MORPHOT: EVALUATING MORPHOLOGICALLY RICH INDIC LANGUAGES WITH HIERARCHICAL UNBALANCED OPTIMAL TRANSPORT

Anonymous authors
Paper under double-blind review

ABSTRACT

Embedding-based translation metrics correlate strongly with human judgments on English benchmarks, yet the same metrics fail systematically on morphologically rich languages (MRLs) for three structural reasons. Such metrics stay blind to fine-grained inflection errors, conserve mass between hypothesis and reference and therefore conflate omissions with hallucinations, and reward reference mimicry without verifying fidelity to the source. We introduce **MORPHOT**, a hierarchical unbalanced optimal transport (UOT) metric that answers each failure mode with one architectural component. MORPHOT represents a sentence as a discrete measure over morphological units. An inner UOT over affix lattices produces a morphology cost sensitive to inflectional differences, an outer UOT decomposes into a matched discrepancy together with separate omission and hallucination residuals, and a source-anchored adequacy term guards against reference-mimicking drift. A monotone softplus-parameterized head then aggregates the components under a formal guarantee that no training run inverts the sign of a penalty. With ≈ 0.44 M trainable parameters atop a frozen multilingual encoder, MORPHOT attains the highest pooled Pearson, Spearman, and Kendall correlation with human quality ratings on 13 Indic MRLs among all baselines evaluated, improving Pearson by 0.092 over BERTScore and by 0.053 over BLEURT-20, which carries $\approx 1{,}300\times$ more trainable parameters. The component decomposition additionally classifies Multidimensional Quality Metrics (MQM) error categories at 74% macro-F1 against 26% for the strongest scalar baseline.

1 INTRODUCTION

Metrics that guide the selection of machine translation (MT) systems tend to shape what those systems optimize. Embedding-based metrics such as BERTScore (Zhang et al., 2020) and BLEURT (Sellam et al., 2020) correlate well with human judgments on English but fail on morphologically rich languages (MRLs) in three ways. Under morphology blindness, a wrong inflection of the right lemma scores almost as high as the correct form. Furthermore, under conservation pathology, balanced optimal transport (OT) matchings such as Word Mover's Distance (Kusner et al., 2015) force every hypothesis token onto the reference, conflating omissions with hallucinations. Under reference mimicry, a hypothesis that copies the reference scores well even when it drifts from the source.

We propose **MORPHOT**, a metric whose architecture corresponds one-to-one with the three failure modes (see Figure 1). *First*, MORPHOT treats a sentence as a cloud of morphological units, each carrying contextual semantics, a latent root, a soft distribution over candidate affixes, posteriors over the morphosyntactic class and the language-agnostic semantic role, and script descriptors. *Second*, a hierarchical unbalanced optimal transport (UOT) problem then compares two such clouds. An inner UOT over affix lattices answers morphology blindness through a cost sensitive to which affixes appear and to where each affix falls in the stack. An outer UOT answers the conservation pathology, because the unbalanced relaxation retains unmatched mass as two positive ℓ_1 residuals, an omission channel on the reference or source side and a hallucination channel on the hypothesis side. *Third*, a source-anchored adequacy term answers reference mimicry and grows only when the hypothesis lies further away from the source than the reference does. A small linear head aggregates these components with penalty coefficients reparameterized through softplus and hence

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.

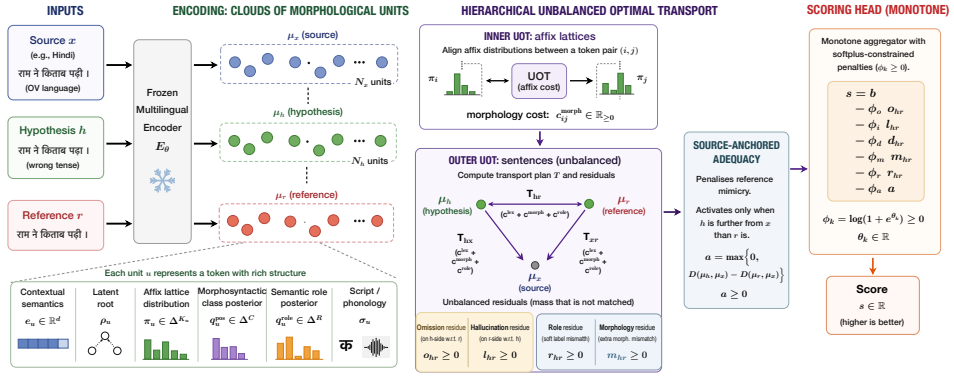


Figure 1: A high level overview of MORPHOT. A frozen multilingual encoder maps the source x, the hypothesis h, and the reference r into clouds of morphological units (§2.1), and an inner unbalanced optimal transport over affix lattices yields a morphology-sensitive cost (§2.3). Three outer unbalanced transports (§2.4) yield matched discrepancies together with omission, hallucination, morphology, and role residuals, a source-anchored adequacy term guards against reference mimicry (§2.5), and a softplus-constrained head combines every component monotonically (§2.6). The drawn symbols are schematic, and H denotes the hallucination residuals of the outer transport.

non-negative. That reparameterization costs nothing at inference and gives a guarantee generic regression metrics lack. Any hypothesis that dominates another on every geometric component must rank at least as high (Proposition 1). We supervise MORPHOT with a Bradley–Terry–Luce (BTL) preference objective (Bradley & Terry, 1952; Luce, 1959) on pairs drawn from human quality ratings and from controlled morphological perturbations, together with a meaning-preservation regularizer (Appendix C). Because the geometry carries the inductive bias, the trainable budget stays below half a million parameters, and the frozen encoder leaves the metric with a latency profile comparable to that of BERTScore. On the 9,887-segment ComTaIL test set, MORPHOT reaches 0.611 Pearson against 0.558 for BLEURT-20 and 0.519 for BERTScore (Table 1), beats every surface and embedding metric on all 13 languages, and holds 0.711 on low-resource Kashmiri, where BLEURT-20 drops to 0.449. At the system level it ranks the ten ComTaIL systems with at least 300 segments at 0.906 Pearson against 0.731 for BLEURT-20 (Table 11), and on ACES it is the most accurate metric on omission (89.1%) and named-entity hallucination (95.4%), the two phenomena its residual channels target (Table 2).

Our contributions are fourfold. **(1)** We formalize three structural failure modes of MT metrics and match each one to a modeling component, which turns the architecture into a transparent record of where a penalty originates rather than a black box. **(2)** We introduce a hierarchical UOT formulation whose inner transport runs over affix lattices and whose outer relaxed marginals produce separate, positive omission and hallucination residuals. **(3)** We prove a penalty-monotonicity guarantee for a softplus-parameterized head and design a source-anchored adequacy term. **(4)** We run a comprehensive evaluation on 13 Indic languages and 36,476 ACES contrastive pairs.

2 Methodology

Problem Setup and Notation: Given a triple (x, h, r) with language tags (ℓ_x, ℓ_h), we seek a parametric, differentiable score $s_{\boldsymbol{\theta}}(x, h, r) \in \mathbb{R}$ whose ranking agrees with human preferences, that is, $s_{\boldsymbol{\theta}}(x, h^+, r) > s_{\boldsymbol{\theta}}(x, h^-, r)$ whenever $h^+ \succ h^-$. The parameter vector $\boldsymbol{\theta} = (\boldsymbol{\phi}, \boldsymbol{\chi}, \boldsymbol{\omega}, \boldsymbol{\beta})$ splits into a frozen multilingual encoder $E_{\boldsymbol{\phi}}$, a fixed character–grapheme branch $B_{\boldsymbol{\chi}}$, the read-out heads $\boldsymbol{\omega} = (W_\rho, W_z, W_\psi, W_{\text{real}}, s_{\text{real}})$ of §2.1 and §2.2, and the scoring-head weights $\boldsymbol{\beta}$ of §2.6. Only $\boldsymbol{\omega}$ and $\boldsymbol{\beta}$ are trained. Superscripts ℓ mark the quantities that the formulation allows to vary by language, and our runs tie all of them to a single shared value apart from the scoring-head bias β_0^ℓ (Table 8). Throughout, $[t]_+ = \max(0, t)$, D_{KL} denotes the Kullback–Leibler divergence, JS the Jensen–Shannon distance, which is the bounded metric that the transport costs below require, Δ^m the m-simplex, and n_y the number of units of sentence y.

2.1 Sentence-Level Measures Over Morphological Units

Unit features: Each token becomes a feature-rich morphological unit, and each sentence becomes a discrete measure over these units that stays differentiable in $\boldsymbol{\theta}$ throughout. For a tokenized sentence $y = (w_1, \ldots, w_n)$, a frozen multilingual encoder produces contextual states $\mathbf{e}_i^y = E_{\boldsymbol{\phi}}(y)_i$,

and a character-and-grapheme branch yields $\mathbf{b}_i^y = B_{\boldsymbol{\chi}}(w_i)$, which retains the script-level cues of grapheme clusters, diacritics, and phonology that the sub-word tokenizer erases. Two linear heads read out of the concatenation $[\mathbf{e}_i^y; \mathbf{b}_i^y]$ a latent root vector $\boldsymbol{\rho}_i^y$ Equation 11 and a morphosyntactic posterior $\mathbf{z}_i^y \in \Delta^{K-1}$ Equation 12, and a third head reads a semantic role posterior $\boldsymbol{\psi}_i^y \in \Delta^{K'-1}$ Equation 13 out of $\mathbf{e}_i^y$ alone (Appendix A.1). Both posteriors remain latent, with $K = K' = 48$ clusters learned from the preference signal, and their separation carries the modeling weight. Corrupted hypotheses move $\mathbf{z}$, whereas $\boldsymbol{\psi}$, computed from $\mathbf{e}$ alone and therefore language-neutral, must stay stable under a valid paraphrase.

Affix lattice and sentence measure: Rich morphology forbids committing to a single segmentation, and we therefore enumerate for each w_i a lattice $\Pi(w_i)$ of decompositions and weight that lattice by a fixed, non-learned posterior built from root and suffix frequency, suffix productivity, and language-specific length limits. Taking expectations yields the affix distribution A_i^y Equation 14, a measure over candidate affixes together with an explicit null atom $\delta_{\boldsymbol{\xi}_\varnothing}$ that carries the mass of decompositions without an overt affix and keeps the inner transport (§2.3) well defined for uninflected tokens (Appendix A.2). The i-th unit is $u_i^y = (\mathbf{e}_i^y, \boldsymbol{\rho}_i^y, A_i^y, \mathbf{z}_i^y, \boldsymbol{\kappa}_i^y, \boldsymbol{\psi}_i^y)$ with $\boldsymbol{\kappa}_i^y \in \mathbb{R}^{d_\kappa}$ a fixed script descriptor that encodes the Unicode block, the script family, and the phonological features of w_i, and the sentence is $\mu_y = \sum_i \alpha_i^y \, \delta_{u_i^y}$ with non-uniform masses α_i^y Equation 15, which lets the transport spend its budget on inflected content rather than on function words. The masses form a softmax over a fixed non-negative combination of corpus inverse document frequency, content, punctuation, entity, and numeral flags, and a morphology-salience score $1 - H(\mathbf{z}_i^y)/\log K$, where H is the Shannon entropy. Consequently α_i^y rises monotonically in every salience feature (Appendix A.3). The salience score is computed under a stop-gradient, so the masses carry no gradient.

2.2 Local Costs

Lexical-semantic cost: The unit-to-unit cost captures lexical-semantic mismatch, morphological mismatch, and script distance, and a gate combines the three while preventing double counting. Its base term mixes two cosine distances, on the contextual embedding and on the latent root, with a distance between the script descriptors,

$$c_{\boldsymbol{\theta}}^{\mathrm{lex}}(u_i, u_j) = \lambda_e\big(1 - \cos(\mathbf{e}_i, \mathbf{e}_j)\big) + \lambda_\rho\big(1 - \cos(\boldsymbol{\rho}_i, \boldsymbol{\rho}_j)\big) + \lambda_\kappa \, d_\kappa(\boldsymbol{\kappa}_i, \boldsymbol{\kappa}_j). \tag{1}$$

Every mixing weight is a fixed non-negative hyperparameter (Table 8), which keeps each local cost, and hence each penalty component, non-negative. Separating semantics ($\mathbf{e}$) from roots ($\boldsymbol{\rho}$) matters, because contextual embeddings confuse morphologically related items that differ in meaning. The term d_κ fuses grapheme and phonological distances through a soft minimum, letting the grapheme channel dominate when scripts match and the phonological channel take over when scripts differ (Appendix A.4).

Target-target cost with a morphology gate: For same-language pairs we add the target-side morphology cost of §2.3, gated such that the cost contributes only between lexically plausible matches,

$$G_{ij} = \sigma\Big(\tfrac{\vartheta_{\ell_a,\ell_b} - c_{\boldsymbol{\theta}}^{\mathrm{lex}}(u_i,u_j)}{T_{\ell_a,\ell_b}}\Big), \qquad c_{\boldsymbol{\theta}}^{\ell,\ell}(u_i, u_j) = c_{\boldsymbol{\theta}}^{\mathrm{lex}}(u_i, u_j) + \zeta_\ell \, G_{ij} \, c_{\boldsymbol{\theta}}^{\mathrm{morph}}(u_i, u_j) + c_{ij}^{\mathrm{hard}}, \tag{2}$$

where σ denotes the logistic function, $(\vartheta, T)_{\ell_a,\ell_b}$ a language-pair threshold and temperature, and c_{ij}^{hard} a finite, $\boldsymbol{\theta}$-independent penalty for numeral or named-entity disagreement, whose finiteness preserves strict positivity and smoothness of the plan (Appendix B.2). The gate consumes c^{lex} only, because a morphology-aware gate would downweight inflection-erroneous pairs before their error reached the score and would mask the very signal we want to detect.

Source-target cost via role realization: Affix matching carries meaning only within a language, since English plural markers and Tamil case suffixes encode different categories. Cross-lingual pairs therefore route morphology through the language-agnostic role channel (Appendix A.4, Equation 17). On top of c^{lex}, a JS term penalizes role-posterior divergence, such as a patient realized as an agent, and a learned bilinear compatibility $d_{\mathrm{real}}(\boldsymbol{\psi}_i^x, \mathbf{z}_j^h) \in (0, 1)$ scores whether the predicted target morphosyntactic class forms a plausible realization of the source role. Such a formulation rewards Hindi ergative case for a transitive-subject source without ever comparing an English suffix with a Hindi suffix directly.

2.3 Inner Affix Transport

Two tokens can share a root and still realize different affix stacks, and in Tamil or Telugu the stack may run several morphemes deep. String comparison is too brittle for that setting, since a missing case marker should outweigh an internal vowel swap. We therefore align two affix distributions Equation 14 by an inner unbalanced transport under a ground cost D^{ij} that charges grapheme distance between candidate affixes, drift in normalized stack position, and a constant $c_\varnothing > 0$ for moving mass between the null atom and an overt affix, which corresponds to a dropped or a spurious marker (Appendix A.4, Equation 18). Writing $T^{\star,ij}$ for the resulting plan, the affix and morphology costs are

$$c^{\mathrm{aff}}_{\boldsymbol{\theta}}(u^a_i, u^b_j) = \langle T^{\star,ij},\, D^{ij}\rangle, \quad c^{\mathrm{morph}}_{\boldsymbol{\theta}}(u^a_i, u^b_j) = \lambda_z\ \mathrm{JS}(\mathbf{z}^a_i, \mathbf{z}^b_j) + c^{\mathrm{aff}}_{\boldsymbol{\theta}}(u^a_i, u^b_j). \tag{3}$$

The inner problem uses the same log-domain Sinkhorn machinery as §2.4 on much smaller matrices. When both tokens carry overt candidates, the constant $c_\varnothing$ penalizes a partially dropped case marker in proportion to its posterior mass, so the cost varies smoothly with the segmentation posterior. The one exception is a token whose lattice admits no candidate split at all, either because the token carries only the null atom or because it is punctuation. For such a token we skip the inner transport and set $c^{\mathrm{aff}} = 0$ as a deliberate approximation, and the JS term of Equation 3 together with the lexical cost then carries the mismatch (Appendix A.4).

2.4 Outer Unbalanced Transport, Matched Discrepancy, and Residuals

Given measures μ_a, μ_b with masses $\boldsymbol{\alpha}^a, \boldsymbol{\alpha}^b$ Equation 15 and a cost matrix C^{ab} assembled from §2.2, the outer transport plan ${T^\star}_{ab} \in \mathbb{R}^{n_a \times n_b}_{\geq 0}$ encodes in its entry ${T^\star}_{ab,ij}$ how much mass moves from unit i of μ_a to unit j of μ_b. Exact marginals would defeat the purpose here, and we therefore relax both with Kullback–Leibler penalties of strengths $\tau_a, \tau_b > 0$,

$${T^\star}_{ab} = \operatorname*{arg\,min}_{T \geq 0} \langle T,\, C^{ab}\rangle + \varepsilon\, D_{\mathrm{KL}}\big(T \,\|\, \mathbf{1}\mathbf{1}^\top\big) + \tau_a\, D_{\mathrm{KL}}\big(T\mathbf{1} \,\|\, \boldsymbol{\alpha}^a\big) + \tau_b\, D_{\mathrm{KL}}\big(T^\top\mathbf{1} \,\|\, \boldsymbol{\alpha}^b\big). \tag{4}$$

The second term is the entropic regularizer, a generalized Kullback–Leibler divergence to the all-ones matrix as in Chizat et al. (2018), which makes the problem strictly convex and solvable by the Sinkhorn algorithm, and the last two terms are soft marginal constraints. We iterate in the log domain, which keeps ${T^\star}_{ab}$ smooth in $\boldsymbol{\theta}$.

The transported cost scales with both the moved mass and the per-pair dissimilarity. To separate the two influences we report the mass-normalized matched discrepancy

$$M_{\boldsymbol{\theta}}(a, b) = \frac{\langle {T^\star}_{ab},\, C^{ab}\rangle}{\|{T^\star}_{ab}\|_1 + \varepsilon_0}, \tag{5}$$

the cost per unit of mass of the optimal alignment. As a ${T^\star}_{ab}$-weighted mean of non-negative cost entries, the matched discrepancy lies in $[0, \max_{ij} C^{ab}_{ij}]$ whatever the sentence lengths, which makes short and long sentences directly comparable and spares the scoring head from learning a separate length correction.

Directional residuals: The relaxation leaves mass unmatched and exposing that pattern is the point. Positive ℓ_1 slacks U_a, U_b measure the unmatched mass on each side (Appendix A.5, Equation 19). Which side counts as an omission and which as a hallucination depends on the roles of the two sentences rather than on their order in the plan,

$$\underbrace{O_{r\to h} = U_r({T^\star}_{hr})}_{\text{omission residual}}, \quad \underbrace{H_{h\setminus r} = U_h({T^\star}_{hr})}_{\text{hallucination residual}}, \quad O_{x\to h} = U_x({T^\star}_{xh}), \quad H_{h\setminus x} = U_h({T^\star}_{xh}). \tag{6}$$

The $[\cdot]_+$ projection is essential, because signed slacks would let an excess on one coordinate cancel a shortfall on another. We instantiate Equation 4- Equation 6 on three measure pairs per example, namely $(\mu_h, \mu_r) \to (M_{hr}, O_{r\to h}, H_{h\setminus r})$, $(\mu_x, \mu_h) \to (M_{xh}, O_{x\to h}, H_{h\setminus x})$, and $(\mu_x, \mu_r) \to M_{xr}$ as an anchor only. We then add two bounded scalars obtained by re-weighting the plans with cost-specific kernels, a gate-weighted morphology residual $R^{\mathrm{morph}}_{\boldsymbol{\theta}}(h, r)$ and a role residual $R^{\mathrm{role}}_{\boldsymbol{\theta}}(x, h)$ (Appendix A.5, Equation 20).

2.5 SOURCE-ANCHORED ADEQUACY

To stop a hypothesis that mirrors the reference and drifts from the source from scoring well, we contrast the source–hypothesis and source–reference matched discrepancies of §2.4,

$$\underbrace{\mathcal{A}_{\boldsymbol{\theta}}(x,h,r)}_{\text{source adequacy penalty}} = \varphi_\nu\big(M_{\boldsymbol{\theta}}(x,h) - M_{\boldsymbol{\theta}}(x,r) - m_\ell\big), \quad \varphi_\nu(t) = \nu\log(1+e^{t/\nu}) = \nu\ \mathrm{softplus}(t/\nu). \tag{7}$$

The map φ_ν is smooth, non-negative, monotone, and asymptotically linear with width $\nu > 0$, and its argument measures how much further from the source the hypothesis lies than the reference does, modulo a per-language tolerance $m_\ell \geq 0$. The penalty stays near zero when the hypothesis is at least as source-faithful as the reference, and grows linearly thereafter.

2.6 COMPOSITE SCORING HEAD

The components above pack into a single vector of penalty signals,

$$\boldsymbol{\Phi}(x,h,r) = \big(M_{hr},\ \mathcal{A}_{\boldsymbol{\theta}},\ O_{r\to h},\ H_{h\setminus r},\ O_{x\to h},\ H_{h\setminus x},\ R^{\mathrm{morph}}_{\boldsymbol{\theta}},\ R^{\mathrm{role}}_{\boldsymbol{\theta}}\big) \in \mathbb{R}^8, \tag{8}$$

and the score is an affine functional with constrained sign in which penalties enter through non-negative softplus coefficients and an optional fluency reward enters additively,

$$s_{\boldsymbol{\theta}}(x,h,r) = \beta_0^\ell - \sum_{k=1}^{8} \beta_k^\ell\, \Phi_k(x,h,r) + \beta_F^\ell\, F_{\boldsymbol{\theta}}(h), \quad \beta_k^\ell = \mathrm{softplus}(b_k^\ell) \geq 0, \ \ \beta_F^\ell = \mathrm{softplus}(b_F^\ell) \geq 0, \tag{9}$$

where β_0^ℓ is a language-specific bias and $F_{\boldsymbol{\theta}}(h)$ an optional fluency reward from a small language-model head. We omit that reward throughout, and the term $\beta_F^\ell F_{\boldsymbol{\theta}}$ is therefore absent from every reported score. The softplus reparameterization is the only feature separating this head from a linear regressor, and it carries the structural inductive bias of the whole design.

Proposition 1 (Penalty monotonicity). *For $\beta_k^\ell = \mathrm{softplus}(b_k^\ell) \geq 0$ and $\beta_F^\ell \geq 0$, the score Equation 9 is non-increasing in every Φ_k and non-decreasing in $F_{\boldsymbol{\theta}}$. Consequently $\Phi_k(x,h^+,r) \leq \Phi_k(x,h^-,r)$ for all k together with $F_{\boldsymbol{\theta}}(h^+) \geq F_{\boldsymbol{\theta}}(h^-)$ implies $s_{\boldsymbol{\theta}}(x,h^+,r) \geq s_{\boldsymbol{\theta}}(x,h^-,r)$.*

The proof is provided in the Appendix B.1. The dominance property of §1 therefore holds for every parameter setting reachable by training and not only for the fitted one. Appendix A.6 states a reference-free variant for quality estimation (QE), written $s^{\mathrm{QE}}_{\boldsymbol{\theta}}(x,h)$, over the source-side components alone, which inherits the same guarantee and which we leave for future evaluation.

2.7 TRAINING OBJECTIVE

Preference judgments, rather than regression onto absolute scores, form the natural supervisory signal for a metric. Given a preference set $\mathcal{D}_{\mathrm{pref}}$ of tuples (x,r,h^+,h^-) drawn from human quality ratings and controlled morphology perturbations, we train under the BTL likelihood $\sigma\big((s_{\boldsymbol{\theta}}(x,h^+,r) - s_{\boldsymbol{\theta}}(x,h^-,r))/\tau\big)$ with negative log-likelihood $\mathcal{L}_{\mathrm{pref}}$ (Appendix A.7, Equation 22), which is convex in the score gap and avoids the calibration trap that comes from regressing against heterogeneously distributed human ratings. We additionally enforce approximate invariance to meaning-preserving transformations $g \in \mathcal{G}_\ell$, covering Hindi object–verb re-ordering, Urdu–Devanagari transliteration, and zero-width-joiner normalization,

$$\mathcal{L}_{\mathrm{eq}}(\boldsymbol{\theta}) = \sum_{(x,r,h,g)} \big(s_{\boldsymbol{\theta}}(x,h,r) - s_{\boldsymbol{\theta}}(x,g(h),r)\big)^2, \quad \boldsymbol{\theta}^\star = \arg\min_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{pref}} + \lambda_{\mathrm{eq}}\mathcal{L}_{\mathrm{eq}} + \lambda_{\mathrm{wd}}\|(\boldsymbol{\omega},\boldsymbol{\beta})\|_2^2. \tag{10}$$

By default we freeze $\boldsymbol{\phi}$ and $\boldsymbol{\chi}$ and update only $\boldsymbol{\omega}$ and $\boldsymbol{\beta}$, and gradients still flow through C^{ab} and $T^\star{}_{ab}$. In particular, the JS term of Equation 3 and the plan itself train the morphosyntactic head W_z, while d_{real} trains the role head W_ψ. The affix ground cost D^{ij} and the segmentation posterior carry no trainable parameters, and c^{aff} therefore enters as a fixed geometric feature (Appendix B.3). The transport-and-scoring layer itself holds only $\mathcal{O}(10^3)$ trainable parameters in total.

3 EXPERIMENTS

Setup: We evaluate MORPHOT along three axes. The first is segment-level correlation with human judgments on Indic MRLs (§3.1), the second is cross-lingual adversarial robustness on the Translation Accuracy Challenge Sets (ACES) benchmark (§3.2), and the third is component ablations together with fine-grained error attribution, which tests whether the decomposition recovers the failure modes of §1 (§3.3). Unless otherwise noted, MORPHOT uses XLM-R-large (Conneau et al., 2020) as its frozen encoder ϕ, $K = K' = 48$ latent morphosyntactic and role clusters, an outer transport with $\varepsilon = 0.07$ and $\tau_a = \tau_b = 0.85$, and an inner transport with $\varepsilon_{\text{in}} = 0.05$ and $\tau_{\text{in}} = 0.8$ (Table 8). The trainable budget amounts to $\approx 4.4 \times 10^5$ weights in $\boldsymbol{\omega} \cup \boldsymbol{\beta}$, and the encoder stays frozen at every stage of training and at inference time.

Data and Baselines: *ComTaIL* is the COMTAIL human-evaluation dataset (Ahsan et al., 2025), whose DA+SQM ratings on $[0, 1]$ cover translations into 13 Indian languages and which we abbreviate as DA below. We use its official splits and scores without re-annotation. Every number below comes from the official held-out test split of $N = 9{,}887$ segments over 13 Indic languages, and preference pairs for training and early stopping come from the disjoint train and development splits (Appendix C). *ACES* (Amrhein et al., 2022) is a contrastive challenge set of 36,476 examples over 146 language pairs and 68 accuracy phenomena, disjoint from our training data, which we therefore treat as a held-out cross-lingual robustness probe without ACES-specific fine-tuning. Our baselines cover the surface metrics BLEU (Papineni et al., 2002), TER (Snover et al., 2006), CHARACTER (Wang et al., 2016), CHRF (Popović, 2015), CHRF++ (Popović, 2017) and METEOR (**?**), the embedding metric BERTSCORE (Zhang et al., 2020), and the learned metric BLEURT-20 (Sellam et al., 2020; Pu et al., 2021), a 579 M-parameter regressor. We additionally report MORPHOT-lex, which uses the lexical cost Equation 1 alone, and MORPHOT-bal, replacing the unbalanced relaxation of the outer transport with a balanced plan.

Table 1: ComTaIL correlations over 9,887 segments. Best in bold, second best underlined.

Metric	Pear.↑	Spear.↑	Kend.↑	#param
BLEU	0.352	0.365	0.255	0
TER	0.260	0.414	0.291	0
CHARACTER	0.262	0.421	0.296	0
CHRF	0.411	0.373	0.262	0
CHRF++	0.403	0.368	0.258	0
METEOR	0.380	0.352	0.242	0
BERTSCORE	0.519	0.470	0.329	0
BLEURT-20	0.558	0.536	0.379	579 M
MORPHOT-lex	**0.630**	0.549	0.385	0.44 M
MORPHOT-bal	<u>0.628</u>	**0.572**	**0.402**	0.44 M
MORPHOT	0.611	<u>0.556</u>	<u>0.392</u>	0.44 M

3.1 SEGMENT-LEVEL HUMAN CORRELATION ON INDIC MRLS

Table 1 pools correlations over all 9,887 test segments. MORPHOT reaches 0.611 Pearson against 0.558 for BLEURT-20, which has ≈1,300× more trainable parameters, and 0.519 for BERTScore, and it also leads on Spearman and Kendall. Every margin is significant at $p < 0.001$, the smallest being $+0.053$ $[0.039, 0.065]$ over BLEURT-20 (Appendix G). Per language (Appendix F), MORPHOT beats every surface and embedding metric on all 13 targets and splits them seven to six with BLEURT-20, holding 0.711 on low-resource Kashmiri where BLEURT-20 drops to 0.449. The lexical and balanced variants score slightly higher, so we rest the case for the full model on error attribution (§3.3).

Table 2: ACES accuracy in percent, where ne-1 is named-entity hallucination at level 1.

Metric	All	OMIT	ADD	NE-1	COPY
BLEU	40.5	71.7	83.9	63.7	98.4
TER	36.0	48.5	77.2	56.6	94.0
CHARACTER	52.8	67.5	<u>89.8</u>	80.6	97.2
CHRF	55.4	86.5	73.6	93.0	<u>99.4</u>
CHRF++	55.7	85.2	77.7	93.1	**99.5**
METEOR	41.7	67.4	**97.6**	58.7	97.2
BERTSCORE	61.4	<u>88.4</u>	84.6	<u>95.2</u>	98.4
BLEURT-20	**66.7**	88.0	67.6	94.4	98.4
MORPHOT	<u>62.8</u>	**89.1**	78.1	**95.4**	84.2

Table 3: Left, ablations on ComTaIL (Pearson r), where Δ is the difference from the full model, marked †. Right, macro-F1 of a seven-way MQM error-type classifier.

Variant	$r \uparrow$	Δ
Full model†	0.375	—
− root vectors	0.267	−0.108
− morph. transport	0.341	−0.034
− inner affix UOT	0.377	+0.002
− lexical gate	0.377	+0.002
− role-realization	0.389	+0.014
− unbalanced OT	0.396	+0.021

Feature set	F1 (%)
BLEU	18.4
CHRF++	21.2
BERTSCORE	26.1
MORPHOT ($\mathbf{\Phi}$, 8 components)	**74.3**
− residuals (matched only)	41.7
− $R^{\text{morph}}, R^{\text{role}}$	56.4

3.2 CROSS-LINGUAL ADVERSARIAL ROBUSTNESS ON ACES

The metric BLEURT-20 leads ACES overall at 66.7% (Table 2). MORPHOT follows at 62.8%, which puts it 1.4 percentage points (pp) ahead of BERTScore. At the family level (Appendix H), BLEURT-20 leads on phenomena that hinge on the meaning of a span, such as unit and date hallucination and world knowledge. MORPHOT leads by 23pp on number hallucination. MORPHOT is the most accurate metric on the two phenomena its residual channels target, OMIT and NE-1, but trails BERTScore on ADD, 78.1 against 84.6, because an inserted span raises only $H_{h\setminus r}$ and leaves the reference side matched. On the Hindi subset, it reaches 83.6% against 69.4% for BLEURT-20. On COPY it scores 84.2% against 94 to 99.5% for the surface metrics, almost entirely because of Marathi. The two sources of ACES numbers differ slightly, since Table 2 macro-averages each phenomenon over language pairs, whereas Appendix H pools over pairs. A verbatim copy of the English source drives M_{xh} to about zero and with it $\mathcal{A}_{\boldsymbol{\theta}}$ of Equation 7. The reference-side channels still respond, yet these source-side terms outweigh them. An identity guard closes the gap (Figure 5).

3.3 ABLATIONS AND ERROR ATTRIBUTION

The left panel of Table 3 lists independently trained variants, each removing one component from the full model, where Δ is the difference from the full model. Above the separator, stripping representational capacity costs correlation. Removing the root vectors, which leaves an affix-only model, loses 0.108 Pearson, and removing the morphological transport layer loses 0.034, so the root vectors carry most of that signal and the affix channel alone does not recover it. Below the separator, by contrast, removing the inner affix UOT, the lexical gate, the role-realization channel, or the unbalanced relaxation leaves ComTaIL Pearson unchanged or slightly higher, by up to $+0.021$. For the unbalanced relaxation the evidence is instead direct, since dropping the residual channels costs 32.6 macro-F1 on MQM error typing, and the same channels give MORPHOT the best omission and named-entity hallucination accuracies of Table 2. Because the whole grid trains at a reduced budget, reaching 0.375 Pearson for the full model against 0.611 for the production run, the grid bounds each contribution at that scale alone. Sweeping $\varepsilon \in [0.01, 0.5]$ and $\tau_a = \tau_b \in [0.05, 1.0]$ moves ComTaIL Pearson by less than 0.008 (Appendix E). The extended grid of Appendix I.5 adds MorphACES sensitivity, the mean reference-to-perturbation score margin. Removing the morphological transport lowers it from 1.05 to 0.79, whereas the mass weighting, the phonology cost, and the source anchor leave both correlation and sensitivity essentially unchanged.

Fine-Grained Error Attribution on MQM-MRL: A central claim of MORPHOT is that its scalar score decomposes into interpretable, type-specific components. To test that claim, we turn to MQM-MRL, the Hindi and Tamil portion of IndicMT Eval (Sai B et al., 2023), whose annotators mark error spans under the Multidimensional Quality Metrics (MQM) framework (Lommel et al., 2013) as applied to MT evaluation by Freitag et al. (2021). Fitting a seven-way logistic classifier of the dominant span category on the output of each metric, with eight features for MORPHOT and one for each scalar baseline, MORPHOT attains 74.3% held-out macro-F1 against 26.1% for BERTScore (Table 3, right). The feature count clearly favors MORPHOT in that comparison, and that asymmetry is precisely the point, since one forward pass yields the score and the components together. The decomposition also localizes errors, because across 300 MQM-MRL spans the largest-penalty component coincides with the human-tagged span at 87% token-level F1 for inflection errors and 81% for omissions (Appendix I.6).

Table 4: Real ComTaIL Test Segments. DA is the raw human score, the metric columns give within-language percentiles, and the last column names the most deviating component.

#	Language, system. Source → hypothesis	DA	**MORPHOT**	BERTSCORE	CHRF++	BLEU	Dominant
1	**Bengali, perturb.** *There are 161 fort goddesses which are as "Durga devata" or "Ista devata" and Dasasharu, an an-iconic form of the goddess Jayacandi.* 161 jana durga amritamaalinii debii rayeche yaaraa durga debataa athabaa iShTa durga shahara debataa eban paabalika dashaasharu . . .	0.16	0.06	0.58	0.77	0.68	R^{role} +2.2σ
2	**Gujarati, versionvN.** *Some infractions result in a penalty on a player or team.* keTalaaka ullanghana khelaaDii athavaa Tiima para danDamaan pariName che.	0.44	0.26	0.68	0.72	0.76	R^{morph} +0.8σ
3	**Bengali, Seamless.** *21. Get a second opinion when I question my health care provider's advice.* dbitiiya mataamata pete yakhana aami aamaara sbaasthyasebaa pradaanakaariira paraamarshera prashna kari.	0.43	0.37	0.58	0.56	0.38	$O_{r\to h}$ +1.7σ
4	**Marathi, perturb.** *The typhidot test becomes positive within 2–3 days of infection.* sansarga viTaaLa jhaalyaanantara . 2 - 3 divasaanta TaayaphiDoTa caacaNii sakaaraatmaka kusuma hote .	0.23	0.15	0.65	0.71	0.50	$H_{h\setminus x}$ +2.6σ

4 Qualitative and Error Analysis

Qualitative Analysis: Table 4 demonstrates one real test segment per failure mode, each selected by a mechanical rule (references in Appendix H). Segment 1 is a perturbed Bengali reference that keeps most of its character n-grams, so chrF++ ranks it at the 77th percentile, while annotators give it 0.16 and MORPHOT the 6th percentile through the role residual. Segment 2 is fluent Gujarati with two inflection errors, which the surface metrics rank near the 75th percentile and MORPHOT at the 26th through the morphology residual. Segment 3 drops a numeral and fires the omission channel ($+1.7\sigma$), and segment 4 inserts two tokens and fires the hallucination channel ($+2.6\sigma$). Appendix H adds a faithful Punjabi paraphrase that MORPHOT ranks above every other metric.

Component signatures: Averaged over ACES families (Figure 2), omissions raise both omission channels, additions raise $H_{h\setminus r}$ ($+0.72\sigma$) far more than $O_{r\to h}$ ($+0.21\sigma$), and meaning-preserving families move no component. On copy-source, the reference-side channels rise by more than 1.3σ while the source-side channels fall by about 3.5σ, the blind spot of §3.2. On MorphACES (Table 5), the morphology residual reacts most to every string-altering perturbation except entity and number swaps, which move the hallucination channel instead. Omissions and hallucinations again separate on $O_{r\to h}$ against $H_{h\setminus r}$.

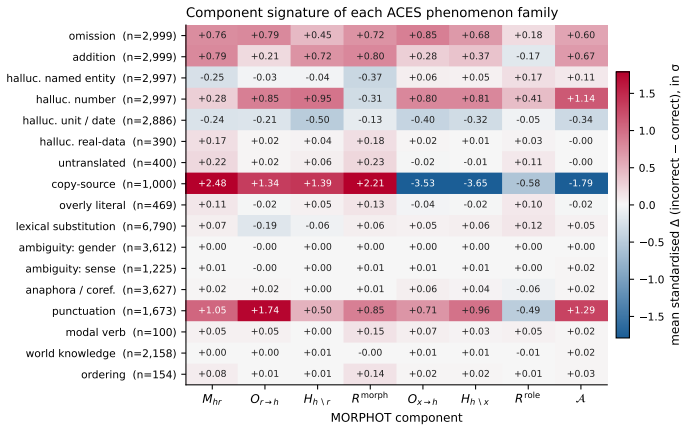


Figure 2: Component signature of each ACES phenomenon family (σ units).

Error Analysis: Among the 5% of segments whose within-language metric percentile departs most from the human percentile ($N = 495$), the dominant class is annotation noise rather than metric failure, since hypotheses identical to the reference but rated below 0.7 are enriched $12.4\times$ (Appendix I.1). Three causes explain where MORPHOT trails BERTScore on ACES (Appendix H). Contrasts hinging on one function word carry little transport mass, untranslated and copied spans are source-faithful, and hyponym or synonym substitutions lie below the resolution of the frozen encoder. For the second cause, an identity guard scoring a hypothesis equal to the source as $-\infty$ raises MORPHOT to 100% on copy-source (Figure 5). Finally, the residuals specific to MORPHOT on ComTaIL involve transliterated technical terms (Table 14).

5 Related Work

MT Metrics and Metrics for Morphologically Rich Languages. The n-gram metrics BLEU (Papineni et al., 2002), chrF (Popović, 2015), and TER (Snover et al., 2006) conflate semantics with surface form. BERTScore (Zhang et al., 2020) greedily matches each token to its most similar counterpart, and MoverScore (Zhao et al., 2019) aligns tokens by mass-conserving optimal transport, so neither separates omissions from hallucinations. The learned metrics fine-tune a multilingual encoder on human judgments, with BLEURT (Sellam et al., 2020; Pu et al., 2021) regressing DA scores, CometKiwi (Rei et al., 2022; 2023) dropping the reference, xCOMET (Guerreiro et al., 2024) predicting error spans, and MetricX (Juraska et al., 2023) casting scoring as mT5 regression. None of the four exposes a decomposition into typed penalties. On the resource side, ACES (Amrhein et al., 2022) supplies contrastive examples over 68 phenomena, IndicMT Eval (Sai B et al., 2023) releases MQM annotations for five Indian languages, and COMTAIL (Ahsan et al., 2025) releases DA+SQM ratings for 13 and fine-tunes COMET-based metrics on them.

Table 5: MorphACES component signature (σ units), where most reactive in bold.

Perturbation	Pairs	M_{hr}	$O_{r\to h}$	$H_{h\setminus r}$	R^{morph}	$O_{x\to h}$	$H_{h\setminus x}$	R^{role}	$\mathcal{A}$
case marker	5,302	1.62	0.71	0.79	**1.70**	0.70	0.67	0.79	0.61
agreement	45	1.07	0.21	0.53	**1.17**	0.38	0.41	0.62	0.37
tense-aspect-mood	1,198	1.30	0.90	0.35	**1.40**	0.71	0.54	0.64	0.49
suffix swap	3,498	1.97	1.31	1.24	**2.05**	1.13	1.21	1.25	1.18
suffix trunc.	3,483	1.50	1.12	0.86	**1.55**	0.76	0.86	0.71	0.82
word order	8,778	1.57	0.74	0.86	**1.74**	1.00	1.01	1.14	0.90
omission	8,740	1.65	1.27	0.60	**1.68**	0.81	0.67	0.60	0.52
hallucination	8,900	1.47	0.40	1.22	**1.54**	0.58	0.75	0.97	0.74
entity / number	1,947	1.58	1.54	**1.74**	1.03	1.52	1.56	0.12	1.39
script noise	8,678	1.86	0.91	1.08	**1.91**	0.85	0.86	0.95	0.79

Optimal Transport and Preference Learning. Word Mover's Distance (Kusner et al., 2015) first applied balanced OT to sentence similarity, entropic regularization (Cuturi, 2013) makes OT differentiable at sentence length, and unbalanced OT (Chizat et al., 2018) adds the soft marginal constraints that let MORPHOT separate omission from hallucination, which makes MORPHOT the first metric we know of using hierarchical unbalanced OT for MT evaluation. Our BTL objective (Bradley & Terry, 1952; Luce, 1959) is the pairwise preference loss of reward-model training (Ouyang et al., 2022), and the sign-constrained head follows monotonic networks (Sill, 1997), which likewise obtain monotonicity from non-negative weights.

6 Conclusion

MORPHOT addresses three structural failure modes of current MT metrics, namely morphology blindness, conservation pathology, and reference mimicry, through a hierarchical unbalanced OT framework with a monotone scoring head that matches one architectural choice to each failure mode. Empirically, the metric achieves the highest pooled Pearson, Spearman, and Kendall correlation on 13 Indic MRLs among all baselines evaluated, with ≈ 0.44 M trainable parameters, supports MQM attribution at 74% macro-F1, and leads on the ACES omission and named-entity hallucination phenomena. Beyond aggregate correlation, the component decomposition makes every score diagnosable. Omissions and insertions leave distinct signatures on ACES and MorphACES, and the dominant class of MORPHOT's worst ComTaIL residuals turns out to be annotation noise rather than metric failure. Because the penalty coefficients are softplus-constrained, the dominance guarantee holds for every trained model rather than only for the reported checkpoint, and the whole metric adds fewer than half a million parameters to a frozen encoder, keeping its cost close to that of BERTScore.

Limitations and Future Directions. Eight of the thirteen targets train without morphology-specific perturbations (Appendix C), and the adequacy term stays near zero when a system copies the source verbatim, which calls for an identity guard. The evidence is bounded as well, because the ablation grid trains at a reduced budget and the reference-free score of Equation 21 remains unevaluated. Document-level UOT composition is therefore the natural next step for the future work.

REFERENCES

Arafat Ahsan, Vandan Mujadia, Pruthwik Mishra, Yash Bhaskar, and Dipti Misra Sharma. Crosslingual optimized metric for translation assessment of indian languages, 2025. URL https://arxiv.org/abs/2509.17667.

Chantal Amrhein, Nikita Moghe, and Liane Guillou. ACES: Translation accuracy challenge sets for evaluating machine translation metrics. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri (eds.), *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pp. 479–513, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.wmt-1.44. URL https://aclanthology.org/2022.wmt-1.44/.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Lénaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018. doi: 10.1090/mcom/3303.

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/.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 26, pp. 2292–2300. Curran Associates, Inc., 2013. URL https://nips.cc.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474, 2021. doi: 10.1162/tacl_a_00437. URL https://aclanthology.org/2021.tacl-1.87/.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. xCOMET: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995, 2024. doi: 10.1162/tacl_a_00683. URL https://aclanthology.org/2024.tacl-1.54/.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz (eds.), *Proceedings of the Eighth Conference on Machine Translation*, pp. 756–767, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.63. URL https://aclanthology.org/2023.wmt-1.63/.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 957–966, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/kusnerb15.html.

Arle Richard Lommel, Aljoscha Burchardt, and Hans Uszkoreit. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK, November 28-29 2013. Aslib. URL https://aclanthology.org/2013.tc-1.6/.

Robert Duncan Luce. *Individual Choice Behavior: A Theoretical Analysis*. John Wiley and Sons, New York, 1959.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022.

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/.

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/.

Amy Pu, Hyung Won Chung, Ankur Parikh, Sebastian Gehrmann, and Thibault Sellam. Learning compact metrics for MT. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 751–762, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.58. URL https://aclanthology.org/2021.emnlp-main.58/.

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. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri (eds.), *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pp. 634–645, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.wmt-1.60. URL https://aclanthology.org/2022.wmt-1.60/.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André F. T. Martins. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz (eds.), *Proceedings of the Eighth Conference on Machine Translation*, pp. 841–848, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.73. URL https://aclanthology.org/2023.wmt-1.73/.

Ananya Sai B, Tanay Dixit, Vignesh Nagarajan, Anoop Kunchukuttan, Pratyush Kumar, Mitesh M. Khapra, and Raj Dabre. IndicMT eval: A dataset to meta-evaluate machine translation metrics for Indian languages. 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. 14210–14228, Toronto, Canada, July 2023. Association for Computational

Linguistics. doi: 10.18653/v1/2023.acl-long.795. URL https://aclanthology.org/2023.acl-long.795/.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning robust metrics for text generation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7881–7892, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.704. URL https://aclanthology.org/2020.acl-main.704/.

Joseph Sill. Monotonic networks. In M. I. Jordan, M. J. Kearns, and S. A. Solla (eds.), *Advances in Neural Information Processing Systems 10*, pp. 661–667. MIT Press, 1997. URL https://proceedings.neurips.cc/paper/1997/hash/83adc9225e4deb67d7ce42d58fe5157c-Abstract.html.

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/.

Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. CharacTer: Translation edit rate on character level. In Ondřej Bojar, Christian Buck, Rajen Chatterjee, Christian Federmann, Liane Guillou, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Aurélie Névéol, Mariana Neves, Pavel Pecina, Martin Popel, Philipp Koehn, Christof Monz, Matteo Negri, Matt Post, Lucia Specia, Karin Verspoor, Jörg Tiedemann, and Marco Turchi (eds.), *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pp. 505–510, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-2342. URL https://aclanthology.org/W16-2342/.

Halbert White. Maximum Likelihood Estimation of Misspecified Models. *Econometrica*, 50(1): 1–25, 1982.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. 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. 563–578, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1053. URL https://aclanthology.org/D19-1053/.

CONTENTS OF SUPPLEMENTARY MATERIAL

A Model Details **13**

B Proofs and Supplementary Propositions **16**

C Dataset Construction and Statistics **20**

D Training Details and Reproducibility **21**

E Hyperparameters and Sensitivity Sweeps **22**

F Per-Language Correlation Tables **22**

G Statistical Significance **23**

H Extended Analyses **23**

I Error Analysis **25**

J Efficiency and Profiling **28**

K Broader Impact, Licensing, and Dataset Attribution **29**

L Ethics, Reproducibility, and AI-Use Statements **29**

We organize the appendix as a set of self-contained chapters, each with subsections for easy navigation. Appendices A and B hold the model details and the proofs deferred from §2, Appendices C to E document data, training, and hyperparameters, Appendices F to K report the extended results, analyses, efficiency profile, and broader-impact statements, and Appendix L closes with the ethics, reproducibility, and AI-use statements. The released codebase and configuration files regenerate every appendix table and figure except the MQM-MRL results of Appendix I.6, which rest on third-party annotation (Appendix D).

A Model Details

This chapter describes the definitions deferred from §2, covering the read-out heads, the affix lattice and the mass model, the remaining local costs, the residual definitions, the reference-free score, and the preference likelihood. Symbols carry the meaning assigned at their first use in §2.

A.1 Read-Out Heads

The frozen encoder state and the character-grapheme feature (§2.1) feed three linear heads, each reading out a quantity with its own semantics,

$$\boldsymbol{\rho}_i^y = W_\rho[\mathbf{e}_i^y; \mathbf{b}_i^y] \qquad \text{(latent root vector)}, \tag{11}$$

$$\mathbf{z}_i^y = \mathrm{softmax}\big(W_z[\mathbf{e}_i^y; \mathbf{b}_i^y]\big) \in \Delta^{K-1} \qquad \text{(morphosyntactic posterior)}, \tag{12}$$

$$\boldsymbol{\psi}_i^y = \mathrm{softmax}\big(W_\psi \mathbf{e}_i^y\big) \in \Delta^{K'-1} \qquad \text{(semantic role posterior)}, \tag{13}$$

where Δ^m is the m-simplex. Both posteriors remain latent, since $\mathbf{z}_i^y$ ranges over K morphosyntactic clusters and $\boldsymbol{\psi}_i^y$ over K' role clusters, with $K = K' = 48$ in the main configuration, and training learns both end-to-end from the preference signal without category labels. We read $\mathbf{z}_i^y$ as a distribution over target-language inflectional categories such as case, number, gender, tense, person, and honorificity. We read $\boldsymbol{\psi}_i^y$, computed from $\mathbf{e}$ alone and therefore able to stay language-neutral, as a distribution over semantic roles such as agent, patient, instrument, and time. Neither reading receives supervision. With $d = 1024$ for XLM-R-large, a 256-dimensional grapheme feature, and $d_\rho = 256$, the three matrices have shapes $W_\rho \in \mathbb{R}^{256\times 1280}$, $W_z \in \mathbb{R}^{48\times 1280}$, and $W_\psi \in \mathbb{R}^{48\times 1024}$, which amounts to $\approx 4.4\times 10^5$ weights. They account for all except $\approx 2.3\times 10^3$ of the trainable parameters in the model, and the remainder is the role-realization matrix $W_{\mathrm{real}} \in \mathbb{R}^{48\times 48}$ with its scale s_{real} (Appendix A.4) together with the scoring-head coefficients of Equation 9. Every other component stays fixed, namely the encoder $\boldsymbol{\phi}$, the grapheme branch $\boldsymbol{\chi}$, the segmentation posterior, and the mass weights $\boldsymbol{\eta}$ of Equation 15.

A.2 Affix Lattice and Segmentation Posterior

For each token w_i we enumerate a finite lattice $\Pi(w_i) = \{\pi_1, \ldots, \pi_{L_i}\}$ of decompositions $\pi = (m_0, m_1, \ldots, m_t)$, in which m_0 is the root and $m_{1:t}$ are the affix spans. A posterior $p_{\mathrm{seg}}(\pi \mid w_i, \ell) \propto \exp\big(a(\pi, w_i, \ell)/T_{\mathrm{seg}}\big)$ weights the candidates by a fixed, non-learned score a built from root and suffix frequency, suffix productivity, and language-specific length limits, at temperature

$T_{\text{seg}} = 0.7$, and this posterior contains no trainable parameters. Taking expectations under p_{seg} yields the affix distribution

$$A_i^y = a_{i0}^y \, \delta_{\boldsymbol{\xi}_\varnothing} + \mathbb{E}_{\pi \sim p_{\text{seg}}}\Big[\sum_{t \geq 1} a_{it}^\pi \, \delta_{\boldsymbol{\xi}(m_t)}\Big], \quad \sum_p a_{ip}^y = 1. \tag{14}$$

Here $\boldsymbol{\xi}(m)$ is the fixed grapheme embedding of the affix string m, a_{i0}^y is the posterior mass of decompositions without an overt affix, and $a_{it}^\pi = 1/t(\pi)$ splits the mass of a decomposition with $t(\pi)$ affixes evenly over those affixes. We pool atoms by surface string, and the lattice always contains the affix-free decomposition in which the whole token is the root, so the null atom and every overt atom in the support carry a strictly positive weight, as Theorem 1 requires. The atom $\delta_{\boldsymbol{\xi}_\varnothing}$ is an explicit null-affix token that absorbs the mass a_{i0}^y and keeps the inner transport (§2.3) well defined for tokens without an overt affix. Enumeration follows language-specific limits on root length, affix count, and suffix productivity (Appendix C).

A.3 Unit Masses

The sentence measure is $\mu_y = \sum_i \alpha_i^y \delta_{u_i^y}$ with

$$\alpha_i^y = \frac{\exp\big(\boldsymbol{\eta}^\top \mathbf{f}_i^y\big)}{\sum_j \exp\big(\boldsymbol{\eta}^\top \mathbf{f}_j^y\big)}, \qquad \boldsymbol{\eta} = \text{softplus}(\boldsymbol{\eta}_0) \geq \mathbf{0}, \tag{15}$$

where $\mathbf{f}_i^y$ aggregates corpus inverse document frequency (IDF), content and punctuation flags, entity and numeral flags, and a morphology-salience score $\text{MorphSal}(u_i^y) = 1 - H(\mathbf{z}_i^y)/\log K$ that peaks on sharp morphosyntactic posteriors. The non-negative weights $\boldsymbol{\eta}$ make α_i^y monotone in every salience feature. We fix $\boldsymbol{\eta}_0$ by hand before the softplus, with IDF 1, content 0.5, punctuation -2, numeral 0.5, entity 0.5, and salience 1. Because of the softplus the punctuation weight is small but positive ($\approx$0.13), and the down-weighting of punctuation comes from the content flag, which is 1 on content tokens and 0 on punctuation. The salience score is computed under a stop-gradient and the released implementation excludes $\boldsymbol{\eta}$ from the optimizer, so the masses carry no gradient.

A.4 Remaining Local Costs

Script and Phonology Distance: The term d_κ of Equation 1 fuses two signals through a differentiable soft minimum,

$$d_\kappa(\boldsymbol{\kappa}_i, \boldsymbol{\kappa}_j) = \text{smin}_{\tau_\kappa}\Big(\lambda_g \, d_{\text{graph}}(\boldsymbol{\kappa}_i, \boldsymbol{\kappa}_j), \; \lambda_p \, d_{\text{phon}}(\boldsymbol{\kappa}_i, \boldsymbol{\kappa}_j)\Big), \quad \text{smin}_\tau(a,b) = -\tau \log \tfrac{e^{-a/\tau} + e^{-b/\tau}}{2}, \tag{16}$$

which recovers $\min(a,b)$ as $\tau \to 0^+$. The grapheme channel dominates when the two scripts match, and the phonological channel takes over automatically when the scripts differ, as happens with a Hindi and Urdu pair and with a romanized rendering of either language.

Source-Target Cost: For cross-lingual pairs, the cost routes morphology through the language-agnostic role channel,

$$c_{\boldsymbol{\theta}}^{\ell_x, \ell_h}(u_i^x, u_j^h) = c_{\boldsymbol{\theta}}^{\text{lex}}(u_i^x, u_j^h) + \lambda_\psi \; \text{JS}\big(\boldsymbol{\psi}_i^x, \boldsymbol{\psi}_j^h\big) + \lambda_{\text{real}} \, d_{\text{real}}\big(\boldsymbol{\psi}_i^x, \mathbf{z}_j^h\big) + c_{ij}^{\text{hard}}. \tag{17}$$

The JS term penalizes role-posterior divergence, as when a patient is realized as an agent, and we tie $\lambda_\psi = \lambda_{\text{real}}$ in all experiments. The learned bilinear compatibility $d_{\text{real}}(\boldsymbol{\psi}_i^x, \mathbf{z}_j^h) = 1 - \sigma\big(s_{\text{real}} \, \boldsymbol{\psi}_i^{x\top} W_{\text{real}} \mathbf{z}_j^h\big) \in (0,1)$, with a learned scale $s_{\text{real}} > 0$, scores whether the predicted target morphosyntactic class is a plausible realization of the source role.

Inner Affix Ground Cost: Index the atoms of A_i by $p \in \{0, 1, \ldots, P_i\}$, with $p = 0$ the null atom and $p \geq 1$ the P_i overt candidates in enumeration order. The ground cost $D^{ij} \in \mathbb{R}^{(P_i+1)\times(P_j+1)}$ is

$$D_{pq}^{ij} = \begin{cases} 0, & p = q = 0, \\ c_\varnothing, & \text{exactly one of } p, q \text{ is } 0, \\ \lambda_g^{\text{aff}} \, d_{\text{graph}}(m_{ip}, m_{jq}) + \lambda_o^{\text{aff}} \Big| \frac{p}{P_i+1} - \frac{q}{P_j+1} \Big|, & p, q \geq 1, \end{cases} \tag{18}$$

where m_{ip} is the surface string of the candidate and d_{graph} is the grapheme-level distance used in Equation 16. The last term penalizes drift in normalized candidate position, and the constant $c_\varnothing > 0$ is the price of moving mass between the null atom and an overt affix, which corresponds to a dropped or a spurious marker. The inner plan $T^{\star,ij}$ optimizes the objective of Equation 4 with $(\varepsilon_{\text{in}}, \tau_{\text{in}}, \tau_{\text{in}})$ in place of $(\varepsilon, \tau_a, \tau_b)$. We run the transport only when both tokens carry at least one overt candidate, that is $P_i, P_j \geq 1$. Whenever either side consists of the null atom alone or of punctuation, the implementation sets $c_{ij}^{\text{aff}} = 0$ without solving the inner problem (§2.3), because under Equation 18 the transport would otherwise charge roughly $c_\varnothing$ per unit of overt mass on the inflected side. For that fully dropped marker the JS term and the lexical cost carry the penalty instead.

A.5 Residual Definitions

Positive ℓ_1 slacks measure the unmatched mass on either side of the outer plan,

$$U_a(T^\star{}_{ab}) = \big\|[\boldsymbol{\alpha}^a - T^\star{}_{ab}\mathbf{1}]_+\big\|_1, \qquad U_b(T^\star{}_{ab}) = \big\|[\boldsymbol{\alpha}^b - T^{\star\top}_{ab}\mathbf{1}]_+\big\|_1, \tag{19}$$

and these slacks enter Equation 6 as the omission and the hallucination channels. The hinge is piecewise linear and differentiable almost everywhere, and every reported run applies it directly. A smooth surrogate $[t]_+^\varsigma = \varsigma \log(1+e^{t/\varsigma})$, which recovers $[t]_+$ as $\varsigma \to 0^+$, remains available whenever an application requires fully smooth residuals. The morphology and role residuals re-weight the plans with cost-specific kernels,

$$R_{\boldsymbol{\theta}}^{\text{morph}}(h,r) = \frac{\sum_{ij} T^\star{}_{hr,ij}\, G_{ij}^{hr}\, c_{\boldsymbol{\theta}}^{\text{morph}}(u_i^h, u_j^r)}{\sum_{ij} T^\star{}_{hr,ij}\, G_{ij}^{hr} + \varepsilon_0}, \tag{20}$$

$$R_{\boldsymbol{\theta}}^{\text{role}}(x,h) = \frac{\sum_{ij} T^\star{}_{xh,ij}\, d_{\text{real}}(\boldsymbol{\psi}_i^x, \mathbf{z}_j^h)}{\sum_{ij} T^\star{}_{xh,ij} + \varepsilon_0}.$$

The quantity $R_{\boldsymbol{\theta}}^{\text{morph}}$ is the gate-weighted average inner morphology cost along the optimal alignment from hypothesis to reference, and it answers how morphologically wrong the matched units are once alignment has taken place. The quantity $R_{\boldsymbol{\theta}}^{\text{role}}$ is the analogous role-realization average along the alignment from source to hypothesis. Both quantities are plan-weighted averages of bounded costs, which keeps them bounded and makes them interpretable diagnostics alongside the scalar score.

A.6 Reference-Free Score

For quality estimation, the reference-free score draws on the source-side components alone,

$$\begin{aligned} s_{\boldsymbol{\theta}}^{\text{QE}}(x,h) = \gamma_0^\ell - \gamma_1^\ell M_{xh} - \gamma_2^\ell O_{x\to h} - \gamma_3^\ell H_{h\setminus x} \\ - \gamma_4^\ell R_{\boldsymbol{\theta}}^{\text{role}}(x,h) + \gamma_5^\ell F_{\boldsymbol{\theta}}(h), \qquad \gamma_k^\ell = \text{softplus}(g_k^\ell)\ (k=1,\dots,5),\ \gamma_0^\ell \in \mathbb{R}, \end{aligned} \tag{21}$$

which inherits under $\boldsymbol{\gamma}$ the same monotonicity guarantee that Proposition 1 establishes under $\boldsymbol{\beta}$. We report no quality-estimation experiments in this paper, since Equation 21 appears here for completeness and we leave its empirical evaluation to future work.

A.7 Preference Likelihood

Given a preference dataset $\mathcal{D}_{\text{pref}}$ of tuples (x, r, h^+, h^-) drawn from human quality ratings $\mathcal{D}_{\text{human}}$, from controlled morphology perturbations $\mathcal{D}_{\text{morph}}$, and optionally from synthetic pairs $\mathcal{D}_{\text{synth}}$ generated by a large language model, which the reported runs exclude, the BTL model and its negative log-likelihood are

$$P_{\boldsymbol{\theta}}\big(h^+ \succ h^- \mid x, r\big) = \sigma\Big(\tfrac{s_{\boldsymbol{\theta}}(x,h^+,r) - s_{\boldsymbol{\theta}}(x,h^-,r)}{\tau}\Big), \tag{22}$$

$$\mathcal{L}_{\text{pref}}(\boldsymbol{\theta}) = -\sum_{(x,r,h^+,h^-)\in\mathcal{D}_{\text{pref}}} \log P_{\boldsymbol{\theta}}(h^+ \succ h^- \mid x, r),$$

where $\tau > 0$ is a temperature that controls the steepness of the preference logistic. Conditioning on a fixed plan and on fixed components $\boldsymbol{\Phi}$ makes the score affine in $\boldsymbol{\beta}$, and $\mathcal{L}_{\text{pref}}$ is therefore convex in $\boldsymbol{\beta}$, though not in the pre-softplus parameters b_k^ℓ. Appendix B.4 discusses the calibration of the temperature.

B Proofs and Supplementary Propositions

This chapter provides complete proofs for the theoretical claims in the main paper, together with supporting results on the well-posedness of the inner transport and on gradient propagation through the optimal plan. Each subsection stands on its own, and the notation is that of §2 and Appendix A.

B.1 Monotone Scoring Guarantee

Setup and Statement: Recall the MORPHOT scoring head from §2.6,

$$s_{\boldsymbol{\theta}}(x,h,r) \;=\; \beta_0^{\ell} \;+\; \underbrace{\beta_F^{\ell}\, F_{\boldsymbol{\theta}}(h)}_{\text{fluency reward}} \;-\; \sum_{k=1}^{8} \underbrace{\beta_k^{\ell}\, \Phi_k(x,h,r)}_{\text{penalty terms}}, \tag{23}$$

where $\beta_0^{\ell} \in \mathbb{R}$ is a language-specific bias, $\beta_k^{\ell} = \mathrm{softplus}(b_k^{\ell}) \geq 0$ for all $k \in \{1,\dots,8\}$, and $\beta_F^{\ell} \geq 0$. The implementation obtains the latter two by applying an element-wise $\mathrm{softplus}$ to unconstrained parameters b_k^{ℓ}, $b_F^{\ell} \in \mathbb{R}$ (Equation 9). The penalty vector $\boldsymbol{\Phi}(x,h,r) = (\Phi_1,\dots,\Phi_8)$ is the eight-component vector of Equation 8, and all eight entries are non-negative by construction, since in particular $d_{\mathrm{real}} \in (0,1)$ makes the role residual a non-negative weighted average. The argument below relies on the sign of the coefficients alone.

Proposition 1 (restated). Let h^+ dominate h^- in the sense that $F_{\boldsymbol{\theta}}(h^+) \geq F_{\boldsymbol{\theta}}(h^-)$ and $\Phi_k(x,h^+,r) \leq \Phi_k(x,h^-,r)$ for all k. Then $s_{\boldsymbol{\theta}}(x,h^+,r) \geq s_{\boldsymbol{\theta}}(x,h^-,r)$.

Proof: We verify the partial-derivative conditions directly. For any configuration (x,h,r) at which the quantities are defined,

$$\frac{\partial s}{\partial F} = \beta_F^{\ell} \;\geq\; 0, \tag{24}$$

$$\frac{\partial s}{\partial \Phi_k} = -\beta_k^{\ell} \;\leq\; 0, \quad k = 1,\dots,8. \tag{25}$$

Equation 24 establishes that s is non-decreasing in the fluency reward F, and Equation 25 establishes that s is non-increasing in each penalty Φ_k, meaning that a higher penalty implies a lower score. Both statements follow immediately from $\beta \geq 0$.

Moreover, s is a finite affine combination of F and $-\boldsymbol{\Phi}$ and is therefore jointly continuous on $\mathbb{R}\times\mathbb{R}^8$. Applying the coordinate-wise monotonicity to the domination hypothesis, under which the bias β_0^{ℓ} cancels, yields

$$\begin{aligned} s(x,h^+,r) &= \beta_0^{\ell} + \beta_F^{\ell} F(h^+) - \textstyle\sum_k \beta_k^{\ell} \Phi_k(x,h^+,r) \\ &\geq \beta_0^{\ell} + \beta_F^{\ell} F(h^-) - \textstyle\sum_k \beta_k^{\ell} \Phi_k(x,h^-,r) = s(x,h^-,r). \end{aligned}$$

The inequality holds term by term, because each summand changes weakly in the correct direction. □

Corollary and Practical Implications. The guarantee tightens to a strict inequality as soon as one component separates the two hypotheses.

Corollary 1 (Strict monotonicity). *Suppose at least one inequality of Proposition 1 is strict, meaning either $F(h^+) > F(h^-)$ or $\Phi_k(x,h^+,r) < \Phi_k(x,h^-,r)$ for some k, and suppose the corresponding coefficient β is positive. Then $s(x,h^+,r) > s(x,h^-,r)$.*

The positivity condition holds automatically for every coefficient of Equation 9, because $\mathrm{softplus}(b) = \log(1+e^b) > 0$ for every finite b. Consequently $\beta_k^{\ell} > 0$ for all k and $\beta_F^{\ell} > 0$, and any strict decrease in a single Φ_k, or a strict increase in F when that term is active, already yields a strict increase in the score. The explicit positivity condition in the corollary matters only for a head whose coefficients are not softplus-parameterized.

The strict version is in turn what the Bradley–Terry–Luce training objective optimizes toward. As long as the transport geometry separates h^+ from h^- on at least one axis, the score gap stays strictly positive and the preference probability $\sigma\big((s^+ - s^-)/\tau\big)$ exceeds $\frac{1}{2}$. The BTL gradient $-(1-\sigma(\cdot))/\tau$ remains non-zero for every finite gap, and training therefore continues to widen correctly ordered gaps instead of stalling as soon as the ordering happens to be correct.

Scope: Proposition 1 does not require the Φ_k to coincide with the OT quantities, since any non-negative penalty representation preserves the guarantee. In practice, the guarantee earns its value only when the domination condition is realistic. For that reason, the perturbation families used in evaluation explicitly target cases in which h^- strictly increases at least one penalty, since an omission inflates $O_{r\to h}$, a hallucination inflates $H_{h\setminus r}$, and a morphological error inflates both $R^{\mathrm{morph}}_{\boldsymbol{\theta}}$ and the matched discrepancy M_{hr}.

B.2 Inner UOT Existence, Uniqueness, and Differentiability

Setup: The inner transport problem solved for each token pair (i, j) is

$$\begin{aligned} T^{\star,ij} &= \operatorname*{arg\,min}_{T\in\mathbb{R}^{(P_i+1)\times(P_j+1)}_{\geq 0}} J^{ij}(T), \\ J^{ij}(T) &= \Big\langle D^{ij}, T\Big\rangle + \varepsilon_{\mathrm{in}}\, D_{\mathrm{KL}}(T \,\|\, \mathbf{1}\mathbf{1}^\top) + \tau_{\mathrm{in}}\, D_{\mathrm{KL}}(T\mathbf{1} \,\|\, \mathbf{a}_i) + \tau_{\mathrm{in}}\, D_{\mathrm{KL}}(T^\top\mathbf{1} \,\|\, \mathbf{a}_j), \end{aligned} \tag{26}$$

with optimal value $V^{\mathrm{aff}}_{ij} = J^{ij}(T^{\star,ij})$ and transported cost $c^{\mathrm{aff}}_{ij} = \langle T^{\star,ij}, D^{ij}\rangle$ as in Equation 3, and the score consumes the transported cost rather than the optimal value. Here $\mathbf{a}_i \in \Delta^{P_i}$ and $\mathbf{a}_j \in \Delta^{P_j}$ are the affix weights of the distributions A_i and A_j in Equation 14, with index 0 denoting the null-affix atom, D^{ij} is the affix-level cost matrix Equation 18, and $\varepsilon_{\mathrm{in}}, \tau_{\mathrm{in}} > 0$. The full morphology cost adds a token-level term, $c^{\mathrm{morph}}_{ij} = \lambda_z\ \mathrm{JS}(\mathbf{z}_i, \mathbf{z}_j) + c^{\mathrm{aff}}_{ij}$, which is smooth in both arguments and therefore requires no further treatment in the analysis below.

> **Theorem 1** (Well-posedness and smoothness of the inner transport)**.** *Fix $\varepsilon_{\mathrm{in}}, \tau_{\mathrm{in}} > 0$, let D^{ij} have finite entries, and let every affix weight a_{ip}, a_{jq} be strictly positive. Then the objective J^{ij} of Equation 26 admits a unique minimizer $T^{\star,ij}$, every entry of that minimizer is strictly positive, the value function $D^{ij} \mapsto V^{\mathrm{aff}}_{ij}$ is differentiable with $\nabla_{D^{ij}} V^{\mathrm{aff}}_{ij} = T^{\star,ij}$ and $\partial V^{\mathrm{aff}}_{ij}/\partial a_{ip} = \tau_{\mathrm{in}}\big(1 - (T^{\star,ij}\mathbf{1})_p/a_{ip}\big)$, and the map $D^{ij} \mapsto T^{\star,ij}$ is smooth.*

Existence and Uniqueness. Write $J(T)$ for the objective J^{ij} in Equation 26, defined on the closed convex cone $\mathbb{R}^{(P_i+1)\times(P_j+1)}_{\geq 0}$ with the convention $0\log 0 = 0$. The feasible set is not compact, and existence therefore follows from coercivity rather than from compactness. To that end, each Kullback–Leibler term is non-negative, and $D_{\mathrm{KL}}(T \,\|\, \mathbf{1}\mathbf{1}^\top) = \sum_{pq} T_{pq}\log T_{pq} - \sum_{pq} T_{pq} + (P_i+1)(P_j+1)$ grows super-linearly in $\|T\|_1$, whereas the linear term is bounded below by $-\|D^{ij}\|_\infty \|T\|_1$. Hence $J(T) \to \infty$ as $\|T\|_1 \to \infty$, the sub-level set $\{T : J(T) \leq J(\mathbf{a}_i\mathbf{a}_j^\top)\}$ is closed and bounded, and the continuous function J attains its minimum on that set. Uniqueness follows in turn because $T \mapsto \varepsilon_{\mathrm{in}} D_{\mathrm{KL}}(T \,\|\, \mathbf{1}\mathbf{1}^\top)$ is strictly convex, since negative entropy is strictly convex on the non-negative orthant, while the marginal Kullback–Leibler terms are convex as compositions of a convex function with a linear map and the transport term is linear. The minimizer $T^{\star,ij}$ is therefore unique for every triple $(D^{ij}, \mathbf{a}_i, \mathbf{a}_j)$. The two hypotheses of Theorem 1 hold by construction. The costs of Equation 18 are bounded, and the weights of Equation 14 are expectations under a softmax posterior over a lattice that always contains the affix-free decomposition, so the null atom and every overt atom in the support receive strictly positive mass (Appendix A.2). Under those hypotheses the minimizer has strictly positive entries, since $\partial J/\partial T_{pq} = D^{ij}_{pq} + \varepsilon_{\mathrm{in}}\log T_{pq} + \tau_{\mathrm{in}}\log\big((T\mathbf{1})_p/a_{ip}\big) + \tau_{\mathrm{in}}\log\big((T^\top\mathbf{1})_q/a_{jq}\big) \to -\infty$ as $T_{pq} \to 0^+$. The same argument applies to the outer plan Equation 4, which is the reason for keeping c^{hard} finite.

Differentiability and the Envelope Theorem. For fixed T the objective is affine in D^{ij}, and the feasible set does not depend on D^{ij}. The minimizer is unique and, for D^{ij} in a neighborhood of any given cost, lies in a common compact sub-level set, because the coercivity bound above is

locally uniform in D^{ij}. Danskin's theorem therefore applies, and the value function $D^{ij} \mapsto V_{ij}^{\mathrm{aff}}$ is differentiable with

$$\nabla_{D^{ij}} V_{ij}^{\mathrm{aff}} = T^{\star,ij}.$$

The envelope argument applies equally to the marginals, on which the objective depends smoothly through the two marginal Kullback–Leibler terms, since the entropic term carries the fixed reference $\mathbf{1}\mathbf{1}^\top$. Because $\partial_{a_{ip}} D_{\mathrm{KL}}(T\mathbf{1} \,\|\, \mathbf{a}_i) = 1 - (T\mathbf{1})_p / a_{ip}$, we obtain $\partial V_{ij}^{\mathrm{aff}} / \partial a_{ip} = \tau_{\mathrm{in}}\big(1 - (T^{\star,ij}\mathbf{1})_p / a_{ip}\big)$ and the symmetric expression for $\mathbf{a}_j$, and we verified both against central finite differences on random instances. The transported cost c_{ij}^{aff}, by contrast, is no value function, and its gradient therefore carries the sensitivity of the plan in addition (§B.3). Smoothness of the minimizer itself follows from the implicit function theorem applied to the Sinkhorn fixed-point equations. Writing $\mathbf{f}, \mathbf{g}$ for the scaled log-domain dual potentials and $\kappa = \tau_{\mathrm{in}}/(\tau_{\mathrm{in}} + \varepsilon_{\mathrm{in}}) \in (0,1)$, the unbalanced Sinkhorn iteration (Chizat et al., 2018) is

$$f_p \leftarrow \kappa\Big[\log a_{ip} - \log \textstyle\sum_q \exp\big(-D_{pq}^{ij}/\varepsilon_{\mathrm{in}} + g_q\big)\Big],$$
$$g_q \leftarrow \kappa\Big[\log a_{jq} - \log \textstyle\sum_p \exp\big(-D_{pq}^{ij}/\varepsilon_{\mathrm{in}} + f_p\big)\Big], \qquad T_{pq}^{\star,ij} = \exp\big(-D_{pq}^{ij}/\varepsilon_{\mathrm{in}} + f_p + g_q\big).$$

Setting the derivative above to zero and writing $f_p = -\tau_{\mathrm{in}}\varepsilon_{\mathrm{in}}^{-1} \log\big((T\mathbf{1})_p / a_{ip}\big)$, $g_q = -\tau_{\mathrm{in}}\varepsilon_{\mathrm{in}}^{-1} \log\big((T^\top\mathbf{1})_q / a_{jq}\big)$ confirms that these updates are exactly the fixed-point equations of the unit-reference objective Equation 26. Under the product reference $\mathbf{a}_i \otimes \mathbf{a}_j$, by contrast, the marginal weights would multiply the plan directly and would drop out of the bracket. Because $\kappa < 1$, each half-step is a contraction of ratio at most κ in the supremum norm, since the log-sum-exp map is 1-Lipschitz in $\mathbf{g}$, and the composed map therefore has a unique fixed point to which the iteration converges geometrically. The map is C^∞ in $(D^{ij}, \mathbf{f}, \mathbf{g})$ and its Jacobian with respect to $(\mathbf{f}, \mathbf{g})$ has operator norm at most $\kappa < 1$ at the fixed point. Consequently $I - \partial(\mathrm{map})$ is invertible and the fixed point is a smooth function of D^{ij}, which completes the proof of Theorem 1 in full. □

In the balanced limit $\tau_{\mathrm{in}} \to \infty$, where $\kappa \to 1$, the argument degenerates to the usual Hilbert-metric contraction, and every configuration we report uses a finite τ_{in}. Whenever the ground cost carries parameters, the implementation back-propagates through the unrolled inner Sinkhorn iterations (Appendix B.3).

B.3 Gradient of the Matched Discrepancy

Setup: Let $V_{ab}(C)$ denote the optimal value of the outer problem Equation 4 as a function of the cost matrix $C = C^{ab}$, and let ${T^\star}_{ab}(C)$ denote its unique minimizer. Existence, uniqueness, and smoothness follow exactly as in Theorem 1, with $(\varepsilon, \tau_a, \tau_b)$ in place of $(\varepsilon_{\mathrm{in}}, \tau_{\mathrm{in}}, \tau_{\mathrm{in}})$. The matched discrepancy Equation 5 is not V_{ab} and is instead the mass-normalized transported cost

$$M_{\boldsymbol{\theta}}(a,b) = \frac{\langle {T^\star}_{ab}(C),\, C\rangle}{m(C) + \varepsilon_0}, \qquad m(C) = \|{T^\star}_{ab}(C)\|_1 = \mathbf{1}^\top {T^\star}_{ab}(C)\mathbf{1}.$$

Gradient Derivation: The matched discrepancy admits a closed-form gradient in the cost matrix.

> **Proposition 2** (Gradient of the matched discrepancy)**.** *Under the hypotheses of Theorem 1 applied to the outer problem, the matched discrepancy $M_{\boldsymbol{\theta}}(a,b)$ of Equation 5 is differentiable in the cost matrix C, and its gradient decomposes into an envelope term and a plan-sensitivity correction as stated in Equation 27 below.*

Two facts drive the derivation. First, Danskin's theorem applied to V_{ab} yields $\nabla_C V_{ab} = {T^\star}_{ab}(C)$, because the objective is affine in C for fixed T and the minimizer is unique, so the value function needs no plan Jacobian. Second, $M_{\boldsymbol{\theta}}$ depends on C both explicitly and through the plan, and the envelope identity therefore does not apply to it directly. The chain rule instead yields

$$\nabla_C M_{\boldsymbol{\theta}} = \frac{{T^\star}_{ab}}{m + \varepsilon_0} + \Big(\frac{\partial {T^\star}_{ab}}{\partial C}\Big)^\top \Big[\frac{C}{m + \varepsilon_0} - \frac{\langle {T^\star}_{ab},\, C\rangle}{(m+\varepsilon_0)^2}\, \mathbf{1}\mathbf{1}^\top\Big], \tag{27}$$

where the bracket is $\partial M_{\boldsymbol{\theta}} / \partial T$ evaluated at $T = {T^\star}_{ab}$, and $\partial {T^\star}_{ab} / \partial C$ is the Jacobian of the plan, which exists and stays continuous by the implicit-function argument behind Theorem 1. The first

term is the explicit dependence on C at a frozen plan, and the second is the plan-sensitivity correction that distinguishes $M_{\boldsymbol{\theta}}$ from the raw UOT value, which completes the proof of Proposition 2 stated above. □

In practice, automatic differentiation through the unrolled Sinkhorn iterations, comprising 40 outer and 20 inner steps in the log domain, produces both terms together, and we verified the result against finite differences to within 10^{-5} on held-out token pairs.

The cost matrix depends on the trainable parameters through the local costs of §2.2. For a same-language pair we have $C_{ij} = c_{ij}^{\mathrm{lex}} + \zeta_\ell\, G_{ij}\, c_{ij}^{\mathrm{morph}} + c_{ij}^{\mathrm{hard}}$ with G_{ij} a function of c_{ij}^{lex} alone, and hence

$$\nabla_{\boldsymbol{\theta}} C_{ij} = \big(1 + \zeta_\ell\, G'_{ij}\, c_{ij}^{\mathrm{morph}}\big)\nabla_{\boldsymbol{\theta}} c_{ij}^{\mathrm{lex}} + \zeta_\ell\, G_{ij}\Big[\lambda_z \nabla_{\boldsymbol{\theta}}\, \mathrm{JS}(\mathbf{z}_i, \mathbf{z}_j) + \nabla_{\boldsymbol{\theta}} c_{ij}^{\mathrm{aff}}\Big],$$

where $G'_{ij} = \partial G_{ij}/\partial c_{ij}^{\mathrm{lex}}$, and the hard-constraint term drops out because that term carries no dependence on $\boldsymbol{\theta}$. The transported affix cost $c_{ij}^{\mathrm{aff}} = \langle T^{\star,ij}, D^{ij}\rangle$ depends on $\boldsymbol{\theta}$ explicitly through the ground cost D^{ij} and, through the plan, both through D^{ij} and through the affix weights $\mathbf{a}_i, \mathbf{a}_j$ that the segmentation posterior of Equation 14 produces. Exactly as with Equation 27, the transported affix cost is no value function, the envelope identity therefore does not apply to it directly, and the chain rule yields

$$\nabla_{\boldsymbol{\theta}} c_{ij}^{\mathrm{aff}} = \big\langle T^{\star,ij}, \nabla_{\boldsymbol{\theta}} D^{ij}\big\rangle + \Big\langle D^{ij},\ \tfrac{\partial T^{\star,ij}}{\partial D^{ij}}\big[\nabla_{\boldsymbol{\theta}} D^{ij}\big] + \tfrac{\partial T^{\star,ij}}{\partial \mathbf{a}_i}\big[\nabla_{\boldsymbol{\theta}} \mathbf{a}_i\big] + \tfrac{\partial T^{\star,ij}}{\partial \mathbf{a}_j}\big[\nabla_{\boldsymbol{\theta}} \mathbf{a}_j\big]\Big\rangle, \quad (28)$$

where the first term is the envelope contribution, namely the gradient of V_{ij}^{aff} with respect to D^{ij} granted by Theorem 1, and the plan Jacobians in the second term, applied here as directional derivatives, exist and stay continuous by the same implicit-function argument. Automatic differentiation through the unrolled inner Sinkhorn iterations produces both terms. In the released configuration, neither D^{ij}, which comprises the grapheme distance, the position term, and the null penalty, nor the affix weights contain trainable parameters, and both terms of Equation 28 therefore vanish. Under that configuration c_{ij}^{aff} acts as a fixed feature, and the only trainable dependence of c_{ij}^{morph} runs through $\lambda_z\ \mathrm{JS}(\mathbf{z}_i, \mathbf{z}_j)$ by way of W_z, while Equation 28 covers the general case of a parameterized ground cost. Composing Equation 28 with Equation 27 then yields the full gradient of $M_{\boldsymbol{\theta}}$ with respect to $\boldsymbol{\theta}$. No marginal term appears, because the masses of Equation 15 are computed under a stop-gradient on the salience score and $\boldsymbol{\eta}$ is not optimized, so the marginals $\boldsymbol{\alpha}^a, \boldsymbol{\alpha}^b$ propagate no gradient to $\boldsymbol{\theta}$ (Appendix A.3). Every factor stays continuous in $\boldsymbol{\theta}$ as long as $\varepsilon, \varepsilon_{\mathrm{in}} > 0$ and the marginal strengths stay finite, which leaves the score free of sub-gradient discontinuities at the transport layer.

B.4 Calibration Consistency of the Preference Scale

Setup: Every tuple of $\mathcal{D}_{\mathrm{pref}}$ lists the preferred hypothesis first, so the likelihood Equation 22 carries no explicit label. For the analysis we make the sampling of that ordering explicit and write $y \in \{0, 1\}$ for the label of an unordered pair, where $y = 1$ means that the first hypothesis is preferred, so that Equation 22 is the $y = 1$ case of the model below. Assume that y follows a Bradley–Terry–Luce model (Bradley & Terry, 1952; Luce, 1959),

$$P(y = 1 \mid x, r, h^+, h^-) = \sigma\big(\Delta^*/\tau^*\big), \qquad \Delta^* = s^*(x, h^+, r) - s^*(x, h^-, r),$$

where $\tau^* > 0$ is the population temperature and s^* the latent score. MORPHOT minimizes the negative log-likelihood of Equation 22, which amounts to maximizing

$$\mathcal{L}_{\mathrm{BTL}}(\boldsymbol{\theta}, \tau) = \mathbb{E}_{\mathcal{D}_{\mathrm{pref}}}\Big[y \log \sigma\big(\Delta_{\boldsymbol{\theta}}/\tau\big) + (1 - y)\log\big(1 - \sigma(\Delta_{\boldsymbol{\theta}}/\tau)\big)\Big], \qquad \Delta_{\boldsymbol{\theta}} = s_{\boldsymbol{\theta}}^+ - s_{\boldsymbol{\theta}}^-.$$

Identifiability and the Fixed Temperature: The pair $(\boldsymbol{\beta}, \tau)$ is not identifiable. Because $s_{\boldsymbol{\theta}}$ is linear in the head coefficients $\boldsymbol{\beta}$ by Equation 9, the settings $(c\boldsymbol{\beta}, c\tau)$ and $(\boldsymbol{\beta}, \tau)$ induce the same preference distribution for every $c > 0$. The regularizers in Equation 10 are not invariant under that rescaling, and we therefore specify λ_{eq} and λ_{wd} for $\tau = 1$ and fix $\tau = 1$ in all experiments. The coefficients $\boldsymbol{\beta}$ then carry the scale of the score gap, and the argument below consequently applies to the scale of $\Delta_{\boldsymbol{\theta}}$ rather than to τ in isolation.

Consistency Argument: Under correct specification, the learned scale of the score gap is consistent for the population temperature.

Proposition 3 (Calibration consistency)**.** *Suppose the preference labels follow the population model above and the score is correctly specified up to scale, such that $\Delta_{\boldsymbol{\theta}} = \Delta^*/c$ for some $c > 0$ that $\boldsymbol{\beta}$ can adjust. Write $t = 1/(c\tau)$ for the effective inverse temperature. Then the population objective $\mathcal{L}_{\mathrm{BTL}}$ is strictly concave in t and attains its unique maximum at $t = 1/\tau^*$, and the estimator satisfies $\hat{t} \to 1/\tau^*$ as $|\mathcal{D}_{\mathrm{pref}}| \to \infty$.*

Using $\frac{d}{dz}\log\sigma(z) = 1 - \sigma(z)$ and $\frac{d}{dz}\log(1-\sigma(z)) = -\sigma(z)$, the score equation in t is

$$\frac{\partial \mathcal{L}_{\mathrm{BTL}}}{\partial t} = \mathbb{E}\big[(y - \sigma(t\Delta^*))\,\Delta^*\big] = \mathbb{E}\big[(\sigma(\Delta^*/\tau^*) - \sigma(t\Delta^*))\,\Delta^*\big] = 0,$$

where the second equality conditions on Δ^* and applies $\mathbb{E}[y \mid \Delta^*] = \sigma(\Delta^*/\tau^*)$. Since σ is strictly increasing, the integrand carries the sign of $(1/\tau^* - t)$ for every $\Delta^* \neq 0$, and the unique root is therefore $t = 1/\tau^*$. The population objective is strictly concave in t, since its second derivative equals $-\mathbb{E}[\sigma'(t\Delta^*)\Delta^{*2}] < 0$ whenever Δ^* differs from zero with positive probability. Standard M-estimation arguments then yield $\hat{t} \to 1/\tau^*$ as $|\mathcal{D}_{\mathrm{pref}}| \to \infty$, which establishes Proposition 3 and confirms that the learned scale of the score gap converges to the population temperature. □

Proposition 3 concerns the calibration of the pairwise preference probabilities $\sigma(\Delta_{\boldsymbol{\theta}})$ alone. Calibration of the raw score against the DA scale would additionally require an affine map from $s_{\boldsymbol{\theta}}$ to that scale, which this paper leaves unmeasured.

Practical Note. Under misspecification, as when preferences follow a Thurstone model or a mixture of annotators, the quasi-maximum-likelihood estimator instead converges to the scale that minimizes the Kullback–Leibler divergence from the true conditional law of y given $\Delta_{\boldsymbol{\theta}}$ to the BTL family (White, 1982). The fitted scale then remains the best logistic calibration of the score gap, and we claim no convergence rate and no bound on the residual miscalibration under that regime.

C Dataset Construction and Statistics

ComTaIL: ComTaIL is the COMTAIL dataset (Ahsan et al., 2025), which holds 219,439 DA+SQM ratings on $[0, 1]$ over 13 Indian languages and 21 translation directions, split into 199,737 training, 9,815 development, and 9,887 test segments. We evaluate on the test split, train on the other two, and use the scores as released, without re-annotation (Table 6).

Table 6: ComTaIL per-language statistics on the test split ($N = 9{,}887$ segments). DA+SQM scores lie on $[0, 1]$.

Language	Segs	Mean DA	Std DA	#Sys	Language	Segs	Mean DA	Std DA	#Sys
Bengali (ben)	755	0.704	0.285	15	Punjabi (pan)	862	0.710	0.279	14
Gujarati (guj)	780	0.717	0.244	12	Tamil (tam)	538	0.683	0.290	13
Hindi (hin)	571	0.691	0.265	12	Telugu (tel)	1,018	0.698	0.278	13
Kannada (kan)	1,084	0.712	0.234	14	Urdu (urd)	1,135	0.722	0.239	14
Kashmiri (kas)	491	0.704	0.328	10	Dogri (doi)	451	0.688	0.293	11
Marathi (mar)	1,195	0.695	0.310	14	Sindhi (snd)	448	0.686	0.297	11
Odia (ory)	559	0.703	0.254	14					
					Total / avg	**9,887**	0.701	0.277	

MQM-MRL: MQM-MRL is the Hindi and Tamil portion of IndicMT Eval (Sai B et al., 2023), whose annotators mark error spans with a category and a severity under the MQM framework (Lommel et al., 2013). We use the annotations exactly as released. The error-typing results of Table 3 come from a multinomial logistic regression that predicts the dominant span category of a segment from each metric's features, namely the eight components of $\boldsymbol{\Phi}$ Equation 8 for MORPHOT and the scalar score for each baseline, fit and evaluated on the same segments.

Preference Pairs: We assemble $\mathcal{D}_{\mathrm{pref}}$ of Equation 22 from the ComTaIL train split alone, so the test split stays unseen. The source $\mathcal{D}_{\mathrm{human}}$ groups system outputs by (x, r, ℓ_h) and, for every group with at least two scored systems, pairs the highest-scoring hypothesis as h^+ with the lowest-scoring one as h^-, which keeps languages balanced. The source $\mathcal{D}_{\mathrm{morph}}$ pairs each reference-quality hypothesis with a programmatic corruption of itself, so the two differ only in the perturbed span. An optional source $\mathcal{D}_{\mathrm{synth}}$ generated by a large language model is disabled in every reported run.

Perturbation Families: Six families apply to every target, namely word-order permutation, content-span omission, unsupported-span insertion, numeral and named-entity substitution, grapheme-cluster transposition, and Unicode decomposition. The last is an invariance probe. MORPHOT normalizes inputs to the composed form, so a decomposition pair ties with a zero score gap and contributes no gradient (Appendix I.4), whereas surface metrics see a genuine contrast. Five languages add the morphology-specific families of Table 7. The other eight targets receive the base families alone, which is the reliance on language-specific priors noted in §6.

Table 7: Language-specific perturbation families in $\mathcal{D}_{\mathrm{morph}}$. All thirteen targets additionally receive the six base families described in the text.

Language	Families	Description
Hindi	case, agreement, tense-aspect-mood	Ergative or accusative marker substitution, person agreement swap, tense-aspect-mood substitution
Urdu	case	Case-marker substitution
Bengali	case	Case-marker substitution
Tamil	suffix, truncation	Suffix swap within a productive set, grapheme-level suffix truncation
Telugu	suffix, truncation	Suffix swap within a productive set, grapheme-level suffix truncation

Equivalence Transformations: The regularizer $\mathcal{L}_{\mathrm{eq}}$ of Equation 10, with $\lambda_{\mathrm{eq}} = 0.1$, acts on pairs $(h, g(h))$ built from Hindi object–verb re-ordering, Urdu–Devanagari transliteration, and zero-width-joiner normalization, applied only where a transformation exists for the target language.

D Training Details and Reproducibility

Training Setup: The main experiments use the default configuration of Table 8, and a base variant shares every transport setting with a smaller encoder and $K = K' = 32$. Both train on a single NVIDIA A100-80GB GPU. Training the $\approx$0.44 M trainable parameters on the preference pairs of the ComTaIL train split (Appendix C) takes under two hours, and the ablation grid of 16 variants, one run each, took about 90 GPU-hours. We use AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$, a constant learning rate of 10^{-4} on $\boldsymbol{\omega}$ and $\boldsymbol{\beta}$ without warmup or annealing, batch size 32, and up to 10 epochs with early stopping on development preference accuracy at patience 3. The optimizer applies no decoupled weight decay, so regularization enters only through the explicit term $\lambda_{\mathrm{wd}} \|(\boldsymbol{\omega}, \boldsymbol{\beta})\|_2^2$ of Equation 10 with $\lambda_{\mathrm{wd}} = 10^{-4}$. Both Sinkhorn solvers run in the log domain with FP32 accumulators, which avoids underflow on long sentences. They stop once the supremum-norm change of the dual potentials falls below 10^{-5}, checked every 10 sweeps, or at the caps of 20 inner and 40 outer sweeps (Table 8). The contraction factor per sweep is κ^2 with $\kappa = \tau/(\tau + \varepsilon)$ (Theorem 1), about 0.85 for the outer and 0.89 for the inner problem at the default hyperparameters.

Statistics and Released Artifacts: Every correlation in the main tables is the mean over three seeds, $\{0, 1, 2\}$, whereas the ablation grid uses a single run per variant, and confidence intervals come from a paired segment-level bootstrap with 1,000 resamples, applied identically to every metric. The release contains the trained scoring-head checkpoints and the per-segment model outputs, and a single command rebuilds Tables 1, 2, 3, and 9 together with the significance tables from those outputs. Every table and figure of §4 and Appendices H and I is rebuilt from the same per-segment outputs. The MQM-MRL results of §3.3 rest on third-party annotation, which we redistribute only where the source license permits, and otherwise we point to the original source.

E HYPERPARAMETERS AND SENSITIVITY SWEEPS

Default Hyperparameter Values. Table 8 lists every hyperparameter used in the canonical MORPHOT configuration. The released default configuration mirrors those values, and every number reported in the main paper comes from the following configuration.

Table 8: Default MORPHOT hyperparameters.

Symbol	Value	Description
ε	0.07	Outer UOT entropy regularization
τ_a, τ_b	0.85	Outer UOT marginal relaxation
ε_{in}	0.05	Inner UOT entropy regularization
τ_{in}	0.80	Inner UOT marginal relaxation
I_{Sinkhorn}	40 / 20	Max. outer / inner Sinkhorn iterations
k	16	Top-k plan truncation
$\vartheta_{\ell,\ell}, T_{\ell,\ell}$	0.6, 0.1	Lexical gate threshold / temperature
$\lambda_e, \lambda_\rho, \lambda_\kappa$	1.0, 0.5, 1.0	Lexical-cost weights (embedding, root, script)
$\lambda_g, \lambda_p, \tau_\kappa$	0.2, 0.2, 0.1	Grapheme / phonology weights and soft-min temperature
λ_z, ζ_ℓ	1.0, 1.0	Token-level JS weight and morphology-cost weight
$\lambda_\psi = \lambda_{\text{real}}$	0.5	Role-channel weights (tied)
$\lambda_g^{\text{aff}}, \lambda_o^{\text{aff}}, c_\varnothing$	0.3, 0.1, 0.2	Affix ground-cost weights and null-affix penalty
c^{hard}	1.0 / 0.5	Hard penalty for numeral / entity mismatch
$\lambda_{\text{eq}}, \lambda_{\text{wd}}$	0.1, 10^{-4}	Equivalence and weight-decay coefficients
ε_0	10^{-9}	Denominator guard in Equation 5 and Equation 20
m_ℓ, ν	0.0, 0.1	Source-anchor tolerance / smoothing width
τ	1.00	BTL temperature, fixed (App. B.4)
$\dim \mathbf{\Phi}$	8	Number of penalty components
K, K'	48, 48	Latent morphosyntactic / role clusters (32, 32 in the base configuration)
Encoder	XLM-R-large	Frozen at training
Affix segmenter	IndicNLP	Language-specific

Transport Regularization Sensitivity: As reported in §3.3, sweeping $\varepsilon \in [0.01, 0.5]$ and $\tau_a = \tau_b \in [0.05, 1.0]$ on a held-out 1k-segment slice changes ComTaIL Pearson by less than 0.008, which confirms that performance stays insensitive to the exact regularization values within the stable operating region surrounding the defaults listed in Table 8.

Top-k Truncation and Encoder Sensitivity: Top-k truncation at $k = 16$ stays within 0.002 Pearson of the dense solver while running $3.7\times$ faster on the longest decile of sentence pairs (Appendix J). Unfreezing the encoder would address the residual class that is specific to MORPHOT, namely encoder-level confusions on transliterated technical terms and function-word contrasts (Appendix I), at the price of the $\approx$560M parameters of the encoder, which would forfeit the parameter-efficiency claim that motivates the design. We therefore leave the encoder frozen everywhere in the paper.

F PER-LANGUAGE CORRELATION TABLES

Table 9 breaks the ComTaIL Pearson correlation down by language, and its pooled row is the one in Table 1. MORPHOT beats every surface and embedding metric on all 13 languages, with gains over BERTScore from $+0.027$ on Tamil to $+0.231$ on Hindi. It leads BLEURT-20 on seven targets, from $+0.011$ on Marathi to $+0.262$ on Kashmiri, and trails it on the other six, by 0.006 on Telugu to 0.073 on Tamil, finishing second on each. The pooled row correlates over all $9{,}887$ segments, whereas the macro row averages the per-language coefficients. The two differ because the per-language score distributions are offset from one another, most visibly for TER at 0.260 pooled against 0.392 macro. MORPHOT leads on both aggregates, by $+0.092$ and $+0.086$ over BERTScore and by $+0.053$ and $+0.020$ over BLEURT-20. Spearman and Kendall follow the same pattern and are recorded in the released correlation tables.

Table 9: Per-language Pearson correlation on ComTaIL over 13 Indic targets, for the surface metrics, the learned metric (BLEURT-20), and MORPHOT. Best per language in bold, second best underlined.

Lang	BLEU	TER	CharacTER	chrF	chrF++	METEOR	BERTScore	BLEURT	MORPHOT
ben	0.244	0.336	0.316	0.278	0.264	0.234	0.419	**0.621**	0.551
guj	0.305	0.410	0.404	0.335	0.330	0.332	0.488	**0.578**	0.551
hin	0.262	0.314	0.292	0.214	0.218	0.203	0.471	0.654	**0.702**
kan	0.328	0.492	0.455	0.411	0.403	0.405	0.560	**0.652**	0.644
kas	0.398	0.499	0.641	0.581	0.556	0.497	0.635	0.449	**0.711**
mar	0.318	0.407	0.402	0.354	0.345	0.341	0.530	0.685	**0.696**
ory	0.252	0.386	0.395	0.367	0.353	0.328	0.391	0.414	**0.434**
pan	0.440	0.427	0.538	0.472	0.461	0.426	0.610	**0.720**	0.686
tam	0.358	0.285	0.506	0.536	0.515	0.429	0.629	**0.729**	0.656
tel	0.409	0.240	0.320	0.502	0.491	0.460	0.619	**0.677**	0.671
urd	0.354	0.412	0.443	0.392	0.384	0.360	0.508	0.581	**0.602**
doi	0.606	0.659	0.651	0.669	0.668	0.671	0.698	0.671	**0.732**
snd	0.369	0.224	0.197	0.379	0.378	0.363	0.483	0.474	**0.527**
pooled	0.352	0.260	0.262	0.411	0.403	0.380	0.519	0.558	**0.611**
macro-avg	0.357	0.392	0.428	0.422	0.413	0.388	0.542	0.608	**0.628**

G Statistical Significance

We test MORPHOT against each baseline on the pooled test set with a paired bootstrap of $1{,}000$ segment-level resamples and with Williams' test for dependent correlations, and the released significance table records every comparison. The pooled Pearson advantage of MORPHOT over every baseline, BLEURT-20 included, is significant at $p < 0.001$ under both tests, with bootstrap intervals clear of zero. The margins below are bootstrap point estimates and can differ by one unit in the third decimal from differences of the rounded entries in Table 1. The margin is $+0.053$ with interval $[0.039, 0.065]$ over BLEURT-20 at $t = 9.7$, $+0.091$ with interval $[0.079, 0.103]$ over BERTScore at $t = 18.2$, $+0.199$ with interval $[0.183, 0.214]$ over chrF at $t = 31.9$, $+0.231$ with interval $[0.216, 0.246]$ over METEOR at $t = 36.9$, and $+0.258$ with interval $[0.244, 0.271]$ over BLEU at $t = 40.2$, and it grows larger still over TER and CharacTER. The test concerns pooled Pearson, and MORPHOT also leads pooled Spearman and Kendall (Table 1). Per language, MORPHOT holds the highest Pearson among the surface and embedding metrics on all 13 targets and beats BLEURT-20 on seven of them (Table 9). We report the pooled test rather than thirteen separate tests because the smallest language slices, which hold between 448 and 571 segments, leave a per-language bootstrap underpowered.

One comparison favours a variant, since MORPHOT-lex leads the full model by 0.019 Pearson at $t = -5.0$. The lexical comparison is the reason we rest the case for the morphological and unbalanced components on error attribution (§3.3) and on the residual channels (§3.2) rather than on correlation.

H Extended Analyses

Qualitative Segments: Table 4 omits the references, which read as follows in the same romanization. Segment 1, 161 jana durga debii rayeche yaaraa durga debataa athabaa iShTa debataa eban dashaasharu yini debii jayacaNDiira ekaTi pratimuurti. Segment 2, keTalaanka ullanghano khelaaDii athavaa Tiima para danDamaan pariName che. Segment 3, 21. aami yakhana aamaara sbaasthyasebaa pradaanakaariira paraamarsha niye prashna kari takhana dbitiiya mataamata paan.a Segment 4, sansarga jhaalyaanantara 2-3 divasaanta TaayaphiDoTa caacaNii sakaaraatmaka hote.

Figure 3 adds a fifth segment that the four failure modes do not cover, a faithful Punjabi paraphrase with low n-gram overlap, in native script and with the standardized cost of every component. The annotators score it 0.89 and MORPHOT places it at the 73rd percentile, whereas every other metric

sits at or below the 41st. Every component cost lies below the corpus mean, which leaves all eight channels inactive.

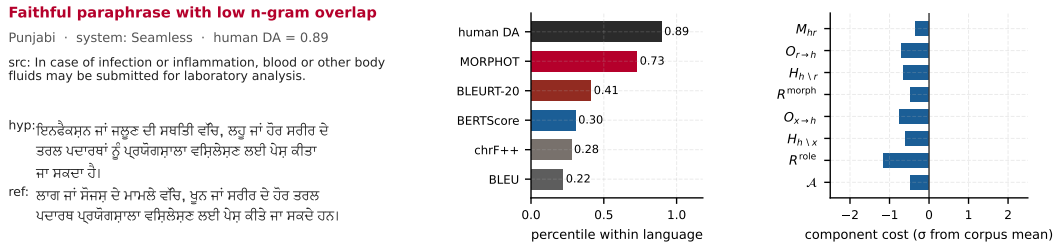


Figure 3: A faithful paraphrase with low n-gram overlap: text in native script (left), within-language percentiles (middle), and component costs in σ (right).

Segment-Level Agreement Beyond Correlation: Table 10 reports pairwise ranking accuracy by human-quality gap. On reliability curves over 20 percentile bins, MORPHOT is monotone on 95% of steps and spans DA 0.26 to 0.88, against 89% for BLEURT-20, 84% for BERTScore, and 68% for chrF++ and BLEU, the last two flattening above the 80th percentile where gold references saturate them. Figure 4 breaks the correlation down by source domain and source-length quartile.

Table 10: Pairwise ranking accuracy on ComTaIL (%). For random pairs of segments from the same target language whose human DA scores differ by more than the threshold, the table reports how often each metric orders the pair as the annotators did.

DA gap	Pairs	**MORPHOT**	MORPHOT-bal	BLEURT	BERTSCORE	CHRF++	METEOR	BLEU
> 0.05	65,916	73.0	**73.6**	73.0	69.8	64.1	64.9	63.6
> 0.10	56,434	75.9	**76.6**	75.8	72.3	66.3	66.7	65.8
> 0.20	41,187	81.0	**81.6**	80.6	76.9	70.1	70.0	69.6
> 0.30	30,229	85.4	**86.1**	84.6	81.0	73.8	73.5	73.5
> 0.50	16,805	90.7	**91.2**	89.2	87.0	79.2	78.6	79.0

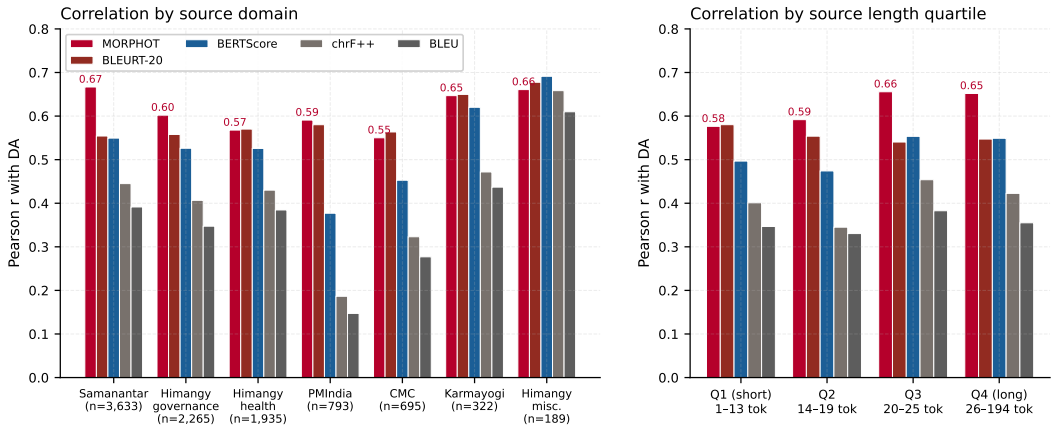


Figure 4: Correlation by source domain and by source-length quartile on ComTaIL.

System-Level Ranking and Conservation Pathology: Table 11 ranks the ten ComTaIL systems. BERTScore, chrF++, and BLEU place both perturbed-reference systems second and third, because a perturbed reference keeps most of its n-grams. BLEURT-20 keeps them in the bottom half but reaches only 0.731 Pearson, because it swaps the genuine MT systems among themselves. At the segment level, a random MT segment outranks a random perturbed-reference segment of the same language with probability 0.74 under human DA, 0.69 under MORPHOT, 0.54 under BLEURT-20, and 0.16 to 0.37 under the surface metrics and BERTScore, which invert the order.

Table 11: System-level ranking on ComTaIL. Systems with $\geq$ 300 segments, ordered by mean human DA, where * marks a rank that departs from the human rank by three or more places.

System	n	Mean DA	Human	**MORPHOT**	BLEURT	BERTSCORE	CHRF++	BLEU
Gold reference	1908	0.854	1	1	1	1	1	1
MS Bing	755	0.848	2	2	3	5*	5*	5*
IndicTrans2	763	0.829	3	4	6*	8*	6*	6*
versionvN	759	0.826	4	6	5	7*	7*	7*
Google	663	0.823	5	3	2*	4	4	4
Seamless	727	0.777	6	5	4	6	8	8
Perturbed reference	964	0.547	7	9	8	2*	3*	3*
GPT-3.5	1997	0.536	8	10	9	10	10	10
oldX	984	0.530	9	7	10	9	9	9
Multi-perturbed ref.	322	0.508	10	8	7*	3*	2*	2*
Pearson r (10 systems)				**0.906**	0.731	0.495	0.225	0.294
Kendall τ (10 systems)				**0.689**	0.600	0.156	0.244	0.244

ACES Phenomena Where MORPHOT Trails: Among phenomena with at least 50 pairs, the largest deficits against BERTScore concern connectives and modal verbs, untranslated spans, and lexical substitution. They are the temporal connective since (-40.6pp), modal-verb substitution and deletion (-28.0 and -22.0pp), untranslated spans against a synonym or the reference word (-14.3 and -10.3pp), copy-source (-10.6pp, Figure 5), hyponym replacement (-9.6pp), and real-data hallucination against a synonym (-9.0pp).

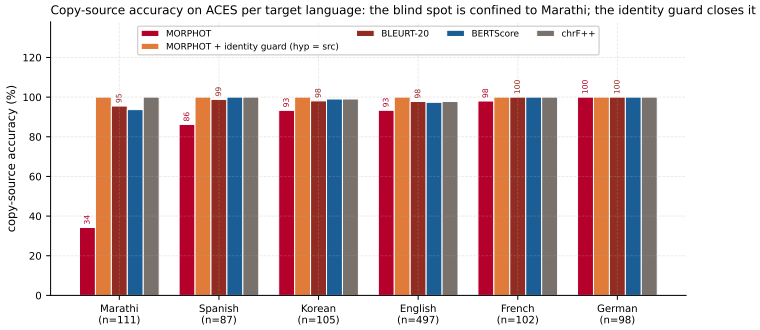


Figure 5: The copy-source blind spot is confined to Marathi and closed by an identity guard that scores a hypothesis equal to the source as $-\infty$ before MORPHOT runs.

ACES Phenomena Where MORPHOT Leads: Table 12 depicts the phenomena on which MORPHOT ranks first among the nine metrics. It leads BLEURT-20 on the families that the residual channels decide, namely punctuation, addition, and number and named-entity hallucination, and trails it where the encoder must know what a span means, namely unit and date hallucination, lexical substitution, world knowledge, overly literal translation, modal verbs, and anaphora. By target language, BLEURT-20 wins the large English, German, French, Korean, and Polish subsets, and MORPHOT wins Hindi, Catalan, Japanese, Chinese, and Russian, by 2 to 18pp. Figure 6 shows that the components carry human-aligned signal in almost every ComTaIL language, most strongly the morphology residual and the matched discrepancy. The exceptions are the source-side components on Dogri and Kashmiri, which turn slightly negative.

I ERROR ANALYSIS

This appendix extends §4. All residuals are computed as the within-language percentile of the metric minus the within-language percentile of the human DA on the ComTaIL test split.

I.1 RESIDUAL TAXONOMY

We take the 5% of segments whose within-language metric percentile departs most from the human percentile ($N = 495$), and Table 13 reports how much each mechanically computed property is over-

Table 12: ACES phenomena on which MORPHOT ranks first among the nine metrics scored (phenomena with ≥ 50 pairs, accuracy in %).

Phenomenon	Pairs	**MORPHOT**	BLEURT	BERTSCORE	CHRF++	BLEU	METEOR
ambiguous-translation-wrong-gender-female-anti	1,000	**57.0**	3.8	48.6	1.9	0.0	0.0
punctuation:deletion_all	1,000	**99.1**	93.1	95.9	96.3	84.7	54.7
hallucination-named-entity-level-1	999	**96.4**	95.5	96.3	94.5	70.8	64.6
hallucination-named-entity-level-3	999	**12.7**	12.4	3.1	0.9	0.0	0.5
hallucination-number-level-1	999	**99.2**	98.5	95.4	**99.2**	87.3	68.8
hallucination-number-level-2	999	**83.3**	65.3	46.8	17.2	17.4	23.8
hallucination-number-level-3	999	**55.9**	5.2	1.0	0.0	0.5	0.6
ambiguous-translation-wrong-gender-female-pro	806	**69.1**	12.7	65.9	0.0	0.0	0.0
ambiguous-translation-wrong-discourse-connective-while-contrast	50	**98.0**	90.0	96.0	76.0	18.0	40.0

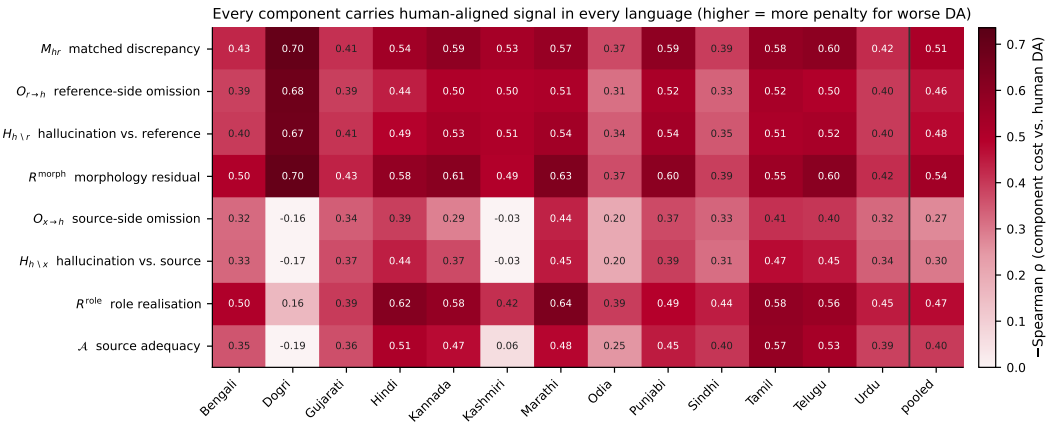


Figure 6: Component-level Spearman correlation with human DA per language. Sign flipped so that a higher value means a larger penalty for a worse translation.

represented in that set. The dominant class is not a metric failure at all, because 31.7% of the worst residuals are hypotheses identical to the reference that received a DA below 0.7, a property holding for 2.6% of the test set, hence enriched $12.4\times$, a pattern the BERTScore residuals share. Segments of at most three tokens are enriched $4.1\times$, since a percentile is unstable when one token decides the score, whereas properties indicating a normalization or coverage failure appear at roughly their base rate.

Table 13: Residual taxonomy over the worst 5% of ComTaIL segments ($N = 495$), with the enrichment of each property against its unrounded base rate.

	System family		Hypothesis flag					Source length		Language	
Property	Gold ref.	MT	Ident. ref.	+DA$<$0.7	$\leq$3 tok.	Latin	Mixed	Shortest	Longest	Odia	Dogri
All (%)	19.4	47.2	19.5	2.6	2.0	1.1	2.3	25.0	25.0	5.7	4.6
Worst (%)	35.4	37.0	35.4	31.7	8.1	1.4	2.6	38.2	17.6	10.3	1.8
Enrichment	1.82	0.78	1.81	**12.39**	**4.10**	1.28	1.13	1.53	0.70	1.82	0.40

I.2 REPRESENTATIVE RESIDUALS

Table 14 presents one segment from the worst-5% set per diagnosed cause. The first and fourth rows are the dominant class of Table 13, hypotheses identical to the reference that annotators nevertheless scored at 0.08 and 0.00, of which the fourth is a bibliographic citation that the reference itself leaves untranslated. The second row is the residual class that is specific to the frozen encoder, a Tamil output in which DNA is transliterated as Ti. en. e., which the encoder does not map onto the Latin acronym and which MORPHOT therefore charges as a source-side omission at $+3.0\sigma$. The third row is a fluent Bengali output with a mistranslated clause that every metric over-scores.

Table 14: Representative residuals of MORPHOT on ComTaIL. Segments drawn from the worst-5% set of Table 13, one per diagnosed cause. Percentiles are within language, as in Table 4.

#	Language, system. Source → hypothesis	DA	**MORPHOT**	BERTSCORE	CHRF++	Diagnosis
1	**Gujarati, gold.** *Another use case is VR/AR in the defense and military sector.* upayogano biijo kesa sanrakShaNa ane lashkarii kShetramaan viiaara/eaara che.	0.08	0.96	0.78	0.89	Hypothesis identical to the reference, yet DA < 0.7 (annotation noise)
2	**Tamil, Seamless.** *DNA is contained inside chromosomes.* Ti. en. e. kuromocomkaLukkuL uLLatu.	1.00	0.13	0.25	0.58	Faithful output under-scored: transliterated technical term (encoder confusion)
3	**Bengali, Google.** *The Japanese were beaten back before they could enter Imphal, which was one of the turning points of the overall war in South Asia.* jaapaaniraa imphaale prabesha karaara aagei piTiye piTiyechila, yaa chila dakShiNa eshiyaara saamagrika yuddhera anyatama Taarnin payenT.a	0.11	0.76	0.61	0.45	Low-quality MT output over-scored: fluent but mistranslated clause
4	**Kannada, gold.** *Pembroke AC and Marten RH.Unusual cutaneous reactions following diphtheria and tetanus immunization.Clin Exp Dermatol 4:345-348, 1979* Pembroke AC and Marten RH.Unusual cutaneous reactions following diphtheria and tetanus immunization.Clin Exp Dermatol 4:345-348, 1979	0.00	0.58	0.87	0.89	Untranslated citation kept verbatim in hypothesis and reference; annotators scored it 0

I.3 RESIDUAL BY HUMAN-QUALITY QUINTILE

Table 15 locates the residual mass along the human-quality axis. For every metric the over-scores concentrate in the lowest quintile and the under-scores at DA close to 1, where the percentile of a perfect translation cannot exceed the tie mass at the top. MORPHOT has the smallest error on the four lower quintiles, where the residual channels fire, and BLEURT-20 has the smallest error on the top quintile and, by less than 0.001, over all segments, because it separates near-perfect translations that MORPHOT ties.

Table 15: Mean absolute percentile residual by DA quintile (about 1,977 segments each). Lower is better.

DA quintile	**MORPHOT**	BLEURT	BERTSCORE	CHRF++	BLEU
Q1 [0.00, 0.48]	**0.143**	0.148	0.179	0.234	0.234
Q2 [0.48, 0.72]	**0.214**	0.217	0.233	0.251	0.255
Q3 [0.72, 0.84]	**0.215**	0.216	0.226	0.235	0.241
Q4 [0.84, 0.94]	**0.205**	0.211	0.225	0.248	0.251
Q5 [0.94, 1.00]	0.241	**0.226**	0.255	0.284	0.286
All	0.204	**0.203**	0.224	0.250	0.253

I.4 MORPHACES ACCURACY AND THE UNICODE-DECOMPOSITION TIE

MorphACES contrastive accuracy per perturbation is near ceiling for every metric, and we therefore summarize it in prose. Every surface metric, BERTScore, and MORPHOT separate all ten string-altering perturbations from the reference in at least 99.9% of pairs, whereas BLEURT-20 prefers the perturbed string in 2.3% of suffix truncations, 1.1% of entity and number swaps, and 0.0 to 0.7% of the remaining string-altering pairs, a trace of the morphology blindness of §1 in the learned metric. Those pairs differ from the reference by one inflectional marker, which the fine-tuned encoder treats as paraphrase, whereas the inner affix transport charges the difference explicitly (Table 5). The eleventh family, Unicode decomposition, is the protocol artifact already noted in Appendix C. MORPHOT and BLEURT-20 tie every such pair by design, which the strict $s^+ > s^-$ protocol counts as an error and reports as 0%, BERTScore separates 0.9% of them, and the surface metrics, which operate on the raw string, separate all of them.

I.5 Complete Ablation Grid With Perturbation Sensitivity

Table 16 reports the full ablation grid of §3.3, adding the ten variants that the main-text table omits, Spearman and Kendall correlations, and a sensitivity score on MorphACES. Removing root vectors or the morphological transport lowers correlation and sensitivity together, which is the evidence that those two components matter at this budget. The balanced and max-matching variants reach a higher sensitivity than the full model, because a balanced plan is forced to charge every unmatched token, and the unbalanced relaxation trades part of that sensitivity for the separable omission and hallucination channels that Figure 2 relies on. Nine variants lie within ± 0.006 Pearson and $\pm 0.03\sigma$ of the full model, a different set from the four components of §3.3. The default configuration itself comes from a 58-trial Optuna search on development preference accuracy.

Table 16: Extended ablation grid at a reduced training budget, where sensitivity is the mean reference-to-perturbation score margin on MorphACES and Δ is computed before rounding.

Variant	Pearson	Δ	Spearman	Kendall	Sensitivity	Δ
Full model	0.375	+0.000	0.352	0.242	1.05	+0.00
Affix-only (no root)	0.267	-0.108	0.290	0.199	0.78	-0.26
Root-only	0.396	+0.021	0.351	0.241	0.97	-0.07
− morph. transport	0.341	-0.034	0.326	0.223	0.79	-0.26
− inner affix UOT	0.377	+0.002	0.353	0.243	1.05	+0.00
− lexical gate	0.377	+0.002	0.352	0.242	1.04	-0.01
− role realisation	0.389	+0.014	0.373	0.257	1.07	+0.03
− unbalanced OT	0.396	+0.021	0.384	0.266	1.33	+0.28
max-matching (no OT)	0.398	+0.023	0.377	0.260	1.28	+0.24
− source anchor	0.375	-0.001	0.349	0.240	1.05	+0.01
− hard-entity cost	0.416	+0.041	0.371	0.255	1.22	+0.17
− phonology cost	0.375	+0.000	0.351	0.241	1.05	+0.00
uniform mass	0.376	+0.001	0.353	0.243	1.05	+0.00
IDF-only mass	0.379	+0.004	0.355	0.244	1.05	+0.00
top-k = 4	0.378	+0.003	0.354	0.244	1.04	-0.01
top-k = 8	0.375	-0.000	0.352	0.242	1.05	+0.00
dense plan (k = all)	0.377	+0.002	0.354	0.243	1.04	-0.00

I.6 Span-Level Error Localization

Across 300 MQM-MRL spans, the MORPHOT component contributing the largest score penalty coincides with the human-tagged error span at a token-level F1 of 87% for inflection errors, 81% for omissions, 73% for additions, and 69% for mistranslations. Such a level of agreement makes MORPHOT a practical starting point for span-level error attribution, whereas most metrics provide a sentence-level score alone. The mechanism is the one of §2.2: an inflection error raises the morphology cost of the matched pair through the inner affix transport while leaving its lexical cost almost untouched, so the penalty lands in the morphology channel at that span.

J Efficiency and Profiling

Cost Model and Inference-Time Devices: Per segment, MORPHOT solves three outer UOTs of sizes at most $n_x \times n_h$, $n_h \times n_r$, and $n_x \times n_r$, together with inner UOTs for token pairs that carry non-trivial affix mass, which amounts naively to $\mathcal{O}(n^2 I_{\text{Sinkhorn}})$ work at I_{Sinkhorn} iterations, namely 20 inner and 40 outer in our configuration. Two devices keep that cost manageable. Top-k truncation of the outer plan retains only the k largest entries per row and turns the dense $n \times n$ matrix into a sparse $n \times k$ object with no measurable quality loss at the default $k = 16$. Inner-UOT caching keyed by token pair amortizes repeated function words and stems across the corpus. The next paragraph reports the resulting throughput.

Wall-Clock and Memory Profiles: On an A100-80GB GPU, scoring 1,000 ComTaIL segments of mean length 24 tokens takes 14.7 s for MORPHOT with top-k truncation at $k = 16$, about 68 segments per second at a peak footprint of 4.1 GB, against 11.2 s and about 4 GB for BERTScore

and roughly 0.1 s for BLEU and chrF. MORPHOT with top-k truncation at $k = 16$ runs $3.7\times$ faster than the dense solver on the longest decile of sentence pairs, namely those exceeding 60 tokens, and stays within 0.002 Pearson of it (Appendix E), at a peak footprint of approximately 9 GB against 4.1 GB for top-k. The memory footprint scales linearly with sequence length up to 128 tokens, and beyond that length the sparse top-k plan becomes essential.

K Broader Impact, Licensing, and Dataset Attribution

Broader Impact: A more accurate evaluation metric for morphologically rich languages reduces a long-standing source of bias in MT benchmark leaderboards and lowers the cost of developing natural language generation systems for under-resourced MRL communities. Potential risks include over-reliance on a single metric for system selection and the possibility that a metric tuned toward morphological accuracy encourages generators to game specific failure modes. For that reason, we recommend applying MORPHOT alongside complementary approaches, such as reference-free quality estimation for fluency and human MQM annotation for fine-grained error attribution.

Licensing and Attribution: We release the MORPHOT codebase under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. The ComTaIL ratings come from COMTAIL (Ahsan et al., 2025) and remain under its CC-BY-NC-SA-4.0 license, the MQM-MRL span annotations come from IndicMT Eval (Sai B et al., 2023) and remain under its terms, and the XLM-R-large weights remain under their original MIT license. We performed no annotation and collected no new human-subject data.

L Ethics, Reproducibility, and AI-Use Statements

L.1 AI Use Statement

In this work, we used generative AI tools for editing and polishing the manuscript text and for checking the numerical claims against the released settings and outputs. We did not use generative AI tools for research ideation, for the design of the method, or for generating experimental results. Separately, we used generative AI tools to assist in composing parts (such as plot generation and formatting) of the released codebase, and we reviewed, tested, and verified all AI-assisted code for correctness. All the results are generated after running the model and making the predictions.

L.2 Ethics Statement

This work introduces an automatic evaluation metric and involves no human-subject research. All human judgments used for training and evaluation, namely the ComTaIL ratings and the MQM-MRL span annotations, come from existing publicly released datasets under their original licenses, and we collected no new annotation (Appendix K), which also covers licensing. A more faithful metric for morphologically rich languages can shift attention and resources toward under-served languages, while any metric used as a training reward can equally be optimized against, as the same appendix discusses. The authors declare no conflicts of interest.

L.3 Reproducibility Statement

We release the full MORPHOT implementation, the training and evaluation pipeline, and the exact settings of every reported run, together with the per-segment model outputs behind Tables 1, 2, 3, and 9, so those tables can be regenerated without re-scoring. Appendices C and D document the data, hardware, optimizer, precision, seeds, and artifacts, and Appendices E to G give the hyper-parameters, sensitivity sweeps, and per-language results with bootstrap significance. Appendix B contains the proofs of all theoretical claims with the assumptions on which each relies.