Parameter-Efficient Quality Estimation *via* Frozen Recursive Models

Umar Abubacar **Roman Bauer** **Diptesh Kanojia**

NICE Research Group, University of Surrey, UK

Surrey Institute for People-Centred AI (PAI), University of Surrey, UK

{ua00104, r.bauer, d.kanojia}@surrey.ac.uk

Abstract

Tiny Recursive Models (TRM) achieve strong results on reasoning tasks through iterative refinement of a shared network. We investigate whether these recursive mechanisms transfer to Quality Estimation (QE) for low-resource languages using a three-phase methodology. Experiments on 8 language pairs on a low-resource QE dataset reveal three findings. First, TRM's recursive mechanisms do not transfer to QE. External iteration hurts performance, and internal recursion offers only narrow benefits. Next, representation quality dominates architectural choices, and lastly, *frozen* pre-trained embeddings match fine-tuned performance while reducing trainable parameters by 37× (7M vs 262M). TRM-QE with frozen XLM-R embeddings achieves a Spearman's correlation of 0.370, matching fine-tuned variants (0.369) and outperforming an equivalent-depth standard transformer (0.336). On Hindi and Tamil, frozen TRM-QE outperforms MonoTransQuest (560M parameters) with 80× fewer trainable parameters, suggesting that weight sharing combined with frozen embeddings enables parameter efficiency for QE. We release the code publicly for further research[1].

1 Introduction

While the dominant paradigm for enhancing reasoning relies on massive parameter scaling, Tiny Recursive Models (TRM) present a resource-efficient counter-narrative (Jolicoeur-Martineau, 2025). By applying iterative refinement, *i.e.*, looping data through a shared network multiple times, small models can match the performance of larger single-pass models on reasoning tasks. This raises an intriguing question: can these recursive reasoning mechanisms be transferred to evaluation tasks like Quality Estimation (QE)?

QE is fundamentally a problem of *cross-lingual reasoning*. Unlike translation, which focuses on generation, QE requires a model to align two semantic spaces, *source* and machine translation output (*hypothesis*), to perform granular error-based reasoning. The model must not only detect deviations but also quantify the severity of them, *i.e.,* numerical reasoning, to predict a quality score. In the current paradigm, large cross-encoders[2] like XLM-R-XXL (Conneau et al., 2020; Goyal et al., 2021) are continually pre-trained, and additionally fine-tuned on large-scale QE data for multiple language pairs, which is expensive for low-resource languages. Given *QE aims to predict translation quality without any reference translations*, it is critical for low-resource languages with limited data.

We investigate whether TRM's recursive efficiency can improve cross-lingual reasoning in resource-constrained scenarios using a QE dataset (Zerva et al., 2022; Blain et al., 2023; Sindhujan et al., 2025b) that challenges the model to predict translation quality for English to Indo-Aryan (Hindi, Marathi) and Dravidian (Tamil, Telugu) language families. These translation language pairs involve major typological divergences, such as aligning English's subject–verb–object (SVO) structure with the SOV order of Indic languages, across different scripts. The 'reasoning' required here is challenging, as the model must maintain context across long-range dependencies and overcome the relatively poor alignment in pre-trained spaces.

Our investigation demonstrates that while the specific 'reasoning' loops of TRM do not directly outperform standard approaches, we observe a "frozen efficiency" phenomenon where *frozen* pre-trained embeddings combined with a lightweight, weight-shared TRM head can match the performance of fully fine-tuned baselines while reducing trainable parameters. Our work suggests that,

[1]Code is available at https://github.com/surrey-nlp/TRMQE

[2]multilingual language models

across distinct language families, the key to efficient QE is not adapting the entire large encoder but learning a recursive reasoning mechanism over fixed representations.

2 Background

QE as reasoning. Quality Estimation has evolved from feature-engineered heuristics to deep learning architectures that learn directly from data. Sentence-level QE is typically modelled as a regression task, predicting a score y given a source s and translation t. The task is challenging and complex due to the necessary cross-lingual and numerical reasoning. The model must map qualitative errors to a quantitative penalty. For example, a *mistranslated named entity* may incur a higher penalty as compared to a *disfluent* word. Frameworks like TransQuest (Ranasinghe et al., 2020) tackle this by fine-tuning all parameters of massive encoders. While effective, this couples *representation learning* with *task reasoning*, leading to the compute requirements for fine-tuning and potential overfitting in low-resource settings. Existing work also leverages decoder-based large language models and proposes LoRA-based approaches (Sindhujan et al., 2025b), and reasoning via constraints like "annotation guidelines" within prompts (Sindhujan et al., 2025a).

TRM's architecture attempts to decouple computational depth from parameter count (Jolicoeur-Martineau, 2025). Unlike standard models where each layer l possesses unique weights W_l, TRM utilises a shared transformer block parameterised by θ applied repeatedly. TRM combines two mechanisms: *internal recursion* through L passes of the shared block (with 2 layers per pass, yielding $2L$ effective layers) and *external iteration* through refinement steps with adaptive halting. On reasoning tasks, small recursive models match larger single-pass models, but this benefit may not transfer to tasks where solutions are directly pattern-matched from inputs. We hypothesise that this recursive structure is relevant for QE as it theoretically allows the model to iteratively re-align source s and translation t, verifying semantic consistency akin to human re-reading. We probe if this recurrence can substitute for parameter width, allowing models to 'ponder' and refine internal representations before output. TRM achieves strong reasoning performance through a shared transformer applied recursively. We suggest that on reasoning tasks, small recursive models can match larger single-pass models, but this benefit may not transfer to tasks where solutions are directly pattern-matched from inputs.

Pair	Train	Test	Family
en-gu	7K	1K	Indo-Aryan
en-hi	7K	1K	Indo-Aryan
en-mr	26K	699	Indo-Aryan
en-ta	7K	1K	Dravidian
en-te	7K	1K	Dravidian
et-en	7K	1K	Uralic
ne-en	7K	1K	Indo-Aryan
si-en	7K	1K	Indo-Aryan

Table 1: Surrey Low-Resource QE Dataset statistics.

3 Experimental Setup

3.1 Dataset

Table 1 summarises the Surrey Low-Resource QE Dataset[3], which provides Direct Assessment annotations[4] (Specia et al., 2018; Fonseca et al., 2019) for 8 English-centric language pairs spanning three language families. The dataset totals approximately 75,000 training examples, with most pairs containing 7,000 training and 1,000 test examples. English-Marathi is notably larger with 26,000 training examples.

3.2 Model Architecture

TRM-QE adapts the TRM architecture for regression by repurposing the adaptive halting output for quality prediction. The original TRM uses a 2-dimensional output head (q_{halt}, $q_{continue}$) at the first sequence position for Adaptive Computation Time (Graves, 2016). We repurpose sigmoid(q_{halt}) as our quality score prediction. The model processes concatenated source-translation pairs through pretrained embeddings, applies L-cycles of recursive refinement through a shared transformer block, and produces the final prediction. Our baseline configuration uses 512-dimensional hidden states, 6 L-cycles, 2 layers per cycle, with a 7M parameter transformer core and 255M embedding parameters (262M total). Experiments were conducted on a single NVIDIA A100 GPU, with training taking approximately 5.5 hours.

[3] huggingface.co/surrey-nlp/Low-resource-QE-DA-dataset

[4] quality score per instance ranging from 0 to 100

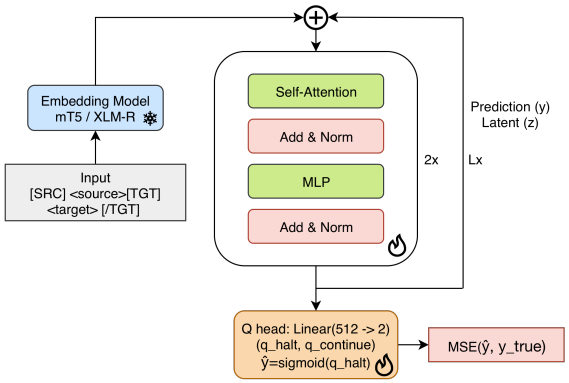

Figure 1: TRM-QE architecture. Source-translation pairs are marked with special tokens and encoded by a pretrained model (frozen or fine-tuned). The TRM applies L weight-shared cycles, then the Q head outputs a quality score via sigmoid activation.

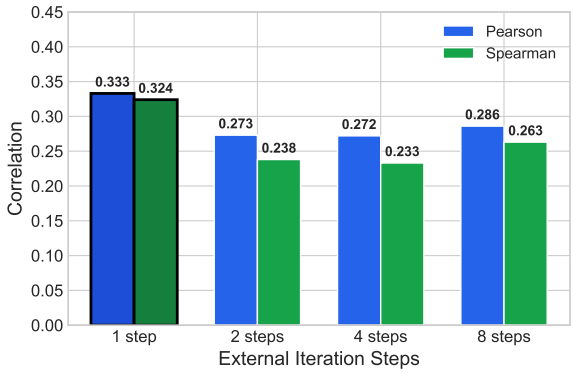

Figure 2: External iteration ablation: single-step models outperform multi-step variants across all step counts tested (1–16).

We evaluate on sentence-level QE, predicting z-normalised Direct Assessment scores transformed to $[0, 1]$ via sigmoid to match the model's output layer. Given source-translation pairs without references, we measure performance via Pearson and Spearman correlations.

3.3 Three-Phase Methodology

To disentangle their effects from input representation quality and training efficiency, our work adopts a three-phase methodology testing recursion with fixed representations, varying representations with fixed architecture, and comparing frozen *versus* fine-tuned embeddings.

Phase 1 fixes the input representation using fine-tuned mT5-small embeddings (512 dimensions) and systematically varies recursion parameters. We test external iteration steps from 1 to 16 and L-cycles from 1 to 6.

Phase 2 fixes the architecture (L=2 cycles, 1 external step) and varies representation quality with fine-tuned embeddings. We compare mT5-small, mT5-base, and XLM-R. For XLM-R (1024-dim), we use SVD projection to 512-dim to maintain consistent model capacity. This phase quantifies how much representation quality matters relative to architectural choices.

Phase 3 compares frozen versus fine-tuned embeddings to isolate the contribution of embedding adaptation. By *frozen*, we mean the pretrained XLM-R encoder weights remain fixed during QE training—only the 7M TRM transformer parameters are updated. By *fine-tuned*, we mean all 262M parameters (XLM-R encoder + TRM) are jointly optimised. We also compare fine-tuned TRM against a fine-tuned standard 8-layer transformer without weight sharing to isolate the effect of weight sharing.

4 Results

4.1 Phase 1: Recursion Effects

External Iteration Figure 2 presents results for varying external iteration steps while keeping L-cycles fixed at 6. Contrary to intuitions from reasoning tasks, single-step models achieve the highest correlation (0.333 Pearson, 0.324 Spearman), outperforming multi-step models. Performance degrades sharply from 1 to 2 steps (Spearman drops from 0.324 to 0.238), with no recovery at higher step counts.

Internal Recursion Figure 3 shows results for varying L-cycles with external iteration fixed at 1 step. The relationship between depth and performance is non-monotonic. L=1 (2 effective layers) underperforms at 0.295 Pearson, suggesting insufficient capacity for cross-lingual alignment. L=4 (8 layers) achieves peak performance (0.336 Pearson, 0.333 Spearman), while deeper configurations (L=6) degrade. This suggests a sweet spot around L=4 for mT5-small embeddings.

4.2 Phase 2: Representation Effects

Using L=2, 1-step configuration, we compare different embedding sources (we revisit optimal L-cycles for XLM-R with frozen embeddings in Phase 3). Table 2 shows that XLM-R (Conneau et al., 2020) embeddings achieve the best results at 0.387 Pearson and 0.369 Spearman, a 32% improvement over mT5-small (Xue et al., 2021)

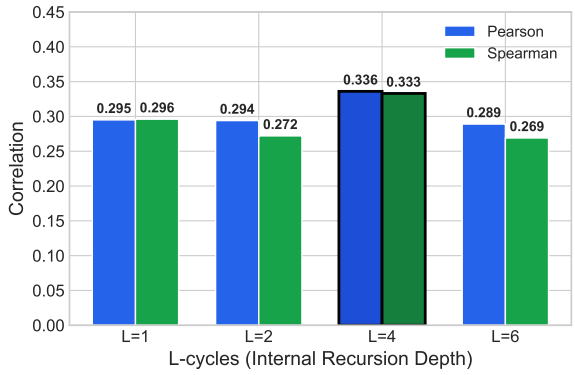

Figure 3: L-cycle ablation: performance peaks at L=4 (8 effective layers), with degradation at both shallower (L=1, L=2) and deeper (L=6) configurations.

Encoder	Dim	Pearson	Spearman
mT5-small	512	0.294	0.272
XLM-R	**1024→512**	**0.387**	**0.369**

Table 2: Embedding comparison with L=2, 1-step architecture.

(0.294 Pearson), confirming that representation quality dominates architectural choices for QE.

4.3 Phase 3: Frozen vs Fine-tuned Embeddings

Table 3 presents an insightful finding: *frozen XLM-R embeddings match or exceed fine-tuned performance while reducing trainable parameters by* 37×. TRM-QE with frozen embeddings (7M trainable) achieves 0.370 Spearman, matching the fine-tuned variant (0.369 with 262M trainable). *However, recursive depth is essential* as frozen L=1 achieves only 0.321 and L=6 degrades to 0.288, confirming L=4 as optimal, in our case. External iteration also hurts frozen models (0.356 with 4 steps vs 0.370 with 1 step), matching the fine-tuned pattern. *Weight sharing is critical for frozen embeddings.* A standard 8-layer transformer with frozen XLM-R achieves only 0.290 Spearman versus TRM's 0.370, a larger gap than for fine-tuned models (0.336 vs 0.369). Frozen embeddings also require strong representations: frozen mT5-small achieves only 0.297 versus frozen XLM-R's 0.370.

The frozen model's success suggests that fine-tuning large embeddings on limited QE data may cause overfitting, while frozen representations combined with TRM's weight-shared architecture provide better generalisation.

Model	Train.	Pears.	Spear.
Frozen XLM-R (L-cycle ablation)			
TRM (frz, L=4)	**7M**	**0.381**	**0.370**
TRM (frz, L=1)	7M	0.336	0.321
TRM (frz, L=6)	7M	0.296	0.288
Frozen Ablations			
TRM (frz, L=4, 4 steps)	7M	0.369	0.356
Std 8-layer (frz)	27M	0.313	0.290
TRM (frz mT5, L=4)	7M	0.327	0.297
Fine-tuned Baselines			
TRM (ft, L=2)	262M	0.387	0.369
Std 8-layer (ft)	262M	0.361	0.336
TRM (ft, L=4)	262M	0.339	0.323

Table 3: Frozen vs fine-tuned comparison. frz=frozen, ft=fine-tuned. Frozen XLM-R with TRM (7M) matches fine-tuned (0.370 vs 0.369).

4.4 Comparison with TransQuest

Tables 4 and Table 5 compare TRM-QE against MonoTransQuest (Ranasinghe et al., 2020). While MonoTransQuest achieves higher overall correlation (0.494 vs 0.370), frozen TRM-QE outperforms it on Hindi (+0.097) and Tamil (+0.039) with 80× fewer trainable parameters.

The performance gap across languages reflects XLM-R's pre-training data distribution (Conneau et al., 2020): Hindi and Tamil have substantially more Common Crawl data than Telugu and Sinhala. With frozen embeddings, lower-resource languages cannot compensate through task-specific adaptation. *Per-example error analysis confirms this pattern.* Tamil and Hindi achieve both strong correlation (0.556, 0.462) and low mean absolute error (0.098, 0.093) while Telugu predictions are near-random (0.164 Spearman, 0.161 MAE).

Interestingly, translation direction shows the opposite pattern to prior work: language pairs with English as source (en→X) achieve higher average Spearman (0.405) than those with English as target (X→en, 0.324). This differs from Sindhujan et al. (2025b), who found English-as-target performed better with fine-tuned models. With frozen embeddings, the model may rely more on source-side features for quality assessment, benefiting when the source is in the better-represented language (English).

5 Related Work

TransQuest (Ranasinghe et al., 2020) established XLM-R fine-tuning as the dominant QE approach. ALOPE (Sindhujan et al., 2025b) extends this by adaptively combining Transformer layers from fine-tuned InfoXLM. Our work explores a different

Model	Trainable	Total	Spearman
TRM-QE (frozen)	**7M**	262M	0.370
TRM-QE (fine-tuned)	262M	262M	0.369
MonoTransQuest	560M	560M	0.494
TransQuest†	560M	560M	0.592

Table 4: Comparison with baselines. †Average from Sindhujan et al. (2025b). Frozen TRM-QE matches fine-tuned with 37× fewer trainable parameters.

Language	TRM-QE	MonoTQ
en-ta (Tamil)	**0.556**	0.517
en-hi (Hindi)	**0.462**	0.365
en-gu (Gujarati)	0.423	**0.434**
en-mr (Marathi)	0.418	**0.458**
et-en (Estonian)	0.368	**0.741**
ne-en (Nepali)	0.333	**0.593**
si-en (Sinhala)	0.270	**0.527**
en-te (Telugu)	0.164	**0.199**
Overall	0.370	**0.494**

Table 5: Per-language Spearman correlation. TRM-QE (frozen, 7M trainable) vs MonoTransQuest (560M). Bold indicates best per row. TRM-QE outperforms on Tamil and Hindi.

direction: frozen pretrained embeddings with a lightweight task head, achieving comparable performance with drastically fewer trainable parameters. TRM (Jolicoeur-Martineau, 2025) and Universal Transformers (Dehghani et al., 2019) showed recursive depth can substitute for model size. Concurrent analysis of TRM on ARC-AGI (Roye-Azar et al., 2025) found that recursion is effectively shallow and most accuracy is achieved at the first step, corroborating our finding that recursion provides limited benefit for QE.

6 Conclusion

We find that TRM's recursive mechanisms do not transfer to QE: external iteration hurts performance, internal recursion shows narrow benefits, and representation quality dominates architectural choices. Frozen pretrained embeddings match fine-tuned performance (0.370 vs 0.369 Spearman) while reducing trainable parameters by 37× (7M vs 262M). However, sufficient recursive depth is essential. Frozen L=1 achieves only 0.321 and L=6 degrades to 0.288, confirming L=4 as optimal. Weight sharing is critical for frozen embeddings as a standard 8-layer transformer with frozen XLM-R achieves only 0.290 versus TRM's 0.370, a larger gap than for fine-tuned models (0.336 vs 0.369).

On Hindi and Tamil, frozen TRM-QE outperforms MonoTransQuest (560M parameters) with 80× fewer trainable parameters, demonstrating that parameter-efficient QE is viable for language pairs where fine-tuning large models is impractical.

Limitations

We evaluate only TRM's recursive architecture on the Surrey Low-Resource dataset; other recursive approaches (Universal Transformers, PonderNet) may behave differently, and validation on WMT QE shared tasks would strengthen generalisation claims. Architectural hyperparameters (L-cycles, external steps) were tuned on mT5-small in Phase 1; optimal settings may differ for other embeddings. Indeed, fine-tuned XLM-R performs best at L=2 rather than L=4. However, the key finding that single-step models outperform multi-step variants holds for both mT5-small and frozen XLM-R. Our adaptation reuses TRM's halting head for quality prediction; we tested a decoupled regression head which showed the same patterns (1-step best) but slightly worse performance. We report bootstrap confidence intervals but not multiple training seeds due to computational constraints; consistent patterns across ablations suggest robust findings. We test only sentence-level QE; word-level or document-level QE may exhibit different recursion dynamics.

Potential Risks. QE models that underestimate translation errors could propagate low-quality translations in downstream applications. Our frozen approach reduces compute requirements but inherits any biases present in pretrained XLM-R embeddings. Performance varies substantially across languages (Telugu: 0.164 vs Tamil: 0.556), risking unequal benefit across language communities.

Future Work

Key directions include investigating why Hindi and Tamil show competitive performance while Telugu and Sinhala lag, testing more permutations of architectures with frozen embeddings and identifying other NLP tasks recursion may benefit.

Acknowledgments

This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) DTP Studentship 2753922 for the University of Surrey. We thank the Surrey-NLP team for the low-resource QE dataset. Generative AI tools were used to assist with drafting text and code; all content was reviewed by the authors.

References

Frédéric Blain, Chrysoula Zerva, Ricardo Rei, Nuno M. Guerreiro, Diptesh Kanojia, José G. C. de Souza, Beatriz Silva, Tânia Vaz, Jingxuan Yan, Fatemeh Azadi, Constantin Oraşan, and André Martins. 2023. Findings of the WMT 2023 shared task on quality estimation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 629–653, Singapore. Association for Computational Linguistics.

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

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Łukasz Kaiser. 2019. Universal transformers. In *International Conference on Learning Representations*.

Erick Fonseca, Lisa Yankovskaya, André F. T. Martins, Mark Fishel, and Christian Federmann. 2019. Findings of the WMT 2019 shared tasks on quality estimation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 1–10, Florence, Italy. Association for Computational Linguistics.

Naman Goyal, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau. 2021. Larger-scale transformers for multilingual masked language modeling. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 29–33, Online. Association for Computational Linguistics.

Alex Graves. 2016. Adaptive computation time for recurrent neural networks. *arXiv preprint arXiv:1603.08983*.

Alexia Jolicoeur-Martineau. 2025. Less is more: Recursive reasoning with tiny networks. *arXiv preprint arXiv:2510.04871*.

Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2020. TransQuest: Translation quality estimation with cross-lingual transformers. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5070–5081, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Antonio Roye-Azar, Santiago Vargas-Naranjo, Dhruv Ghai, Nithin Balamurugan, and Rayan Amir. 2025. Tiny recursive models on ARC-AGI-1: Inductive biases, identity conditioning, and test-time compute. *arXiv preprint arXiv:2512.11847*.

Archchana Sindhujan, Diptesh Kanojia, Constantin Orasan, and Shenbin Qian. 2025a. When LLMs struggle: Reference-less translation evaluation for low-resource languages. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 437–459, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Archchana Sindhujan, Shenbin Qian, Chi Chun Matthew Chan, Constantin Orasan, and Diptesh Kanojia. 2025b. ALOPE: Adaptive layer optimization for translation quality estimation using large language models. *arXiv preprint arXiv:2508.07484*.

Lucia Specia, Frédéric Blain, Varvara Logacheva, Ramón F. Astudillo, and André F. T. Martins. 2018. Findings of the WMT 2018 shared task on quality estimation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 689–709, Brussels, Belgium. Association for Computational Linguistics.

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

Chrysoula Zerva, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José G. C. de Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Oraşan, Marina Fomicheva, André F. T. Martins, and Lucia Specia. 2022. Findings of the WMT 2022 shared task on quality estimation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 69–99, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.