LAYER-WISE CAUSAL EVIDENCE OF SCRIPT-DEPENDENT ADAPTATION IN LOW-RESOURCE TRANSLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Parameter-efficient adaptation of translation models is judged by corpus-level metrics, while the question of where adaptation occurs is answered with a familiar set of layer-wise instruments: adapter norms, representational similarity, linear probes and layer ablations. We fine-tune IndicTrans2-1B with rank-64 LoRA under rsLoRA scaling, one adapter per translation direction shared by four language–script pairs, across eight pipelines covering English, Manipuri in two scripts, Assamese and Bodo. Adaptation is uneven and unordered by data volume: English→Bodo, the smallest corpus, gains 8.69 points chrF++ and English→Meetei Mayek 4.10 points, whereas Bengali-script Manipuri is unchanged and Assamese, half of the training mixture, loses 3.34 points. Retranslating the training corpus for synthetic data changes chrF++ by less than one point in all forty comparisons and does not compound over rounds. The instruments prove less informative than they appear. Cross-lingual CKA mainly reflects whether two pipelines saw the same sentences, reaching 0.932 in the decoder for the one pair with identical inputs and at most 0.09 for the eleven pairs without; two probes reduce to one subword-fertility statistic; a part-of-speech probe reaches 0.94 accuracy at 0.03 selectivity; a script probe is perfect at every layer because its labels encode pipeline identity; and adapter norms never significantly predict ablation effects, whose load-bearing count ranges from 2 to 32 layers on differences of at most 1.05 chrF++. We release our code and resources[1].

1 INTRODUCTION

Parameter-efficient fine-tuning is the standard means of adapting pre-trained translation models to under-served languages, and the case for it is strong for Northeast India, where parallel corpora are scarce, several languages use more than one script, and IndicTrans2 (Gala et al., 2023) already gives non-trivial zero-shot quality. Practitioners combine two ingredients: a shared low-rank adapter (Hu et al., 2022) and synthetic data generated by the model itself (Sennrich et al., 2016; Hoang et al., 2018), then a layer-wise analysis that informs where capacity should go next, drawing on adapter update norms, representational similarity, linear probes, and layer ablations. Each is reported as beneficial in isolation, yet little is known about their joint behavior in the regime that motivates them: few script-diverse languages, tens of thousands of sentence pairs, and one shared adapter per direction.

We study that regime with eight pipelines covering Manipuri in two scripts, Assamese and Bodo, each paired with English in both directions (Table 1). The four English→X pipelines share one rank-64 LoRA adapter with rank-stabilised scaling (Kalajdzievski, 2023), and the four X→English pipelines share another. The English side of the test data is common to all eight, so differences within a direction group reflect language and script rather than content. We then evaluate filtered synthetic-data variants on top of the adapted models, and apply the layer-wise instruments as objects of study rather than settled measurements. Figure 1 summarises the study.

[1] https://anonymous.4open.science/r/NE_Indic_Layerwise_Analysis-5BC6/

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

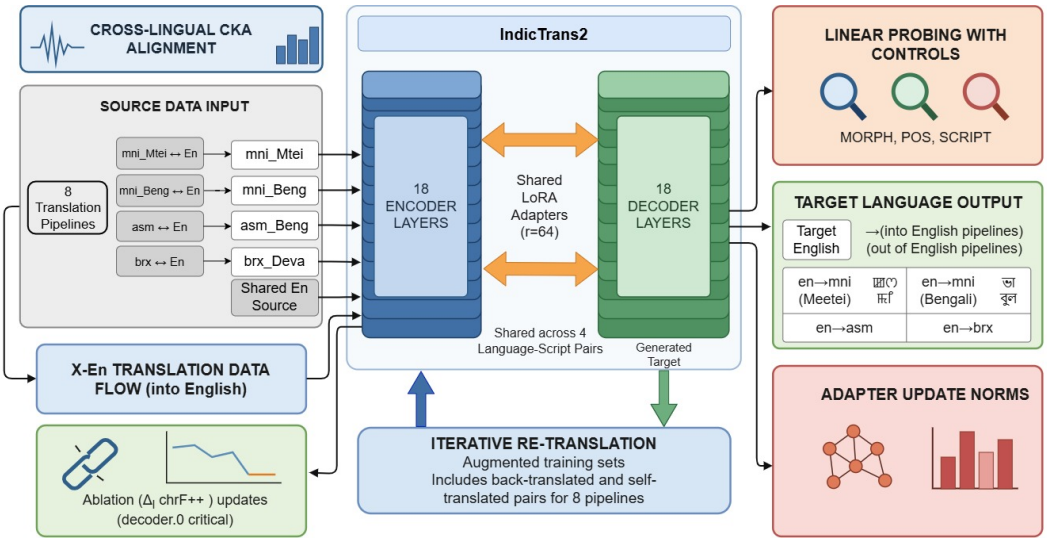

Figure 1: Overview of the proposed direction-level LoRA adaptation and layer-wise diagnostic framework for IndicTrans2-1B. Shared LoRA adapters are trained for English→Indic and Indic→English directions across four language–script pairs.

Adaptation proves uneven and is not ordered by data volume (Table 2): Bodo, the smallest corpus, gains most (8.69 points chrF++ out of English, 4.20 into English), Meetei Mayek gains 4.10, Bengali-script Manipuri is unchanged, and Assamese loses 3.34 despite supplying half of its group's mixture and being the pipeline the base model handles best. Synthetic pairs obtained by back-translation and by three rounds of teacher re-translation change chrF++ by less than one point in all forty comparisons and do not compound.

Several of the instruments measure the experiment rather than the model. Cross-lingual CKA is at most 0.09 for the eleven pipeline pairs shown different sentences and high only for the pair shown the same ones (Table 3). Two probes share one subword-fertility label and return identical accuracies, and a script probe is perfect at every layer because its labels restate pipeline identity (Table 12). Adapter norms never significantly predict ablation effects, and the load-bearing count runs from 2 to 32 layers on differences of at most 1.05 chrF++ (Table 5).

Our key contributions are the following.

- A direction-level multilingual LoRA adaptation (rsLoRA) of IndicTrans2-1B for four Northeast Indian language–script pairs in both directions, on a content-matched multi-way test set.
- A controlled evaluation of filtered synthetic augmentation from in-domain parallel data, including three rounds of iterative teacher re-translation, showing negligible and non-compounding effects.
- An audit of a layer-wise toolkit (norms, CKA drift and cross-lingual CKA, probes with shuffled-label controls, single-layer ablation) that identifies concrete confounds and the conditions under which each signal can be trusted.
- Evaluation observations for these languages, including a systematic disagreement between word-level and character-level metrics on the two pipelines translating into Manipuri.

2 Method

2.1 Problem Statement

Let e denote English and let $\mathcal{L} = \{Manipuri(\textit{Meetei}), Manipuri(\textit{Bengali}), Assamese, bodo\}$ be the four Indic language–script pairs. A *pipeline* $p = (s_p, t_p)$ is an ordered source–target pair with $\{s_p, t_p\} = \{e, \ell\}$, $\ell \in \mathcal{L}$, giving the eight pipelines A–H as shown in Table 1; they form two

Algorithm 1 Iterative teacher re-translation for direction group g

Require: LoRA-FT (LoRA-FineTune) adapters $\phi^g_{\mathrm{FT}}, \phi^{\bar{g}}_{\mathrm{FT}}$; training sets $\{\mathcal{D}^{\mathrm{tr}}_p\}_{p\in\mathcal{G}^g}$; ratios $\rho_{\mathrm{BT}}, \rho_{\mathrm{IT}}$; thresholds $\kappa_{\mathrm{BT}}, \kappa_{\mathrm{IT}}$; rounds K

for $p \in \mathcal{G}^g$ **do**
  build $\mathcal{D}^{\mathrm{BT}}_p$ with the opposite-group adapter $\phi^{\bar{g}}_{\mathrm{FT}}$ ▷ Eq. 5
$\phi^{(0)} \leftarrow \phi^g_{\mathrm{FT}}$
for $k = 1, \ldots, K$ **do**
  for $p \in \mathcal{G}^g$ **do**
    build $\mathcal{D}^{(k)}_p$ with teacher $\phi^{(k-1)}$ ▷ Eq. 6
    $\widehat{\mathcal{D}}_p \leftarrow \mathrm{dedup}\big(\mathcal{D}^{\mathrm{tr}}_p \uplus \mathcal{D}^{\mathrm{BT}}_p \uplus \mathcal{D}^{(k)}_p\big)$
  initialise $\phi \leftarrow \phi^g_{\mathrm{FT}}$ ▷ student restarts from LoRA-FT
  train ϕ on $\biguplus_p \widehat{\mathcal{D}}_p$ with Eq. 3; $\phi^{(k)} \leftarrow$ checkpoint selected by Eq. 4
return $\phi^{(1)}, \ldots, \phi^{(K)}$

Table 1: Pipelines and dataset statistics. Pipelines in $\mathcal{G}^{\rightarrow}$ represent En→Indic and $\mathcal{G}^{\leftarrow}$ represent Indic→En. Training, dev, and test sizes are identical for both directions of each pair, and π_p is the pipeline mixture weight.

		Pipeline ID		Dataset Size			
Indic Language(X)	**Script**	$\mathcal{G}^{\rightarrow}$ (En→X)	$\mathcal{G}^{\leftarrow}$ (X→En)	Train ($\lvert\mathcal{D}^{\mathrm{tr}}\rvert$)	Dev($\lvert\mathcal{D}^{\mathrm{dev}}_p\rvert$)	Test($\lvert\mathcal{D}^{\mathrm{te}}_p\rvert$)	π_p
Manipuri	Meetei Mayek	A	B	16,000	1,000	2,000	0.148
Manipuri	Bengali	C	D	22,687	1,000	2,000	0.210
Assamese	Assamese	E	F	54,000	1,000	2,000	0.500
Bodo	Devanagari	G	H	15,214	1,000	2,000	0.141

direction groups, $\mathcal{G}^{\rightarrow} = \{A, C, E, G\}$ and $\mathcal{G}^{\leftarrow} = \{B, D, F, H\}$, with $\bar{g}$ the group opposite to g. Each pipeline has sets $\mathcal{D}^{\mathrm{tr}}_p$, $\mathcal{D}^{\mathrm{dev}}_p$ and $\mathcal{D}^{\mathrm{te}}_p$ of pairs (x, y), the two directions of a language–script pair using the same pairs with source and target swapped. The English sides of the test sets are nearly identical across all eight pipelines (22,164 whitespace tokens for A–D and 22,160 for E–H), so the test data is content-matched; development sets differ across languages, except that E/F and G/H share English token statistics.

The base model for group g is an encoder–decoder Transformer θ^g_0 with 18 encoder and 18 decoder layers, namely `indictrans2-en-indic-1B` for $\mathcal{G}^{\rightarrow}$ and `indictrans2-indic-en-1B` for $\mathcal{G}^{\leftarrow}$ (Gala et al., 2023). The IndicTrans toolkit supplies a preprocessing map $\tau(x; s, t)$, which normalises the sentence and prepends source- and target-language tags, and a postprocessing map τ_t^{-1} applied after decoding; targets are preprocessed with the target language in both tag positions, and sequences are truncated to 256 tokens. All splits, test references included, pass through a deterministic cleaning map κ, which edits many sentences but removes no pairs and does not deduplicate, so duplicates in the raw data remain in $\mathcal{D}^{\mathrm{tr}}_p$. Training sizes differ by a factor of up to 3.6 across languages, so a comparison between two languages, or between two scripts of one language, is not a controlled comparison of that variable alone.

2.2 Direction-level low-rank adaptation

Let $\mathcal{M}$ be the adapted linear maps: the query, key, value and output projections of every attention block (encoder and decoder self-attention, decoder cross-attention) and both feed-forward projections, in all 36 layers. For $m \in \mathcal{M}$ with frozen weight $W^m_0 \in \mathbb{R}^{d^m_{\mathrm{out}} \times d^m_{\mathrm{in}}}$, LoRA (Hu et al., 2022) with rank-stabilised scaling (Kalajdzievski, 2023) replaces $u \mapsto W^m_0 u$ with

$$u \;\mapsto\; W^m_0 u \;+\; \frac{\alpha}{\sqrt{r}}\, B^m A^m \,\mathrm{drop}_q(u), \qquad A^m \in \mathbb{R}^{r \times d^m_{\mathrm{in}}}, \quad B^m \in \mathbb{R}^{d^m_{\mathrm{out}} \times r}, \tag{1}$$

with dropout rate q and B^m initialised to zero, so the adapted model $\theta^g_0 \oplus \phi$, $\phi = \{(A^m, B^m)\}_m$, initially equals the base model. With $r = 64$ and the IndicTrans2-1B dimensions $d = 1024$, $d_{\mathrm{ff}} = 8192$, the adapter has $|\phi| = r \sum_{m\in\mathcal{M}} (d^m_{\mathrm{in}} + d^m_{\mathrm{out}}) = 70{,}778{,}880$ parameters, 5.97% ($\mathcal{G}^{\rightarrow}$) and 6.47% ($\mathcal{G}^{\leftarrow}$) of the total.

One adapter ϕ^g is trained per group on the concatenation of its pipelines' training sets,

$$\mathcal{D}_g^{\mathrm{tr}} = \biguplus_{p \in \mathcal{G}^g} \{(x, y, p) : (x, y) \in \mathcal{D}_p^{\mathrm{tr}}\}, \qquad \pi_p = \frac{|\mathcal{D}_p^{\mathrm{tr}}|}{\sum_{q \in \mathcal{G}^g} |\mathcal{D}_q^{\mathrm{tr}}|}, \tag{2}$$

where $\biguplus$ keeps duplicates. Each epoch visits every example once, so pipeline p contributes a fraction π_p with no temperature re-weighting; batches are length-grouped. With $\mathcal{T}(\mathcal{B})$ the non-padding target positions of a micro-batch, $\tilde{x}_i = \tau(x_i; s_{p_i}, t_{p_i})$ and $p_{ij}(\cdot) = p_{\theta_0^g \oplus \phi}(\cdot \mid \mathbf{y}_{i,<j}, \tilde{x}_i)$, we minimise the token-averaged label-smoothed cross-entropy

$$\mathcal{L}_{\mathcal{B}}(\phi) = \frac{1}{|\mathcal{T}(\mathcal{B})|} \sum_{(i,j) \in \mathcal{T}(\mathcal{B})} \Big[-(1-\varepsilon) \log p_{ij}(\mathrm{y}_{ij}) - \frac{\varepsilon}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \log p_{ij}(v) \Big], \tag{3}$$

over ϕ only, with $\varepsilon = 0.1$ and output vocabulary $\mathcal{V}$. Every Δ_{eval} steps a checkpoint ϕ_k is scored by corpus chrF on the pooled development set $\mathcal{D}_g^{\mathrm{dev}}$, decoding greedily from token ids without τ_t^{-1} or κ:

$$v_k = \mathrm{chrF}\big(\{\hat{\mathbf{y}}_{\phi_k}^{\mathrm{greedy}}(x)\}, \{\mathbf{y}\}\big)_{(x,y) \in \mathcal{D}_g^{\mathrm{dev}}}. \tag{4}$$

A checkpoint improves if $v_k > \max_{k'<k} v_{k'} + \delta$; training stops after P evaluations without improvement or at the epoch limit, and we keep $\phi_{\mathrm{FT}}^g = \phi_{k^\star}$, $k^\star = \arg\max_k v_k$, one checkpoint for all four pipelines of the group. We call this system LoRA-FT (LoRA-FineTune).

2.3 Self-generated training data

Both procedures re-translate a random subset of a pipeline's own training data; no external monolingual text is used. Let $T_\phi^{s \to t}(x)$ be beam-search translation followed by τ_t^{-1}, tag stripping and the output-side normalisation of κ; $\mathcal{S}_\rho(\mathcal{D})$ a uniformly random subset of $\lfloor \rho |\mathcal{D}| \rfloor$ elements; and $\sigma(a, b)$ the cosine similarity of LaBSE embeddings (Feng et al., 2022). For $p = (s, t) \in \mathcal{G}^g$, reverse-model re-translation (BT) back-translates targets with the opposite-group adapter, and round k of iterative teacher re-translation (Iter-k) forward-translates sources with a teacher $\phi^{(k-1)}$:

$$\mathcal{D}_p^{\mathrm{BT}} = \big\{(\hat{x}, y) : (x, y) \in \mathcal{S}_{\rho_{\mathrm{BT}}}(\mathcal{D}_p^{\mathrm{tr}}),\ \hat{x} = T_{\phi_{\mathrm{FT}}^{\bar{g}}}^{t \to s}(y),\ |\hat{x}|_w \geq 3,\ \sigma(\hat{x}, x) \geq \kappa_{\mathrm{BT}}\big\}, \tag{5}$$

$$\mathcal{D}_p^{(k)} = \big\{(x, \tilde{y}) : (x, y) \in \mathcal{S}_{\rho_{\mathrm{IT}}}(\mathcal{D}_p^{\mathrm{tr}}),\ \tilde{y} = T_{\phi^{(k-1)}}^{s \to t}(x),\ |\tilde{y}|_w \geq 3,\ \tilde{y} \neq y,\ \sigma(\tilde{y}, y) \geq \kappa_{\mathrm{IT}}\big\}, \tag{6}$$

with $\phi^{(0)} = \phi_{\mathrm{FT}}^g$ and $\phi^{(k)}$ the adapter selected for Exp-$(k{+}1)$. Since every y already has a human source, Eq. 5 yields machine paraphrases of existing sources (Sennrich et al., 2016), while Eq. 6 is distillation and self-training restricted to in-domain sources with a reference-based filter (Kim & Rush, 2016; He et al., 2020). With $\mathrm{dedup}(\cdot)$ removing pairs whose whitespace-stripped source and target coincide, the training sets are

$$\widehat{\mathcal{D}}_p^{\mathrm{Exp1}} = \mathrm{dedup}\big(\mathcal{D}_p^{\mathrm{tr}} \uplus \mathcal{D}_p^{\mathrm{BT}}\big), \tag{7}$$

$$\widehat{\mathcal{D}}_p^{\mathrm{Exp}(k+1)} = \mathrm{dedup}\big(\mathcal{D}_p^{\mathrm{tr}} \uplus \mathcal{D}_p^{\mathrm{BT}} \uplus \mathcal{D}_p^{(k)}\big), \qquad k \in \{1, 2, 3\}. \tag{8}$$

Deduplication also drops duplicate original pairs (15,812 of 16,000 remain for A/B, 12,768 of 15,214 for G/H). Each experiment trains the group union of these sets with Eq. 3 and Eq. 4, initialised at ϕ_{FT}^g with a fresh optimiser, so teachers supply pseudo-labels only. Exp-5 instead continues from the Exp-2 adapter on $\widehat{\mathcal{D}}^{\mathrm{Exp2}}$; its intended segmentation sampling (Kudo, 2018) did not activate, so it is reported as continued training.

2.4 Evaluation

Hypotheses are $\hat{y} = T_\phi^{s_p \to t_p}(x)$ and references are the cleaned test targets, both NFC-normalised before scoring; note that the output-side normalisation of κ can modify hypotheses, for instance by collapsing immediately repeated English words. We report BLEU (Papineni et al., 2002) and chrF++ ($\beta = 2$, character order 6, word order 2) (Popović, 2015; 2017) at corpus level with sacreBLEU defaults (Post, 2018) and the `13a` tokenizer for all languages. TER (Snover et al., 2006). All systems, the zero-shot baseline included, share the decoding settings of §2.6.

2.5 LAYER-WISE DIAGNOSTICS

All diagnostics analyse ϕ^g_{FT}, pipeline by pipeline, on independent samples X_p of development sources drawn with a fixed seed: $n = 500$ for probes (of which the first 100 for ablation), $n = 300$ for CKA and $n = 50$ for cross-attention. Inputs are tokenised in batches of 16 with padding to the longest sequence, and for layer ℓ of either stack we pool the layer output $H_\ell(x) \in \mathbb{R}^{T_B \times d}$ over all T_B positions, padding included,

$$z_\ell(x) = \frac{1}{T_B} \sum_{t=1}^{T_B} H_\ell(x)_t, \qquad Z_\ell(X) = \big[z_\ell(x)\big]_{x \in X} \in \mathbb{R}^{n \times d}. \tag{9}$$

Decoder states come from a forward pass whose decoder input is the first token of the tokenised encoder input, so $T_B = 1$ and z^{dec}_ℓ describes a single decoding position under that fixed prefix.

Update norms. For layer ℓ with adapted modules $\mathcal{M}_\ell$ (six per encoder layer, ten per decoder layer),

$$\bar{n}_\ell = \frac{1}{|\mathcal{M}_\ell|} \sum_{m \in \mathcal{M}_\ell} \|B^m A^m\|_F, \qquad \iota^{\mathrm{norm}}_\ell = \frac{\bar{n}_\ell}{\max_{\ell'} \bar{n}_{\ell'}}, \tag{10}$$

the maximum running over all 36 layers. The norm omits $\alpha/\sqrt{r}$, and $\bar{n}_\ell$ is shared by all pipelines of a group.

Representational similarity. With $K = ZZ^\top$, $K' = Z'Z'^\top$ and the centring matrix $C = I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$, linear CKA (Kornblith et al., 2019) is

$$\mathrm{CKA}(Z, Z') = \frac{\langle CKC,\, CK'C \rangle_F}{\|CKC\|_F \, \|CK'C\|_F}, \tag{11}$$

which presumes that row i of Z and of Z' describe the same stimulus. *Drift* applies it to base and adapted states on identical inputs; *cross-lingual CKA* compares two pipelines of one group, each on its own sample with rows matched by index:

$$\mathrm{Drift}_\ell(p) = \mathrm{CKA}\big(Z^{\theta_0}_\ell(X_p),\, Z^{\theta_0 \oplus \phi_{\mathrm{FT}}}_\ell(X_p)\big), \tag{12}$$

$$\mathrm{XCKA}_\ell(p, q) = \mathrm{CKA}\big(Z_\ell(X_p),\, Z_\ell(X_q)\big). \tag{13}$$

Eq. 13 meets the pairing assumption only when X_p and X_q hold the same, or translation-equivalent, sentences in the same order (§4).

Linear probes. For each layer we fit multinomial logistic regression on standardised pooled states, using a stratified 80/20 split and inverse regularisation strength 1, and report test accuracy. Three labellings are used, all computed from the source sentence. The *fertility tercile* is

$$c^{\mathrm{fer}}(x) = 1 + \sum_{j=1}^{2} \mathbb{1}\big[\mathrm{fer}(x) \geq q_{j/3}\big], \qquad \mathrm{fer}(x) = \frac{|\mathrm{SP}(x)|}{|x|_w}, \tag{14}$$

for the tokenizer's SentencePiece segmentation $\mathrm{SP}(\cdot)$ and empirical terciles $q_{j/3}$ over X_p; both the "morphological complexity" and the "subword-to-word reconstruction" probe use Eq. 14 and therefore coincide. The *coarse POS* label is the tag class most frequent in x, on English-source pipelines only, and *script identity* labels two pooled pipelines of one group by dominant Unicode script, skipped when the pooled label has one class (C/E and D/F). Following control tasks (Hewitt & Liang, 2019), a second probe f^π is fitted on permuted training labels, and we report selectivity (Sel)

$$\mathrm{Sel}_\ell = \mathrm{Acc}(f, c) - \mathrm{Acc}(f^\pi, c), \tag{15}$$

both probes being scored against the true labels c.

Ablation and combined importance. Let $\phi^{-\ell}$ set $B^m \leftarrow 0$ for all $m \in \mathcal{M}_\ell$, removing that layer's update in Eq. 1. On the first 100 sentences of X_p with cleaned development references Y_p, translated with beam search and without the output-side normalisation of κ, we measure

$$\Delta_\ell(p) = \mathrm{chrF{+}{+}}\big(\hat{Y}_{\phi_{\mathrm{FT}}}, Y_p\big) - \mathrm{chrF{+}{+}}\big(\hat{Y}_{\phi^{-\ell}}, Y_p\big), \tag{16}$$

so $\Delta_\ell > 0$ means removing the update hurts. The two signals are combined as

$$\iota_\ell(p) = \tfrac{1}{2}\iota_\ell^{\mathrm{norm}} + \tfrac{1}{2}\iota_\ell^{\mathrm{abl}}(p), \qquad \iota_\ell^{\mathrm{abl}}(p) = \begin{cases} \Delta_\ell(p) \big/ \max_{\ell'} \Delta_{\ell'}(p), & \max_{\ell'} \Delta_{\ell'}(p) > 0, \\ 0, & \text{otherwise}, \end{cases} \tag{17}$$

and layer ℓ is *load-bearing* for p if $\iota_\ell(p) \geq 0.5$; since Δ_ℓ may be negative, $\iota_\ell^{\mathrm{abl}}$ is not confined to $[0, 1]$. We also report the Spearman correlation ρ_S between $\{\bar{n}_\ell\}$ and $\{\Delta_\ell(p)\}$ over the 36 layers, and inspect cross-attention weights from the same forward pass qualitatively.

2.6 Experimental Setup

Data and implementation. We use the WMT26-Indic-MT data on low-resource Indic translation (Pakray et al., 2026), covering four English-centred language–script pairs and hence eight pipelines, with sizes and mixture weights π_p (Eq. 2) in Table 1. All models are implemented in PyTorch with Hugging Face *transformers* and *PEFT* (rsLoRA), using *IndicProcessor* for τ and τ_t^{-1} (§2.1), and trained with *Seq2SeqTrainer* (AdamW, cosine schedule with linear warm-up, seed 16) in *bf16* with SDPA attention on a single NVIDIA RTX PRO 6000 Blackwell GPU (96 GB). LoRA-FT and Exp-1–Exp-5 share the configuration of Table 7; only the peak learning rate and epoch budget differ by group (4×10^{-5} over 10 epochs for $\mathcal{G}_{\rightarrow}$, 6×10^{-5} over 5 for $\mathcal{G}_{\leftarrow}$). Every system decodes with beam search (size 5, length penalty 1.0) and is scored on the same test set under the same output-side normalisation κ (§2.4), so differences reflect the training data alone.

Systems. Per pipeline, we compare seven systems. The unmodified IndicTrans2-1B checkpoint is the *Zero-shot* baseline. *LoRA-FT* uses one adapter per group, trained on the original mixture (Eq. 2), and serves as the reference for all augmentation runs. *Exp-1* (**Back-Translation**) retrains LoRA-FT after adding back-translated pairs (Eq. 7). *Exp-2/3/4* (**BT + Iterative Pseudo-Labelling**) add one further round of teacher pseudo-labels, with $k = 1, 2, 3$, respectively (Eq. 8). Each round restarts from ϕ^g_{FT} rather than the previous round. *Exp-5* (**BT + Best Iterative Round + Subword Regularisation**) continues training from the strongest of Exp-2/3/4, selected by group-averaged test BLEU (Exp-2 in both groups). It uses that round's training data. We report it as continued training because the intended subword regularisation did not activate (§2.3).

Layer-wise diagnostics. We run the diagnostics of §2.5 on the group-level adapter ϕ^g_{FT}, separately per pipeline, drawing an independent sample from D_p^{dev} for each ($n = 500$ for probing and ablation, $n = 300$ for CKA, $n = 50$ for cross-attention). Section 3.2 reports cross-lingual CKA (Eq. 13), causal adapter ablation (Eq. 16) and combined importance with its Spearman correlation ρ_S (Eq. 17). Appendix A.3 reports update norms (Eq. 10, §A.3.1), representational drift (Eq. 12, §A.3.2), the fertility, coarse-POS and script-identity probes with control-task selectivity (Eqs. 14 and 15, §A.3.3), the encoder/decoder split of the ablation (§A.3.4), and cross-attention entropy and peak mass (§A.3.5).

3 Results & Analysis

3.1 Translation Quality

Table 2 reports BLEU/chrF++ for all eight pipelines across the zero-shot Baseline, LoRA-FT and Exp-1–Exp-5, with the two metrics bolded independently against their own row maximum; they do not always agree on which system is best.

Adaptation is uneven across directions. LoRA-FT gains most on the smallest pipeline, *en→brx_Deva* (G, +10.81 BLEU / +8.69 chrF++), and on *en→mni_Mtei* (A, +4.10 chrF++ at unchanged BLEU). For *en→mni_Beng* (C) and *en→asm_Beng* (E), the baseline instead holds the row's highest BLEU; this is clearest on C, where baseline BLEU (7.35) exceeds every adapted system while chrF++ moves the other way (34.41 vs. 35.27 for Exp-3). The four *indic→eng* pipelines improve on both metrics.

Augmentation changes little. Across all $8 \times 5 = 40$ pipeline×experiment comparisons, chrF++ moves by less than a point from LoRA-FT (largest swing +0.86 on C, Exp-3) and BLEU by less than two (largest −1.83 on B, Exp-5). Exp-2 (BT + Iter-1) is the strongest variant, taking the highest chrF++ on D, F and G and the highest BLEU on F and G, which matches the group-averaged BLEU ranking used to pick the Exp-5 starting checkpoint (§2.3); Exp-1 and Exp-5 win no pipeline on either metric. No variant improves on LoRA-FT across all eight pipelines.

Table 2: BLEU/chrF++ results for the eight pipelines. Bold denotes the best value for each metric within a pipeline

ID	Baseline	LoRA-FT	Exp-1	Exp-2	Exp-3	Exp-4	Exp-5
A	7.30/34.80	**7.36**/38.91	6.66/38.62	7.12/39.18	7.14/**39.19**	6.26/39.01	6.63/38.79
B	23.78/57.28	**28.22/59.37**	27.07/58.92	27.35/59.24	27.28/59.01	26.92/58.87	26.39/58.55
C	**7.35**/34.41	6.13/34.41	5.67/34.74	5.90/34.89	5.83/**35.27**	5.79/35.05	5.74/34.84
D	23.34/57.06	**24.84**/57.60	24.27/57.39	**24.84/57.69**	24.50/57.45	24.64/57.56	24.24/57.40
E	**24.14/59.79**	21.65/56.46	21.66/56.62	22.27/56.97	22.34/57.05	22.11/56.99	21.55/56.42
F	32.41/66.42	38.39/66.97	37.96/67.10	**38.94/67.30**	38.30/67.06	38.36/67.21	37.94/67.19
G	24.25/57.12	35.06/65.81	34.94/65.70	**35.33/65.98**	35.04/65.70	34.62/65.63	34.88/65.74
H	31.13/63.11	**40.58**/67.33	39.93/67.23	40.25/67.18	40.36/67.44	40.09/**67.48**	39.69/67.15

Table 3: Cross-lingual CKA (Eq. 13) within direction groups ($\mathcal{G}_{\rightarrow}$: A, C, E, G; $\mathcal{G}_{\leftarrow}$: B, D, F, H) (Table.1). Languages: Manipuri in Meetei Mayek (*Mtei*) and Bengali (*Beng*) scripts, Assamese (*Asm*), and Bodo (*Bodo*). The last column in each group lists layers with CKA ≥ 0.85 (out of 36).

	Group $\mathcal{G}_{\rightarrow}$ (En→X)				**Group $\mathcal{G}_{\leftarrow}$ (X→En)**			
Language Pair	**Pair**	Enc.	Dec.	≥ 0.85	**Pair**	Enc.	Dec.	≥ 0.85
Mtei ↔ Beng	A↔C	0.002	0.001	0/36	B↔D	0.005	0.004	0/36
Mtei ↔ Asm	A↔E	0.025	0.013	0/36	B↔F	0.003	0.004	0/36
Mtei ↔ Bodo	A↔G	0.015	0.010	0/36	B↔H	0.019	0.005	0/36
Beng ↔ Asm	C↔E	0.001	0.001	0/36	D↔F	0.003	0.005	0/36
Beng ↔ Bodo	C↔G	0.002	0.001	0/36	D↔H	0.003	0.002	0/36
Asm ↔ Bodo	**E↔G**	**0.481**	**0.932**	**19/36**	**F↔H**	0.086	0.087	0/36

3.2 Layer-wise analysis

We now ask *where* adaptation takes effect, reporting cross-lingual CKA, causal ablation and combined importance here and the remaining diagnostics in Appendix A.3.

Cross-lingual representations align for only one language pair. Table 3 gives layer-wise CKA (Eq. 13) for every pair within a direction group. Alignment is absent everywhere except Assamese↔Bodo: across the other eleven pairs mean CKA never exceeds 0.09 and no layer reaches 0.85, whereas E↔G reaches 0.932 in the decoder (17/18 layers ≥ 0.85) against 0.481 in the encoder. E and G are also the one pair whose development sets share an English side, so both receive identical inputs; §4 shows that this, rather than any property of the languages, accounts for the value.

Causal ablation gives an unstable count of load-bearing layers. Ablating each LoRA-updated layer in turn changes chrF++ on 100 development sentences (Eq. 16), so these values are not comparable to the test scores of Table 2. In Table 4 the count of load-bearing layers ($\Delta_\ell > 0$) ranges from 5/36 on A to 34/36 on D, on effects of at most 1.05 chrF++, and the most critical layer is a decoder layer in seven of eight pipelines (the exception is B, encoder layer 12) even though the update norms of Appendix A.3 concentrate in the encoder.

Adapter-norm importance does not predict causal importance. Combining both signals (Eq. 17) flags between 2/36 (A) and 32/36 (D) layers as load-bearing (Table 5), but the Spearman correlation between $\{\bar{n}_\ell\}$ and $\{\Delta_\ell(p)\}$ over the 36 layers is never significantly positive: negative for A ($\rho_S = -0.446$, $p = 0.006$) and D (-0.361, $p = 0.030$), and indistinguishable from zero elsewhere ($|\rho_S| \leq 0.28$, $p > 0.10$). Update-norm magnitude is therefore not a reliable proxy for a layer's causal contribution in either direction.

Statistical significance. Table 6 reports paired bootstrap significance (1,000 resamples, Koehn (2004)) for LoRA-FT vs. the zero-shot baseline, on BLEU and chrF++, across all eight pipelines. Three pipelines disagree between metrics. A improves significantly on chrF++ but not BLEU. C regresses significantly on BLEU while chrF++ does not change. F improves significantly on BLEU, but its chrF++ gain is not significant. E is the only pipeline where LoRA-FT regresses on both metrics, both significantly. B, D, G, and H improve significantly on both.

Table 4: Causal ablation (Eq. 16) on 100 dev sentences per pipeline without κ-normalization. Side-by-side alignment of $\mathcal{G}_{\rightarrow}$ (En→X) and $\mathcal{G}_{\leftarrow}$ (X→En) highlights directional asymmetries. "Hurts/Helps" counts layers with $\Delta_\ell > 0$ vs. $\Delta_\ell < 0$ (out of 36 layers total). Languages: Manipuri (*Mtei*/*Beng*), Assamese (*Asm*), and Bodo (*Bodo*).

	Group $\mathcal{G}_{\rightarrow}$ (En→X)				**Group $\mathcal{G}_{\leftarrow}$ (X→En)**			
Language Pair	**ID**	**chrF++**	**Hurts/Helps**	**Top Layer (Δ_ℓ)**	**ID**	**chrF++**	**Hurts/Helps**	**Top Layer (Δ_ℓ)**
Manipuri (Meetei)	A	42.15	5 / 31	dec.0 (+0.83)	B	59.03	12 / 24	enc.12 (+0.33)
Manipuri (Bengali)	C	53.59	13 / 23	dec.7 (+0.37)	D	63.26	34 / 2	dec.0 (+0.69)
Assamese (Assamese)	E	50.08	6 / 30	dec.4 (+0.73)	F	69.72	14 / 22	dec.6 (+0.64)
Bodo (Devanagari)	G	60.42	24 / 12	dec.16 (+0.67)	H	68.16	25 / 11	dec.6 (+1.05)

Table 5: Combined importance (Eq. 17) and the norm/ablation Spearman correlation ρ_S over 36 layers per pipeline. No ρ_S is significantly positive.

Pipelines A–D					**Pipelines E–H**				
ID	**Load-bearing (enc/dec)**	**Total/36**	ρ_S	p	**ID**	**Load-bearing(enc/dec)**	**Total/36**	ρ_S	p
A	0/2	2/36	-0.446	0.006	E	2/4	6/36	-0.277	0.101
B	5/6	11/36	-0.055	0.751	F	4/8	12/36	-0.092	0.593
C	7/4	11/36	-0.038	0.824	G	10/8	18/36	0.038	0.827
D	15/17	32/36	-0.361	0.030	H	8/10	18/36	-0.067	0.698

Robustness of the A-vs-C comparison. A and C differ in training size by a factor of 1.4 (Table 1), so their comparison is confounded with corpus size. Size-matching to the smaller count leaves the chrF++ gap essentially unchanged ($+4.05$ pp unbalanced vs. $+4.01$ pp balanced), the balanced–unbalanced difference is not significant for either pipeline ($p = 0.22$ for A, $p = 0.26$ for C; see Appendix Figure 2), and training each pipeline alone moves chrF++ by less than 0.2 pp. Across five seeds (Appendix Figure 3) both pipelines are stable (A: 38.41 ± 0.15 chrF++; C: 34.52 ± 0.04), yet A's gain over baseline is significant in all five seeds ($p < 0.001$) and C's in none ($p > 0.4$). The asymmetry is therefore a property of the language–script pair, not of corpus size or training-run variance. We ran this check for A and C only, the pair behind the script contrast of §1.

4 Discussion

Why self-generated data does not help. Reported gains from back-translation and self-training come mostly from monolingual text absent from the parallel data (Sennrich et al., 2016; Hoang et al., 2018; He et al., 2020), whereas ours re-translate sentences the model has already seen; data diversification (Nguyen et al., 2020) gains from this recipe, but with several independently trained models trained from scratch. Our pseudo-labels come from one adapter and must pass a LaBSE filter against the gold target ($\sigma \geq 0.78$, Eq. 6), which admits precisely the outputs that add least to what the student knows. We see no collapse either, since every set retains the real data and every student restarts from ϕ_{FT} (Shumailov et al., 2024; Gerstgrasser et al., 2024).

Probes and norms need controls that match the label. The two "linguistic" probes share the fertility label of Eq. 14: they are one probe run twice, and their identical accuracies say nothing about morphology as distinct from subword reconstruction. For the script probe the control cannot help at all, because the confound lies in the labels: script is a function of pipeline identity, which every input carries as a language tag. Adapter norms fail differently: the profile is flat enough that every layer exceeds 0.5 normalised norm (Appendix A.3.1), so $\iota_\ell \geq 0.5$ becomes a weak condition on Δ_ℓ and the count swings from 2 layers on A to 32 on D. Methods allocating capacity by such scores (Zhang et al., 2023; Zhou et al., 2025) assume they track a layer's contribution.

Where adaptation takes effect. Norms and drift point to the encoder, whose CKA against the base model falls to 0.23–0.49 in seven pipelines while the decoder stays at 0.65–0.83 (Table 9); ablation places the largest effect in the decoder for seven of eight. Some disagreement is expected, since norms and drift measure how much a layer changes, not what it contributes (Hase et al., 2023). Drift also fails to track quality: E drifts most in the encoder and loses 3.3 points, while G drifts less

Table 6: Paired bootstrap significance of LoRA-FT vs. baseline (BLEU, chrF++), all eight pipelines. Δ is LoRA-FT $-$ baseline; *marks $p < 0.05$.

ID	Direction	ΔBLEU	p	ΔchrF++	p
A	eng→mni_Mtei	$+0.06$	0.85	$+4.11^*$	< 0.001
B	mni_Mtei→eng	$+4.44^*$	< 0.001	$+2.09^*$	< 0.001
C	eng→mni_Beng	-1.22^*	< 0.001	-0.00	0.97
D	mni_Beng→eng	$+1.50^*$	< 0.001	$+0.54^*$	0.018
E	eng→asm_Beng	-2.49^*	< 0.001	-3.33^*	< 0.001
F	asm_Beng→eng	$+5.98^*$	< 0.001	$+0.55$	0.076
G	eng→brx_Deva	$+10.81^*$	< 0.001	$+8.69^*$	< 0.001
H	brx_Deva→eng	$+9.45^*$	< 0.001	$+4.22^*$	< 0.001

and gains most. For A, removing 31 of 36 updates even *improves* development chrF++ although the full adapter adds 4.1 points on test.

What cross-lingual CKA is measuring. Eq. 11 assumes row i of both matrices describes the same stimulus, which is why cross-lingual comparisons rely on parallel sentences (Kudugunta et al., 2019). Eq. 13 meets that assumption only when two pipelines draw the same development sentences in the same order; for eleven of the twelve pairs they do not, and the near-zero values of Table 3 are what CKA returns for unpaired inputs. The exception, E↔G, is the pair whose development sets share an English side (§2.1), so fixed-seed sampling gives both identical inputs and the decoder CKA of 0.93 needs no appeal to lexical borrowing between Assamese and Bodo. That this pair is nonetheless perfectly separable by a script probe shows representations can be globally similar under CKA yet trivially distinguishable along one direction (Davari et al., 2023).

5 Conclusion

We adapted IndicTrans2-1B to four Northeast Indian language-script pairs in both directions with one shared LoRA adapter per direction group, and applied six layer-wise diagnostics to the result. Adaptation proved uneven and unordered by data volume: Bodo, the smallest corpus, improved most (+8.69 chrF++ out of English) followed by Meetei Mayek +4.10, while Bengali-script Manipuri was unchanged and Assamese decreased 3.34 chrF++ despite supplying half of its training mixture; augmenting with the model's own re-translations (BT+FT+Iter) moved chrF++ by less than a point in all forty comparisons and did not compound over rounds. The diagnostics were more cautionary still. Cross-lingual CKA reproduced input overlap rather than model structure, two probes turned out to share one fertility label, the script probe was perfect at every layer because its labels encode pipeline identity, and adapter norms never significantly predicted ablation effects, whose load-bearing count ranged from 2 to 32 layers on differences of at most 1.05 chrF++ over 100 sentences. Such analyses are therefore interpretable only on content-matched inputs, with probe labels that are not determined by tokenizer statistics or pipeline identity, with resampled confidence intervals, and with orthography and metric coverage checked for the script at hand. Repeating these comparisons across seeds, mixtures and model scales, and with external monolingual signal, is left to future work.

Reproducibility Statement

All experiments use the public IndicTrans2-1B checkpoints *indictrans2-en-indic-1B* and *indictrans2-indic-en-1B* (Gala et al., 2023) and WMT 2026 data. The data splits, preprocessing, LoRA configuration, training objective, synthetic-data procedures, evaluation settings, and layer-wise diagnostics are fully specified in Sections 3.1–3.6 and Appendix A. Experiments use a random seed of 16 and a single NVIDIA RTX PRO 6000 Blackwell GPU (96 GB).

Ethics Statement

We use publicly released WMT26 Indic-MT corpora for four Northeast Indian language–script pairs; they contain no personal data beyond what is public in their sources. The work aims to serve under-resourced languages, but it also shows how easily quality is misread here, since metrics disagree across scripts and neural metrics inherit their encoders' language coverage. Adapted models reproduce the biases and orthographic conventions of their training data, so we release adapters and outputs for research use and recommend evaluation with native speakers before deployment.

REFERENCES

Chantal Amrhein and Rico Sennrich. On Romanization for model transfer between scripts in neural machine translation. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 2461–2469, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.223. URL https://aclanthology.org/2020.findings-emnlp.223/.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. Massively multilingual neural machine translation in the wild: Findings and challenges. 2019. URL https://arxiv.org/abs/1907.05019.

Ankur Bapna and Orhan Firat. Simple, scalable adaptation for neural machine translation. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 1538–1548, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1165. URL https://aclanthology.org/D19-1165/.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, March 2022. doi: 10.1162/coli_a_00422. URL https://aclanthology.org/2022.cl-1.7/.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. What do neural machine translation models learn about morphology? In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 861–872, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1080. URL https://aclanthology.org/P17-1080/.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747/.

MohammadReza Davari, Stefan Horoi, Amine Natik, Guillaume Lajoie, Guy Wolf, and Eugene Belilovsky. Reliability of CKA as a similarity measure in deep learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=8HRvyxc606.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT sentence embedding. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 878–891, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.62. URL https://aclanthology.org/2022.acl-long.62/.

Baban Gain, Trilok Nath Singh, and Asif Ekbal. One model to translate them all? a journey to mount doom for multilingual model merging. 2026. URL https://arxiv.org/abs/2604.02881.

Jay Gala, Pranjal A Chitale, A K Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M, Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M Khapra, Raj Dabre, and Anoop Kunchukuttan. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=vfT4YuzAYA.

Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Tomasz Korbak, Henry Sleight, Rajashree Agrawal, John Hughes, Dhruv Bhandarkar Pai, Andrey Gromov, Dan Roberts, Diyi Yang, David L. Donoho, and Sanmi Koyejo. Is model collapse inevitable? breaking the curse of recursion by accumulating real and synthetic data, 2024. URL https://openreview.net/forum?id=5B2K4LRgmz.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Junxian He, Jiatao Gu, Jiajun Shen, and Marc'Aurelio Ranzato. Revisiting self-training for neural sequence generation. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SJgdnAVKDH.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. Iterative back-translation for neural machine translation. In Alexandra Birch, Andrew Finch, Thang Luong, Graham Neubig, and Yusuke Oda (eds.), *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pp. 18–24, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-2703. URL https://aclanthology.org/W18-2703/.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Damjan Kalajdzievski. A rank stabilization scaling factor for fine-tuning with lora. 2023. URL https://arxiv.org/abs/2312.03732.

Yash Khemchandani, Sarvesh Mehtani, Vaidehi Patil, Abhijeet Awasthi, Partha Talukdar, and Sunita Sarawagi. Exploiting language relatedness for low web-resource language model adaptation: An Indic languages study. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1312–1323, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.105. URL https://aclanthology.org/2021.acl-long.105/.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In Jian Su, Kevin Duh, and Xavier Carreras (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1317–1327, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1139. URL https://aclanthology.org/D16-1139/.

Philipp Koehn. Statistical significance tests for machine translation evaluation. In Dekang Lin and Dekai Wu (eds.), *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 388–395, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-3250/.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019.

Taku Kudo. Subword regularization: Improving neural network translation models with multiple subword candidates. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 66–75, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1007. URL https://aclanthology.org/P18-1007/.

Sneha Kudugunta, Ankur Bapna, Isaac Caswell, and Orhan Firat. Investigating multilingual NMT representations at scale. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 1565–1575, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1167. URL https://aclanthology.org/D19-1167/.

Xuan-Phi Nguyen, Shafiq Joty, Wu Kui, and Ai Ti Aw. Data diversification: a simple strategy for neural machine translation. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Partha Pakray, Santanu Pal, Advaitha Vetagiri, Kshetrimayum Boynao Singh, Sandeep Kumar Dash, Arnab Kumar Maji, Saralin A. Lyngdoh, Lenin Laitonjam, Anupam Jamatia, Ajit Das, Sunita Warjri, Uzzal Sharma, Koj Sambyo, and Riyanka Manna. Findings of WMT 2026 shared task on low-resource Indic languages translation. In *Proceedings of the Eleventh Conference on Machine Translation (WMT)*, Budapest, Hungary, 2026.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin (eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

Telmo Pires, Robin Schmidt, Yi-Hsiu Liao, and Stephan Peitz. Learning language-specific layers for multilingual machine translation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14767–14783, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.825. URL https://aclanthology.org/2023.acl-long.825/.

Maja Popović. chrF: character n-gram F-score for automatic MT evaluation. In Ondřej Bojar, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, and Pavel Pecina (eds.), *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pp. 392–395, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-3049. URL https://aclanthology.org/W15-3049/.

Maja Popović. chrF++: words helping character n-grams. In Ondřej Bojar, Christian Buck, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, and Julia Kreutzer (eds.), *Proceedings of the Second Conference on Machine Translation*, pp. 612–618, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4770. URL https://aclanthology.org/W17-4770/.

Matt Post. A call for clarity in reporting BLEU scores. In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor (eds.), *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 186–191, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6319. URL https://aclanthology.org/W18-6319/.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In Katrin Erk and Noah A. Smith (eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 86–96, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1009. URL https://aclanthology.org/P16-1009/.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pp. 223–231, Cambridge, Massachusetts, USA, August 8-12 2006. Association for Machine Translation in the Americas. URL https://aclanthology.org/2006.amta-papers.25/.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5797–5808, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1580. URL https://aclanthology.org/P19-1580/.

Haoran Xu, Kenton Murray, Philipp Koehn, Hieu Hoang, Akiko Eriguchi, and Huda Khayrallah. X-ALMA: Plug & play modules and adaptive rejection for quality translation at scale. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=csbf1p8xUq.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Hf17y6u9BC.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adaptive budget allocation for parameter-efficient fine-tuning. volume abs/2303.10512, 2023. URL https://api.semanticscholar.org/CorpusID:257631760.

Hongyun Zhou, Xiangyu Lu, Wang Xu, Conghui Zhu, Tiejun Zhao, and Muyun Yang. LoRA-drop: Efficient LoRA parameter pruning based on output evaluation. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 5530–5543, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.371/.

A Appendix

A.1 Related Work

Translation for Northeast Indian languages. IndicTrans2 (Gala et al., 2023) covers all 22 scheduled Indian languages, including Manipuri in both scripts, Assamese and Bodo, which are also the focus of the WMT low-resource Indic task (Pakray et al., 2026). Script is a recurring factor in transfer, since sharing a script, or transliterating into a common one, tends to help related languages (Amrhein & Sennrich, 2020; Khemchandani et al., 2021). Manipuri, written in two scripts, is therefore a natural testbed, and our test set shares its English side across pipelines so that differences within a direction group reflect language and script rather than content.

Placing adapter capacity by layer. LoRA (Hu et al., 2022) with rank-stabilised scaling (Kalajdzievski, 2023) is now standard for specialising translation models, following earlier adapters (Bapna & Firat, 2019). Several methods decide *where* that capacity should go, by parameter sensitivity (Zhang et al., 2023), output magnitude (Zhou et al., 2025), or per-language layers (Pires et al., 2023) and modules (Xu et al., 2025), all presuming that layer importance can be measured reliably. The diagnostics used for that purpose have known limits: CKA compares representations across languages (Kornblith et al., 2019; Kudugunta et al., 2019) but can be manipulated without changing behaviour (Davari et al., 2023); probe accuracy alone is not evidence of encoding (Belinkov et al., 2017; Hewitt & Liang, 2019; Belinkov, 2022); and ablation depends on the metric and

Table 7: Hyperparameters. Schedule entries are given as $\mathcal{G}_{\rightarrow}$ / $\mathcal{G}_{\leftarrow}$; all other entries apply to both groups, with slashes listing the named quantities in order.

LoRA-FT (shared)		**Augmentation (Exp-1–Exp-5)**	
rank r / α	64 / 128	ρ_{BT} / ρ_{IT}	0.40 / 0.25
dropout q	0.1	κ_{BT} / κ_{IT}	0.75 / 0.78
scaling γ_r	$\alpha/\sqrt{r}$	rounds K / max epochs	3 / 3
label smoothing ε	0.1	peak LR / warm-up	2×10^{-5} / 2%
weight decay / grad. clip	0.01 / 1.0	Δ_{eval} / patience P	300 / 8
batch (device $\times$ accum.)	32×2	initialisation	ϕ_{FT} (Exp-5: Exp-2)
max sequence length	256	α_{sr} (Exp-5, inactive)	0.1
Schedule ($\mathcal{G}_{\rightarrow}$ / $\mathcal{G}_{\leftarrow}$)		**Decoding and diagnostics**	
max epochs	10 / 5	beam / length penalty	5 / 1.0
peak LR	4×10^{-5} / 6×10^{-5}	max new tokens	256
warm-up ratio	0.10 / 0.05	dev-selection decoding	greedy, ≤ 128
Δ_{eval} (steps)	150 / 300	dev batch size	16
patience P	10 / 8	n: probe (ablation) / CKA / attn.	500 (100) / 300 / 50
threshold δ (chrF)	0.05	probe split / load-bearing	80/20 / $\iota_\ell \geq 0.5$

intervention chosen (Voita et al., 2019; Zhang & Nanda, 2024; Hase et al., 2023). Sharing parameters across languages also invites interference (Arivazhagan et al., 2019; Conneau et al., 2020). Closest to our design, Gain et al. (2026) find that merging fine-tuned models works better under a shared target language than a shared source; we ask the same question of one jointly trained adapter per direction group.

Synthetic data from the model itself. Back-translation (Sennrich et al., 2016; Hoang et al., 2018), self-training (He et al., 2020) and sequence-level distillation (Kim & Rush, 2016) enlarge low-resource training sets, but most gains come from monolingual text the parallel data lacks. Closest to our setting, data diversification (Nguyen et al., 2020) re-translates the parallel data itself; we follow that recipe with a LaBSE filter (Feng et al., 2022) on an adapter over a strong pretrained model, repeated over rounds, which also relates our setting to collapse under recursive training (Shumailov et al., 2024) and to the real data that prevents it (Gerstgrasser et al., 2024).

A.2 Hyperparameters Table

Table 7 lists every hyperparameter used for LoRA-FT, the five augmentation experiments, and decoding and diagnostics. Group-level schedule entries are given as $\mathcal{G}_{\rightarrow}$ / $\mathcal{G}_{\leftarrow}$; all other entries are shared across both groups.

A.3 Additional layer-wise diagnostics

This appendix reports the remaining four diagnostics from §2.5: LoRA update norms, representational drift, linear probing, and cross-attention inspection. Together with §3.2, this covers all six analyses.

A.3.1 LoRA update norms

Each direction group trains one shared adapter, so update norms are identical across pipelines within a group. We therefore report one row per group, not per pipeline. Table 8 lists the five layers with the highest mean $\|B^m A^m\|_F$ (Eq. 10) in each group; Figure 4 plots the full per-layer profile for every pipeline. Both groups' top-5 layers are all encoder layers. This is the source of the mismatch flagged in §3.2: norm magnitude alone flags 36/36 layers as load-bearing in both groups, so it cannot on its own identify which layers matter causally.

A.3.2 Representational drift

For each pipeline we compare base-model and LoRA-FT representations on the same inputs (Eq. 12). Low CKA means the adapter moved that layer's representation further from the base

Table 8: Top-5 layers by LoRA update norm (Eq. 10), grouped by direction. Importance is normalized to each group's maximum.

Group $G_{\rightarrow}$ (A, C, E, G)			**Group** $G_{\leftarrow}$ (B, D, F, H)		
Layer	$\|B^m A^m\|_F$	Importance	Layer	$\|B^m A^m\|_F$	Importance
enc.16	1.128	1.000	enc.13	0.746	1.000
enc.15	1.123	0.995	enc.14	0.745	0.999
enc.14	1.108	0.982	enc.10	0.740	0.992
enc.13	1.099	0.974	enc.11	0.734	0.983
enc.12	1.078	0.955	enc.12	0.733	0.983

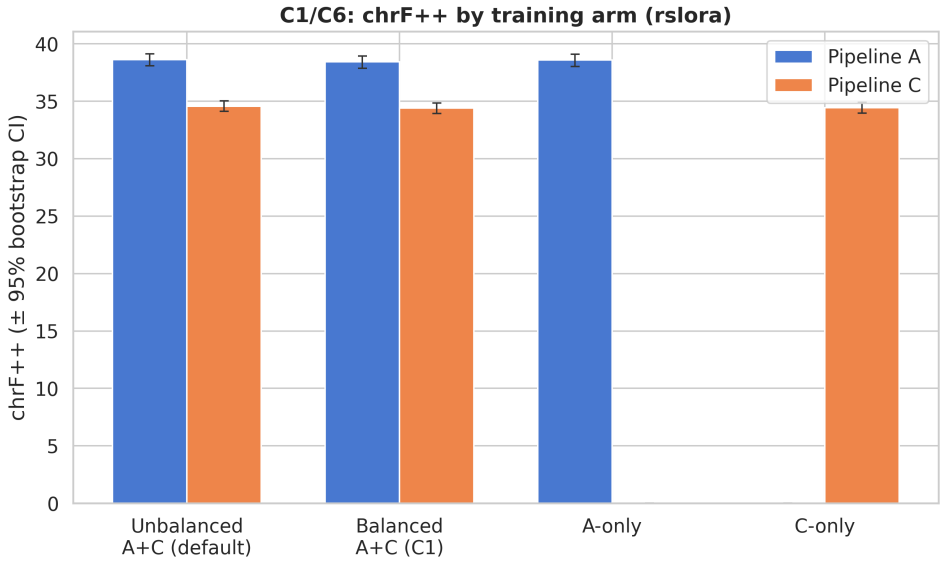

Figure 2: chrF++ (rsLoRA) for pipelines A and C under four training arms, covering two reviewer-requested robustness checks: C1, size-matching A and C to the smaller pipeline's training count (balanced vs. unbalanced), and C6, training each pipeline jointly vs. alone (solo). The A-C gap is stable across all four arms.

model. Table 9 reports the encoder and decoder means, and Figure 5 shows the per-layer curves behind those means. The encoder drifts more than the decoder in seven of eight pipelines. The one exception is A, where encoder and decoder drift by a similar amount.

Table 9: Representational drift (Eq. 12): mean CKA between base and LoRA-FT states, 18 encoder / 18 decoder layers per pipeline. Lower means more drift.

ID	Encoder mean CKA	Decoder mean CKA
A	0.687	0.785
B	0.240	0.728
C	0.290	0.767
D	0.310	0.648
E	0.228	0.814
F	0.388	0.726
G	0.360	0.831
H	0.492	0.757

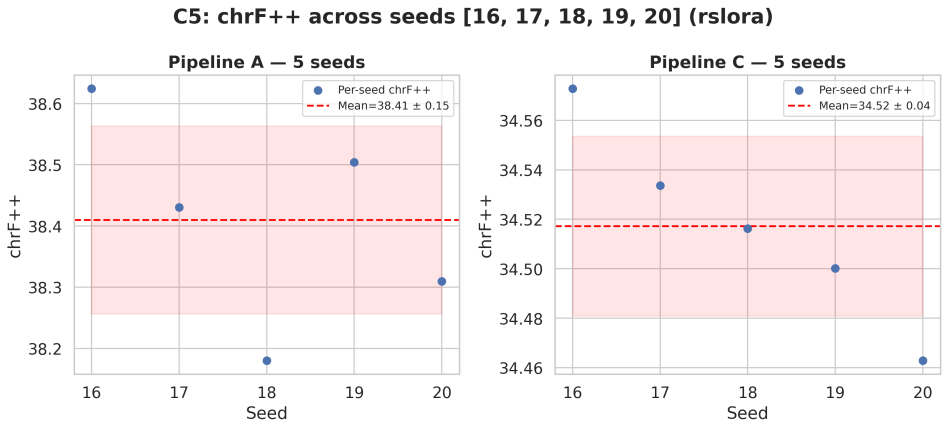

Figure 3: chrF++ (rsLoRA) for pipelines A and C across five random seeds (16–20). Dashed line: five-seed mean; shaded band: ± 1 std. Per-seed variation is small for both pipelines.

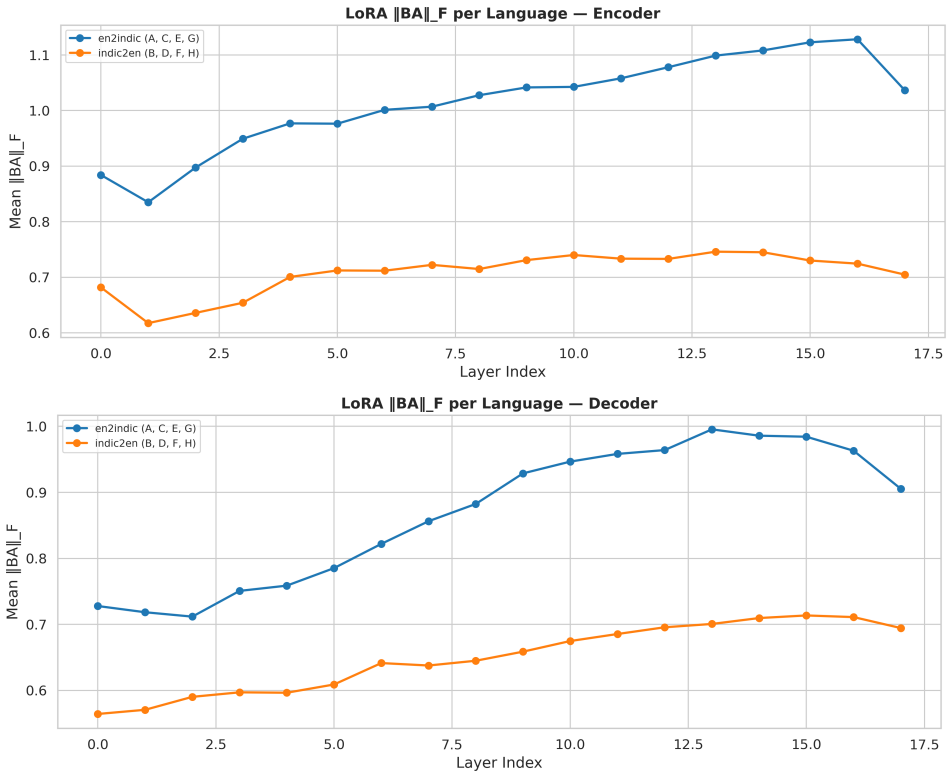

Figure 4: LoRA update norm $\bar{n}_\ell$ (Eq. 10) by layer, encoder (left) and decoder (right), one line per pipeline.

A.3.3 Linear probing

Within-pipeline probes (morphological complexity, POS, reconstruction). Morphological complexity and subword-to-word reconstruction both use the same fertility label (Eq. 14). We confirm they give identical accuracy at every layer; only selectivity differs, since the control task uses an independent label permutation. Table 10 reports the best layer per pipeline and side, and Figure 6 plots accuracy and selectivity across all layers. Table 11 and Figure 7 give the same breakdown for



Figure 5: Representational drift (Eq. 12) per layer for pipelines A–H. Each panel shows the encoder and decoder curves.

the coarse-POS probe, run only on the four English-source pipelines (A, C, E, G), as specified in §!2.5.

Table 10: Best-layer accuracy and selectivity (Eq. 15) for the morphological-complexity / reconstruction probe (identical accuracy by construction; best layer may differ because selectivity does not).

	Encoder		Decoder	
ID	Layer (Acc.)	Sel.	Layer (Acc.)	Sel.
A	1 (0.65)	0.43	16 (0.46)	0.21
B	11 (0.61)	0.37	7 (0.52)	0.21
C	0 (0.70)	0.45	2 (0.51)	0.25
D	3 (0.61)	0.32	3 (0.50)	0.23
E	9 (0.70)	0.41	6 (0.55)	0.33
F	17 (0.62)	0.44	15 (0.53)	0.23
G	1 (0.71)	0.41	9 (0.82)	0.26
H	1 (0.67)	0.35	6 (0.56)	0.30

Table 11: Best-layer accuracy and selectivity for the coarse-POS probe, English-source pipelines only.

	Encoder		Decoder	
ID	Layer (Acc.)	Sel.	Layer (Acc.)	Sel.
A	10 (0.94)	0.03	11 (0.92)	0.05
C	11 (0.88)	0.46	5 (0.83)	0.40
E	0 (0.79)	0.30	12 (0.82)	0.30
G	11 (0.81)	0.31	9 (0.82)	0.26

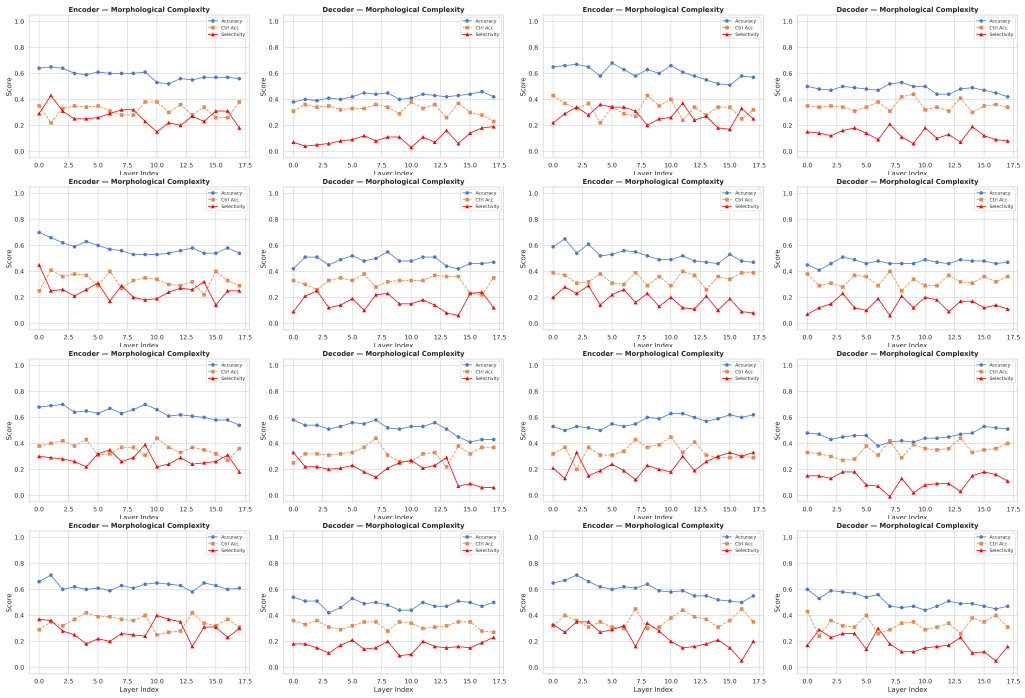

Figure 6: Morphological-complexity probe accuracy and selectivity by layer for pipelines A–H.

Cross-pipeline script-identity probing (Phase 4B). For pairs of pipelines in the same direction group, we pool their states and probe script identity. C↔E and D↔F are skipped: both pairs use the same script (Bengali), so the pooled label has only one class. For all 10 remaining pairs, accuracy is 1.000 at *every* layer, not just the best one; only selectivity varies with depth. Table 12 gives the best-selectivity layer per pair, and Figure 8 shows the flat accuracy curve against the varying selectivity for two representative pairs.

Table 12: Script-identity probe. Accuracy is 1.000 at every layer for all 10 valid pairs; the table gives the layer with the highest selectivity.

Side	Pair	Best layer	Selectivity
Decoder	A↔C	3	0.595
Decoder	A↔E	13	0.610
Decoder	A↔G	12	0.610
Decoder	C↔G	6	0.555
Decoder	E↔G	1	0.570
Encoder	B↔D	4	0.555
Encoder	B↔F	12	0.560
Encoder	B↔H	7	0.555
Encoder	D↔H	1	0.600
Encoder	F↔H	2	0.555

A.3.4 Causal ablation: encoder/decoder split

The encoder/decoder split shows the load is not just concentrated in one layer, but in one *stack*. Table 13 breaks the same ablation results down by encoder and decoder separately (18 layers each), rather than pooling all 36, and Figure 9 plots the full per-layer Δ_ℓ profile behind each row. Two findings sharpen the picture from Table 4. First, pipeline A is an extreme case: *no* encoder layer is causally load-bearing (0/18 hurt, 18/18 help), and its single largest encoder Δ_ℓ is actually negative (-0.08 at layer 0) — removing any encoder layer's LoRA update either helps or is neutral, so for this pipeline the entire adaptation effect lives in the decoder. Pipeline D sits at the opposite extreme,

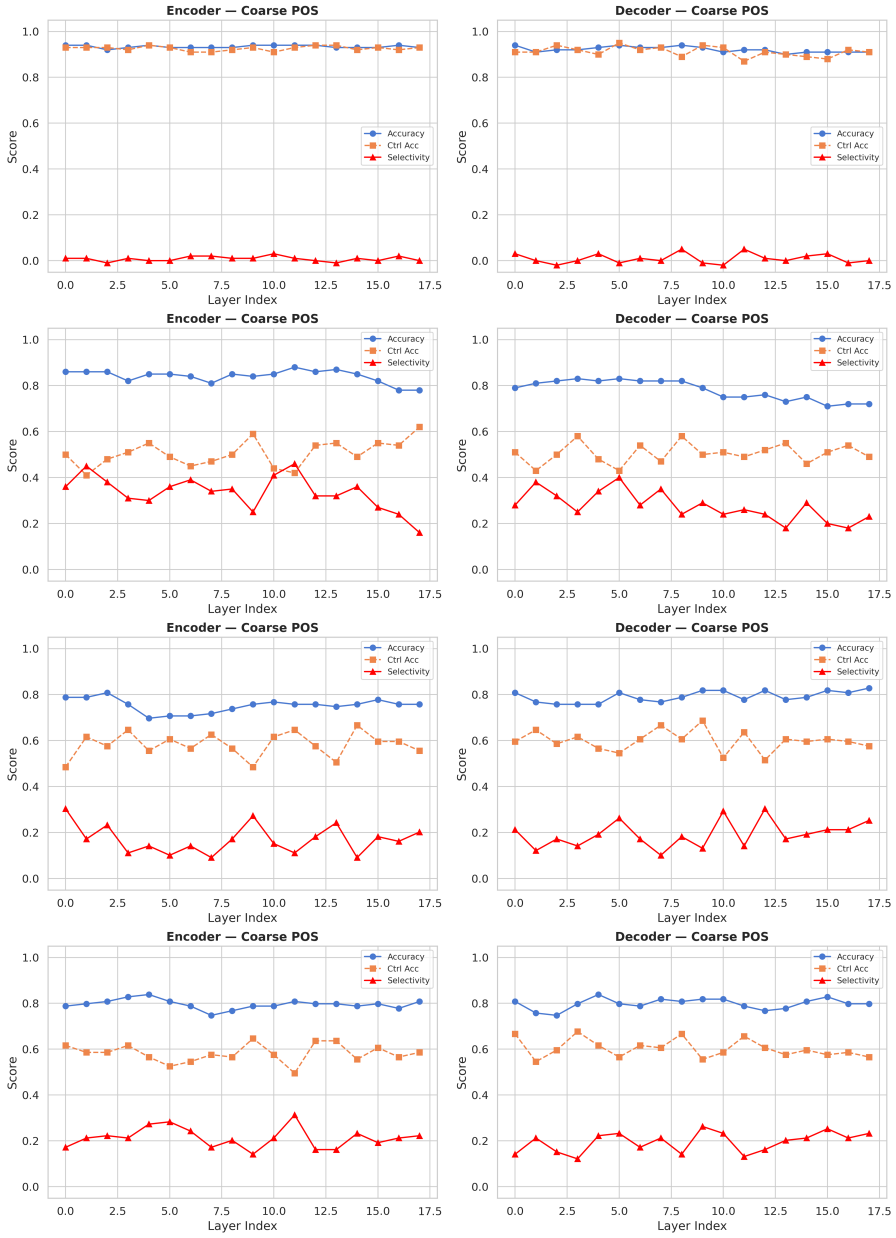

Figure 7: Coarse-POS probe accuracy and selectivity by layer for pipelines A, C, E, and G.

with its decoder stack fully saturated (18/18 layers hurt when removed) and its encoder stack also majority-load-bearing (16/18). Second, the asymmetry we noted for the single most-critical layer holds at the stack level too: the decoder's hurt-count is greater than or equal to the encoder's in seven of eight pipelines (the sole exception is C, 7/18 encoder vs. 6/18 decoder), and the decoder's largest Δ_ℓ exceeds the encoder's largest Δ_ℓ in seven of eight pipelines (the sole exception is B, where encoder layer 12 is the global maximum). The two exceptions are not the same pipeline, so no single direction group is driving the pattern.

A.3.5 Cross-attention inspection

We inspect encoder-decoder cross-attention from the same one-token-prefix forward pass used for decoder states (§ 2.5), reporting per-layer entropy and peak attention mass. Table 14 summarises the mean and endpoint values, and Figure 10 plots the full per-layer curve for every pipeline. There is

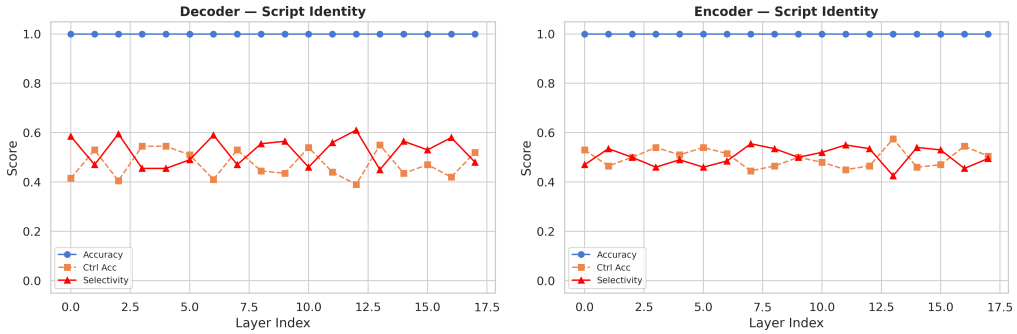

Figure 8: Script-identity probe accuracy (flat at 1.0) and selectivity by layer, for A↔G (decoder) and B↔H (encoder) as representative pairs.

Table 13: Causal ablation (Eq. 16), split by encoder and decoder stack (18 layers each, no ties). "Hurts"/"Helps" count layers with $\Delta_\ell > 0$/$\Delta_\ell < 0$ within that stack. The last two columns give the single largest Δ_ℓ within the encoder and within the decoder respectively; a negative value means no layer in that stack is load-bearing.

			Encoder (/18)		Decoder (/18)		Most critical layer (Δ_ℓ)	
ID	Direction	Full chrF++	Hurts	Helps	Hurts	Helps	Encoder	Decoder
A	eng→mni_Mtei	42.15	0/18	18/18	5/18	13/18	0 (−0.08)	0 (+0.83)
B	mni_Mtei→eng	59.03	5/18	13/18	7/18	11/18	12 (+0.33)	7 (+0.28)
C	eng→mni_Beng	53.59	7/18	11/18	6/18	12/18	17 (+0.26)	7 (+0.37)
D	mni_Beng→eng	63.26	16/18	2/18	18/18	0/18	9 (+0.49)	0 (+0.69)
E	eng→asm_Beng	50.08	2/18	16/18	4/18	14/18	12 (+0.31)	4 (+0.73)
F	asm_Beng→eng	69.72	5/18	13/18	9/18	9/18	9 (+0.31)	6 (+0.64)
G	eng→brx_Deva	60.42	11/18	7/18	13/18	5/18	16 (+0.59)	16 (+0.67)
H	brx_Deva→eng	68.16	12/18	6/18	13/18	5/18	10 (+0.44)	6 (+1.05)

no consistent trend with depth: entropy neither rises nor falls monotonically from layer 0 to layer 17 in any pipeline. E and G stand out with markedly lower entropy and higher peak attention than the other six pipelines — the same two pipelines whose decoder representations showed near-perfect cross-lingual CKA in § 3.2. We do not claim a causal link between the two observations, only that they point the same way.

Table 14: Encoder-decoder cross-attention, 18 decoder layers per pipeline. Entropy in nats; max attention is the peak probability mass on a single source token.

ID	Mean entropy	Entropy (L0 → L17)	Max attn. (L0 → L17)
A	3.354	3.402 → 3.323	0.243 → 0.220
B	3.747	3.921 → 3.772	0.102 → 0.119
C	3.356	3.354 → 3.337	0.246 → 0.219
D	3.671	3.743 → 3.802	0.130 → 0.079
E	2.682	2.712 → 2.674	0.318 → 0.301
F	3.313	3.439 → 3.427	0.119 → 0.083
G	2.651	2.700 → 2.604	0.324 → 0.305
H	3.144	3.220 → 3.129	0.158 → 0.124

A.3.6 Combined layer importance

Figure 11 shows the per-layer combined importance score (Eq. 17) for all eight pipelines, encoder (top) and decoder (bottom), with the load-bearing threshold (0.5) dashed. Load-bearing layers range from 2/36 for Pipeline A to 32/36 for Pipeline D (Table 5), and are scattered across depth rather than



Figure 9: Causal ablation Δ_ℓ per layer (Eq. 16) for pipelines A–H. Encoder layers 0–17 are followed by decoder layers 0–17. Positive bars indicate layers whose removal decreases chrF++ (load-bearing), whereas negative bars indicate layers whose removal increases chrF++.

concentrated in any band, consistent with the weak/negative correlation between adapter-norm and causal importance reported in the main text.

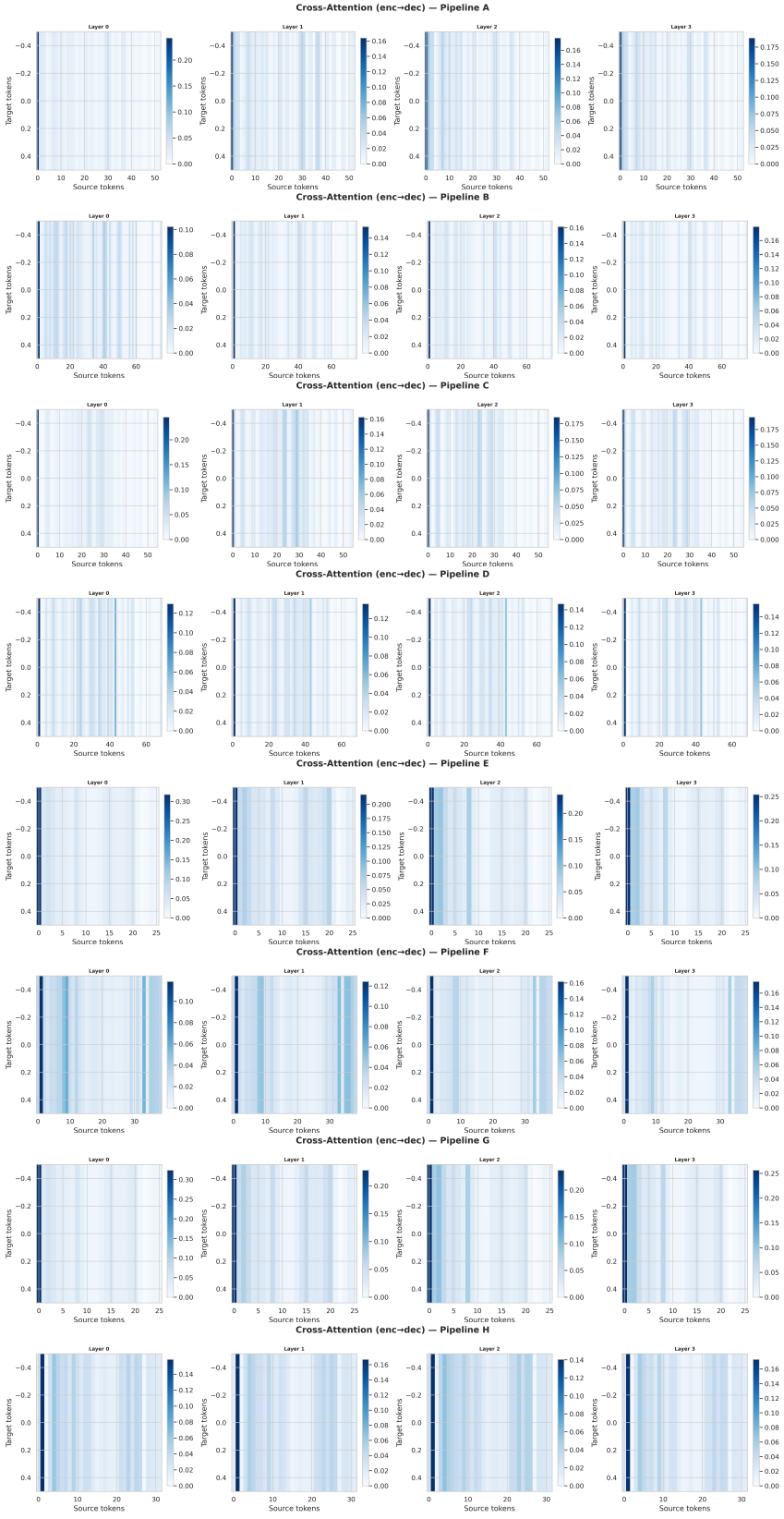

Figure 10: Encoder-decoder cross-attention entropy and peak mass by decoder layer for pipelines A–H.

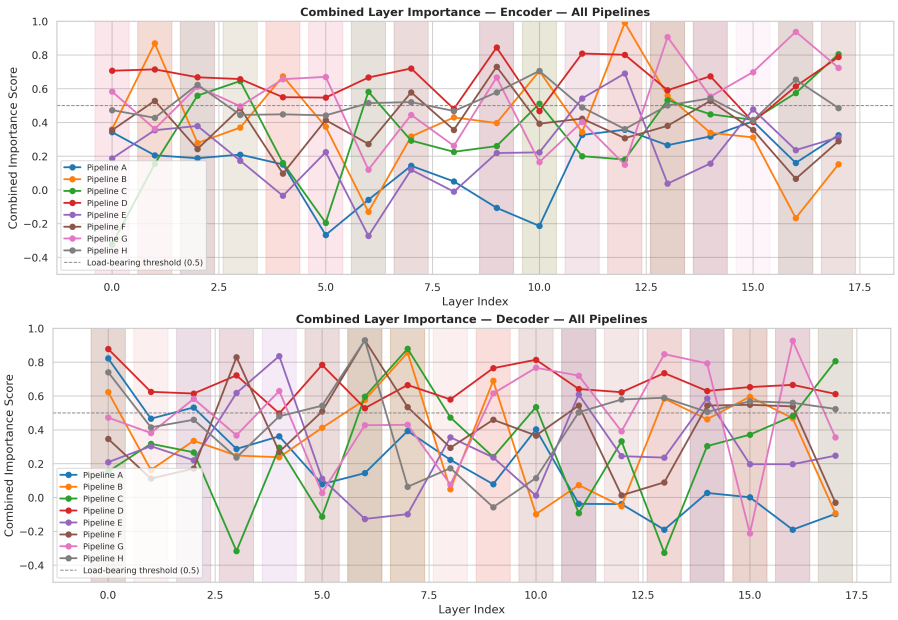

Figure 11: Combined layer importance score (Eq. 17) per layer for all eight pipelines, encoder (top) and decoder (bottom). Dashed line marks the load-bearing threshold (0.5); shaded bands mark layers flagged load-bearing for at least one pipeline.