MACE: AFFINITY AND CONFLICT-AWARE DATA SELECTION FOR MULTILINGUAL INSTRUCTION TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multilingual instruction tuning improves the ability of large language models to follow instructions across languages. The training mixture shapes both the supervision available to each language and interactions between languages. Data selection is therefore central to effective multilingual training. Through shared parameters, however, an example aligned with learning directions in its own language may induce updates that oppose those of other languages. Selection requires reliable references for diverse within-language directions and an explicit assessment of cross-language opposition. We introduce MACE (Multilingual Affinity and Conflict-aware Example Selection), a gradient-based data selection framework addressing both requirements. MACE constructs multiple gradient prototypes for each language and measures a candidate's in-language affinity through alignment with its closest prototype. It assesses cross-language conflict exposure using the Magnitude-weighted Gradient Conflict Ratio (MGCR), which measures the share of gradient-product magnitude contributed by opposing coordinates. Combining affinity with an exposure penalty yields a score for ranking and retaining candidates within each language. Across five multilingual benchmarks, MACE + KMC improves the average score over LangGPS + KMC by a relative 1.99% on Qwen2.5-7B and 4.59% on Llama-3.1-8B.

1 INTRODUCTION

Multilingual instruction tuning improves performance across languages, extending the reach of large language models to users with different linguistic backgrounds (Muennighoff et al., 2023; Weber et al., 2024). A shared model can enable cross-language transfer, but its effectiveness depends on the composition of the training mixture. Constructing an effective multilingual training set from a larger candidate corpus is therefore a central question. Under a limited data budget, selection must retain examples that support learning in their own language while accounting for updates that may oppose learning in others through shared parameters (Choenni et al., 2023; Wang et al., 2020).

Existing selectors consider quality, complexity, diversity, target-task influence, and language separability (Li et al., 2024b; Liu et al., 2024b; Xia et al., 2024; Ye et al., 2026). Selecting for both within-language learning and cross-language interaction raises two further challenges.

Challenge I: Establishing reliable in-language references. Instructions within a language span different tasks and response forms, so their gradients need not follow one direction. Assessing a candidate's alignment requires references that capture this diversity.

Reliable in-language references establish how a candidate aligns with its own language, but that alignment alone cannot establish its suitability for multilingual training. The same update also interacts with other languages through shared parameters.

Challenge II: Uncovering cross-language conflict masked by net alignment. A candidate gradient can align positively with another language's reference gradient while opposing it on individual parameter coordinates. Agreeing contributions can offset these opposing components in the net inner product. Assessing cross-language interactions therefore requires separating the two contributions and quantifying the share of interaction magnitude associated with opposition.

Figure 1 illustrates both challenges. With disjoint held-out Aya gradients, four prototypes increase average in-language alignment from 0.3488 to 0.4466 across 16 languages, compared with a single

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.

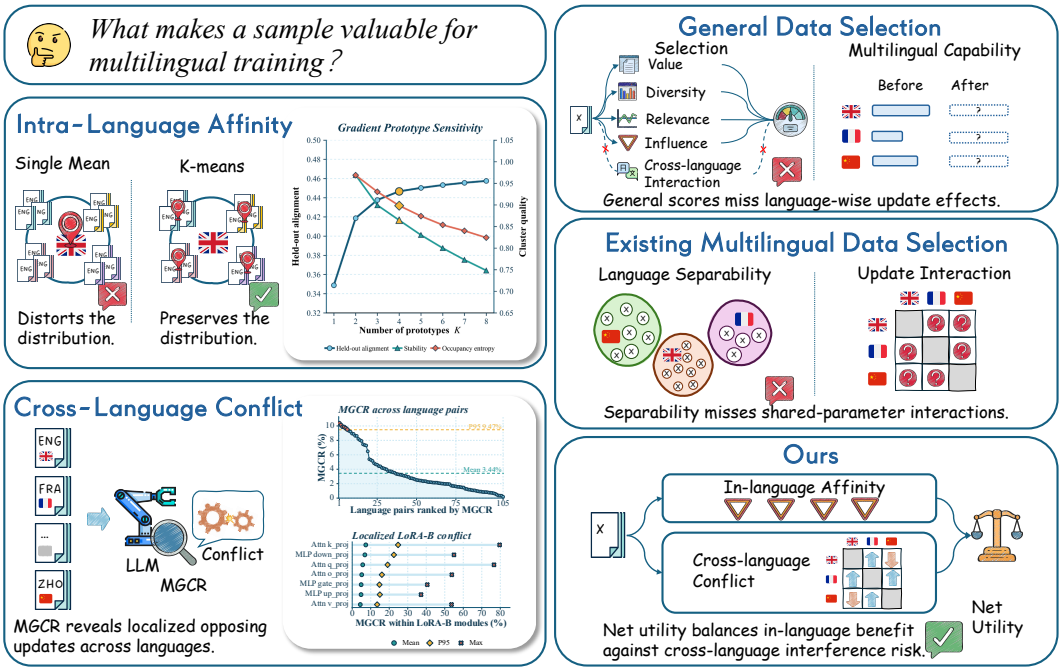

Figure 1: Motivation and overview of MACE. Left: multiple gradient prototypes improve held-out coverage over a single mean across 16 Aya languages; coordinate-level analysis reveals localized opposition across 15 Belebele languages (105 pairs). Right: prototype affinity and cross-language conflict exposure score candidates for multilingual selection.

mean (Singh et al., 2024). Semantically parallel Belebele examples show opposition that varies by language pair and concentrates in particular parameter subsets (Bandarkar et al., 2024).

To address these challenges, we propose MACE (Multilingual Affinity and Conflict-aware Example Selection), a selection framework. Multiple gradient prototypes provide in-language references, and alignment with the closest prototype defines affinity. The Magnitude-weighted Gradient Conflict Ratio (MGCR) measures the fraction of gradient-product magnitude contributed by opposing coordinates. MGCR between language means identifies conflict reference languages; each candidate is then scored against their mean gradients. MACE combines affinity and an exposure penalty to rank candidates within each language.

Across five multilingual benchmarks, MACE improves both K-means Clustering (KMC) and LESS as downstream selectors. With KMC, MACE exceeds LangGPS by a relative 1.99% on Qwen2.5-7B and 4.59% on Llama-3.1-8B in the five-benchmark average.

Our contributions are threefold:

- We provide gradient-based evidence for both challenges: multiple prototypes improve held-out in-language coverage, and cross-language opposition concentrates in parameter subsets.
- We introduce MACE, a multilingual data selection framework that combines local prototype affinity with MGCR-based conflict exposure to rank candidates within each language.
- We evaluate MACE using instruction data in 16 languages, two backbones, and five multilingual benchmarks, showing improvements across backbones and downstream selectors.

2 Related Work

Joint multilingual instruction tuning. Multitask fine-tuning supports cross-lingual generalization (Muennighoff et al., 2023), with performance depending on language composition (Weber et al., 2024; Ji & Chen, 2024; Kew et al., 2024). Multilingual examples improve English-centered tuning (Shaham et al., 2024) and can outperform monolingual tuning under a fixed budget (Chen et al.,

2024b). Data-sharing and interference studies characterize how jointly trained languages interact through shared parameters (Choenni et al., 2023; Wang et al., 2020).

Instruction data selection. Quality and coverage are common selection objectives (Albalak et al., 2024). Quality is assessed through manual curation (Zhou et al., 2023), model ratings, uncertainty, or linguistic indicators (Chen et al., 2024a; Liu et al., 2024a; Cao et al., 2023); difficulty and length provide additional signals (Li et al., 2024b; Zhao et al., 2024). Small-model scoring reduces assessment cost (Li et al., 2024a; Mekala et al., 2024). To balance quality with coverage, methods combine quality and diversity (Liu et al., 2024b; Bukharin et al., 2024; Zhang et al., 2025a), representing diversity through tags, clusters, sparse features, or novelty (Lu et al., 2024; Ge et al., 2024; Yang et al., 2025b;c). Submodular objectives and label-graph information gain offer further ways to construct diverse instruction subsets (Renduchintala et al., 2024; Chen et al., 2025a).

Target-oriented selection assesses relevance to a distribution (Xie et al., 2023), task, or validation objective (Liu et al., 2024c; Feng et al., 2025; Jain et al., 2026). Influence functions and TracIn relate examples to target predictions or losses (Koh & Liang, 2017; Pruthi et al., 2020), motivating instruction selectors based on optimizer-aware or capability-balanced influence (Xia et al., 2024; Dai et al., 2025). Other objectives use preference-validation rewards (Wu et al., 2025) or non-differentiable metrics (Wang et al., 2025a). Example contribution can also be assessed across subsets through Shapley values (Ghorbani & Zou, 2019; Schoch et al., 2023), or through an example's value as an in-context demonstration (Agarwal et al., 2025; Yang et al., 2026; Han & Huang, 2026).

Optimization-based selection uses gradient approximation, gradient matching, or validation performance (Mirzasoleiman et al., 2020; Killamsetty et al., 2021a;b). Gradient clustering captures coverage (Zhang et al., 2025b; Pan et al., 2024; Wang et al., 2025c), while gradient graphs, joint influence, and instruction-graph propagation model sample interactions (Zhao et al., 2025; Wei et al., 2025; Han et al., 2025). Selection can also follow training dynamics, using small-model trajectories (Yang et al., 2024) or updating subsets from losses and gradients during training (Qin et al., 2024; Wang et al., 2025b; Lin et al., 2025). Extensions select individual tokens or data–layer combinations (Pang et al., 2025; Yao et al., 2026); large-scale studies examine how candidate-pool size and diversity affect selection (Ivison et al., 2025; Xia et al., 2025).

Multilingual data selection. MuRating scores data quality for multilingual pretraining, while M-DaQ combines instruction quality and cross-lingual diversity for multilingual instruction tuning (Chen et al., 2025b; Zhao et al., 2026). For multilingual instruction tuning, LangGPS performs data selection based on language separability (Ye et al., 2026), whereas MACE jointly models local in-language alignment and cross-language opposition. Related optimization methods group languages, project conflicting gradients, or coordinate updates (Wang et al., 2023; Yu et al., 2020; Wang et al., 2021). CONGRAD uses cross-lingual consensus gradients for preference-data selection (Li et al., 2026).

3 METHODOLOGY

3.1 OVERVIEW AND PRELIMINARIES

Let $\mathcal{D} = \bigcup_{\ell \in \mathcal{L}} \mathcal{D}_\ell$ be the candidate dataset, where $\mathcal{D}_\ell$ contains examples in language ℓ and $|\mathcal{L}| \geq 2$. For a budget $1 \leq P \leq |\mathcal{D}|$, MACE returns $\mathcal{S} \subseteq \mathcal{D}$ with $|\mathcal{S}| = P$ by scoring in-language prototype affinity and cross-language conflict exposure (Figure 2).

For each language, we uniformly sample a nonempty reference-anchor set $\mathcal{A}_\ell$ that is disjoint from the candidate pool $\mathcal{D}_\ell$. Let $\theta \in \mathbb{R}^d$ denote the shared scoring parameters, instantiated as LoRA adapter parameters in our experiments, with all other model parameters fixed. At a common scoring state, we compute anchor and candidate gradients $g_x = \nabla_\theta L(x; \theta)$. Here, L is the negative log-likelihood averaged over response tokens; prompt tokens are excluded from the loss. We write $\hat{g}_x = \mathrm{norm}(g_x)$, where $\mathrm{norm}(v) = v/\|v\|_2$ for $v \neq 0$ and $\mathrm{norm}(0) = 0$. The anchors provide local prototypes for affinity scoring and language mean gradients for exposure scoring.

In our implementation, scoring uses a newly initialized rank-8 LoRA adapter on each base model. The adapter acts on attention and feed-forward projections and remains at its initial state throughout gradient collection. Appendix A specifies the adapter initialization and gradient extraction.

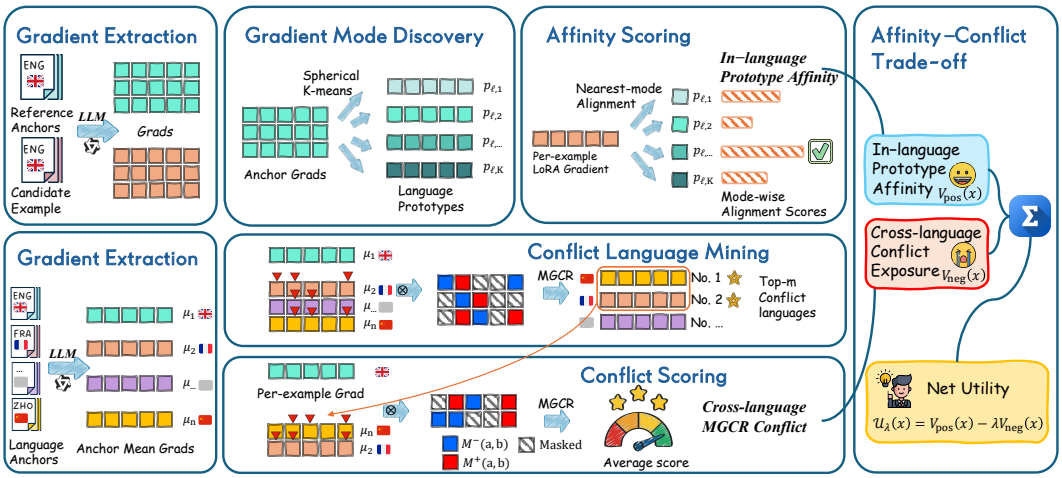

Figure 2: MACE overview for instruction data in 16 Aya languages. Language-specific gradient prototypes provide local references for affinity estimation. MGCR between language mean gradients identifies conflict reference languages; candidate gradients are then compared with these references to compute conflict exposure. An exposure penalty combines with prototype affinity to score candidates for selection.

The reference prototypes and language means are constructed once and reused across candidates. MGCR between language means identifies a conflict reference set for each language. Each candidate gradient is compared with local prototypes for affinity and with the selected reference-language means for conflict exposure, at a common scoring state.

3.2 In-Language Prototype Affinity

We represent each language with K local gradient prototypes. We apply a sparse random projection (Clarkson & Woodruff, 2013) to the anchor gradient directions, normalize the projected vectors, and cluster them using spherical K-means (Dhillon & Modha, 2001). Given the resulting groups $\{\mathcal{A}_{\ell,j}\}_{j=1}^{K}$, we construct each prototype from the mean of the raw gradients within each group in the original parameter coordinates:

$$p_{\ell,j} = \mathrm{norm}\left(\frac{1}{|\mathcal{A}_{\ell,j}|}\sum_{a\in\mathcal{A}_{\ell,j}} g_a\right). \tag{1}$$

For a candidate x in language ℓ_x, we define its prototype affinity as

$$V_{\mathrm{pos}}(x) = \max_{1\le j\le K} \hat{g}_x^{\top} p_{\ell_x,j}. \tag{2}$$

This cosine-based score lies in $[-1, 1]$ and measures alignment with the candidate's most compatible local reference; Appendix A specifies how empty groups are handled.

Taking the maximum lets a candidate align with one local direction even when it differs from the language mean. Averaging across prototypes would combine compatible and incompatible directions. The projection determines cluster membership, while affinity is evaluated in the original scoring coordinates.

Clustering before averaging retains references for distinct groups of instructions. A candidate can therefore receive high affinity through one group without requiring agreement with the other groups represented in its language.

3.3 Cross-Language Conflict Exposure

To assess cross-language interactions, we first identify conflict reference languages using their mean anchor gradients, $\mu_k = |\mathcal{A}_k|^{-1}\sum_{a\in\mathcal{A}_k} g_a$, and then compare each candidate gradient with the selected language means. Here, $-\mu_k$ is the descent direction for that language's mean anchor loss.

Let Ω_k index the D largest entries of μ_k in absolute value, with $1 \leq D \leq d$ and ties resolved by a fixed coordinate order. This language-specific coordinate set focuses on the strongest reference components and is shared by all candidates scored against language k. Because each reference language selects its own coordinates, the same candidate can exhibit different patterns of agreement and opposition across languages, even when its in-language affinity remains unchanged.

Magnitude-weighted conflict. Agreeing and opposing coordinate contributions can offset one another in the gradient inner product. We separate their magnitudes to quantify the opposing share of the interaction. For gradients a, b, define these magnitudes on Ω as

$$M_\Omega^-(a,b) = \sum_{i\in\Omega}[-a_i b_i]_+, \qquad M_\Omega^+(a,b) = \sum_{i\in\Omega}[a_i b_i]_+, \tag{3}$$

where $[z]_+ = \max(z, 0)$. The Magnitude-weighted Gradient Conflict Ratio (MGCR) is

$$C_\Omega(a,b) = \frac{M_\Omega^-(a,b)}{M_\Omega^-(a,b) + M_\Omega^+(a,b)}. \tag{4}$$

MGCR quantifies the fraction of interaction magnitude contributed by opposing gradient components and is invariant to positive rescaling of either gradient. A zero total interaction magnitude receives an MGCR score of zero.

For coordinate products 3 and -1, the net inner product is positive, but MGCR is $1/(3+1) = 0.25$. It records the opposing share; total interaction magnitude enters separately in the reference-loss analysis below.

Exposure scoring. For each language ℓ, we rank other languages by the mean-gradient conflict $C_{\Omega_k}(\mu_\ell, \mu_k)$. For an integer $m \geq 1$, let $\mathcal{T}_m(\ell)$ contain the $q = \min(m, |\mathcal{L}| - 1)$ highest-scoring languages. This reference set is fixed for all candidates in language ℓ. For candidate x, define $C_k(x) = C_{\Omega_k}(g_x, \mu_k)$ for $k \in \mathcal{T}_m(\ell_x)$. Raw exposure averages these scores:

$$R(x) = \frac{1}{|\mathcal{T}_m(\ell_x)|} \sum_{k\in\mathcal{T}_m(\ell_x)} C_k(x). \tag{5}$$

Language-level conflict determines the reference set, while candidate-level MGCR measures each example's opposition to these references.

3.4 Affinity–Exposure selection

We allocate a selection budget of P examples in proportion to language data sizes, using integer quotas summing to P. Affinity and exposure determine which examples fill each quota. Within each original language pool $\mathcal{D}_\ell$ with at least two examples, we convert raw exposure scores into normalized ranks:

$$V_{\text{neg}}(x) = \frac{\text{rank}_{\mathcal{D}_\ell}(R(x)) - 1}{|\mathcal{D}_\ell| - 1}. \tag{6}$$

Ranks are ascending, with average ranks for ties; singleton pools receive zero. This places exposure on a $[0, 1]$ scale within each language.

We combine prototype affinity with an exposure penalty to define the selection score:

$$\mathcal{U}_\lambda(x) = V_{\text{pos}}(x) - \lambda V_{\text{neg}}(x), \qquad \lambda \geq 0. \tag{7}$$

MACE retains the highest-scoring candidates within each language quota and merges them into $\mathcal{S}$ of size P. The weight λ controls the exposure penalty. Appendix A gives projection, clustering, quota-allocation, and parameter settings.

At $\lambda > 0$, equal affinity favors lower exposure, and equal exposure favors higher affinity. Fixed quotas determine each language's allocation, while the score determines which examples are retained.

3.5 Optimization Interpretation of MGCR

To connect MGCR with cross-language optimization, we analyze how a candidate-gradient step changes the mean anchor loss of another language. Let $a = g_x$ be the candidate gradient and $b = \mu_k$ the mean anchor gradient of reference language k. With $\Omega = \Omega_k$, their inner product on the selected coordinates satisfies

$$\sum_{i\in\Omega} a_i b_i = \big(M_\Omega^+ + M_\Omega^-\big)\big(1 - 2C_\Omega(a,b)\big), \tag{8}$$

where $M_\Omega^\pm = M_\Omega^\pm(a,b)$.

Along the update direction $-g_x$, the negative of this inner product gives the selected-coordinate contribution to the directional derivative of the reference loss. Proposition 1 uses this relation to bound the loss change after a finite step.

Proposition 1 (Reference-loss interval). Let $F_k^{\mathcal{A}}(\theta) = |\mathcal{A}_k|^{-1}\sum_{z\in\mathcal{A}_k} L(z;\theta)$ be the mean anchor loss, so that $\nabla F_k^{\mathcal{A}}(\theta) = \mu_k$. Consider $\theta^+ = \theta - \eta g_x$ with $\eta > 0$, and write $\Delta F_k^{\mathcal{A}}(x) = F_k^{\mathcal{A}}(\theta^+) - F_k^{\mathcal{A}}(\theta)$. Define $S_k(x) = \sum_{i\in\Omega_k} |\mu_{k,i} g_{x,i}|$ and $T_k(x) = \sum_{i\notin\Omega_k} \mu_{k,i} g_{x,i}$. If $\nabla F_k^{\mathcal{A}}$ is β_k-Lipschitz along the update segment in these parameter coordinates, then

$$\big|\Delta F_k^{\mathcal{A}}(x) - \eta S_k(x)\big(2C_k(x) - 1\big)\big| \le \eta|T_k(x)| + \frac{\beta_k\eta^2}{2}\|g_x\|_2^2. \tag{9}$$

Proposition 1 shows that, for a fixed positive interaction magnitude $S_k(x)$ and step size η, MGCR directly characterizes the selected-coordinate first-order contribution to reference-loss change. This contribution increases linearly with MGCR and becomes positive when $C_k(x) > 1/2$. The bound accounts for the remaining coordinates and loss curvature. Appendix D gives the proof and a candidate-ordering condition that compares the MGCR gap with the error bounds for both candidates.

4 Experiments

4.1 Experimental Setup

Models and data budget. We evaluate Qwen2.5-7B (Yang et al., 2025a) and Llama-3.1-8B (Grattafiori et al., 2024) using Aya instruction data from 16 languages. MACE retains 25% of the candidate corpus; a separate downstream selector retains 20% of that subset, yielding a 5% final budget. We also train on the full corpus for reference. All methods use three epochs of LoRA (Hu et al., 2021) fine-tuning in LLaMA-Factory (Zheng et al., 2024) on two RTX 4090 GPUs per run.

Evaluation and reporting. We evaluate XNLI (Conneau et al., 2018), XStoryCloze (Lin et al., 2022), mMMLU (multilingual MMLU; Hendrycks et al., 2020), XQuAD (Artetxe et al., 2020), and XLSum (Hasan et al., 2021). Each benchmark evaluates ten languages with equal numbers of examples per language; Appendix A lists the languages and sample counts. Metrics are accuracy for the first three, normalized containment accuracy for XQuAD following LangGPS (Ye et al., 2026), and ROUGE-L F1 (Lin, 2004) for XLSum. Across runs, we vary seeds for reference sampling, random projection, clustering, stochastic downstream selection, and fine-tuning. Scores are three-run means on a 0–100 scale; Avg. is the unweighted five-benchmark mean. Gains are absolute score-point differences unless explicitly marked as relative percentages.

Comparison methods. Final-subset baselines are Random, K-means Clustering (KMC), LESS (Xia et al., 2024), MTLD-based lexical-diversity selection (McCarthy & Jarvis, 2010), and IFD (Li et al., 2024b). We compare MACE before KMC and LESS with each selector alone, and MACE + KMC with LangGPS + KMC as a multilingual selection baseline.

Implementation details. We sample up to 512 reference anchors per language to construct the prototypes and mean gradients. These anchors are separate from the candidate pool used for pre-selection. We use $K = 4$ prototypes per language, $D = 1024$ reference coordinates, $m = 3$ conflict reference languages per language, and an exposure-penalty weight of $\lambda = 2.5$ across both backbones and selectors. Settings and sensitivity appear in Appendix A.6 and Section 4.6.

4.2 Main Results

Table 1 compares 5% selection with full-data training. Among 5% methods, MACE + KMC leads on Qwen2.5-7B and MACE + LESS on Llama-3.1-8B in five-benchmark average.

Against LangGPS + KMC, MACE + KMC improves all five benchmarks on both backbones, with average gains of 1.212 points on Qwen2.5-7B and 2.102 points on Llama-3.1-8B (relative increases of 1.99% and 4.59%). Adding MACE to KMC raises its average by 0.891 and 2.314 points, respectively; adding it to LESS raises the average by 1.590 and 1.664 points.

We pair each selection method with the same downstream selector and final data budget.

Under the shared three-epoch setup, MACE + KMC at the 5% budget exceeds full-data training on Qwen2.5-7B (62.036 versus 58.692). On Llama-3.1-8B, full-data training gives the higher average (51.362 versus 47.858).

Selection cost. Appendix A.8 compares selection time and downstream performance for MACE + LESS and standalone LESS.

Table 1: Main results. Selection methods use 5% of the data and report three-run mean ± SD; Full uses 100% and shows means only. Avg. is the unweighted five-benchmark mean. Bold marks the best 5% mean per backbone.

Method	XNLI	XStoryCloze	mMMLU	XQuAD	XLSum	Avg.
Qwen2.5-7B						
Full (100%)	74.3	71.11	55.3	73.1	19.62	58.692
Random	74.8 ±0.71	74.28 ±1.24	55.0 ±0.63	81.0 ±0.51	18.93 ±0.28	60.820 ±0.53
KMC	75.2 ±1.33	75.61 ±0.56	55.5 ±0.79	80.7 ±1.22	18.72 ±0.17	61.145 ±0.64
LESS	74.9 ±0.15	72.15 ±0.19	55.3 ±0.13	80.1 ±0.21	19.35 ±0.11	60.368 ±0.12
MTLD	74.0 ±0.22	74.85 ±0.27	55.0 ±0.19	79.7 ±0.29	18.51 ±0.16	60.420 ±0.18
IFD	74.5 ±1.14	73.90 ±0.65	55.8 ±0.27	80.1 ±0.68	18.81 ±0.23	60.614 ±0.47
LangGPS + KMC	74.8 ±0.18	73.74 ±0.22	55.5 ±0.15	80.8 ±0.24	19.34 ±0.13	60.824 ±0.14
MACE + KMC	**76.0** ±0.45	77.36 ±0.52	55.7 ±0.40	**81.7** ±0.58	**19.38** ±0.32	**62.036** ±0.23
MACE + LESS	75.9 ±0.36	**77.64** ±0.42	**55.9** ±0.32	81.1 ±0.46	19.26 ±0.27	61.958 ±0.18
Llama-3.1-8B						
Full (100%)	46.3	79.98	46.1	73.3	11.11	51.362
Random	39.9 ±1.22	57.94 ±0.85	46.4 ±0.49	72.2 ±1.21	10.42 ±0.29	45.370 ±0.64
KMC	40.6 ±0.44	59.44 ±0.56	46.5 ±1.16	72.4 ±0.83	8.76 ±0.17	45.544 ±0.50
LESS	40.1 ±0.14	62.39 ±0.18	47.0 ±0.13	72.7 ±0.20	10.01 ±0.10	46.438 ±0.12
MTLD	**41.0** ±0.26	62.69 ±0.33	46.3 ±0.22	71.6 ±0.19	9.92 ±0.06	46.304 ±0.17
IFD	39.6 ±0.15	52.86 ±1.07	44.6 ±0.38	62.2 ±0.97	10.34 ±0.14	41.918 ±0.44
LangGPS + KMC	40.4 ±0.22	60.64 ±0.24	46.9 ±0.12	71.4 ±0.25	9.43 ±0.09	45.756 ±0.14
MACE + KMC	40.5 ±0.36	68.05 ±0.44	47.3 ±0.26	72.9 ±0.50	**10.63** ±0.23	47.858 ±0.28
MACE + LESS	40.8 ±0.33	**68.56** ±0.28	**47.6** ±0.31	**73.2** ±0.38	10.38 ±0.22	**48.102** ±0.24

4.3 Gains Across Tasks and Languages

MACE + KMC improves all five benchmark scores over LangGPS + KMC on both backbones (Table 1). These improvements span multiple evaluation languages within each task: Figure 3 shows gains in 39 of 50 benchmark–language cells on Qwen2.5-7B and 40 of 50 on Llama-3.1-8B.

The gains also extend to evaluation languages absent from the Aya instruction-tuning mixture. Across the ten benchmark–language cells in this group, MACE + KMC improves over LangGPS + KMC by 1.561 points on Qwen2.5-7B and 1.164 points on Llama-3.1-8B, with each benchmark weighted equally. For training languages, the corresponding gains over LangGPS + KMC are 1.148 points on Qwen2.5-7B and 2.197 points on Llama-3.1-8B. Both training and non-training language groups show positive average gains over the corresponding KMC and LESS selectors as well (Appendix C.2).

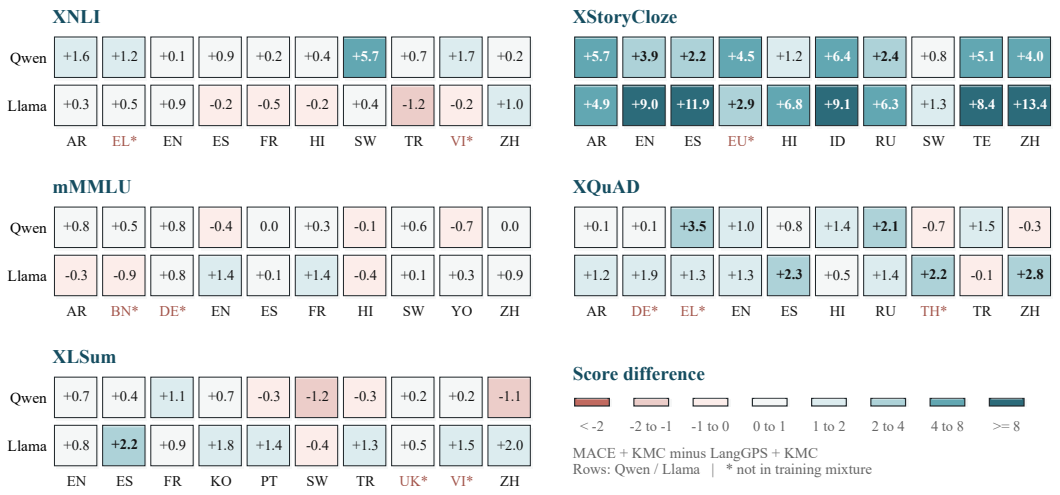

Figure 3: Per-language score differences between MACE + KMC and LangGPS + KMC. Each panel shows a benchmark; rows denote backbones, and cells show score differences rounded to one decimal place. Asterisks mark evaluation languages absent from the instruction-tuning mixture. MACE + KMC improves 39 of 50 cells on Qwen2.5-7B and 40 of 50 on Llama-3.1-8B; two Qwen cells are unchanged. Positive values favor MACE + KMC; negative values favor LangGPS + KMC.

Table 2: Exposure-penalty ablation at the 5% data budget. Affinity only sets $\lambda = 0$. Entries report three-run mean $\pm$ SD, including the five-benchmark average (Avg.); bold marks the better mean within each backbone.

Variant	XNLI	XStoryCloze	mMMLU	XQuAD	XLSum	Avg.
Qwen2.5-7B						
Affinity only + KMC	75.5 $\pm$0.12	**77.82** $\pm$0.15	55.3 $\pm$0.10	79.4 $\pm$0.17	18.36 $\pm$0.09	61.272 $\pm$0.10
MACE + KMC	**76.0** $\pm$0.45	77.36 $\pm$0.52	**55.7** $\pm$0.40	**81.7** $\pm$0.58	**19.38** $\pm$0.32	**62.036** $\pm$0.23
Llama-3.1-8B						
Affinity only + KMC	38.4 $\pm$0.20	56.37 $\pm$0.15	46.7 $\pm$0.09	70.5 $\pm$0.12	8.23 $\pm$0.07	44.050 $\pm$0.10
MACE + KMC	**40.5** $\pm$0.36	**68.05** $\pm$0.44	**47.3** $\pm$0.26	**72.9** $\pm$0.50	**10.63** $\pm$0.23	**47.858** $\pm$0.28

4.4 Component Analysis

The coverage diagnostic in Figure 1 clusters the projected gradients of 256 reference examples per language and evaluates 64 disjoint held-out examples. Each held-out gradient receives its maximum cosine similarity to a fitted center in the shared 1,024-dimensional projected space. We average these scores over held-out examples, languages, and three clustering initializations.

In downstream fine-tuning, four prototypes improve the five-benchmark average over a single prototype on Qwen2.5-7B (Figure 4). With four prototypes, adding conflict exposure to affinity-only selection raises the average by 0.764 points on Qwen2.5-7B and 3.808 points on Llama-3.1-8B. Gains occur on four of five Qwen benchmarks and all five Llama benchmarks (Table 2).

The exposure ablation varies only the conflict term while keeping the remaining selection settings fixed. The prototype-count ablation varies K at a fixed exposure-penalty weight.

4.5 Conflict-Measure Comparison

We compare five exposure measures within the same MACE + KMC pipeline, using the coordinate sets, aggregation, and rank transform specified in Appendix A.4. MGCR achieves the highest average on both backbones (Table 3). Residual is the closest alternative, trailing by 0.192 points on Qwen and 0.988 points on Llama. The measures differ in what they retain: MGCR counts opposing coordinate products even under positive net alignment, where the cosine and normalized negative-dot penalties are zero, and weights them by magnitude rather than counting signs equally.

Table 3: Conflict measures within MACE + KMC at the 5% data budget. Entries give three-run mean ± SD, including the five-benchmark average (Avg.); bold marks column bests within each backbone.

Conflict measure	XNLI	XStoryCloze	mMMLU	XQuAD	XLSum	Avg.
Qwen2.5-7B						
Cosine	75.8 ±0.17	**77.76** ±0.21	55.2 ±0.14	79.7 ±0.23	18.44 ±0.12	61.374 ±0.14
Sign	75.4 ±0.41	74.80 ±0.48	55.3 ±0.36	81.2 ±0.52	19.29 ±0.30	61.186 ±0.32
Negative dot	75.8 ±0.21	74.02 ±0.26	55.3 ±0.18	**81.9** ±0.28	18.62 ±0.15	61.120 ±0.17
Residual	**76.1** ±0.30	76.41 ±0.36	55.6 ±0.26	81.7 ±0.40	**19.46** ±0.22	61.844 ±0.24
MGCR	76.0 ±0.45	77.36 ±0.52	**55.7** ±0.40	81.7 ±0.58	19.38 ±0.32	**62.036** ±0.23
Llama-3.1-8B						
Cosine	38.9 ±0.20	65.70 ±0.18	47.1 ±0.14	72.0 ±0.17	8.64 ±0.08	46.476 ±0.12
Sign	40.3 ±0.24	64.17 ±0.28	46.9 ±0.37	71.8 ±0.33	9.65 ±0.20	46.566 ±0.22
Negative dot	37.4 ±0.36	60.23 ±0.17	47.0 ±0.23	71.0 ±0.55	9.44 ±0.30	45.006 ±0.25
Residual	40.1 ±0.28	66.91 ±0.38	**47.6** ±0.25	70.4 ±0.34	9.34 ±0.22	46.870 ±0.23
MGCR	**40.5** ±0.36	**68.05** ±0.44	47.3 ±0.26	**72.9** ±0.50	**10.63** ±0.23	**47.858** ±0.28

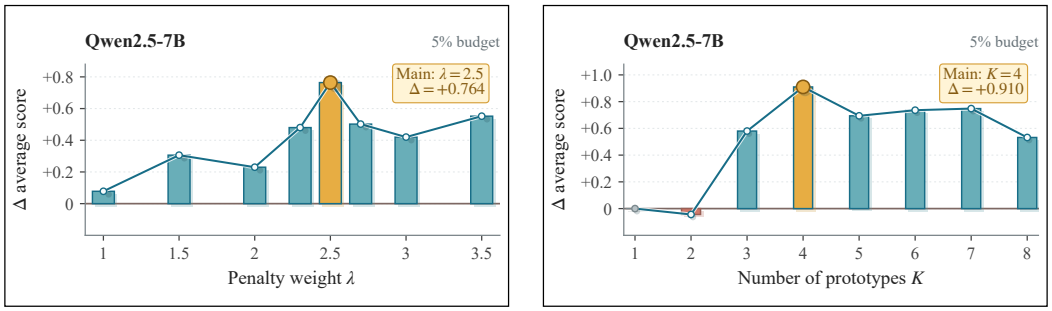

Figure 4: Hyperparameter sensitivity for Qwen2.5-7B + KMC at the 5% data budget. The left panel shows $\lambda = 1$ to 3.5 as changes in the three-run five-benchmark mean against affinity-only selection; the right shows changes against $K = 1$. Highlighted points mark the main settings.

4.6 Hyperparameter Sensitivity

Figure 4 examines exposure-penalty weights from 1 to 3.5 and prototype counts on Qwen2.5-7B with KMC at the 5% data budget. With $K = 4$ fixed, all eight displayed weights improve the five-benchmark average over affinity-only selection. At the main setting $\lambda = 2.5$, the average is 62.036, a gain of 0.764 points over affinity-only selection. At this weight, $K = 4$ improves the average by 0.910 points over a single prototype. Tables 8 and 9 report the individual benchmark scores.

For $K = 5$ to 8, the average ranges from 61.658 to 61.874, above the single-prototype score of 61.126. Tasks favor different counts: XNLI peaks at $K = 6$ and XQuAD at $K = 7$, whereas XStoryCloze, mMMLU, and XLSum peak at $K = 4$.

5 Conclusion

MACE ranks multilingual instruction examples by combining in-language prototype affinity with cross-language conflict exposure. Multiple gradient prototypes capture diverse local directions, while MGCR measures opposing coordinate contributions. Placing MACE before KMC and LESS improves both selectors on Qwen2.5-7B and Llama-3.1-8B and outperforms LangGPS under the same KMC backend. Component analyses support both signals, with gains extending to evaluation languages outside the instruction-tuning mixture.

AI USE STATEMENT

Generative AI tools assisted with language polishing of the manuscript. The authors take responsibility for the final content of this work, including material produced with AI assistance.

ETHICS STATEMENT

MACE allocates selection quotas in proportion to language data sizes. We report per-language results for the evaluated languages. Applications should consider performance for their intended language communities and biases inherited from the source instruction data.

REPRODUCIBILITY STATEMENT

Section 3 defines the selection procedure. Appendix A records scoring, fine-tuning, and evaluation settings. Appendix B reports component and hyperparameter comparisons, and Appendix C analyzes gains across tasks and language groups. Appendix D contains the mathematical analysis; Appendix E provides the complete per-language scores.

REFERENCES

Ishika Agarwal, Krishnateja Killamsetty, Lucian Popa, and Marina Danilevsky. DELIFT: Data efficient language model instruction fine-tuning. In *International Conference on Learning Representations*, volume 2025, pp. 100488–100509, 2025.

Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, Colin Raffel, Shiyu Chang, Tatsunori Hashimoto, and William Yang Wang. A survey on data selection for language models, 2024. URL https://arxiv.org/abs/2402.16827.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4623–4637, 2020.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. The Belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 749–775. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.44.

Alexander Bukharin, Shiyang Li, Zhengyang Wang, Jingfeng Yang, Bing Yin, Xian Li, Chao Zhang, Tuo Zhao, and Haoming Jiang. Data diversity matters for robust instruction tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 3411–3425. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.195.

Yihan Cao, Yanbin Kang, Chi Wang, and Lichao Sun. Instruction mining: Instruction data selection for tuning large language models. *arXiv preprint arXiv:2307.06290*, 2023.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, et al. AlpaGasus: Training a better alpaca with fewer data. In *International Conference on Learning Representations*, 2024a.

Pinzhen Chen, Shaoxiong Ji, Nikolay Bogoychev, Andrey Kutuzov, Barry Haddow, and Kenneth Heafield. Monolingual or multilingual instruction tuning: Which makes a better alpaca. In *Findings of the Association for Computational Linguistics: EACL 2024*, pp. 1347–1356. Association for Computational Linguistics, 2024b. doi: 10.18653/v1/2024.findings-eacl.90.

Yicheng Chen, Yining Li, Kai Hu, Ma Zerun, 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, 2025a.

Zhixun Chen, Ping Guo, Wenhan Han, Yifan Zhang, Binbin Liu, Haobin Lin, Fengze Liu, Yan Zhao, Bingni Zhang, Taifeng Wang, et al. MuRating: A high quality data selecting approach to multilingual large language model pretraining. *Advances in Neural Information Processing Systems*, 38:76166–76198, 2025b.

Rochelle Choenni, Dan Garrette, and Ekaterina Shutova. How do languages influence each other? studying cross-lingual data sharing during LM fine-tuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13244–13257, 2023.

Kenneth L. Clarkson and David P. Woodruff. Low rank approximation and regression in input sparsity time. In *Proceedings of the Forty-Fifth Annual ACM Symposium on Theory of Computing*, pp. 81–90, 2013. doi: 10.1145/2488608.2488620.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2475–2485, 2018.

Qirun Dai, Dylan Zhang, Jiaqi W Ma, and Hao Peng. Improving influence-based instruction tuning data selection for balanced learning of diverse capabilities. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 7079–7102, 2025.

Inderjit S. Dhillon and Dharmendra S. Modha. Concept decompositions for large sparse text data using clustering. *Machine Learning*, 42:143–175, 2001. doi: 10.1023/A:1007612920971.

Lan Feng, Fan Nie, Yuejiang Liu, and Alexandre Alahi. TAROT: Targeted data selection via optimal transport. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 16837–16852. PMLR, 2025.

Yuan Ge, Yilun Liu, Chi Hu, Weibin Meng, Shimin Tao, Xiaofeng Zhao, Mahong Xia, Zhang Li, Boxing Chen, Hao Yang, et al. Clustering and ranking: Diversity-preserved instruction selection through expert-aligned quality estimation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 464–478, 2024.

Amirata Ghorbani and James Zou. Data Shapley: Equitable valuation of data for machine learning. In *International Conference on Machine Learning*, pp. 2242–2251. PMLR, 2019.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 herd of models, 2024. URL `https://arxiv.org/abs/2407.21783`.

Guangzeng Han and Xiaolei Huang. What makes good instruction-tuning data? an in-context learning perspective. *arXiv preprint arXiv:2604.25132*, 2026.

Jindong Han, Hao Liu, Jun Fang, Naiqiang Tan, and Hui Xiong. Automatic instruction data selection for large language models via uncertainty-aware influence maximization. In *Proceedings of the ACM on Web Conference 2025*, pp. 4969–4979, 2025. doi: 10.1145/3696410.3714817.

Tahmid Hasan, Abhik Bhattacharjee, Md Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M Sohel Rahman, and Rifat Shahriyar. XL-Sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 4693–4703, 2021.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

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. *arXiv preprint arXiv:2106.09685*, 2021.

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`.

Ayush Jain, Andrea Montanari, and Eren Sasoglu. Train on validation (ToV): Fast data selection with applications to fine-tuning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=fWHd3yYicX.

Shaoxiong Ji and Pinzhen Chen. How many languages make good multilingual instruction tuning? a case study on BLOOM, 2024. URL https://arxiv.org/abs/2404.04850.

Tannon Kew, Florian Schottmann, and Rico Sennrich. Turning English-centric LLMs into polyglots: How much multilinguality is needed? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 13097–13124, 2024. doi: 10.18653/v1/2024.findings-emnlp.766.

Krishnateja Killamsetty, Durga S, Ganesh Ramakrishnan, Abir De, and Rishabh Iyer. GRADMATCH: Gradient matching based data subset selection for efficient deep model training. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5464–5474. PMLR, 2021a.

Krishnateja Killamsetty, Durga Sivasubramanian, Ganesh Ramakrishnan, and Rishabh Iyer. GLISTER: Generalization based data subset selection for efficient and robust learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8110–8118, 2021b. doi: 10.1609/aaai.v35i9.16988.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 1885–1894. PMLR, 2017.

Jiangnan Li, Thuy-Trang Vu, Christian Herold, Amirhossein Tebbifakhr, Shahram Khadivi, and Gholamreza Haffari. CONGRAD: Conflicting gradient filtering for multilingual preference alignment. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6371–6387, 2026.

Ming Li, Yong Zhang, Shwai He, Zhitao Li, Hongyu Zhao, Jianzong Wang, Ning Cheng, and Tianyi Zhou. Superfiltering: Weak-to-strong data filtering for fast instruction-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14255–14273, 2024a.

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: Human Language Technologies (Volume 1: Long Papers)*, pp. 7602–7635. Association for Computational Linguistics, 2024b. doi: 10.18653/v1/2024.naacl-long.421.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, 2004.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9019–9052, 2022.

Xiaotian Lin, Yanlin Qi, Yizhang Zhu, Themis Palpanas, Chengliang Chai, Nan Tang, and Yuyu Luo. LEAD: Iterative data selection for efficient LLM instruction tuning. *Proceedings of the VLDB Endowment*, 19(3):426–439, 2025.

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. *Advances in Neural Information Processing Systems*, 37:97800–97825, 2024a.

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. In *International Conference on Learning Representations*, 2024b.

Zifan Liu, Amin Karbasi, and Theodoros Rekatsinas. TSDS: Data selection for task-specific model finetuning. *Advances in Neural Information Processing Systems*, 37:10117–10147, 2024c.

Keming Lu, Hongyi Yuan, Zheng Yuan, Runji Lin, Junyang Lin, Chuanqi Tan, Chang Zhou, and Jingren Zhou. #InsTag: Instruction tagging for analyzing supervised fine-tuning of large language models. In *International Conference on Learning Representations*, 2024.

Philip M McCarthy and Scott Jarvis. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2):381–392, 2010.

Dheeraj Mekala, Alex Nguyen, and Jingbo Shang. Smaller language models are capable of selecting instruction-tuning training data for larger language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 10456–10470, 2024.

Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 6950–6960. PMLR, 2020.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, et al. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15991–16111, 2023.

Xingyuan Pan, Luyang Huang, Liyan Kang, Zhicheng Liu, Yu Lu, and Shanbo Cheng. G-DIG: Towards gradient-based DIverse and hiGh-quality instruction data selection for machine translation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15395–15406. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.821.

Jinlong Pang, Na Di, Zhaowei Zhu, Jiaheng Wei, Hao Cheng, Chen Qian, and Yang Liu. Token cleaning: Fine-grained data selection for LLM supervised fine-tuning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 47837–47858. PMLR, 2025. URL `https://proceedings.mlr.press/v267/pang25a.html`.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Ziheng Qin, Kai Wang, Zangwei Zheng, Jianyang Gu, Xiangyu Peng, Zhaopan Xu, Daquan Zhou, Lei Shang, Baigui Sun, Xuansong Xie, et al. InfoBatch: Lossless training speed up by unbiased dynamic data pruning. In *International Conference on Learning Representations*, volume 2024, pp. 39723–39744, 2024.

HSVNS Kowndinya Renduchintala, Sumit Bhatia, and Ganesh Ramakrishnan. SMART: Submodular data mixture strategy for instruction tuning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12916–12934, 2024.

Stephanie Schoch, Ritwick Mishra, and Yangfeng Ji. Data selection for fine-tuning large language models using transferred Shapley values. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pp. 266–275, 2023.

Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. Multilingual instruction tuning with just a pinch of multilinguality. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2304–2317. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.136.

Shivalika Singh, Freddie Vargus, Daniel D'souza, et al. Aya dataset: An open-access collection for multilingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11521–11567. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.620.

Jingtan Wang, Xiaoqiang Lin, Rui Qiao, Pang Wei Koh, Chuan-Sheng Foo, and Bryan Kian Hsiang Low. NICE data selection for instruction tuning in LLMs with non-differentiable evaluation metric. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 63662–63689. PMLR, 2025a.

Mingyang Wang, Heike Adel, Lukas Lange, Jannik Strötgen, and Hinrich Schuetze. Gradsim: Gradient-based language grouping for effective multilingual training. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4631–4646, 2023.

Zige Wang, Qi Zhu, Fei Mi, Yasheng Wang, Haotian Wang, and Lifeng Shang. Dynamic data selection with normalized gradient-based influence approximation for targeted fine-tuning of LLMs. *Knowledge-Based Systems*, pp. 114144, 2025b.

Zige Wang, Qi Zhu, Fei Mi, Minghui Xu, Ruochun Jin, and Wenjing Yang. ClusterUCB: Efficient gradient-based data selection for targeted fine-tuning of LLMs. *arXiv preprint arXiv:2506.10288*, 2025c.

Zirui Wang, Zachary C Lipton, and Yulia Tsvetkov. On negative interference in multilingual models: Findings and a meta-learning treatment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4438–4450, 2020.

Zirui Wang, Yulia Tsvetkov, Orhan Firat, and Yuan Cao. Gradient vaccine: Investigating and improving multi-task optimization in massively multilingual models. In *International Conference on Learning Representations*, 2021.

Alexander Arno Weber, Klaudia Thellmann, Jan Ebert, Nicolas Flores-Herr, Jens Lehmann, Michael Fromm, and Mehdi Ali. Investigating multilingual instruction-tuning: Do polyglot models demand for multilingual instructions? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 20829–20855, 2024.

Jingyu Wei, Bo Liu, Tianjiao Wan, Baoyun Peng, Xingkong Ma, and Mengmeng Guo. JI2S: Joint influence-aware instruction data selection for efficient fine-tuning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 514–527, 2025.

Yang Wu, Huayi Zhang, Yizheng Jiao, Lin Ma, Xiaozhong Liu, Jinhong Yu, Dongyu Zhang, Dezhi Yu, and Wei Xu. ROSE: A reward-oriented data selection framework for LLM task-specific instruction tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 13200–13219, 2025. doi: 10.18653/v1/2025.findings-emnlp.710.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. LESS: Selecting influential data for targeted instruction tuning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54104–54132. PMLR, 2024.

Tingyu Xia, Bowen Yu, Kai Dang, An Yang, Yuan Wu, Yuan Tian, Yi Chang, and Junyang Lin. Rethinking data selection at scale: Random selection is almost all you need. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 2698–2711. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.146. URL https://aclanthology.org/2025.findings-emnlp.146/.

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy S Liang. Data selection for language models via importance resampling. *Advances in Neural Information Processing Systems*, 36: 34201–34227, 2023.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report, 2025a.

Xianjun Yang, Shaoliang Nie, Lijuan Liu, Suchin Gururangan, Ujjwal Karn, Rui Hou, Madian Khabsa, and Yuning Mao. Diversity-driven data selection for language model tuning through sparse autoencoder, 2025b. URL https://arxiv.org/abs/2502.14050.

Yixin Yang, Qingxiu Dong, Linli Yao, Fangwei Zhu, Weilin Luo, Bin Wang, and Zhifang Sui. RICo: Refined in-context contribution for automatic instruction-tuning data selection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 34349–34357, 2026.

Yu Yang, Siddhartha Mishra, Jeffrey Chiang, and Baharan Mirzasoleiman. SmallToLarge (S2L): Scalable data selection for fine-tuning large language models by summarizing training trajectories of small models. *Advances in Neural Information Processing Systems*, 37:83465–83496, 2024.

Yuming Yang, Yang Nan, Junjie Ye, Shihan Dou, Xiao Wang, Shuo Li, Huijie Lv, Tao Gui, Qi Zhang, and Xuanjing Huang. Measuring data diversity for instruction tuning: A systematic analysis and a reliable metric. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18530–18549, 2025c.

Kai Yao, Zhenghan Song, Kaixin Wu, Mingjie Zhong, Danzhao Cheng, Zhaorui Tan, Yixin Ji, and Penglei Gao. GAST: Gradient-aligned sparse tuning of large language models with data-layer selection. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4401–4416, 2026.

Yangfan Ye, Xiaocheng Feng, Xiachong Feng, Lei Huang, Weitao Ma, Qichen Hong, Yunfei Lu, Duyu Tang, Dandan Tu, and Bing Qin. LangGPS: Language separability guided data pre-selection for joint multilingual instruction tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 34450–34458, 2026. doi: 10.1609/aaai.v40i40.40743.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Jia Zhang, Chen-Xi Zhang, Yao Liu, Yi-Xuan Jin, Xiao-Wen Yang, Bo Zheng, Yi Liu, and Lan-Zhe Guo. D3: Diversity, difficulty, and dependability-aware data selection for sample-efficient LLM instruction tuning. *arXiv preprint arXiv:2503.11441*, 2025a.

Jipeng Zhang, Yaxuan Qin, Renjie Pi, Weizhong Zhang, Rui Pan, and Tong Zhang. TAGCOS: Task-agnostic gradient clustered coreset selection for instruction tuning data. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 4686–4701, 2025b.

Chunguang Zhao, Yilun Liu, Pufan Zeng, Yuanchang Luo, Shimin Tao, Minggui He, Weibin Meng, Song Xu, Chen Liu, Hongxia Ma, Li Zhang, Boxing Chen, and Daimeng Wei. M-DaQ: Retrieving samples with multilingual diversity and quality for instruction fine-tuning datasets. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 4361–4366, 2026. doi: 10.1145/3805712.3809946.

Hao Zhao, Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. Long is more for alignment: A simple but tough-to-beat baseline for instruction fine-tuning. *arXiv preprint arXiv:2402.04833*, 2024.

Yang Zhao, Li Du, Xiao Ding, Yangou Ouyang, Hepeng Wang, Kai Xiong, Jinglong Gao, Zhouhao Sun, Dongliang Xu, Qing Yang, Dongchen Li, Bing Qin, and Ting Liu. Beyond similarity: A gradient-based graph method for instruction tuning data selection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24391–24404. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1189. URL https://aclanthology.org/2025.acl-long.1189/.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. LLaMAFactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 400–410, 2024.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan 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 Experimental and Reproducibility Details

A.1 Data and Selection Budgets

We use multilingual instruction–response pairs from Aya (Singh et al., 2024) in the 16 languages listed in Table 4. Each example retains its language identifier throughout scoring and selection.

For each language, we sample up to 512 reference anchors disjoint from the candidate pool. These anchors provide the language mean gradient and local prototypes. We compare candidates with

Table 4: Languages in the Aya instruction-tuning mixture.

Amharic	Arabic	Chinese	English
French	Hindi	Indonesian	Japanese
Korean	Portuguese	Russian	Spanish
Swahili	Telugu	Turkish	Yoruba

Table 5: Gradient extraction and MACE scoring settings. Sequence lengths refer to scoring inputs.

Setting	Value
Maximum anchors per language	512
Adapter rank / scaling parameter	8 / 16
Adapter dropout	0
Model loading precision	bfloat16
Maximum sequence length	1,024
Maximum prompt / response length	768 / 256
Sparse projection dimension	256
Maximum spherical clustering iterations	8
Prototypes per language (K)	4
Reference coordinates (D)	1,024
Conflict reference languages per language (m)	3
Exposure penalty (λ)	2.5

prototypes of their own language and mean reference gradients of other languages. Selection quotas are proportional to language data sizes, subject to integer allocation and data availability.

MACE retains $P = 0.25|\mathcal{D}|$ examples during pre-selection. KMC or LESS retains 20% of these, giving $B = 0.20P = 0.05|\mathcal{D}|$ selected examples. The Full reference uses the entire corpus. The affinity-only ablation sets $\lambda = 0$ and uses the same pre-selection budget, language quotas, and downstream KMC stage.

A.2 Gradient Extraction and Prototype Construction

We compute scoring gradients with Qwen2.5-7B and Llama-3.1-8B in their base-model configurations. The backbone weights remain frozen, and gradients are taken with respect to LoRA adapter parameters attached to the attention query, key, value, and output projections and the feed-forward gate, up, and down projections. Table 5 summarizes the scoring configuration.

The model remains at a fixed state during scoring. Each example contributes the mean negative log-likelihood over response tokens; prompt tokens are excluded. We use the tokenizer's chat template when available and a user–assistant prompt otherwise. Long prompts are truncated from the left, and responses from the right. After backpropagation, adapter gradients are flattened in a fixed parameter order and converted to float32. Gradients are cleared between examples.

Scoring parameterization. The scoring implementation attaches a newly initialized adapter to the base model and performs no warmup updates. For a layer with effective weight $W + sBA$, standard initialization sets $B = 0$ and randomly initializes A (Hu et al., 2021). Writing $G_x = \partial L(x)/\partial W_{\text{eff}}$, the adapter gradients at this state are

$$\nabla_A L(x) = sB^\top G_x = 0, \qquad \nabla_B L(x) = sG_x A^\top. \tag{10}$$

The informative scoring coordinates lie in the B blocks, where the scoring feature is a fixed linear sketch of the full weight gradient. The projection A is shared across anchors and candidates. The zero A-gradient blocks leave inner products, norms, and MGCR unchanged.

We normalize each anchor gradient, apply a shared sparse random projection, and normalize the result. Each projection column has one nonzero entry with a uniformly sampled row and a sign sampled from $\{-1, +1\}$ (Clarkson & Woodruff, 2013). Spherical clustering assigns projected vectors to their most aligned centers and updates centers with normalized cluster means (Dhillon & Modha, 2001). The first center is sampled uniformly; later centers are sampled in proportion to cosine distance from the closest existing center. Final assignments determine the raw

anchor-gradient averages in the original parameter space, which are normalized as prototypes. The language reference gradient is the raw mean over its anchors.

A.3 Conflict Scoring and Numerical Conventions

For reference language k, Ω_k indexes the D largest absolute entries of its mean gradient, on which MGCR is evaluated. For each language ℓ, we rank other languages by $C_{\Omega_k}(\mu_\ell, \mu_k)$ and retain the top $m = 3$ as $\mathcal{T}_m(\ell)$; if fewer than m references are available, all are used. Every candidate in $\mathcal{D}_\ell$ uses this fixed set, averaging its MGCR against all selected language means. Exposure is converted to an ascending normalized rank within the candidate's language pool, with ties assigned average ranks.

A numerical threshold of $\varepsilon = 10^{-12}$ is used for normalization and conflict denominators. Vectors below the normalization threshold are mapped to zero, and a vanishing interaction magnitude receives zero MGCR. Singleton pools receive zero normalized exposure; an all-tied pool of more than one example receives 0.5 for every example. Empty clusters are reinitialized with the least well-aligned anchor during clustering, and an empty final cluster uses the normalized language mean.

A.4 Alternative Exposure Measures

For a candidate and reference, let $u = g_{x,\Omega_k}$ and $v = \mu_{k,\Omega_k}$, define $S = \sum_i |u_i v_i|$, and write $c = u^\top v/(\|u\|_2\|v\|_2)$. Let $I = \{i : u_i \neq 0,\ v_i \neq 0\}$ be the jointly active coordinates. The implemented alternatives are

$$C_{\cos}(u, v) = [-c]_+, \tag{11}$$

$$C_{\text{sign}}(u, v) = \frac{1}{|I|} \sum_{i \in I} \mathbf{1}\{u_i v_i < 0\}, \tag{12}$$

$$C_{\text{dot}}(u, v) = \frac{[-u^\top v]_+}{S}, \tag{13}$$

$$C_{\text{res}}(u, v) = 1 - [c]_+^2. \tag{14}$$

An empty active set or denominator below the numerical threshold gives zero. The normalized negative-dot measure (Negative dot in figures and tables) uses a normalized negative net product; residual scoring measures the squared residual after projection onto the nonnegative reference ray, relative to $\|u\|_2^2$. For $S > 0$, Equation 8 gives $C_{\text{dot}} = [2C_\Omega - 1]_+$. Each measure uses reference-selected coordinates, averaging over the selected reference languages, and the within-language rank transform.

A.5 Fine-Tuning Configuration

All methods, backbones, and ablations use the same fine-tuning configuration in LLaMA-Factory (Table 6). Each run uses two NVIDIA GeForce RTX 4090 GPUs. LoRA adapters are applied to the attention query, key, value, and output projections and the feed-forward gate, up, and down projections, with bias parameters frozen.

For each method, 5% of its instruction data is reserved for validation during fine-tuning. The reported data budgets refer to the subsets before this split. Validation and checkpoint saving occur every 100 steps, with a per-device validation batch size of 2.

A.6 Evaluation Protocol

We evaluate fine-tuned models on the five benchmarks in Table 7, each covering ten languages with equal numbers of examples. XNLI evaluates Arabic, Greek, English, Spanish, French, Hindi, Swahili, Turkish, Vietnamese, and Chinese. XStoryCloze evaluates Arabic, English, Spanish, Basque, Hindi, Indonesian, Russian, Swahili, Telugu, and Chinese. mMMLU evaluates Arabic, Bengali, German, English, Spanish, French, Hindi, Swahili, Yoruba, and Chinese. XQuAD evaluates Arabic, German, Greek, English, Spanish, Hindi, Russian, Thai, Turkish, and Chinese. XLSum evaluates English, Spanish, French, Korean, Portuguese, Swahili, Turkish, Ukrainian, Vietnamese,

Table 6: Shared fine-tuning settings; the effective batch size is $2 \times 8 \times 2 = 32$ examples.

Setting	Value
Training framework	LLaMA-Factory
Fine-tuning method	LoRA SFT
Optimizer	AdamW (`adamw_torch`)
Learning rate	5×10^{-5}
Learning-rate schedule	Cosine
Warmup steps	0
Epochs	3
GPUs per run	2
Per-device training batch size	2
Gradient accumulation steps	8
Effective training batch size	32
Maximum sequence length	2,048
LoRA rank / scaling parameter	8 / 16
LoRA dropout	0
Precision	BF16
Maximum gradient norm	1.0
Sequence packing	Disabled
Prompt template	`alpaca`
Distributed training	DeepSpeed

Table 7: Evaluation metrics and total example counts across evaluated languages.

Benchmark	Metric	Examples
XNLI	Accuracy	10,000
XStoryCloze	Accuracy	15,110
mMMLU	Accuracy	10,000
XQuAD	Normalized containment accuracy	10,000
XLSum	ROUGE-L F1	1,000

and Chinese. We disable sampling during generation. XNLI, XStoryCloze, and mMMLU use task-specific answer extraction and comparison with the reference label.

For XQuAD, normalized containment accuracy is computed after normalizing text and English number expressions. A nonempty prediction matches a nonempty reference if they are equal or either normalized string contains the other; a match to any reference answer is accepted. For XLSum, ROUGE-L F1 is computed from the longest common subsequence of the tokenized prediction and reference (Lin, 2004). The tokenizer lowercases text and uses the regular-expression alternatives for individual Chinese characters, Japanese kana, Korean syllables, and Unicode word sequences.

Each benchmark score averages its ten language scores, equivalent to pooling the equally sized language sets. Scores use a 0–100 scale and are averaged across three runs. Avg. weights the five benchmarks equally; gains are absolute score differences unless labeled as relative percentages.

Scoring settings and randomization. The three runs vary seeds for fine-tuning, reference sampling, random projection, clustering initialization, and stochastic downstream selection. We use $K = 4$ prototypes, the $D = 1024$ largest absolute coordinates of each reference-language mean gradient, $m = 3$ reference languages selected by mean-gradient MGCR, and $\lambda = 2.5$ across both backbones and selectors. The prototype-count sweep evaluates $K = 1, \ldots, 8$ at this weight.

A.7 Gradient Diagnostics in Figure 1

The Aya diagnostic uses 256 reference examples and 64 disjoint held-out examples per language. A shared 1,024-dimensional signed projection represents the gradients. For each $K \in \{1, 2, 3, 4, 5, 6, 7, 8\}$, spherical clustering uses initialization seeds 13, 42, and 2026. Held-out alignment is the maximum cosine similarity to fitted centers, averaged across examples, languages, and clustering runs. Appendix B.4 reports the downstream prototype-count comparison.

The Belebele diagnostic begins with the 16 languages used for Aya selection and compares their mean gradients using semantically parallel reading-comprehension examples (Bandarkar et al., 2024), with up to 900 examples per language. It reports full-coordinate MGCR and its distribution across parameter blocks. Telugu is excluded from this diagnostic because its mean response-token loss is anomalously high relative to the other languages. The analysis covers 15 languages and 105 unordered pairs; Telugu remains in the Aya selection and training experiments.

A.8 Selection Efficiency and Performance

On 62,919 Aya candidates with Qwen2.5-7B and two NVIDIA GeForce RTX 4090 GPUs, MACE pre-selection takes approximately 15.6% of the recorded selection time of standalone LESS. MACE retains 25% of the candidate pool for subsequent selection. Including this pre-selection cost, MACE + LESS is estimated to use 49.0% of the standalone LESS selection time, a reduction of 51.0%.

We estimate the combined cost by keeping the non-candidate-gradient portion of LESS unchanged and scaling candidate-gradient extraction with the retained pool size. Using the cost breakdown reported by LESS (Xia et al., 2024), we approximate the gradient-extraction share as $\rho = 8/9$. With $r = T_{\mathrm{MACE}}/T_{\mathrm{LESS}} \approx 0.156$, the normalized cost is

$$\frac{T_{\mathrm{MACE+LESS}}}{T_{\mathrm{LESS}}} \approx r + (1 - \rho) + 0.25\rho \approx 0.490. \tag{15}$$

This estimate concerns data selection, excluding downstream fine-tuning and evaluation.

At the same 5% final data budget, MACE + LESS achieves a five-benchmark average of 61.958 on Qwen2.5-7B, compared with 60.368 for standalone LESS (Table 1). Thus, the estimated 51.0% reduction in selection time accompanies a gain of 1.590 score points, or 2.63% relative improvement.

B Component Ablations and Hyperparameter Sensitivity

This appendix examines the conflict-exposure component, alternative conflict measures, and sensitivity to λ and K. A paired-gain summary collects the differences used in the main text.

B.1 Contribution of Conflict Exposure

Table 2 in the main text compares affinity-only pre-selection ($\lambda = 0$) with MACE using the same KMC refinement stage.

Conflict exposure improves four benchmarks on Qwen2.5-7B and all five on Llama-3.1-8B. The largest increase on Qwen2.5-7B is on XQuAD (+2.38 points).

B.2 Comparison of Conflict Measures

To examine how the choice of conflict measure affects exposure scoring, we compare MGCR with cosine-based, sign-based, normalized negative-dot, and residual-based alternatives on both backbones. Figures and tables abbreviate these alternatives as Cosine, Sign, Negative dot, and Residual. Appendix A.4 defines their coordinate-wise calculations. Table 3 in the main text reports the corresponding MACE + KMC results.

MGCR achieves the highest five-benchmark average on both backbones, exceeding the closest alternative, Residual, by 0.192 points on Qwen2.5-7B and 0.988 points on Llama-3.1-8B. It improves three of five tasks on Qwen and four of five on Llama relative to Residual.

The normalized negative-dot measure equals $[2C_\Omega - 1]_+$ and assigns zero whenever $C_\Omega \leq 0.5$; the cosine penalty also assigns zero to nonnegative net alignment. Sign scoring counts opposing coordinates equally, whereas MGCR weights them by gradient-product magnitude. Residual scoring depends on the squared cosine of positive alignment and also responds to orthogonal components (Appendix A.4). These definitions distinguish the information retained by each exposure measure.

B.3 Exposure-Penalty Strength

Table 8 reports the displayed $\lambda = 1$ to 3.5 range on Qwen2.5-7B with KMC and $K = 4$ fixed. The affinity-only result ($\lambda = 0$) in Table 2 provides the reference for the left panel of Figure 4.

Table 8: Sensitivity to $\lambda = 1$–3.5 on Qwen2.5-7B with KMC (5% data). Scores are three-run means. Avg. is the five-benchmark mean; bold marks column bests. Table 2 gives the $\lambda = 0$ reference.

λ	XNLI	XStoryCloze	mMMLU	XQuAD	XLSum	Avg.
1	75.4	75.14	55.1	**82.2**	18.99	61.350
1.5	**76.1**	76.39	55.3	81.4	18.70	61.578
2	76.0	76.14	55.1	81.1	19.19	61.502
2.3	75.5	77.38	55.7	81.0	19.27	61.752
2.5	76.0	77.36	**55.7**	81.7	**19.38**	**62.036**
2.7	75.8	77.25	55.2	81.5	19.14	61.774
3	75.8	77.28	55.6	80.6	19.08	61.692
3.5	75.6	**77.64**	55.4	81.2	19.25	61.824

Within the displayed range, XNLI peaks at $\lambda = 1.5$, XStoryCloze at $\lambda = 3.5$, XQuAD at $\lambda = 1$, and mMMLU and XLSum at $\lambda = 2.5$. XQuAD and XLSum exceed their affinity-only scores at every displayed weight. The five-benchmark average varies non-monotonically across these weights.

B.4 Prototype-Count Sensitivity

We vary the number of language-specific prototypes from $K = 1$ to $K = 8$ on Qwen2.5-7B with KMC, fixing $\lambda = 2.5$ and all other scoring settings. The right panel of Figure 4 summarizes the average changes, and Table 9 reports the individual benchmark scores.

The five-benchmark average increases from 61.126 with a single prototype to 62.036 at $K = 4$, the highest average among the tested settings. For $K = 5$ to 8, the average remains between 61.658 and 61.874, while individual tasks attain their highest scores at different prototype counts.

B.5 Task-Level Paired Gains

Table 10 reports benchmark-level differences and average gains for the framework and conflict-measure comparisons. Each difference follows the comparison direction stated in its row.

C Distribution of Gains

We examine how aggregate improvements are distributed across benchmarks and across evaluation languages present in or absent from the instruction-tuning mixture.

C.1 Benchmark Contributions

We decompose the paired gains in Table 1 by benchmark. The five-benchmark difference is the mean of the five task-wise differences, and each leave-one-benchmark-out column averages the remaining four. Table 11 shows how omitting each task changes the aggregate.

C.2 Training and Non-Training Evaluation Languages

The training mixture covers the 16 languages listed in Appendix A. The five benchmarks contribute 50 benchmark–language cells: 40 evaluate a training language and 10 do not. The latter are Greek and Vietnamese on XNLI; Basque on XStoryCloze; Bengali and German on mMMLU; German, Greek, and Thai on XQuAD; and Ukrainian and Vietnamese on XLSum. We recompute paired differences from the per-language scores in Appendix E, averaging within each partition and then across benchmarks so that each benchmark receives one-fifth of the weight.

All twelve partition-level differences are positive. On Llama-3.1-8B, gains are larger for training languages in all three comparisons. On Qwen2.5-7B, their relative size depends on the baseline:

Table 9: Sensitivity to the number of prototypes on Qwen2.5-7B with KMC at the 5% final data budget. Scores are three-run means. Avg. reports the five-benchmark mean. Bold marks the best mean in each column.

K	XNLI	XStoryCloze	mMMLU	XQuAD	XLSum	Avg.
1	75.2	76.05	54.9	80.6	18.89	61.126
2	75.8	76.42	54.8	80.2	18.09	61.082
3	75.7	77.04	55.6	81.1	19.21	61.706
4	76.0	**77.36**	**55.7**	81.7	**19.38**	**62.036**
5	76.1	77.30	55.1	81.5	19.15	61.820
6	**76.2**	77.34	55.6	81.3	18.98	61.862
7	75.9	77.35	55.2	**81.8**	19.08	61.874
8	75.7	77.09	55.4	81.0	19.09	61.658

Table 10: Paired gains from three-run means. Framework rows compare MACE against the listed baseline; conflict rows compare MGCR against another measure within MACE + KMC. Values are absolute score differences; Avg. compares unweighted five-benchmark means. Negative dot denotes the normalized negative-dot measure.

Comparison	XNLI	XStoryCloze	mMMLU	XQuAD	XLSum	Avg.
Qwen2.5-7B: framework comparisons						
MACE + KMC vs. KMC	+0.8	+1.75	+0.2	+1.0	+0.66	**+0.891**
MACE + LESS vs. LESS	+1.0	+5.49	+0.6	+1.0	−0.09	**+1.590**
MACE + KMC vs. LangGPS + KMC	+1.3	+3.62	+0.2	+1.0	+0.04	**+1.212**
MACE + KMC vs. affinity only	+0.5	−0.46	+0.4	+2.4	+1.02	**+0.764**
Qwen2.5-7B: MGCR vs. alternative conflict measure						
MGCR vs. Cosine	+0.2	−0.40	+0.5	+2.1	+0.94	**+0.662**
MGCR vs. Sign	+0.7	+2.56	+0.4	+0.6	+0.09	**+0.850**
MGCR vs. Negative dot	+0.2	+3.34	+0.4	−0.2	+0.76	**+0.916**
MGCR vs. Residual	−0.1	+0.95	+0.1	+0.1	−0.08	**+0.192**
Llama-3.1-8B: framework comparisons						
MACE + KMC vs. KMC	−0.1	+8.61	+0.8	+0.5	+1.87	**+2.314**
MACE + LESS vs. LESS	+0.8	+6.17	+0.5	+0.5	+0.37	**+1.664**
MACE + KMC vs. LangGPS + KMC	+0.1	+7.41	+0.3	+1.5	+1.20	**+2.102**
MACE + KMC vs. affinity only	+2.1	+11.68	+0.5	+2.4	+2.40	**+3.808**
Llama-3.1-8B: MGCR vs. alternative conflict measure						
MGCR vs. Cosine	+1.6	+2.35	+0.1	+0.8	+1.99	**+1.382**
MGCR vs. Sign	+0.2	+3.88	+0.3	+1.1	+0.98	**+1.292**
MGCR vs. Negative dot	+3.1	+7.82	+0.3	+1.9	+1.19	**+2.852**
MGCR vs. Residual	+0.4	+1.14	−0.4	+2.5	+1.29	**+0.988**

Table 11: Paired average-score gains and leave-one-benchmark-out averages. Each row compares MACE with a baseline using the same KMC or LESS backend; omitted-task columns exclude only the named benchmark.

Baseline	All five	Omit XNLI	Omit XStoryCloze	Omit mMMLU	Omit XQuAD	Omit XLSum
Qwen2.5-7B						
KMC	+0.891	+0.903	+0.676	+1.063	+0.863	+0.948
LESS	+1.590	+1.748	+0.615	+1.828	+1.750	+2.010
LangGPS + KMC	+1.212	+1.198	+0.610	+1.470	+1.278	+1.505
Llama-3.1-8B						
KMC	+2.314	+2.925	+0.740	+2.703	+2.778	+2.425
LESS	+1.664	+1.893	+0.538	+1.948	+1.955	+1.988
LangGPS + KMC	+2.102	+2.608	+0.775	+2.543	+2.258	+2.328

gains over KMC and LangGPS + KMC are larger for non-training languages, while gains over

Table 12: Paired average-score gains on training and non-training evaluation languages, recomputed from per-language results with the same downstream selector in each comparison.

	Qwen2.5-7B		Llama-3.1-8B	
Comparison	**Training**	**Non-training**	**Training**	**Non-training**
MACE + KMC vs. KMC	+0.893	+1.029	+2.404	+1.264
MACE + LESS vs. LESS	+1.840	+0.121	+1.715	+1.086
MACE + KMC vs. LangGPS + KMC	+1.148	+1.561	+2.197	+1.164

LESS are larger for training languages. The benefit therefore extends to both language groups, with its distribution depending on the backbone and selector.

D MGCR Analysis

All statements use scoring parameters $\theta \in \mathbb{R}^d$, with other model parameters fixed. Gradients and coordinate sets are evaluated at a common scoring state and held fixed in each one-step comparison. We prove Proposition 1 for mean anchor loss, then extend it in Proposition 2 to a reference objective with a different gradient and an update with a residual term. Finally, we derive the coordinate-budget and aggregation properties used by MACE.

D.1 Relation to Reference-Loss Changes

Fix a reference language k. Its average anchor loss is $F_k^{\mathcal{A}}(\theta) = |\mathcal{A}_k|^{-1} \sum_{a \in \mathcal{A}_k} L(a;\theta)$, so the mean gradient used by MACE is $\mu_k = \nabla F_k^{\mathcal{A}}(\theta)$. For a candidate gradient g_x, define the total interaction magnitude and raw MGCR on Ω_k as

$$S_k(x) = \sum_{i \in \Omega_k} |\mu_{k,i} g_{x,i}|, \qquad C_k(x) = \frac{\sum_{i \in \Omega_k} [-\mu_{k,i} g_{x,i}]_+}{S_k(x)}. \tag{16}$$

The ratio is defined for $S_k(x) > 0$; when $S_k(x) = 0$, set $C_k(x) = 0$ as in the scoring rule. Separating agreeing and opposing products yields

$$-\sum_{i \in \Omega_k} \mu_{k,i} g_{x,i} = S_k(x)\big(2C_k(x) - 1\big). \tag{17}$$

The left-hand side is the selected coordinates' contribution to the directional derivative of the anchor loss along $-g_x$. The right-hand side separates interaction magnitude from the balance of opposing and agreeing components. At a fixed positive $S_k(x)$, this contribution increases linearly with MGCR and changes sign at $C_k(x) = 1/2$. The identity also holds when $S_k(x) = 0$.

MGCR is invariant to separate positive rescaling of the candidate and reference gradients:

$$C_{\Omega_k}(s g_x, t \mu_k) = C_{\Omega_k}(g_x, \mu_k), \qquad s, t > 0. \tag{18}$$

Both the opposing magnitude and the total magnitude scale by st. Thus C_Ω is homogeneous of degree zero in each gradient for a fixed coordinate set. Positive rescaling of μ_k also preserves its largest-magnitude coordinate set under the fixed tie rule.

Proof of Proposition 1. The integral form of the first-order expansion gives

$$\begin{aligned} \Delta F_k^{\mathcal{A}}(x) &= -\eta \mu_k^\top g_x + r_k(x) \\ &= \eta S_k(x)(2C_k(x) - 1) - \eta T_k(x) + r_k(x), \end{aligned} \tag{19}$$

where

$$r_k(x) = -\eta \int_0^1 \left[\nabla F_k^{\mathcal{A}}(\theta - t\eta g_x) - \mu_k\right]^\top g_x \, dt. \tag{20}$$

The Lipschitz condition yields $|r_k(x)| \le \beta_k \eta^2 \|g_x\|_2^2 \int_0^1 t \, dt = \beta_k \eta^2 \|g_x\|_2^2/2$. Applying the triangle inequality to Equation 19 proves Equation 9. □

Retaining the signed tail gives center $\eta S_k(x)(2C_k(x)-1) - \eta T_k(x)$ with error radius $\beta_k \eta^2 \|g_x\|_2^2/2$.

Proposition 2 (General finite-step reference-loss interval). Let F_k be a differentiable reference-language loss with $h_k = \nabla F_k(\theta)$. Write an update as $\theta^+ = \theta - u_x$, where $u_x = \eta g_x + e_x$ and $\eta > 0$. All gradients and updates use the same parameterization. Assume ∇F_k is β_k-Lipschitz along the update segment. Define

$$T_k(x) = \sum_{i \notin \Omega_k} \mu_{k,i} g_{x,i} \tag{21}$$

and

$$\begin{aligned} E_k(x) = \eta |T_k(x)| + \eta \|h_k - \mu_k\|_2 \|g_x\|_2 \\ + \|h_k\|_2 \|e_x\|_2 + \frac{\beta_k}{2} \|u_x\|_2^2. \end{aligned} \tag{22}$$

Then $\Delta F_k(x) = F_k(\theta - u_x) - F_k(\theta)$ satisfies

$$\left| \Delta F_k(x) - \eta S_k(x) \big(2C_k(x) - 1\big) \right| \le E_k(x). \tag{23}$$

Proof. Smoothness gives $\Delta F_k(x) = -h_k^\top u_x + r_k(x)$, where $|r_k(x)| \le \beta_k \|u_x\|_2^2/2$. Substituting $u_x = \eta g_x + e_x$ and $\mu_k^\top g_x = S_k(x)(1 - 2C_k(x)) + T_k(x)$ gives

$$\begin{aligned} \Delta F_k(x) = \eta S_k(x)(2C_k(x) - 1) - \eta T_k(x) \\ - \eta (h_k - \mu_k)^\top g_x - h_k^\top e_x + r_k(x). \end{aligned} \tag{24}$$

The triangle and Cauchy–Schwarz inequalities bound the last four terms by $E_k(x)$. □

The four terms in $E_k(x)$ account for omitted coordinates, anchor–reference gradient differences, the update residual, and curvature. Setting $F_k = F_k^A$ and $e_x = 0$ recovers Proposition 1. If Ω_k also contains all coordinates, the radius becomes $\beta_k \eta^2 \|g_x\|_2^2/2$.

Numerical scores. Let $\widetilde{C}_k(x)$ be the numerically computed score. The exact ratio $C_k(x)$ uses the zero-mass convention defined above. The interval for $\widetilde{C}_k(x)$ has center $\eta S_k(x)(2\widetilde{C}_k(x) - 1)$ and radius $E_k(x) + 2\eta S_k(x)|\widetilde{C}_k(x) - C_k(x)|$. This follows directly from the triangle inequality and quantifies the effect of numerical thresholding and rounding on the loss interval.

Corollary (Candidate ordering). Consider two candidates evaluated at the same scoring state against the same reference objective and coordinate set, with a common $\eta > 0$ and valid bounds $E_k(x)$ and $E_k(y)$ from Proposition 2. If

$$\begin{aligned} \eta \big[S_k(y)(2C_k(y) - 1) \\ - S_k(x)(2C_k(x) - 1) \big] > E_k(x) + E_k(y), \end{aligned} \tag{25}$$

then $\Delta F_k(x) < \Delta F_k(y)$. In particular, for equal interaction magnitudes $S_k(x) = S_k(y) = S > 0$, the condition becomes

$$2\eta S \big(C_k(y) - C_k(x) \big) > E_k(x) + E_k(y). \tag{26}$$

Proof. Subtract the upper bound for $\Delta F_k(x)$ in Proposition 2 from the lower bound for $\Delta F_k(y)$. The stated condition makes this difference positive. □

At matched positive interaction magnitude, the MGCR gap determines the difference between the interval centers. Equation 26 ensures that the finite-step loss intervals have the same strict ordering.

D.2 Reference-Coordinate Selection

The omitted-coordinate term in Proposition 2 connects the analysis to the reference-specific coordinate sets used in MACE. For any $\Omega \subseteq \{1, \dots, d\}$, with complement Ω^c,

$$\left| \sum_{i \notin \Omega} \mu_{k,i} g_{x,i} \right| \le \|\mu_{k,\Omega^c}\|_2 \|g_x\|_2. \tag{27}$$

Coordinate-budget property. For $1 \le D \le d$, retaining the D largest absolute entries minimizes $\|\mu_{k,\Omega^c}\|_2$ among sets of size D, since its square equals $\|\mu_k\|_2^2 - \sum_{i\in\Omega} \mu_{k,i}^2$. Consequently, for any $G > 0$,

$$\min_{|\Omega|=D} \sup_{\|g\|_2 \le G} \left| \sum_{i \notin \Omega} \mu_{k,i} g_i \right| = G \|\mu_{k,\Omega_k^c}\|_2. \tag{28}$$

For fixed Ω with $\mu_{k,\Omega^c} \neq 0$, equality holds for $g_\Omega = 0$ and $g_{\Omega^c} = G\mu_{k,\Omega^c}/\|\mu_{k,\Omega^c}\|_2$. If the omitted reference coordinates are zero, both sides vanish. Thus the chosen coordinates minimize the worst-case omitted inner product over $\|g\|_2 \le G$ and bound its contribution to the loss interval by $\eta G \|\mu_{k,\Omega_k^c}\|_2$.

D.3 From Pairwise Conflict to Candidate Ranking

Fix a language ℓ and its reference set $\mathcal{T} = \mathcal{T}_m(\ell)$, selected using language mean gradients, with $q = |\mathcal{T}|$. For $x \in \mathcal{D}_\ell$, raw exposure averages $C_k(x)$ over this set, so $R(x) \in [0,1]$ and is nondecreasing in every selected pairwise score. With the reference set fixed, if $|C'_k(x) - C_k(x)| \le \delta$ for every $k \in \mathcal{T}$, then

$$|R'(x) - R(x)| \le \frac{1}{q} \sum_{k\in\mathcal{T}} |C'_k(x) - C_k(x)| \le \delta. \tag{29}$$

Here, $R'(x)$ averages the perturbed scores over the same reference set. The bound follows from the triangle inequality.

Within a fixed language pool of size at least two, average ranks assign equal values to tied exposures and strictly preserve the order of distinct exposures. Thus, for candidates x, y in that pool with equal affinity and $\lambda > 0$,

$$R(x) < R(y) \quad \Longrightarrow \quad \mathcal{U}_\lambda(x) - \mathcal{U}_\lambda(y) = \lambda\big(V_{\mathrm{neg}}(y) - V_{\mathrm{neg}}(x)\big) > 0. \tag{30}$$

Thus, averaging over the fixed reference set preserves monotonicity in its pairwise conflict scores, and lower exposure gives a higher selection score when affinity is equal and $\lambda > 0$.

E Detailed Per-Language Results

This section reports the per-language scores behind the aggregate results. Entries are three-run means on the 0–100 scale. Each benchmark uses equal numbers of examples per language; its Avg. column uses unrounded language scores and matches the aggregate table. Table averages and boldface are determined before display rounding. Main and conflict-measure comparisons cover both backbones, while the λ and K sweeps use Qwen2.5-7B with KMC.

E.1 MAIN RESULTS

Table 13: Per-language XNLI results for the main methods. Scores are means of three runs.

Setting	AR	EL	EN	ES	FR	HI	SW	TR	VI	ZH	Avg.
Qwen2.5-7B											
Full (100%)	77.0	70.5	88.7	84.4	83.4	69.9	38.8	72.0	77.0	80.9	74.3
Random	75.8	71.6	89.9	83.3	82.6	70.3	47.1	71.2	77.9	78.5	74.8
KMC	76.9	72.7	90.5	83.6	82.4	70.6	46.7	71.3	78.3	78.8	75.2
LangGPS + KMC	76.3	71.5	90.5	83.3	82.5	70.0	45.8	71.1	77.6	78.9	74.8
LESS	75.2	72.3	90.9	83.0	81.8	69.3	50.9	70.7	77.4	77.6	74.9
MTLD	75.2	70.3	88.8	83.0	81.8	69.3	42.1	71.9	77.2	80.7	74.0
IFD	75.6	71.6	90.8	83.0	82.2	69.6	45.5	70.7	76.7	79.3	74.5
Affinity only + KMC ($\lambda = 0$)	77.2	72.6	91.2	84.1	82.9	69.6	48.0	72.1	77.7	79.6	75.5
MACE + KMC	77.9	72.7	90.6	84.2	82.7	70.4	51.5	71.8	79.3	79.1	76.0
MACE + LESS	78.0	73.1	90.8	83.7	82.4	71.4	49.6	71.8	78.0	79.9	75.9
Llama-3.1-8B											
Full (100%)	45.9	49.4	47.7	47.8	47.2	45.9	39.9	45.2	47.7	46.4	46.3
Random	40.1	39.9	41.3	42.1	43.7	36.6	35.0	38.5	38.8	43.1	39.9
KMC	40.7	41.5	40.0	43.2	44.4	36.4	35.5	41.0	40.6	43.1	40.6
LangGPS + KMC	40.6	40.7	40.8	42.7	43.2	37.8	35.3	40.4	39.7	43.1	40.4
LESS	39.1	40.3	43.6	42.3	42.9	36.1	34.5	39.2	39.2	43.4	40.1
MTLD	41.8	43.2	39.2	42.4	42.9	40.0	35.1	39.9	40.2	45.1	41.0
IFD	42.3	39.0	35.1	41.1	43.2	39.1	36.7	40.4	39.6	39.9	39.6
Affinity only + KMC ($\lambda = 0$)	38.1	41.5	40.0	39.6	39.3	38.1	31.7	37.3	40.0	38.6	38.4
MACE + KMC	40.9	41.2	41.7	42.5	42.7	37.6	35.7	39.2	39.5	44.1	40.5
MACE + LESS	39.6	41.8	43.5	42.7	43.5	38.2	36.5	39.4	39.1	43.8	40.8

Table 14: Per-language XStoryCloze results for the main methods. Scores are means of three runs.

Setting	AR	EN	ES	EU	HI	ID	RU	SW	TE	ZH	Avg.
Qwen2.5-7B											
Full (100%)	81.40	85.24	86.23	56.19	69.29	74.92	82.04	55.46	50.89	69.42	71.11
Random	83.30	86.52	90.20	51.69	81.75	78.38	87.93	47.52	57.98	77.54	74.28
KMC	84.18	88.95	91.26	51.42	82.20	81.47	89.01	47.39	59.23	81.01	75.61
LangGPS + KMC	81.14	87.16	90.87	49.97	80.94	77.63	87.95	47.32	55.66	78.76	73.74
LESS	82.20	78.56	86.37	57.84	80.68	74.65	82.79	54.20	55.59	68.63	72.15
MTLD	81.01	88.15	91.13	53.81	80.48	79.22	88.02	47.78	61.28	77.63	74.85
IFD	80.34	88.55	90.07	50.30	80.34	77.30	87.76	47.45	58.57	78.29	73.90
Affinity only + KMC ($\lambda = 0$)	86.96	91.00	92.72	54.93	83.32	84.71	90.21	49.57	61.61	83.19	77.82
MACE + KMC	86.87	91.10	93.09	54.50	82.17	84.02	90.31	48.08	60.72	82.76	77.36
MACE + LESS	84.86	90.15	92.40	57.62	82.94	83.75	90.52	50.37	61.45	82.34	77.64
Llama-3.1-8B											
Full (100%)	78.49	91.13	87.76	68.76	79.68	83.92	84.78	65.19	73.46	86.63	79.98
Random	54.05	70.13	60.80	53.06	63.38	57.36	54.12	52.86	53.19	60.47	57.94
KMC	54.40	74.19	63.07	53.21	67.24	59.30	54.67	52.95	53.34	62.08	59.44
LangGPS + KMC	55.26	77.96	65.19	53.28	68.03	58.31	54.67	52.95	54.14	66.64	60.64
LESS	56.79	79.24	67.29	55.37	69.88	59.87	56.37	54.36	56.64	68.07	62.39
MTLD	56.85	81.14	68.83	53.87	66.38	62.48	57.38	52.95	53.87	73.13	62.69
IFD	52.81	53.08	52.88	52.81	52.81	52.88	52.88	52.81	52.81	52.81	52.86
Affinity only + KMC ($\lambda = 0$)	54.96	67.60	64.29	44.90	56.15	60.45	61.18	41.26	49.87	63.04	56.37
MACE + KMC	60.18	86.96	77.04	56.19	74.85	67.42	60.95	54.27	62.51	80.08	68.05
MACE + LESS	60.80	87.46	75.88	58.14	77.02	68.67	61.65	56.36	64.50	75.09	68.56

Table 15: Per-language mMMLU results for the main methods. Scores are means of three runs.

Setting	AR	BN	DE	EN	ES	FR	HI	SW	YO	ZH	Avg.
Qwen2.5-7B											
Full (100%)	55.3	45.7	65.1	73.9	65.5	64.6	46.7	34.3	34.3	67.9	55.3
Random	55.1	46.6	63.3	72.4	66.8	64.7	45.8	34.4	33.9	67.4	55.0
KMC	55.0	47.3	63.8	73.0	66.4	65.0	46.2	34.9	36.0	67.2	55.5
LangGPS + KMC	54.8	46.7	63.4	73.0	67.1	65.3	46.0	34.5	36.5	67.7	55.5
LESS	54.9	48.0	62.8	71.9	66.2	64.3	46.1	35.4	35.6	67.7	55.3
MTLD	54.9	46.3	62.9	72.9	67.0	64.7	45.4	34.9	33.2	68.1	55.0
IFD	54.6	47.2	64.1	73.6	67.3	65.9	47.8	35.8	33.9	67.9	55.8
Affinity only + KMC ($\lambda = 0$)	54.6	47.8	63.6	72.6	66.2	64.2	46.0	36.5	35.1	66.6	55.3
MACE + KMC	55.6	47.2	64.2	72.6	67.1	65.6	45.9	35.1	35.8	67.7	55.7
MACE + LESS	54.8	47.9	64.3	73.2	67.1	64.8	46.9	37.3	35.6	67.4	55.9
Llama-3.1-8B											
Full (100%)	43.7	37.4	52.1	62.7	52.2	52.6	43.7	34.4	33.4	49.1	46.1
Random	44.9	40.7	53.2	61.8	54.1	52.0	42.3	35.8	29.8	49.0	46.4
KMC	44.0	41.8	53.9	61.3	53.1	50.4	43.5	36.6	31.5	48.8	46.5
LangGPS + KMC	45.1	42.5	54.2	61.7	53.8	52.5	43.2	36.0	31.6	48.5	46.9
LESS	44.1	41.7	53.9	62.1	52.6	53.2	42.3	37.6	34.1	48.6	47.0
MTLD	42.3	41.6	52.8	60.0	53.6	53.6	43.5	35.1	31.4	49.1	46.3
IFD	42.6	40.3	51.7	58.0	51.9	44.7	43.4	35.1	30.9	47.0	44.6
Affinity only + KMC ($\lambda = 0$)	44.7	41.2	53.9	62.1	53.4	53.6	43.0	35.5	30.5	49.5	46.7
MACE + KMC	44.8	41.6	55.0	63.1	53.9	53.9	42.8	36.1	31.9	49.4	47.3
MACE + LESS	45.5	42.0	54.7	62.9	54.2	54.4	43.8	36.3	31.4	50.3	47.6

Table 16: Per-language XQuAD results for the main methods. Scores are means of three runs.

Setting	AR	DE	EL	EN	ES	HI	RU	TH	TR	ZH	Avg.
Qwen2.5-7B											
Full (100%)	75.2	76.8	55.8	87.4	80.3	58.1	68.2	77.9	74.7	77.0	73.1
Random	77.2	85.6	69.7	89.9	84.1	71.4	76.5	86.6	78.7	90.6	81.0
KMC	77.1	85.0	69.8	89.2	84.3	72.1	76.8	86.1	78.8	88.2	80.7
LangGPS + KMC	77.1	85.2	69.4	88.5	84.8	71.3	76.8	86.3	78.6	89.9	80.8
LESS	78.4	83.3	72.1	86.1	83.6	72.9	74.4	85.9	76.6	88.1	80.1
MTLD	77.0	83.0	69.2	87.3	83.7	71.6	75.1	84.1	77.5	88.3	79.7
IFD	78.6	82.9	70.3	88.3	84.9	70.7	75.5	85.2	77.0	87.1	80.1
Affinity only + KMC ($\lambda = 0$)	77.8	82.2	71.1	86.2	80.0	71.8	76.9	84.2	78.1	85.3	79.4
MACE + KMC	77.2	85.3	72.9	89.5	85.6	72.7	78.9	85.6	80.1	89.6	81.7
MACE + LESS	78.3	84.9	70.7	89.1	84.9	72.2	77.1	85.1	79.3	89.3	81.1
Llama-3.1-8B											
Full (100%)	68.1	74.6	66.7	85.2	77.7	72.6	61.5	74.5	72.9	79.0	73.3
Random	66.3	74.1	66.6	83.0	76.5	70.2	63.2	72.0	72.5	77.8	72.2
KMC	67.8	74.2	66.8	83.4	76.3	70.1	63.0	70.7	72.3	79.3	72.4
LangGPS + KMC	65.7	71.8	66.0	82.5	75.8	70.5	62.6	71.0	71.9	75.9	71.4
LESS	66.5	73.5	65.9	82.8	76.9	72.7	61.5	74.3	72.5	80.5	72.7
MTLD	66.1	71.7	64.5	83.0	77.0	71.5	61.6	72.0	71.0	77.9	71.6
IFD	55.1	60.4	54.7	80.9	68.3	60.2	53.4	61.3	64.5	63.1	62.2
Affinity only + KMC ($\lambda = 0$)	64.5	71.6	65.2	81.8	75.6	68.5	61.3	70.4	70.0	76.0	70.5
MACE + KMC	66.9	73.7	67.3	83.8	78.1	71.0	64.0	73.2	71.8	78.7	72.9
MACE + LESS	67.2	74.3	67.8	84.5	78.3	71.2	63.9	73.1	72.9	78.9	73.2

Table 17: Per-language XLSum results for the main methods. Scores are means of three runs.

Setting	EN	ES	FR	KO	PT	SW	TR	UK	VI	ZH	Avg.
Qwen2.5-7B											
Full (100%)	21.15	17.29	21.05	25.96	18.76	17.16	14.28	13.69	20.14	26.76	19.62
Random	18.07	17.58	20.76	25.67	19.15	16.43	14.91	12.88	18.54	25.29	18.93
KMC	18.10	17.59	20.09	25.42	19.06	16.02	14.17	12.45	18.68	25.61	18.72
LangGPS + KMC	18.27	18.29	20.44	26.51	19.56	18.11	14.47	12.82	18.79	26.17	19.34
LESS	18.21	18.45	20.47	26.32	19.64	16.89	14.84	12.95	19.71	25.99	19.35
MTLD	19.04	17.59	20.54	24.37	19.02	15.33	13.53	12.03	18.19	25.50	18.51
IFD	18.96	17.28	20.20	25.56	19.01	15.98	13.89	12.83	18.52	25.87	18.81
Affinity only + KMC ($\lambda = 0$)	17.74	17.11	19.49	25.65	16.95	16.53	13.98	10.74	18.84	26.61	18.36
MACE + KMC	18.95	18.64	21.53	27.24	19.27	16.90	14.15	12.99	19.04	25.08	19.38
MACE + LESS	18.66	18.67	20.95	26.05	19.24	17.28	14.30	12.52	19.39	25.52	19.26
Llama-3.1-8B											
Full (100%)	14.22	9.21	9.97	15.52	6.91	14.90	8.69	4.09	12.85	14.73	11.11
Random	12.83	10.37	10.79	11.50	8.92	13.85	7.69	4.85	14.26	9.18	10.42
KMC	14.51	7.12	7.38	9.87	6.34	15.41	6.76	3.59	10.28	6.34	8.76
LangGPS + KMC	13.45	9.64	10.63	7.66	7.33	14.60	6.59	4.57	12.79	7.05	9.43
LESS	12.38	11.22	11.79	9.31	8.62	13.26	7.74	4.46	13.16	8.13	10.01
MTLD	12.92	9.16	9.06	10.79	7.41	14.70	8.42	4.18	13.10	9.49	9.92
IFD	11.65	10.69	10.56	14.16	7.26	13.68	7.56	5.45	13.48	8.92	10.34
Affinity only + KMC ($\lambda = 0$)	12.08	9.14	8.29	8.68	5.95	12.02	5.70	2.70	12.44	5.33	8.23
MACE + KMC	14.21	11.88	11.55	9.41	8.77	14.24	7.85	5.07	14.31	9.02	10.63
MACE + LESS	14.23	11.29	10.44	10.83	8.10	14.17	7.85	4.85	14.59	7.48	10.38

E.2 Conflict-Measure Comparison

Table 18: Per-language XNLI results for alternative conflict measures.

Setting	AR	EL	EN	ES	FR	HI	SW	TR	VI	ZH	Avg.
Qwen2.5-7B											
Cosine	76.8	71.7	90.3	85.0	83.3	70.5	49.6	71.8	78.7	80.3	75.8
Sign	77.3	71.9	89.9	83.7	83.3	70.8	47.3	71.4	78.1	79.8	75.4
Negative dot	77.3	73.0	91.0	83.8	82.2	71.0	52.4	70.8	77.6	78.9	75.8
Residual	77.7	72.5	90.6	84.8	83.7	70.9	50.5	71.2	79.3	79.7	76.1
MGCR	77.9	72.7	90.6	84.2	82.7	70.4	51.5	71.8	79.3	79.1	76.0
Llama-3.1-8B											
Cosine	38.5	40.0	41.8	40.8	40.9	36.5	33.5	38.0	37.2	41.9	38.9
Sign	40.1	40.7	40.6	42.1	42.3	38.6	34.5	39.5	39.0	45.6	40.3
Negative dot	37.0	38.5	39.8	38.7	39.3	35.1	33.2	36.9	35.7	39.9	37.4
Residual	39.9	41.2	41.5	42.6	42.2	37.8	34.0	39.3	38.9	43.7	40.1
MGCR	40.9	41.2	41.7	42.5	42.7	37.6	35.7	39.2	39.5	44.1	40.5

Table 19: Per-language XStoryCloze results for alternative conflict measures.

Setting	AR	EN	ES	EU	HI	ID	RU	SW	TE	ZH	Avg.
Qwen2.5-7B											
Cosine	86.64	90.28	92.79	55.14	84.17	84.45	90.41	47.66	62.48	83.59	77.76
Sign	82.93	82.41	87.56	51.49	82.86	81.01	89.15	47.45	62.21	80.94	74.80
Negative dot	82.06	81.07	85.18	53.87	81.87	78.89	88.55	47.58	61.61	79.55	74.02
Residual	85.57	89.28	91.40	51.22	83.45	83.19	89.68	47.52	60.69	82.13	76.41
MGCR	86.87	91.10	93.09	54.50	82.17	84.02	90.31	48.08	60.72	82.76	77.36
Llama-3.1-8B											
Cosine	59.31	86.25	70.96	55.87	73.01	66.19	58.85	54.75	59.84	71.95	65.70
Sign	58.39	83.15	69.38	56.74	72.16	64.42	57.67	55.95	56.14	67.73	64.17
Negative dot	56.86	73.87	63.35	55.28	65.20	58.39	54.81	55.13	55.01	64.36	60.23
Residual	61.79	84.29	71.39	59.94	73.64	66.75	61.20	59.08	60.07	70.99	66.91
MGCR	60.18	86.96	77.04	56.19	74.85	67.42	60.95	54.27	62.51	80.08	68.05

Table 20: Per-language mMMLU results for alternative conflict measures.

Setting	AR	BN	DE	EN	ES	FR	HI	SW	YO	ZH	Avg.
Qwen2.5-7B											
Cosine	55.0	47.1	63.6	72.4	65.9	64.8	45.4	35.6	34.6	67.4	55.2
Sign	55.6	46.9	63.7	72.7	66.5	65.3	46.0	35.3	34.0	67.1	55.3
Negative dot	55.3	46.3	63.6	72.2	66.2	65.9	45.4	35.5	34.9	67.4	55.3
Residual	55.3	46.0	63.7	72.6	67.1	65.9	46.6	35.8	35.2	67.5	55.6
MGCR	55.6	47.2	64.2	72.6	67.1	65.6	45.9	35.1	35.8	67.7	55.7
Llama-3.1-8B											
Cosine	45.4	41.6	55.1	63.0	53.5	53.4	43.4	36.2	30.5	49.1	47.1
Sign	45.3	41.4	54.7	62.2	53.3	52.9	42.5	35.7	32.8	48.5	46.9
Negative dot	45.3	41.7	54.7	61.6	53.7	53.2	42.3	36.0	32.5	48.6	47.0
Residual	45.5	42.2	55.4	62.5	53.5	54.3	44.1	37.5	31.7	49.3	47.6
MGCR	44.8	41.6	55.0	63.1	53.9	53.9	42.8	36.1	31.9	49.4	47.3

Table 21: Per-language XQuAD results for alternative conflict measures.

Setting	AR	DE	EL	EN	ES	HI	RU	TH	TR	ZH	Avg.
Qwen2.5-7B											
Cosine	77.0	82.7	69.3	88.5	81.9	71.1	75.2	84.6	79.0	87.6	79.7
Sign	77.3	85.2	70.3	89.3	84.8	72.8	77.6	85.5	79.8	89.2	81.2
Negative dot	78.0	85.8	72.5	90.2	85.0	73.0	78.4	86.6	80.6	88.8	81.9
Residual	76.9	85.7	72.0	89.4	86.5	72.5	77.3	86.0	81.1	89.5	81.7
MGCR	77.2	85.3	72.9	89.5	85.6	72.7	78.9	85.6	80.1	89.6	81.7
Llama-3.1-8B											
Cosine	66.3	73.4	65.2	82.5	76.9	71.0	61.8	72.8	72.2	78.4	72.0
Sign	65.8	73.7	65.1	83.1	76.6	69.4	62.1	72.6	72.0	77.4	71.8
Negative dot	64.9	73.0	63.8	82.6	76.9	69.0	61.3	70.8	70.1	77.5	71.0
Residual	64.8	72.1	63.9	82.4	75.8	69.1	60.8	69.7	69.5	75.8	70.4
MGCR	66.9	73.7	67.3	83.8	78.1	71.0	64.0	73.2	71.8	78.7	72.9

Table 22: Per-language XLSum results for alternative conflict measures.

Setting	EN	ES	FR	KO	PT	SW	TR	UK	VI	ZH	Avg.
Qwen2.5-7B											
Cosine	17.51	17.53	20.06	25.86	18.11	16.83	13.49	11.04	18.14	25.86	18.44
Sign	18.35	18.13	21.31	27.00	19.84	16.97	13.58	12.88	19.26	25.60	19.29
Negative dot	18.31	17.68	19.92	25.16	18.22	16.23	13.96	11.73	18.93	26.01	18.62
Residual	18.77	18.48	21.13	26.83	20.00	17.51	13.85	12.75	19.60	25.71	19.46
MGCR	18.95	18.64	21.53	27.24	19.27	16.90	14.15	12.99	19.04	25.08	19.38
Llama-3.1-8B											
Cosine	13.34	10.14	9.60	5.54	7.01	12.64	6.60	3.51	12.25	5.72	8.64
Sign	13.46	10.33	10.97	8.80	7.34	14.62	7.03	3.94	12.70	7.32	9.65
Negative dot	12.87	10.41	10.33	8.90	7.53	14.23	7.36	3.81	13.27	5.65	9.44
Residual	13.70	9.28	10.29	6.33	6.62	14.48	7.73	4.33	13.08	7.54	9.34
MGCR	14.21	11.88	11.55	9.41	8.77	14.24	7.85	5.07	14.31	9.02	10.63

E.3 Exposure-Penalty Sensitivity

Table 23: Per-language XNLI sensitivity for $\lambda = 1$–3.5.

Setting	AR	EL	EN	ES	FR	HI	SW	TR	VI	ZH	Avg.
Qwen2.5-7B											
1	77.2	72.3	90.4	84.0	82.9	70.4	48.3	71.3	78.5	78.8	75.4
1.5	78.1	73.2	90.6	84.0	82.2	71.8	51.7	71.6	78.1	80.1	76.1
2	77.7	72.2	91.0	84.4	83.2	71.0	50.1	71.6	79.1	79.2	76.0
2.3	77.1	72.2	90.6	84.5	83.0	71.2	47.5	71.4	78.0	79.1	75.5
2.5	77.9	72.7	90.6	84.2	82.7	70.4	51.5	71.8	79.3	79.1	76.0
2.7	77.7	72.6	90.8	85.1	83.3	70.6	48.4	71.5	78.8	79.1	75.8
3	77.2	72.6	90.7	84.1	83.1	70.7	50.8	71.6	78.7	78.7	75.8
3.5	77.6	73.7	90.5	84.3	82.3	71.1	48.4	71.1	78.3	78.9	75.6

Table 24: Per-language XStoryCloze sensitivity for $\lambda = 1$–3.5.

Setting	AR	EN	ES	EU	HI	ID	RU	SW	TE	ZH	Avg.
Qwen2.5-7B											
1	82.40	88.29	88.88	50.89	82.93	80.61	89.74	47.52	60.03	80.15	75.14
1.5	85.11	90.01	91.93	53.41	83.32	82.53	89.08	47.51	60.89	80.15	76.39
2	84.84	87.49	91.00	53.74	83.26	82.73	89.81	47.72	61.42	79.42	76.14
2.3	87.66	87.51	89.70	55.03	84.87	82.63	91.48	50.52	64.09	80.31	77.38
2.5	86.87	91.10	93.09	54.50	82.17	84.02	90.31	48.08	60.72	82.76	77.36
2.7	86.23	90.93	92.52	54.40	83.06	83.45	90.07	47.58	62.34	81.87	77.25
3	85.97	90.80	92.52	54.20	82.66	83.59	90.21	47.58	62.41	82.86	77.28
3.5	85.96	91.25	93.11	52.80	84.50	83.25	91.19	49.23	62.07	83.05	77.64

Table 25: Per-language mMMLU sensitivity for $\lambda = 1$–3.5.

Setting	AR	BN	DE	EN	ES	FR	HI	SW	YO	ZH	Avg.
Qwen2.5-7B											
1	55.6	45.9	63.3	72.6	66.2	65.3	46.0	33.8	34.3	67.6	55.1
1.5	55.5	48.1	64.2	73.0	66.4	64.8	46.5	34.2	32.8	67.6	55.3
2	55.2	46.5	63.7	72.5	66.6	64.9	45.4	35.0	34.1	67.5	55.1
2.3	55.2	46.8	64.2	72.7	67.2	65.6	46.3	34.9	35.3	68.5	55.7
2.5	55.6	47.2	64.2	72.6	67.1	65.6	45.9	35.1	35.8	67.7	55.7
2.7	55.6	46.5	64.5	72.3	66.9	64.7	45.0	34.7	34.6	67.3	55.2
3	55.3	47.1	64.2	72.7	67.3	65.3	46.4	35.3	35.2	67.6	55.6
3.5	54.7	46.4	63.8	72.3	67.3	65.2	46.1	35.3	35.6	67.4	55.4

Table 26: Per-language XQuAD sensitivity for $\lambda = 1$–3.5.

Setting	AR	DE	EL	EN	ES	HI	RU	TH	TR	ZH	Avg.
Qwen2.5-7B											
1	78.1	86.1	72.8	89.6	85.7	73.8	78.4	86.4	81.1	89.5	82.2
1.5	78.8	84.4	72.6	89.3	83.8	73.1	77.2	85.8	79.5	89.0	81.4
2	77.2	84.8	71.2	88.9	85.0	72.4	76.7	86.1	78.8	89.8	81.1
2.3	75.9	84.9	72.2	88.7	85.1	72.6	78.5	84.2	79.5	88.2	81.0
2.5	77.2	85.3	72.9	89.5	85.6	72.7	78.9	85.6	80.1	89.6	81.7
2.7	78.9	85.8	71.5	88.8	84.9	72.4	77.2	85.9	80.3	89.1	81.5
3	78.3	84.8	70.0	88.0	83.9	72.3	76.5	85.0	79.4	88.2	80.6
3.5	78.4	85.3	71.2	89.1	85.5	71.8	76.9	85.6	80.0	88.2	81.2

Table 27: Per-language XLSum sensitivity for $\lambda = 1$–3.5.

Setting	EN	ES	FR	KO	PT	SW	TR	UK	VI	ZH	Avg.
Qwen2.5-7B											
1	18.18	17.89	20.16	26.32	19.29	17.34	14.10	12.10	18.63	25.88	18.99
1.5	18.01	17.84	19.95	26.02	18.42	16.53	13.66	12.01	18.90	25.63	18.70
2	18.70	18.64	20.65	26.49	19.43	16.81	13.44	12.46	19.18	26.12	19.19
2.3	17.99	17.91	21.32	26.58	19.80	17.15	13.55	12.76	19.31	26.35	19.27
2.5	18.95	18.64	21.53	27.24	19.27	16.90	14.15	12.99	19.04	25.08	19.38
2.7	18.63	18.37	20.81	26.13	18.88	17.26	14.06	13.19	19.08	24.99	19.14
3	18.49	18.32	21.19	26.31	18.68	16.67	13.64	12.70	19.64	25.13	19.08
3.5	18.89	18.14	20.50	26.47	19.60	17.91	13.52	12.59	19.11	25.78	19.25

E.4 PROTOTYPE-COUNT SENSITIVITY

Table 28: Per-language XNLI sensitivity to the number of prototypes K.

Setting	AR	EL	EN	ES	FR	HI	SW	TR	VI	ZH	Avg.
Qwen2.5-7B											
1	76.5	72.2	90.4	82.2	82.7	70.7	46.8	71.9	78.6	79.9	75.2
2	77.1	73.4	91.2	83.6	82.6	69.9	50.3	71.8	78.8	79.5	75.8
3	77.7	73.2	90.4	83.5	82.9	71.0	47.9	71.2	79.0	79.8	75.7
4	77.9	72.7	90.6	84.2	82.7	70.4	51.5	71.8	79.3	79.1	76.0
5	77.7	72.9	90.6	84.3	82.6	71.0	50.2	71.9	79.3	80.0	76.1
6	77.2	72.6	91.1	84.5	82.6	71.6	51.4	71.6	78.8	80.1	76.2
7	77.6	73.0	90.8	84.1	83.2	71.8	48.1	72.0	79.0	79.6	75.9
8	78.0	73.2	90.3	84.0	82.7	71.7	47.0	71.4	78.9	79.7	75.7

Table 29: Per-language XStoryCloze sensitivity to the number of prototypes K.

Setting	AR	EN	ES	EU	HI	ID	RU	SW	TE	ZH	Avg.
Qwen2.5-7B											
1	84.71	87.89	89.94	51.62	83.32	81.01	89.34	47.58	62.81	82.26	76.05
2	84.81	87.92	89.97	54.63	83.35	83.22	88.45	47.81	62.57	81.50	76.42
3	85.78	87.10	90.48	54.41	83.93	83.46	90.61	48.32	63.94	82.34	77.04
4	86.87	91.10	93.09	54.50	82.17	84.02	90.31	48.08	60.72	82.76	77.36
5	85.01	91.03	92.16	54.70	83.55	83.68	90.04	47.62	62.77	82.43	77.30
6	86.20	87.53	92.49	54.24	85.28	82.83	90.44	48.94	63.90	81.50	77.34
7	85.58	89.35	92.57	52.22	84.19	82.93	90.87	48.65	63.81	83.33	77.35
8	85.55	88.93	91.90	53.19	84.69	82.31	89.85	48.95	62.85	82.70	77.09

Table 30: Per-language mMMLU sensitivity to the number of prototypes K.

Setting	AR	BN	DE	EN	ES	FR	HI	SW	YO	ZH	Avg.
Qwen2.5-7B											
1	54.2	45.8	63.6	73.0	66.5	65.1	45.4	34.4	34.2	67.2	54.9
2	53.8	46.5	63.1	72.6	66.0	64.7	45.6	34.9	34.1	67.1	54.8
3	55.3	47.3	64.5	72.9	66.7	65.3	45.4	35.4	35.1	67.8	55.6
4	55.6	47.2	64.2	72.6	67.1	65.6	45.9	35.1	35.8	67.7	55.7
5	55.0	46.1	63.8	72.7	66.0	64.9	45.9	34.5	35.0	67.2	55.1
6	55.7	46.4	63.7	73.0	67.1	65.5	46.4	35.7	34.2	68.0	55.6
7	54.9	46.2	63.7	72.5	66.8	65.7	45.8	34.6	34.3	67.4	55.2
8	55.0	46.1	64.0	72.8	66.9	65.5	46.3	35.7	34.4	67.5	55.4

Table 31: Per-language XQuAD sensitivity to the number of prototypes K.

Setting	AR	DE	EL	EN	ES	HI	RU	TH	TR	ZH	Avg.
Qwen2.5-7B											
1	77.3	85.1	70.7	88.6	84.1	71.5	76.6	85.1	78.3	88.3	80.6
2	78.8	83.2	70.7	88.3	83.3	71.6	77.1	85.2	78.4	85.8	80.2
3	77.9	84.7	70.3	89.3	84.8	71.9	78.3	85.5	78.7	89.1	81.1
4	77.2	85.3	72.9	89.5	85.6	72.7	78.9	85.6	80.1	89.6	81.7
5	77.5	85.5	72.7	89.1	85.6	71.9	78.6	85.7	80.7	87.6	81.5
6	77.6	84.8	70.9	89.2	85.4	72.7	77.0	85.7	79.8	89.6	81.3
7	77.0	85.6	71.8	90.1	85.8	72.5	78.9	86.0	81.2	89.4	81.8
8	77.3	85.3	69.9	89.0	84.7	71.6	77.5	85.4	79.6	89.7	81.0

Table 32: Per-language XLSum sensitivity to the number of prototypes K.

Setting	EN	ES	FR	KO	PT	SW	TR	UK	VI	ZH	Avg.
Qwen2.5-7B											
1	18.94	17.93	20.47	26.11	18.24	17.13	13.98	11.82	18.44	25.81	18.89
2	17.33	17.22	18.67	25.86	16.47	16.45	13.32	11.14	18.61	25.80	18.09
3	18.69	18.59	20.89	26.23	19.37	17.25	14.17	12.68	19.13	25.12	19.21
4	18.95	18.64	21.53	27.24	19.27	16.90	14.15	12.99	19.04	25.08	19.38
5	18.31	18.06	20.77	26.34	18.10	17.27	14.47	13.08	19.03	26.04	19.15
6	17.90	17.98	20.65	25.42	19.48	17.67	13.59	11.90	19.65	25.54	18.98
7	18.76	18.06	20.06	26.43	19.64	17.01	13.60	12.55	19.05	25.66	19.08
8	18.31	18.36	20.61	26.05	19.77	16.31	13.89	12.77	18.97	25.90	19.09