# Quality-Aware Self-Correcting Speech Translation, on an Edge Device!

Zubair Ajmal Farooq · Diptesh Kanojia
School of Computer Science and Electronic Engineering, University of Surrey

NATURE INSPIRED COMPUTING AND ENGINEERING RESEARCH GROUP

UNIVERSITY OF SURREY

**No cloud dependency** | **Fully offline on 4 GB Jetson Nano** | **QE-triggered self-correction** | **Demo: EN→ES, FR, DE · Evaluated: EN→ES**

## RESEARCH QUESTIONS

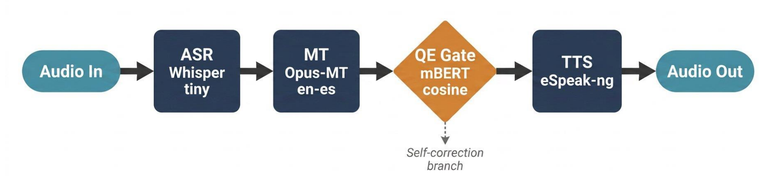
- **RQ1** Can a 4 GB Jetson Nano deliver fully offline, quality-aware, self-correcting speech translation without cloud dependency?
- **RQ2** Does QE-triggered N-best reranking produce measurable improvements in BLEU, ChrF, and COMET over greedy decoding?
- **RQ3** Does Minimum Bayes Risk selection outperform direct QE reranking as the candidate selection criterion?

## THE PROBLEM

Cloud-based speech translation exposes voice, speaker identity, and context to external servers, unacceptable in healthcare, legal, and low-bandwidth settings.

No published work has evaluated a fully integrated, quality-aware, self-correcting pipeline under a hard 4 GB memory ceiling.

## SYSTEM PIPELINE




*Dashed branch = conditional self-correction. If QE < τ, N-best beam triggered; MBR selects consensus.*

## BASELINE PERFORMANCE (EN→ES)

| 31.6% | 26.09 | 54.85 | 0.8485 |
|---|---|---|---|
| WER, Whisper tiny | BLEU, Opus-MT | ChrF baseline | COMET baseline |

*1,012 FLORES-200. Contextualised against NLLB-200 (34.6 BLEU, 3.3B params): no NLLB fits within 4 GB.*

## METHODOLOGY

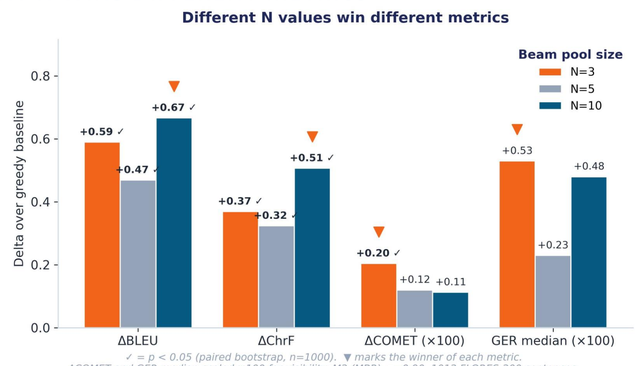
1. **Baseline phase.** Greedy decode 1,012 FLORES-200; record BLEU, ChrF, COMET; WER from Whisper tiny on FLEURS.
2. **Grid search.** Sweep τ ∈ {0.75–0.90} × N ∈ {3,5,10} × three methods (M1, M2, M3). 36 configurations on-device.
3. **Statistical validation.** Paired bootstrap resampling (n=1,000, seed=42) for corpus-level significance.
4. **Off-device cross-check.** COMET and GER computed on Google Colab; mBERT QE used on-device as a proxy.

## HEADLINE RESULT (EN→ES, M2 MBR, T=0.90)

*1,012 FLORES-200 · paired bootstrap n=1,000, seed=42 · no model retraining*

| ΔBLEU | ΔChrF | ΔCOMET |
|---|---|---|
| **+0.67** | **+0.51** | **+0.0020** |
| *best at N=10* | *best at N=10* | *best at N=3* |
| p < 0.001 | p < 0.001 | p = 0.002 |
| 95% CI [+0.35, +1.03] | 95% CI [+0.30, +0.74] | 95% CI [+0.001, +0.004] |

## DIFFERENT N WINS DIFFERENT METRICS




✓ = p < 0.05 (paired bootstrap, n=1000). ▼ marks the winner of each metric.
ΔCOMET and GER median scaled ×100 for visibility. M2 (MBR), τ=0.90, 1012 FLORES-200 sentences.

*N=10 maximises BLEU/ChrF. N=3 wins COMET and GER median — the recommended config for quality-per-edit.*

## KEY FINDING: QE IS A GATE, NOT A RANKER

| Method | ΔBLEU | ΔChrF | ΔCOMET | Sig. |
|---|---|---|---|---|
| M1, QE-rerank | −0.13 | −0.11 | −0.0004 | n.s. |
| M2, MBR (best) | **+0.67** | **+0.50** | **+0.0011** | ✓ |
| M3, Constrained Beam Search | −0.95 | −0.86 | −0.0061 | worse |

Removing 680 MB mBERT from candidate selection (M2 vs M1) does not hurt — it improves performance. MBR peer-consensus ChrF outperforms direct QE reranking. M3 anchors propagate first-pass errors into the search space.

## TRIGGER RATE VS THRESHOLD (T)

| 0.1% | 1.3% | 22.7% | 86.2% |
|---|---|---|---|
| τ = 0.75 | τ = 0.80 | τ = 0.85 | τ = 0.90 ★ |

Live demo QE scores (0.82–0.88) all fall below τ=0.90, confirming the threshold suits natural speech. QE-COMET Pearson r=0.41 (p<0.001).

## ADAPTED METRIC: GAIN-TO-EDIT RATIO (GER)

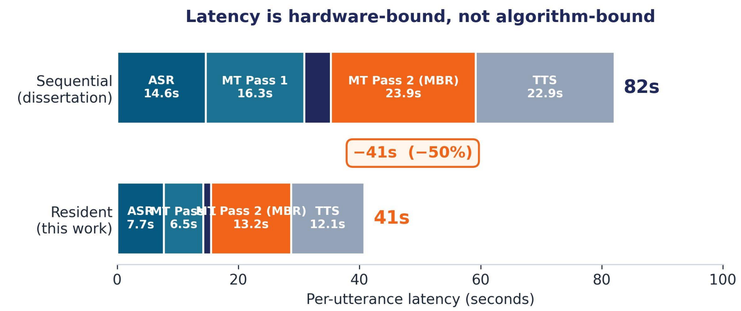
Adapted from HTER. Measures semantic gain relative to edit distance, surfacing surgical vs blunt rewrites:

```
GER = ΔCOMET / TER(original, corrected)
```

| Config | GER (median) | N_corr | Interpretation |
|---|---|---|---|
| M2, N=3 | **+0.0053 ★** | 485 | Highest gain per typical edit |
| M2, N=5 | **+0.0023** | 542 | Diminishing per-edit gains |
| M2, N=10 | **+0.0048** | 580 | More edits, smaller typical gain |

*★ N=3 chosen: highest median gain per correction, aligns with best ΔCOMET. Mean-GER picks N=5 but is skewed by a few large rewrites; median is the robust per-sentence measure.*

## LATENCY: HARDWARE, NOT ALGORITHM




*Sequential: per-utterance load+unload around each stage (dissertation Table 5.3 averages).*
*Resident: all four models held in memory simultaneously (measured on Jetson Nano, en-de utterance).*

Sequential (measured, 4 GB Nano): per-utterance model reload dominates overhead. Phase 3B baseline (resident, no correction): ≈1,840 ms.
Projected M2 on 8 GB Orin: ≈8–12 s

## MEMORY BUDGET (SEQUENTIAL LOAD)

| 160 MB | 180 MB | 680 MB | <10 MB |
|---|---|---|---|
| Whisper tiny | Opus-MT | mBERT (QE) | eSpeak-ng |

*Bar above shows actual proportions. Peak RAM 3.33 GB. OS + runtime ~2–2.5 GB.*

## FUTURE WORK

- → **8 GB Jetson Orin Nano:** Eliminates model-reload overhead; M2 latency ≈8–12 s.
- → **INT8 + TurboQuant:** Quantise Opus-MT & mBERT for simultaneous loading.
- → **Trained QE (COMET-Kiwi INT8):** Stronger gate; current r=0.41 leaves headroom.
- → **Encoder-level QE injection:** Embed quality into Opus-MT encoder (Koneru et al. 2025).
- → **Additional language pairs:** EN-Finnish, EN-Turkish to test MBR generalisation.

## LIMITATIONS

- Single language pair (EN→ES); generalisation unverified
- WER 31.59% propagates ASR errors downstream
- mBERT cosine similarity is a proxy QE signal (r=0.41 vs COMET)
- FLORES-200 is favourable; noisier domains would trigger more often

**01 — Privacy by Architecture**
Audio never leaves the device. Temp WAV files deleted post-transcription. No cloud, no logs.

**02 — QE = Gate, Not Ranker**
Removing mBERT from selection (M2 vs M1) improves performance. 680 MB saved on critical path.

**03 — GER Surfaces Efficiency**
N=10 maximises corpus BLEU. N=3 wins COMET and GER median — the recommended config for quality-per-edit.

**04 — Bottleneck is RAM, Not Research**
23,873 ms overhead is model-reload cost under 4 GB ceiling. On 8 GB hardware, near real-time.

**KEY REFERENCES**

Radford et al. (2023) Whisper.
Tiedemann & Thottingal (2020) Opus-MT.
Devlin et al. (2019) BERT.
Rei et al. (2020) COMET.
Goyal et al. (2022) FLORES-200.
Fernandes et al. (2022) QA decoding.

**CONTACT**

za00640@surrey.ac.uk

University of Surrey
Dept. of Computer Science