Donors and Recipients: On Asymmetric Transfer Across Tasks and Languages with Parameter-Efficient Fine-Tuning

Anonymous ACL submission

Abstract

Large language models (LLMs) perform strongly across tasks and languages, yet how improvements in one task or language affect other tasks and languages and their combinations remains poorly understood. We conduct a controlled PEFT/LoRA study across multiple open-weight LLM families and sizes, treating task and language as transfer axes while conditioning on model family and size; we fine-tune each model on a single task–language *source* and measure transfer as the percentage-point change versus its baseline score when evaluated on all other task–language *target* pairs. We decompose transfer into (i) Matched-Task (Cross-Language), (ii) Matched-Language (Cross-Task), and (iii) Cross-Task (Cross-Language) regimes. We uncover two consistent general patterns. First, a pronounced on-task vs. off-task asymmetry: Matched-Task (Cross-Language) transfer is reliably positive, whereas off-task transfer often incurs collateral degradation. Second, a stable *donor–recipient* structure across languages and tasks (hub donors vs. brittle recipients). We outline implications for risk-aware fine-tuning and model specialisation.

1 Introduction

Large Language Models (LLMs) have become a cornerstone of modern AI, exhibiting impressive capabilities across a wide range of tasks (Achiam et al., 2023; Brown et al., 2020). In parallel, parameter-efficient fine-tuning (PEFT) methods such as Low-Rank Adaptation (LoRA) have become a widely used way to adapt and specialize pre-trained models (Mangrulkar et al., 2022; Hu et al., 2022a). Yet, it remains unclear and not systematically studied how improving a model on a specific task, capability, or language via PEFT methods, serving as a practical standard, propagates to other settings. Prior work documents sizeable cross-lingual performance variation even for strong multilingual models (Hu et al., 2020). Research on multi-task and instruction-tuning shows that they can induce negative transfer, where optimizing for some tasks reduces performance elsewhere (Mueller et al., 2024). In sequential (continual) settings, new fine-tuning can also cause catastrophic forgetting, overwriting previously acquired competencies (Goodfellow et al., 2015). While negative transfer and catastrophic forgetting are well documented in multi-task and sequential regimes—and some studies touch both task and language axes—they do not isolate single-source (one task–language pair) effects or provide a controlled, comparable map across model families and sizes spanning the full task–language grid. Most studies examine either cross-lingual transfer within a fixed task or cross-task transfer within a fixed language. This motivates the need for predictive, risk-aware guidance about adaptation side effects—i.e., when fine-tuning on one task or language will improve, leave unchanged, or harm other capabilities.

We address this gap by uncovering a *donor–recipient* structure that supports source selection for a given target task–language combination while limiting performance degradation on other tasks and languages. We present evidence of a pronounced transfer asymmetry: Matched-Task (Cross-Language) transfer tends to be positive, whereas off-task transfer carries elevated risk of regressions. In order to systematically investigate these transfer dynamics, we define our experimental space along four primary dimensions: task, language, model family, and model size. We construct a comprehensive evaluation space composed of numerous task–language pairings. For each of these individual settings, we fine-tune models from two prominent open-weight LLM families at various scales. This results in a multi-dimensional study where every task–language combination is trained and evaluated for every model, allowing us to decompose transfer into (i) Matched-Task (Cross-Language), (ii) Matched-Language (Cross-Task),

and (iii) Cross-Task (Cross-Language) regimes. Our study is designed to understand the patterns of this transfer in terms of its impact (the direction and magnitude of performance changes) and the stability of these changes. Specifically, we seek to answer the following research questions:

- **Impact:** When a model is fine-tuned on a specific task–language pair, what are the consequences for its performance on other tasks and in other languages? Are there specific tasks that benefit others? Conversely, are there specific tasks or languages causing harm to others when used for fine-tuning?

- **Stability of transfer patterns:** How stable are transfer patterns across specific dimensions? How much of the variation in transfer is driven by the model we use (family and size) versus by properties of the source and target (which language or which task)? Does that balance shift across the different settings?

Contributions. (i) We quantify a pronounced transfer asymmetry—Matched-Task (Cross-Language) transfer is reliably positive, whereas off-task transfer is typically negative with substantially higher harm rates—establishing the core on-task vs. collateral trade-off; (ii) we reveal a stable donor–recipient structure across languages and task types, and formalise its predictability via mixed-effects variance decomposition and a Consistency Index; (iii) we translate these patterns into practical heuristics for risk-aware fine-tuning (see §6), including when to prefer matched-task sources and when to use multi-source or regularised updates to mitigate collateral harm.

2 Related Work

Research on cross-lingual transfer in LLMs has accelerated, with a growing emphasis on how model and data design choices—such as pre-training language selection and fine-tuning setup—affect both the strength and the stability of transfer across languages. Malkin et al. (2022) show that pre-training languages can act as donors or recipients in zero-shot transfer, while (Chua et al., 2025) identify a cross-lingual knowledge barrier: models often handle translation/alignment, yet fail to carry over task-specific knowledge without multilingual fine-tuning. Work on zero-shot instruction tuning (e.g. (Chirkova and Nikoulina, 2024)) shows that English-only instruction-tuned models generalise cross-lingually but with degraded factuality. Complementary findings suggest that the factual accuracy itself can vary depending on the language of inquiry (Aggarwal et al., 2025), underscoring that cross-lingual generalisation and factual reliability do not always go hand in hand. Another line of work develops multilingual, language-parallel benchmarks. The BenchMAX (Huang et al., 2025) offers a comprehensive, post-edited, multiway parallel benchmark that spans 10 tasks across 17 languages and reports persistent cross-language gaps that scaling alone does not close. Beyond evaluation, Okapi (Lai et al., 2023b) contributes a multilingual framework (instruction data, preference data, translated evals) for 26 languages; results highlight that preference-based instruction tuning yields larger multilingual gains than supervised fine-tuning and that behaviour varies substantially by language. Coverage and culture also matter: (Joshi et al., 2020) document stark under-representation of most world languages, and (Choenni et al., 2024) show that multilingual fine-tuning can transfer cultural values across languages. Beyond cross-linguality, cross-task transfer reveals important trade-offs. Single-task fine-tuning can rapidly erode generality via format specialisation, while multitask instruction tuning often yields positive transfer to unseen tasks (Wei et al., 2022). Parameter-efficient methods such as LoRA aim to add capabilities with less interference (Hu et al., 2022b). Finally, interactions among capability dimensions (e.g., reasoning, coding, factuality, fairness) are not trivial: code-heavy pretraining is frequently linked to improved reasoning, yet optimising solely for accuracy can degrade fairness (Zhang et al., 2024). Our study builds on these insights and introduce a controlled, cross-model framework that isolates single-source PEFT and maps task–language transfer, revealing a strong on- vs off-task asymmetry and stable donor–recipient roles.

3 Methodology

Our methodology is designed to systematically evaluate cross-dimensional transfer in a controlled and reproducible manner. We follow a three-stage process for each model: (1) we establish its baseline performance through zero-shot evaluation on a suite of multilingual benchmarks; (2) we fine-tune the model on a specific task in a single language;

Table 1: Datasets and benchmarks grouped by evaluated property.

Property	Benchmarks
Common-sense reasoning	XWinograd[a], HellaSwag[b]
Knowledge-intensive reasoning	ARC-Challenge[c], Global MMLU[d]
Factuality	TruthfulQA[e]
Mathematical reasoning	MGSM[f]
Coding	HumanEval[g]
Fairness & bias	MBBQ[h]

[a] Muennighoff et al. (2023); based on (Tikhonov and Ryabinin, 2021). [b] Zellers et al. (2019); multilingual translation via (Lai et al., 2023a). [c] Clark et al. (2018); multilingual translation via (Lai et al., 2023a). [d] Singh et al. (2024). [e] Lin et al. (2022) multilingual translation via (Lai et al., 2023a). [f] Shi et al. (2023); based on GSM8K (Cobbe et al., 2021). [g] Huang et al. (2025); based on (Liu et al., 2023). [h] Neplenbroek et al. (2024); based on (Parrish et al., 2022).

and (3) we conduct a comprehensive cross-lingual and cross-task evaluation of the fine-tuned model to measure the impact of the targeted adaptation (LoRA fine-tuning on exactly one dataset–language source; see Appendix. A.1 for details).

3.1 Models

Our selection includes models from two model families-Llama 3 (Dubey et al., 2024) and Qwen 2.5 (Qwen et al., 2025) - allowing us to analyse the stability of cross-dimensional transfer patterns across different architectural designs and sizes. Specifically, we fine-tune and evaluate seven instruction-tuned open-weight models: Llama 3 (3.2 1B, 3.2 3B, 3.1 8B) and Qwen 2.5 (0.5B, 1.5B, 3B, 7B).

3.2 Benchmarks & Languages

We evaluate seven public benchmarks spanning knowledge-intensive QA, commonsense inference, factuality, coreference, math reasoning, code generation, and social bias. Table 1 groups benchmarks by evaluated property and provides references. Detailed per-dataset entries are consolidated in the Appendix (Table 14).

3.3 Experimental Setup

Our experimental procedure is divided into three stages: Baseline Evaluation, Fine-tuning, and Cross-Evaluation.

3.3.1 Baseline Evaluation

First, we evaluate the performance of each original, pre-trained model on all languages for every benchmark.[1] The results from this stage serve as a baseline, representing the model's out-of-the-box multilingual capabilities before any task-specific fine-tuning.

3.3.2 Fine-tuning and Cross-Evaluation

We study transfer along four dimensions: model family, model size, task, and language. For each combination of these dimensions, we fine-tune the model on a single source task–language pair and then evaluate it across all target tasks and languages. **Fine-tuning:** We use PEFT via LoRA with a fixed configuration and fine-tune each model for 3 epochs on the training split of one benchmark in one language (see App. A.1 for full implementation details). **Transfer evaluation:** After fine-tuning on a source (e.g., `ARC-Challenge`, French), we evaluate zero-shot on all remaining target benchmarks and languages to quantify transfer over the task–language grid.

3.3.3 Evaluation Protocol and Metrics

For a base model m and its fine-tuned variant m_{ft} trained on a single dataset–language source (d, ℓ^*), our primary metric is the absolute percentage-point change $\Delta\% = 100\,[s(m_{\text{ft}}) - s(m)]$ (pp), where $s(\cdot) \in [0, 1]$ is the task score. We also report *win rate* (fraction of targets with $\Delta\% > 0$) and *harm rate* (fraction with $\Delta\% < -1.0$ pp). To isolate transfer, we exclude the trained cell (d, ℓ^*) and partition results into three *transfer regimes*: **Matched-Task, Cross-Language** (MT–CL; same dataset, other languages $\{(d, \ell) : \ell \neq \ell^*\}$); **Matched-Language, Cross-Task** (ML–CT; same language, other datasets $\{(d', \ell^*) : d' \neq d\}$); **Cross-Task, Cross-Language** (CT–CL; other datasets and other languages $\{(d', \ell) : d' \neq d,\ \ell \neq \ell^*\}$). We use these abbreviations henceforth; the same short forms are used in Tables.

4 Results and Analysis

We analyse cross-dimensional transfer along the two axes introduced in Section 1: impact (what is the direction and magnitude of fine-tuning effects across the evaluation matrix?) and stability (do patterns replicate across tasks, model families,

[1] All evaluations are run with the *LM Evaluation Harness* v0.4.9.1 (Gao et al., 2024). The harness provides a unified codebase with consistent prompting and decoding, metric computation, and task versioning. All prompts/decoding follow lm-evaluation harness defaults unless stated. Implementation specifics (tokenizer policy, padding/truncation, decoding) are detailed in Appendix. A.1.

Bucket	Mean Δ	Median Δ	Win %	Harm %
MT–CL	1.63	0.79	67.16	10.98
ML–CT	−1.57	−0.04	38.81	30.97
CT–CL	−1.23	−0.09	35.40	27.95

Table 2: Cross-lingual vs. cross-task (global). Δ is change in accuracy (percentage points). Harm if $\Delta < -1.0$ pp. Abbreviations: MT–CL = Matched-Task (Cross-Language), ML–CT = Matched-Language (Cross-Task), CT–CL = Cross-Task (Cross-Language). The same abbreviations are used in Tables 3–5.

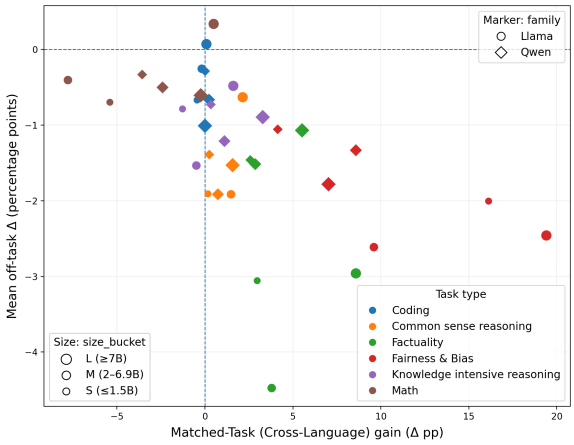

Figure 1: Matched-Task (Cross-Language) vs. off-task Pareto frontier. Each point is a fine-tuning run. The x-axis shows the gain (pp): mean $\Delta_{\%}$ on the trained dataset aggregated over all languages excluding the trained language (d, ℓ^*). The y-axis shows the mean off-task change (pp) across all other dataset–language pairs. Colours indicate source task type; marker shape encodes model family; marker size encodes size bucket **S/M/L** ($\leq$ 1.5B / 2–6.9B / $\geq$ 7B). Dashed lines mark zero gain/impact.

and scales?). We use the percentage-point change metric and the three transfer regimes defined in Sec. 3.3.3.

4.1 Impact: A Macro View of Transfer

We begin with the aggregate effect across all models, sources, and evaluation settings. By aggregating all runs, fine-tuning yields a mean change of $\mathbf{-0.75}$ pp with a win rate of $\mathbf{40.86}\%$. While many outcomes are neutral, the negative mean and sub-50% win rate underscore an asymmetry: degradations, when they occur, tend to be larger in magnitude than the typical gains.

4.1.1 Matched-Task (Cross-Language) vs. off-task trade-offs

Matched-Task (Cross-Language) transfer is reliably positive (mean gain $\approx +1.6$ pp) with a high win rate and low harm. By contrast, both off-task regimes are negative on average and markedly riskier (harm rate $\approx 30\%$) (Table 2). Building on this framing, for each fine-tuning run we compute (i) the Matched-Task (Cross-Language) gain $\Delta_{\text{on-task}}$ and (ii) the collateral impact as the mean off-task $\Delta_{\%}$ over all other dataset–language cells. We also bucket model sizes as S ($\leq$ 1.5B), M (2–6.9B), and L ($\geq$ 7B). Figure 1 situates each run in this trade-off space by plotting $\Delta_{\text{on-task}}$ versus collateral impact. Consistent with Table 2, the largest on-task gains tend to co-occur with more negative off-task effects, and visible clustering is driven primarily by source task type rather than model family or size. As shown in Table 6, high-return sources tend to be more destructive off-task. For example, Fairness & Bias and Factuality deliver large on-task improvements (+10.50 and +4.16 pp; on-task win rates 86.96% and 86.40%) but incur non-trivial collateral losses (mean off-task −1.74 and −2.21 pp; harm rate 35.69% and 34.29%). Common-sense reasoning yields modest on-task lift (+0.98 pp; 72.00% win rate) yet still imposes off-task costs (mean −1.52 pp; 36.30% harm rate). By contrast, "safer" sources trade return for stability: Knowledge-intensive reasoning is near-neutral on-task (+0.75 pp) with milder collateral (mean −0.91 pp; 24.24% harm rate), while coding and math are comparatively benign off-task (harm rate 16.19% and 18.07%) but do not meaningfully lift on-task performance (means −0.06 and −3.21 pp; coding on-task win rate 7.07%). Overall, Pareto "sweet spots" (positive on-task gain, near-zero collateral, low harm rate) remain rare under single-source LoRA. Within types, collateral varies by model family and size. For Factuality, Qwen-L shows lower collateral (−1.07 pp; 29.31% harm rate) than Llama-M (−4.48 pp; 48.79% harm rate). For Fairness & Bias, Qwen-S is less intrusive (−1.05 pp; 22.57% harm rate) than Llama-L (−2.46 pp; 46.70% harm rate). Coding (Llama-L) is especially benign off-task (collateral +0.07 pp; 12.80% harm rate).

4.1.2 Donor–Recipient Structure

We break transfer down into two complementary roles: donors, which export performance gains to others, and recipients, which absorb them. For a model fine-tuned on a dataset–language pair (d, ℓ^*):

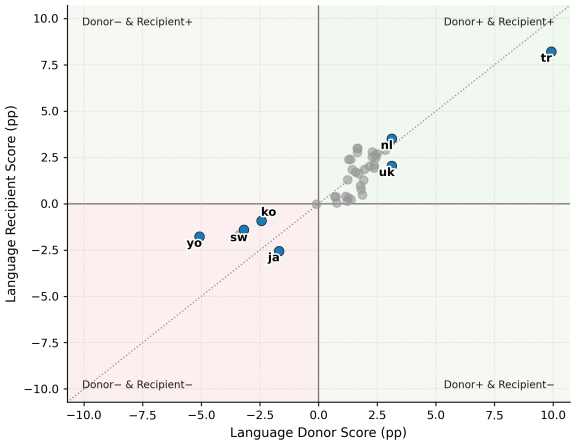

Figure 2: Language donor vs. recipient roles. Each code marks a language positioned by its Donor Score (x-axis) and Recipient Score (y-axis), computed within–task and cross-lingually while excluding the trained source cell. Quadrants (shaded): Donor+ & Recipient+ (green), Donor- & Recipient- (red), with intermediate tones for the mixed-sign quadrants. Highlighted languages: tr, nl, uk (strong donors and recipients) and yo, sw, ko, ja (weak donors and recipients). Scores are in percentage points (pp).

- The **Language Donor Score** of ℓ^* is the average Δ (percentage-point change) observed when evaluating the same dataset d in all other languages $\ell \neq \ell^*$.
- The **Language Recipient Score** of a target language t is the average incoming Δ that t receives from other source languages on the same dataset d.

Task-level Donor/Recipient scores are defined in the same way, but with languages held fixed instead of tasks. In all cases, we exclude the trained cell itself, average scores across models, and then aggregate across the full set of models. Cross-lingual transfer is dominated by a small set of hub languages (Figure 2; Table 12). Turkish is the clearest hub: as a source it yields the largest within-task gains for other languages and, as a target, it also benefits strongly from others. Dutch and Ukrainian follow as donors, with Dutch also ranking among the top recipients. On the recipient side, Spanish, Catalan, Vietnamese, Romanian, and Italian form a cluster of high-gain targets with moderate donor strength. Several additional languages—e.g., Swedish, Croatian, Danish and Indonesian—sit firmly in the positive–positive quadrant, reinforcing the hub structure. By contrast, Japanese, Korean, Swahili, Thai, and Yoruba are negative in both roles, consistently harming other languages as fine-tuning sources and showing limited benefit when receiving transfer. Overall, donor and recipient scores move together—no language flips sign—suggesting that cross-lingual compatibility is largely a property of the language–data locus rather than a one-directional effect.

Holding language fixed, task types split into two regimes (Table 7). Benign donors—math and coding—exert comparatively small negative spillover on other tasks. In contrast, intrusive donors—knowledge-intensive reasoning, common sense, factuality, and fairness/bias—produce progressively larger collateral drops. Recipient behaviour is asymmetric: coding is the only clear net beneficiary of incoming cross-type transfer; factuality and common sense are near-neutral slight positives, while fairness/bias and knowledge-intensive are net negatives. Math remains uniquely brittle. This suggests that fairness/factuality fine-tuning tends to buy on-task gains at the cost of broad in-language degradation, whereas math and coding are safer donors but rarely help other types (with coding the notable exception as a recipient).

4.1.3 Task–Task Transfer

Two task types—Coding and Math—stand out as relatively "benign" sources: their mean outgoing impact on other types is small. By contrast, fine-tuning on Factuality, Fairness & Bias, Knowledge-intensive reasoning, or Common-sense reasoning typically depresses performance elsewhere, often by multiple percentage points. As targets, Coding is the most consistently receptive: nearly all donors provide small positive lift, including Fairness→Coding, Knowledge-intensive reasoning→Coding, Common-sense reasoning→Coding, Factuality→Coding, and Math→Coding. Factuality shows a slight positive as a recipient overall, with the largest incoming gain coming from Coding→Factuality. By contrast, Math is uniquely brittle: nearly every donor reduces Math, sometimes substantially. Knowledge-intensive reasoning also absorbs widespread negative transfer. These results highlight pronounced asymmetry: transfer is not symmetric across roles. A task can be a harmful donor yet a (near-)benign recipient (e.g., Factuality: Donor -2.61 pp vs. Recipient $+0.12$ pp), whereas Coding is both a relatively safe donor and a strong recipient (Donor -0.57 pp; Recipient $+0.83$ pp).

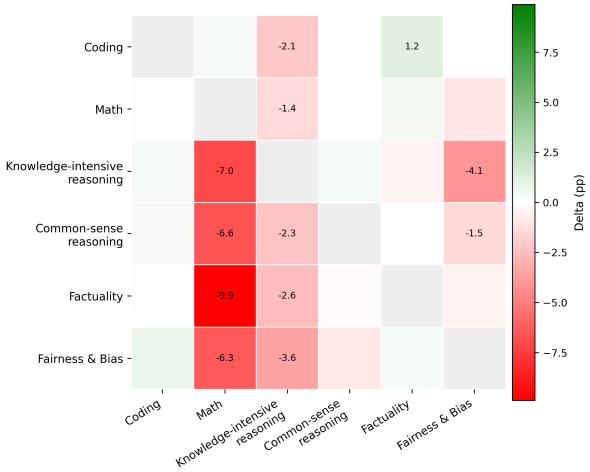

Figure 3: Task-to-task transfer heatmap. Cells show the mean percentage-point change when fine-tuning on the row (donor) task and evaluating the column (recipient) task; the diagonal is masked. Green denotes positive transfer and red denotes negative; numbers mark $|\Delta| \geq 1.0$ pp. See Appendix Table 13 for the full numeric matrix.

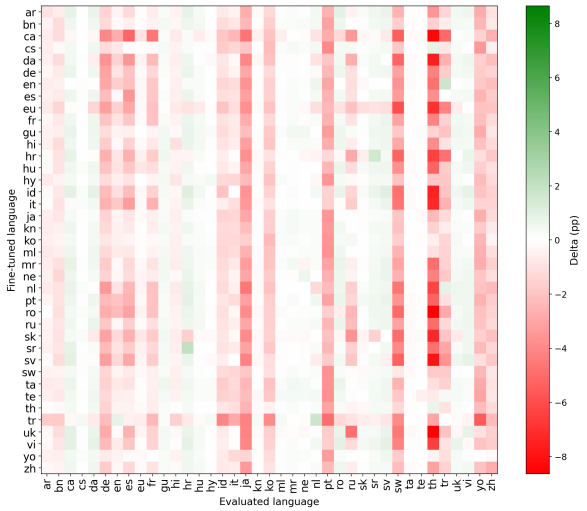

Figure 4: Mean $\Delta_{\%}$ (pp) by fine-tuned language (rows) and evaluated language (columns). Green indicates improvement, red indicates degradation. Per-pair values are listed in Appendix Table 15.

Setup	Intercept (95% CI)			Slope (95% CI)			R^2	Zero-X
MT–CL	0.08	−0.05	0.21	4.81	4.31	5.31	0.023	–
ML–CT	1.35	0.91	1.80	−8.15	−9.72	−6.58	0.046	0.17
CT–CL	0.85	0.80	0.91	−6.13	−6.35	−5.92	0.041	0.14

Table 3: Baseline-adjusted regressions. Ordinary least squares (OLS) with heteroskedasticity-consistent (HC1) standard errors; Δ (pp) $= \alpha + \beta\, x$, where x is the pre-fine-tuning baseline. "Zero-X" is $-\alpha/\beta$; "–" indicates it falls outside $[0, 1]$.

4.1.4 Language–Language Transfer

We quantify cross-lingual transfer with an $|\mathcal{L}| \times |\mathcal{L}|$ matrix whose rows are fine-tuned (source) languages ℓ^* and columns are evaluated (target) languages t. Each cell stores the mean percentage-point change $\Delta_{\%}$ when fine-tuning on ℓ^* and evaluating in t, averaged within model and task and then across models. Figure 4 visualises this matrix (green denotes gains, red denotes harms). Several targets consistently benefit across sources, with columns dominated by light to medium green. This includes Catalan (ca), Swedish (sv), Danish (da), Croatian (hr), and Gujarati (gu). These languages receive helpful transfer from a broad range of donors rather than relying on a single source. In contrast, some targets are consistently hurt across sources, showing widespread red with multiple large-magnitude negatives. Thai (th), Japanese (ja), and Swahili (sw) fall into this brittle group. The matrix is notably asymmetric: languages that receive well are not necessarily good donors. For example, sv and ca are strong recipients yet export substantial harm when used as fine-tuning sources. Turkish (tr) stands out as a solid receiver with many positive incoming edges, but as a donor its effects are mixed and often negative. The landscape features clear recipient hubs and brittle targets, with donor strength not mirroring recipient strength.

4.1.5 Baseline-adjusted transfer

To check whether transfer effects are driven by starting points, we regress Δ (percentage points) on the pre-fine-tuning baseline $x \in [0, 1]$ separately for each regime. For MT–CL, Δ increases with baseline ($\beta \approx +4.8$ pp per unit), with no clear intercept at $x{=}0$. For ML–CT and CT–CL, Δ decreases as baseline rises ($\beta \approx -6$ to -8), with small positive intercepts and a zero-crossing at $x \approx 0.14$–0.17 (Table 3). Overall, baseline explains little variance ($R^2 \approx 0.02$–0.05), so the observed transfer patterns are not artifacts of baseline differences.

4.2 Stability

4.2.1 Variance decomposition

To quantify the sources of variation in transfer effects, we fit a linear mixed model that partitions variance across model characteristics (family, size), source (fine-tuning dataset and language), and target (evaluated dataset and language). The resulting variance shares are reported in Table 4. In the

	Overall	MT–CL	ML–CT	CT–CL
Model	47.7	9.0	50.9	1.9
Source	20.3	29.0	8.3	4.0
Target	6.2	30.6	9.1	24.9
Residual	25.8	31.4	31.8	69.3

Table 4: REML variance shares (%) of Δ by model-, source-, and target-level components, overall and by transfer regime. Columns sum to ≈100% (rounding ±0.1).

	Dataset level	Task-type level
Regime b	**Median CI [IQR]**	**Median CI [IQR]**
MT–CL	0.235 [0.107–0.409]	0.195 [0.100–0.358]
ML–CT	0.263 [0.067–0.429]	0.284 [0.136–0.477]
CT–CL	0.194 [0.148–0.234]	0.209 [0.156–0.259]

Table 5: Consistency index (Kendall's τ_b) by transfer regime. Higher is more stable (1 = identical orderings, 0 = no stable ordering, -1 = reverse).

MT–CL regime, the model component is comparatively small (9.0%), whereas source and target account for larger, comparable shares (29.0% and 30.6%). This indicates that cross-lingual transfer regularities are organised primarily by donor data and the evaluation language rather than by model identity. For ML–CT, the model component dominates (50.9%), exceeding both source (8.3%) and target (9.1%), suggesting that off-task interference is strongly model-dependent and varies across architectures and scales. When both task and language change (CT–CL), the residual term is largest (69.3%) with a sizeable target contribution (24.9%) and minimal model/source shares (1.9%/4.0%), indicating limited predictable structure in this regime. Aggregated across regimes, the model accounts for the largest share (47.7%); however, the regime-wise decomposition clarifies where transfer is structured (Matched-Task, Cross-Language) versus noisy (Cross-Task, Cross-Language).

4.2.2 Consistency Index

For each source s (e.g., a specific task or language) and transfer regime $b \in \{$ Matched-Task (Cross-Language), Matched-Language (Cross-Task), Cross-Task (Cross-Language) $\}$, we collect, for every model m, the vector of transfer effects across recipients

$$\mathbf{\Delta}_m^{(s,b)} = \left(\Delta_{m,s\to r}\right)_{r\in\mathcal{R}_{s,b}}.$$

We define a per-source Consistency Index as the mean pairwise τ_b,

$$\mathrm{CI}(s,b) \;=\; \frac{\sum_{i<j} \tau_b\big(\mathbf{\Delta}_{m_i}^{(s,b)}, \mathbf{\Delta}_{m_j}^{(s,b)}\big)}{\binom{M}{2}},$$

so that $\mathrm{CI} \in [-1, 1]$, with larger values indicating more stable cross-model ordering. At the dataset level we summarise $\mathrm{CI}(s,b)$ across sources using the median and interquartile range; at the task-type level we aggregate recipients within task types and repeat the same computation.

The results reveal that models agree on recipient rankings, but only to a limited degree. At the dataset level (Table 5), the highest stability appears in the ML–CT regime (median $\mathrm{CI} = 0.26$, IQR 0.07–0.43), followed by MT–CL (0.24, 0.11–0.41) and CT–CL (0.19, 0.15–0.23). For intuition, these values correspond to roughly 60–63% pairwise concordance in the ordering of recipients. Aggregating by task type yields a similar picture: Matched-Language (Cross-Task) remains most stable (median $\mathrm{CI} = 0.28$), while Matched-Task (Cross-Language) and Cross-Task (Cross-Language) fall lower (0.20 and 0.21, respectively). Side-by-side summaries at both dataset and task-type levels are provided in Appendix Table 8. Factuality tasks show notably high stability in the Matched-Task (Cross-Language) regime, with CI values around 0.66–0.70 across multiple source languages. By contrast, Coding and Fairness & Bias display unstable orderings (CI ≈ -0.21 to -0.16), indicating that models often disagree on which recipients benefit most. In the Cross-Task (Cross-Language) regime, consistency is lowest: for several knowledge-intensive tasks in Indic and Dravidian languages, CI values approach zero, suggesting almost random variation across models. Top/bottom sources by consistency for each regime are listed in Appendix Tables 9, 10, and 11.

These findings reinforce the mixed-effects analysis: cross-task transfers within the same language are shaped by task-type structure, Matched-Task (Cross-Language) transfers show stable donor and recipient roles, and Cross-Task (Cross-Language) transfers remain dominated by target-specific noise.

5 Discussion

Our study reveals a transfer landscape that is strongly asymmetric and structured rather than uniformly positive. Matched-Task (Cross-Language) transfer is consistently helpful, while off-task transfer often imposes collateral damage. Aggregated across runs, single-source LoRA fine-tuning is slightly negative on average, with benefits concentrated in MT–CL settings and losses concentrated in ML–CT and CT–CL regimes. This asymmetry should be treated as a first-order constraint on fine-tuning strategies.

5.1 Donor–recipient structure is uneven

Our analysis reveals a highly uneven donor-recipient structure for both languages and tasks, a finding with direct implications for practitioners. The key takeaway is that transfer is not symmetric; the best source for a given target must be chosen carefully, as universal donors are rare.

For language transfer, a few hubs like Turkish and Dutch are effective at exporting within-task gains, while others like Japanese and Korean are consistently poor donors and recipients. Critically, strong recipients such as Catalan and Swedish are not necessarily safe sources, meaning source selection must be tailored to the specific target language.

For task transfer, the roles are similarly specialized. Coding and Math are safe donors with limited side effects, whereas Factuality or Fairness and Bias are intrusive donors that improve their own performance but harm others. As recipients, Coding is consistently receptive, while Math is brittle. This demonstrates that specializing a model on one task can create negative interference with other capabilities, a crucial consideration when designing for broad competence. This highlights a critical trade-off between achieving peak performance on a single metric and maintaining a model's generalist capabilities.

5.2 Stability: where regularities hold—and where they do not

Our stability analysis reveals that the predictability of transfer depends heavily on the transfer regime. We find that Matched-Task (Cross-Language) gains are not only larger on average but also more predictable across models. Conversely, outcomes in off-task or cross-task settings are less consistent and hinge more on specific model properties. Two key analyses support this conclusion. First, a mixed-effects variance decomposition shows that in Matched-Task (Cross-Language) scenarios, source properties dominate, indicating a stable phenomenon. In Matched-Language (Cross-Task) transfers, however, model factors like architecture and scale become the decisive drivers of interference. Second, rank-based consistency indices confirm this disparity. While there is moderate agreement on the donor-recipient relationships, some sources (e.g., Factuality) are highly stable across models, whereas others (e.g., Coding, Fairness and Bias) produce inconsistent recipient rankings.

6 Conclusion and Future Directions

We presented a controlled, model-spanning map of task–language transfer under single-source PEFT/LoRA across two open-weight LLM families and multiple scales. Our experiments isolate on-task vs. off-task effects and reveal a pronounced asymmetry: Matched-Task (Cross-Language) transfer is reliably positive, whereas off-task transfer tends to be negative with substantially higher harm rates. We further uncover a stable donor–recipient structure across languages and task types, quantify regime-dependent variance shares via mixed-effects models, and show moderate cross-model agreement using a Consistency Index. These results motivate risk-aware fine-tuning: choose sources for specific targets, avoid naive single-task specialization when broad competence matters, and consider multi-source or regularized schedules to balance on-task gains against collateral impact.

Future work should investigate the scalability of our findings by replicating these experiments on yet larger models and across a more diverse set of architectures. It would be particularly insightful to examine whether the observed patterns of variance-share and stability persist in models with mixture-of-experts, dense versus sparse routing mechanisms, or under varied instruction-tuning paradigms. In order to broaden the evaluation, future studies should incorporate a wider range of tasks, including generative, program-of-thought, and safety/robustness benchmarks. Additionally, to counteract potential translation artifacts and ensure more reliable cross-lingual comparisons, the integration of human-calibrated multilingual judgments is recommended.

Limitations

Our conclusions are bounded by (i) model coverage (two open-weight families at 0.5B–8B scales), (ii) the task suite and its multilingual construction (a mixture of translated, post-edited, or localised benchmarks), and (iii) a single adaptation regime (one-source PEFT with a fixed LoRA recipe and merged adapters). These choices could bias donor–recipient maps and harm rates; larger or different architectures, full fine-tuning or alternative adapter placements/merging, and richer prompting/decoding policies may shift magnitudes even if the qualitative asymmetries persist. Beyond scope, several design decisions may affect measured effects. First, many evaluations rely on translation or post-editing, which can introduce artefacts favouring certain typologies, scripts, or register; while this enables broad language coverage, cross-lingual differences may partly reflect translation choices rather than purely model transfer. Second, our evaluation protocol fixes decoding in a zero-shot setting with pinned seeds, which stabilises comparisons but may interact with transfer (few-shot, Chain of thought (CoT) prompting, or different decoding could change outcomes). Third, our primary metric is absolute percentage-point change with specific harm rate threshold; alternative thresholds or utility-weighted objectives could alter reported harm rates. Finally, we do not sweep hyper-parameters, we study single-source specialisation rather than multi-source or regularised schedules, and our benchmarks emphasise short-form QA and classification tasks.

References

OpenAI Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haim ing Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, and 260 others. 2023. Gpt-4 technical report.

Tushar Aggarwal, Kumar Tanmay, Ayush Agrawal, Kumar Ayush, Hamid Palangi, and Paul Pu Liang. 2025. Language models' factuality depends on the language of inquiry. *Preprint*, arXiv:2502.17955.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Nadezhda Chirkova and Vassilina Nikoulina. 2024. Zero-shot cross-lingual transfer in instruction tuning of large language models. *Preprint*, arXiv:2402.14778.

Rochelle Choenni, Anne Lauscher, and Ekaterina Shutova. 2024. The echoes of multilinguality: Tracing cultural value shifts during language model fine-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15042–15058, Bangkok, Thailand. Association for Computational Linguistics.

Lynn Chua, Badih Ghazi, Yangsibo Huang, Pritish Kamath, Ravi Kumar, Pasin Manurangsi, Amer Sinha, Chulin Xie, and Chiyuan Zhang. 2025. Crosslingual capabilities and knowledge barriers in multilingual large language models. *Preprint*, arXiv:2406.16135.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*.

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

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 82 others. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. A framework for few-shot language model evaluation.

Ian J. Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. 2015. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *Preprint*, arXiv:1312.6211.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022a. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022b. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: a massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Xu Huang, Wenhao Zhu, Hanxu Hu, Conghui He, Lei Li, Shujian Huang, and Fei Yuan. 2025. Benchmax: A comprehensive multilingual evaluation suite for large language models. *Preprint*, arXiv:2502.07346.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Viet Lai, Chien Nguyen, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. 2023a. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 318–327, Singapore. Association for Computational Linguistics.

Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2023b. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. *Preprint*, arXiv:2307.16039.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. 2023. Is your code generated by chatGPT really correct? rigorous evaluation of large language models for code generation. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Dan Malkin, Tomasz Limisiewicz, and Gabriel Stanovsky. 2022. A balanced data approach for evaluating cross-lingual transfer: Mapping the linguistic blood bank. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4903–4915, Seattle, United States. Association for Computational Linguistics.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. PEFT: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

David Mueller, Mark Dredze, and Nicholas Andrews. 2024. Multi-task transfer matters during instruction-tuning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14880–14891, Bangkok, Thailand. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Vera Neplenbroek, Arianna Bisazza, and Raquel Fernández. 2024. MBBQ: A dataset for cross-lingual comparison of stereotypes in generative LLMs. In *First Conference on Language Modeling*.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I. Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Wei-Yin Ko, Madeline Smith, Antoine Bosselut, Alice Oh, Andre F. T. Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2024. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. *Preprint*, arXiv:2412.03304.

Alexey Tikhonov and Max Ryabinin. 2021. It's All in the Heads: Using Attention Heads as a Baseline for Cross-Lingual Transfer in Commonsense Reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3534–3546, Online. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. *Preprint*, arXiv:2109.01652.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Qingquan Zhang, Qiqi Duan, Bo Yuan, Yuhui Shi, and Jialin Liu. 2024. Exploring accuracy-fairness trade-off in large language models. *Preprint*, arXiv:2411.14500.

A Appendix

A.1 Implementation Details

We load models with HuggingFace Transformers (v4.54.1) as `AutoModelForCausalLM` in FP16 and use each model's default tokenizer; if no pad token exists we set it to `UNK`, else `EOS`, otherwise add a new `[PAD]` and resize embeddings. Text is tokenized with truncation and padding to fixed task-specific lengths. Fine-tuning uses PEFT/LoRA (r=32, α=64) on attention projections (`q_proj`, `k_proj`, `v_proj`, `o_proj`) and MLP blocks (`gate_proj`, `up_proj`, `down_proj`), optimizing a causal-LM objective over "prompt + gold answer" with labels equal to inputs (no prompt masking). Training runs for 3 epochs in FP16 with AdamW (learning rate 5×10^{-5}, β_1=0.9, β_2=0.999, ϵ=10^{-8}), a linear schedule with 10% warmup, gradient clipping at 1.0, automatic batch-size discovery, and epoch-end evaluation and checkpoint saving. For evaluation we use `lm-evaluation-harness` v0.4.9.1 with fixed seeds: `random_seed=0`, `numpy_seed=1234`, `torch_seed=1234`; task-specific scoring/decoding follows harness defaults (e.g., log-likelihood for multiple choice; otherwise greedy).

Compute resources and budget. All experiments were run on a multi-GPU research cluster with NVIDIA GH200 GPUs (120 GB each). The total compute budget across fine-tuning and evaluation was $\approx$ 12,000 GPU-hours.

A.2 Licenses and Terms of Use

We use only publicly available datasets, models, and tools under their original licenses, and we do not redistribute any third-party datasets or model weights. All third-party artifacts are cited in the main text and Appendix. We will release our code under the Apache License 2.0 and include third-party license notices in the repository. All third-party datasets, models, and tools are used strictly for research in accordance with their intended use and access conditions as stated by their creators. We do not repurpose research-only resources for non-research contexts.

Source type	On-task		Off-task	
	Δ (pp)	Win %	Δ (pp)	Harm %
Common-sense reasoning	0.98	72.00	−1.52	36.30
Factuality	4.16	86.40	−2.21	34.29
Fairness & bias	10.50	86.96	−1.74	35.69
Knowledge-intensive reasoning	0.75	62.66	−0.91	24.24
Coding	−0.06	7.07	−0.42	16.19
Math	−3.21	36.97	−0.36	18.07

Table 6: On-task vs. off-task by source task type (aggregated across families and sizes). On-task: mean cross-lingual Δ on the trained dataset excluding the trained language; off-task: mean Δ over all other task–language cells. Harm if $\Delta < -1.0$ pp.

Task type	Score (pp)	
	Donor	Recipient
Common-sense reasoning	−2.17	0.10
Factuality	−2.61	0.12
Fairness & bias	−3.90	−1.83
Knowledge-intensive reasoning	−1.30	−2.53
Coding	−0.57	0.83
Math	−0.18	−9.51

Table 7: Task-type donor vs. recipient scores (percentage points).

(a) Dataset level

Regime	Median CI [Q1–Q3]
Matched-Task (Cross-Language)	0.235 [0.107–0.409]
Matched-Language (Cross-Task)	0.263 [0.067–0.429]
Cross-Task (Cross-Language)	0.194 [0.148–0.234]

Table 8: Consistency index (CI) summaries at dataset and task-type levels.

(b) Task-type level

Regime	Median CI [Q1–Q3]
Matched-Task (Cross-Language)	0.195 [0.100–0.358]
Matched-Language (Cross-Task)	0.284 [0.136–0.477]
Cross-Task (Cross-Language)	0.209 [0.156–0.259]

Segment	Source task type	Source lang.	Mean Kendall's τ_b
Top	Factuality	id	0.704 659
Top	Factuality	es	0.700 154
Top	Factuality	vi	0.668 612
Top	Factuality	de	0.664 107
Top	Factuality	en	0.658 372
Bottom	Coding	zh	−0.207 112
Bottom	Fairness & bias	en	−0.200 000
Bottom	Fairness & bias	tr	−0.200 000
Bottom	Coding	hu	−0.194 776
Bottom	Fairness & bias	nl	−0.155 556

Table 9: Task-type level: top/bottom 5 sources by consistency (Kendall's τ_b) for the *Matched-Task (Cross-Language)* regime.

Segment	Source task type	Source lang.	Mean Kendall's τ_b
Top	Common-sense reasoning	eu	0.349 620
Top	Factuality	ca	0.337 053
Top	Knowledge-intensive reasoning	ca	0.334 225
Top	Knowledge-intensive reasoning	sk	0.324 542
Top	Factuality	ro	0.322 308
Bottom	Knowledge-intensive reasoning	te	−0.004 819
Bottom	Knowledge-intensive reasoning	gu	0.004 207
Bottom	Knowledge-intensive reasoning	ml	0.009 508
Bottom	Knowledge-intensive reasoning	kn	0.009 649
Bottom	Knowledge-intensive reasoning	ta	0.029 872

Table 10: Task-type level: top/bottom 5 sources by consistency (Kendall's τ_b) for the *Cross-Task (Cross-Language)* regime.

Segment	Source task type	Source lang.	Mean Kendall's τ_b
Top	Common-sense reasoning	nl	0.777 778
Top	Factuality	hu	0.733 333
Top	Knowledge-intensive reasoning	nl	0.733 333
Top	Knowledge-intensive reasoning	zh	0.682 540
Top	Factuality	nl	0.666 667
Bottom	Common-sense reasoning	vi	−0.200 000
Bottom	Common-sense reasoning	sr	−0.155 556
Bottom	Common-sense reasoning	hu	−0.133 333
Bottom	Coding	sr	−0.066 667
Bottom	Knowledge-intensive reasoning	de	−0.044 444

Table 11: Task-type level: top/bottom 5 sources by consistency (Kendall's τ_b) for the *cross-task, same-language* regime.

Lang	Donor score (pp)	Recipient score (pp)
ar	-0.02	1.64
bn	0.65	0.96
ca	0.74	2.49
cs	0.24	0.31
da	1.98	2.14
de	1.62	2.06
en	1.26	3.25
es	0.92	3.93
eu	1.63	-0.06
fr	1.84	1.91
gu	2.08	0.89
hi	2.41	1.37
hr	2.77	2.02
hu	1.75	1.05
hy	1.33	0.28
id	1.56	2.55
it	0.67	1.70
ja	-2.75	-1.36
kn	2.08	0.41
ko	-3.18	-0.72
ml	1.29	0.44
mr	1.29	0.81
ne	2.07	0.43
nl	4.85	3.90
pt	1.91	1.21
ro	1.47	2.07
ru	2.82	1.87
sk	1.74	1.61
sr	2.50	1.91
sv	2.05	2.49
sw	-4.97	-2.19
ta	1.58	0.50
te	1.51	0.33
th	0.06	-0.17
tr	7.14	9.59
uk	3.49	1.31
vi	3.95	2.20
yo	-5.35	-1.31
zh	1.68	3.02

Table 12: Language Donor vs. Recipient Scores (percentage points; pp).

Donor task	Recipient task	Avg. transfer (pp)
Coding	Common-sense reasoning	0.012
Coding	Factuality	2.042
Coding	Fairness & bias	0.003
Coding	Knowledge-intensive reasoning	−2.717
Coding	Math	−0.186
Common-sense reasoning	Coding	0.510
Common-sense reasoning	Factuality	1.090
Common-sense reasoning	Fairness & bias	−1.536
Common-sense reasoning	Knowledge-intensive reasoning	−2.628
Common-sense reasoning	Math	−7.927
Factuality	Coding	0.273
Factuality	Common-sense reasoning	−0.144
Factuality	Fairness & bias	0.713
Factuality	Knowledge-intensive reasoning	−2.867
Factuality	Math	−15.828
Fairness & bias	Coding	0.724
Fairness & bias	Common-sense reasoning	−0.872
Fairness & bias	Factuality	0.082
Fairness & bias	Knowledge-intensive reasoning	−4.047
Fairness & bias	Math	−5.691
Knowledge-intensive reasoning	Coding	0.523
Knowledge-intensive reasoning	Common-sense reasoning	0.288
Knowledge-intensive reasoning	Factuality	0.168
Knowledge-intensive reasoning	Fairness & bias	−3.778
Knowledge-intensive reasoning	Math	−8.944
Math	Coding	0.173
Math	Common-sense reasoning	0.157
Math	Factuality	0.734
Math	Fairness & bias	−1.069
Math	Knowledge-intensive reasoning	−2.232

Table 13: Directed task-to-task transfer effects. Positive values indicate beneficial transfer (accuracy gains, in percentage points), negative values indicate detrimental transfer.

Dataset	Task type	#Langs	Language codes (ISO 639-1)	Construction type	Short description / rationale
ARC-Challenge	Science QA (commonsense + background knowledge)	32	ar, bn, ca, da, de, es, en, eu, fr, gu, hi, hr, hu, hy, id, it, kn, ml, mr, ne, nl, pt, ro, ru, sk, sr, sv, ta, te, uk, vi, zh	*Translation-based* (LLM translation)	Knowledge-intensive multiple choice requiring background/world knowledge beyond surface cues; used to probe cross-lingual transfer for reasoning that relies on external knowledge under parallelized content.
TruthfulQA	Factuality / truthful QA	32	ar, bn, ca, da, de, en, es, eu, fr, gu, hi, hr, hu, hy, id, it, kn, ml, mr, ne, nl, pt, ro, ru, sk, sr, sv, ta, te, uk, vi, zh	*Translation-based* (LLM translation)	Stress-tests truthfulness against common misconceptions and misleading prompts; included to examine whether gains in other tasks/languages spill over or harm factual reliability when content is held constant across languages.
HellaSwag	Commonsense inference	32	ar, bn, ca, da, de, en, es, eu, fr, gu, hi, hr, hu, hy, id, it, kn, ml, mr, ne, nl, pt, ro, ru, sk, sr, sv, ta, te, uk, vi, zh	*Translation-based* (LLM translation)	Adversarial commonsense MCQ designed to reduce annotation artifacts; used to test everyday reasoning transfer across scripts and typologies under parallel scenarios.
Global-MMLU	Knowledge-intensive multiple choice	15	ar, bn, de, en, es, fr, hi, id, it, ja, ko, pt, sw, yo, zh	*Translation-based* (human translated/post-edited; localised)	Broad subject knowledge with localization to mitigate Anglocentrism; we evaluate the 15 languages from Global-MMLU-Lite that were fully human-translated or post-edited. Curated by professional annotators and Cohere Labs contributors.
xWinograd	Coreference / pronoun resolution	6	en, fr, ja, pt, ru, zh	*Authored multilingual (non-translated; aggregated monolingual sources)*	Coreference resolution targeting pronominal/lexical ambiguity across languages; offers a genuinely multilingual signal to validate transfer patterns beyond translation artifacts.
MGSM	Math word-problem reasoning	11	bn, de, en, es, fr, ja, ru, sw, te, th, zh	*Translation-based* (human translated; parallelized)	Math word problems translated from GSM8K with numeric/semantic alignment; probes whether cross-lingual transfer covers compositional reasoning and arithmetic grounding.
HumanEval+	Code generation (unit-test pass@k)	17	ar, bn, cs, de, en, es, fr, hu, ja, ko, ru, sr, sw, te, th, vi, zh	*Translation-based* (LLM translation + human post-edit; parallelized)	Code generation evaluated by unit tests; English set extended to 16 non-English languages via GPT–4o translation with native-speaker post-editing, enabling clean execution-level comparisons across languages.
MBBQ	Social bias / stereotypes (MCQ)	4	en, es, nl, tr	*Translation-based* (machine translation + human post-edit; parallelized)	Bias/stereotype QA with templated scenarios; items translated with Google Translate and NLLB-200, then hand-checked by native speakers (choose best or rewrite). Measures whether fine-tuning along other axes affects fairness and induces collateral degradation cross-lingually.

Table 14: Datasets, task types, language coverage, construction type, and short descriptions summarizing task focus and motivation. We use *Authored multilingual (non-translated)* for datasets created directly in multiple languages, and *Translation-based* for datasets built by translating a source set (LLM or machine translation, often with human post-editing). "Parallelized" indicates the same items across languages; "localized" indicates culturally adapted while maintaining comparability.

Source \ Eval	ar	bn	ca	cs	da	de	en	es	eu	fr	gu	hi	hr	hu	hy	id	it	ja	kn	ko	ml	mr	ne	nl	pt	ro	ru	sk	sr	sv	sw	ta	te	th	tr	uk	vi	yo	zh
ar	−0.20	−0.49	0.97	0.12	0.73	−1.80	−1.05	−1.55	−0.10	−0.81	0.14	−0.14	0.72	0.20	−0.09	−0.05	−0.57	−2.95	−0.15	−0.87	0.00	0.07	0.03	0.05	−1.47	0.75	−0.27	0.07	0.46	0.95	−2.76	0.05	−0.28	−4.11	−1.20	0.18	0.00	−2.72	−0.71
bn	−0.16	−1.05	0.74	0.00	0.63	−1.36	−0.89	−0.91	−0.08	−0.78	0.39	−0.15	0.46	0.34	−0.03	−0.52	−1.24	−2.54	−0.18	−2.16	0.16	0.28	0.17	0.19	−2.09	0.36	0.01	−0.03	0.51	0.82	−2.38	−0.01	−0.18	−1.83	−1.42	0.19	−0.06	−1.95	−0.56
ca	−1.09	−1.74	0.41	0.30	−0.96	−5.58	−8.04	−7.65	−1.17	−6.73	0.15	−0.75	−0.45	−1.20	−0.14	−2.04	−3.67	−7.54	−0.29	−3.70	−0.37	0.04	−0.18	−2.03	−2.89	−2.59	−5.39	−0.71	−0.52	−0.89	−8.95	−0.26	−0.48	−16.22	−3.29	−0.81	−1.05	−2.37	−4.40
da	−0.35	−1.44	0.21	0.12	−0.37	−4.91	−3.42	−4.34	−0.50	−3.07	0.38	−0.24	0.34	−0.29	−0.06	−0.51	−0.56	−5.35	−0.20	−1.76	−0.30	−0.09	−0.06	−1.22	−1.68	0.04	−3.43	−0.09	0.27	−0.25	−7.35	−0.10	−0.24	−10.45	−1.64	0.20	0.00	−1.17	−2.59
de	−0.19	−0.43	1.31	0.20	1.07	−5.01	−3.67	−3.18	−0.15	−2.73	0.27	0.20	0.82	0.40	−0.10	0.09	−0.36	−4.09	−0.12	−1.19	0.07	0.19	0.14	−0.13	−1.96	0.96	−1.66	0.27	0.75	1.37	−5.99	−0.06	0.13	−8.42	−1.21	0.48	0.55	−2.37	−2.05
en	−0.47	−0.86	0.24	0.26	−0.05	−4.23	−3.96	−4.06	−0.42	−3.44	0.02	−0.19	−0.20	0.04	0.16	−0.78	−1.46	−6.08	−0.05	−0.97	−0.02	0.22	0.09	0.26	−2.49	−0.13	−1.93	−0.56	−0.31	0.12	−6.27	−0.14	−0.23	−9.93	7.75	−0.32	−0.38	−2.18	−3.42
es	−0.40	−0.88	0.65	0.06	0.38	−3.31	−0.21	−5.05	−0.34	−3.21	0.10	−0.38	0.60	0.09	−0.07	−1.40	−1.58	−3.94	0.04	−2.04	0.03	0.27	−0.02	0.23	−2.96	0.02	−1.35	−0.13	0.38	0.60	−5.07	0.05	−0.06	−5.22	3.26	0.06	−0.26	−3.68	−2.30
eu	−1.06	−0.55	0.46	0.00	−0.88	−4.20	−4.80	−3.79	−0.18	−3.45	0.11	−0.48	−0.60	−1.07	−0.39	−1.81	−2.12	−4.42	−0.08	−3.26	0.20	0.13	−0.42	−0.55	−2.79	−1.17	−2.39	−0.74	−0.55	−0.37	−6.90	−0.11	−0.14	−8.20	−0.81	−1.08	−0.68	−2.65	−2.46
fr	−0.38	−0.50	0.86	0.00	0.97	−2.45	−2.71	−2.47	−0.28	−2.82	0.37	−0.07	0.71	0.26	0.11	−0.38	−1.18	−3.08	−0.08	−1.97	0.13	0.33	0.12	0.53	−2.64	0.38	−0.92	0.22	0.58	1.37	−4.41	−0.02	−0.06	−4.30	−0.85	0.42	0.21	−3.29	−1.29
gu	−0.67	−0.72	0.89	0.00	0.63	−2.27	−1.29	−1.65	0.05	−1.14	0.44	−0.22	0.56	0.24	0.19	−0.95	−1.48	−3.32	0.28	−2.69	0.10	0.41	0.42	0.19	−2.78	0.30	0.07	0.01	0.41	0.73	−3.68	0.15	−0.31	−2.13	−0.78	0.17	0.10	−3.33	−1.25
hi	−0.55	−1.37	0.84	0.00	0.57	−3.40	−1.47	−3.27	−0.22	−2.01	0.61	−1.51	0.45	0.30	0.24	−0.59	−1.35	−5.57	0.13	−3.32	0.11	0.45	0.01	0.52	−2.31	0.43	−1.15	−0.25	0.31	0.88	−5.74	0.14	−0.19	−6.51	3.40	0.09	0.16	−2.68	−1.70
hr	−0.12	−0.95	0.96	0.30	0.04	−3.46	−2.46	−3.09	−0.71	−2.64	0.55	−0.11	0.11	0.27	0.16	−0.36	−0.28	−4.54	−0.11	−1.75	0.00	0.17	−0.22	−0.20	−1.79	−0.30	−3.76	−0.51	2.16	0.80	−8.16	0.04	−0.05	−8.34	−2.32	−0.38	0.55	−1.59	−2.29
hu	0.07	−0.58	1.12	0.20	1.24	−3.14	−4.03	−2.75	−0.96	−2.20	0.23	0.06	0.85	−0.01	0.09	0.16	−0.23	−4.06	−0.43	−0.49	0.27	−0.07	−0.17	0.72	−1.52	0.81	−2.16	0.43	0.93	1.50	−6.96	−0.21	−0.32	−10.70	0.35	0.17	0.40	−2.15	−1.43
hy	−0.48	0.14	0.94	0.10	0.73	−1.49	−0.76	−0.89	−0.07	−0.90	0.35	−0.30	0.35	0.32	0.18	−0.41	−1.43	−2.28	0.14	−2.00	0.35	0.16	0.18	0.00	−3.31	0.46	0.37	0.11	0.46	1.13	−2.52	0.00	0.03	−1.07	−0.35	0.24	0.00	−2.53	−1.12
id	0.05	−0.23	1.34	0.20	1.34	−3.85	−4.28	−3.38	−0.45	−2.98	0.32	0.17	0.85	0.50	0.02	−0.77	−0.10	−4.83	0.01	−0.82	0.35	0.34	0.03	0.98	−2.01	0.86	−2.24	0.02	0.59	1.92	−8.42	−0.07	0.13	−11.71	−1.88	0.37	0.50	−2.45	−1.97
it	−0.18	−0.40	−0.36	0.37	0.74	−3.50	−4.73	−4.26	−0.29	−3.76	0.24	0.11	0.52	0.26	0.01	0.05	−0.57	−4.54	−0.21	−0.13	0.14	0.29	0.10	0.29	−2.05	−0.42	−2.65	0.07	0.32	1.13	−6.70	−0.11	0.14	−10.46	−1.51	0.00	0.15	−1.86	−1.94
ja	−0.11	−0.30	0.73	0.12	0.66	−1.42	−1.25	−0.30	−0.26	−1.29	0.31	−0.06	0.24	0.28	0.14	−0.98	−1.42	−4.00	0.02	−2.10	0.27	0.45	0.36	0.13	−4.55	0.41	0.80	0.03	0.26	1.03	−2.75	0.07	−0.19	0.00	−0.60	0.28	−0.11	−4.38	−0.41
kn	−0.08	−0.04	1.13	0.00	0.99	−2.23	−1.22	−1.10	−0.09	−0.98	0.62	0.22	0.45	0.32	0.13	−0.27	−1.20	−2.70	−0.38	−1.86	0.20	0.62	0.22	1.08	−2.77	0.60	0.20	0.03	0.42	1.26	−4.07	0.29	−0.34	−1.13	6.17	0.44	0.08	−2.10	−0.94
ko	0.05	0.74	0.74	0.00	1.01	−1.21	−0.33	0.45	−0.20	−0.99	0.38	0.29	−0.12	0.64	0.42	−0.17	−2.05	−1.38	0.06	−2.37	0.41	0.72	0.68	0.95	−5.29	1.60	1.00	0.29	0.24	1.42	−2.46	0.29	0.38	1.47	−0.42	0.63	0.49	−3.97	0.32
ml	−0.41	0.62	1.22	0.00	1.25	−1.76	−1.59	−0.92	−0.17	−0.71	0.34	0.23	0.45	0.42	0.29	−0.51	−1.33	−3.41	−0.25	−1.85	0.24	0.35	0.35	1.02	−3.21	0.71	0.61	0.37	0.64	1.52	−4.24	0.16	0.26	0.36	−0.57	0.29	0.28	−3.13	−1.57
mr	−0.70	−1.35	1.26	0.00	0.98	−3.58	−1.70	−3.12	−0.51	−2.89	0.63	−0.06	0.06	0.07	0.33	−0.12	−1.00	−6.70	0.14	−2.39	0.59	0.77	−0.28	1.34	−2.61	0.45	−1.61	−0.14	0.21	1.25	−7.26	−0.05	−0.65	−11.45	3.04	0.05	0.53	−1.95	−3.38
ne	−0.06	−0.89	1.30	0.41	0.90	−2.44	−0.29	−1.74	−0.10	−1.34	0.69	0.59	0.90	0.50	0.27	0.11	0.13	−4.12	−0.03	−1.16	0.42	0.16	0.26	0.69	−2.06	0.86	−0.08	0.34	0.72	1.67	−4.92	0.34	0.41	−7.85	−1.70	0.36	0.48	−2.12	−1.28
nl	−0.71	−1.21	0.59	0.23	−0.24	−4.87	−2.45	−3.52	−0.49	−3.36	0.32	−0.53	0.44	0.31	0.10	−1.25	−1.05	−6.70	−0.06	−2.70	0.14	0.23	−0.13	−1.90	−2.34	−0.01	−2.71	−0.07	0.24	0.47	−7.32	0.06	0.04	−9.07	7.42	0.17	0.10	−2.41	−3.63
pt	0.06	−0.75	0.68	0.30	0.98	−4.04	−6.52	−4.72	−0.43	−4.12	0.34	−0.15	0.72	0.36	0.03	−0.28	−2.08	−4.60	−0.05	−1.55	0.18	0.38	−0.01	0.48	−2.53	0.63	−2.10	0.22	0.64	1.60	−6.89	−0.07	−0.01	−9.33	−1.62	0.41	0.34	−3.02	−2.41
ro	0.09	−0.76	−0.15	0.46	0.58	−3.96	−4.81	−4.20	−0.46	−3.97	0.27	−0.08	0.46	0.28	−0.03	−0.76	−0.68	−4.60	−0.20	−0.98	−0.04	0.22	0.06	0.38	−1.68	−0.34	−4.03	0.14	0.62	0.99	−8.39	−0.06	0.27	−13.33	−0.40	−0.17	−0.02	−3.19	−2.25
ru	−0.53	−1.02	1.08	0.00	1.26	−4.38	−2.09	−3.90	−0.18	−3.26	0.56	0.11	0.56	0.42	0.15	−0.23	−0.70	−5.87	0.11	−2.47	0.12	0.34	0.16	1.13	−2.49	0.75	−4.99	0.05	0.64	1.60	−8.11	−0.14	−0.27	−13.23	−0.15	0.34	0.87	−2.63	−2.92
sk	0.04	−0.12	0.60	0.61	−0.65	−4.14	−3.34	−3.46	−0.94	−3.27	0.66	0.11	−1.77	−0.30	0.13	−0.25	0.02	−5.38	−0.21	−0.57	0.19	−0.02	−0.20	0.67	−1.69	−0.29	−5.75	0.05	−1.63	0.47	−8.23	0.04	0.44	−12.49	1.14	−1.07	0.34	−2.34	−2.34
sr	−0.10	−0.55	1.51	0.30	0.92	−3.85	−3.24	−2.84	−0.78	−3.16	0.51	0.33	2.37	0.41	0.06	0.56	−0.17	−4.45	−0.31	−1.26	0.15	0.41	−0.01	0.63	−1.40	0.50	−3.39	0.23	0.18	1.56	−9.49	0.08	0.13	−12.11	−4.12	0.16	0.62	−2.76	−2.44
sv	−0.10	−1.43	1.05	0.30	0.00	−4.94	−4.32	−4.68	−0.43	−3.55	0.48	0.02	0.45	−0.17	−0.08	−0.26	−0.28	−6.79	−0.15	−1.19	−0.06	0.29	−0.06	−0.32	−1.61	0.13	−3.15	−0.05	0.33	0.10	−8.78	−0.07	−0.29	−13.00	−1.31	0.21	0.29	−0.85	−3.72
sw	0.10	0.13	1.31	0.30	1.06	−1.51	−2.51	−0.60	−0.09	−1.59	0.41	−0.07	0.27	0.30	0.07	−0.33	−2.19	−1.26	−0.11	−0.98	0.32	0.35	0.22	0.76	−4.00	0.36	0.90	−0.30	0.45	1.59	−2.14	−0.12	−0.23	0.00	−0.33	0.33	0.16	−2.99	0.12
ta	−0.29	0.04	1.10	0.15	0.92	−2.26	−1.21	−1.17	−0.11	−1.33	0.86	−0.21	0.38	0.23	0.31	−1.05	−1.18	−2.74	0.09	−2.33	0.15	0.55	0.42	0.53	−3.44	0.76	0.71	0.30	0.33	1.28	−3.16	0.38	0.17	−0.11	0.94	0.39	0.23	−3.56	−0.89
te	0.15	0.35	1.06	0.00	1.42	−1.89	−0.96	−0.42	−0.15	−1.37	0.64	0.42	0.57	0.41	0.23	−0.35	−0.91	−2.77	−0.26	−1.74	0.56	0.54	0.49	1.14	−4.00	0.86	0.88	0.37	0.64	1.76	−2.33	0.22	−0.75	0.06	2.46	0.56	0.38	−4.46	−0.58
th	−0.35	−0.42	−0.29	0.00	0.54	−0.90	0.00	0.06	−0.21	−0.97	−0.04	0.87	0.40	0.08	−0.09	−0.86	−1.23	−0.35	−0.31	−0.24	−0.04	0.18	0.74	−0.18	−3.88	0.52	0.18	0.39	0.32	0.77	−0.22	0.01	−0.24	1.60	−1.94	0.13	−0.56	−6.58	−0.17
tr	−1.59	−1.94	−0.37	0.00	−1.18	−2.29	−0.01	−1.15	−0.36	−2.35	0.02	−1.96	−0.87	−0.58	0.46	−4.23	−2.99	−4.02	0.45	−3.85	0.31	−0.27	−0.26	0.99	−4.83	−1.82	−1.55	−0.75	−0.51	−1.05	−3.38	−0.02	−0.38	−3.15	0.00	−0.37	−0.86	−4.22	−1.37
uk	−0.44	−0.39	1.12	0.30	1.52	−4.00	−0.83	−3.60	−0.37	−2.81	0.54	0.44	0.89	0.66	0.11	0.19	0.11	−6.45	−0.21	−1.90	0.02	0.15	−0.04	1.47	−1.48	1.07	−5.86	0.67	0.80	1.55	−7.02	−0.14	0.16	−14.12	2.29	−0.17	1.13	−1.59	−2.69
vi	0.10	−0.92	1.64	0.61	1.59	−4.12	−3.48	−3.74	−0.26	−3.78	0.49	0.95	1.28	0.89	0.17	0.68	0.34	−7.13	−0.03	−2.18	0.22	0.45	0.39	1.55	−0.56	1.62	−3.38	0.65	0.96	2.14	−9.56	0.04	0.11	−17.32	1.93	0.62	−0.17	−1.31	−3.42
yo	−0.16	−0.08	0.51	0.00	0.45	−1.24	−1.26	−0.11	−0.14	−0.65	0.13	0.10	0.27	0.30	0.00	−0.94	−1.47	−1.37	−0.38	−1.79	0.34	0.01	−0.09	0.16	−3.69	−0.39	0.33	−0.18	0.34	0.72	−3.74	−0.03	−0.04	−1.20	0.00	−0.03	−0.58	0.00	0.07
zh	−0.23	−0.39	0.89	0.08	0.79	−2.34	−0.73	−1.92	−0.21	−2.09	0.33	0.32	0.42	0.47	0.13	−0.43	−0.91	−3.64	0.01	−1.53	0.18	0.32	0.22	0.65	−2.90	0.63	−0.72	0.15	0.35	1.01	−3.52	0.01	−0.07	−4.58	1.72	0.31	0.43	−2.69	−3.51

Table 15: Language-to-language transfer matrix (Δ in percentage points). Rows are *source* (fine-tuned) languages; columns are *evaluation* languages.