

Balancing Global Quality and Pronoun-Specific Feedback for Context-Aware Machine Translation

Harshit Dhankhar^{1*}, Baban Gain², Asif Ekbal², Yogesh Mani Tripathi³

¹Department of Computer Science & Engineering, University of California San Diego, USA

²Department of Computer Science & Engineering, Indian Institute of Technology Patna, India

³Department of Mathematics, Indian Institute of Technology Patna, India

Abstract

Context-aware machine translation can expose the evidence needed for pronoun choice, but standard fine-tuning does not explicitly prioritize these sparse discourse-sensitive decisions. We study PRONMT, a reward-guided iterative self-training method that combines sentence-level quality estimation with a signed confidence signal at generated pronoun positions. For each current sentence and its preceding source context, PRONMT samples candidate translations, scores them using reference-free quality estimation together with a reference-derived pronoun label, and fine-tunes on the highest-scoring candidate. On filtered English–German Europarl and English–French News Commentary data, PRONMT improves over context-aware supervised fine-tuning on BLEU and COMET. Ablations show that pronoun-only feedback can severely degrade sentence-level translation quality on these pronoun-focused data, while hard binary feedback underperforms confidence-weighted feedback. These results indicate that targeted linguistic feedback is most useful when combined with both a global quality signal and the context relevant to the targeted decision. We make the code publicly available at <https://github.com/Harshit2807161/ProNMT>.

1 Introduction

Pronoun translation is a compact but difficult test of context-aware machine translation. A pronoun in the current sentence may be ambiguous without an antecedent or discourse cue from a preceding sentence, and a system that accepts additional context does not necessarily use that context for the relevant decision (Li et al., 2020; Appicharla et al., 2023). Figure 1 illustrates an English–German example in which the translation of *it* depends on the antecedent *advertising* in the preceding sentence.

EN This advertising cheapens all women.
It cheapens every one of us and our daughters.

DE Eine solche Art von Werbung erniedrigt alle Frauen.
Sie erniedrigt uns alle und unsere Töchter.

Figure 1: A preceding source sentence provides evidence for pronoun choice in the current sentence. Pronouns are italicized and antecedents are underlined. Data: Europarl English–German.

Context-aware architectures and training objectives improve access to information beyond the sentence boundary (Tiedemann and Scherrer, 2017; Junczys-Dowmunt, 2019; Voita et al., 2018). However, broad sentence-level objectives do not explicitly prioritize sparse linguistic decisions such as pronoun choice. Quality estimation (QE) offers a practical global signal for feedback-guided MT training because it can score candidate translations without human judgments or a reference at scoring time (Rei et al., 2021; He et al., 2024). Yet a global QE score may overlook a locally important pronoun error.

A natural response is to add a pronoun-specific signal. Such a signal introduces a second problem: optimizing a narrow local objective may reward candidates that satisfy the pronoun criterion while damaging the rest of the translation. The interaction among context, global quality, and targeted feedback is therefore central. We study this interaction through PRONMT, a reward-guided iterative self-training framework. The translation model samples candidates, a combined reward ranks them, and the highest-scoring candidate becomes the target for a supervised update. The reward contains two components: reference-free QE for global translation quality and a reference-conditioned signed confidence signal at a generated pronoun position.

Our experiments attempt to answer the following four questions:

*Work done while at Indian Institute of Technology Patna.

1. Does pronoun-specific feedback add value beyond context-aware supervised fine-tuning?
2. Are global QE and local pronoun feedback complementary?
3. Is the local feedback more useful when preceding-sentence context is available?
4. Does confidence weighting rank candidates more effectively than a hard binary pronoun reward?

Our contributions are threefold. First, we formulate pronoun-aware QE-guided self-training that combines a global reference-free quality signal with reference-derived pronoun supervision. Second, we provide a controlled comparison of context-aware and context-agnostic conditions together with QE-only, pronoun-only, combined, and binary-reward variants. Third, experiments on English–German and English–French show that local pronoun feedback is harmful in isolation but complementary to global QE when the model can access potentially relevant preceding context.

2 Related Work

2.1 Context-Aware and Pronoun-Focused MT

Context-aware MT commonly incorporates preceding sentences through concatenation (Tiedemann and Scherrer, 2017; Junczys-Dowmunt, 2019), multiple encoders (Li et al., 2020), or auxiliary objectives (Appicharla et al., 2024). Prior work shows that context can improve anaphora-sensitive translation (Voita et al., 2018), but also that models may remain insensitive to whether the supplied context is relevant (Li et al., 2020; Appicharla et al., 2023). Preceding-sentence information has also been used in multilingual and document-aware MT systems (Wang et al., 2020; Sun et al., 2022; Wang et al., 2023).

ContraPro provides a large contrastive test suite for English–German pronoun translation (Müller et al., 2018). It is designed primarily for controlled evaluation rather than as a natural parallel corpus for iterative candidate-selection training. We therefore train on filtered parallel corpora and analyze the preliminary ContraPro training attempt in Section 4.2.

2.2 Feedback-Guided Training for MT

MT systems have been improved using human bandit feedback and other limited-feedback settings

(Kreutzer et al., 2018a,b). Reinforcement learning from human feedback can optimize learned preferences, but commonly requires human annotations and additional optimization components (Ouyang et al., 2022; Santacrose et al., 2023; Ramamurthy et al., 2023). In contrast, our procedure does not optimize a policy with PPO. It uses reward-based candidate selection followed by iterative supervised updates.

2.3 Quality Estimation as a Training Signal

QE estimates translation quality without a reference at inference time (Rei et al., 2020b, 2021). Recent work has used QE as a reward model for MT improvement (He et al., 2024). Our focus is complementary: we test whether a global QE signal benefits from an additional linguistically targeted signal for a sparse contextual decision, and whether that local signal remains safe when optimized jointly with sentence-level quality.

3 Pronoun-Aware QE-Guided Self-Training

3.1 Task and Context Representation

Let x_i be the current source sentence and x_{i-1} its preceding source sentence. In the context-aware condition, the translation model receives

$$\tilde{x}_i = \langle \text{context} \rangle x_{i-1} \langle / \text{context} \rangle x_i, \quad (1)$$

while the desired output is the translation of x_i only. In the context-agnostic condition, the model receives x_i . The target reference y_i^{ref} supplies the pronoun label used by the local reward. Thus, only the QE component is reference-free; the complete training objective combines reference-free QE with reference-derived pronoun supervision.

3.2 Candidate Selection and Iterative Updates

Let $M_i = P(y \mid x; \theta_i)$ denote the translation model at iteration i . For every input in a mini-batch, we sample up to k distinct candidate translations using multinomial sampling. Each candidate is scored by the reward in Equation 4. The highest-scoring candidate is then used as the target for an iterative supervised fine-tuning update. Algorithm 1 summarizes the procedure.

This is reward-guided self-training rather than reinforcement learning: the reward ranks sampled candidates, but the model update is a supervised update on the selected candidate.

Algorithm 1 PRONMT iterative self-training

Require: Training inputs $\mathcal{X}$, references $\mathcal{Y}$, initial model M_0 , reward r , batch size B , temperature T , candidate limit k

```

1: for  $i = 0, 1, \dots, N - 1$  do
2:    $D_i \leftarrow \text{SampleBatch}(\mathcal{X}, \mathcal{Y}, B)$ 
3:    $\mathcal{B} \leftarrow \emptyset$ 
4:   for each  $(x, y^{\text{ref}}) \in D_i$  do
5:      $y_1, \dots, y_k \sim P_T(y \mid x; \theta_i)$ 
6:      $y^* \leftarrow \arg \max_{y_j} r(x, y_j, y^{\text{ref}})$ 
7:      $\mathcal{B} \leftarrow \mathcal{B} \cup \{(x, y^*)\}$ 
8:   end for
9:   Update  $\theta_i$  on  $\mathcal{B}$  to obtain  $M_{i+1}$ 
10: end for

```

3.3 Global Translation-Quality Signal

For a candidate y , $R_{\text{QE}}(x_i, y)$ is produced by a reference-free COMET-QE model. In the context-aware condition, the translation model receives $\tilde{x}_i$, but QE scores the candidate against the current source sentence x_i . This prevents the context prefix from being treated as content that should appear in the translation.

3.4 Signed Pronoun-Position Confidence

For a generated candidate $y = (y_1, \dots, y_m)$, let j be a position at which the decoded candidate contains a recognized target-language pronoun. We define pronoun-position confidence as

$$C_{\text{pro}}(y, j) = \max_{v \in \mathcal{V}} P(v \mid x, y_{<j}; \theta), \quad (2)$$

where $\mathcal{V}$ is the target vocabulary. This quantity measures model confidence at the position where a pronoun is generated; it is not an independent pronoun-accuracy metric.

Let $p(y)$ be the generated pronoun and p^{ref} the aligned reference pronoun. If such a pronoun is not generated we reward that case 0. The signed local reward is then:

$$R_{\text{pro}}(y) = \begin{cases} +C_{\text{pro}}(y, j), & p(y) = p^{\text{ref}}, \\ -C_{\text{pro}}(y, j), & p(y) \neq p^{\text{ref}}, \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

The combined candidate-selection reward is

$$r(x_i, y, y_i^{\text{ref}}) = \alpha R_{\text{QE}}(x_i, y) + \beta R_{\text{pro}}(y). \quad (4)$$

QE provides a global constraint on sentence-level translation quality, while R_{pro} , also referred to as ProConf when averaged across multiple sentences, prioritizes the targeted pronoun decision.

Table 1: Source-sentence length statistics for ContraPro and filtered Europarl.

Dataset	Mean	Median	Var.	Std.	Max	Min
ContraPro	12.533	11.0	53.031	7.282	67	2
Europarl	33.378	29.0	404.219	20.105	212	1

3.5 Binary Pronoun Reward

To test whether confidence weighting is useful, we replace Equation 3 with

$$R_{\text{bin}}(y) = \begin{cases} +1, & p(y) = p^{\text{ref}}, \\ -1, & p(y) \neq p^{\text{ref}}, \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

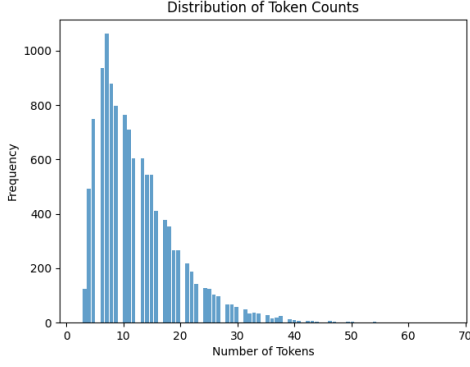
The binary variant assigns the same magnitude to every matching or non-matching candidate. The confidence-weighted variant therefore provides a finer ranking among candidate translations.

4 Experimental Setup**4.1 Pronoun-Focused Parallel Corpora**

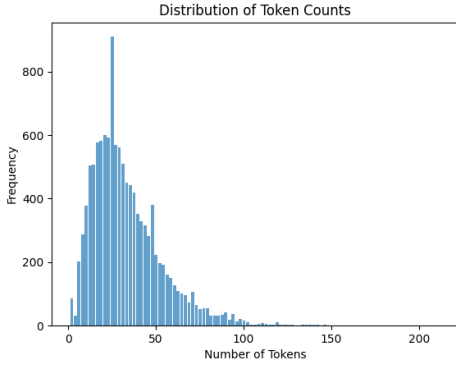
For English–German, we start from 1,920,209 Europarl sentence pairs. We retain a pair when the English sentence contains *it* exactly once, the German sentence contains exactly one singular third-person pronoun from *er*, *sie*, and *es*, and the two pronouns are aligned by fast_align (Dyer et al., 2013). Alignment filtering reduces noise from alternatives such as *das* and *dies* (Müller et al., 2018). The filtered data contains 117,834 *es*, 17,447 *er*, and 39,439 *sie* instances. We sample 15,000 instances per class and construct 42,750 training, 750 validation, and 1,500 test examples.

For English–French, we apply the same procedure to News Commentary v16 from OPUS (Tiedemann, 2009), retaining English *they* and its French counterparts *ils* and *elles*. The English–German setting therefore targets grammatical-gender alternatives for *it*, while English–French targets the masculine and feminine alternatives of *they*. The English–German reward weights are transferred to English–French without retuning.

For each corpus, the context-aware input prefixes the current source sentence with the preceding source sentence. The same data split and random seed are used for corresponding context-aware and context-agnostic conditions.



(a) ContraPro



(b) Filtered Europarl

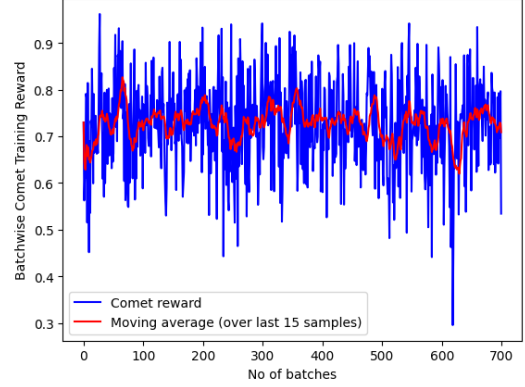
Figure 2: Source token-count distributions.

4.2 Choice of Training Data and ContraPro Analysis

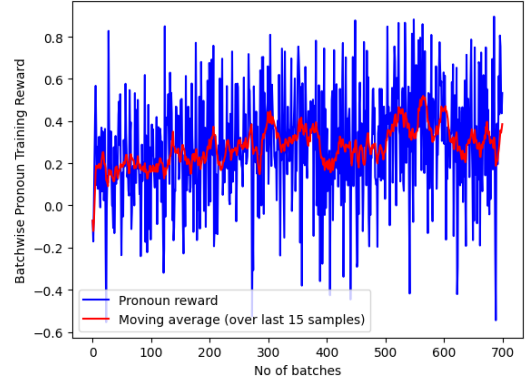
ContraPro is a controlled contrastive evaluation suite rather than a natural parallel corpus designed for iterative self-training. We nevertheless conducted a preliminary training attempt. The source sentences are shorter on average than those in our filtered Europarl corpus, as summarized in Table 1 and Figure 2. Before training, the average validation COMET-QE score was 0.7636 on ContraPro and 0.6323 on filtered Europarl. This left less QE headroom in the ContraPro training attempt (see Figure 3). Sentence length may contribute to this difference, but the current experiments do not establish a causal explanation.

4.3 Model and Training

All experiments use the distilled 600M-parameter NLLB-200 model. Before reward-guided training, the context-aware model is adapted through supervised fine-tuning so that it learns to consume the context prefix while translating only the current sentence. This context-aware SFT model is the principal baseline for evaluating the subsequent



(a) QE reward



(b) Pronoun reward

Figure 3: Preliminary ContraPro training curves over the first 700 updates.

reward-guided training stage after adaptation to the context-prefixed input format. We do not include a separately continued SFT control for the context-agnostic setting. Consequently, comparisons between context-aware and context-agnostic PRONMT should be interpreted as comparisons between the two training conditions rather than as a controlled estimate of the effect of preceding context alone.

For PRONMT, the mini-batch size is $B = 4$, and up to $k = 10$ distinct candidates are generated for each input by multinomial sampling. Iterative supervised training runs for at most 700 updates with learning rate 6×10^{-5} and gradient accumulation of 16. The binary-reward comparison uses the same setup and is trained for 750 updates. The supervised adaptation and reward-guided stages use the settings summarized below.

4.4 Compared Systems

We compare the following configurations:

- **Baseline:** the NLLB-200 model before task-specific reward-guided training;

Table 2: Hyperparameters for initial context-aware supervised fine-tuning.

Hyperparameter	Value
Learning rate	8×10^{-3}
Per-device train batch size	8
Gradient accumulation steps	16
Evaluation accumulation steps	16
Per-device evaluation batch size	8
Weight decay	0.01
Number of epochs	10

- **Context-aware SFT**: supervised adaptation to the context-prefixed input format;
- **QE only**: candidate selection using R_{QE} ;
- **Pronoun only**: candidate selection using R_{pro} ;
- **PRONMT**: joint QE and confidence-weighted pronoun feedback;
- **Binary**: joint QE and binary pronoun feedback.

4.5 Validation and Evaluation

For each reward-weight configuration, checkpoints are evaluated every 100 updates, and the checkpoint with the highest combined validation reward is retained. The reward weights α and β are then selected using COMET and BLEU on the validation set. Table 3 reports the corresponding English–German validation results. The held-out test set is used only for the final evaluations reported in Tables 4 and 5.

We report BLEU (Papineni et al., 2002) and reference-based COMET using *wmt22-comet-da* (Rei et al., 2020a). COMET is presented on a 0–100 scale. We also report the mean QE reward and mean signed pronoun-position confidence used for candidate selection. These internal reward values are diagnostic quantities, not independent measures of pronoun accuracy. English–German and English–French QE values are not directly comparable because the two experiments use different QE models.

5 Results and Analysis

5.1 Does Reward-Guided Training Improve over Context-Aware SFT?

Table 4 reports English–German test results. Context-aware SFT reaches 81.19 COMET and

25.57 BLEU. Joint QE and pronoun feedback improves these scores to 81.92 COMET and 26.96 BLEU, gains of 0.73 COMET and 1.39 BLEU. Although the COMET difference is modest, the improvement is consistent across both a learned reference-based metric and an overlap-based metric. The combined system also raises the internal pronoun-position confidence from 0.1406 to 0.4183. We do not treat this internal increase as an independent accuracy result, but it confirms that the selected candidates are more strongly aligned with the targeted reward while external translation quality is preserved.

The comparison with context-aware SFT distinguishes the initial adaptation to the context-prefixed input format from the subsequent reward-guided refinement stage. However, because PRONMT receives additional parameter updates after SFT, this comparison does not by itself isolate the effect of reward-guided candidate selection from that of additional training. We therefore interpret the gain over SFT as evidence that the complete PRONMT training procedure further improves the adapted model, rather than as a controlled estimate of the contribution of candidate selection alone.

The English–German improvement should nevertheless be interpreted as a refinement of an already strong system rather than a replacement for supervised context adaptation. Context-aware SFT accounts for most of the improvement over the raw context baseline, while the subsequent PRONMT stage is associated with a further improvement. This layered interpretation is more faithful to the training pipeline: SFT first establishes the ability to consume the extended input, and reward-guided self-training then encourages the model to prefer candidates that jointly satisfy sentence-level quality and the targeted pronoun criterion.

5.2 Why Are Both Reward Components Necessary?

The clearest result is the failure of pronoun-only optimization. In the context-aware English–German condition, pronoun-only training reaches a high internal ProConf value of 0.9675 but collapses to 16.79 COMET and 0.07 BLEU. The same conflict appears without context: pronoun-only training raises ProConf to 0.8998, yet COMET and BLEU fall below even the unadapted baseline. The local objective can therefore be satisfied while the rest of the translation deteriorates. ProConf should not be interpreted as pronoun accuracy, especially when

Table 3: English–German validation performance under different reward weights. #tokens is the sentence-specific token count and #avg-len ≈ 30 is the average source length. QE and ProConf are validation-set internal rewards.

Reward weights (α, β)	COMET	BLEU	QE	ProConf
<i>Without preceding context</i>				
(1, 1/#tokens)	72.88	15.2200	0.1189	0.3769
(1, 1/#avg-len)	73.27	15.2318	0.1026	0.3524
(1.2, 1/#tokens)	63.30	10.8723	0.0069	0.3225
(1.2, 1/#avg-len)	64.03	11.5233	0.0551	0.3115
<i>With preceding context</i>				
(1, 1/#tokens)	78.66	21.2250	0.0127	0.8270
(1, 1/#avg-len)	78.98	21.9542	0.0121	0.8543
(1.2, 1/#tokens)	80.72	26.4251	0.0223	0.4213
(1.2, 1/#avg-len)	80.21	25.2356	0.0311	0.4246

Table 4: English–German test performance. QE and ProConf are the mean internal QE reward and signed pronoun-position confidence. The reward weights are selected on the validation set.

Method	COMET	BLEU	QE	ProConf
<i>Without preceding context</i>				
Baseline	56.68	7.8904	-0.1154	0.0902
Pronoun only	50.30	1.8290	-0.1534	0.8998
QE only	58.30	2.6723	0.0480	0.3798
PRONMT	73.95	15.2630	0.1186	0.3672
Binary	63.02	12.2501	0.1052	0.7500
<i>With preceding context</i>				
Baseline	66.03	11.5783	0.0598	0.3370
Context-aware SFT	81.19	25.5680	0.0258	0.1406
Pronoun only	16.79	0.0703	-0.6024	0.9675
QE only	80.06	21.5588	0.0303	0.2262
PRONMT	81.92	26.9574	0.0362	0.4183
Binary	78.58	21.5808	0.0173	0.7500

considered without global translation metrics.

This failure is a candidate-selection analogue of reward over-optimization. The local term examines a single targeted position, so a candidate can receive a favorable local score even when its lexical choice, word order, coverage, or fluency elsewhere is poor. Repeatedly training on such candidates can amplify the mismatch between the optimized signal and the broader translation task. The near-zero BLEU score in the context-aware pronoun-only condition shows that this is not a minor trade-off. It is a qualitative collapse of the sentence-level objective. The result motivates using the pronoun signal as a preference among otherwise plausible translations, rather than as a standalone definition of quality.

QE-only selection avoids this collapse and

reaches 80.06 COMET and 21.56 BLEU, but it remains below the combined system at 81.92 COMET and 26.96 BLEU. The gap is particularly visible in BLEU, where the combined objective improves by 5.40 points over QE only. A global score can distinguish broadly adequate candidates, but a sparse pronoun error may have limited influence on the sentence-level score, especially when the remaining content is translated well. The local term changes the ordering among such candidates by assigning explicit importance to the pronoun position.

The two components therefore constrain different failure modes. QE discourages candidates that satisfy the local criterion by damaging the rest of the sentence, while pronoun feedback prevents a high global score from making the targeted deci-

sion effectively invisible. Their combination gives the best overall balance in the context-aware setting. Importantly, the combined reward does not require both component values to rise monotonically or to have directly comparable scales. Its role is to produce a useful ordering of sampled candidates, and the held-out COMET and BLEU results indicate that this ordering is more effective than either component alone.

5.3 How Does Preceding Context Affect Targeted Feedback?

The combined method also improves the context-agnostic model, from 56.68 to 73.95 COMET and from 7.89 to 15.26 BLEU. This result shows that the improvement is not dependent on access to a preceding-sentence cue. The complete PRONMT training procedure also improves the model when only the current sentence is available. At the same time, the highest overall performance is obtained when reward-guided training starts from a model adapted to the context-prefixed input. Relative to no-context PRONMT, the context-aware system is higher by 7.97 COMET and 11.69 BLEU.

This pattern is consistent with targeted feedback becoming more informative when the model can access evidence relevant to the targeted decision. Without preceding context, some pronoun choices remain underdetermined, and the reward can only favor agreement with the reference rather than teach a reproducible contextual mapping. With context, the sampled candidates can differ in whether they exploit the available antecedent information, giving the local reward a more meaningful basis for ranking them. The result does not establish that every filtered example requires cross-sentence resolution, nor does it directly measure context usage. It instead shows that the strongest trade-off between global quality and the targeted reward occurs in the condition where potentially relevant context is available.

The raw context baseline is not the main comparison because the base model has not yet learned the artificial context format. This issue is especially visible in English–French, where simply adding the prefix lowers COMET from 74.29 to 60.58 and BLEU from 26.23 to 16.42. After context-aware SFT, performance recovers to 76.84 COMET and 28.92 BLEU. These numbers show that access to additional text is not automatically beneficial: the model first needs to learn the structural distinction between contextual evidence and content to trans-

late. Context-aware SFT is therefore the fair baseline for assessing the contribution of PRONMT, while the raw context baseline mainly quantifies the cost of introducing an unfamiliar input format.

The different baseline behavior across English–German and English–French also cautions against treating context concatenation as a universally positive architectural intervention. Its value depends on adaptation, data characteristics, and how the prefix interacts with the pretrained model. The empirical pattern is narrower and clearer: once the model can process the context-prefixed format, the subsequent PRONMT stage is associated with further improvements in the produced outputs.

5.4 Confidence Weighting versus Binary Feedback

The binary variant obtains a larger raw pronoun reward but lower external translation scores. In context-aware English–German, binary feedback reaches 78.58 COMET and 21.58 BLEU, compared with 81.92 and 26.96 for confidence-weighted feedback. The same ordering holds without context, where the confidence-weighted system exceeds the binary variant by 10.93 COMET and 3.01 BLEU. Thus, the apparent advantage of the binary system on its own local reward does not transfer to the external metrics.

A hard $+1/-1$ signal gives every matching or non-matching candidate the same magnitude. Consequently, candidates with very different model support can become indistinguishable under the local component, and a single pronoun match can dominate small differences in QE depending on the reward scale. Pronoun-position confidence provides a finer candidate ranking because the magnitude reflects the model’s confidence at the generated pronoun position. In this selection-based setting, the benefit should not be described as reducing gradient variance, since gradients are taken through the supervised update rather than through the reward. The empirical advantage is instead that confidence weighting offers more resolution when candidates are ordered before the update.

The comparison also illustrates why raw values from differently defined rewards should not be compared as though they were the same metric. A binary score is structurally encouraged to approach its extreme values, whereas confidence-weighted scores depend on the model distribution. The relevant evidence is not that one internal number is numerically larger, but that the candidate ranking

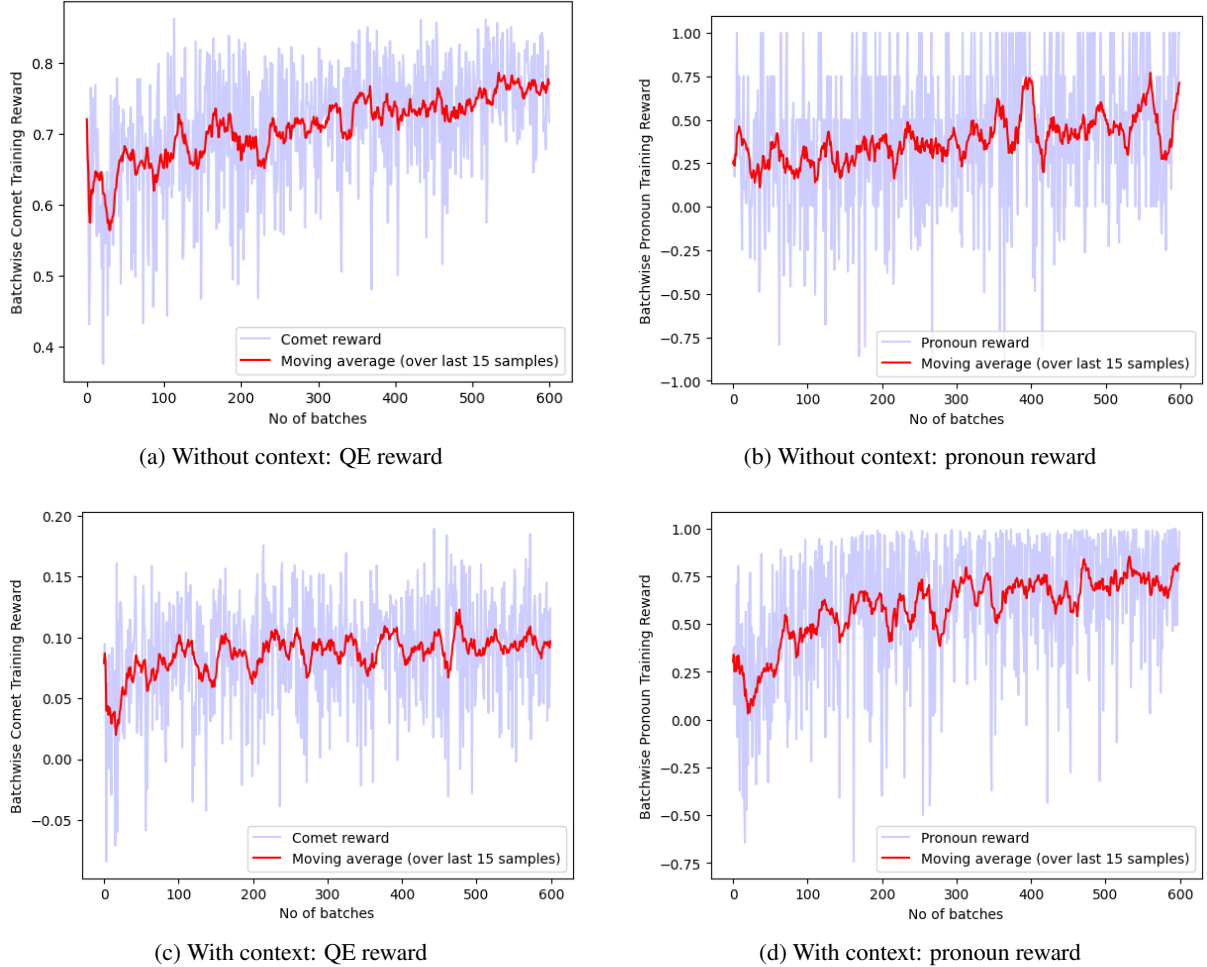


Figure 4: Batchwise reward trajectories for combined-reward training. Curves show raw values and moving averages.

induced by confidence weighting leads to better held-out translations.

5.5 Transfer to English–French

Table 5 reports English–French results using the English–German reward weights without retuning. In the context-aware condition, PRONMT improves over SFT from 76.84 to 79.37 COMET and from 28.92 to 30.54 BLEU. The corresponding gains are 2.53 COMET and 1.62 BLEU. The no-context condition also improves over its baseline, from 74.29 to 76.16 COMET and from 26.23 to 28.03 BLEU.

This transfer is informative because the French setting is not a direct replication of the German pronoun inventory. English–German focuses on *it* and the alternatives *er*, *sie*, and *es*, whereas English–French focuses on *they* and the gendered alternatives *ils* and *elles*. Reusing the same reward weights therefore tests whether the balance

between global and local feedback remains useful when both the language pair and the targeted mapping change. The positive result suggests that the central design choice is not tied only to the numerical optimum of the English–German validation set and can transfer to a second language pair and target mapping without retuning.

The evidence is still limited to two related target languages and a single translation model, so it does not establish broad multilingual generality. It does, however, strengthen the interpretation of the English–German result. The improvement is unlikely to be solely an idiosyncrasy of one filtered corpus or one particular set of German pronouns, because the same combined objective also improves a separately constructed French setting without additional tuning. The larger COMET gain over SFT in English–French further suggests that the value of reward-guided refinement can depend on how much room remains after supervised con-

Table 5: English–French test performance. English–German reward weights are reused without retuning. QE and ProConf are internal rewards and use a different QE model from Table 4.

Method	COMET	BLEU	QE	ProConf
<i>Without context</i>				
Baseline	74.29	26.2314	0.7610	0.3197
PRONMT	76.16	28.0341	0.7914	0.3846
<i>With context</i>				
Baseline	60.58	16.4238	0.7571	0.2806
Context-aware SFT	76.84	28.9204	0.7638	0.1678
PRONMT	79.37	30.5399	0.8007	0.4908

text adaptation.

5.6 What Does Reward-Weight Selection Reveal?

Table 3 provides an additional view of the trade-off between the two reward components. In the context-aware condition, the configuration with the highest internal ProConf is $(1, 1/\text{\#avg-len})$, which reaches 0.8543, but its COMET and BLEU scores are 78.98 and 21.95. The selected configuration, $(1.2, 1/\text{\#tokens})$, has a substantially lower ProConf of 0.4213 but the best external scores of 80.72 COMET and 26.43 BLEU. This mirrors the component ablation: maximizing the local diagnostic is not equivalent to producing the best translation system.

Sentence-specific normalization of the pronoun term is the better choice at the selected QE weight, exceeding average-length normalization by 0.51 COMET and 1.19 BLEU at $\alpha = 1.2$, though the two are within 0.32 COMET at $\alpha = 1$. The comparison suggests that the balance is affected not only by the nominal values of α and β , but also by whether the local contribution is normalized according to each sentence. Since a pronoun-level term is combined with a sentence-level QE score, sentence-specific scaling can prevent the relative influence of the local term from changing simply because inputs differ in length.

The no-context validation results show a different optimum: $(1, 1/\text{\#avg-len})$ gives the highest COMET and BLEU, while increasing α to 1.2 hurts both metrics. This difference supports selecting reward weights on the validation set separately for the two input conditions rather than assuming that a single balance must be optimal everywhere. More broadly, the sensitivity table reinforces the main narrative that global and local signals are complementary but not interchangeable. Their usefulness

depends on maintaining a scale at which the local term can influence candidate choice without overwhelming sentence-level quality.

5.7 Relationship between Internal Rewards and External Metrics

Across the ablations, internal rewards and external metrics are related but clearly not interchangeable. Pronoun-only and binary training can obtain large local values while producing lower COMET and BLEU, so the validation configuration with the highest ProConf is not the selected system. Conversely, context-aware SFT has a relatively low ProConf of 0.1406 but strong external performance. These cases show that ProConf is best understood as a description of the candidate-selection objective rather than a standalone evaluation measure.

QE is also an internal training signal, even though it is derived from a learned MT evaluation model. Optimizing the same QE model used for candidate selection may improve its score without guaranteeing equivalent gains under a separate evaluator. For that reason, the main results include reference-based COMET and BLEU, and the paper bases its conclusions on agreement between these held-out metrics rather than on the training rewards alone. The combined system’s advantage is strongest when it improves both external metrics while avoiding the local-reward collapse observed in the ablations.

5.8 Training Dynamics

Figure 4 shows the batchwise QE and pronoun rewards for the English–German combined-reward runs. Both raw trajectories fluctuate substantially because each update is based on sampled candidates and a small mini-batch. The moving averages are therefore more informative than individual batch values. In the no-context run, the QE moving average generally rises over training, while the pronoun trajectory improves more gradually and remains highly variable. In the context-aware run, the pronoun moving average increases markedly during the early updates and then fluctuates around a higher range, while the QE trajectory improves early and later remains comparatively stable.

The trajectories are compatible with the two reward components learning at different rates. The targeted signal can change quickly when candidate pronouns are re-ranked, whereas sentence-level quality aggregates many decisions and may move more conservatively. At the same time, the occa-

sional negative or low batch rewards show that sampling continues to expose difficult cases throughout training. A smooth monotonic curve is not expected because the procedure repeatedly changes the model that generates the candidate pool.

These plots are diagnostic rather than the basis for model selection. Checkpoints are selected using the combined validation reward, and final claims are drawn from held-out COMET and BLEU. This separation is important because an apparently improving training component can still correspond to poorer external quality, as demonstrated directly by the pronoun-only and binary ablations. The curves are therefore most useful for showing that optimization is active and noisy, while the tables determine whether the resulting model is actually better.

6 Conclusion

We presented PRONMT, a reward-guided iterative self-training method for balancing global translation quality with pronoun-specific feedback. The experiments support three conclusions. First, context-aware supervised fine-tuning is a strong starting point, but combined reward-guided training improves further in both English–German and English–French. Second, pronoun-only feedback is unsafe as an isolated objective because a high local reward can coincide with severe deterioration in COMET and BLEU. Third, confidence-weighted feedback provides a more useful candidate ranking than a hard binary signal. More broadly, targeted linguistic feedback is most effective when it is constrained by a global quality objective and paired with access to the information relevant to the targeted decision.

Limitations

The study covers one model and two pronoun-focused language-pair settings. The filtered corpora identify aligned pronoun alternatives, but they do not establish that every example requires preceding context. Because the evaluation sets are constructed around the targeted pronoun phenomena, improvements in BLEU and COMET on these sets do not establish improvements on naturally distributed MT data or on sentences that do not contain the targeted phenomenon. We also do not compare against contemporary LLM-based contextual MT systems; accordingly, our experiments characterize the behavior of the proposed training objective rather than establish state-of-the-art contextual MT

performance. The method also depends on the reliability of the QE model, accurate pronoun identification, word alignment during data construction, and validation-based selection of the reward weights. Because English–German and English–French use different QE models, their raw QE values are not directly comparable. Finally, the method targets phenomena that can be represented through explicit token-position signals; it may not transfer directly to discourse properties that lack such localized supervision.

Acknowledgments

Baban Gain acknowledges the Prime Minister’s Research Fellowship (PMRF), Ministry of Education, Government of India, for supporting this research under Grant No. 2703563. The authors gratefully acknowledge support from the grant titled “COIL-D: Centre of Indian Language Data,” sponsored by MeitY, Government of India (Grant No. 11(14)/2018-HCC(TDIL)Voll-III-Part(21)).

References

- Ramakrishna Appicharla, Baban Gain, Santanu Pal, and Asif Ekbal. 2023. [A case study on context encoding in multi-encoder based document-level neural machine translation](#). In *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 160–172, Macau SAR, China. Asia-Pacific Association for Machine Translation.
- Ramakrishna Appicharla, Baban Gain, Santanu Pal, Asif Ekbal, and Pushpak Bhattacharyya. 2024. [A case study on context-aware neural machine translation with multi-task learning](#). In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 246–257, Sheffield, UK. European Association for Machine Translation (EAMT).
- Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. [A simple, fast, and effective reparameterization of IBM model 2](#). In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.
- Zhiwei He, Xing Wang, Wenxiang Jiao, Zhuosheng Zhang, Rui Wang, Shuming Shi, and Zhaopeng Tu. 2024. [Improving machine translation with human feedback: An exploration of quality estimation as a reward model](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8164–8180, Mexico City, Mexico. Association for Computational Linguistics.

- Marcin Junczys-Dowmunt. 2019. [Microsoft translator at WMT 2019: Towards large-scale document-level neural machine translation](#). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 225–233, Florence, Italy. Association for Computational Linguistics.
- Julia Kreutzer, Shahram Khadivi, Evgeny Matusov, and Stefan Riezler. 2018a. [Can neural machine translation be improved with user feedback?](#) In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 92–105, New Orleans - Louisiana. Association for Computational Linguistics.
- Julia Kreutzer, Joshua Uyheng, and Stefan Riezler. 2018b. [Reliability and learnability of human bandit feedback for sequence-to-sequence reinforcement learning](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1777–1788, Melbourne, Australia. Association for Computational Linguistics.
- Bei Li, Hui Liu, Ziyang Wang, Yufan Jiang, Tong Xiao, Jingbo Zhu, Tongran Liu, and Changliang Li. 2020. [Does multi-encoder help? a case study on context-aware neural machine translation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3512–3518, Online. Association for Computational Linguistics.
- Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 2018. A Large-Scale Test Set for the Evaluation of Context-Aware Pronoun Translation in Neural Machine Translation. In *WMT 2018*, Brussels, Belgium. Association for Computational Linguistics.
- 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. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS ’22*, Red Hook, NY, USA. Curran Associates Inc.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Rajkumar Ramamurthy, Prithviraj Ammanabrolu, Kianté Brantley, Jack Hessel, Rafet Sifa, Christian Bauckhage, Hannaneh Hajishirzi, and Yejin Choi. 2023. [Is reinforcement learning \(not\) for natural language processing: Benchmarks, baselines, and building blocks for natural language policy optimization](#). Preprint, arXiv:2210.01241.
- Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. 2021. [Are references really needed? unbabel-IST 2021 submission for the metrics shared task](#). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, Online. Association for Computational Linguistics.
- Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020a. [COMET: A neural framework for MT evaluation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.
- Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020b. [Unbabel’s participation in the WMT20 metrics shared task](#). In *Proceedings of the Fifth Conference on Machine Translation*, pages 911–920, Online. Association for Computational Linguistics.
- Michael Santacrose, Yadong Lu, Han Yu, Yuanzhi Li, and Yelong Shen. 2023. Efficient rlhf: Reducing the memory usage of ppo. *arXiv preprint arXiv:2309.00754*.
- Zewei Sun, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Lei Li. 2022. [Re-thinking document-level neural machine translation](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3537–3548, Dublin, Ireland. Association for Computational Linguistics.
- Jörg Tiedemann. 2009. *News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces*, volume V, pages 237–248.
- Jörg Tiedemann and Yves Scherrer. 2017. [Neural machine translation with extended context](#). In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.
- Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. [Context-aware neural machine translation learns anaphora resolution](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1264–1274, Melbourne, Australia. Association for Computational Linguistics.
- Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. [Document-level machine translation with large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore. Association for Computational Linguistics.
- Longyue Wang, Zhaopeng Tu, Xing Wang, Li Ding, Liang Ding, and Shuming Shi. 2020. [Tencent AI lab machine translation systems for WMT20 chat translation task](#). In *Proceedings of the Fifth Conference on Machine Translation*, pages 483–491, Online. Association for Computational Linguistics.