Translation: When should AI take a detour?

Paulo Rodrigues
The University of Chicago College
paulorossi@uchicago.edu

Jianghuai Li
The University of Chicago College
jianghuai@uchicago.edu

Abstract

This paper investigates two interconnected questions in neural machine translation: (1) Can intermediate pivot languages improve translation quality? (2) Can translation quality be predicted without human references? We evaluate pivot-based translation across high and low-resource language pairs, examining how pivot language selection and chain configuration affect output quality. Simultaneously, we assess the effectiveness of RT metrics as reference-free quality signals, comparing them against state-of-the-art neural estimators like COMETkiwi. We find that pivot languages provide benefits only under specific circumstances such as linguistic proximity, which must outweigh the cost of information loss across hops. We further find that Round Trip (RT) metrics behave as coarse quality classifiers: they are effective at flagging bad translations but have limited utility separating good from great. These findings have practical implications for translation pipeline design. Pivot languages should be used sparingly and selected to minimize linguistic distance. RT metrics are best deployed as safety rails for flagging bad translations for human review.

Keywords

Machine Translation, Pivot Languages, Quality Estimation, Round-Trip Translation, COMETkiwi, Low-Resource Translation

ACM Reference Format:
Paulo Rodrigues and Jianghuai Li. 2025. Translation: When should AI take a detour?. In . ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

1.1 Background

Large language models (LLMs) have demonstrated unprecedented translation capabilities in recent years, particularly between high-resource language pairs. However, significant challenges remain in two critical areas: translating for low-resource language settings where parallel training data is scarce, and evaluating translation quality without access to human references or judgments.

The concept of using intermediate or "pivot" languages to facilitate translation has long had history in machine translation research. The intuition is straightforward: if direct translation from language A to language C is difficult due to limited parallel data, perhaps translating A → B → C through a well-resourced pivot language B can yield better results. This approach has shown promise for certain language pairs, particularly when the pivot language shares linguistic features with either the source or target. [1]

Simultaneously, the field has seen significant advances in translation quality estimation (QE): the task of predicting translation quality without human references. Neural approaches such as COMETkiwi have tailored training techniques and data to achieve strong correlations with human judgments, enabling reference-free evaluation at scale [2]. However, these models require substantial training data and computational resources. An alternative approach, round-trip translation (RT), offers a fully unsupervised signal: if translating a sentence forward and back preserves its meaning, then intuitively the forward translation is likely to be good. [3]

Nonetheless, numerous questions remain about optimal translation strategies. When should a model take a "detour" through an intermediate language? How should pivots be selected and ordered? And can simple metrics like RT serve as reliable quality indicators? These are questions this paper seeks to explore.

1.2 Problem Statement

Despite extensive research on pivot-based translation and quality estimation, several gaps in knowledge remain. First, the conditions under which pivot languages help versus hurt translation quality remain poorly characterized for modern LLM-based systems. While prior work has demonstrated benefits for specific language pairs, systematic analysis of when pivoting provides net benefit, and when the information loss at each hop outweighs potential gains, is lacking.

Second, the importance of pivot language ordering in multi-hop translation chains has received limited attention. When multiple intermediates are used, does the sequence matter? If so, what principles should guide ordering decisions?

Third, while Moon et al. (2020) has demonstrated that RT metrics offer good reference-free quality signal at a system level, their study showed limited use at a sentence level. This study aims to build on their work by investigated the utility of RT metrics at a sentence or paragraph level. The practical utility of RT metrics relative to neural quality estimators like COMETkiwi has also not been thoroughly evaluated. Understanding the strengths and limitations of RT metrics will help determine their appropriate role in translation pipelines.

These gaps are important because they directly impact the design of production translation systems. Suboptimal pivot selection wastes computational resources and degrades user experience while poor quality metrics can allow poor translations to reach end users.

1.3 Objectives and Contributions

Our study evaluates pivot-based translation behavior (including pivot choice and chain length, and low-resource target settings)

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

and investigates whether RT metrics can predict quality without references and function as "safety rails." We make the following contributions: (1) demonstrate trends in pivot language selection and translation performance; (2) show that pivoting is highly sensitive to linguistic proximity and that costs can outweigh benefits; (3) study RT metrics as reference-free quality signals, including where they behave as coarse failure detectors; and (4) compare RT metrics against COMETkiwi and test whether combining them yields additional predictive value.

2 Literature Review

2.1 Pivot-Based Machine Translation

The concept of pivot languages in machine translation has been around since the era of statistical MT systems and has recently been adapted for neural approaches. Cheng et al. [4] introduced joint training algorithms for pivot-based NMT, demonstrating that simultaneously training source-to-pivot and pivot-to-target models leads to significant improvements over independent training. Their work on Europarl and WMT corpora established that pivot-based methods can boost NMT performance, particularly for resource-scarce language pairs.

Kim et al. extended this line of research with transfer learning strategies using pivot languages, proposing step-wise training and adapter components to connect pre-trained encoders and decoders. [5] Their methods achieved improvements of up to +2.6% BLEU on WMT 2019 French→German tasks and demonstrated effectiveness in zero-shot scenarios. Recent studies by Sulistyo et al. examined pivot-based NMT for low-resource Indonesian languages, showing that Indonesian as a pivot improved Javanese-Madurese translation quality across all BLEU n-gram scores. [6]

A consistent finding across this literature is that linguistic proximity matters: pivots that share features with either source or target tend to be more effective. However, most prior work has focused on dedicated NMT systems rather than modern LLMs, and systematic analysis of when pivot costs outweigh benefits remains limited.

2.2 Quality Estimation Without References

Quality estimation (QE) aims to predict translation quality without human references. The field has most recently evolved to large language model-based approaches. Rei et al. introduced COMET, a neural framework for MT evaluation that achieves high correlation with human judgments. [7] The reference-free variant, COMETkiwi, combines the COMET framework with the predictor-estimator architecture of OpenKiwi, achieving state-of-the-art results.

An alternative approach to reference-free evaluation is round-trip translation (RT). Moon et al. revisited RT-based QE, showing that while BLEU scores on round-trip translations were poor quality predictors, semantic embeddings applied to RT achieved the highest correlations with human judgments among WMT 2019 QE submissions. Our work builds on Moon et al.'s foundation by comparing RT metrics against neural estimators like COMETkiwi, and characterizing the conditions under which RT metrics provide useful signal.

2.3 Cross-Lingual Sentence Embeddings

Semantic RT metrics rely on cross-lingual sentence embeddings to measure similarity. Feng et al. introduced LaBSE (Language-agnostic BERT Sentence Embedding), a model that produces language-agnostic embeddings for 109 languages, [8] which achieved 83.7% accuracy on the 112-language Tatoeba benchmark – a substantial improvement over prior approaches like LASER (65.5%).

However, LaBSE suffers from embedding space anisotropy – the phenomenon where cosine similarity may not linearly correspond to semantic similarity. To resolve this issue, Wang et al. extended contrastive learning approaches to multilingual settings with mSimCSE, demonstrating that SimCSE-style training can yield isotropic cross-lingual embeddings without parallel data. [9] Our study uses LaBSE for semantic RT metrics while acknowledging this limitation, and we analyze its implications for quality prediction. This choice was made in view of COMETkiwi's inability to do monolingual comparisons.

2.4 Evaluation Benchmarks

Rigorous MT evaluation requires high-quality benchmarks with professional translations. Goyal et al. introduced FLORES-101, consisting of 3,001 sentences from English Wikipedia professionally translated into 101 languages. [10] Unlike prior benchmarks limited to specific domains or constructed semi-automatically, FLORES provides many-to-many aligned translations enabling evaluation of multilingual systems. The dataset has since been extended to FLORES+ covering over 200 languages and is now managed by the Open Language Data Initiative.

For quality metrics, we employ both surface-level (chrF) and semantic (COMET) measures. Popović introduced chrF as a character n-gram F-score that correlates better with human judgments and handles morphologically rich languages better than BLEU. [11] Our study uses FLORES+ as the dataset for multiple language pairs while employing both chrF and COMET to capture different aspects of translation quality.

2.5 Research Gaps

Several gaps remain from existing literature. While pivot-based translation has been extensively studied, most work predates modern LLMs and focuses on demonstrating benefits rather than characterizing when pivoting helps or perhaps hurts. Second, the effect of pivot ordering in multi-hop chains has received limited attention, despite practical importance for system design. Third, while Moon et al. showed promise for semantic RT metrics, their study was limited to showing utility of RT at a system level rather than a sentence or paragraph level. Systematic comparison against neural estimators like COMETkiwi and an analysis of where and when RT can complement COMETkiwi is also a novel contribution. These are gaps our study aims to address.

3 Methodology

3.1 Datasets and Models

All experiments use the FLORES-200 development set, a massively multilingual benchmark with 996 parallel sentences across 200+ languages. For the chain length experiment (RQ1), we translate

from English to Spanish via various pivot configurations. For the low-resource experiment (RQ2), we target six languages spanning three resource levels: Scottish Gaelic and Faroese (~20k speakers), Romanian and Tagalog (~2M speakers), and Swahili and Hindi (~20M+ speakers).

Translation is performed using OSS 120B via the CELS academic API infrastructure. Each translation step uses an independent context window to prevent the model from "peeking" at previous steps.

3.2 Evaluation Metrics

We employ three complementary evaluation metrics. BLEU measures n-gram overlap (0–100, higher is better) and serves as the standard surface-level metric. chrF computes character n-gram F-score (0–100, higher is better), which is more robust for morphologically rich languages. COMET is a neural MT evaluation metric trained on human judgments (0–1, higher is better), providing semantic quality assessment.

4 RQ1: Chain Length Experiment

This experiment evaluates how translation quality degrades as chain length increases when translating from English to Spanish. We compare two pivot selection strategies: a family-based approach using pivots from the same language family as the target (French, Italian, Portuguese), and a diverse approach using pivots from unrelated language families (Chinese, Arabic, Russian, Hindi, Swahili, Japanese, Korean).

4.1 Results by Chain Length and Pivot Type

Table 1: Chain Length Experiment Results (EN → ES)

k	Pivot Group	n	BLEU	chrF	COMET
0 (direct)	family	527	27.40	55.69	0.871
0 (direct)	diverse	461	29.18	56.87	0.871
1	family	389	26.13	54.77	0.865
1	diverse	401	22.33	51.02	0.853
3	family	387	25.66	53.65	0.862
3	diverse	385	18.40	47.78	0.842
5	family	384	25.38	53.47	0.866
5	diverse	381	15.06	43.93	0.798

Table 1 presents the complete results across all chain configurations. The data reveal a stark contrast between the two pivot strategies. Family-based pivots maintain remarkably stable quality across chain lengths, with BLEU scores remaining in the 25–27 range even with 5 intermediate languages and COMET scores hovering around 0.86. In contrast, diverse pivots show dramatic degradation: BLEU drops from 29.18 at baseline to 15.06 at k=5, a loss of 14 BLEU points, while COMET falls from 0.871 to 0.798.

Table 2: Average Scores by Chain Length

k	BLEU	chrF	COMET
0	28.29	56.28	0.871
1	24.23	52.90	0.859
3	22.03	50.72	0.852
5	20.22	48.70	0.832

Table 3: Average Scores by Pivot Group

Pivot Group	BLEU	chrF	COMET
Family (Romance)	26.14	54.40	0.866
Diverse	21.24	49.90	0.841

4.2 Key Findings for RQ1

The results demonstrate that family pivots maintain quality across chain lengths. Even with 5 intermediate languages, BLEU scores remain stable in the 25–27 range, and COMET stays around 0.86. This suggests that when pivots share linguistic features with the target language, the information loss at each hop is minimized.

Diverse pivots, by contrast, degrade significantly with chain length. The 14-point BLEU drop and the COMET decrease from 0.871 to 0.798 indicate substantial semantic drift when translations pass through typologically distant languages.

Perhaps the most striking finding is that linguistic similarity of pivots matters more than chain length. A 5-hop chain through Romance languages performs better than a 1-hop chain through Chinese. This contradicts a naive "error accumulation" model and suggests that the compatibility between pivot and target languages is the dominant factor.

Direct translation (k=0) achieves the best overall scores, though family-based pivoting shows minimal degradation. This confirms that for modern LLMs, the traditional rationale for pivot translation—leveraging stronger translation paths—does not apply when direct translation is already well-supported.

5 RQ2: Low-Resource Language Evaluation

This experiment evaluates translation quality to low-resource target languages grouped by approximate speaker population, comparing direct translation against pivot-based approaches. Six target languages were selected across three resource buckets: Scottish Gaelic (gla_Latn) and Faroese (fao_Latn) representing approximately 20,000 speakers; Romanian (ron_Latn) and Tagalog (tgl_Latn) representing approximately 2 million speakers; and Swahili (swh_Latn) and Hindi (hin_Deva) representing 20 million or more speakers. Four pivot languages were tested: French (fra_Latn), Spanish (spa_Latn), German (deu_Latn), and Chinese (zho_Hans).

5.1 Results by Language and Method

Table 4: Low-Resource Bucket Experiment Results

Bucket	Lang	Method	n	BLEU	chrF	COMET
20k	gla	direct	379	17.63	49.22	0.794
20k	gla	via_fra	382	15.58	47.57	0.786
20k	gla	via_spa	379	16.22	47.89	0.792
20k	gla	via_deu	383	15.38	47.16	0.789
20k	gla	via_zho	380	12.68	45.54	0.788
20k	fao	direct	385	17.49	44.35	0.638
20k	fao	via_fra	383	14.52	41.59	0.628
20k	fao	via_spa	390	14.60	41.98	0.629
20k	fao	via_deu	382	14.54	41.51	0.629
20k	fao	via_zho	389	11.00	38.11	0.621
2m	ron	direct	435	39.96	64.81	0.915
2m	ron	via_fra	388	35.60	62.32	0.909
2m	ron	via_spa	391	37.83	63.18	0.905
2m	ron	via_deu	388	34.43	60.91	0.906
2m	ron	via_zho	398	26.87	55.81	0.897
2m	tgl	direct	406	34.62	62.46	0.864
2m	tgl	via_fra	382	29.42	59.52	0.855
2m	tgl	via_spa	398	30.65	60.14	0.857
2m	tgl	via_deu	393	27.98	59.22	0.856
2m	tgl	via_zho	390	23.35	54.79	0.845
20m	swh	direct	413	32.15	62.47	0.852
20m	swh	via_fra	389	26.44	58.42	0.841
20m	swh	via_spa	392	29.47	60.69	0.848
20m	swh	via_deu	386	24.44	57.16	0.842
20m	swh	via_zho	380	19.48	53.59	0.833
20m	hin	direct	473	27.53	54.53	0.816
20m	hin	via_fra	389	23.97	51.01	0.806
20m	hin	via_spa	391	24.50	50.88	0.799
20m	hin	via_deu	387	21.98	49.74	0.795
20m	hin	via_zho	396	18.63	46.23	0.792

5.2 Summary Statistics

Table 5: Average Scores by Translation Method (across all 6 languages)

Method	BLEU	chrF	COMET
direct	28.23	56.31	0.813
via_spa	25.55	54.13	0.805
via_fra	24.26	53.41	0.804
via_deu	23.13	52.62	0.803
via_zho	18.67	49.01	0.796

Table 6: Average Scores by Population Bucket (direct only)

Bucket	BLEU	chrF	COMET
~20k speakers	17.56	46.79	0.716
~2M speakers	37.29	63.64	0.890
~20M+ speakers	29.84	58.50	0.834

5.3 Key Findings for RQ2

The results demonstrate that direct translation consistently outperforms all pivot-based approaches. Across all six target languages, direct EN→target translation achieves the highest BLEU, chrF, and COMET scores. This finding contradicts the traditional rationale for pivot translation in low-resource settings.

Pivoting via Chinese degrades quality the most severely. On average, pivoting through Chinese loses approximately 10 BLEU points compared to direct translation (28.23 → 18.67). This aligns with the chain length findings: typologically distant languages introduce more semantic drift per hop.

Romance pivots (Spanish, French) perform best among pivot options, which is consistent with the chain length experiment. Linguistically similar pivots preserve more semantic content, even when the target language is not in the Romance family.

Translation quality correlates strongly with language resource availability. The 2M speaker bucket (Romanian, Tagalog) achieves BLEU 37.29 and COMET 0.89, while the 20k bucket (Scottish Gaelic, Faroese) achieves only BLEU 17.56 and COMET 0.72. This suggests that model training data availability remains the dominant factor in translation quality.

Faroese presents a particularly challenging case. Despite similar speaker counts to Scottish Gaelic, Faroese achieves much lower COMET (0.64 vs 0.79), possibly due to less representation in training data or greater typological distance from English.

Romanian achieves the best overall scores (BLEU 39.96, COMET 0.92), likely due to its Romance language family connection to the predominantly Romance-heavy training data that most LLMs are exposed to.

6 RQ3: Round-Trip Consistency Analysis

Building upon previous work by Moon et al. (2020), we investigate whether round-trip (RT) translation metrics, a fully unsupervised signal, can predict translation quality without access to references. For this experiment, the FLORES+ dataset was used, for n = 100 for each unique source-pivot-target language combination. chrF score (*quality_chrf*) and COMET (*quality_comet*) were used to evaluate source-output translation and are known as surface and semantic quality metrics respectively (more in Appendix A.2.1).

Table 7: All translations are from English, total 1200 samples

Target	Pivots	Samples	Mean COMET
TUR	fra, deu, rus, cmn	400	0.8888
JPN	fra, deu, rus, cmn	400	0.9077
NPI	fra, deu, rus, cmn	400	0.8083

6.1 Predictive Metrics: Definitions and Calculations

Table 8: List of key predictive metrics

Metric	Calculation
rt_source	src → pivot → src
rt_hop2	pivot → tgt → pivot
rt_target	tgt → pivot → tgt
rt_min_st	min(rt_src, rt_hop2, rt_out)
rt_geom_st	$\sqrt[3]{(\text{rt\_src} \times \text{rt\_hop2} \times \text{rt\_out})}$
direct_src_out_comet	COMETkiwi(src, out)
combined_rt_direct	$\sqrt{(\text{rt\_geom\_sem\_st} \times \text{direct\_src\_out})}$

Semantic variants of the rt_* metrics listed above also exist, calculated using LaBSE-cosine similarity. The importance of the distinction between surface and semantic predictive and quality metrics will become apparent in later sections.

COMETkiwi is a neural based reference free translation evaluator belonging to the same family as COMET. It is of motivating interest to also study whether any combination of RT metrics can outperform COMETkiwi, or whether RT metrics combined with COMETkiwi can outperform either.

6.2 Analysis Methods for RQ3

Pearson correlation analysis was conducted to measure the correlation between predictor and quality methods. ROC AUC Analysis was conducted to measure how well the predictive metrics distinguish failures from successes, and stratified correlation analysis was used to understand whether predictive metrics work equally well across translation quality levels. Granularity analysis was done using spearman correlation to test how well the predictive metrics can rank translations within quality bands. Per-pivot decomposition was conducted to analyse how predictor performance varies across pivot languages, and finally multiple regression analysis tested whether RT predictive metrics adds predictive value beyond the COMETkiwi based direct_source_output_comet. More information about each of these methods are available in Appendix A.3.x.

6.3 Pearson Correlation and Implications

Table 9: Correlation to surface/semantic quality respectively

Predictor	TUR	JPN	NPI	Mean
rt_geom_sem_st	.48/.41	.34/.31	.31/.38	.38/.37
direct_src_out_comet	.34/.74	.40/.63	.31/.68	.35/.68
combined_rt_direct	.44/.73	.45/.58	.35/.64	.41/.65

The table show that there is significant correlation between the predictive metrics and semantic quality, and less so for surface quality. Since surface and semantic quality seem to produce different most-correlated metrics, it's important to keep in mind the effect of comparing surface/semantic based predictive metrics against surface/semantic based quality metrics, any combination of which would likely produce a different result.

For most meaningful analysis and to minimise potential complications, such as surface-based mechanisms not matching semantic based mechanisms, the rest of this study typically compares semantic predictive metrics to semantic quality metrics, avoiding the use of surface-based metrics unless insightful or meaningful. The claim that surface-based mechanisms behave differently to semantic based mechanisms is evidenced in Table B1.T1 in Appendix B, and by tables B4.T1-B4.T3 and B5.T1.

6.4 Smoke Detector Pattern Suggested by ROC AUC Analysis

Table 10: ROC AUC for quality_comet < 0.90/0.85/0.80

Predictor	TUR	JPN	NPI	Mean
rt_geom_sem_st	.67/.70/.81	.64/.70/.74	.82/.66/.68	.71/.69/.74
direct_src_out_comet	.86/.89/.91	.81/.81/.86	.91/.77/.83	.86/.82/.87
combined_rt_direct	.84/.86/.90	.77/.80/.84	.91/.75/.81	.84/.81/.85

As the quality of the translation drops, the discriminative power of the predictive metrics increases for TUR and JPN. Intuitively, this suggests that the predictive metrics are better at identifying when something goes wrong (a smoke detector), rather than between good and great. It's of note that rt_geometric_semantic_st (no influence of direct_source_output_comet) sees the greatest increase, suggesting that this trend has a greater correlation with the nature of RTs.

When combined with the Pearson correlation data from table 3, it's clear that RT based metrics and direct_source_output_comet correlate with quality_comet through different mechanisms. It is likely that RTs provide a more indirect measure of quality because RTs directly measure consistency and information loss rather than quality (which is what direct_source_output_comet measures).

Another insight to note is that the data for NPI doesn't follow the expected trend, and as a result the mean is skewed as well. This is likely due to extreme class imbalance caused by NPI's fundamentally different quality distribution (see Table B1.T2). If we change the quality thresholds to produce a similar distribution for NPI as with TUR and JPN with original thresholds, we get the expected trend which provides strong evidence for this explanation:

Table 11: ROC AUC for NPI after threshold adjustment for quality_comet

Adj. NPI threshold	Equiv. TUR	Failures	AUC
0.820	@0.90	198	0.657
0.750	@0.85	73	0.722
0.687	@0.80	21	0.803

6.5 Stratified Correlation Analysis

Table 12: Correlation with rt_geom_sem_st

Language	Low Stratum r	High Stratum r	Δ
TUR	0.369	0.004	+0.364
JPN	0.304	0.041	+0.264
NPI	0.390	0.134	+0.256
Mean	0.354	0.060	+0.295

These values essentially confirm the smoke detector hypothesis, and shows that it is robust across target languages despite NPI previously showing a deviant trend in ROC AUC analysis.

Table 13: Correlation with direct_src_out_comet

Language	Low Stratum r	High Stratum r	Δ
TUR	0.671	0.367	+0.303
JPN	0.472	0.351	+0.121
NPI	0.595	0.346	+0.249
Mean	0.579	0.355	+0.224

This table then suggests that direct_source_output_comet also shows a smoke detector like behavior, and in fact displays the same drop in correlation moving from low to high stratum, though displaying meaningful correlation even at high stratum compared to that of RT metrics which completely disappear.

In terms of practical implications, it's thus recommended to use RT metrics as alert flagging bad translations for human review rather than a ranker for translation quality since it fails at the higher-quality stratum. For differentiating higher-quality samples, direct_source_output_comet is the clear superior metric.

6.6 Granularity Analysis

Table 14: Using rt_geom_sem_st, quartiles for quality_comet

Quartile	TUR	JPN	NPI	Mean	Interpretation
Q1 (lowest)	0.347	0.250	0.320	0.306	Moderate
Q2	-0.037	0.039	0.100	0.034	None
Q3	-0.017	0.053	-0.055	-0.006	None
Q4 (highest)	0.054	0.286	0.130	0.157	Weak

It's clear that in Q2 and Q3 rt_geometric_semantic_st is essentially ranking randomly. A higher RT score provides no information about relative ranking, so this confirms that RT metrics are very coarse rankers and likely works best with a very small number of quality buckets. It is notable that in Q4 RT has a slight ranking ability, and this is an unexpected finding. However, the effect is weak and inconsistent across target languages.

One possible concern about this analysis is that low ranking ability is due to anisotropy in the embedding space of the LaBSE model, such that there isn't much difference between the RT values within Q2 or Q3. Calculations suggest RT standard deviation in Q1 is ~1.7x larger than in Q2-Q4, meaning anisotropy exists, but the translation quality against RT score regression slopes provide a potentially stronger suggestion: Q1 slopes are 0.3-0.8, while Q2 and Q3 slopes range from -0.05 to 0.08, and Q4 slopes are variable (which explains the weak ranking ability). So there is a genuinely flat relationship which suggests a more fundamental granularity limitation of the quality differentiation capability in RT metrics. See the scatter plot B6.P1 in Appendix B for a visualization of the slopes mentioned.

6.7 Per-Pivot Analysis

From Table B7.T1 and B7.T2 in Appendix B, the order of pivot language by translation quality is fra > deu > rus > cmn.

Table 15: Pivot language performance breakdown

Criterion	Best	2nd	3rd	Worst
ROC AUC	rus (.74)	fra (.69)	deu (.67)	cmn (.65)
Correlation	fra (.42)	rus (.39)	cmn (.37)	deu (.31)
Smoke Det. Δ	fra (.36)	cmn (.35)	deu (.28)	rus (.20)

Several patterns emerge from this analysis: French performs consistently well across all metrics, but it is likely not due to language family closeness (German is closer to English than French but performs basically worst). Russian achieves the highest ROC AUC but the lowest Smoke Detector Δ. This is a contradiction, but it reflects what these metrics measure. Rus maintains relatively consistent correlation with quality across both low-quality (r = 0.344) and high-quality strata. Mandarin exhibits the opposite pattern of having the worst AUC ROC but a good Smoke Detector Δ, showing reasonable correlation in the low-quality stratum (r = 0.365) but near-zero correlation in the high-quality stratum (r = 0.018).

These inconsistencies suggest that pivot language effectiveness for RT-based quality is not based on translation quality, language family closeness or other simple factors. It is potentially influenced by model characteristics and training data, the types of errors each pivot language tends to generate, as well as interactions with the target language which are not analyzed here. These results also demonstrate that ROC AUC, correlation and stratified correlation likely measure fundamentally different properties of prediction metrics. A pivot language that excels at one criterion may perform poorly on another, and the choice of evaluation metrics needs to be guided by intended application. For coarse quality filtering tasks, stratified analysis is most relevant.

6.8 Does RT Add Value Beyond COMETkiwi?

COMETkiwi and RT are only moderately correlated ($r \approx 0.45$), meaning they likely still measure different things. However, what RT uniquely captures does not predict COMET quality well. The partial correlation r(direct|RT) = 0.626 shows that after controlling for RT, direct comparison still strongly predicts quality, however the converse r(RT|direct) = 0.096, which suggests that RT provides almost no additional information towards the COMET quality metric. This means whatever RT captures by measuring what COMETkiwi doesn't is largely redundant towards COMET quality. This is supported by the multiple regression analysis below.

Table 16: Multiple regression for RT value beyond COMETkiwi

Lang	COMETkiwi R^2	Combined R^2	Impr.	COMETkiwi Weight	RT Weight
TUR	0.554	0.563	+0.88%	0.698	0.105
JPN	0.391	0.398	+0.66%	0.592	0.088
NPI	0.464	0.456	+0.10%	0.662	0.037

An important caveat is that the effect of RT beyond COMETkiwi should also depend on the semantic quality metric used (in this

discussion quality_comet). If instead a different quality metric were used, such as LaBSE with cosine similarity, it is entirely possible for RT metrics to produce a meaningful boost when combined with COMETkiwi. In fact this was observed in a previous pilot test, the results of which are omitted for brevity.

6.9 Safety Rail Method Validation

The concept of using RT metrics as a Safety Rail (flagging out bad translations for human review) was tested using n = 100 Europarl EN-ES sentence pairs across 2 translation chain types and using min(semantic RT) (LaBSE cosine similarity). The objective of the task was to correctly classify each sample into their respective Good/Bad labels. Only 2 quality bins were used in this analysis in view of the correlation analysis suggesting RT metrics are coarse classifiers.

Table 17: Chain type and path

Type	Path	Min RT	Mean RT
Good	EN→ES	0.970	0.980
Bad	EN→ZH→JA→DE→ES	0.912	0.956

Classification performance using min(semantic RT) yielded the following results:

ROC AUC: 0.943 (Excellent)
Cohen's d: 2.01 (Large)
Optimal Threshold: min_rt < 0.96
Precision: 85%
Recall: 94%

RT metrics work as a safety rail. They can reliably distinguish Good from Bad translations as defined above with 94.3% AUC.

7 Discussion

Our results across three research questions paint a consistent picture: modern LLMs do not benefit from pivot-based translation strategies, and round-trip consistency serves as an effective but limited quality signal.

7.1 Pivot Translation in the LLM Era

Unlike traditional SMT/NMT systems where pivoting through high-resource languages could improve low-resource translation, our experiments demonstrate that modern LLMs perform best with direct translation. This finding holds across chain lengths (RQ1), resource levels (RQ2), and pivot language choices. The mechanism appears straightforward: each translation hop introduces information loss, and LLMs are sufficiently capable that direct translation typically outperforms any indirect route.

The one exception—that linguistically similar pivots degrade less than distant ones—does not contradict this conclusion. Rather, it suggests that when chains are unavoidable, selecting pivots that share structural features with the target language minimizes the "telephone game" effect of cumulative translation errors.

7.2 Round-Trip Metrics as Quality Signals

The RQ3 analysis reveals that RT metrics function as effective "smoke detectors" but poor quality rankers. With ROC AUC of 0.94 for distinguishing good from catastrophically bad translations, RT consistency provides a practical reference-free signal for flagging translations that require human review. However, the near-zero correlation in high-quality strata indicates that RT cannot differentiate between good and excellent translations.

7.3 Practical Implications

For production translation systems, our findings suggest: (1) use direct translation as the default strategy; (2) if pivot translation is required for architectural reasons, select linguistically similar pivots and minimize chain length; (3) deploy RT metrics as a quality filter to catch catastrophic failures, but do not use them for fine-grained quality ranking.

8 Limitations and Future Work

8.1 Limitations

There are several limitations to our work. First, all experiments were conducted using a single model architecture (OSS 120B). While this provides consistency for our study, results may vary across different LLM architectures, sizes, and training regimes.

Second, our RT metric analysis relies on LaBSE embeddings for semantic similarity computation. LaBSE embeddings have been observed to exhibit anisotropy, the tendency for embeddings to cluster in a narrow cone rather than being evenly distributed. The effect of anisotropy has been discussed in section 6.6 Granularity Analysis, and we acknowledge the limitations it brings to quality evaluation, and provide arguments suggesting its limited effects towards our findings. Alternative embedding models (multilingual BERT variants) or different similarity metrics may yield different patterns.

Third, the choice of quality metric significantly influences our findings. RT metrics correlate differently with surface-level metrics (BLEU, chrF) versus semantic metrics (COMET, LaBSE similarity). The "smoke detector" pattern is strongest when using semantic quality definitions, but weaker or absent when using BLEU. This suggests that our conclusions about RT metric utility are conditional on the quality metric of interest.

Fourth, each research question was evaluated on a single dataset. RQ1 and RQ2 used FLORES-200 for English-to-Spanish and low-resource translation, while RQ3 used FLORES+ for English-to-Turkish/Japanese/Nepali. While FLORES is a well-established benchmark, results may not generalize to other domains (e.g., technical documentation, conversational text, literary translation) or other language pairs. The external validity of our findings requires validation across diverse datasets and domains.

Fifth, our pivot language selection was limited to four languages (French, German, Russian, Chinese) for RQ3 and four languages (French, Spanish, German, Chinese) for RQ2. Other pivots may behave differently, and the finding that Romance pivots preserve quality better than distant pivots may not generalize to other language families.

Finally, our low-resource language evaluation used discrete resource buckets (~20k, ~2M, ~20M+ speakers) rather than a continuous spectrum. This limits our ability to identify the precise resource threshold at which direct translation becomes unreliable, or to attribute quality differences to concrete factors like training data availability, typological distance, or script complexity.

8.2 Future Work

Several improvements would strengthen our findings. First, multi-model evaluation would test whether our conclusions hold across different LLM architectures (e.g., GPT-4, Claude, Llama, multilingual specialized models).

Second, exploring alternative RT metric definitions could identify formulations that complement COMETkiwi more effectively. Our multiple regression analysis showed minimal improvement when combining RT with COMETkiwi, but different RT aggregations (e.g. learned fusion, or weighted combinations as shown in Appendix B Tabe B8.T2) or different similarity metrics might yield more meaningful gains. Investigating which aspects of translation quality RT captures that COMETkiwi misses could inform better quality estimation systems.

Third, testing chain-of-thought style translation within a single context window would directly test the analogy to reasoning tasks. Instead of separate API calls for each hop, a prompt could instruct the model to "translate from English to Spanish via French, showing your intermediate translation." This would allow the model to leverage its reasoning capabilities and potentially correct errors in intermediate steps, potentially yielding different results than our independent-hop design.

Fourth, systematic exploration of language family effects could identify cases where pivot translation actually helps. Our Romance family findings suggest that linguistic similarity matters, but we did not test whether certain low-resource language pairs might benefit from routing through related high-resource languages. For example, translating to a low-resource Dravidian language via Tamil, or to a low-resource Bantu language via Swahili, might reveal scenarios where pivot translation provides genuine benefits.

Fifth, continuous resource level analysis would strengthen the validity of RQ2 findings. Rather than discrete buckets, testing across a spectrum of resource levels (e.g., 10k, 50k, 100k, 500k, 1M, 5M, 20M+ speakers) could allow identification of the precise threshold at which direct translation quality degrades, and enable attribution to concrete factors like training data availability, model coverage, or typological distance.

Sixth, domain-specific evaluation would test whether our findings generalize beyond news/encyclopedic text. Technical documentation, conversational dialogue, literary text, or code comments may exhibit different patterns, particularly if certain domains benefit more or less from the lexical priming effects we observed with Romance pivots.

Seventh, investigating alternative embedding models for RT computation could address anisotropy concerns observed for LaBSE.

Finally, human evaluation would provide ground truth validation. While COMET is trained on human judgments, direct human assessment of translation quality would allow us to test whether RT metrics correlate with human-perceived quality, and whether the "smoke detector" pattern holds when quality is defined by human raters rather than automatic metrics.

9 Conclusions

This study investigated pivot-based translation strategies and round-trip consistency as a quality predictor for LLM-based machine translation. Our experiments yield four main conclusions.

First, pivot translation does not help for LLM-based MT. Unlike traditional SMT/NMT systems where pivoting through high-resource languages could improve low-resource translation, modern LLMs perform best with direct translation.

Second, chain length matters, but pivot selection matters more. A long chain through linguistically similar languages (Romance family) degrades less than a short chain through distant languages (Chinese).

Third, language resource level remains a strong predictor of quality. Even with powerful LLMs, translation to truly low-resource languages (~20k speakers) remains challenging.

Fourth, RT metrics function as effective "smoke detectors." They can reliably distinguish catastrophic translation failures (AUC = 0.94) but cannot rank good translations against excellent ones. COMET, chrF, and BLEU show consistent trends throughout our experiments, with all metrics generally agreeing on the relative ranking of methods, although the limitations of BLEU for languages like Japanese have been noted and BLEU avoided for such lagnuages as in RQ3. It's also of note that COMET shows smaller degradation for pivot-based approaches due to semantic similarity being preserved even when surface forms differ.

10 Acknowledgements

We would like to thank Professor Rick Stevens, The University of Chicago Computer Science deparment, and Argonne ALCF for providing access to the OSS 120b model used in this study, and for providing the opportunity for this study to be conducted.

11 Data Availability

Experimental results for RQ1 and RQ2 are stored in `data/experiments/`:

- `run1_chain_length_results.csv` – Raw translations for chain length experiment
- `run1_chain_length_comet_scores.csv` – COMET scores for chain length
- `run1_low_resource_bucket_results.csv` – Raw translations for low-resource experiment
- `run1_low_resource_comet_scores.csv` – COMET scores for low-resource
- `run1_pivot_family_results.csv` – Raw translations for pivot family experiment

Experimental results for RQ3 are displayed in Appendix B.

12 References

References

[1] Paul, M., Finch, A., & Sumita, E. (2013). How to choose the best pivot language for the automatic translation of low-resource languages. ACM Transactions on Asian Language Information Processing, 12(4), Article 14. https://doi.org/10.1145/2539995

[2] Rei, R., Treviso, M., Guerreiro, N. M., Zerva, C., Farinha, A. C., Maroti, C., de Souza, J. G. C., Glushkova, T., Alves, D., Coheur, L., Lavie, A., & Martins, A. F. T. (2022). CometKiwi: IST-Unbabel 2022 submission for the quality estimation shared task. Proceedings of the Seventh Conference on Machine Translation (WMT), 634–645. https://aclanthology.org/2022.wmt-1.60

[3] Moon, J., Cho, H., & Park, E. L. (2020). Revisiting round-trip translation for quality estimation. Proceedings of the 22nd Annual Conference of the European Association for Machine Translation (EAMT), 91–104. https://aclanthology.org/2020.eamt-1.10

[4] Cheng, Y., Yang, Q., Liu, Y., Sun, M., & Xu, W. (2017). Joint training for pivot-based neural machine translation. Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI 2017), 3974–3980. https://doi.org/10.24963/ijcai.2017/555

[5] Kim, Y., Petrov, P., Petrushkov, P., Khadivi, S., & Ney, H. (2019). Pivot-based transfer learning for neural machine translation between Non-English languages. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 866–876. https://doi.org/10.18653/v1/D19-1080

[6] Sulistyo, D. A., Wibawa, A. P., Prasetya, D. D., & Ahda, F. A. (2025). An enhanced pivot-based neural machine translation for low-resource languages. International Journal of Advances in Intelligent Informatics, 11(2), 258–274. https://doi.org/10.26555/ijain.v11i2.2115

[7] Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2020). COMET: A neural framework for MT evaluation. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2685–2702. https://doi.org/10.18653/v1/2020.emnlp-main.213

[8] Feng, F., Yang, Y., Cer, D., Arivazhagan, N., & Wang, W. (2022). Language-agnostic BERT sentence embedding. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 878–891. https://doi.org/10.18653/v1/2022.acl-long.62

[9] Wang, Y., Wu, A., & Neubig, G. (2022). English contrastive learning can learn universal cross-lingual sentence embeddings. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 9122–9133. https://doi.org/10.18653/v1/2022.emnlp-main.621

[10] Goyal, N., Gao, C., Chaudhary, V., Chen, P.-J., Wenzek, G., Ju, D., Krishnan, S., Ranzato, M., Guzmán, F., & Fan, A. (2022). The FLORES-101 evaluation benchmark for low-resource and multilingual machine translation. Transactions of the Association for Computational Linguistics, 10, 522–538. https://doi.org/10.1162/tacl_a_00474

[11] Popović, M. (2015). chrF: character n-gram F-score for automatic MT evaluation. Proceedings of the Tenth Workshop on Statistical Machine Translation, 392–395. https://doi.org/10.18653/v1/W15-3049

APPENDIX A: Calculations, Context and Code Samples

A.1.1 COMET Quality Score (quality_comet)

Model: Unbabel/wmt22-comet-da
Type: Reference-based neural quality estimation
Range: 0.0 to 1.0 (higher = better)
Calculation:

```
from comet import download_model, load_from_checkpoint

model = load_from_checkpoint(download_model("Unbabel/wmt22-comet-da"))
data = [{
    "src": source_text,   # English source
    "mt": output_text,    # System translation
    "ref": target_reference  # Human reference
}]
output = model.predict(data, batch_size=32)
quality_comet = output.scores[0]
```

Remark: COMET is trained on human quality judgments and correlates highly with human evaluation. Scores below 0.85 typically indicate noticeable quality issues; scores below 0.80 indicate significant problems

A.1.2 chrF Quality Score (quality_chrf)

Type: Character n-gram F-score
Range: 0 to 100 (higher = better)
Calculation:

$$\text{chrF} = \left(1 + \beta^2\right) \cdot \frac{\text{chrP} \cdot \text{chrR}}{\beta^2 \cdot \text{chrP} + \text{chrR}} \tag{1}$$

```
from sacrebleu import sentence_chrf
quality_chrf = sentence_chrf(output_text, [target_reference]).score
```

Remark: chrF was used rather than BLEU because BLEU completely fails for languages like JPN.

A.2.1 chrF Surface RT predictive metrics

Remark: Same scoring process as A.1.2 chrF Quality Score, but now instead comparing e.g. source and source' for rt_source. This generates **surface** RT metrics that are those without the 'semantic' term in their names. Semantic RT metrics on the other hand have the term 'semantic' in their names, with the exception of combined_rt_direct_geometric, which is defined to use semantic RT metrics, and instead combined_rt_direct_geometric (surface) is defined to use surface RT metrics.

A.2.2 LaBSE Semantic RT predictive metrics

Model: sentence-transformers/LaBSE
Type: Cross-lingual embedding cosine similarity
Range: 0.0 to 1.0 (higher = better)
Calculation:

```
from sentence_transformers import SentenceTransformer

model = SentenceTransformer('sentence-transformers/LaBSE')
embed_output = model.encode(output_text)
embed_reference = model.encode(target_reference)

quality_semantic = cosine_similarity(embed_output, embed_reference)
```

Remark: LaBSE and cosine similarity was used for generation of semantic RT predictive metrics because the COMET family doesn't have a lightweight model that supports same language comparisons. In fact when given the exact same text in the same language for src and mt, COMETkiwi fails to return 1.0, which suggests COMETkiwi possesses such a limitation due to how it was trained.

Remark: Using LaBSE instead of a COMET family model is non-ideal because LaBSE with cosine similarity may have the problem of anisotropy, where a small shift in LaBSE score could be a very big change in quality or vice versa. The prominence of this issue is discussed under the limitations section. Using LaBSE as the generator of predictive metrics is perhaps still somewhat tolerable, but it is not tolerable for quality metrics (which is why COMET was used).

A.2.3 COMETkiwi predictive metrics (direct_source_output_comet)

Model: Unbabel/wmt22-cometkiwi-da

Type: Reference-free neural quality estimation

Calculation:

```
from comet import download_model, load_from_checkpoint

model = load_from_checkpoint(download_model("Unbabel/wmt22-cometkiwi-da"))
data = [{
    "src": source_text,   # English source
    "mt": output_text     # System translation (no reference needed)
}]
output = model.predict(data, batch_size=32)
direct_source_output_comet = output.scores[0]
```

Remark: Here we use COMETkiwi to create the non-RT

A.3.1 Pearson Correlation Analysis

Purpose: Measure linear relationship between predictor and quality metrics.

Definition: Pearson correlation coefficient (r) measures the strength and direction of linear association.

Calculation:

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} \tag{2}$$

```
from scipy import stats
r, p_value = stats.pearsonr(df[predictor], df[quality_metric])
```

A.3.2 ROC AUC Analysis

Purpose: Measure how well a predictor metric can distinguish translation failures from successes.

Definition: ROC AUC (Receiver Operating Characteristic - Area Under Curve) quantifies the probability that a randomly chosen failure has a lower predictor score than a randomly chosen success.

Calculation:

```
from sklearn.metrics import roc_auc_score

# Define failure as quality below threshold
labels = (df[quality_metric] < threshold).astype(int)  # 1 = failure, 0 = success

# RT scores (negate because higher RT should predict success)
# We want: high RT -> low failure probability
rt_scores = -df[rt_metric].values

auc = roc_auc_score(labels, rt_scores)
```

Interpretation:

- AUC = 1.0: Perfect discrimination (all failures have lower RT than all successes)
- AUC = 0.5: Random chance (RT provides no information)
- AUC > 0.7: Moderate discriminative power
- AUC > 0.8: Good discriminative power
- AUC > 0.9: Excellent discriminative power

Thresholds used:

- quality_comet: 0.80, 0.85, 0.90
- quality_chrf: 50, 55, 60

Bootstrap Confidence Intervals Calculation:

```
def bootstrap_auc_ci(labels, scores, n_bootstrap=1000, alpha=0.05):
    aucs = []
    for _ in range(n_bootstrap):
        indices = np.random.choice(len(labels), size=len(labels), replace=True)
        boot_labels = labels[indices]
        boot_scores = scores[indices]
        if len(np.unique(boot_labels)) == 2:  # Need both classes
            aucs.append(roc_auc_score(boot_labels, boot_scores))
    return np.percentile(aucs, [100*alpha/2, 100*(1-alpha/2)])
```

A.3.3 Stratified Correlation Analysis

Purpose: Test whether the predictor works equally well across quality levels.

Hypothesis: The "smoke detector" hypothesis states that RT is better at detecting failures (low quality) than distinguishing good from excellent (high quality).

Method:

(1) Split data at median quality into two strata
(2) Compute correlation separately in each stratum
(3) Calculate $\Delta = r_{\text{low}} - r_{\text{high}}$

Calculation:

```
median_quality = df[quality_metric].median()

# Low-quality stratum (bottom 50%)
low_df = df[df[quality_metric] <= median_quality]
r_low = stats.pearsonr(low_df[predictor], low_df[quality_metric])[0]

# High-quality stratum (top 50%)
high_df = df[df[quality_metric] > median_quality]
r_high = stats.pearsonr(high_df[predictor], high_df[quality_metric])[0]

delta = r_low - r_high  # Smoke detector strength
```

Interpretation:

- $\Delta > 0$: Predictor is more useful in low-quality stratum (smoke detector pattern confirmed)
- $\Delta \approx 0$: Predictor works equally well at all quality levels
- $\Delta < 0$: Predictor is more useful in high-quality stratum (unexpected)

A.3.4 Granularity Analysis

Purpose: Test whether the predictor can rank translations within quality bands, not just separate good from bad.

Method:

(1) Split data into quartiles by quality, since n = 400 per target language, we have n = 100 per quartile
(2) Compute Spearman rank correlation within each quartile
(3) Compare correlation across quartiles

Calculation:

```
# Create quartile labels
df['quartile'] = pd.qcut(df[quality_metric], 4, labels=['Q1', 'Q2', 'Q3', 'Q4'])

results = {}
for q in ['Q1', 'Q2', 'Q3', 'Q4']:
    q_df = df[df['quartile'] == q]
```

```
    rho, p = stats.spearmanr(q_df[predictor], q_df[quality_metric])
    results[q] = {'spearman_rho': rho, 'n': len(q_df)}
```

Interpretation:

- High ρ in Q1 (low quality): Predictor can rank bad translations
- Low ρ in Q2-Q3 (mid quality): Predictor cannot make fine distinctions
- This pattern confirms RT as a coarse classifier, not a fine-grained scorer

A.3.5 Per-Pivot Decomposition

Purpose: Analyze how predictor performance varies across pivot languages.

Method: Run all analyses (ROC, correlation, stratified correlation) separately for each pivot language, then compare.

Metrics compared (ranked) across pivots:

(1) ROC AUC at threshold 0.85
(2) Pearson correlation with quality
(3) Stratified Correlation Smoke detector Δ
(4) Mean quality achieved

A.3.6 Multiple Regression Analysis

Purpose: Test whether RT adds predictive value beyond COMET-QE.

Interpretation: %score for improvement, small % score indicates that RT adds minimal value.

Calculation:

```
from sklearn.linear_model import LinearRegression

# COMET-QE alone
X_comet = df['direct_source_output_comet'].values.reshape(-1, 1)
model_comet = LinearRegression().fit(X_comet, df['quality_comet'])
r2_comet = model_comet.score(X_comet, df['quality_comet'])

# COMET-QE + RT combined
X_both = df[['direct_source_output_comet', 'rt_geometric_semantic_st']].values
model_both = LinearRegression().fit(X_both, df['quality_comet'])
r2_both = model_both.score(X_both, df['quality_comet'])

improvement = r2_both - r2_comet  # Additional variance explained by RT
optimal_weights = model_both.coef_  # [weight_comet, weight_rt]
```

APPENDIX B: Complete Data Tables

B.1 Quality Metrics Summary (Aggregated)

Table B1.T1. All relevant quality metrics used.

Metric	TUR Mean	TUR Std	JPN Mean	JPN Std	NPI Mean	NPI Std
quality_comet	0.8888	0.0547	0.9077	0.0415	0.8083	0.0676
quality_chrf	55.14	11.39	38.76	12.80	43.72	11.52
quality_semantic	0.9305	0.0507	0.9083	0.0601	0.8910	0.0577

Table B1.T2. Failure rates by language.

Threshold	TUR n	TUR %	JPN n	JPN %	NPI n	NPI %
quality_comet < 0.80	21	5.3%	11	2.8%	151	37.8%
quality_comet < 0.85	73	18.3%	34	8.5%	275	68.8%
quality_comet < 0.90	198	49.5%	122	30.5%	386	96.5%
quality_chrf < 60	264	66.0%	384	96.0%	365	91.2%
quality_chrf < 55	189	47.2%	349	87.2%	330	82.5%
quality_chrf < 50	130	32.5%	327	81.8%	292	73.0%

B.2 Predictive Metrics Summary (Aggregated)

Table B2.T1. chrF-based surface RT metrics, scaled 0-100.

Metric	TUR Mean	TUR Std	JPN Mean	JPN Std	NPI Mean	NPI Std
rt_source	68.76	13.14	68.91	13.55	68.76	12.83
rt_hop2	69.40	15.47	65.87	14.24	69.56	15.03
rt_output	82.72	10.77	70.98	15.60	74.60	12.53
rt_min_st	62.22	13.55	57.71	12.49	60.60	12.16
rt_geometric_st	72.65	10.12	67.72	10.43	70.15	9.62

Table B2.T2. LaBSE-based semantic RT metrics, scaled 0-1.

Metric	TUR Mean	TUR Std	JPN Mean	JPN Std	NPI Mean	NPI Std
rt_source_semantic	0.9589	0.0444	0.9604	0.0406	0.9590	0.0444
rt_hop2_semantic	0.9673	0.0298	0.9552	0.0402	0.9639	0.0366
rt_output_semantic	0.9804	0.0230	0.9732	0.0267	0.9676	0.0294
rt_min_semantic_st	0.9468	0.0450	0.9363	0.0457	0.9396	0.0462
rt_geometric_semantic_st	0.9685	0.0253	0.9625	0.0262	0.9631	0.0269

Table B2.T3 Direct and combined metrics, scaled 0-1.

Metric	TUR Mean	TUR Std	JPN Mean	JPN Std	NPI Mean	NPI Std
direct_source_output_comet	0.8604	0.0442	0.8723	0.0294	0.8807	0.0406
combined_rt_direct_geometric	0.9126	0.0311	0.9162	0.0235	0.9208	0.0303
combined_rt_direct_geometric_surface*	0.7889	0.0673	0.7664	0.0659	0.7844	0.0657

Remark: combined_rt_direct_geometric_surface = sqrt(direct_source_output_comet × (rt_geometric_st / 100)), where rt_geometric_st is the surface-level round-trip metric (chrF-based, scale 0-100).

B.3 Pearson Correlation Analysis

Table B3.T1. Correlation with quality_chrf, aggregated.

Predictor	TUR	JPN	NPI	Mean
rt_geometric_st	0.465	0.321	0.386	0.391
rt_geometric_semantic_st	0.481	0.341	0.311	0.378
direct_source_output_comet	0.338	0.400	0.308	0.349
combined_rt_direct_geometric	0.444	0.448	0.351	0.414
combined_rt_direct_geometric (surface)	0.494	0.377	0.405	0.425

Table B3.T2. Correlation with quality_comet, aggregated.

Predictor	TUR	JPN	NPI	Mean
rt_geometric_st (surface)	0.405	0.097	0.388	0.297
rt_geometric_semantic_st	0.413	0.313	0.377	0.368
direct_source_output_comet	0.744	0.625	0.681	0.683
combined_rt_direct_geometric	0.729	0.583	0.644	0.652
combined_rt_direct_geometric (surface)	0.565	0.217	0.512	0.432

B.4 ROC AUC Analysis

Table B4.T1. ROC AUC for quality_chrf < 60, aggregated.

Predictor	TUR	JPN	NPI	Mean
rt_geometric_st (surface)	0.694	0.665	0.704	0.687
rt_geometric_semantic_st	0.687	0.594	0.639	0.640
direct_source_output_comet	0.610	0.667	0.658	0.645
combined_rt_direct_geometric	0.661	0.657	0.673	0.664
combined_rt_direct_geometric (surface)	0.688	0.684	0.719	0.697

Table B4.T2 ROC AUC for quality_chrf < 55, aggregated.

Predictor	TUR	JPN	NPI	Mean
rt_geometric_st (surface)	0.729	0.646	0.691	0.688
rt_geometric_semantic_st	0.724	0.658	0.673	0.685
direct_source_output_comet	0.647	0.703	0.663	0.671
combined_rt_direct_geometric	0.705	0.720	0.693	0.706
combined_rt_direct_geometric (surface)	0.734	0.671	0.708	0.704

Table B4.T3. ROC AUC for quality_chrf < 50, aggregated.

Predictor	TUR	JPN	NPI	Mean
rt_geometric_st (surface)	0.704	0.590	0.692	0.662
rt_geometric_semantic_st	0.724	0.619	0.643	0.662
direct_source_output_comet	0.645	0.650	0.654	0.649
combined_rt_direct_geometric	0.704	0.667	0.678	0.683
combined_rt_direct_geometric (surface)	0.710	0.613	0.709	0.677

Remark: Notice how ROC AUC doesn't follow the expected trend when using the chrF quality metric. Note that this doesn't invalidate the smoke detector hypothesis, rather it may be either a symptom of the extremely high failure rate at these chosen values of chrF (which is a limitation of the study) or it might just show that chrF is not a meaningful quality metric for the purpose of this study.

Table B4.T4 ROC AUC for quality_comet < 0.90, aggregated.

Predictor	TUR	JPN	NPI	Mean
rt_geometric_st (surface)	0.647	0.577	0.676	0.633
rt_geometric_semantic_st	0.672	0.635	0.818	0.709
direct_source_output_comet	0.859	0.813	0.909	0.860
combined_rt_direct_geometric	0.841	0.772	0.912	0.841
combined_rt_direct_geometric (surface)	0.713	0.627	0.726	0.689

Table B4.T5. ROC AUC for quality_comet < 0.85, aggregated.

Predictor	TUR	JPN	NPI	Mean
rt_geometric_st (surface)	0.705	0.582	0.669	0.652
rt_geometric_semantic_st	0.698	0.701	0.657	0.685
direct_source_output_comet	0.885	0.814	0.774	0.824
combined_rt_direct_geometric	0.863	0.802	0.754	0.806
combined_rt_direct_geometric (surface)	0.784	0.643	0.709	0.712

Table B4.T6. ROC AUC for quality_comet < 0.80, aggregated.

Predictor	TUR	JPN	NPI	Mean
rt_geometric_st (surface)	0.767	0.487	0.666	0.640
rt_geometric_semantic_st	0.811	0.738	0.678	0.742
direct_source_output_comet	0.905	0.863	0.826	0.865
combined_rt_direct_geometric	0.902	0.839	0.805	0.849
combined_rt_direct_geometric (surface)	0.840	0.597	0.721	0.719

B.5 Stratified Correlations

Table B5.T1. Using rt_geometric_st, stratifying quality_chrf

Language	r_low	r_high	Δ (r_low - r_high)
TUR	0.363	0.211	+0.152
JPN	0.360	0.075	+0.286
NPI	0.171	0.239	-0.069
Mean	0.298	0.175	+0.123

Remark: This provides significant evidence that chrF is not a good quality metric for this study, because it differs significantly from the results using semantic quality (COMET) shown below in Table B15.

Table B5.T2. Using rt_geometric_semantic_st, stratifying quality_comet

Language	Low Stratum r	High Stratum r	Δ (Low - High)
TUR	0.369	0.004	**+0.364**
JPN	0.304	0.041	**+0.264**
NPI	0.390	0.134	**+0.256**
Mean	**0.354**	**0.060**	**+0.295**

Table B5.T3 Using direct_source_output_comet, stratifying quality_comet

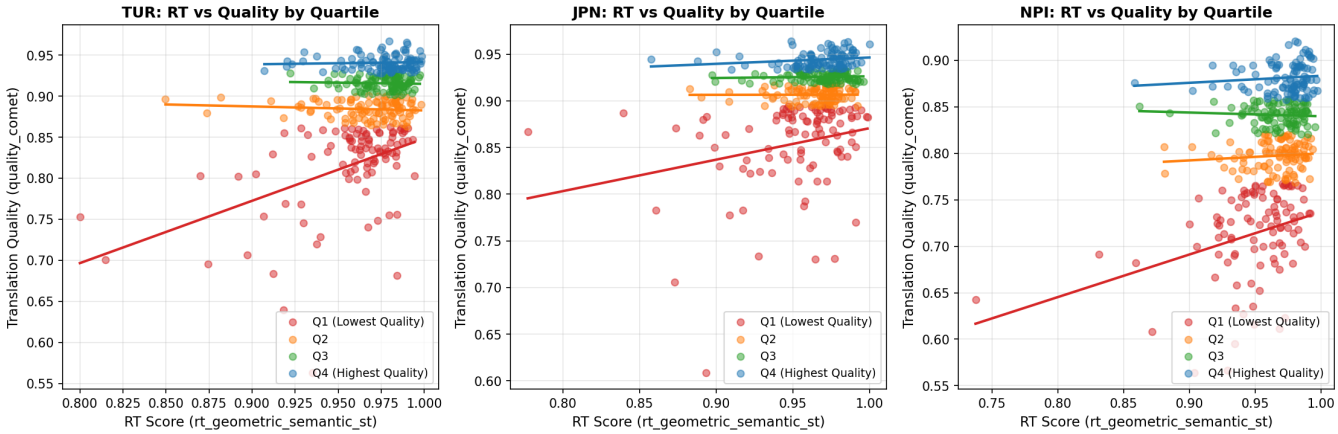
Language	Low Stratum r	High Stratum r	Δ (Low - High)
TUR	0.671	0.367	+0.303
JPN	0.472	0.351	+0.121
NPI	0.595	0.346	+0.249
Mean	0.579	0.355	+0.224

B.6 Granularity Analysis

Table B6.T1. Using rt_geometric_semantic_st, quartiles for quality_comet

Quartile	TUR	JPN	NPI	Mean	Interpretation
Q1 (lowest)	0.347	0.250	0.320	**0.306**	Moderate ranking
Q2	-0.037	0.039	0.100	0.034	No ranking
Q3	-0.017	0.053	-0.055	-0.006	No ranking
Q4 (highest)	0.054	0.286	0.130	0.157	Weak ranking

Plot B6.P1. Scatter plots of translation quality against RT score



Scatter plots showing RT vs Quality by Quartile for TUR, JPN, and NPI languages. Q1 (Lowest Quality) shown in red, Q2 in orange, Q3 in green, and Q4 (Highest Quality) in blue.

B.7 Per-Pivot Analysis

Table B7.T1. quality_chrf scores by pivot

Pivot	TUR	JPN	NPI	Mean
fra	57.91	40.66	45.66	48.08
deu	56.51	39.76	43.92	46.73
rus	54.78	38.06	42.94	45.26
cmn	51.36	36.56	42.36	43.43

Table B7.T2. quality_comet scores by pivot

Pivot	TUR	JPN	NPI	Mean
fra	0.8930	0.9098	0.8147	0.873
deu	0.8878	0.9095	0.8108	0.869
rus	0.8903	0.9070	0.8041	0.867
cmn	0.8842	0.9046	0.8036	0.864

Table B7.T3. rt_geometric_semantic_st scores by pivot

Pivot	TUR	JPN	NPI
fra	0.9774	0.9665	0.9695
deu	0.9733	0.9648	0.9660
rus	0.9659	0.9602	0.9638
cmn	0.9574	0.9586	0.9533

Table B7.T4. direct_source_output_comet scores by pivot

Pivot	TUR	JPN	NPI
fra	0.8653	0.8751	0.8850
deu	0.8627	0.8719	0.8825
rus	0.8584	0.8719	0.8786
cmn	0.8553	0.8705	0.8768

Table B7.T5. Per-Pivot Pearson correlation, rt_geometric_semantic_st with quality_comet

Pivot	TUR	JPN	NPI	Mean
fra	0.342	0.458	0.469	0.423
deu	0.268	0.298	0.368	0.311
rus	0.438	0.385	0.332	0.385
cmn	0.575	0.160	0.382	0.372

Table B7.T6 Per-Pivot ROC AUC at comet_quality < 0.85, using rt_geometric_semantic_st

Pivot	TUR	JPN	NPI	Mean
fra	0.723	0.681	0.666	0.690
deu	0.583	0.724	0.704	0.670
rus	0.760	0.836	0.608	0.735
cmn	0.737	0.591	0.612	0.647

Table B7.T5. Ranking of pivot languages

Criterion	Best	2nd	3rd	Worst
ROC AUC (mean)	rus (0.735)	fra (0.690)	deu (0.670)	cmn (0.647)
Correlation (mean)	fra (0.423)	rus (0.385)	cmn (0.372)	deu (0.311)
Smoke Detector Δ	fra (0.364)	cmn (0.347)	deu (0.275)	rus (0.202)

B.8 Multiple Regression Analysis for investigating RT utility beyond COMETkiwi

Table B8.T1. Multiple regression for RT value beyond COMETkiwi

Language	COMETKiwi R^2	Combined R^2	Improvement	COMETkiwi Weight	RT Weight
TUR	0.554	0.563	+0.88%	0.698	0.105
JPN	0.391	0.398	+0.66%	0.592	0.088
NPI	0.464	0.456	+0.10%	0.662	0.037

Table B8.T2. Alternative combinations correlation with quality_comet

Strategy	TUR	JPN	NPI	Mean
COMET-QE alone	**0.744**	**0.625**	**0.681**	**0.683**
RT alone	0.413	0.313	0.377	0.368
Geometric mean	0.729	0.583	0.644	0.652
Arithmetic mean	0.724	0.575	0.639	0.646
Weighted (0.8/0.2)	**0.750**	**0.630**	0.679	**0.686**
Min (conservative)	0.744	0.622	**0.678**	0.681