Minibatch Selection for Language Models via Partition Matroid Constrained Gradient Matching

Prayas Agrawal*, Prateek Chanda*, Ishita Khatri, Ganesh Ramakrishnan, Bamdev Mishra, Pratik Jawanpuria

ICML 2026 Poster 64010

Joint cross-domain minibatch selection under per-domain budgets, with an RSC/RSM-based approximation guarantee.

* Equal Contribution

Background: minibatch selection for heterogeneous fine-tuning

Large minibatches improve LLM convergence but incur substantial GPU memory *and* wall-clock cost per step.

- Recent work (GREATS, CoLM) selectively trains on a *subset* of each minibatch – shrinking the backward pass to cut both memory and time while preserving update quality.
- Fine-tuning corpora are heterogeneous: $\mathcal{D}_{\mathrm{tr}} = \biguplus_{c=1}^{C} \mathcal{V}_c$ (C domains).
- Per-step selection must therefore be effective, domain-aware, *and* cheap (it runs every step).

Question

At each step t, given a candidate batch $\mathcal{B}_t = \biguplus_{c=1}^{C} \mathcal{B}_t^c$, which subset $\mathcal{S}_t \subseteq \mathcal{B}_t$ should we train on?

Motivation: how existing methods handle domain heterogeneity

Two families of approaches:

- **Per-domain independent selection** (e.g., per-domain GREATS / CoLM). Each domain runs its own selector; selections across domains are uncoupled.
- **Continuous mixture learning** via auxiliary proxy LLMs (DoReMi, RegMix, CLIMB). Learns continuous weights $\boldsymbol{w} \in \Delta^C$ over domains, but the proxy must mirror the main LLM and adds substantial compute / wall-clock overhead.

Gap

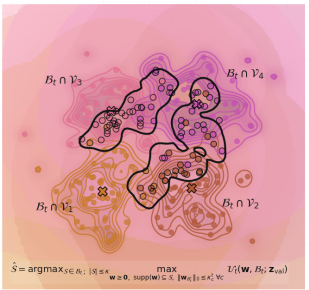
Neither family selects jointly across domains, and neither is discrete and proxy-free with guarantees.

Problem formulation

At training step t, $\mathcal{B}_t = \biguplus_{c=1}^{C} \mathcal{B}_t^c$ with $\mathcal{B}_t^c = \mathcal{B}_t \cap \mathcal{V}_c$. Select $\mathcal{S}_t \subseteq \mathcal{B}_t$ to solve

$$\max_{\mathcal{S} \subseteq \mathcal{B}_t} \mathcal{U}_t(\mathcal{S}) \quad \text{s.t.} \quad |\mathcal{S} \cap \mathcal{B}_t^c| \leq \kappa_c^t \;\; \forall c, \quad \sum_{c=1}^{C} \kappa_c^t = \kappa.$$

- $\mathcal{U}_t(\mathcal{S})$: a time-dependent set utility.
- $\{\kappa_c^t\}$: per-domain capacities, κ the total per-step budget.
- The per-domain capacity constraints form a partition matroid on $\mathcal{B}_t$ (formalized later).



Validation-guided utility and its marginal gain

Following Koh & Liang and the GREATS line of work, define the utility as the validation-loss reduction after one SGD step using $\mathcal{S}$:

$$\mathcal{U}_t(\mathcal{S}) \;:=\; \ell(\boldsymbol{z}_{\mathrm{val}};\boldsymbol{\theta}_t) \;-\; \ell\big(\boldsymbol{z}_{\mathrm{val}};\,\tilde{\boldsymbol{\theta}}_{t+1}(\mathcal{S})\big), \qquad \tilde{\boldsymbol{\theta}}_{t+1}(\mathcal{S}) = \boldsymbol{\theta}_t - \eta_t \sum_{\boldsymbol{z}\in\mathcal{S}} \boldsymbol{g}_{\boldsymbol{\theta}_t}(\boldsymbol{z}).$$

A second-order Taylor expansion yields the marginal gain

$$\boldsymbol{\Delta}\mathcal{U}_t(\boldsymbol{z}_i \mid \mathcal{S}) \;\approx\; \eta_t \,\langle \boldsymbol{g}_i, \boldsymbol{g}_{\mathrm{val}} \rangle \;-\; \eta_t^2 \,\langle \boldsymbol{g}_i,\; \boldsymbol{\mathcal{H}}_{\boldsymbol{z}_{\mathrm{val}}}(\boldsymbol{\theta}_t) \textstyle\sum_{\boldsymbol{z}\in\mathcal{S}} \boldsymbol{g}_{\boldsymbol{\theta}_t}(\boldsymbol{z}) \rangle.$$

First term: importance score of $\boldsymbol{z}_i$ wrt. $\boldsymbol{z}_{\mathrm{val}}$. Second term: Hessian-weighted similarity to $\mathcal{S}$ (promotes diversity).

GREATS as a binary-weight special case, and its limitations

Under $\mathcal{H} \approx \boldsymbol{I}$, GREATS (Wang et al., 2024) greedily selects

$$\boldsymbol{z}^* = \arg\max_{\boldsymbol{z} \in \mathcal{B}_t \setminus \mathcal{S}} \langle \boldsymbol{g}_{\theta_t}(\boldsymbol{z}),\, \boldsymbol{g}_{\theta_t}(\boldsymbol{z}_{\mathrm{val}}) \rangle \;-\; \eta_t \, \langle \boldsymbol{g}_{\theta_t}(\boldsymbol{z}),\, \textstyle\sum_{\boldsymbol{z}_j \in \mathcal{S}} \boldsymbol{g}_{\theta_t}(\boldsymbol{z}_j) \rangle.$$

- Greedy with *binary* $\{0, 1\}$ selection: each sample is either in $\mathcal{S}$ or out.
- Marginal gain can be negative $\Rightarrow$ non-monotone objective.
- Only *conjectured* weakly submodular; no approximation guarantee for greedy.
- Single overall cardinality budget; per-domain variants (`ID`) run GREATS inside each domain $\Rightarrow$ no cross-domain coupling.

Two issues to fix

(i) non-monotone objective; (ii) no joint selection across domains.

PartitionSel: relax to non-negative weights over batch samples

Replace the binary selection with non-negative weights $\boldsymbol{w} \in \mathbb{R}_+^{|\mathcal{B}_t|}$ over batch samples. Define

$$\mathcal{U}_t(\boldsymbol{w}) \;=\; \langle \boldsymbol{w},\, \bar{\boldsymbol{\mu}}_{\theta_t} \rangle \;-\; \tfrac{\eta_t}{2} \langle \boldsymbol{w},\, \boldsymbol{K}_{\theta_t} \boldsymbol{w} \rangle,$$

with per-sample gradient matrix

$$\boldsymbol{G}_{\theta_t} = \left[\boldsymbol{g}_{\theta_t}(\boldsymbol{z}_1), \ldots, \boldsymbol{g}_{\theta_t}(\boldsymbol{z}_{|\mathcal{B}_t|}) \right]^\top, \quad \boldsymbol{K}_{\theta_t} = \boldsymbol{G}_{\theta_t} \boldsymbol{G}_{\theta_t}^\top, \quad \bar{\boldsymbol{\mu}}_{\theta_t} = \boldsymbol{G}_{\theta_t} \bar{\boldsymbol{g}}_{\theta_t} + \tfrac{\eta_t}{2} \boldsymbol{\nu}_{\theta_t},$$

where $\boldsymbol{\nu} = \mathrm{diag}(\boldsymbol{K})$ and $\bar{\boldsymbol{g}}_{\theta_t}$ averages $\boldsymbol{g}_{\theta_t}(\boldsymbol{z}_{\mathrm{val}})$ over the validation set. The induced set function $f(\mathcal{S}) = \max_{\boldsymbol{w} \geq \boldsymbol{0},\, \mathrm{supp}(\boldsymbol{w}) \subseteq \mathcal{S}} \mathcal{U}_t(\boldsymbol{w})$ is

- monotone (enlarging $\mathcal{S}$ enlarges the feasible region; $\mathcal{U}_t(\boldsymbol{0}) = 0$);
- strictly concave in $\boldsymbol{w}$ on its support since $\boldsymbol{K} \succ 0$ (high-dimensional LLM gradients);
- recovers GREATS exactly when $\boldsymbol{w} \in \{0, 1\}^{|\mathcal{B}_t|}$ (the $\frac{\eta_t}{2}\boldsymbol{\nu}$ correction cancels self-interaction).

Cross-domain coupling via a partition matroid

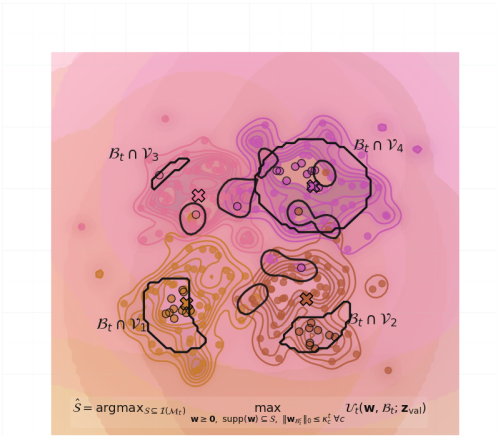
Define the partition matroid $\mathcal{M}_t = (\mathcal{B}_t, \mathcal{I}_t)$, with

$$\mathcal{I}_t = \{\mathcal{S} \subseteq \mathcal{B}_t : |\mathcal{S} \cap \mathcal{B}_t^c| \leq \kappa_c^t \ \forall c\}.$$

The domain-aware selection problem is

$$\max_{\mathcal{S} \in \mathcal{I}_t} f_{\mathcal{M}}(\mathcal{S}), \qquad f_{\mathcal{M}}(\mathcal{S}) = \max_{\substack{\boldsymbol{w} \geq \mathbf{0} \\ \mathrm{supp}(\boldsymbol{w}) \subseteq \mathcal{S}}} \mathcal{U}_t(\boldsymbol{w}).$$

A *single* utility couples all domains; the per-domain budgets are exactly the matroid's independence constraint.



Joint, matroid-constrained selection across all domains.

Algorithm and theoretical guarantee

OMP for weakly submodular maximization under a matroid:

1. Init. support $\mathcal{S}_0 \leftarrow \emptyset$, $\boldsymbol{w} \leftarrow \mathbf{0}$; $\boldsymbol{g} \leftarrow \nabla_{\boldsymbol{w}} \mathcal{U}_t(\boldsymbol{w})$.
2. For $i = 1, \ldots, \kappa$:
   - Pick M_i, a maximal indep. set of $\mathcal{M}_t / \mathcal{S}_{i-1}$, maximizing $\sum_{\boldsymbol{z} \in M_i} \max(0, [\boldsymbol{g}]_{\boldsymbol{z}})$.
   - Sample $\boldsymbol{z} \sim \mathrm{Unif}(M_i)$; set $\mathcal{S}_i \leftarrow \mathcal{S}_{i-1} \cup \{\boldsymbol{z}\}$.
   - Refit $\boldsymbol{w}_{\mathcal{S}_i} \leftarrow \arg\max_{\boldsymbol{w} \geq \mathbf{0},\, \mathrm{supp}(\boldsymbol{w}) \subseteq \mathcal{S}_i} \mathcal{U}_t(\boldsymbol{w})$ via APGA; refresh $\boldsymbol{g}$.

Theorem (Thm. 4.4)

$f_{\mathcal{M}}$ is monotone and γ-weakly submodular under $\mathcal{M}_t$ ($\boldsymbol{K}_{\boldsymbol{\theta}_t} \succ 0$ via RSC/RSM). The OMP iterate $\hat{\mathcal{S}}_t$ satisfies

$$\mathbb{E}\Big[f_{\mathcal{M}}(\hat{\mathcal{S}}_t)\Big] \;\geq\; (1+\widetilde{C}_1/c_{2\kappa})^{-2} f_{\mathcal{M}}(\mathcal{S}^*),$$

with $c_{2\kappa}$, $\widetilde{C}_1$ the RSC/RSM constants; ratio independent of η_t.

Empirical results

Math reasoning

Method	Avg. acc.
Random	60.14
CoLM	61.19
GREATS	62.40
Indep. per-domain	62.63
PartitionSel	**63.63**

Qwen2.5-3B, MetaMathQA, 50% subset.

Molecule generation

Method	BLEU	Edit
Random	0.62	30.77
CoLM	0.59	30.64
GREATS	0.62	30.84
IWD	0.63	31.07
PartitionSel	**0.66**	**29.71**

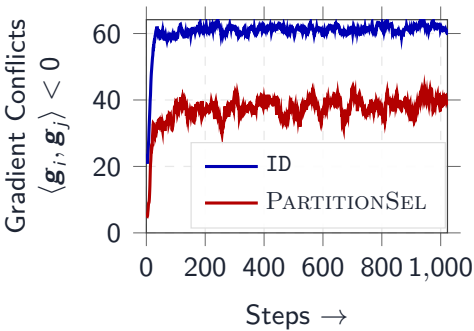
Qwen2.5-3B, Mol-Instructions, 25% subset.

vs. DoReMi

Dataset	DoReMi	Ours
MMLU-Math	53.08	**60.57**
AQuA	58.66	**61.02**
SAT	72.27	**76.36**
Average	61.34	**65.99**

+4.65 avg. over DoReMi proxy mixture.

Mechanism: fewer conflicting gradients in the chosen batch



Count of selected pairs with $\langle g_i, g_j \rangle < 0$ across training steps.

Joint, matroid-constrained selection sharply reduces gradient conflicts relative to per-domain independent selection.

Cross-domain coupling translates into more compatible training updates.

Take-aways

- Subset selection shrinks the backward pass: lower GPU memory *and* wall-clock per step, at no cost to convergence.
- Joint subset selection across domains via a partition matroid — past work selects per-domain or learns a proxy-LLM mixture.
- Our objective is monotone and weakly submodular; OMP has an approximation guarantee (Thm. 4.4).
- Beats GREATS, CoLM, ID, IWD, and DoReMi — with no auxiliary proxy LLM.

Thank you.

Poster 64010.