Cross-Lingual Transfer for Machine Translation in Turkic Languages

Anonymous ACL submission

Abstract

Cross-lingual transfer is central to low-resource machine translation, but its behavior within closely related language families remains insufficiently characterized. We study transfer among five Turkic languages—Turkish, Azerbaijani, Uzbek, Kazakh, and Kyrgyz—using pairwise transfer matrices. In this setting, each model is fine-tuned with one transfer source and evaluated on a different transfer target while the translation target remains the same. Across mT5 experiments, we find that transfer is strongest between closely related Turkic pairs, especially Turkish–Azerbaijani and Kazakh–Kyrgyz. We also show that transfer direction matters, and that the same transfer source–transfer target pair can behave differently when the translation target changes. Latinization improves BLEU and chrF in several script-mismatched settings, but its effect is not uniform across metrics. Additional analyses show that transfer sources are mostly stable across different datasets and model settings.

1 Introduction

Neural Machine Translation (NMT) (Sutskever et al., 2014) has largely replaced earlier statistical and rule-based machine translation systems. Attention mechanisms (Bahdanau et al., 2015) and the Transformer architecture (Vaswani et al., 2017) have substantially improved translation quality, but low-resource settings still remain difficult because high-quality parallel data is scarce (Koehn and Knowles, 2017). Transfer learning is therefore widely used to share knowledge across languages (Zoph et al., 2016), especially when languages are typologically related and share vocabulary or morphology (Kocmi, 2020).

The Turkic family is an unusually useful test bed for pairwise transfer. It combines typological relatedness with noticeable asymmetries in resource level and script: Turkish and Azerbaijani are written in Latin script, while Kazakh and Kyrgyz are commonly written in Cyrillic; Uzbek is mixed in practice but often processed in Latin. This combination creates a controlled setting in which transfer can be studied across related languages while still exposing the effects of script alignment and resource imbalance.

Because the Turkic language family is less-resourced (i.e. parallel data between Turkic language pairs is limited), cross-lingual transfer can be a practical strategy for building MT systems in this family. This makes it important to understand which Turkic languages transfer well to each other, and whether these transfer relations change with the target language. Much existing evidence in multilingual MT comes from English-centric settings, where English is often used as the main source, target, or pivot language in training and evaluation (Fan et al., 2021). When English is used as the fixed target language in evaluation pairs, transfer from one source language to another is observed only under English as the target. This does not show how the choice of target language changes cross-lingual transfer. To study this effect within the Turkic family, we use Turkic languages as target languages and ask a more fine-grained question: for a fixed target language, which source language is the best fine-tuning donor for each unseen evaluation source, and does that preference remain stable when the evaluation source, script representation, dataset, or pretraining regime changes? Although target-aware data selection has been studied in multilingual NMT (Wang and Neubig, 2019), pairwise transfer between source languages under different Turkic target languages remains less explored. We address this gap by constructing fixed-target transfer matrices among five Turkic languages.

Throughout the paper, we distinguish between translation direction and transfer direction. The **translation source** is the language from which a sentence is translated, and the **translation target** is the language into which it is translated. For

example, in Turkish-to-Azerbaijani translation, Turkish is the translation source and Azerbaijani is the translation target. In our fixed-target transfer setting, we additionally define a **transfer source** and a **transfer target**. The transfer source is the source language used during fine-tuning, while the transfer target is the source language used only at evaluation time. For instance, if a model is fine-tuned on Turkish→Azerbaijani and evaluated on Uzbek→Azerbaijani, then Turkish is the transfer source, Uzbek is the transfer target, and Azerbaijani remains the fixed translation target. This design measures the contribution of the transfer source to the transfer target under the same translation target.

Our study is organized around three research questions:
RQ1. Which Turkic languages transfer most effectively to unseen languages in cross-lingual machine translation?
RQ2. How does the choice of translation target affect cross-lingual transfer?
RQ3. How does transliteration into the Latin alphabet affect Turkic cross-lingual transfer?

Our main contributions are as follows:
C1. To the best of our knowledge, we present the first MT-based pairwise transfer matrix for five Turkic languages.
C2. Unlike prior transfer studies that rely on non-Turkic pivot languages, we use Turkic languages as both translation sources and targets, enabling a controlled analysis of transfer behavior within the same language family.
C3. We release our best-performing Turkic machine translation models together with the book-aligned parallel pairs created for this study.[1]

2 Related Work

Research in multilingual neural machine translation (MNMT) has extensively studied how transfer emerges across languages, when it is beneficial or harmful, and how data composition shapes these dynamics. We summarize the literature along three axes: (i) transfer learning and cross-lingual transfer, (ii) multilingual interference and directionality, and (iii) machine translation for low-resource and Turkic languages.

[1] https://anonymous.4open.science/r/Cross-LingualTransferforMachineTranslationinTurkicLanguages-40D8/

2.1 Transfer Learning and Cross-Lingual Transfer

Transfer learning plays a central role in low-resource machine translation. Zoph et al. (2016) show that transferring parameters from a high-resource parent model to a low-resource child model yields significant improvements. Subsequent work demonstrates that transfer is strengthened by linguistic similarity and shared sub-word vocabularies (Nguyen and Chiang, 2017), and that even simple continued training can yield strong gains without architectural changes (Kocmi and Bojar, 2018). Later analyses confirm that transfer is most effective for low-resource and closely related language pairs (Kocmi, 2020).

Recent work has focused on measuring transfer more explicitly. Representational Transfer Potential (RTP) (Stap et al., 2023) captures positive and negative transfer via representation similarity rather than surface metrics. Large-scale studies such as ATLAS (Longpre et al., 2025) and Interference Matrix (Alastruey et al., 2025) further show that transfer is asymmetric and strongly influenced by language family and script similarity.

2.2 Multilingual Interference and Directionality

Although multilingual training enables parameter sharing across languages, adding more languages or mixing imbalanced data does not always improve performance. Prior work has identified capacity limitations and negative transfer in multilingual models, often described as the curse of multilinguality (Conneau et al., 2020). This issue is especially relevant for low-resource language families, where dominant high-resource languages may help some transfer directions while hurting others because of inter-language parameter competition (Blevins et al., 2024; Liu and Niehues, 2025).

Directionality is also important for cross-lingual transfer. Prior work on joint multilingual training shows that translation tasks can interact positively, negatively, or asymmetrically, where one task improves while the other degrades (Wang and Zhang, 2022). Other studies investigate target-aware data selection in multilingual NMT (Wang and Neubig, 2019) and the role of translation direction in multilingual training (Luo et al., 2025). However, less attention has been given to the direction from transfer source to transfer target, especially how this direction changes under different translation

targets. Our work focuses on this setting by constructing pairwise transfer matrices among Turkic languages.

2.3 Low-Resource and Turkic Machine Translation

Low-resource machine translation is often constrained by the scarcity of high-quality parallel data. Backtranslation (Sennrich et al., 2016) is commonly used to create synthetic parallel data by translating monolingual target-language data into the source language, and high-resource pivot languages such as English are also used to support data construction (Kim et al., 2019). Beyond data creation, low-resource MT also depends on methods for finding and assessing useful parallel sentence pairs. Sentence embedding models such as LaBSE are effective for low-resource bitext mining (Chimoto and Bassett, 2022), while large language models can support translation quality assessment and filtering decisions (Kocmi and Federmann, 2023a).

Turkic languages present a particularly challenging low-resource setting because many language pairs have limited direct parallel data, and the family also includes differences in script and resource availability. Prior work has demonstrated the usefulness of Turkish as a high-resource pivot for improving NLU performance across other low-resource Turkic languages (Senel et al., 2024). In machine translation, multilingual training for Turkic languages has also been shown to improve performance through joint training (Mirzakhalov et al., 2021). Building on this line of research, we provide a comprehensive cross-lingual analysis of Turkic MT transfer, examining how transfer patterns change across Latinization, model architectures, datasets, and continual pretraining settings.

3 Data

We use three types of data: monolingual corpora for continual pretraining (CPT), bilingual corpora for fine-tuning, and held-out evaluation datasets. The study covers Turkish (`tr`), Azerbaijani (`az`), Uzbek (`uz`), Kazakh (`kk`), and Kyrgyz (`ky`), denoted by their respective ISO 639-1 language codes.

For CPT, we collect monolingual data from publicly available sources, including Wikipedia, CC-100 (Conneau et al., 2020), Leipzig Corpora (Goldhahn et al., 2012), OSCAR (Abadji et al., 2022), mC4 (Xue et al., 2021), MADLAD-400 (Kudugunta et al., 2023), and HPLT (Aulamo et al., 2023). These corpora are used to adapt `mT5-small` to Turkic text before translation fine-tuning.

For bilingual fine-tuning, we combine publicly available parallel data with synthetic and mined parallel pairs. Public resources include KazParC (Yeshpanov et al., 2024), NTREX (Federmann et al., 2022), FLORES+ (Gordeev et al., 2024), and OPUS (Tiedemann, 2012). Because direct Turkic–Turkic parallel data is highly imbalanced and not very suitable for training due to short sentence length, we additionally construct training pairs through back-translation with `facebook/nllb-200-distilled-600M` (Costa-jussà et al., 2022) and LaBSE-based bitext mining from multilingual book translations (Feng et al., 2022). We combine book-aligned and back-translated pairs to create a 100K-pair fine-tuning dataset for each translation pair, with a total of 2-million pairs. Details are provided in Appendix A.1.

This data setting is imperfect but realistic for low-resource Turkic MT. Since many fine-tuning pairs are synthetic, we interpret the results primarily as evidence about *relative transfer structure* under a controlled training pipeline, not as a claim that the absolute MT quality is optimal. We discuss this limitation explicitly in the Limitations section.

For evaluation, we use XWMT as the main benchmark because it provides comparable test pairs across the five Turkic languages (Mirzakhalov et al., 2021). We use Tatoeba (Tiedemann, 2020) as a secondary benchmark to test whether donor preferences and transfer patterns remain stable across datasets.

4 Methodology

Continual Pretraining (CPT) We build on `mT5-small` (Xue et al., 2021), a 300M-parameter multilingual extension of T5 pretrained with a span-corruption objective on mC4. We continue pretraining on Turkic monolingual corpora with the same objective to adapt the model to Turkic text before translation fine-tuning. This follows prior work showing that continued pretraining can improve low-resource translation performance for Uyghur and language adaptation for Turkish (Lu et al., 2025; Toraman, 2024). In the allCPT setting, we sample one million approximately 512-token chunks from each of the five languages. Each training sample is prepended with a language tag, such as `<tr>`, to preserve language identity.

Fine-tuning For each fine-tuning run, we select one translation pair $(i \to t)$, where i is the source language and t is the target language. We use explicit translation-source and translation-target tags, following the text-to-text formulation of T5-style models (Raffel et al., 2020) and multilingual translation tagging conventions (Johnson et al., 2017). Each input is formatted as:

```
<src_lang> <tgt_lang>: src sentence (1)
```

where <src_lang> denotes the translation source and <tgt_lang> denotes the translation target. Both are one of <tr>, <az>, <uz>, <kk>, and <ky>.

Evaluation Evaluation is conducted in a zero-shot and fixed-translation-target setting. After fine-tuning on one translation pair $(i \to t)$, we evaluate the model on translation pairs $(j \to t)$, where the translation target t remains fixed and j varies over the Turkic languages.

When $i \neq j$, i is the *transfer source* and j is the *transfer target*: by keeping the translation target fixed, we measure how fine-tuning on $i \to t$ affects performance on the unseen pair $j \to t$. In other words, we measure how much the transfer source i contributes to the transfer target j when the translation target is t.

When $i = j$, the model is evaluated on the same translation pair used during fine-tuning. This is the easiest evaluation case, because the model has seen fine-tuning data for the exact source–target direction being tested. We use this as a reference score for the transfer cases where $i \neq j$.

Recovery rate To quantify cross-lingual transfer, we use recovery rate adapted from prior transfer-analysis work (Turc et al., 2021; Toraman et al., 2022). For an evaluation metric m, let $S_m(i, j, t)$ denote the score of a model fine-tuned on $i \to t$ and evaluated on $j \to t$, where i is the transfer source, j is the transfer target, and t is the translation target. Recovery rate measures the percentage of performance retained when the transfer source i is different from the transfer target j for a certain translation target t. It is computed as:

$$R_m(i, j, t) = \frac{S_m(i, j, t)}{S_m(j, j, t)} \times 100. \quad (2)$$

To obtain a language-pair-level transfer score from transfer source i to transfer target j, we average recovery over all valid translation targets that are different from both i and j:

$$T_m(i, j) = \frac{1}{|\mathcal{L} \setminus \{i, j\}|} \sum_{t \in \mathcal{L} \setminus \{i,j\}} R_m(i, j, t), \quad (3)$$

where $\mathcal{L} = \{\mathrm{tr}, \mathrm{az}, \mathrm{uz}, \mathrm{kk}, \mathrm{ky}\}$. Thus, $T_m(i, j)$ summarizes how strongly transfer source i supports transfer target j across possible Turkic translation targets.

We also include English as a high-resource external transfer source from a different language family. English provides a useful contrast for observing how transfer behaves when the fine-tuning source is not linguistically close to the Turkic languages. For Turkic transfer sources, recovery is averaged over translation targets $t \in \mathcal{L} \setminus \{i, j\}$, so that the translation target differs from both the transfer source i and the transfer target j. For English, this condition is relaxed because English is included only as an external transfer source and not as a translation target in $\mathcal{L}$. Therefore, the English row is averaged over translation targets $t \in \mathcal{L} \setminus \{j\}$.

Evaluation Metrics We report four evaluation metrics. BLEU (Papineni et al., 2002) measures word-level n-gram overlap between the model output and the reference translation, while chrF (Popović, 2015) measures character-level n-gram overlap with an F-score. We also use neural evaluation metrics: COMET (Rei et al., 2020) is a reference-based metric that estimates translation quality using the source sentence, model output, and reference translation, whereas COMETKiwi (Rei et al., 2022) is a reference-free quality estimation metric that evaluates the model output using only the source sentence. We report all metrics consistently across the transfer-matrix experiments.

5 Experiments and Results

5.1 Cross-Lingual Transfer Matrices (RQ1)

We first evaluate cross-lingual transfer in the all-CPT mT5 setting, where all five Turkic languages are represented in Latin script. Using the recovery rate defined in Section 4, we construct transfer matrices where rows correspond to transfer sources and columns correspond to transfer targets. These matrices answer RQ1 by showing how transfer dynamics are shaped by the relationship between the transfer source and the transfer target.

Closer Turkic pairs show stronger transfer than English. Table 1 shows that transfer strength differs clearly across transfer source–transfer target pairs. The strongest BLEU recovery appears

BLEU recovery rate (%)						
FT \ Eval	**tr**	**az**	**uz**	**kk**	**ky**	**Avg.**
tr	–	57.19	35.03	26.59	38.64	**39.36**
az	41.78	–	40.82	30.65	43.15	**39.10**
uz	33.05	44.59	–	37.20	42.48	**39.33**
kk	37.15	42.77	39.98	–	48.99	**42.22**
ky	32.54	38.41	41.82	39.32	–	**38.02**
en	30.86	34.16	28.08	29.47	32.29	**30.97**

chrF recovery rate (%)						
FT \ Eval	**tr**	**az**	**uz**	**kk**	**ky**	**Avg.**
tr	–	79.90	70.17	59.61	69.55	**69.81**
az	71.39	–	73.16	66.08	74.47	**71.28**
uz	65.55	73.60	–	71.66	73.34	**71.04**
kk	66.31	71.37	73.35	–	79.87	**72.73**
ky	63.05	69.96	72.42	68.86	–	**68.57**
en	63.11	63.87	62.65	63.05	65.17	**63.57**

Table 1: allCPT mT5 Latin recovery-rate transfer matrix. Rows indicate transfer sources and columns indicate transfer targets; non-diagonal cells report recovery averaged over translation targets different from both languages. The English row is a non-Turkic baseline, diagonal cells are omitted, and the Avg. column reports row-wise averages.

from Turkish to Azerbaijani, where tr→az reaches 57.19%, and from Kazakh to Kyrgyz, where kk→ky reaches 48.99%. The same pattern appears in chrF: tr→az reaches 79.90%, while kk→ky reaches 79.87%. These pairs also align with the subgroups of the Turkic family: Turkish and Azerbaijani are both Oghuz languages, while Kazakh and Kyrgyz are both Kipchak languages.

English is included as a high-resource non-Turkic transfer source to test whether the strongest transfer results come from general high-resource training or from linguistic relatedness within the Turkic family. In BLEU, English recovers only 28.08–34.16%, which is below the strongest Turkic transfer pairs. In chrF, English is more stable, reaching 62.65–65.17%, but it still remains below the intra-Turkic pairs. This suggests that high-resource training alone does not explain the strongest transfer results, and linguistic relatedness within the Turkic family plays an important role.

This is consistent with recent work suggesting that intrinsic language similarity and domain match can support cross-lingual transfer (Eronen et al., 2023), with similar observations reported for Uralic languages (Tars et al., 2021). While recent studies have established the effectiveness of cross-lingual transfer across the Turkic language family (Nguyen and Chiang, 2017; Yazar and Kiliç, 2025), our results further provide evidence that the degree of similarity, family subgroup proximity (e.g., within Oghuz or Kipchak branches), is associated with transfer efficiency.

Kazakh is the strongest transfer source on average. The Avg. column in Table 1 summarizes the overall strength of each language as a transfer source across all transfer targets. In BLEU, Kazakh has the highest average recovery with 42.22%, followed by Turkish, Uzbek, and Azerbaijani with similar averages around 39%. The same pattern appears in chrF, where Kazakh again has the highest average recovery with 72.73%, followed by Azerbaijani with 71.28% and Uzbek with 71.04%.

English has the lowest average in both metrics, with 30.97% BLEU and 63.57% chrF, showing that a high-resource non-Turkic transfer source is weaker on average than the Turkic transfer sources. Full recovery matrices for all targets and metrics are provided in Appendix A.3.1.

5.2 Translation Target Effect (RQ2)

Transfer depends on the translation target. Figure 1 shows that the same transfer source–transfer target pair can produce different recovery rates depending on the translation target. When Azerbaijani is the transfer source and Kyrgyz is the transfer target, recovery changes from 35.64% with Turkish as the translation target to 48.23% with Uzbek and 45.59% with Kazakh. Therefore, transfer cannot be described only as a fixed relationship between two source languages. The translation target also shapes how much knowledge transfers from one source language to another.

Transfer is directional and shaped by transfer-target–translation-target similarity. Another interpretation of Figure 1 is that transfer is directional: transferring from Azerbaijani to Kyrgyz is not equivalent to transferring from Kyrgyz to Azerbaijani under the same translation target. For instance, with Turkish as the translation target, az→ky recovery is 35.64%, whereas ky→az recovery is 52.49%. A similar asymmetry appears for az→kk and kk→az with Turkish as the translation target, where recovery increases from 32.76% to 51.01% in the reverse direction. These results suggest that recovery rate tends to be higher when the transfer target and the translation target belong to the same Turkic subgroup, either Oghuz or Kipchak.

A related study suggests that similar target languages can provide stronger positive transfer in one-to-many multilingual MT by using the same

translation source for the fine-tuning and evaluation pairs while varying their translation targets (Meng and Monz, 2024). Our setting differs from this design: we keep the translation target fixed and reverse the transfer source–transfer target direction. Therefore, our results provide a complementary view by showing that transfer is also affected by the relationship between the transfer target and the fixed translation target. Additional transfer source–transfer target pairs are provided in Appendix A.3.2

5.3 Latinization Effects on Transfer (RQ3)

Latinization is a secondary but important factor because the five languages differ in script usage. Turkish and Azerbaijani are written in Latin script, Kazakh and Kyrgyz are commonly written in Cyrillic, and Uzbek is mixed in practice. Therefore, original-script experiments combine linguistic transfer with orthographic mismatch, while Latinized experiments reduce that script mismatch and make surface overlap easier for the model to use (Sun et al., 2022).

Transfer	ΔBLEU (%)	ΔchrF (%)	ΔCOMET (%)
az→kk	+33.00	+30.81	-19.04
az→ky	+63.64	+46.08	-24.49
uz→kk	+41.03	+64.06	-24.91
uz→ky	+59.63	+40.79	-21.27
ky→kk	-21.01	-8.20	-11.06

Table 2: Relative effect of Latinization on allCPT XWMT scores with tr as the translation target. The Transfer column denotes transfer source → transfer target. Positive values indicate that Latinization improves the score over the original-script setting, while negative values indicate a decrease.

Latinization improves surface metrics for script-mismatched transfer targets. Table 2 reports percent changes from the original-script setting to the Latinized setting. The largest gains appear when the transfer target is Kazakh or Kyrgyz, which are written in Cyrillic in the original-script setting. For example, when Azerbaijani is the transfer source and Kyrgyz is the transfer target, Latinization improves BLEU by 63.64% and chrF by 46.08%. Similarly, when Uzbek is the transfer source and Kazakh is the transfer target, Latinization improves BLEU by 41.03% and chrF by 64.06%. These gains suggest that Latinization helps the model exploit surface-form similarities that are partly hidden when the transfer source and transfer target are represented in different scripts.

Latinization is not uniformly beneficial across settings and metrics. ky→kk row shows that Latinization does not always improve performance: BLEU decreases by 21.01% and chrF decreases by 8.20%. This indicates that Latinization is most useful when it reduces script mismatch between transfer languages. The same rows show negative COMET deltas, which means that better surface-form overlap does not always translate into better semantic metric scores (Purkayastha et al., 2023; Soni and Bhattacharyya, 2024). We therefore treat Latinization as a factor that reshapes the matrix, especially for BLEU and chrF, rather than as a uniform improvement across all evaluation criteria. Full Latinized-minus-original matrices are reported in Appendix A.3.3.

6 Ablation Studies

The previous section presented the main transfer pattern. In this section we collect correlation-based evidence showing that the transfer patterns we observe are not artifacts of a single dataset, CPT regime, or architecture. These analyses further support the reliability of the observed intra-Turkic transfer relations.

The transfer patterns are mostly stable across different test datasets. We first examine the stability of transfer source preferences across XWMT and Tatoeba test datasets. For each fixed evaluation pair $j \rightarrow t$, where j is the transfer target and t is the translation target, we compare all models that were fine-tuned toward the same translation target language t but with different transfer source languages. Each transfer source i defines one candidate donor model, fine-tuned on $i \rightarrow t$ and evaluated on the same pair $j \rightarrow t$. The score obtained on this evaluation pair is treated as the donor score for language i.

For example, for the evaluation pair az→tr, we compare models fine-tuned on az→tr, uz→tr, kk→tr, and ky→tr, all evaluated on az→tr. The donor with the highest score is selected as the best donor for that evaluation pair. Repeating this procedure for every evaluation pair gives a donor ranking for each metric and dataset.

We compare the donor rankings obtained from XWMT and Tatoeba using two criteria. *Same best donor* means that the highest-scoring donor is identical on both datasets for the same evaluation pair. *Same full order* means that the complete ranking of donors is identical across the two datasets. Since there are five target languages and four non-target evaluation sources for each target, each metric is compared over 20 donors.

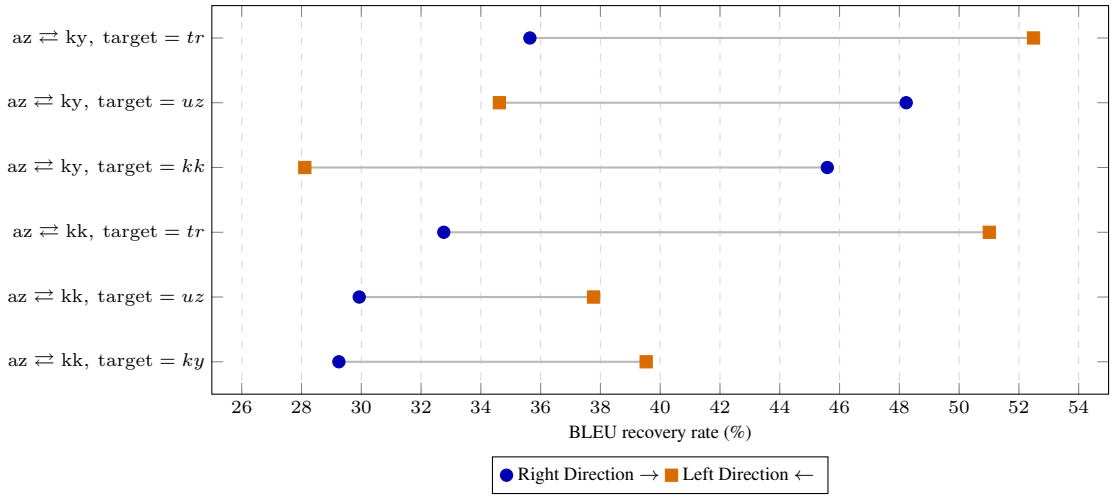


Figure 1: Paired BLEU recovery-rate comparison for mT5 Latin. Each y-axis label shows a pair of possible transfer source–transfer target directions under a fixed translation target. Direction → denotes transfer from the left language to the right language in the y-axis label, while Direction ← denotes the reverse transfer direction. In the first row, right direction measures az as the transfer source and ky as the transfer target with tr as the translation target The left direction measures the reverse direction, where ky is the transfer source and az is the transfer target.

Setting	Metric	Same best donor	Same full order
Orig.	BLEU	20/20	13/20
Orig.	chrF	20/20	17/20
Orig.	COMET	9/20	6/20
Latin	BLEU	19/20	11/20
Latin	chrF	20/20	14/20
Latin	COMET	17/20	11/20

Table 3: Donor-ranking stability between XWMT and Tatoeba for noCPT mT5. "Same best donor" means that the top-ranked fine-tuning donor is the same on both datasets for a given target and evaluation source. "Same full order" means that the complete donor ranking is identical across the two datasets.

Table 3 shows that BLEU and chrF donor rankings are highly stable across datasets. In the original-script setting, both metrics select the same best donor in all 20 columns; in the Latinized setting, chrF remains perfect and BLEU differs in only one column. Full-order stability is lower, but still strong for BLEU and chrF. COMET is less consistent, especially in the original-script setting, where it preserves the same best donor in only 9/20 columns. Therefore, we treat BLEU and chrF as the main signals for donor-order stability, while COMET serves as complementary evidence. Even though absolute scores differ across XWMT and Tatoeba, the highest-ranked donor usually remains the same for BLEU and chrF. This supports the view that the observed transfer structure is not specific to a single test set, but reflects a more general intra-family transfer pattern among the Turkic languages. Detailed correlation results are provided in Appendix A.3.4.

Transfer scores remain mostly correlated across back-translated training datasets. We test the effect of the back-translation model used to create the fine-tuning data on the observed transfer patterns. We select az and kk as translation targets. For this analysis, we compare model results obtained from the original NLLB-generated fine-tuning data with results obtained from a second fine-tuning dataset generated using `google/madlad400-3b-mt` (Kudugunta et al., 2023).

Test set	Target	Metric	r	Best donor
XWMT	az	BLEU	0.884	4/4
XWMT	az	chrF	0.772	4/4
XWMT	kk	BLEU	0.796	4/4
XWMT	kk	chrF	0.843	4/4
Tatoeba	az	BLEU	0.740	4/4
Tatoeba	az	chrF	0.500	4/4
Tatoeba	kk	BLEU	0.601	4/4
Tatoeba	kk	chrF	0.744	3/4

Table 4: Stability between NLLB-based and MADLAD-based fine-tuning data for translation targets az and kk. r denotes Pearson correlation computed over all donor×evaluation cells for each target-specific matrix. Best donor reports how often the same top-scoring transfer source is selected in both settings.

Table 4 shows positive Pearson correlations between NLLB-based and MADLAD-based fine-tuning matrices in all BLEU and chrF settings. The correlations are strongest on XWMT, ranging from 0.772 to 0.884, which suggests that the relative transfer scores are largely preserved when the back-translation model changes. The correlations on Tatoeba are lower, especially for chrF with Azerbaijani as the translation target, but they remain positive. This indicates that the exact score distribution is affected by the back-translation model and evaluation set, while the overall transfer pat-

tern remains partially consistent. The best-donor agreement gives a similar but coarser signal: the top donor is preserved in almost all cases, with the only one exception. These results further support the generalizability of our findings, while future validation on human-translated data from broader domains would provide an even stronger basis for generalization. Other correlation metrics produced results that were similar to, or lower than, those reported here. This suggests that there is an underlying transfer structure, but also that bias from the back-translation model might affect the results.

The transfer patterns are largely similar with or without CPT. We then examine how continual pretraining affects the structure of cross-source transfer. If CPT substantially changed the transfer behavior, we would expect the noCPT and allCPT settings to produce very different off-diagonal recovery patterns. Instead, Table 5 shows high off-diagonal correlations between the noCPT and allCPT settings.

Comparison	Metric	r_{off}	Best donor
allCPT–noCPT	BLEU	0.967	16/20
allCPT–noCPT	chrF	0.915	17/20
allCPT–noCPT	COMET	0.964	11/20
allCPT–noCPT	COMETKiwi	0.977	15/20

Table 5: Stability of cross-source transfer patterns on Latinized XWMT. r_{off} denotes off-diagonal Pearson correlation and best donor denotes same best donor. High off-diagonal correlation means that the relative pattern of zero-shot transfer is largely preserved.

The high off-diagonal correlations in Table 5 suggest that CPT does not create a completely new transfer map. For example, allCPT–noCPT reaches $r_{\text{off}} = 0.967$ for BLEU and 0.915 for chrF. This means that CPT can change absolute scores, but the relative organization of transfer remains largely stable. Detailed allCPT-minus-noCPT matrices are provided in Appendix A.3.5.

The transfer patterns are stable across encoder-decoder and decoder-only architectures. Finally, we compare the Latinized mT5 matrix with an auxiliary Qwen3 0.6B experiment. This tests whether the raw transfer structure is specific to an encoder–decoder model or whether a similar pattern also appears in a decoder-only model.

Table 6 shows strong alignment between mT5 and Qwen3 on raw transfer matrices. For all raw cells, Pearson correlation reaches 0.953 for BLEU and 0.929 for chrF; for off-diagonal raw cells, it remains 0.883 for BLEU and 0.820 for chrF. Recovery rate correlations are weaker but still positive.

Level	Metric	r	ρ
All raw cells (80)	BLEU	0.953	0.897
All raw cells (80)	chrF	0.929	0.896
Off-diagonal raw (60)	BLEU	0.883	0.784
Off-diagonal raw (60)	chrF	0.820	0.775
Recovery rate (20)	BLEU	0.649	0.522
Recovery rate (20)	chrF	0.529	0.439

Table 6: mT5 Latin vs Qwen3 Latin correlations on Latinized XWMT. r and ρ denote Pearson and Spearman correlation coefficients, respectively. Raw donor×evaluation matrices are strongly aligned; recovery summaries are less aligned but still moderately correlated.

This suggests that architecture changes may rescale or smooth the transfer signal, but they do not erase the family-internal structure observed in the main mT5 matrices. Full Qwen3 recovery tables are reported in Appendix A.4.

7 Conclusion

This paper examines cross-lingual transfer among Turkish, Azerbaijani, Uzbek, Kazakh, and Kyrgyz through fixed-target transfer matrices. The results show that transfer within the Turkic family depends on both the transfer direction and the translation target. The strongest recovery patterns are concentrated around the Turkish–Azerbaijani and Kazakh–Kyrgyz blocks. At the same time, the same transfer source does not behave uniformly for every transfer target.

Latinization provides a second source of evidence for this interpretation. In script-mismatched directions, especially those involving Kazakh and Kyrgyz, Latinization often improves BLEU and chrF, while these gains are not always mirrored by COMET. This indicates that surface-level improvements and semantic-quality estimates can diverge. Therefore, Latinization should be interpreted as a factor that reshapes transfer patterns rather than as a uniform improvement strategy.

The stability analyses further support the reliability of the observed structure. Donor preferences are highly consistent between XWMT and Tatoeba for BLEU and chrF, and the off-diagonal correlations across noCPT and allCPT show that CPT mostly preserves the transfer pattern even when absolute scores change. The auxiliary Qwen3 experiment leads to a similar conclusion: raw donor–evaluation matrices are strongly correlated with the mT5 results, although recovery summaries are less aligned. Overall, these findings suggest that cross-lingual transfer is a multi-factor phenomenon shaped by the transfer source, transfer target, translation target, script representation, and model setting.

Limitations

The main limitation of this study is the use of mostly back-translated bilingual pairs for fine-tuning. This choice was necessary because high-quality direct parallel data between low-resource Turkic languages is limited, and manually creating a balanced human-translated dataset was outside the scope of this work. However, synthetic data can introduce artifacts from the back-translation model and may affect the absolute scores reported in the transfer matrices. For this reason, our claims focus on relative donor preferences, matrix structure, and stability under a fixed experimental pipeline, rather than on absolute translation quality.

A second limitation is that the analysis is restricted to five Turkic languages. The observed correlations suggest that the transfer structure is systematic within this family, but the same conclusions may not directly generalize to other language families with different typological, script, or resource profiles. Repeating the same fixed-target analysis on other low-resource families and on more human-translated data would be an important direction for future work.

Ethical Considerations

This work aims to support research on low-resource Turkic machine translation, but the resulting models should not be treated as production-ready systems. Because much of the fine-tuning data is synthetic or automatically mined, translations may contain errors, artifacts, or biases inherited from the back-translation and filtering pipeline. These risks are especially important in high-stakes domains such as legal, medical, or governmental translation.

We use publicly available or automatically generated data and do not intentionally include private or personally identifiable information. However, web-derived corpora may still contain noisy or biased content. We therefore recommend that any released models or datasets be used primarily for research, accompanied by documentation of their construction process and limitations, and carefully evaluated before deployment, especially for minority-language communities where translation errors or uneven data coverage may have disproportionate effects.

Licensing and Copyright. For any released data, models, or scripts, we will preserve the required attribution information and release only materials that can be redistributed under the corresponding source licenses. Our code and model releases will be distributed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0, while data releases will follow the licenses of the original sources. When a source does not permit redistribution of the original text, we will provide the processing scripts or derived metadata instead of the restricted content. This is intended to support reproducibility while respecting the intellectual property rights of the original data providers and authors.

Reproducibility. To support reproducibility, we release the dataset construction scripts, Latinization rules, fine-tuning and evaluation scripts, and result matrices used in our experiments. The repository also includes the configuration files and instructions needed to reproduce the main transfer-matrix results. All reported findings should be interpreted with respect to our specific experimental setup, including the selected models, preprocessing pipeline, decoding parameters, evaluation datasets, and metric implementations.

Use of Generative AI. Generative AI tools were used only to assist with language editing and improving the clarity of the manuscript. All scientific contributions, experimental design, data construction decisions, result analysis, and interpretations were conducted and verified by the authors.

References

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a cleaner document-oriented multilingual crawled corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4344–4355, Marseille, France. European Language Resources Association.

Belen Alastruey, João Maria Janeiro, Alexandre Allauzen, Maha Elbayad, Loïc Barrault, and Marta R. Costa-jussà. 2025. Interference matrix: Quantifying cross-lingual interference in transformer encoders. *arXiv preprint arXiv:2508.02256*.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. In *Transactions of the Association for Computational Linguistics*.

Mikko Aulamo, Nikolay Bogoychev, Shaoxiong Ji, Graeme Nail, Gema Ramírez-Sánchez, Jörg Tiedemann, Jelmer van der Linde, and Jaume Zaragoza. 2023. HPLT: High performance language technologies. In *Proceedings of the 24th Annual Conference*

of the European Association for Machine Translation, pages 517–518, Tampere, Finland. European Association for Machine Translation.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations*.

Terra Blevins, Tomasz Limisiewicz, Suchin Gururangan, Margaret Li, Hila Gonen, Noah A. Smith, and Luke Zettlemoyer. 2024. Breaking the curse of multilinguality with cross-lingual expert language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10822–10837, Miami, Florida, USA. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K. Bergen. 2024. When is multilinguality a curse? language modeling for 250 high- and low-resource languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4074–4096, Miami, Florida, USA. Association for Computational Linguistics.

Everlyn Asiko Chimoto and Bruce A. Bassett. 2022. Very low resource sentence alignment: Luhya and Swahili. In *Proceedings of the Fifth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2022)*, pages 1–8, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Francisco Guzmán, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Ona de Gibert, Joseph Attieh, Teemu Vahtola, Mikko Aulamo, Zihao Li, Raúl Vázquez, Tiancheng Hu, and Jörg Tiedemann. 2025. Scaling low-resource MT via synthetic data generation with LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27674–27692, Suzhou, China. Association for Computational Linguistics.

Juuso Eronen, Michal Ptaszynski, Karol Nowakowski, and Zheng Lin Chia. 2023. Improving Polish to English neural machine translation with transfer learning: Effects of data volume and language similarity. In *Proceedings of the 1st International Workshop on Multilingual, Multimodal and Multitask Language Generation*, pages 40–46, Tampere, Finland. European Association for Machine Translation.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, and 1 others. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).

Isai Gordeev, Sergey Kuldin, and David Dale. 2024. FLORES+ translation and machine translation evaluation for the Erzya language. In *Proceedings of the Ninth Conference on Machine Translation*, pages 614–623, Miami, Florida, USA. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Yunsu Kim, Petre Petrov, Pavel Petrushkov, Shahram Khadivi, and Hermann Ney. 2019. Pivot-based transfer learning for neural machine translation between non-English languages. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 866–876, Hong Kong, China. Association for Computational Linguistics.

Tom Kocmi. 2020. Exploring benefits of transfer learning in neural machine translation. *arXiv preprint arXiv:2001.01622*.

Tom Kocmi and Ondřej Bojar. 2018. Trivial transfer learning for low-resource neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 244–252, Brussels, Belgium. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023a. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023b. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Sneha Kudugunta and 1 others. 2023. Madlad-400: A multilingual and document-level large audited dataset. *arXiv preprint arXiv:2309.04662*.

Danni Liu and Jan Niehues. 2025. Conditions for catastrophic forgetting in multilingual translation. In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 347–359, Suzhuo, China. Association for Computational Linguistics.

Shayne Longpre, Sneha Kudugunta, Niklas Muennighoff, I-Hung Hsu, Isaac Caswell, Alex Pentland, Sercan Arik, Chen-Yu Lee, and Sayna Ebrahimi. 2025. Atlas: Adaptive transfer scaling laws for multilingual pretraining, finetuning, and decoding the curse of multilinguality. *arXiv preprint arXiv:2510.22037*.

Kaiwen Lu, Yating Yang, Fengyi Yang, Rui Dong, Bo Ma, Aihetamujiang Aihemaiti, Abibilla Atawulla, Lei Wang, and Xi Zhou. 2025. Low-resource language expansion and translation capacity enhancement for LLM: A study on the Uyghur. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8360–8373, Abu Dhabi, UAE. Association for Computational Linguistics.

Yingfeng Luo, Ziqiang Xu, Yuxuan Ouyang, Murun Yang, Dingyang Lin, Kaiyan Chang, Tong Zheng, Bei Li, Peinan Feng, Quan Du, Tong Xiao, and Jingbo Zhu. 2025. Beyond english: Toward inclusive and scalable multilingual machine translation with llms. *arXiv preprint arXiv:2511.07003*.

Yan Meng and Christof Monz. 2024. Disentangling the roles of target-side transfer and regularization in multilingual machine translation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1828–1840, St. Julian's, Malta. Association for Computational Linguistics.

Jamshidbek Mirzakhalov, Anoop Babu, Duygu Ataman, Sherzod Kariev, Francis Tyers, Otabek Abduraufov, Mammad Hajili, Sardana Ivanova, Abror Khaytbaev, Antonio Laverghetta Jr., Bekhzodbek Moydinboyev, Esra Onal, Shaxnoza Pulatova, Ahsan Wahab, Orhan Firat, and Sriram Chellappan. 2021. A large-scale study of machine translation in Turkic languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5876–5890, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Toan Q. Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 296–301, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Sukannya Purkayastha, Sebastian Ruder, Jonas Pfeiffer, Iryna Gurevych, and Ivan Vulić. 2023. Romanization-based large-scale adaptation of multilingual language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7996–8005, Singapore. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):5485–5551.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti,

José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Lütfi Kerem Senel, Benedikt Ebing, Konul Baghirova, Hinrich Schuetze, and Goran Glavaš. 2024. Kardeş-NLU: Transfer to low-resource languages with the help of a high-resource cousin – a benchmark and evaluation for Turkic languages. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1672–1688, St. Julian's, Malta. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Govind Soni and Pushpak Bhattacharyya. 2024. RoMantra: Optimizing neural machine translation for low-resource languages through Romanization. In *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*, pages 157–168, AU-KBC Research Centre, Chennai, India. NLP Association of India (NLPAI).

David Stap, Vlad Niculae, and Christof Monz. 2023. Viewing knowledge transfer in multilingual machine translation through a representational lens. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14973–14987, Singapore. Association for Computational Linguistics.

Simeng Sun, Angela Fan, James Cross, Vishrav Chaudhary, Chau Tran, Philipp Koehn, and Francisco Guzmán. 2022. Alternative input signals ease transfer in multilingual machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5291–5305, Dublin, Ireland. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, pages 3104–3112.

Maali Tars, Andre Tättar, and Mark Fišel. 2021. Extremely low-resource machine translation for closely related languages. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 41–52, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Jörg Tiedemann. 2020. The tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Cagri Toraman. 2024. Adapting open-source generative large language models for low-resource languages: A case study for Turkish. In *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, pages 30–44, Miami, Florida, USA. Association for Computational Linguistics.

Cagri Toraman, Furkan Şahinuç, and Eyup Yilmaz. 2022. Large-scale hate speech detection with cross-domain transfer. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2215–2225, Marseille, France. European Language Resources Association.

Iulia Turc, Kenton Lee, Jacob Eisenstein, Ming-Wei Chang, and Kristina Toutanova. 2021. Revisiting the primacy of English in zero-shot cross-lingual transfer. *Preprint*, arXiv:2106.16171.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010.

Qian Wang and Jiajun Zhang. 2022. Addressing asymmetry in multilingual neural machine translation with fuzzy task clustering. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5129–5141, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Xinyi Wang and Graham Neubig. 2019. Target conditioned sampling: Optimizing data selection for multilingual neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5823–5828, Florence, Italy. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Bilge Kağan Yazar and Erdal Kiliç. 2025. Improving low-resource Kazakh-English and Turkish-English neural machine translation using transfer learning and part of speech tags. *IEEE Access*, 13:32341–32356.

Rustem Yeshpanov, Alina Polonskaya, and Huseyin Atakan Varol. 2024. KazParC: Kazakh parallel corpus for machine translation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9633–9644, Torino, Italia. ELRA and ICCL.

Brian Yu, Hansen Lillemark, and Kurt Keutzer. 2023. Simple and effective input reformulations for translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10322–10334, Singapore. Association for Computational Linguistics.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.

A Appendix

A.1 Dataset Creation Details

A.1.1 Monolingual Corpus

Large-scale monolingual corpora were constructed for five Turkic languages (Turkish, Azerbaijani, Kazakh, Kyrgyz, and Uzbek) by collecting data from publicly available resources, including Wikipedia, CC-100 (Conneau et al., 2020), Leipzig Corpora (Goldhahn et al., 2012), OSCAR (Abadji et al., 2022), mC4 (Xue et al., 2021), MADLAD-400 (Kudugunta et al., 2023), and HPLT (Aulamo et al., 2023). Additional language-specific datasets were incorporated where available. The collected raw text was cleaned through line-level filtering (removal of short, numeric, or noisy lines), paragraph reconstruction, and MD5-based deduplication. The final datasets were stored in JSONL format for each language. The amount of data collected for each language is shown in Table 7. For continual pre-training (CPT), the data was segmented into approximately 480-token chunks using the `google/mt5-small` tokenizer (Xue et al., 2021), while preserving sentence boundaries. Cyrillic-based languages were transliterated into their official Latin alphabet versions.

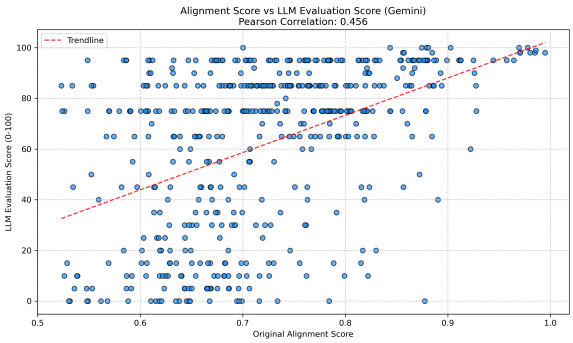
Language	Number of Chunks	File Size
Turkish	69,848,795	102GB
Azerbaijani	16,960,000	25GB
Kazakh	6,638,029	27GB
Kyrgyz	3,710,962	13GB
Uzbek	8,265,191	17GB

Table 7: Distribution of Monolingual Data

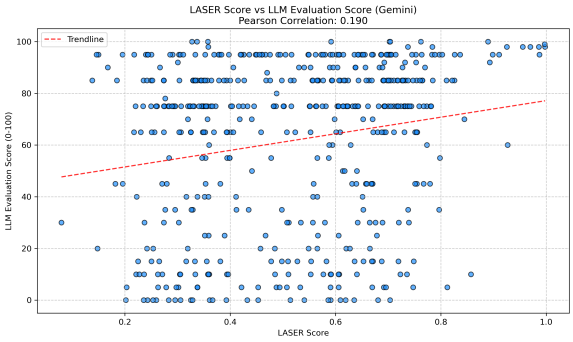
A.1.2 Bilingual Corpus

Parallel corpora were obtained from three main sources: pivot-based extraction on publicly available datasets, back-translation, and open-source books available in multiple Turkic languages. Direct and pivot-based bilingual pairs were obtained from KazParC (Yeshpanov et al., 2024), NTREX (Federmann et al., 2022), and FLORES+ (Gordeev et al., 2024), OPUS (Tiedemann, 2012) including corpora such as OpenSubtitles, TED2020, and CCAligned. All parallel data underwent multi-stage filtering, including language detection via FastText (Bojanowski et al., 2017), script normalization, removal of noisy or misaligned pairs, length ratio constraints, and deduplication. Since the collected data were highly imbalanced and contained few direct pairs between low-resource Turkic languages, synthetic parallel data was generated via back-translation using `facebook/nllb-200-distilled-600M` (Costa-jussà et al., 2022), with quality filtering based on LaBSE (Feng et al., 2022) similarity scores and LLM-as-a-judge evaluation method. Additionally, parallel data was extracted from multilingual book translations using LaBSE-based bitext mining, applying both paragraph- and sentence-level alignment with cosine similarity thresholds. Comparative evaluation against LASER (Artetxe and Schwenk, 2019) using an LLM-based scoring framework (Kocmi and Federmann, 2023b) demonstrated stronger correlation for LaBSE, which was therefore used for semantic filtering (Figure 2). These results also correlate with the work by Chimoto and Bassett (2022) which stated that LaBSE works better than LASER for low-resource languages with a study on African languages. The prompt given to our judge LLM (Gemini 3.1 Pro) to score the quality of the machine translation is taken from (Kocmi and Federmann, 2023b) and can be found below.

LLM Evaluation Prompt *Score the following translation from {Source Language} to {Target Language} on a continuous scale from 0 to 100, where a score of zero means "no meaning preserved" and a score of one hundred means "perfect meaning and grammar".*
{Source Language} source: {Source Sentence}
{Target Language} translation: {Translated Sentence}
Score:



(a) LaBSE Score vs. LLM Score



(b) LASER Score vs. LLM Score

Figure 2: Comparison of Dataset Filtering Metrics

In this work, we used pairs from publicly available datasets for CPT in transfer-coefficient sampling experiments, and we combined back-translated data with book-aligned pairs to create the FT data for transfer-matrix calculations. We used all aligned book pairs and completed the FT dataset size to 100K pairs for each language using back-translation (de Gibert et al., 2025), following the 100K training-size setting used by Eronen et al. (2023). The number of manually curated book-aligned pairs among Turkic languages is shown in Table 8. For inference / test data we used a subset of `turkic-interlingua/turkic_xwmt` test dataset (Mirzakhalov et al., 2021) which contains 400 pairs between each language pairs and as the secondary test set for correlation experiments we used a subset of Tatoeba dataset (Tiedemann, 2020) which also contains 400 pairs for each language pair.

Language Pair (Source-Target)	Total
tr-kk	7605
tr-az	6982
az-uz	2104
tr-uz	1999
kk-ky	275
tr-ky	255
Total	**19220**

Table 8: Distribution of Book Pairs between Languages

A.2 Implementation Details

A.2.1 CPT and FT Scripts for mT5-small

The experiments use `mT5-small` (300M parameter) as the base sequence-to-sequence model. Continual pretraining follows the T5 span-corruption objective: each monolingual chunk is tokenized, random spans are replaced with sentinel tokens, and the target sequence consists of the removed spans. The same CPT hyperparameters are used across all CPT scripts, as reported in Table 9.

Fine-tuning uses a fixed-translation-target transfer design. For each translation target language t, the model is fine-tuned on one translation source–target pair $i \rightarrow t$ and evaluated on all source languages $j \rightarrow t$ for transfer target $j \neq t$. Inputs are formatted as explicit source and target tags followed by the source sentence. The same fine-tuning and decoding hyperparameters are used across the FT scripts, and these shared settings are reported in Table 10. Because GPU resources were limited, we did not perform an exhaustive hyperparameter search and instead adopted the learning-rate setting from Yu et al. (2023).

For all CPT and FT experiments, a single Nvidia A100 80GB GPU is used via Google Cloud Console Virtual Machine Instances. Fine-tuning mT5-small model with a single translation source–target pair $i \rightarrow t$ took 1 hour 40 minutes to 2 hours and CPT with 5M total chunks each of approximately 512 tokens took 5 to 6 hours.

A.2.2 FT and Evaluation Scripts for Qwen3 0.6B

To test whether the transfer structure observed with mT5 is specific to an encoder–decoder architecture or not, we also run an auxiliary decoder-only experiment with `Qwen/Qwen3-0.6B`. The Qwen3 experiment follows the same fixed-translation-target transfer logic as the mT5 transfer-matrix experiments: for each pivot translation target language t, the model is fine-tuned on one donor direction (both transfer and translation source) $i \rightarrow t$ and then evaluated on all test directions $j \rightarrow t$ where transfer target $j \neq t$. The resulting matrices use fine-tuning donor (transfer source) languages as rows and evaluation source (transfer target) languages as columns.

Because Qwen3 is a causal language model, the translation input and target are concatenated into a single sequence. Each training example is formatted as a natural-language translation prompt

CPT setting	Value
Base architecture	`google/mt5-small` / local `mT5-small` copy
Objective	T5 span corruption
Maximum examples	1,000,000 per language for allCPT
Maximum token length	512
Noise density	0.15
Mean noise span length	3.0
Training epochs	1
Batch size	32
Gradient accumulation	1
Learning rate	5×10^{-5}
Weight decay	0.01
Warmup ratio	0.03
Optimizer	`adamw_torch`
Precision	bf16
Save policy	step-based saving with `save_total_limit=1`
Seed	42

Table 9: Shared CPT hyperparameters used across the CPT scripts.

Fine-tuning/evaluation setting	Value
Languages	`tr`, `az`, `uz`, `kk`, `ky`
Input format	`<src> <tgt> source sentence`
Training epochs	5.0
Train batch size	8
Eval batch size	8
Gradient accumulation	4
Learning rate	3×10^{-4}
Weight decay	0.01
Warmup ratio	0.1
Scheduler	linear
Validation split	0.05
Max source / target length	256 / 256
Generation	beam size 5, maximum length 256, early stopping
Maximum test samples	400
Metrics	sacreBLEU, chrF, COMET, COMETKiwi
COMET models	`Unbabel/wmt22-comet-da`, `Unbabel/wmt22-cometkiwi-da`
Seed	42

Table 10: Shared fine-tuning and evaluation hyperparameters used in the transfer-matrix scripts.

followed by the reference translation:

```
Translate from SourceLanguage to
TargetLanguage: source sentence \n
target sentence <eos>
```

The loss is computed only on the target/completion tokens. Prompt tokens are masked with `-100` in the label sequence, so that they do not contribute to the training objective. During generation, the same prompt format is used, and the prompt portion is removed from the decoded output before scoring. Implementation details can be seen on Table 11. Similar to mT5, we did not perform hyperparameter search and used the Qwen3 fine-tuning learning-rate from Luo et al. (2025).

For evaluation, the fine-tuned Qwen3 model is loaded once for each translation source–target run and reused across all evaluation sources (transfer targets) for that fixed translation target. The tokenizer is switched to left padding during generation, which is required for batched causal-LM decoding. Predictions are normalized for whitespace and punctuation before scoring with sacreBLEU and chrF. For each run, the script stores per-pair scores, transfer matrices, logs, and generated predictions in JSONL format. For Qwen3 FT and evaluation experiments dual Nvidia P100 16GB GPU is used via [anonymous due to blind submission]. Fine-tuning Qwen3 0.6B model with a single translation source–target pair $i \rightarrow t$ and obtaining the evaluation results took approximately 14 hours on this GPU setup.

A.2.3 Examples From Training and Test Datasets

The following examples illustrate the JSONL formats used by the scripts. Training files contain explicit source and target language fields, source and target texts, and metadata about pivoting or synthetic generation. Test files follow the common `translation` dictionary format used by

Qwen3 FT/Eval setting	Value
Base model	`Qwen/Qwen3-0.6B`
Model class	`AutoModelForCausalLM`
Tokenizer class	`AutoTokenizer`
Training framework	Hugging Face `Trainer` / `TrainingArguments`
Fine-tuning design	Fixed-translation-target translation source fine-tuning, evaluated for all transfer targets
Prompt format	`Translate from {SourceLanguage} to {TargetLanguage}: source\n`
Training sequence	Prompt + target + `eos`
Loss masking	Prompt labels set to `-100`; loss computed only on target tokens
Training epochs	3.0
Train batch size	4
Eval batch size	4
Gradient accumulation	8
Effective batch size	32
Learning rate	2×10^{-5}
Weight decay	0.01
Warmup ratio	0.1
Scheduler	cosine
Max gradient norm	1.0
Validation split	0.05
Max source / target length	256 / 256
Maximum generated tokens	256
Generation	beam size 5, early stopping, no sampling
Maximum test samples	400 per evaluation pair
Precision	fp16 disabled; bf16 disabled
Metrics	sacreBLEU, chrF
Seed	42

Table 11: Qwen3 fine-tuning and evaluation hyperparameters used in the auxiliary decoder-only transfer experiment.

XWMT/Tatoeba-style datasets.


```
{"src_lang":"az","tgt_lang":"tr","src":"Cengiz
    Hanin hokmranligi dovrunde 1206-1227-ci
    iller arasinda Simali Cinde Bati Xia ve Jin
    Hanedani; Turkistanda Kara Hitay,
    Maveraunnehir; Iranda Harezm, Horasan ve
    Harezmsahlar, Kafkasyada Gurculer, Dest-i
    Kipcakdaki Rus Knezleri, Kipcaklar ve Idil
    Bolqarlar uzerinde seferler", "tgt":"Cengiz
    Han, hukumdarligi doneminde, 1206-1227
    arasinda, Kuzey Çin'deki Bati Xia ve Jin
    Hanedani; Turkistan'daki Kara Hitay,
    Maveraunnehir; Harezm, Horasan ve Iran'daki
    Harezmsahlar, Kafkasya'daki Gurculer, Dest-i
     Kipcak'taki Rus Knezlikleri, Kipcaklar ile
    Idil Bulgarlari uzerine seferler yapti ve
    imparatorlugu doneminde gerceklestirdigi
    hicbir savasi kaybetmedi.", "idx":0, "pivot
    ":"tr", "nllb_model":"facebook/nllb-200-
    distilled-600M", "second":"none"}
```

```
{"translation":{"tr":"Kizim yurt disinda okuyor
    .","ky":"Kizim cet olkodo okuyt."}}
{"translation":{"tr":"Artik kime inanacagimi
    bilmiyorum.","ky":"Emi kimge isenerimdi
    bilbeym."}}
{"translation":{"tr":"Buyuk bir arabam var.","ky
    ":"Menin con masinam bar."}}
```


A.2.4 Example Model Outputs

Table 12 shows example outputs from the allCPT Latinized XWMT evaluation directory. All examples use Turkish as the fixed translation target and the model fine-tuned on `uz`→`tr`; only the evaluation source (transfer target) changes.

A.2.5 Reproducibility Statement

Our anonymous GitHub repository [2] contains:

- Dataset construction scripts;
- Latinization scripts and rules;
- CPT, fine-tuning and evaluation scripts;
- HTML files that contain the full result matrices and comparison summaries.

To reproduce the main mT5 transfer matrices, run the CPT checkpoint preparation step, fine-tune one model for each ordered source–target pair, and then run the fixed-target evaluation script with `MAX_TEST_SAMPLES=400`. For COMET and COMETKiwi, the Hugging Face token must allow access to the corresponding `Unbabel` checkpoints. All runs use seed 42.

[2] https://anonymous.4open.science/r/Cross-LingualTransferforMachineTranslationinTurkicLanguages-40D8/

Eval pair	Source	Reference	Prediction
az→tr	IB rəsmiləri əlavə olaraq qeyd edirlər ki, niyyət terrorçulara sular vasitəsilə sızmaq da olabilər.	Niyet, teroristlerin sudan sizdirilmasi olacagi, IB yetkilileri tarafından ek olarak not edildi.	IB resmileri iptal edilebilir ki , niyet terrorçulara sular vasitəsiyle sızmaq da olabilər.
kk→tr	Maqsat - lañkesterge su arqyly enu , dep qosty aqparattyq agenttik.	Niyet, teroristlerin sudan sizdirilmasi olacagi, IB yetkilileri tarafından ek olarak not edildi.	İlk olarak lañkesterge su aracılığıyla enu , olarak kurduğu amatör agenttik.
ky→tr	"Indiyanın çalgındoo byurosu bildirgendey, terroristter deñizge kirüü ıkmaların üyrötüp jatışat.	Niyet, teroristlerin sudan sizdirilmasi olacagi, IB yetkilileri tarafından ek olarak not edildi.	Ülkede ikinci olarak, İran'ın çalgındoo bürosu bildirdiği , terroristler denize girdik ıkmaların üyrütüp gitti.
uz→tr	Shuningdek , maqsad terroristlar orasiga suv orqali kirib borish deya Axborot agentligi rasmiylari qoshimcha qiladi.	Niyet, teroristlerin sudan sizdirilmasi olacagi, IB yetkilileri tarafından ek olarak not edildi.	Ayrıca amaç terroristler arasında su aracılığıyla erişilmesi olduğuna Güvenlik Enstitüsü resmî resmi sayısı dahil olmak üzere işaret eder.

Table 12: Example generated predictions from the allCPT Latinized XWMT output directory from uz→tr FT.

A.3 Detailed mT5 Results as Complementary Data

A.3.1 Latinized allCPT XWMT Recovery Tables

Figure 3 reports the full Latinized allCPT XWMT recovery matrices used to support the main transfer analysis. The rows of panels correspond to fixed translation targets, while the metric columns show BLEU, chrF, COMET, and COMETKiwi. Within each panel, diagonal cells report the raw same-source score for the corresponding fine-tuning/evaluation pair, and off-diagonal cells report the recovery rate obtained when the fine-tuning source differs from the evaluation source. This organization preserves the complete numerical evidence behind the averaged transfer matrix in the main paper while making the target-conditioned structure visible: strong transfer is concentrated around related-language blocks, but the same donor can behave differently depending on the fixed translation target and evaluation source.

A.3.2 Translation Target Effect

Figure 4 extends the target-effect analysis from the main paper by showing additional bidirectional transfer comparisons under different fixed translation targets. Each line connects two opposite transfer directions for the same language pair and translation target. Large gaps between the two markers indicate directional asymmetry, while changes across rows show that the same transfer pair can behave differently when the translation target changes. These patterns support the main claim that cross-lingual transfer among Turkic languages is target-conditioned rather than reducible to a single donor–recipient similarity score.

A.3.3 Nominal Deltas for Latinized - Original XWMT test dataset Raw Scores on allCPT

Figure 5 reports the raw-score changes obtained by replacing original-script inputs with Latinized inputs in the allCPT XWMT setting. Positive values indicate that Latinization improves the corresponding fine-tuning/evaluation cell, while negative values indicate a decrease. The figure complements the discussion in Section 5.3 by showing that Latinization does not act as a uniform improvement mechanism. Instead, its effect is concentrated in script-sensitive directions, especially those involving Kazakh and Kyrgyz, and the magnitude of the change differs substantially across metrics.

A.3.4 XWMT and Tatoeba test dataset Correlation Details

Tables 13–15 provide the detailed evidence behind the cross-dataset stability analysis. Table 13 reports raw-cell Pearson and Spearman correlations between the XWMT and Tatoeba matrices, while Table 14 summarizes whether the best donor and complete donor ordering are preserved across datasets. Table 15 gives the corresponding target-level donor-order comparisons for BLEU and chrF. Together, these results show that absolute scores may vary between benchmarks, but the donor preference structure is largely stable for surface metrics, especially BLEU and chrF.

A.3.5 Nominal Deltas for noCPT-allCPT Raw Scores

Figure 6 compares the Latinized XWMT raw-score matrices obtained with allCPT against the corresponding noCPT matrices. Positive values indicate that continual pretraining on all five Turkic languages improves a cell, while negative values indicate that the noCPT model performs better. The comparison shows that allCPT is not uniformly beneficial. The largest decreases often appear on diagonal supervised settings, including BLEU drops

Setting	Metric	Target	N cells	Pearson r	Spearman rho
Original	BLEU	tr	16.00	0.9659	0.9324
Original	BLEU	az	9.00	-0.0428	-0.3000
Original	BLEU	uz	9.00	0.2949	0.2667
Original	BLEU	kk	9.00	0.7480	0.7667
Original	BLEU	ky	9.00	0.9542	0.8833
Original	BLEU	overall	52.00	0.2877	0.3401
Original	chrF	tr	16.00	0.9878	0.9882
Original	chrF	az	16.00	0.8445	0.9059
Original	chrF	uz	9.00	0.2097	0.1500
Original	chrF	kk	9.00	0.0188	0.3333
Original	chrF	ky	9.00	0.9340	0.8167
Original	chrF	overall	59.00	0.3037	0.2776
Original	COMET	tr	16.00	0.8236	0.7618
Original	COMET	az	16.00	0.4199	0.3853
Original	COMET	uz	16.00	0.4455	0.5294
Original	COMET	kk	9.00	0.5487	0.5000
Original	COMET	ky	9.00	0.9048	0.8787
Original	COMET	overall	66.00	0.2453	0.2467
Latinized	BLEU	tr	16.00	0.9515	0.8529
Latinized	BLEU	az	9.00	-0.0227	-0.0667
Latinized	BLEU	uz	9.00	0.0985	0.0333
Latinized	BLEU	kk	9.00	0.8810	0.9333
Latinized	BLEU	ky	9.00	0.9706	0.8333
Latinized	BLEU	overall	52.00	0.1792	0.2706
Latinized	chrF	tr	16.00	0.9776	0.9647
Latinized	chrF	az	16.00	0.8847	0.9176
Latinized	chrF	uz	9.00	-0.3764	-0.2510
Latinized	chrF	kk	9.00	0.3287	0.2833
Latinized	chrF	ky	9.00	0.8956	0.7833
Latinized	chrF	overall	59.00	0.1598	0.1403
Latinized	COMET	tr	16.00	0.9480	0.9500
Latinized	COMET	az	16.00	0.5940	0.7412
Latinized	COMET	uz	16.00	0.4127	0.5706
Latinized	COMET	kk	9.00	0.2515	0.4333
Latinized	COMET	ky	9.00	-0.2085	-0.0667
Latinized	COMET	overall	66.00	0.2084	0.2733

Table 13: XWMT–Tatoeba raw-cell correlation coefficients.

Setting	Metric	Compared columns	Same best donor	Same full order	Same self-transfer best donor	Interpretation
Original	BLEU	20.00	20/20	13/20	20/20	Very stable; suitable for coefficient estimation
Original	chrF	20.00	20/20	17/20	20/20	Very stable; suitable for coefficient estimation
Original	COMET	20.00	9/20	6/20	5/20	Unstable; use only as weak/secondary signal
Latinized	BLEU	20.00	19/20	11/20	19/20	Very stable; suitable for coefficient estimation
Latinized	chrF	20.00	20/20	14/20	20/20	Very stable; suitable for coefficient estimation
Latinized	COMET	20.00	17/20	11/20	16/20	Less stable; use only as weak/secondary signal

Table 14: XWMT–Tatoeba donor-ranking stability summary.

Setting	Metric	Target	Eval source	XWMT donor order	Tatoeba donor order	Verdict
Latinized	BLEU	tr	az	az (13.43) > uz (7.48) > ky (7.05) > kk (6.85)	az (14.19) > uz (8.98) > ky (5.35) > kk (5.27)	same full order
Latinized	BLEU	tr	uz	uz (6.29) > ky (2.90) > az (2.84) > kk (2.53)	uz (7.76) > ky (1.50) > az (1.29) > kk (1.16)	same full order
Latinized	BLEU	tr	kk	kk (4.06) > ky (1.88) > uz (1.65) > az (1.33)	ky (3.31) > kk (2.57) > uz (1.59) > az (1.15)	different best donor
Latinized	BLEU	tr	ky	ky (4.04) > kk (1.96) > uz (1.74) > az (1.44)	ky (2.15) > kk (1.39) > uz (0.71) > az (0.54)	same full order
Latinized	BLEU	az	tr	tr (12.86) > ky (4.61) > kk (4.60) > uz (4.36)	tr (12.25) > ky (5.90) > uz (4.46) > kk (4.37)	same best donor
Latinized	BLEU	az	uz	uz (9.17) > ky (4.27) > kk (3.94) > tr (3.12)	uz (3.54) > ky (1.49) > tr (1.21) > kk (1.05)	same best donor
Latinized	BLEU	az	kk	kk (5.50) > ky (2.58) > uz (1.72) > tr (1.34)	kk (4.89) > ky (2.77) > uz (1.53) > tr (1.45)	same full order
Latinized	BLEU	az	ky	ky (5.32) > kk (2.43) > uz (2.28) > tr (1.79)	ky (3.61) > kk (1.99) > uz (1.50) > tr (1.16)	same full order
Latinized	BLEU	uz	tr	tr (5.59) > az (2.23) > kk (2.15) > ky (1.92)	tr (4.66) > az (1.93) > ky (0.61) > kk (0.48)	same best donor
Latinized	BLEU	uz	az	az (6.99) > tr (3.53) > kk (2.64) > ky (2.42)	az (3.47) > tr (1.68) > ky (0.71) > kk (0.61)	same best donor
Latinized	BLEU	uz	kk	kk (6.95) > az (2.08) > tr (1.87) > ky (1.72)	kk (6.62) > tr (1.67) > ky (1.66) > az (1.59)	same best donor
Latinized	BLEU	uz	ky	ky (3.96) > kk (2.09) > az (1.91) > tr (1.53)	ky (3.85) > kk (1.87) > az (1.77) > tr (1.49)	same full order
Latinized	BLEU	kk	tr	tr (3.94) > az (1.50) > ky (1.08) > uz (0.99)	tr (4.30) > az (3.65) > ky (0.63) > uz (0.51)	same full order
Latinized	BLEU	kk	az	az (5.23) > tr (3.18) > uz (1.68) > ky (1.47)	az (4.58) > tr (2.47) > ky (1.82) > uz (1.49)	same best donor
Latinized	BLEU	kk	uz	uz (8.90) > az (3.58) > tr (3.16) > ky (2.92)	uz (6.55) > az (3.10) > tr (2.91) > ky (2.36)	same full order
Latinized	BLEU	kk	ky	ky (6.12) > az (2.79) > tr (2.67) > uz (2.54)	ky (3.85) > tr (1.41) > uz (1.30) > az (0.95)	same best donor
Latinized	BLEU	ky	tr	tr (3.44) > az (1.63) > uz (1.38) > kk (1.28)	tr (1.57) > az (0.95) > uz (0.37) > kk (0.34)	same full order
Latinized	BLEU	ky	az	az (4.68) > tr (2.82) > uz (2.15) > kk (1.85)	az (3.18) > tr (2.15) > uz (1.12) > kk (0.97)	same full order
Latinized	BLEU	ky	uz	uz (5.85) > az (2.17) > kk (2.15) > tr (2.08)	uz (5.24) > kk (2.41) > tr (2.36) > az (2.06)	same best donor
Latinized	BLEU	ky	kk	kk (6.70) > uz (2.66) > az (1.96) > tr (1.91)	kk (3.14) > uz (1.35) > az (1.33) > tr (0.93)	same full order
Latinized	chrF	tr	az	az (48.64) > uz (40.06) > kk (38.65) > ky (38.26)	az (42.86) > uz (35.10) > kk (32.36) > ky (31.57)	same full order
Latinized	chrF	tr	uz	uz (39.45) > az (29.86) > ky (29.68) > kk (29.66)	uz (33.90) > kk (22.40) > ky (22.31) > az (21.48)	same best donor
Latinized	chrF	tr	kk	kk (34.70) > ky (27.32) > uz (25.97) > az (24.67)	kk (24.47) > ky (20.40) > uz (18.59) > az (16.97)	same full order
Latinized	chrF	tr	ky	ky (33.25) > kk (28.25) > uz (25.37) > az (24.41)	ky (22.50) > kk (21.62) > uz (19.44) > az (18.77)	same full order
Latinized	chrF	az	tr	tr (47.74) > kk (32.88) > ky (32.22) > uz (31.87)	tr (41.10) > ky (29.47) > uz (28.82) > kk (28.52)	same best donor
Latinized	chrF	az	uz	uz (43.77) > ky (32.83) > kk (32.52) > tr (31.02)	uz (32.44) > ky (24.87) > kk (24.28) > tr (23.43)	same full order
Latinized	chrF	az	kk	kk (38.07) > ky (28.58) > uz (25.42) > tr (22.96)	kk (34.35) > ky (27.34) > uz (23.66) > tr (22.26)	same full order
Latinized	chrF	az	ky	ky (36.32) > kk (29.26) > uz (24.65) > tr (24.41)	ky (33.68) > kk (27.37) > uz (23.21) > tr (23.09)	same full order
Latinized	chrF	uz	tr	tr (38.31) > az (26.77) > kk (23.68) > ky (23.27)	tr (30.43) > az (22.02) > kk (17.47) > ky (17.21)	same full order
Latinized	chrF	uz	az	az (41.58) > tr (31.28) > ky (27.50) > kk (27.41)	az (30.77) > tr (23.45) > kk (21.69) > ky (21.08)	same best donor
Latinized	chrF	uz	kk	kk (40.31) > az (23.62) > tr (22.05) > ky (21.27)	kk (38.79) > az (23.83) > tr (23.30) > ky (21.74)	same full order
Latinized	chrF	uz	ky	ky (34.83) > kk (25.81) > az (24.71) > tr (23.36)	ky (34.64) > kk (26.66) > az (24.93) > tr (23.63)	same full order
Latinized	chrF	kk	tr	tr (35.25) > az (24.31) > ky (21.47) > uz (21.16)	tr (28.20) > az (21.08) > ky (17.11) > uz (16.60)	same full order
Latinized	chrF	kk	az	az (39.18) > tr (32.02) > uz (25.64) > ky (25.50)	az (35.98) > tr (29.50) > ky (25.36) > uz (24.57)	same best donor
Latinized	chrF	kk	uz	uz (46.07) > az (33.56) > tr (31.93) > ky (30.88)	uz (40.46) > az (29.80) > tr (28.33) > ky (28.15)	same full order
Latinized	chrF	kk	ky	ky (41.58) > az (32.87) > uz (31.54) > tr (30.92)	ky (29.86) > uz (23.76) > az (23.59) > tr (22.15)	same best donor
Latinized	chrF	ky	tr	tr (31.62) > az (23.82) > uz (22.09) > kk (21.58)	tr (22.67) > az (20.17) > uz (18.56) > kk (18.26)	same full order
Latinized	chrF	ky	az	az (34.48) > tr (28.53) > uz (25.17) > kk (23.70)	az (34.98) > tr (28.51) > uz (24.86) > kk (23.49)	same full order
Latinized	chrF	ky	uz	uz (37.44) > az (26.56) > kk (26.42) > tr (26.33)	uz (38.02) > kk (27.63) > tr (26.62) > az (26.51)	same best donor
Latinized	chrF	ky	kk	kk (39.41) > uz (28.91) > az (27.01) > tr (25.15)	kk (27.55) > uz (22.96) > az (21.96) > tr (19.90)	same full order

Table 15: Detailed XWMT–Tatoeba donor-order comparisons for the Latinized mT5 setting using BLEU and chrF.

tr

BLEU	az-tr	uz-tr	kk-tr	ky-tr
az-tr	13.43	45.2%	32.8%	35.6%
uz-tr	55.7%	6.29	40.6%	43.1%
kk-tr	51.0%	40.2%	4.06	48.5%
ky-tr	52.5%	46.1%	46.3%	4.04
Avg.	53.1%	43.8%	39.9%	42.4%

chrF	az-tr	uz-tr	kk-tr	ky-tr
az-tr	48.64	75.7%	71.1%	73.4%
uz-tr	82.4%	39.45	74.8%	76.3%
kk-tr	79.5%	75.2%	34.70	85.0%
ky-tr	78.7%	75.2%	78.7%	33.25
Avg.	80.2%	75.4%	74.9%	78.2%

COMET	az-tr	uz-tr	kk-tr	ky-tr
az-tr	0.7182	80.5%	71.2%	67.2%
uz-tr	91.0%	0.6053	75.1%	69.9%
kk-tr	88.1%	78.4%	0.5429	81.4%
ky-tr	87.0%	78.7%	81.9%	0.5519
Avg.	88.7%	79.2%	76.1%	72.8%

COMETKiwi	az-tr	uz-tr	kk-tr	ky-tr
az-tr	0.6805	94.7%	95.3%	95.5%
uz-tr	93.7%	0.6217	96.0%	97.7%
kk-tr	89.2%	88.2%	0.4594	98.3%
ky-tr	88.5%	89.4%	93.4%	0.4241
Avg.	90.5%	90.8%	94.9%	97.2%

az

BLEU	tr-az	uz-az	kk-az	ky-az
tr-az	12.86	34.0%	24.4%	33.6%
uz-az	33.9%	9.17	31.3%	42.9%
kk-az	35.8%	43.0%	5.50	45.7%
ky-az	35.8%	46.6%	46.9%	5.32
Avg.	35.2%	41.2%	34.2%	40.7%

chrF	tr-az	uz-az	kk-az	ky-az
tr-az	47.74	70.9%	60.3%	67.2%
uz-az	66.8%	43.77	66.8%	67.9%
kk-az	68.9%	74.3%	38.07	80.6%
ky-az	67.5%	75.0%	75.1%	36.32
Avg.	67.7%	73.4%	67.4%	71.9%

COMET	tr-az	uz-az	kk-az	ky-az
tr-az	0.7429	82.8%	69.7%	68.1%
uz-az	103.9%	0.6533	73.7%	66.7%
kk-az	98.7%	80.1%	0.5776	77.7%
ky-az	99.9%	83.1%	84.2%	0.6049
Avg.	100.8%	82.0%	75.9%	70.8%

COMETKiwi	tr-az	uz-az	kk-az	ky-az
tr-az	0.6893	88.0%	92.5%	91.5%
uz-az	107.4%	0.6324	94.5%	92.0%
kk-az	102.5%	87.5%	0.4760	95.5%
ky-az	102.6%	88.7%	93.8%	0.4610
Avg.	104.2%	88.1%	93.6%	93.0%

uz

BLEU	tr-uz	az-uz	kk-uz	ky-uz
tr-uz	5.59	50.5%	26.9%	38.6%
az-uz	39.9%	6.99	29.9%	48.2%
kk-uz	38.5%	37.8%	6.95	52.8%
ky-uz	34.3%	34.6%	24.7%	3.96
Avg.	37.6%	41.0%	27.2%	46.5%

chrF	tr-uz	az-uz	kk-uz	ky-uz
tr-uz	38.31	75.2%	54.7%	67.1%
az-uz	69.9%	41.58	58.6%	70.9%
kk-uz	61.8%	65.9%	40.31	74.1%
ky-uz	60.7%	66.1%	52.8%	34.83
Avg.	64.1%	69.1%	55.4%	70.7%

COMET	tr-uz	az-uz	kk-uz	ky-uz
tr-uz	0.6472	83.0%	62.8%	67.5%
az-uz	97.0%	0.6731	64.6%	67.7%
kk-uz	97.5%	83.0%	0.6121	74.5%
ky-uz	97.6%	82.0%	64.3%	0.5981
Avg.	97.4%	82.7%	63.9%	69.9%

COMETKiwi	tr-uz	az-uz	kk-uz	ky-uz
tr-uz	0.6367	87.2%	84.3%	87.2%
az-uz	107.6%	0.6313	84.0%	89.1%
kk-uz	108.7%	92.4%	0.4998	92.0%
ky-uz	107.7%	91.4%	81.1%	0.4799
Avg.	108.0%	90.3%	83.1%	89.4%

kk

BLEU	tr-kk	az-kk	uz-kk	ky-kk
tr-kk	3.94	60.8%	35.5%	43.6%
az-kk	38.1%	5.23	40.2%	45.6%
uz-kk	25.1%	32.1%	8.90	41.5%
ky-kk	27.4%	28.1%	32.8%	6.12
Avg.	30.2%	40.3%	36.2%	43.6%

chrF	tr-kk	az-kk	uz-kk	ky-kk
tr-kk	35.25	81.7%	69.3%	74.4%
az-kk	69.0%	39.18	72.8%	79.1%
uz-kk	60.0%	65.4%	46.07	75.9%
ky-kk	60.9%	65.1%	67.0%	41.58
Avg.	63.3%	70.7%	69.7%	76.5%

COMET	tr-kk	az-kk	uz-kk	ky-kk
tr-kk	0.5726	93.6%	90.7%	88.3%
az-kk	83.3%	0.6144	92.3%	89.2%
uz-kk	79.5%	85.3%	0.6259	85.1%
ky-kk	81.5%	87.0%	93.4%	0.6383
Avg.	81.4%	88.6%	92.1%	87.5%

COMETKiwi	tr-kk	az-kk	uz-kk	ky-kk
tr-kk	0.4548	98.2%	98.4%	93.2%
az-kk	123.8%	0.4844	102.2%	96.6%
uz-kk	156.4%	110.1%	0.4976	93.6%
ky-kk	151.3%	119.9%	113.7%	0.4472
Avg.	143.8%	109.4%	104.8%	94.5%

ky

BLEU	tr-ky	az-ky	uz-ky	kk-ky
tr-ky	3.44	60.3%	35.6%	28.5%
az-ky	47.4%	4.68	37.1%	29.3%
uz-ky	40.1%	45.9%	5.85	39.7%
kk-ky	37.2%	39.5%	36.8%	6.70
Avg.	41.6%	48.6%	36.5%	32.5%

chrF	tr-ky	az-ky	uz-ky	kk-ky
tr-ky	31.62	82.7%	70.3%	63.8%
az-ky	75.3%	34.48	70.9%	68.5%
uz-ky	69.9%	73.0%	37.44	73.4%
kk-ky	68.2%	68.7%	70.6%	39.41
Avg.	71.1%	74.8%	70.6%	68.6%

COMET	tr-ky	az-ky	uz-ky	kk-ky
tr-ky	0.5123	93.3%	88.6%	83.6%
az-ky	95.6%	0.5416	90.5%	85.7%
uz-ky	98.7%	92.7%	0.5319	89.4%
kk-ky	99.2%	93.9%	95.9%	0.5658
Avg.	97.8%	93.3%	91.7%	86.2%

COMETKiwi	tr-ky	az-ky	uz-ky	kk-ky
tr-ky	0.3997	97.0%	98.7%	91.7%
az-ky	125.7%	0.4241	100.7%	92.8%
uz-ky	172.8%	112.7%	0.4425	95.7%
kk-ky	180.3%	123.2%	115.7%	0.4037
Avg.	159.6%	111.0%	105.0%	93.4%

diagonal raw score recovery rate

Figure 3: Full Latinized allCPT XWMT recovery matrices. Rows of panels correspond to fixed translation targets and metric columns correspond to BLEU, chrF, COMET, and COMETKiwi. Diagonal cells report raw same-source scores, off-diagonal cells report recovery rates, and the Avg. row reports the average recovery for each fine-tuning source.

of -2.93 for az→tr, -3.60 for uz→az, -3.34 for kk→uz, -3.18 for uz→kk, and -2.15 for kk→ky. At the same time, several kk/ky-related cells improve, such as the target=az chrF gains for kk and ky ($+4.34$ and $+2.07$). This pattern is consistent with the "curse of multilinguality", where multilingual training under fixed model capacity can hurt some high-resource or already well-represented directions while benefiting lower-resource ones (Chang et al., 2024). Because mT5-small has limited capacity, shared Turkic adaptation may help underrepresented patterns while degrading stronger supervised representations, especially on diagonal cells.

A.4 Qwen3 0.6B Recovery Tables

Figure 7 reports the full Qwen3 0.6B recovery matrices in the same fixed-target format used for the mT5 experiments. Diagonal cells give the raw same-source score, and off-diagonal cells give the recovery rate for cross-source transfer. These matrices provide architecture-level supporting evidence for the main results: although Qwen3 and mT5 differ in model family and training formulation, the Qwen3 matrices still exhibit structured intra-family transfer rather than random donor–evaluation behavior. The figure therefore supports the claim that the observed transfer structure is not merely an artifact of the mT5 encoder–decoder architecture.

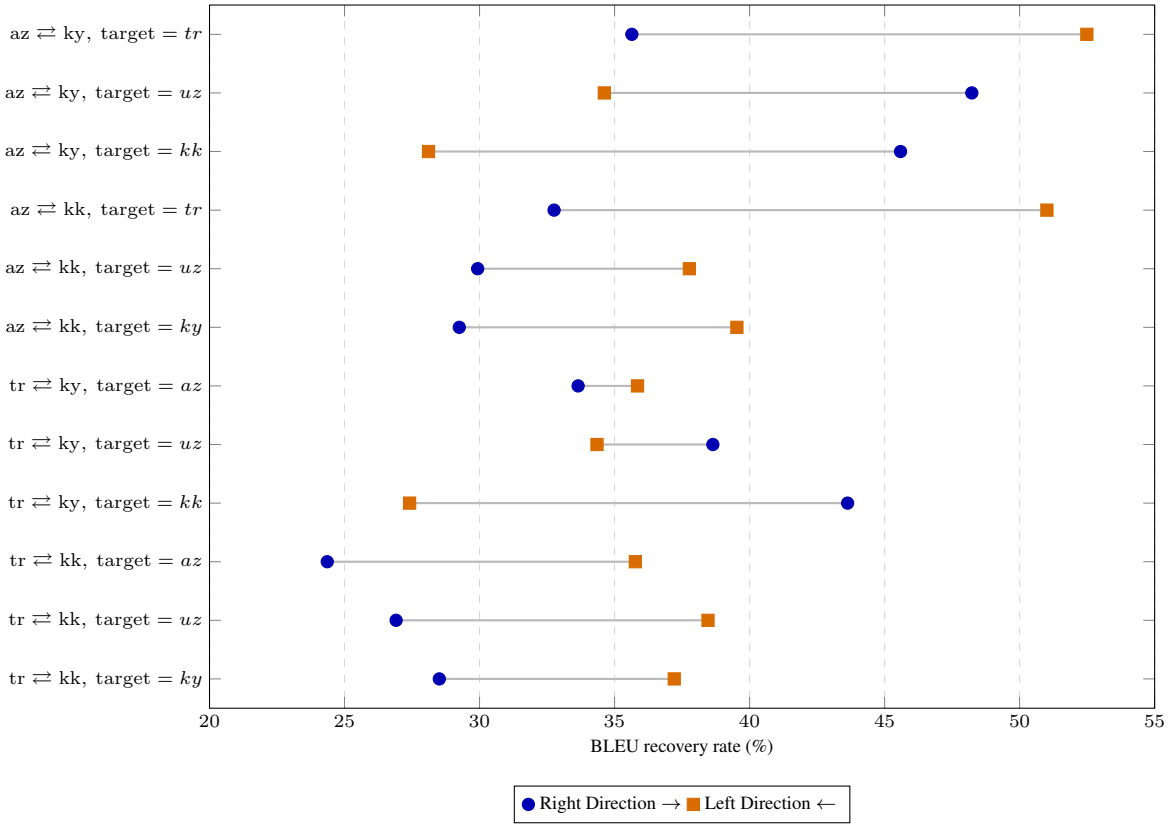


Figure 4: Extended appendix version of the paired BLEU recovery-rate comparison for mT5 Latin. Each y-axis label shows two opposite transfer source–transfer target directions under a fixed translation target. Direction → denotes transfer from the left language to the right language in the y-axis label, while Direction ← denotes transfer in the reverse direction.

tr

BLEU

	az	uz	kk	ky
uz	+0.02	-0.22	+0.48	+0.65
kk	+0.76	+0.35	-0.56	+0.22
ky	+0.31	+0.58	-0.50	+0.54

chrF

	az	uz	kk	ky
uz	-0.14	-0.32	+10.14	+7.35
kk	+0.55	+2.27	-0.53	-0.24
ky	+0.09	+2.17	-2.44	+0.75

az

BLEU

	tr	uz	kk	ky
uz	-0.01	+0.03	+0.47	+0.99
kk	-1.36	+0.41	-0.44	+0.39
ky	-0.45	+0.81	+0.09	+0.43

chrF

	tr	uz	kk	ky
uz	+0.04	+0.16	+14.52	+11.22
kk	-3.06	+1.57	+0.70	+1.50
ky	-0.85	+3.40	-0.16	+1.37

uz

BLEU

	tr	az	kk	ky
az	+0.10	-0.15	+0.73	+1.17
kk	+0.10	-0.15	+1.36	+0.87
ky	+0.19	-0.27	-0.29	+0.47

chrF

	tr	az	kk	ky
az	+0.07	-0.20	+7.19	+11.42
kk	-1.01	-1.98	+1.98	+5.08
ky	+0.94	-0.06	-1.06	+2.17

kk

BLEU

	tr	az	uz	ky
az	+0.27	+0.01	+1.25	+0.44
uz	+0.15	+0.30	+1.60	+0.41
ky	+0.34	+0.48	+1.48	-0.11

chrF

	tr	az	uz	ky
az	+10.25	+2.30	+7.96	+1.30
uz	+16.41	+5.61	+4.29	+0.85
ky	+18.99	+17.90	+19.15	+0.79

ky

BLEU

	tr	az	uz	kk
az	+0.39	+0.69	+0.75	-0.73
uz	+0.46	+0.73	+1.19	+0.06
kk	+0.30	+0.63	+0.83	-0.44

chrF

	tr	az	uz	kk
az	+9.21	+2.48	+4.27	-4.17
uz	+15.87	+5.07	+3.06	-1.86
kk	+16.34	+9.09	+14.71	-0.48

diagonal raw score positive delta negative delta zero delta

Figure 5: Latinized-minus-original allCPT XWMT raw-score deltas. Rows of panels correspond to fixed translation targets and metric columns correspond to BLEU, chrF, and COMET. Positive cells indicate gains from Latinization, while negative cells indicate decreases relative to the original-script setting.

tr

BLEU	az	uz	kk	ky
uz	-2.04	-2.41	+0.37	-0.05
kk	-1.57	-0.94	-2.37	-0.26
ky	-1.55	-0.44	+0.04	-1.33

chrF	az	uz	kk	ky
uz	-1.28	-2.46	+3.73	+1.87
kk	-0.30	-0.47	-2.02	+1.64
ky	-0.85	-0.38	+2.21	-1.50

COMET	az	uz	kk	ky
uz	-0.0270	-0.0562	+0.0626	+0.0313
kk	-0.0121	-0.0259	-0.0536	+0.0435
ky	-0.0062	-0.0216	+0.0245	-0.0411

COMETKiwi	az	uz	kk	ky
uz	-0.0218	-0.0468	+0.0206	+0.0194
kk	-0.0074	-0.0174	-0.0305	+0.0003
ky	+0.0073	-0.0041	-0.0011	-0.0191

az

BLEU	tr	uz	kk	ky
uz	-1.60	-3.60	-0.18	-0.60
kk	-1.23	-0.55	-3.33	-0.44
ky	-1.39	-0.65	+0.12	-2.26

chrF	tr	uz	kk	ky
uz	-1.08	-2.13	+4.34	+2.07
kk	-0.11	+1.37	-1.18	+3.77
ky	-0.57	+0.18	+3.62	-0.23

COMET	tr	uz	kk	ky
uz	-0.0160	-0.0659	+0.0643	+0.0388
kk	-0.0271	-0.0195	-0.0594	+0.0581
ky	-0.0123	-0.0176	+0.0578	-0.0299

COMETKiwi	tr	uz	kk	ky
uz	-0.0245	-0.0511	+0.0192	+0.0281
kk	-0.0230	-0.0048	-0.0299	+0.0199
ky	-0.0078	-0.0129	+0.0087	-0.0161

uz

BLEU	tr	az	kk	ky
az	-0.61	-3.13	-0.44	-0.36
kk	-0.17	-0.95	-3.34	-0.49
ky	-0.54	-1.32	-0.71	-1.79

chrF	tr	az	kk	ky
az	-0.54	-2.70	+0.80	+1.22
kk	+0.37	-1.76	-4.15	-0.10
ky	-0.67	-3.04	-2.27	-2.28

COMET	tr	az	kk	ky
az	-0.0228	-0.0636	+0.0190	+0.0364
kk	-0.0157	-0.0160	-0.0692	+0.0179
ky	+0.0082	-0.0204	-0.0110	-0.0534

COMETKiwi	tr	az	kk	ky
az	-0.0304	-0.0493	-0.0062	+0.0182
kk	-0.0247	-0.0084	-0.0350	-0.0053
ky	-0.0032	-0.0110	-0.0137	-0.0152

kk

BLEU	tr	az	uz	ky
az	-0.67	-1.80	-0.39	-0.04
uz	-0.38	-0.73	-3.18	+0.08
ky	-0.27	-0.63	-0.10	-1.82

chrF	tr	az	uz	ky
az	-2.12	-1.09	+1.63	+4.02
uz	-0.59	-1.46	-1.80	+3.48
ky	-0.23	+0.51	+1.85	-0.50

COMET	tr	az	uz	ky
az	-0.0175	-0.0059	+0.0018	+0.0364
uz	+0.0020	-0.0145	-0.0208	+0.0264
ky	+0.0023	+0.0098	+0.0183	-0.0045

COMETKiwi	tr	az	uz	ky
az	-0.0132	-0.0140	-0.0239	+0.0301
uz	+0.0167	-0.0035	-0.0185	+0.0248
ky	+0.0196	+0.0409	+0.0026	-0.0129

ky

BLEU	tr	az	uz	kk
az	-1.01	-2.08	-0.81	-0.46
uz	-0.53	-0.93	-1.87	+0.26
kk	-0.29	-0.56	-0.40	-2.15

chrF	tr	az	uz	kk
az	-2.86	-2.04	-0.11	+0.33
uz	-1.04	-1.48	-0.58	+2.54
kk	-0.09	-0.06	+1.88	-0.39

COMET	tr	az	uz	kk
az	-0.0167	-0.0072	-0.0134	-0.0603
uz	+0.0092	-0.0078	-0.0103	-0.0259
kk	+0.0078	+0.0087	+0.0157	-0.0073

COMETKiwi	tr	az	uz	kk
az	+0.0063	-0.0130	-0.0358	-0.0468
uz	+0.0708	+0.0060	-0.0129	-0.0291
kk	+0.0433	+0.0300	+0.0067	-0.0119

diagonal raw score positive delta negative delta zero delta

Figure 6: allCPT-minus-noCPT raw-score deltas on Latinized XWMT. Rows of panels correspond to fixed translation targets and metric columns correspond to BLEU, chrF, COMET, and COMETKiwi. Positive cells indicate gains from allCPT, while negative cells indicate stronger noCPT performance.

tr

BLEU

	az-tr	uz-tr	kk-tr	ky-tr
az-tr	13.43	43.6%	21.6%	25.1%
uz-tr	55.5%	6.51	25.3%	31.1%
kk-tr	45.4%	33.5%	4.62	49.7%
ky-tr	50.2%	35.6%	51.5%	3.50
en-tr	39.3%	32.6%	18.8%	22.3%
Avg.	47.6%	36.3%	29.3%	32.1%

chrF

	az-tr	uz-tr	kk-tr	ky-tr
az-tr	48.64	75.1%	53.5%	51.4%
uz-tr	82.7%	39.77	44.9%	55.5%
kk-tr	78.3%	68.9%	35.23	87.7%
ky-tr	78.5%	69.2%	84.5%	32.50
en-tr	72.1%	64.7%	26.2%	23.5%
Avg.	77.9%	69.5%	52.3%	54.5%

az

BLEU

	tr-az	uz-az	kk-az	ky-az
tr-az	12.86	34.1%	23.9%	19.2%
uz-az	34.0%	9.14	21.0%	26.4%
kk-az	46.4%	38.6%	5.94	41.7%
ky-az	39.4%	37.9%	41.9%	4.89
en-az	32.3%	31.9%	15.0%	20.6%
Avg.	38.0%	35.6%	25.4%	27.0%

chrF

	tr-az	uz-az	kk-az	ky-az
tr-az	47.74	71.1%	60.0%	56.3%
uz-az	66.7%	43.61	29.2%	38.4%
kk-az	75.3%	71.0%	37.37	79.4%
ky-az	69.3%	67.5%	76.9%	34.95
en-az	66.5%	63.4%	24.2%	23.7%
Avg.	69.5%	68.2%	47.6%	49.5%

uz

BLEU

	tr-uz	az-uz	kk-uz	ky-uz
tr-uz	5.80	48.7%	22.4%	16.9%
az-uz	36.7%	7.14	24.1%	21.2%
kk-uz	35.3%	39.1%	5.59	35.0%
ky-uz	29.8%	37.7%	36.0%	3.49
en-uz	29.3%	35.1%	21.8%	21.5%
Avg.	32.8%	40.2%	26.1%	23.6%

chrF

	tr-uz	az-uz	kk-uz	ky-uz
tr-uz	38.23	74.8%	47.5%	45.7%
az-uz	69.8%	41.78	42.9%	40.7%
kk-uz	64.6%	70.3%	38.33	63.5%
ky-uz	58.4%	66.0%	58.3%	32.66
en-uz	57.9%	61.9%	22.8%	20.1%
Avg.	62.7%	68.2%	42.9%	42.5%

kk

BLEU

	tr-kk	az-kk	uz-kk	ky-kk
tr-kk	4.08	58.2%	32.7%	39.0%
az-kk	30.1%	5.22	31.9%	37.7%
uz-kk	20.6%	26.4%	7.30	34.2%
ky-kk	18.1%	19.0%	19.7%	6.23
en-kk	17.2%	15.5%	15.1%	26.2%
Avg.	21.5%	29.8%	24.9%	34.3%

chrF

	tr-kk	az-kk	uz-kk	ky-kk
tr-kk	33.71	79.8%	60.6%	72.5%
az-kk	41.7%	36.88	61.3%	77.4%
uz-kk	14.1%	54.3%	41.78	75.2%
ky-kk	7.4%	20.6%	28.1%	40.79
en-kk	12.2%	19.7%	25.4%	69.9%
Avg.	18.9%	43.6%	43.9%	73.8%

ky

BLEU

	tr-ky	az-ky	uz-ky	kk-ky
tr-ky	3.25	59.1%	32.0%	39.2%
az-ky	38.1%	3.99	30.5%	37.7%
uz-ky	28.3%	35.6%	4.66	36.4%
kk-ky	30.1%	30.6%	28.3%	7.14
en-ky	24.3%	32.8%	21.5%	33.3%
Avg.	30.2%	39.5%	28.1%	36.6%

chrF

	tr-ky	az-ky	uz-ky	kk-ky
tr-ky	29.52	80.6%	66.1%	77.7%
az-ky	49.5%	32.00	64.8%	78.2%
uz-ky	21.1%	62.8%	34.38	77.1%
kk-ky	17.8%	45.7%	34.1%	39.89
en-ky	15.3%	30.5%	30.2%	76.1%
Avg.	25.9%	54.9%	48.8%	77.3%

diagonal raw score recovery rate

Figure 7: Qwen3 0.6B recovery matrices. Rows of panels correspond to fixed translation targets and metric columns correspond to BLEU and chrF. Diagonal cells report raw same-source scores, off-diagonal cells report recovery rates, and the Avg. row reports average recovery for each fine-tuning source.