GOOD SOURCES OR GOOD MATCHES: DISENTANGLING SOURCE QUALITY IN LANGUAGE MODEL TRANSFER

Anonymous authors
Paper under double-blind review

ABSTRACT

Data-selection methods seek to predict which training sources will benefit a target, but prediction accuracy alone does not reveal whether they identify generally useful sources or target-specific matches. This distinction determines whether past transfer experiments can guide future selection or whether new target-specific evidence is needed. We investigate this question through adapter transfer on 15 datasets and four language models. We measure relative source usefulness through target-conditioned rankings and use HodgeRank to identify how much transfer asymmetry is explained by a common source ordering. This ordering corresponds to average source quality and explains most squared asymmetry in the decoder-only models studied; T5 shows more directional variation that a single ordering cannot explain. We then evaluate two ways to reuse measured transfer. Historical source quality guides selection for incoming targets from an established model library, outperforming the tested gradient references in the decoder-only settings. Transfer measured on a smaller pilot also supports source selection for Qwen3-8B and LLaMA, while target-model gradients remain stronger for T5. Further analysis shows that the pilot's general source ordering predicts target-model asymmetry more accurately than its full directional structure. These results clarify what makes transfer evidence useful and provide practical ways to select source datasets using observations from previous targets or a smaller model.

1 INTRODUCTION

Selecting training data requires understanding how a source dataset affects performance on the intended targets. A prominent approach estimates this usefulness through model gradients, including influence-based methods and gradient-space dataset valuation (Koh & Liang, 2017; Pruthi et al., 2020; Xia et al., 2024; Zeng et al., 2026). Another approach uses smaller models to choose data-mixture weights for larger-model training (Xie et al., 2023; Liu et al., 2025). These methods address how to select or allocate training data without evaluating every alternative at full scale. We investigate a complementary question: what structure in measured source–target transfer can be reused to make these decisions?

The distinction is between a generally useful source and a good match for a particular target. A dataset may perform well across many targets, making its historical transfer record informative before a new target is examined. Alternatively, its usefulness may depend on a specific relationship that requires target-specific evidence. Both effects can produce accurate transfer predictions, but they imply different selection strategies. The same distinction arises across models: a smaller model may identify the same useful sources as a larger model without reproducing its detailed transfer behavior.

Task-transfer studies provide a foundation for examining these relationships (Zamir et al., 2018; Vu et al., 2020; Zhang et al., 2026). Our focus is on separating the contribution of general source quality from the remaining directional structure, then testing how this distinction guides source selection. This requires a suitable measurement scale. Source-trained adapters can be compared directly on a fixed target, whereas comparisons across targets also reflect differences in evaluation metrics, difficulty, and score ranges. We therefore represent each source–target relationship by the source's rank percentile among the candidates for that target. This target-conditioned representation measures relative source usefulness across heterogeneous tasks.

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.

Target-conditioned rankings reveal directional relationships: a dataset may rank highly as a source for another task without the reverse relationship holding. Such asymmetry can arise from a general difference in source quality. A broadly useful source may rank well on many targets, while those tasks rank poorly as sources for it. Observing directionality therefore raises a further question: how much can be explained by a common source ordering, and how much requires a description of individual relationships?

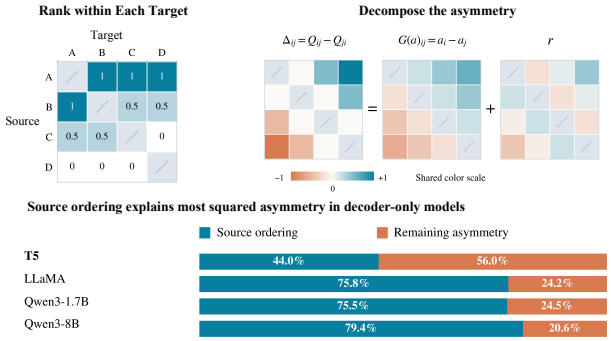


Figure 1: Target-conditioned measurement and decomposition of transfer asymmetry into source-general and pair-specific components.

Figure 1 summarizes how target-conditioned transfer rankings are decomposed into a common source ordering and residual directional structure. We use HodgeRank (Jiang et al., 2011) to perform this decomposition, and separate transfer asymmetry into differences explained by a common source ordering and a remaining residual. Under our target-conditioned measurement, this ordering coincides with the ranking of sources by their average performance across targets. The analysis therefore quantifies how much observed directionality reflects generally useful sources. We evaluate adapter transfer across 15 datasets using T5, LLaMA, Qwen3-1.7B, and Qwen3-8B. A common source ordering explains most squared asymmetry in the decoder-only models, while T5 retains a larger residual share.

We then use this insight in two source-selection settings. In the first, an established target-model library supplies historical transfer observations for selecting sources on incoming targets. This tests whether source quality persists across targets and whether target-specific gradient evidence improves selection within a quality-based shortlist. In the second, adapters trained on a smaller pilot supply measured source–target transfer for selecting datasets on another model. Unlike small-model mixture optimization, this setting reuses observed transfer behavior rather than learned mixing proportions. It supports selecting a source for one target or a common source for several targets.

These settings make measured transfer both an object of study and a practical source of selection evidence. Historical source quality outperforms the tested gradient references in the decoder-only settings, with more limited gains from gradient reranking. Qwen3-1.7B pilot transfer also supports source selection for Qwen3-8B and LLaMA, whereas target-model gradients are stronger for T5. Component analysis reveals a further distinction: retaining only the pilot's source ordering improves target-model asymmetry prediction relative to using its full directional structure. A pilot can therefore provide useful selection evidence even when its residual asymmetry transfers poorly.

Our contributions are threefold:

- **Diagnosing general source quality in transfer asymmetry.** We apply HodgeRank to target-conditioned transfer rankings and connect the fitted source ordering to average source quality. The diagnostic distinguishes directionality explained by generally useful sources from residual asymmetry, revealing the different contributions of these components across the models studied.
- **Reusing historical transfer for incoming-target selection.** We estimate source quality solely within an established target-model library and use it to select sources for targets outside that library. Source-quality screening followed by gradient-cosine reranking tests the additional value of target-specific evidence. The results identify historical source quality as a strong selection signal and quantify the more limited contribution of gradient refinement.
- **Using measured pilot transfer for cross-model source selection.** We use transfer observed on a smaller model to rank candidate sources and select a common source for multiple targets, without target-model transfer observations at selection time. We further compare

the transferability of the pilot's source ordering and residual asymmetry. An exact residual-benefit criterion explains the component ablation and clarifies which part of the pilot's directional structure provides reliable cross-model evidence.

Our central perspective is that identifying good sources and recovering good source–target matches are related but distinct objectives. Studying them separately clarifies what transfer predictors capture and how measured transfer can guide data selection across targets and models.

2 Related Work

Task relationships and transfer prediction. Taskonomy (Zamir et al., 2018) studies directed transfer among visual tasks through target-specific readouts on source representations, using within-target ordinal comparisons to normalize heterogeneous task losses. In NLP, Vu et al. (2020) investigate cross-task transfer and use task embeddings to identify effective intermediate training tasks. The SAE-based Transferability Score of Zhang et al. (2026) uses in-context demonstrations to predict fine-tuning-induced feature shifts and their relevance to downstream domains. Its primary evaluation fixes a training source and predicts performance changes across domains. For comparison data more generally, HodgeRank (Jiang et al., 2011) extracts a least-squares global ordering and an orthogonal cyclic residual, providing a diagnostic of how well a single ordering explains the observations.

Data attribution and dataset valuation. Influence functions (Koh & Liang, 2017) estimate the effect of infinitesimal training-example reweighting on test loss through inverse-Hessian-weighted gradients. TracIn (Pruthi et al., 2020) attributes loss changes using gradient products along training trajectories, while LESS (Xia et al., 2024) uses low-rank gradient representations for targeted instruction-data selection. Datamodels (Ilyas et al., 2022) learn the dependence of model predictions on training-subset membership through repeated training experiments. At the dataset level, Zeng et al. (2026) formulate auxiliary-source valuation through gradient-space kernel mean matching, accounting jointly for target alignment and source redundancy when selecting datasets to combine with target-task data.

Small-model proxies for data selection. DoReMi (Xie et al., 2023) optimizes pretraining domain weights using a small proxy model and distributionally robust optimization of excess loss. RegMix (Liu et al., 2025) fits a regression surrogate to small-model mixture experiments and transfers promising mixtures to larger-model training. Cross-model selection is also studied in LESS (Xia et al., 2024) and convex dataset valuation (Zeng et al., 2026), where model-derived scores guide data selection for another model. These methods use small-scale optimization or valuation to inform more expensive training.

Task-transfer studies predict relationships between datasets, but leave open how much successful prediction comes from identifying sources that work well across many targets. Attribution and valuation methods estimate usefulness for a specified target; they leave open when past transfer performance alone is sufficient for selecting a source for a new target. Small-model mixture and valuation methods show how cheaper models can guide data selection, but do not answer whether measured adapter transfer identifies the same good sources across models or also recovers their remaining directional relationships. We address these questions by applying HodgeRank to transfer asymmetry, testing historical source quality on incoming targets, and using measured pilot transfer to select sources for another model.

3 Measuring Data Relationships

We study how transfer relationships reflect general source quality and target-specific usefulness. Using target-conditioned rankings, we develop a diagnostic that quantifies how much observed directionality is explained by a common source ordering. Our experiments use the 15-task benchmark adopted by Wang et al. (2023): Yelp, Amazon, DBpedia, Yahoo Answers, AG News, MNLI, QQP, RTE, SST-2, WiC, CB, COPA, MultiRC, BoolQ, and IMDB. We evaluate four language models: T5, LLaMA, Qwen3-1.7B, and Qwen3-8B. Rather than following the sequential continual-learning protocol of Wang et al. (2023), we train each source adapter independently from the corresponding

base checkpoint and evaluate it on the other 14 tasks, yielding $15 \times 14 = 210$ directed transfer scores per model.

Target-conditioned measurement. Let A_{ij} denote the performance of an adapter trained on source task i and evaluated on target task j. For a fixed target, source adapters share the same evaluation task and metric. Across targets, differences in metric, difficulty, answer space, and score range change the interpretation of numerical scores. For example, the T5 dbpedia adapter scores 21.00 on MultiRC and 18.09 on Yahoo, yet ranks fourth worst among the 14 sources for MultiRC and first for Yahoo. Across within-source target comparisons, raw scores and within-target ranks disagree on 21.2% of target pairs for T5 and 17.1% for LLaMA.

We therefore evaluate each source relative to the alternatives available for its target. Let $\mathcal{S}_j$ be the candidate-source set for target j, with $n_j = |\mathcal{S}_j| > 1$. Define

$$Q_{ij} = \frac{\mathrm{rank}_j(A_{ij}) - 1}{n_j - 1},$$

where rank 1 denotes the worst source and rank n_j the best; ties receive average ranks. The percentile Q_{ij} places relative source quality on a common ordinal scale and is invariant to strictly increasing transformations of each target's scores. In the full benchmark, each target has the other 14 tasks as candidate sources.

Observed directionality. For distinct tasks, define $\Delta_{ij} = Q_{ij} - Q_{ji}$. Positive Δ_{ij} means that i ranks more highly as a source for j than j ranks as a source for i. Across the four models, mean $|\Delta_{ij}|$ ranges from 0.240 to 0.286. Figure A.1 visualizes this structure through top-3 transfer neighborhoods. Reciprocal edges identify mutual top-3 membership, while one-way edges identify directional membership. To explain these relationships, we distinguish differences in general source quality from asymmetry specific to individual task pairs.

Separating source-general effects. For the complete transfer matrix of N tasks, define

$$s_i = \frac{1}{N-1} \sum_{j \neq i} Q_{ij}, \qquad E_{ij} = Q_{ij} - s_i.$$

The score s_i measures source i's average relative quality across the benchmark, and E_{ij} records its departure from that average on target j. A source with high s_i can induce directional relationships simply by being more useful on average than other sources.

To quantify this contribution, represent asymmetry using one score per task: $G(b)_{ij} = b_i - b_j$. Define the edge inner product $\langle X, Y \rangle_{\mathrm{edge}} = \sum_{i<j} X_{ij} Y_{ij}$ and its induced norm $\| \cdot \|_{\mathrm{edge}}$. The following result identifies the best source-difference representation and characterizes its remaining error.

Proposition 3.1 (Identification of source-general effects)**.** For $N \geq 3$, the unique minimizer of $\|\Delta - G(b)\|^2_{\mathrm{edge}}$ subject to $\sum_i b_i = 0$ is

$$a_i = \frac{1}{N} \sum_{j \neq i} \Delta_{ij} = \frac{N-1}{N} \left(s_i - \frac{1}{2} \right).$$

Writing $r = \Delta - G(a)$, every centered vector b satisfies

$$\|\Delta - G(b)\|^2_{\mathrm{edge}} = \|r\|^2_{\mathrm{edge}} + N\|b - a\|^2_2.$$

The optimal source ordering is therefore exactly the ordering induced by average source quality. The residual r captures the asymmetry that remains after the best such ordering has been removed, with $\|r\|^2_{\mathrm{edge}}$ giving the minimum error attainable by any source-difference representation. Taking $b = 0$ yields

$$\|\Delta\|^2_{\mathrm{edge}} = \|G(a)\|^2_{\mathrm{edge}} + \|r\|^2_{\mathrm{edge}}.$$

This orthogonal decomposition assigns a source-general share $\eta_{\mathrm{src}} = \|G(a)\|^2_{\mathrm{edge}} / \|\Delta\|^2_{\mathrm{edge}}$ whenever $\Delta \neq 0$. Algorithm 1 computes the source scores, transfer deviations, and asymmetry components directly. Its two residuals serve distinct purposes: E describes target-dependent deviations in the full transfer matrix, while r isolates asymmetry unexplained by the global source ordering.

Algorithm 1 Separating Source-General Effects

Require: Complete, finite transfer scores A_{ij} for $N \geq 3$ tasks, $i \neq j$; higher scores are better.
Ensure: Source quality s, target-dependent deviations E, residual asymmetry r, and source-general asymmetry share η_{src}.

for each target j **do**
  Rank $\{A_{ij} : i \neq j\}$ from worst to best, using average ranks for ties.
  $Q_{ij} \leftarrow \big(\mathrm{rank}_j(A_{ij}) - 1\big)/(N-2)$ for all $i \neq j$.
end for
for each source i **do**
  $s_i \leftarrow \frac{1}{N-1}\sum_{j \neq i} Q_{ij}$.
  $a_i \leftarrow \frac{N-1}{N}\big(s_i - \frac{1}{2}\big)$.
end for
for each ordered pair $i \neq j$ **do**
  $E_{ij} \leftarrow Q_{ij} - s_i$.
  $\Delta_{ij} \leftarrow Q_{ij} - Q_{ji}$.
  $r_{ij} \leftarrow \Delta_{ij} - (a_i - a_j)$.
end for
$D \leftarrow \sum_{i<j} \Delta_{ij}^2$.
if $D > 0$ **then**
  $\eta_{\mathrm{src}} \leftarrow \big(\sum_{i<j}(a_i - a_j)^2\big)/D$.
else
  $\eta_{\mathrm{src}} \leftarrow 0$ ▷ Convention when no asymmetry is observed.
end if
return $s, E, r, \eta_{\mathrm{src}}$.

Model	Source share	Residual share	Mean $\lvert\Delta\rvert$	Mean $\lvert r\rvert$
T5	44.0%	56.0%	0.246	0.188
LLaMA	75.8%	24.2%	0.240	0.114
Qwen3-1.7B	75.5%	24.5%	0.286	0.148
Qwen3-8B	79.4%	20.6%	0.267	0.116

Table 1: Decomposition of transfer asymmetry over 105 task pairs per model. Source and residual shares partition squared asymmetry; the final two columns report mean absolute asymmetry before and after removing the source-ordering component.

Observed source-general structure. Table 1 reports the decomposition over all 105 unordered task pairs per model. A single source ordering explains 75.8%, 75.5%, and 79.4% of squared asymmetry in LLaMA, Qwen3-1.7B, and Qwen3-8B, respectively. T5 exhibits a different balance, with 44.0% assigned to the source-general component and 56.0% to the residual.

For comparison, independently permuting source assignments within each target column yields mean source-general shares of 6.6%–6.8% over $1{,}000$ permutations. These permutations preserve each target's percentile values while disrupting consistent source identity across targets. The observed shares therefore reveal substantial structure associated with source identity.

General source quality accounts for the dominant share of observed asymmetry in the three decoder-only models. This finding motivates a practical use of the diagnostic: estimate source quality from historical transfer observations and reuse the resulting ordering to select sources for incoming targets.

4 Source Quality for Incoming Targets

The source-general structure identified above suggests a simple selection rule: reuse historical source quality when a new target arrives. We evaluate this rule within an established library and test whether target-specific gradient alignment improves selection.

Source selection. Partition the tasks into a source library $\mathcal{L}$ and incoming targets $\mathcal{H}$. For $n = |\mathcal{L}|$, define

$$s_i^{\mathcal{L}} = \frac{1}{n-1} \sum_{j \in \mathcal{L} \setminus \{i\}} Q_{ij}^{\mathcal{L}}, \qquad \widehat{i}_{\mathrm{src}} = \arg\max_{i \in \mathcal{L}} s_i^{\mathcal{L}},$$

where $Q_{ij}^{\mathcal{L}}$ ranks source i among $\mathcal{L} \setminus \{j\}$. This rule, *Source Qual.*, assigns the same source ordering to every incoming target. It assumes an existing library of trained adapters and internal transfer evaluations, but requires no incoming-target scoring. Incoming tasks are excluded from the library, and their transfer outcomes are used only for evaluation.

To incorporate target-specific evidence, *Screen+cos.* retains the K highest-quality sources and reranks them by gradient cosine:

$$\mathcal{C}_K = \mathrm{TopK}_{i \in \mathcal{L}}\, s_i^{\mathcal{L}}, \qquad \widehat{i}(t) = \arg\max_{i \in \mathcal{C}_K} G_{it},$$

where G_{it} is the within-target gradient cosine percentile. Cosine measures alignment independently of positive gradient rescaling. Sources outside the shortlist retain their source-quality ordering.

We select K by inner leave-one-library-task-out validation, removing each validation task from both the candidate sources and the targets used to estimate quality. Internal percentiles are recomputed in each fold. The smallest K within one standard error of the best mean selected-source percentile is retained. Thus, $K = 1$ recovers Source Qual., while larger shortlists permit target-specific selection.

Evaluation. We use 50 library–incoming splits for each $n \in \{7, 10, 13\}$ on all four models. Grad Dot and Grad Cos rank the same library sources by global gradient dot product and cosine, respectively. Gradient scores are converted to within-target percentiles and averaged across available gradient seeds. Table 2 reports selected-source percentile, top-1 hit rate, and top-3 set recall. Repeated evaluations are averaged within each incoming task before computing target-balanced means and standard errors.

Results. Historical source quality identifies high-ranking sources for incoming targets in all three decoder-only models. With ten library tasks, Source Qual. attains a selected-source percentile of 0.967 for Qwen3-8B, compared with 0.617 for Grad Cos and 0.391 for Grad Dot. Its top-3 recall also exceeds both gradient baselines at every tested library size.

Screen+cos. yields small and inconsistent gains over Source Qual., indicating that the principal benefit comes from historical source quality. Target-specific evidence nevertheless remains useful in some settings: at library size 13, Grad Cos achieves higher top-1 recovery than Source Qual. for both T5 and LLaMA. The results support reusing source-general effects to identify useful candidates, while distinguishing this objective from exact recovery of the best source for each target.

The cost of obtaining transfer evidence can be reduced through partial evaluation or a smaller pilot. Appendix A shows that adaptive selection using only 50% of internal library evaluations achieves higher mean selected-source percentiles than both gradient references in all nine decoder-only settings. The next section explores the complementary approach of shifting source-adapter training and transfer evaluation to a smaller pilot model.

5 A Smaller Model as a Transfer Pilot

The preceding section uses historical target-model transfer to identify generally useful sources. We now ask whether a smaller model can supply this evidence before target-model source adapters are trained. We examine ranking recovery across several pilot–target model pairs, including T5 as a pilot, and use Qwen3-1.7B as the pilot for source selection on Qwen3-8B, LLaMA, and T5.

Source adapters are trained on the pilot and evaluated on target-task data. Their measured transfer scores guide source selection for the target model, whose transfer observations are reserved for evaluation. This shifts source-adapter training to the pilot. We first examine ranking recovery and source selection, then use our diagnostic to determine whether source-general ordering or residual asymmetry transfers more reliably.

Model	$\lvert\mathcal{L}\rvert$	Grad Dot	Grad Cos	Source Qual.	Screen+cos.
(a) Selected-source percentile					
T5	7	0.703 ± 0.053	0.609 ± 0.054	$\mathbf{0.746} \pm 0.043$	0.738 ± 0.051
	10	0.724 ± 0.051	0.611 ± 0.067	$\mathbf{0.756} \pm 0.043$	$\mathbf{0.756} \pm 0.048$
	13	0.768 ± 0.059	0.618 ± 0.093	0.748 ± 0.044	$\mathbf{0.774} \pm 0.049$
LLaMA	7	0.574 ± 0.047	0.589 ± 0.053	$\mathbf{0.840} \pm 0.024$	0.823 ± 0.022
	10	0.572 ± 0.058	0.587 ± 0.066	$\mathbf{0.859} \pm 0.021$	0.855 ± 0.022
	13	0.602 ± 0.085	0.660 ± 0.081	0.839 ± 0.024	$\mathbf{0.851} \pm 0.027$
Qwen3-1.7B	7	0.590 ± 0.037	0.606 ± 0.049	0.852 ± 0.014	$\mathbf{0.859} \pm 0.011$
	10	0.600 ± 0.037	0.659 ± 0.056	$\mathbf{0.890} \pm 0.022$	0.887 ± 0.021
	13	0.610 ± 0.059	0.694 ± 0.069	0.902 ± 0.028	$\mathbf{0.906} \pm 0.028$
Qwen3-8B	7	0.420 ± 0.027	0.583 ± 0.054	0.919 ± 0.011	$\mathbf{0.920} \pm 0.010$
	10	0.391 ± 0.033	0.617 ± 0.060	$\mathbf{0.967} \pm 0.009$	0.965 ± 0.009
	13	0.364 ± 0.040	0.681 ± 0.070	$\mathbf{0.976} \pm 0.010$	$\mathbf{0.976} \pm 0.010$
(b) Top-1 hit rate					
T5	7	$\mathbf{0.402} \pm 0.078$	0.370 ± 0.081	0.305 ± 0.067	0.320 ± 0.076
	10	0.361 ± 0.086	$\mathbf{0.392} \pm 0.097$	0.150 ± 0.051	0.208 ± 0.069
	13	0.388 ± 0.117	$\mathbf{0.426} \pm 0.122$	0.086 ± 0.048	0.190 ± 0.081
LLaMA	7	0.287 ± 0.068	0.354 ± 0.074	$\mathbf{0.464} \pm 0.058$	0.455 ± 0.057
	10	0.246 ± 0.080	0.307 ± 0.085	0.302 ± 0.062	$\mathbf{0.316} \pm 0.064$
	13	0.321 ± 0.117	$\mathbf{0.399} \pm 0.121$	0.156 ± 0.068	0.209 ± 0.088
Qwen3-1.7B	7	0.149 ± 0.040	0.244 ± 0.059	0.471 ± 0.043	$\mathbf{0.501} \pm 0.035$
	10	0.103 ± 0.041	0.193 ± 0.056	$\mathbf{0.496} \pm 0.083$	0.479 ± 0.077
	13	0.078 ± 0.067	0.122 ± 0.069	0.470 ± 0.116	$\mathbf{0.489} \pm 0.116$
Qwen3-8B	7	0.081 ± 0.018	0.246 ± 0.075	0.725 ± 0.034	$\mathbf{0.731} \pm 0.034$
	10	0.015 ± 0.008	0.171 ± 0.074	$\mathbf{0.781} \pm 0.057$	0.770 ± 0.059
	13	0.000 ± 0.000	0.168 ± 0.091	$\mathbf{0.776} \pm 0.095$	$\mathbf{0.776} \pm 0.095$
(c) Top-3 recall					
T5	7	0.577 ± 0.045	0.564 ± 0.038	0.592 ± 0.043	$\mathbf{0.597} \pm 0.043$
	10	0.474 ± 0.060	0.452 ± 0.053	0.548 ± 0.070	$\mathbf{0.552} \pm 0.070$
	13	0.422 ± 0.068	0.385 ± 0.067	$\mathbf{0.462} \pm 0.071$	$\mathbf{0.462} \pm 0.071$
LLaMA	7	0.471 ± 0.024	0.439 ± 0.029	$\mathbf{0.687} \pm 0.029$	0.682 ± 0.029
	10	0.347 ± 0.031	0.314 ± 0.042	$\mathbf{0.660} \pm 0.042$	0.654 ± 0.041
	13	0.316 ± 0.051	0.296 ± 0.060	$\mathbf{0.607} \pm 0.061$	$\mathbf{0.607} \pm 0.061$
Qwen3-1.7B	7	0.597 ± 0.032	0.577 ± 0.024	$\mathbf{0.765} \pm 0.010$	$\mathbf{0.765} \pm 0.010$
	10	0.392 ± 0.047	0.443 ± 0.042	$\mathbf{0.718} \pm 0.019$	$\mathbf{0.718} \pm 0.019$
	13	0.307 ± 0.058	0.390 ± 0.057	$\mathbf{0.711} \pm 0.040$	$\mathbf{0.711} \pm 0.040$
Qwen3-8B	7	0.377 ± 0.031	0.443 ± 0.030	$\mathbf{0.706} \pm 0.017$	0.705 ± 0.016
	10	0.191 ± 0.032	0.272 ± 0.057	$\mathbf{0.745} \pm 0.032$	$\mathbf{0.745} \pm 0.032$
	13	0.152 ± 0.041	0.226 ± 0.075	$\mathbf{0.775} \pm 0.049$	$\mathbf{0.775} \pm 0.049$

Table 2: Source selection for incoming targets. Source quality and shortlist size are estimated only within $\mathcal{L}$. Values are target-balanced means $\pm$ task-clustered standard errors over 50 splits. Higher is better; all methods tied for the largest reported mean are bold.

Gradient references. We compare pilot transfer with four predictors computed from the target model's gradients: Grad Dot, Grad Cos, Dot Vote, and Cos Vote. Grad Dot and Grad Cos aggregate gradient geometry globally. The voting variants rank candidate sources separately at each layer and average their target-conditioned percentiles, giving equal weight to layer rankings rather than raw score magnitudes.

Recovering source relationships. For a task subset $\mathcal{U}$ and target $j \in \mathcal{U}$, each method ranks the sources $\mathcal{U} \setminus \{j\}$. Pilot transfer uses the measured score A^P_{ij}, while gradient predictors use the corresponding target-model gradient statistic. Percentiles are recomputed within each candidate set and compared with the target model's observed source ranking.

We evaluate subsets of sizes 3 through 15 using macro within-target Spearman correlation. Figure 2 shows that measured pilot transfer outperforms the tested gradient predictors for the three decoder-only pilot–target pairs shown. With T5 as the pilot, the comparison depends on the target: gradients perform better for Qwen3-1.7B, results are mixed for Qwen3-8B, and pilot transfer performs better for LLaMA. Figure A.2 in the appendix reports the corresponding within-target pairwise accuracy.

Selecting one source for multiple targets. Given a target set $\mathcal{T}$ of size k, we select one source dataset to serve all targets. For the full task set $\mathcal{D}$, the candidate sources are $\mathcal{S} = \mathcal{D} \setminus \mathcal{T}$. Let $Q^{e,\mathcal{S}}_{ij}$ denote the percentile assigned by evidence method e to source i for target j, computed among $\mathcal{S}$. We

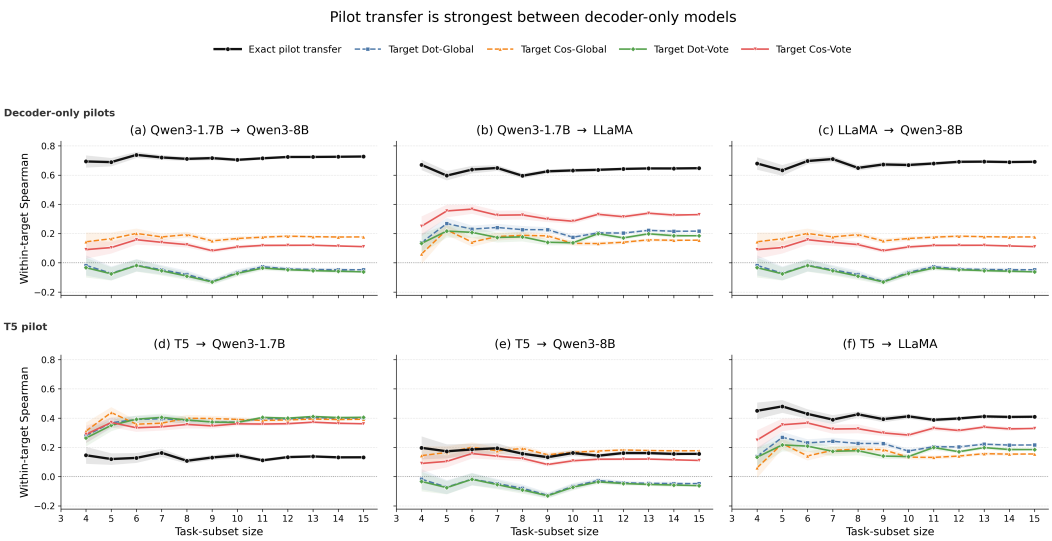


Figure 2: Within-target Spearman recovery across pilot–target model pairs and task-subset sizes. Pilot curves use measured pilot transfer; gradient curves use the target model's gradient geometry.

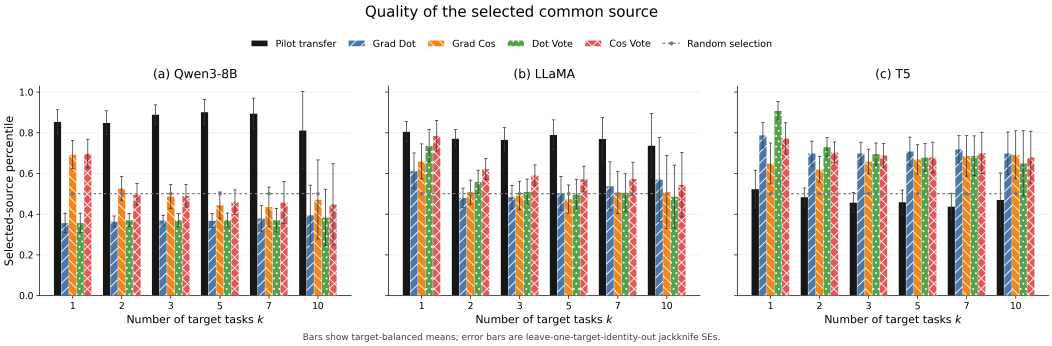


Figure 3: Target-model quality of the common source selected for k target tasks using Qwen3-1.7B pilot transfer or target-model gradient predictors. Bars report its mean target-conditioned transfer percentile; the dotted line denotes random selection.

select

$$U_i^e(\mathcal{T}) = \frac{1}{|\mathcal{T}|} \sum_{j \in \mathcal{T}} Q_{ij}^{e,\mathcal{S}}, \qquad \widehat{i}_e(\mathcal{T}) = \arg\max_{i \in \mathcal{S}} U_i^e(\mathcal{T}).$$

This target-balanced Borda rule maximizes mean within-target percentile. The selected source is evaluated by its actual target-model utility, $\mathrm{SelPct}_M(e, \mathcal{T}) = U^M_{\widehat{i}_e(\mathcal{T})}(\mathcal{T})$.

We evaluate $k \in \{1, 2, 3, 5, 7, 10\}$, enumerating all target sets for $k = 1, 2$ and sampling 200 distinct sets with fixed random seeds for each larger k. Error bars are leave-one-target-identity-out jackknife standard errors.

Figure 3 shows that ranking recovery translates into effective source selection. For Qwen3-8B, the pilot achieves selected-source percentiles of 0.854 for one target and 0.901 for five targets, compared with 0.697 and 0.459 for the best gradient predictor in each setting. For LLaMA, the corresponding pilot values are 0.805 and 0.789, versus 0.785 and 0.572 for gradients. For T5, the pilot obtains 0.523 and 0.459, whereas gradients attain 0.908 and 0.709.

Figures A.3 and A.4 in the appendix report top-1 recovery and top-3 recall. These metrics distinguish selecting a high-quality source from recovering the exact optimum. For example, Cos Vote achieves

Target model	Full pilot	Source only	Reduction
Qwen3-8B	0.0769	**0.0540**	29.8%
LLaMA	0.1037	**0.0642**	38.1%
T5	0.2513	**0.2224**	11.5%

Table 3: Asymmetry prediction MSE over 105 task pairs. Source only retains the pilot's source-general component. Reduction is relative to full-pilot MSE; lower error is better.

higher top-1 recovery for LLaMA with one target, despite the pilot's slightly higher selected-source percentile.

What transfers from the pilot? We next distinguish agreement in source-general ordering from agreement in residual asymmetry. For pilot P and target model M evaluated on the same N tasks, Algorithm 1 gives $\Delta^X = G(a^X) + r^X$, where $G(a^X)_{ij} = a_i^X - a_j^X$. The full pilot predicts Δ^P, while the source-only predictor retains $G(a^P)$. Define their asymmetry prediction errors by

$$L_{\text{src}} = \frac{\|\Delta^M - G(a^P)\|^2_{\text{edge}}}{\binom{N}{2}}, \qquad L_{\text{full}} = \frac{\|\Delta^M - \Delta^P\|^2_{\text{edge}}}{\binom{N}{2}}.$$

The orthogonality in Proposition 3.1 isolates the contribution of the pilot residual.

Corollary 5.1 (Value of pilot residuals)**.** The benefit of retaining the pilot residual is

$$L_{\text{src}} - L_{\text{full}} = \frac{2\langle r^M, r^P\rangle_{\text{edge}} - \|r^P\|^2_{\text{edge}}}{\binom{N}{2}}.$$

Consequently, the full pilot has lower prediction error if and only if $2\langle r^M, r^P\rangle_{\text{edge}} > \|r^P\|^2_{\text{edge}}$.

The criterion separates residual alignment from residual magnitude: positive alignment alone is insufficient for the pilot residual to improve prediction. We evaluate this criterion on the complete 15-task matrices, comparing both pilot predictions against the same 105 target-model task-pair asymmetries.

Table 3 shows that retaining only the source-general component reduces asymmetry prediction error for all three target models. The reductions are 29.8% for Qwen3-8B and 38.1% for LLaMA. Although Qwen3-8B has positive residual alignment with the pilot, this alignment does not offset the pilot residual's squared magnitude.

T5 differs at the source-general level as well. The source-component contribution to full-pilot MSE, $\|G(a^M) - G(a^P)\|^2_{\text{edge}}/\binom{N}{2}$, is 0.1625 for T5, compared with 0.0410 for LLaMA and 0.0299 for Qwen3-8B. Its weaker pilot recovery therefore reflects disagreement in source ordering as well as residual structure.

Together, these experiments connect pilot transfer to the source-general effects identified earlier. Measured pilot transfer supports source selection for Qwen3-8B and LLaMA, while the diagnostic reveals that its source-general ordering predicts target-model asymmetry more reliably than its full directional structure. The distinction explains how a pilot can provide useful source-quality evidence even when its residual asymmetry transfers poorly.

6 Conclusion

We studied whether data transfer reflects generally useful sources or target-specific matches. Target-conditioned rankings place heterogeneous transfer outcomes on a common ordinal scale, while the HodgeRank decomposition separates directionality explained by a common source ordering from pair-specific residual structure. Across the decoder-only models studied, general source quality explains most transfer asymmetry and provides a strong signal for incoming-target selection. This signal remains effective under adaptive partial evaluation, while transfer measured on a smaller related pilot provides an alternative when target-model transfer observations are unavailable. Together, these results distinguish finding good sources from recovering detailed source–target matches and provide a principled basis for reusing transfer evidence across targets and models.

AI USE STATEMENT

Generative AI tools were used for language editing and code debugging. All changes were reviewed and verified by the authors, who take full responsibility for the final manuscript.

ETHICS STATEMENT

This work uses existing public benchmarks and involves no new data collection or human participants.

REPRODUCIBILITY STATEMENT

Experimental details and proofs are provided in the appendix. Code, configurations, and saved data splits will be released with the supplementary material.

REFERENCES

Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. Datamodels: Understanding predictions with data and data with predictions. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9525–9587. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/ilyas22a.html.

Xiaoye Jiang, Lek-Heng Lim, Yuan Yao, and Yinyu Ye. Statistical ranking and combinatorial Hodge theory. *Mathematical Programming*, 127(1):203–244, 2011. doi: 10.1007/s10107-010-0419-x. URL https://doi.org/10.1007/s10107-010-0419-x.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1885–1894. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/koh17a.html.

Qian Liu, Xiaosen Zheng, Niklas Muennighoff, Guangtao Zeng, Longxu Dou, Tianyu Pang, Jing Jiang, and Min Lin. RegMix: Data mixture as regression for language model pre-training. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 38305–38339, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/5f67d864aae6115374fed7beddd119e0-Abstract-Conference.html.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 19920–19930. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/e6385d39ec9394f2f3a354d9d2b88eec-Paper.pdf.

Tu Vu, Tong Wang, Tsendsuren Munkhdalai, Alessandro Sordoni, Adam Trischler, Andrew Mattarella-Micke, Subhransu Maji, and Mohit Iyyer. Exploring and predicting transferability across NLP tasks. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7882–7926, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.635. URL https://aclanthology.org/2020.emnlp-main.635/.

Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuanjing Huang. Orthogonal subspace learning for language model continual learning. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10658–10671, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.715. URL https://aclanthology.org/2023.findings-emnlp.715/.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. LESS: Selecting influential data for targeted instruction tuning. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54104–54132. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/xia24c.html.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy S Liang, Quoc V Le, Tengyu Ma, and Adams Wei Yu. DoReMi: Optimizing data mixtures speeds up language model pretraining. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 69798–69818. Curran Associates, Inc., 2023. doi: 10.52202/075280-3059. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/dcba6be91359358c2355cd920da3fcbd-Paper-Conference.pdf.

Amir R. Zamir, Alexander Sax, William Shen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3712–3722, June 2018. doi: 10.1109/CVPR.2018.00391. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Zamir_Taskonomy_Disentangling_Task_CVPR_2018_paper.html.

Siqi Zeng, Christopher Jung, Rui Li, Zhe Kang, Ming Li, Nima Noorshams, Zhigang Wang, Fuchun Peng, Han Zhao, and Xue Feng. Convex dataset valuation for post-training. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=oAAL0wZSMF.

Qi Zhang, Yifei Wang, Xiaohan Wang, Jiajun Chai, Guojun Yin, Wei Lin, and Yisen Wang. SAE as a crystal ball: Interpretable features predict cross-domain transferability of LLMs without training. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 35869–35888, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/3ca0956ae9d29ab4e009cf9418561bd0-Abstract-Conference.html.

APPENDIX

A SOURCE SELECTION FROM PARTIALLY EVALUATED LIBRARIES

We extend source-quality selection to partially evaluated libraries. Bounds on complete-library source quality guide additional evaluations and provide a stopping criterion. We consider fixed-budget selection and selection with a prescribed library-quality regret tolerance.

A.1 BOUNDS AND ACQUISITION

Let $\mathcal{L}$ contain $n \geq 3$ tasks and Ω denote the observed off-diagonal transfer pairs. We treat recorded scores as fixed, with uncertainty arising from unevaluated pairs. For target j, define $O_j = \{i : (i,j) \in \Omega\}$ and $d_j = |O_j|$. Complete-library source quality is $s_i = (n-1)^{-1} \sum_{j \neq i} Q_{ij}$, where Q_{ij} ranks source i among all $n-1$ candidates for target j.

For an observed pair (i,j), let

$$w_{ij} = \sum_{k \in O_j \setminus \{i\}} \left[\mathbf{1}\{A_{ij} > A_{kj}\} + \frac{1}{2}\mathbf{1}\{A_{ij} = A_{kj}\} \right].$$

Each unobserved competitor contributes between zero and one additional win. Define

$$(\ell_{ij}, u_{ij}) = \begin{cases} \left(\dfrac{w_{ij}}{n-2}, \dfrac{w_{ij} + n - 1 - d_j}{n-2} \right), & i \in O_j, \\ (0,1), & i \notin O_j, \end{cases}$$

and

$$L_i = \frac{1}{n-1}\sum_{j\neq i}\ell_{ij}, \qquad U_i = \frac{1}{n-1}\sum_{j\neq i}u_{ij}.$$

These intervals bound complete-library percentiles rather than ranks recomputed among observed sources.

Proposition A.1 (Source-quality certification)**.** For any observation set Ω, every completion of its unobserved scores satisfies $L_i \le s_i \le U_i$. For any selected source $\widehat{i}$,

$$\max_i s_i - s_{\widehat{i}} \le C(\widehat{i}), \qquad C(\widehat{i}) = \left[\max_{i\neq\widehat{i}} U_i - L_{\widehat{i}}\right]_+ .$$

The bound holds throughout adaptive acquisition. Thus, $C(\widehat{i}) \le \epsilon$ certifies complete-library source-quality regret at most ϵ.

Fixed-budget selection uses $\widetilde{s}_i = (L_i + U_i)/2$. Its error is at most $(U_i - L_i)/2$, so the ordering of two sources is preserved when their quality difference exceeds half the sum of their interval widths. Certified selection instead uses the lower-bound leader $b \in \arg\max_i L_i$ and stops when $C(b) \le \epsilon$.

We initialize with d matching rounds. Each round evaluates every source and target once, without self-pairs or repeated evaluations. This coverage controls the initial interval widths.

Proposition A.2 (Coverage and acquisition)**.** Suppose each target has d observed sources, where $1 \le d \le n-1$, and source i is observed on m_i targets. Then

$$U_i - L_i = 1 - \frac{m_i(d-1)}{(n-1)(n-2)}.$$

Among patterns with d evaluations per target, source-balanced coverage, $m_i = d$ for every i, minimizes the largest interval width.

For arbitrary Ω and fixed nonnegative weights ω, let $W_\omega(\Omega) = \sum_k \omega_k(U_k - L_k)$. Evaluating an unobserved pair (i,j) gives

$$W_\omega(\Omega) - W_\omega\big(\Omega \cup \{(i,j)\}\big) = \frac{d_j\omega_i + \sum_{k\in O_j}\omega_k}{(n-1)(n-2)},$$

independently of the newly observed score.

At each step, we set $\omega_i = \mathbf{1}\{i = b \text{ or } U_i \ge L_b\}$, retaining the leader and unresolved competitors. We then maximize $g_{ij} = d_j\omega_i + \sum_{k\in O_j}\omega_k$. This gives the largest one-step reduction in weighted interval width for the current weights. Bounds and weights are updated after each evaluation.

Fixed-budget runs use the initialization prefix when $B < nd$. Certified runs require $B \ge nd$ and report whether the returned bound meets ϵ. At zero tolerance, any source with maximum library quality is valid, including tied optima.

A.2 Experiments

We reuse the 50 library/incoming splits of the complete-library experiment for each model and $n \in \{7, 10, 13\}$. Acquisition accesses only queried internal scores; complete-library and incoming scores are used afterward for evaluation. Cost counts distinct internal evaluations relative to $n(n-1)$ and excludes source-adapter training.

Adaptive and random allocation share an ordered initialization with $d = 2$. We select its balanced mask from 32 outcome-independent candidates by minimizing the sum of squared source-pair co-occurrence counts. After initialization, random allocation samples remaining pairs uniformly without replacement. Both methods use identical estimators, certificates, budgets, and source-selection tie rules.

Fixed budgets. Table A.1 reports incoming selected-source percentiles for $B = \lfloor \beta n(n-1) \rfloor$, with $\beta \in \{0.5, 0.75\}$. Adaptive allocation exceeds random allocation in mean percentile in 10

Algorithm 2 Adaptive Evaluation of Source Quality

Require: Library $\mathcal{L}$ of $n \geq 3$ tasks; evaluation oracle; budget $n \leq B \leq n(n-1)$; initialization degree $1 \leq d \leq n-1$; tolerance $\epsilon \geq 0$; mode fixed or certified.

Ensure: Selected source $\widehat{i}$, observed pairs Ω, and regret bound $C(\widehat{i})$.

Construct an ordered initialization $\mathcal{I}$ of d edge-disjoint matching rounds.
$\Omega \leftarrow \varnothing$; $O_j \leftarrow \varnothing$, $d_j \leftarrow 0$ for every j.
$L_i \leftarrow 0$, $U_i \leftarrow 1$ for every i.
while $|\Omega| < B$ **do**
  $b \leftarrow \arg\max_i L_i$.
  if mode is certified, $|\Omega| \geq nd$, and $C(b) \leq \epsilon$ **then**
    break
  end if
  if $|\Omega| < nd$ **then**
    $(i, j) \leftarrow \mathcal{I}_{|\Omega|+1}$.
  else
    $\omega_k \leftarrow \mathbf{1}\{k = b \text{ or } U_k \geq L_b\}$ for every k.
    $(i, j) \leftarrow \arg\max_{i \neq j,\, (i,j) \notin \Omega} \left(d_j \omega_i + \sum_{k \in O_j} \omega_k \right)$.
  end if
  Evaluate A_{ij}; add (i, j) to Ω; update O_j, d_j, L, U.
end while
if mode is fixed **then**
  $\widehat{i} \leftarrow \arg\max_i (L_i + U_i)/2$.
else
  $\widehat{i} \leftarrow \arg\max_i L_i$.
end if
return $\widehat{i}, \Omega, C(\widehat{i})$.

Model	n	Complete	50% budget		75% budget	
			Random	Adaptive	Random	Adaptive
T5	7	0.746 ± 0.043	0.611 ± 0.033	$\mathbf{0.638} \pm 0.032$	0.662 ± 0.030	$\mathbf{0.701} \pm 0.037$
	10	0.756 ± 0.043	$\mathbf{0.717} \pm 0.045$	0.715 ± 0.040	$\mathbf{0.734} \pm 0.043$	0.728 ± 0.043
	13	0.744 ± 0.045	$\mathbf{0.762} \pm 0.039$	0.726 ± 0.044	$\mathbf{0.769} \pm 0.038$	0.767 ± 0.038
LLaMA	7	0.840 ± 0.024	0.761 ± 0.017	$\mathbf{0.787} \pm 0.018$	0.796 ± 0.023	$\mathbf{0.819} \pm 0.026$
	10	0.859 ± 0.021	0.790 ± 0.020	$\mathbf{0.834} \pm 0.026$	0.838 ± 0.025	$\mathbf{0.858} \pm 0.019$
	13	0.839 ± 0.024	$\mathbf{0.840} \pm 0.023$	$\mathbf{0.840} \pm 0.027$	0.836 ± 0.022	$\mathbf{0.839} \pm 0.021$
Qwen3-1.7B	7	0.852 ± 0.014	0.779 ± 0.014	$\mathbf{0.829} \pm 0.015$	0.834 ± 0.015	$\mathbf{0.852} \pm 0.014$
	10	0.890 ± 0.022	0.825 ± 0.016	$\mathbf{0.859} \pm 0.013$	0.857 ± 0.019	$\mathbf{0.874} \pm 0.019$
	13	0.901 ± 0.028	0.853 ± 0.018	$\mathbf{0.867} \pm 0.020$	0.840 ± 0.024	$\mathbf{0.888} \pm 0.025$
Qwen3-8B	7	0.919 ± 0.011	0.801 ± 0.008	$\mathbf{0.858} \pm 0.012$	0.891 ± 0.010	$\mathbf{0.905} \pm 0.010$
	10	0.967 ± 0.009	0.928 ± 0.010	$\mathbf{0.946} \pm 0.012$	0.936 ± 0.008	$\mathbf{0.962} \pm 0.009$
	13	0.976 ± 0.010	0.929 ± 0.011	$\mathbf{0.933} \pm 0.012$	0.951 ± 0.010	$\mathbf{0.967} \pm 0.011$

Table A.1: Incoming selected-source percentile: target-balanced mean $\pm$ task-clustered SE. Bold marks the larger reported mean within each budget, including ties. Complete uses all internal evaluations; budgets are rounded down.

of 12 settings at each budget. The equally weighted mean improvements are 0.0197 and 0.0179, respectively. For Qwen3-8B with ten library tasks, it achieves 0.946 with 45 evaluations and 0.962 with 67, compared with 0.967 using all 90 evaluations. Even at the 50% budget, adaptive selection achieves higher mean selected-source percentiles than both Grad Dot and Grad Cos in all nine decoder-only settings (Table 2).

Complete-library selection uses average source quality rather than an oracle for each incoming target. Partial estimates can therefore exceed its incoming performance, as observed for T5 at library size 13. The certification experiment below instead measures recovery of complete-library source quality.

Model	n	Random	Adaptive
T5	7	94.00 ± 0.53	$\mathbf{90.52} \pm 0.65$
	10	93.82 ± 0.43	$\mathbf{90.44} \pm 0.55$
	13	95.41 ± 0.37	$\mathbf{92.21} \pm 0.49$
LLaMA	7	91.19 ± 0.65	$\mathbf{87.14} \pm 0.77$
	10	93.16 ± 0.61	$\mathbf{88.42} \pm 0.78$
	13	94.29 ± 0.38	$\mathbf{90.36} \pm 0.49$
Qwen3-1.7B	7	92.76 ± 0.67	$\mathbf{87.67} \pm 0.89$
	10	94.02 ± 0.50	$\mathbf{90.84} \pm 0.52$
	13	93.85 ± 0.39	$\mathbf{90.50} \pm 0.44$
Qwen3-8B	7	90.33 ± 0.80	$\mathbf{84.00} \pm 1.01$
	10	91.13 ± 0.66	$\mathbf{86.71} \pm 0.70$
	13	92.44 ± 0.48	$\mathbf{87.19} \pm 0.57$

Table A.2: Percentage of internal evaluations required to certify library-quality regret at most 0.05. Values are means $\pm$ descriptive SEs over 50 matched, overlapping splits; lower is better.

Certified stopping. Table A.2 reports the evaluations required to reach $C(b) \leq 0.05$. Every run certifies. Adaptive allocation has lower mean cost than random allocation in every model/library setting, with paired reductions ranging from 3.18 to 6.33 percentage points of the complete matrix.

Adaptive allocation also has lower mean certification cost in all 12 settings at tolerances 0, 0.025, and 0.10. Averaged equally across settings, it saves 8.30%, 9.37%, 11.17%, and 14.54% of complete evaluation at tolerances 0, 0.025, 0.05, and 0.10, respectively; random allocation saves 4.87%, 5.63%, 6.97%, and 9.65%. At zero tolerance, every selection attains maximum library quality. At tolerance 0.05, adaptive selection has mean observed library-quality regret 0.00150. Outcome-dependent acquisition therefore reduces the evaluation cost of reaching the same source-quality certificate, while fixed-budget selection retains strong incoming-target performance with fewer observations.

B Source Quality in the Full Transfer Matrix

We complement the asymmetry decomposition by measuring how much variation in the full transfer matrix is explained by average source quality. For each source i, we predict its target-conditioned rank percentiles using

$$s_i = \frac{1}{N-1} \sum_{j \neq i} Q_{ij}, \qquad \widehat{Q}_{ij} = s_i.$$

The source-explained share is

$$R^2_{\mathrm{src}} = 1 - \frac{\sum_{i \neq j} (Q_{ij} - s_i)^2}{\sum_{i \neq j} (Q_{ij} - \frac{1}{2})^2},$$

where $\frac{1}{2}$ is the mean within-target rank percentile.

Model	Source-explained share	Residual share
T5	30.1%	69.9%
LLaMA	50.6%	49.4%
Qwen3-1.7B	65.6%	34.4%
Qwen3-8B	65.7%	34.3%

Table A.3: Descriptive source-only decomposition of off-diagonal rank-percentile variation in the full transfer matrix.

Average source quality explains approximately two-thirds of transfer-ranking variation in both Qwen models, about half in LLaMA, and 30.1% in T5. Source-general structure therefore extends beyond transfer asymmetry to the full transfer rankings, with its contribution varying across models.

C THEORY

Proof of Proposition 3.1. Average ranks give $\sum_{j\neq i} Q_{ji} = (N-1)/2$ for every target i. Hence

$$\frac{1}{N}\sum_{j\neq i}\Delta_{ij} = \frac{1}{N}\left(\sum_{j\neq i} Q_{ij} - \frac{N-1}{2}\right) = \frac{N-1}{N}\left(s_i - \frac{1}{2}\right).$$

Define a by this expression. Antisymmetry of Δ implies $\sum_i a_i = 0$, and therefore

$$\sum_{j\neq i} r_{ij} = \sum_{j\neq i}\Delta_{ij} - \sum_{j\neq i}(a_i - a_j) = Na_i - Na_i = 0.$$

For any vector c, antisymmetry of r yields

$$\langle r, G(c)\rangle_{\text{edge}} = \sum_{i<j} r_{ij}(c_i - c_j) = \sum_i c_i \sum_{j\neq i} r_{ij} = 0.$$

Moreover,

$$\|G(c)\|^2_{\text{edge}} = \sum_{i<j}(c_i - c_j)^2 = N\|c\|_2^2 - \left(\sum_i c_i\right)^2.$$

For centered b, the vector $a - b$ is centered. Using $\Delta - G(b) = r + G(a-b)$ and orthogonality,

$$\|\Delta - G(b)\|^2_{\text{edge}} = \|r\|^2_{\text{edge}} + \|G(a-b)\|^2_{\text{edge}} = \|r\|^2_{\text{edge}} + N\|a-b\|_2^2.$$

The minimum is attained uniquely at $b = a$. □

Proof of Corollary 5.1. By Proposition 3.1, both residuals are orthogonal to every source-difference matrix. Since $a^M - a^P$ is centered,

$$\binom{N}{2} L_{\text{src}} = N\|a^M - a^P\|_2^2 + \|r^M\|^2_{\text{edge}},$$
$$\binom{N}{2} L_{\text{full}} = N\|a^M - a^P\|_2^2 + \|r^M - r^P\|^2_{\text{edge}}.$$

Subtracting gives

$$\begin{aligned}\binom{N}{2}\left(L_{\text{src}} - L_{\text{full}}\right) &= \|r^M\|^2_{\text{edge}} - \|r^M - r^P\|^2_{\text{edge}} \\ &= 2\langle r^M, r^P\rangle_{\text{edge}} - \|r^P\|^2_{\text{edge}}.\end{aligned}$$

The strict improvement criterion follows immediately. □

Proof of Proposition A.1. Average ranks give the complete-percentile representation

$$Q_{ij} = \frac{1}{n-2}\sum_{k\in\mathcal{L}\setminus\{i,j\}}\left[\mathbf{1}\{A_{ij} > A_{kj}\} + \frac{1}{2}\mathbf{1}\{A_{ij} = A_{kj}\}\right].$$

For an observed pair (i,j), the known terms sum to w_{ij} and the remaining $n-1-d_j$ terms lie in $[0,1]$. Hence $\ell_{ij} \le Q_{ij} \le u_{ij}$. For an unobserved pair, $Q_{ij} \in [0,1]$. Averaging over targets yields $L_i \le s_i \le U_i$.

For any selected source $\widehat{i}$,

$$\begin{aligned}\max_i s_i - s_{\widehat{i}} &= \max\left\{0, \max_{i\neq\widehat{i}}(s_i - s_{\widehat{i}})\right\} \\ &\le \max\left\{0, \max_{i\neq\widehat{i}}(U_i - L_{\widehat{i}})\right\} \\ &= C(\widehat{i}).\end{aligned}$$

These inequalities hold for every observation set and every completion consistent with its scores. They therefore remain valid when observations are chosen adaptively and when acquisition stops.

For the midpoint estimate $\widetilde{s}_i = (L_i + U_i)/2$, write $h_i = U_i - L_i$. Interval containment gives $|\widetilde{s}_i - s_i| \le h_i/2$, and consequently

$$\widetilde{s}_i - \widetilde{s}_k \ge s_i - s_k - \frac{h_i + h_k}{2}.$$

Thus, $s_i - s_k > (h_i + h_k)/2$ ensures $\widetilde{s}_i > \widetilde{s}_k$. □

Proof of Proposition A.2. Suppose each target has d observed sources. An observed percentile interval has width $(n-1-d)/(n-2)$, whereas an unobserved interval has width 1. Source i is observed on m_i targets, so

$$\begin{aligned} U_i - L_i &= \frac{1}{n-1}\left[m_i \frac{n-1-d}{n-2} + n - 1 - m_i\right] \\ &= 1 - \frac{m_i(d-1)}{(n-1)(n-2)}. \end{aligned}$$

Since $\sum_i m_i = nd$, at least one source satisfies $m_i \le d$. Therefore,

$$\max_i (U_i - L_i) \ge 1 - \frac{d(d-1)}{(n-1)(n-2)}.$$

A source-balanced pattern attains equality. Such a pattern exists by taking d distinct nonzero cyclic shifts of the task indices, establishing the minimax-width claim.

Now consider an unobserved pair (i, j) under arbitrary Ω. Evaluating it changes its percentile interval width from 1 to $(n-2-d_j)/(n-2)$, a reduction of $d_j/(n-2)$. For each previously observed source $k \in O_j$, one unknown comparison becomes known, reducing its interval width by $1/(n-2)$. All other percentile interval widths remain unchanged. Averaging across targets and applying the fixed weights yields

$$W_\omega(\Omega) - W_\omega\big(\Omega \cup \{(i,j)\}\big) = \frac{d_j \omega_i + \sum_{k \in O_j} \omega_k}{(n-1)(n-2)}.$$

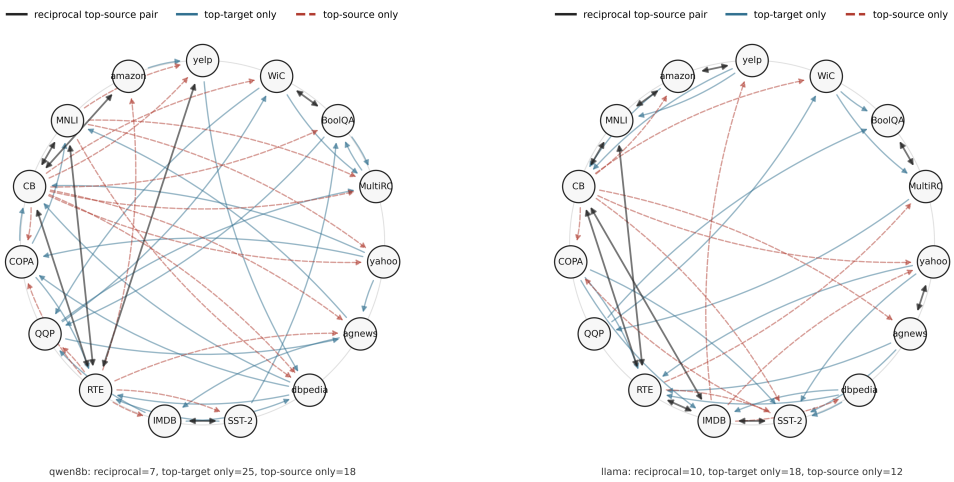
The reduction depends only on the observation pattern and weights, not on the newly observed score. □

D Experimental Setup and Reproducibility

We use the 15-task benchmark adopted by Wang et al. (2023): Yelp, Amazon, DBpedia, Yahoo Answers, AG News, MNLI, QQP, RTE, SST-2, WiC, CB, COPA, MultiRC, BoolQ, and IMDB. Each source-adapter training run starts from the corresponding base checkpoint and uses its source-task configuration. The resulting adapter is evaluated on all other tasks using a shared evaluation configuration, producing 210 directed transfer scores per model.

All launch scripts specify one epoch, learning rate 10^{-3}, a constant schedule, and zero warmup steps. T5-large uses a per-device training batch size of 8 with 4 gradient-accumulation steps; LLaMA and both Qwen models use batch size 1 with 8 accumulation steps. With G data-parallel workers, the nominal effective batch sizes are $32G$ and $8G$, respectively. Maximum source and target lengths are 512 and 50. Instructions use the `single` strategy, with task and dataset names enabled. Both regularization coefficients, `lamda_1` and `lamda_2`, are zero. Training uses model-specific DeepSpeed configurations, and both Qwen scripts explicitly enable BF16.

Evaluation enables generation with `generation_max_length=50`. The per-device evaluation batch size defaults to 128 for T5 and 4 for the other models. Validation during training is disabled, and all-task evaluation loads the final saved adapter. Intermediate checkpointing is disabled for T5, LLaMA, and Qwen3-1.7B. The Qwen3-8B script defaults to saving recovery checkpoints every 100 steps and retaining at most two.



(a) Qwen3-8B (b) LLaMA

Figure A.1: Directed top-3 transfer neighborhoods. An edge from source i to target j indicates that i is among the three highest-ranked sources for j; double arrows indicate reciprocal membership.

Library-based experiments use 50 saved splits for each library size 7, 10, and 13. Sparse acquisition reuses these assignments across all strategies. Source ranks use average ranks for ties, and incoming transfer outcomes are reserved for evaluation. Repeated selection results are averaged within each incoming task before computing target-balanced means and standard errors. Sparse acquisition settings and certification guarantees are given in Appendix A.

E More Experiments

Figure A.1 visualizes the directed transfer neighborhoods underlying our asymmetry analysis. One-way edges show that directional relationships affect membership among the highest-ranked sources, while reciprocal edges identify mutually useful task pairs. Figure A.2 measures whether the pilot correctly orders pairs of candidate sources for each target. Figures A.3 and A.4 evaluate the stricter selection problem, reporting exact recovery of the best common source and coverage of the three best common sources, respectively.

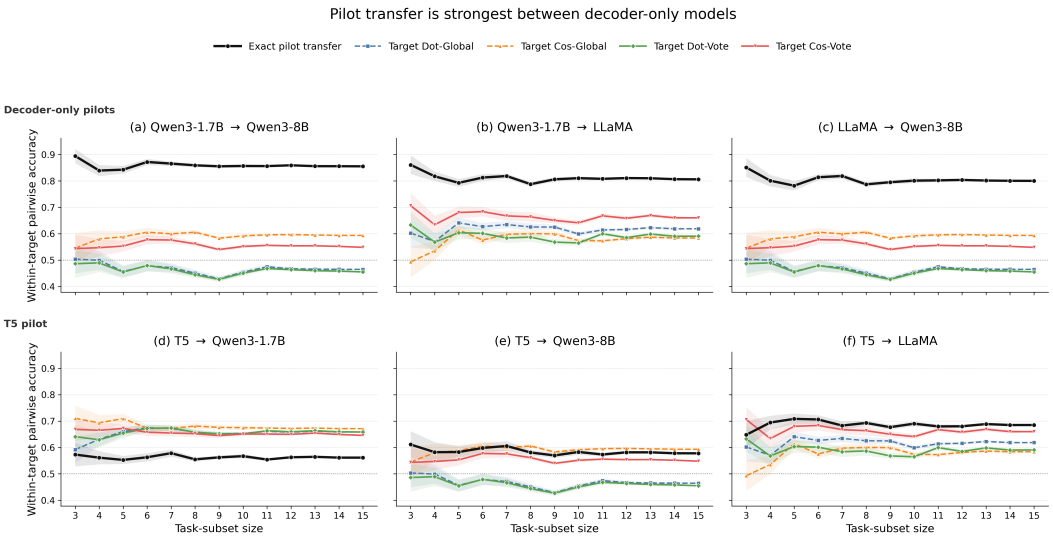


Figure A.2: Within-target pairwise ranking accuracy across model-family pairs and task-subset sizes.

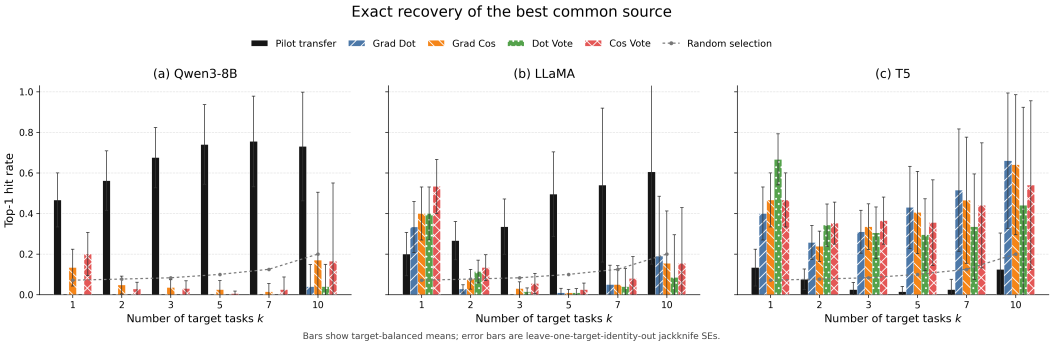


Figure A.3: Exact recovery of the best common source for k target tasks. Top-1 hit rate is the fraction of target sets for which a method selects an oracle-optimal source. The dotted line denotes random selection.

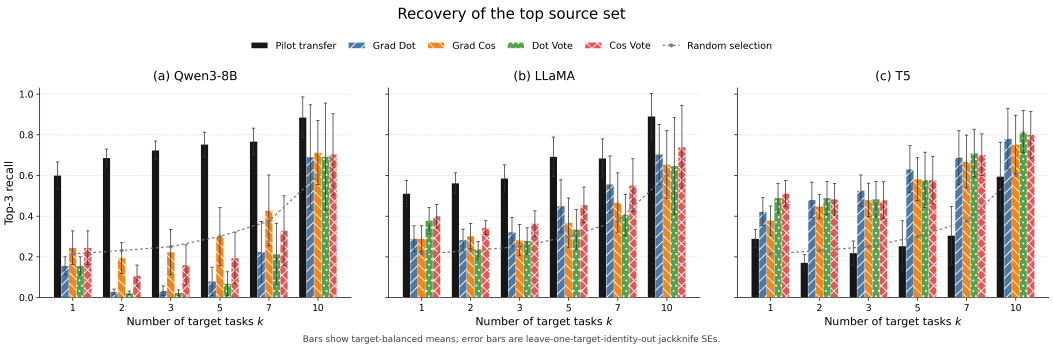


Figure A.4: Recovery of the three best common sources for k target tasks. Top-3 recall measures the overlap between the predicted and oracle top-three source sets. The dotted line denotes random selection.