K selection budget.