LoRA Fine-Tuning of English–Norwegian NMT for the Oil & Gas Industry

Xiaojing Yang, Zhihan Li, Gege Sun, Mengyue Li and **Meriem Beloucif**
Department of Linguistics and Philology, Uppsala University
Uppsala, Sweden
xiaojing.yang.4987@student.uu.se, meriem.beloucif@lingfil.uu.se

Abstract

Adapting large language models to specialized domains remains challenging due to the computational cost of full fine-tuning and the limited availability of domain-specific parallel data. We present a systematic framework for parameter-efficient domain adaptation using Low-Rank Adaptation (LoRA) geared towards efficient learning in low-resource scenarios. Our method combines data-scaling analysis, dual-track hyperparameter optimization, and competitive benchmarking. We evaluate our approach on the low-resource English–Norwegian petroleum translation domain using a distilled version of NLLB and parallel data from the Norwegian Petroleum Directorate. Our adapted model achieves 61.48 BLEU (+24.62 over the base model) and 0.9298 COMET, while updating $<$0.4% of parameters. Our experiments show that the right parameter efficiency helps models achieve high accuracy, outperform evaluated commercial baselines on BLEU, and achieve comparable semantic quality (COMET). Our results provide a reproducible and computationally efficient blueprint for domain adaptation in neural machine translation, particularly for specialized and resource-constrained domains.

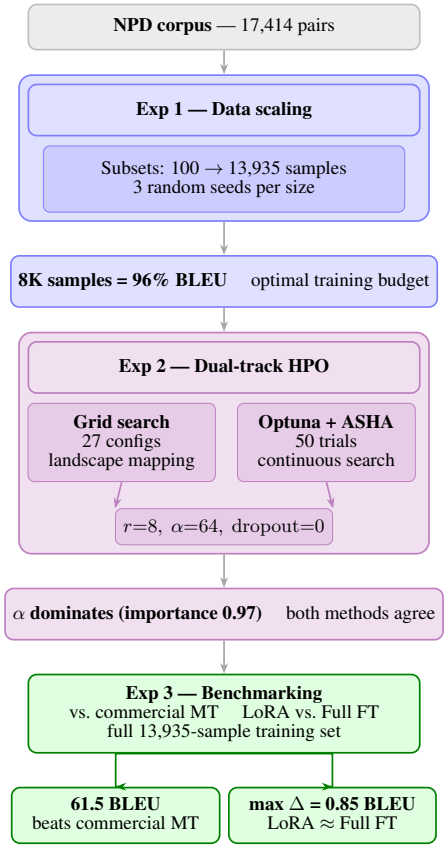

Figure 1: Three-stage experimental framework. Each stage feeds into the next; key findings are shown below each experiment.

1 Introduction

Large Language Models (LLMs) have transformed multilingual NLP, with models like NLLB showing strong general-domain performance. However, in specialized high-stakes domains, such as petroleum engineering, domain terminology is highly specialized, regulatory language is dense,

and error tolerance is low. Full fine-tuning addresses domain mismatch but its GPU cost and engineering overhead remain prohibitive for most practitioners. Parameter-Efficient Fine-Tuning (PEFT) (Houlsby et al., 2019; Li and Liang, 2021; Hu et al., 2022) offers a pragmatic alternative. Low-Rank Adaptation (LoRA) (Hu et al., 2022) freezes pre-trained weights and injects trainable low-rank adapters into Transformer layers, enabling adaptation with minimal parameters. However, LoRA effectiveness depends critically on data volume and hyperparameters (rank r, scaling α). Ad-hoc choices waste compute and yield suboptimal models, yet systematic methodologies balancing scientific rigor with resource constraints remain scarce (He et al., 2022; Ranathunga et al., 2023).

We address this gap with a three-stage framework for efficient LoRA-based domain adaptation (Figure 1): **(1)** data-scaling analysis to identify efficient training budgets, **(2)** dual-method hyperparameter search combining grid search with Optuna-ASHA optimization[1], and **(3)** benchmarking against commercial and open-source baselines. We instantiate this framework on petroleum-domain English→Norwegian Bokmål translation using `facebook/nllb-200-distilled-600M` (NLLB Team et al., 2022) and the Norwegian Petroleum Directorate (NPD) translation memory released via ELRC as **ELRC_559** (European Language Resource Coordination (ELRC), 2017), which we curate and filter into 17,414 pairs. Our contributions include: **(i)** demonstrating that 8,000 training samples can recover 96% of the maximum translation performance achievable with the full dataset, **(ii)** identifying optimal hyperparameter configurations through a dual-track search approach, and **(iii)** achieving a BLEU score of 61.48 while outperforming commercial translation systems on lexical metrics, with only less than 0.4% of model parameters updated. Our approach shows that a high-quality specialized neural machine translation system can be realized under specific resource constraints.

[1]ASHA (Asynchronous Successive Halving Algorithm, (Li et al., 2018)) prunes underperforming trials early based on intermediate validation scores, reallocating compute to more promising configurations.

2 Related Work

Parameter-Efficient Fine-Tuning. PEFT methods enable adaptation with minimal trainable parameters. Adapters (Houlsby et al., 2019) insert bottleneck layers into frozen Transformers, Prefix-Tuning (Li and Liang, 2021) optimizes continuous prompts (≈0.1% parameters), and LoRA (Hu et al., 2022) injects low-rank matrices into attention layers, often matching full fine-tuning performance (He et al., 2022; Ranathunga et al., 2023). These methods make domain adaptation feasible under limited compute, but prior work largely evaluates single hand-picked configurations. *In contrast, we systematically map the LoRA hyperparameter landscape using two independent search methods, providing evidence for which parameters matter and why. We focus on LoRA as a representative PEFT method; empirical comparison with Adapters and Prefix-Tuning is left for future work.*

Low-Resource and Domain-Specific MT. Low-resource MT typically relies on: (i) massive multilingual models for cross-lingual transfer (NLLB Team et al., 2022), (ii) data augmentation such as back-translation (Sennrich et al., 2016), and (iii) domain-specific fine-tuning. Recent surveys (Ranathunga et al., 2023) highlight challenges in data quality, evaluation metrics, and transfer learning. While domain-specific fine-tuning has been explored in specialized fields, systematic frameworks for industrial sectors such as oil & gas remain underexplored. *Our work fills this gap: we provide a reproducible, three-stage methodology for petroleum-domain EN–NO translation that integrates data-scaling analysis with principled hyperparameter search, and validate the resulting system against commercial production MT.*

LoRA for Domain-Specific MT. Most closely related to our work, (Carrión and Casacuberta, 2025) apply LoRA to dedicated NMT architectures for multi-domain adaptation, demonstrating parameter efficiency on par with full fine-tuning. However, their focus is on continual learning and catastrophic forgetting mitigation, and hyperparameters are set heuristically. Similarly, (Zheng et al., 2024) fine-tune LLMs with LoRA for domain-specific MT but do not conduct systematic hyperparameter analysis. In contrast, our work provides a reproducible three-stage framework that explicitly maps the LoRA hyperparameter land-

scape through dual-track search and quantifies data efficiency thresholds.

3 Data

3.1 Source & Licensing

We use the Norwegian Petroleum Directorate (NPD) translation memory, released on ELRC as **ELRC_559** (`npd.no.en-nb.tmx`) (European Language Resource Coordination (ELRC), 2017), containing ≈26,000 EN–NB sentence pairs from official sources (last updated 22 Dec 2017). Content spans drilling/production reports, license announcements, ownership info, and regulatory pages. The corpus is used for research; only code and derived statistics/plots are shared.

3.2 Processing Pipeline

To ensure high-quality data, we *validate, clean, and split* the corpus:

(A) Corpus Diagnostics (Raw). Prior to preprocessing, we conduct corpus-level diagnostics on the raw TMX data to evaluate sentence alignment quality. We analyze source and target token-length distributions as well as EN–NO length correlation. As shown in Figure 2, sentence lengths exhibit a strong linear correlation with a sharply peaked length-ratio distribution around 0.87 (NO/EN), indicating well-aligned sentence pairs with minimal noise.

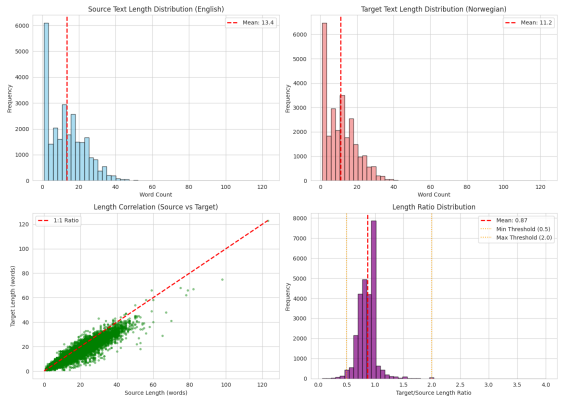

Figure 2: Corpus diagnostics on raw EN–NO pairs: token-length distributions and source–target length correlation. The tight diagonal structure confirms strong sentence-level alignment.

(B) Corpus Quality (OQS 2.0). To quantitatively validate the raw corpus, we compute the *Open Quality Score (OQS 2.0)*, a composite metric aggregating six dimensions of parallel-data quality (Table 1). Semantic alignment is measured using LASER embeddings (Artetxe and Schwenk, 2019); the remaining dimensions are defined in Appendix B. Each component reflects widely used criteria in parallel corpus filtering, providing a practical proxy for overall data quality. The "2.0" designation reflects the version formulated for this study; the equal-weight arithmetic mean is a simplifying assumption, and correlation with human quality judgments is left for future validation.

Dimension	Score
Semantic alignment (LASER)	0.883
Completeness	1.000
Source uniqueness	0.997
Target uniqueness	0.994
Pair uniqueness	0.998
Domain relevance (term coverage)	0.983
Overall OQS	**0.976**

Table 1: OQS 2.0 sub-scores characterizing the raw corpus quality.

(C) Cleaning & De-duplication. We perform minimal preprocessing: punctuation and whitespace normalization, removal of very short or non-linguistic segments ($\leq$2 tokens), and exact de-duplication at the source, target, and pair levels. The high deduplication rate (30.9%) reflects extensive template repetition inherent in regulatory institutional corpora, such as section headings and standardised well-report boilerplate phrases. A small proportion of target sentences contain Nynorsk morphological markers (e.g., *ikkje*, *vart*); Nynorsk is one of the two official written forms of Norwegian alongside Bokmål, and these are retained as they reflect the real-world distribution of NPD documents.

The resulting corpus contains **17,414** EN–NO sentence pairs, randomly split into 13,935 / 1,737 / 1,742 (train/dev/test, 80/10/10) using a fixed random seed (42). Overlap detection confirms zero exact source-sentence overlap across splits.

3.3 Data Limitations

The corpus was last updated in 2017, so recent petroleum terminology may be underrepresented. The dataset exhibits a high template repetition rate (30.9%), meaning that near-duplicate expressions may exist despite exact sentence-level deduplication. As a result, n-gram-level overlap cannot be fully excluded, and BLEU scores may be partially inflated.

Importantly, all systems are evaluated on the same held-out test set, ensuring that relative comparisons remain fair despite potential absolute

score inflation.

Future work should further investigate overlap at the n-gram level and evaluate performance on out-of-domain petroleum texts to better assess generalisability. No personal data is included in the corpus; only code and aggregated statistics are released.

4 Baselines and Pretrained Models

To contextualize the performance of our LoRA-adapted system, we evaluate both commercial MT services and open-source pretrained models on the NPD EN→NO test set (1,742 segments). All NLLB-based systems are decoded with beam size 5 and length penalty 1.0. We evaluate using three automatic metrics: BLEU (Papineni et al., 2002), chrF++ (Popović, 2015), and COMET (`Unbabel/wmt22-comet-da`; (Rei et al., 2022)). COMET is computed independently of decoding on the generated translations.

4.1 Commercial Baselines

We benchmark four widely used commercial systems: Microsoft Translator, DeepL API, Google Cloud Translation, and OpenAI GPT-4o-mini in a zero-shot translation setting. All systems are evaluated using default configurations without domain-specific glossaries. Note that our model is explicitly fine-tuned on in-domain NPD data, whereas commercial systems are general-purpose and trained on substantially larger and broader corpora; this comparison is therefore indicative of practical utility rather than a controlled evaluation under equivalent training conditions. Detailed system descriptions are provided in Appendix C.

Model	BLEU	chrF++	COMET
Microsoft Translator	**57.88**	**75.45**	**0.9313**
DeepL API	54.53	74.29	0.9310
Google Translate	53.50	72.75	0.9288
GPT-4o-mini (zero-shot)	43.23	64.93	0.9115

Table 2: Commercial MT baselines on the NPD EN→NO test set.

Among commercial systems, Microsoft Translator achieves the highest performance across all metrics, while DeepL and Google Translate perform comparably with slightly lower BLEU and chrF++ scores. GPT-4o-mini shows substantially weaker performance in the zero-shot setting, highlighting the limitations of general-purpose models without domain adaptation.

4.2 Open-source Pretrained Models

We consider a range of multilingual pretrained MT models, including NLLB-200, M2M100, and OPUS-MT variants. An initial screening on 100 samples was conducted to rank models; full results are reported in Appendix C. Based on this screening, we select the top three models for full evaluation on the complete test set.

Model	BLEU	chrF++	COMET
NLLB-200-1.3B	39.38	63.60	0.8979
M2M100-1.2B	38.33	62.74	0.8896
NLLB-200-distilled-600M	36.86	61.29	0.8814

Table 3: Open-source pretrained baselines evaluated on the full test set.

We adopt **NLLB-200-distilled-600M** as the base model for LoRA adaptation for three reasons: (i) competitive semantic adequacy despite its smaller size (COMET = 0.8814), (ii) efficient resource usage ($<$ 16 GB VRAM), enabling extensive hyperparameter exploration, and (iii) sufficient headroom for improvement, as indicated by its lower zero-shot performance.

5 Experimental Setup

Following the three-stage framework outlined in Section 1, all experiments use `facebook/nllb-200-distilled-600M` (NLLB Team et al., 2022) with LoRA (Hu et al., 2022) applied to all attention projection layers (Q, K, V, O). Training is implemented in PyTorch with Hugging Face Transformers (Wolf et al., 2020) and the PEFT library (Mangrulkar et al., 2022), using FP16 precision on NVIDIA T4 GPUs. All experiments are tracked via Weights & Biases (Biewald, 2020).

5.1 Experiment 1: Data Scaling

This experiment investigates how translation performance scales with training data size, aiming to identify a data-efficient subset before conducting costly hyperparameter searches.

5.1.1 Dataset and Sampling

We use the English (eng_Latn) to Norwegian Bokmål (nob_Latn) parallel corpus described in Section 3. The training split contains 13,935 sentence pairs, from which subsets of varying sizes (100, 500, 1,000, 2,000, 4,000, 6,000, 8,000, 10,000, and the full set) are randomly sampled without replacement. The validation and test splits

contain 1,737 and 1,742 instances respectively, with the test set strictly held out for final evaluation in Experiment 3. All sequences are tokenized with a maximum length of 128 tokens.

To ensure robustness and account for variability from random initialization and data shuffling, we conduct three independent runs per training subset, using different random seeds (42, 123, 456). Results are reported as the mean and standard deviation across these three runs, providing reliable estimates of model performance and variance.

5.1.2 Training Configuration

Key hyperparameters are summarized in Table 4. All models are trained for 3 epochs with gradient accumulation yielding an effective batch size of 16. Model performance is evaluated on the validation set using BLEU, chrF++, and cross-entropy loss.

Hyperparameter	Setting
LoRA rank (r)	16
LoRA scaling (α)	32
LoRA dropout	0.1
Training epochs	3
Batch size ($\times$ accumulation)	16 (4)
Learning rate	5×10^{-4}
Weight decay	0.01
Precision	FP16
Evaluation interval	every 200 steps

Table 4: Key training hyperparameters used in Experiment 1.

5.1.3 Evaluation

Model performance is evaluated on the validation set using BLEU (Papineni et al., 2002) and chrF (Popović, 2015). Cross-entropy loss is monitored to ensure training convergence. Learning curves and data efficiency plots are generated to visualize performance gains as training data size increases.

5.2 Experiment 2: Hyperparameter Optimization

Building on the data scaling results, we systematically investigate the impact of LoRA rank (r), scaling factor (α), and dropout on translation quality using the 8,000-sample subset identified in Experiment 1.

We explore the hyperparameter space using two complementary strategies: grid search for systematic coverage and Optuna with ASHA pruning (Akiba et al., 2019; Li et al., 2018) for efficient convergence.

Table 5 summarizes the search spaces and fixed parameters shared across both methods.

Hyperparameter	Grid	Optuna
LoRA rank (r)	8,16,32	{8,16,32} (discrete)
LoRA scaling (α)	16,32,64	[16,64] (log-uniform)
LoRA dropout	0.0,0.1,0.2	[0.0,0.2] (uniform)
Learning rate	5e-5,1e-4,5e-4	[5e-5,5e-4] (log-uniform)
Batch size	4,8,16	{4,8,16} (discrete)
Fixed (both methods):		
Grad. accumulation	4	
Training epochs	3	
Weight decay	0.01	
LoRA layers	Q,K,V,O	

Table 5: Hyperparameter search spaces for grid search (27 configs) and Optuna-ASHA (50 trials). Fixed parameters are shared.

5.2.1 Grid Search (2a)

A systematic grid search covers 27 discrete configurations, formed by the full Cartesian product of three ranks ($r \in \{8, 16, 32\}$), three scaling factors ($\alpha \in \{16, 32, 64\}$), and three dropout rates ($p \in \{0.0, 0.1, 0.2\}$). Each configuration is run once on the 8,000-sample subset due to computational constraints, as this stage is intended for coarse-grained landscape mapping rather than precise ranking of configurations. The scaling factor α controls the magnitude of the LoRA update relative to the pre-trained weights: the effective update is scaled by $\frac{\alpha}{r}$, meaning higher α amplifies the adapter's influence on the model output. A larger $\frac{\alpha}{r}$ ratio thus allows the adapter to make more aggressive updates during fine-tuning, which is particularly beneficial when adapting to a domain with substantial lexical shift, as in petroleum-domain translation.

5.2.2 Optuna with ASHA Pruning (2b)

To cross-validate the grid search findings and explore the parameter space more efficiently, we conduct an independent search using Optuna (Akiba et al., 2019) with ASHA pruning (Li et al., 2018). Optuna offers three advantages over grid search in this setting: **(i)** it samples across continuous parameter ranges rather than fixed grid points, enabling finer-grained exploration; **(ii)** its Bayesian-inspired sampler guides subsequent trials toward promising regions based on prior results, improving sample efficiency; and **(iii)** ASHA pruning terminates underperforming trials early, reallocating compute to configurations with higher

potential. Crucially, by treating the two methods as independent searches converging on the same answer, we obtain stronger evidence that the identified optimum is robust rather than an artefact of one particular search strategy.

The search proceeds in two stages. **Stage 1** runs 50 trials on a 2,000-sample subset (33 GPU hours). Using a smaller subset at this stage allows rapid screening of the continuous parameter space while keeping compute tractable. ASHA pruning terminates roughly half of underperforming trials early based on intermediate validation BLEU, focusing resources on promising regions. We adopt a multi-objective criterion jointly maximising BLEU and chrF. Although these two metrics are highly correlated n-gram-based measures, retaining both objectives reduces over-indexing on a single metric and improves stability of the Pareto front under stochastic noise. We select the top three configurations based on their harmonic mean score.

Stage 2 retrains each of the three Pareto-optimal configurations three times on the full 8,000-sample subset using different random seeds (18 GPU hours). This step assesses stability and filters out configurations whose Stage 1 performance was inflated by favourable random initialisation. The final configuration is selected based on mean validation BLEU across the three seeds.

5.3 Experiment 3: Final Model Training

Using the optimal configuration from Experiment 2 ($r = 8, \alpha = 64$, dropout $= 0.0$), the final model is trained on the full 13,935-sample training set for 3 epochs with early stopping (patience=3 evaluation steps). Evaluation is performed on the 1,742-sample held-out test set using BLEU, chrF++, and COMET.

6 Results and Analysis

6.1 Experiment 1: Data Scaling

In the data scaling experiment, we investigated how translation performance scales with the amount of training data. Table 7 reports results for subsets ranging from 100 to 13,935 sentence pairs, while Figure 3 plots the corresponding learning curves. Our analysis reveals a distinct three-phase learning process.

Phase 1: High Efficiency in Low-Resource Regime. The model exhibits exceptional data efficiency at smaller scales. When training data increases from 100 to 2,000 samples, the mean BLEU score jumps dramatically from 8.74 to 53.21, whereas the validation loss decreases significantly from 4.903 to 0.861. This progress shows that the model rapidly acquires domain-specific knowledge. As shown in Figure 3, the marginal BLEU gain exceeds 87 per 1,000 samples in the 100-to-500 sample range.

Hyperparameter	**Value**
Optimized parameters (vs. Exp 1):	
LoRA rank (r)	**8** (was 16)
LoRA scaling (α)	**64** (was 32)
LoRA dropout	**0.0** (was 0.1)
Learning rate	5×10^{-4}
Batch size (per device)	4
Gradient accumulation	4
Effective batch size	16
Training epochs	3
Early stopping patience	3 evaluation steps
LoRA layers	Q, K, V, O
Precision	FP16
Evaluation interval	every 200 steps
Checkpoint interval	every 400 steps

Table 6: Final model configuration. Bold indicates optimized parameters from Experiment 2.

Phase 2: Diminishing Returns. Beyond 2,000 samples, marginal gains decrease substantially. After 4,000 samples, the BLEU gain per 1,000 samples falls below 5. Although the validation loss continues to decline, the rate of improvement slows, implying the model is shifting from learning fundamental patterns to refining more nuanced ones. Increasing the dataset from 10,000 to the full 13,935 pairs, for instance, yields only a 1.5 improvement in BLEU, highlighting a significant reduction in data efficiency.

Phase 3: Cost–Performance Trade-off. Training on the full dataset yields the highest BLEU of 61.62. We set a practical target of 95% of this value (58.54) to balance performance and computational cost. The 8,000-sample subset is the first to surpass this threshold, attaining over 96% of the maximum performance with only 57% of the total training data. This subset thus provides an effective balance between accuracy and computation and was selected for the subsequent hyperparameter analysis in Experiment 2.

6.2 Experiment 2: Parameter Sensitivity

Using the 8,000-sample subset identified in Experiment 1, we examined how LoRA hyperparameters affect translation quality via two complementary search methods: an exhaustive grid search for sys-

Sample Size	BLEU ± Std	chrF ± Std	Loss ± Std
100	8.74 (±0.01)	25.38 (±0.038)	4.903 (±0.0025)
500	43.37 (±0.61)	66.42 (±0.421)	1.209 (±0.0048)
1,000	49.60 (±0.60)	70.57 (±0.436)	1.003 (±0.0089)
2,000	53.21 (±0.15)	73.28 (±0.119)	0.861 (±0.0040)
4,000	54.40 (±0.46)	74.06 (±0.247)	0.814 (±0.0021)
6,000	57.18 (±0.18)	75.71 (±0.109)	0.741 (±0.0024)
8,000	59.13 (±0.37)	76.90 (±0.134)	0.702 (±0.0031)
10,000	60.11 (±0.36)	77.44 (±0.049)	0.675 (±0.0007)
13,935	61.62 (±0.00)	78.26 (±0.000)	0.639 (±0.0000)

Table 7: Performance across training data subsets. Mean values and standard deviations (Std) are calculated over three trials.

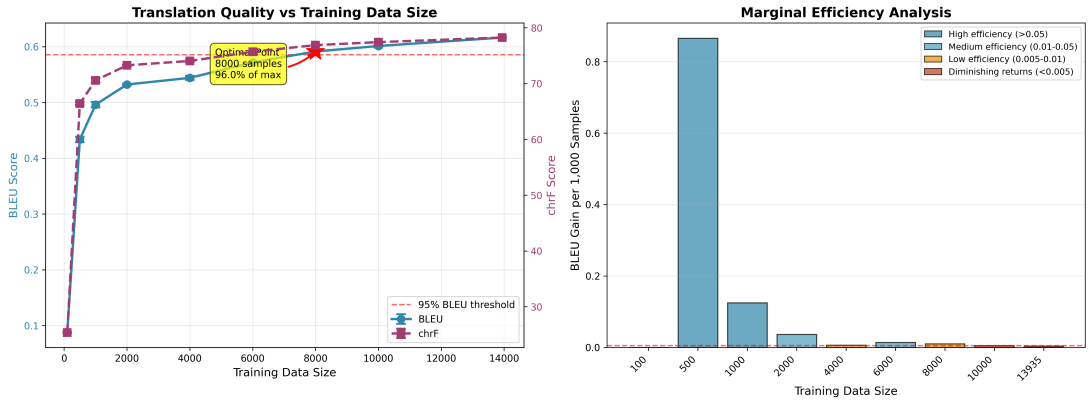

Figure 3: Data Scaling Analysis. Left: BLEU and chrF scores as a function of training data size, with shading indicating standard deviation over three trials. The red dashed line indicates 95% BLEU threshold. Right: Marginal efficiency (BLEU gain per 1,000 samples) with colored bars indicating high, medium, and low efficiency zones.

tematic landscape mapping, and an Optuna-ASHA search for targeted optimization. The two methods required comparable compute (54 vs. 51 GPU hours) and, crucially, converged on the same optimal configuration, providing mutual validation of the results.

6.2.1 Grid Search (2a): Mapping the Hyperparameter Landscape

We first conducted a grid search covering 27 discrete configurations of LoRA rank (r), scaling factor (α), and dropout rate. The objective was not merely to locate a single best-performing setting, but to systematically map the broader hyperparameter landscape and identify dominant and interacting effects among these variables.

Figure 4 visualizes the validation BLEU scores for each (r, α) pair across different dropout levels, while Figure 5 summarizes the aggregated main effects of individual parameters.

From the grid search results, we make three key observations:

- **Scaling factor (α) dominates.** Higher α consistently yields better BLEU scores, with clear improvement from $\alpha = 16$ to $\alpha = 64$.
- **Rank (r) interacts with α.** On average, r has little effect, but the heatmaps show the best single-run BLEU occurs at $r = 8$ with $\alpha = 64$.
- **Dropout has a mild effect.** A dropout of 0.0 produced the highest average BLEU, while larger dropout slightly reduced performance.

The best single-run configuration observed is summarized in Table 8:

r	α	Dropout	BLEU	chrF / Loss
8	64	0.0	57.16	75.46 / 0.7396

Table 8: Grid Search Single-Run Best Configuration.

In summary, the grid search provides a useful but noisy overview of the parameter space. Since each setting was evaluated only once, results remain sensitive to stochastic variation and do not ensure stability.

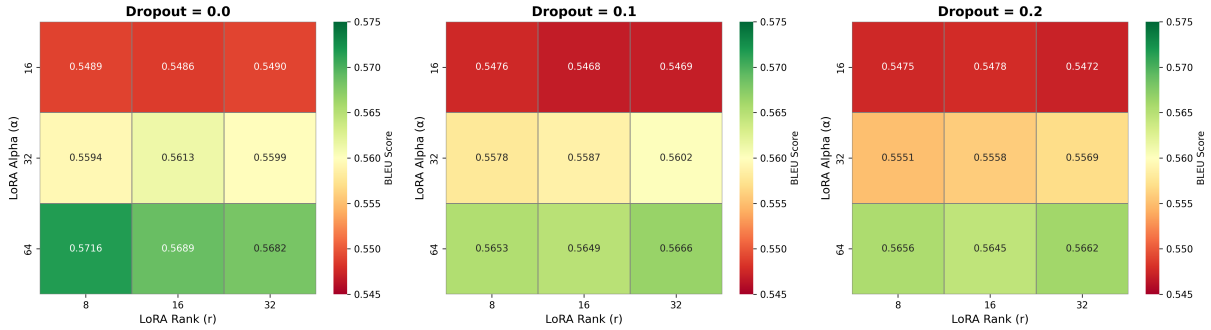

Figure 4: Hyperparameter Sensitivity Heatmaps from the Grid Search. Each subplot shows the validation BLEU score for combinations of LoRA Rank (r) and LoRA Alpha (α) at a fixed dropout rate. Warmer colors (green) indicate higher performance. The consistently strong results for $\alpha = 64$ are clearly visible across all settings.

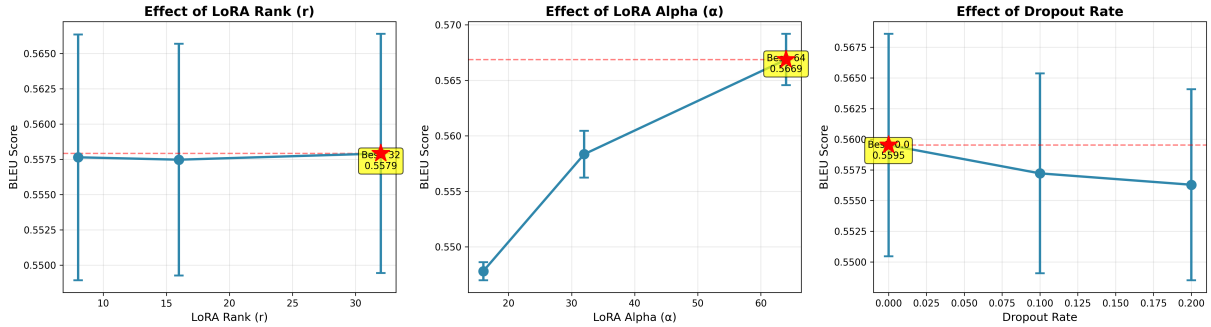

Figure 5: Main Effects of LoRA Hyperparameters on Validation BLEU. Each panel shows the average BLEU score for one hyperparameter, aggregated across all other settings. Error bars represent the 95% confidence interval. The plots quantify the isolated influence of (a) Rank, (b) Alpha, and (c) Dropout.

6.2.2 Optuna with ASHA Pruning (2b): Efficient and Targeted Optimization

To cross-validate the grid search findings, we conducted an independent Optuna-ASHA search as described in Section 5.2. The Optuna study confirmed the grid search trends. As shown in Figure 6, **LoRA Alpha** (α) dominated (importance = 0.973), while Rank (r) and Dropout had minimal impact (0.019 and 0.008, respectively).

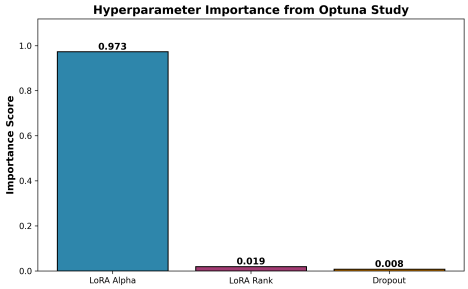

Figure 6: Relative hyperparameter importance from the Optuna study estimated using fANOVA (Hutter et al., 2014). LoRA Alpha (α) dominates, while Rank (r) and Dropout contribute marginally.

To further investigate performance trade-offs, we adopted a **multi-objective optimization** framework to jointly maximize BLEU and chrF scores. Each completed trial is shown in Figure 7, with red points marking non-dominated (Pareto-optimal) solutions and the starred configuration denoting the final model selected based on the harmonic mean of BLEU and chrF.

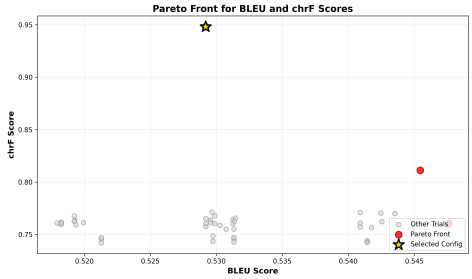

Figure 7: Pareto front of BLEU and chrF scores from the Optuna study. Red points indicate Pareto-optimal solutions, while the starred configuration marks the final selected model.

Stage 2 validation confirmed that the selected configuration ($r = 8, \alpha = 64,$ dropout $= 0.0$) consistently achieved strong performance across three runs with different random seeds, demonstrating stable convergence. The results were in line with the grid search findings. Detailed results are presented in Table 9.

Run	r	α	Dropout	Val BLEU	Val chrF
1	8	64	0.0	60.33	77.64
2	8	64	0.0	60.61	77.76
3	8	64	0.0	60.11	77.51

Table 9: Stage 2 validation results for the selected configuration ($r = 8$, $\alpha = 64$, dropout$= 0.0$) on the 8,000-sample subset. Each run used a different random seed.

In summary, Stage 1 explored the hyperparameter space, producing a Pareto-optimal set of candidates, while Stage 2 validated the top configuration for stability and robustness. The results confirmed consistency with the grid search, providing a well-balanced, computationally efficient configuration for final model training and evaluation.

6.3 Experiment 3: Final Model Performance and Analysis

The final stage involved training the definitive model using the optimal configuration identified in Experiment 2 ($r = 8$, $\alpha = 64$, dropout $= 0.0$). This setup was applied to the full training set of 13,935 English–Norwegian sentence pairs. The resulting model was evaluated on the held-out test set to assess its practical performance against both open-source and commercial baselines.

Table 10 summarizes the performance of our final model against open-source and commercial baselines.

Model	BLEU	chrF++	COMET
Our Fine-tuned Model			
Final Model (Optimized)	**61.48**	**79.19**	0.9298
vs. Open-source Baseline			
NLLB-600M (Zero-shot)	36.86	61.29	0.8814
Δ (absolute)	+24.62	+17.90	+0.048
vs. Commercial Systems			
Microsoft Translator	57.88 (+6.2%)	75.45	**0.9313**
Δ (absolute)	+3.60	+3.74	−0.002
DeepL API	54.53 (+12.8%)	74.29	0.9310
Δ (absolute)	+6.95	+4.90	−0.001
Google Translate	53.50 (+15.1%)	72.75	0.9288
Δ (absolute)	+7.98	+6.44	+0.001
ChatGPT (GPT-4o-mini)	43.23 (+42.2%)	64.93	0.9115
Δ (absolute)	+18.25	+14.26	+0.018

Table 10: Final performance on the NPD EN→NO test set (1,742 segments). Relative BLEU improvements are shown in parentheses.

The optimized model achieved a BLEU score of 61.48, a chrF++ score of 79.19, and a COMET score of 0.9298 on the held-out test set. These results are consistent with the validation outcomes, indicating strong generalization and minimal overfitting. Notably, the model outperforms the open-source baseline by a large margin and reaches or surpasses commercial systems on key lexical and fluency metrics (BLEU and chrF++), while maintaining comparable semantic adequacy (COMET). We note that COMET scores across the top four systems converge within a 0.002 range (0.9288–0.9313), suggesting that semantic adequacy may be near-ceiling for this test set and that BLEU and chrF++ are more discriminative at this performance level. Furthermore, the human error analysis was conducted only on our fine-tuned model; extending it to commercial baselines—particularly Microsoft Translator—would be needed to directly compare critical error rates, and is left for future work.

6.3.1 Comparison with Full Fine-Tuning

To contextualise LoRA's parameter efficiency under controlled conditions, we compare both methods using standard, untuned configurations: LoRA with $r = 16$, $\alpha = 32$, dropout$= 0.1$, lr$= 5 \times 10^{-4}$, and full fine-tuning with lr$= 2 \times 10^{-5}$. Both methods use identical batch size (4) and gradient accumulation (4); neither undergoes dedicated hyperparameter search, ensuring a fair assessment at equivalent tuning effort.

Three patterns emerge from Table 11. At 1,000 samples both methods perform on par ($\Delta = -0.04$), suggesting the adapter bottleneck does not constrain learning in extreme low-resource settings. The gap peaks at 4,000 samples ($\Delta = +0.85$) and narrows consistently thereafter, consistent with full fine-tuning's larger parameter capacity benefiting more from additional data. LoRA also shows lower variance across seeds (e.g. ± 0.02 vs. ± 0.44 at 10,000 samples), indicating greater training stability. The maximum observed gap remains under one BLEU point, confirming LoRA as a viable, parameter-efficient alternative for specialized, resource-constrained domains.

Size	LoRA BLEU	Full FT BLEU	Δ
1,000	49.12 ± 0.12	49.09 ± 0.11	−0.04
2,000	52.98 ± 0.18	53.25 ± 0.11	+0.28
4,000	54.36 ± 0.17	55.22 ± 0.16	+0.85
6,000	57.13 ± 0.30	57.78 ± 0.21	+0.65
8,000	58.65 ± 0.02	59.18 ± 0.03	+0.53
10,000	59.57 ± 0.20	59.64 ± 0.44	+0.07
13,935	61.30 ± 0.15	61.54 ± 0.02	+0.24

Table 11: LoRA vs. Full FT: test BLEU across data sizes (mean ± std over three seeds). Δ = Full FT − LoRA. Both methods use default, untuned configurations; the higher BLEU in Table 10 (61.48) is due to the optimized hyperparameters identified in Experiment 2.

6.4 Human Error Analysis

To complement automatic metrics, we manually reviewed 50 stratified test sentences (2.9% of 1,742) sampled by BLEU score (low/mid/high quality tiers). A domain expert classified errors into eight categories (Word Choice, Terminology, Grammar, Variant Mixing, Numbers/Units, Proper Nouns, Punctuation, Formatting) and three severity levels: Minor (stylistic), Major (clarity reduced), and Critical (meaning distorted). Two annotators labeled the data independently, achieving substantial agreement (Cohen's $\kappa = 0.71$ for error type, $\kappa = 0.61$ for severity) (Cohen, 1960).

6.4.1 Results

41 sentences (82%) contained at least one error, yielding 62 error instances across eight categories (Table 12). Most (83%) were Minor, reflecting acceptable lexical/stylistic variants. However, 12% were Critical—wrong years (2002 vs. 2005), reversed directions (*sørvest* vs. *sørøst*), or incorrect terminology (*nyåpning* "new opening" vs. *gjenåpning* "reopening"). Word Choice (40%) and Variant Mixing (19%) dominated error distribution. BLEU correlated with error presence but failed to detect several Critical semantic errors at high scores, confirming n-gram limitations.

Error Type	**Count**	**%**
Word Choice	25	40.3
Variant Mixing	12	19.4
Grammar	8	12.9
Terminology	7	11.3
Punctuation	4	6.5
Other (Proper Nouns, Formatting, Numbers)	6	9.6
Total	62	100

Table 12: Error type frequency (N=62 instances).

Critical errors altered factual or technical meaning: **Directional error:** model produced *sørvest* (southwest) instead of *sørøst* (southeast), reversing a drilling location. **Terminology confusion:** *nyåpning* (new opening) vs. *gjenåpning* (reopening)—two operationally distinct concepts. **Year errors:** hallucinated year (2002 vs. 2005) or generic temporal reference (*i fjor*) instead of a specific year (*i 2006*). **Nonsense term:** *sokkelhisten* instead of *sokkelskråningen*. Error examples for Critical, Major, and Minor categories are in Appendix D.

6.4.2 Implications

Human evaluation is essential for detecting domain-critical errors that automatic metrics miss. Recommended mitigations include constraints on terminology, control of linguistic variety, and protection of numeric tokens.

7 Conclusion

We propose a three-stage, parameter-efficient adaptation of NLLB-600M for English–Norwegian Bokmål petroleum translation. The framework demonstrates that systematic data selection and hyperparameter tuning can yield production-quality domain-specific MT at a fraction of the cost of full fine-tuning, while remaining competitive with leading commercial systems. Code and a live demo are provided in Appendix A.

8 Limitations and Future Work

Scope and Generalisability. Experiments are limited to EN→NO Bokmål petroleum texts from a single source (17,414 pairs); findings may not generalise to other domains or language pairs. Commercial systems may also have been exposed to in-domain data during training, making comparisons indicative rather than controlled. Future work should validate the framework on additional domains (e.g., legal, medical) and investigate continual adaptation strategies for incorporating updated NPD documents without full retraining.

Linguistic and Evaluation Limitations. A small proportion of target sentences contain Nynorsk morphological markers, contributing to 19.4% of observed errors; variant-aware fine-tuning and glossary-constrained decoding are left for future work. Human evaluation was limited to 50 sentences and not applied to commercial systems, precluding direct comparison of critical error rates. Domain-specific metrics capturing terminology consistency and factual correctness remain an open challenge.

9 Ethical Considerations

ELRC_559 is derived from publicly available NPD materials under open licensing and contains no personal data. Model weights are not released; only code and aggregated results are shared. Practitioners using commercial MT APIs for proprietary data should assess associated privacy risks.

References

[Akiba et al.2019] Akiba, Takuya, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2623–2631.

[Artetxe and Schwenk2019] Artetxe, Mikel and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

[Biewald2020] Biewald, Lukas. 2020. Experiment tracking with Weights and Biases. https://www.wandb.com/. Software available from wandb.com.

[Carrión and Casacuberta2025] Carrión, Salvador and Francisco Casacuberta. 2025. Efficient continual learning in neural machine translation: A low-rank adaptation approach. https://arxiv.org/abs/2512.09910. arXiv preprint arXiv:2512.09910.

[Cohen1960] Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

[European Language Resource Coordination (ELRC)2017] European Language Resource Coordination (ELRC). 2017. Bilingual English–Norwegian parallel corpus from the norwegian petroleum directorate website (ELRC_559). https://data.europa.eu/data/datasets/elrc_559?locale=en. Dataset record on data.europa.eu; Updated: 22.12.2017.

[He et al.2022] He, Junxian, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2022. Towards a unified view of parameter-efficient transfer learning. In *Proceedings of the 10th International Conference on Learning Representations*.

[Houlsby et al.2019] Houlsby, Neil, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2790–2799.

[Hu et al.2022] Hu, Edward J., Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *Proceedings of the 10th International Conference on Learning Representations*.

[Hutter et al.2014] Hutter, Frank, Holger H. Hoos, and Kevin Leyton-Brown. 2014. An efficient approach for assessing hyperparameter importance. In *Proceedings of the 31st International Conference on Machine Learning*, pages 754–762.

[Li and Liang2021] Li, Xiang Lisa and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597.

[Li et al.2018] Li, Lisha, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. 2018. Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research*, 18(185):1–52.

[Mangrulkar et al.2022] Mangrulkar, Sourab, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. PEFT: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft.

[NLLB Team et al.2022] NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, and Al Youngblood. 2022. No language left behind: Scaling human-centered machine translation. https://arxiv.org/abs/2207.04672. arXiv preprint arXiv:2207.04672.

[Papineni et al.2002] Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

[Popović2015] Popović, Maja. 2015. chrF: Character n-gram f-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395.

[Ranathunga et al.2023] Ranathunga, Surangika, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rajneet Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, 55(11):1–37.

[Rei et al.2022] Rei, Ricardo, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585.

[Sennrich et al.2016] Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96.

[Wolf et al.2020] Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

[Zheng et al.2024] Zheng, Jiawei, Hanghai Hong, Feiyan Liu, Xiaoli Wang, Jingsong Su, Yonggui Liang, and Shikai Wu. 2024. Fine-tuning large language models for domain-specific machine translation. https://arxiv.org/abs/2402.15061. arXiv preprint arXiv:2402.15061.

A Resources

All code, training scripts, and evaluation pipelines are publicly available at https://github.com/Entropyobserver/lora-nmt-petroleum. An interactive translation demo of the final model is accessible at https://huggingface.co/spaces/entropy25/mt.

B Open Quality Score (OQS 2.0) Formulation

To provide a reproducible quantitative assessment of parallel corpus quality, we adopt the *Open Quality Score (OQS 2.0)*. This metric evaluates multiple aspects of alignment, completeness, diversity, and domain relevance, producing a scalar in the range $[0, 1]$.

B.1 Formal Definition

Let the parallel corpus be denoted as $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where x_i and y_i are source and target sentences, and N is the total number of sentence pairs. The overall OQS is defined as the arithmetic mean of six normalized dimensions $s_k \in [0, 1]$:

$$\text{OQS}(\mathcal{D}) = \frac{1}{6} \sum_{k=1}^{6} s_k \tag{1}$$

Higher values indicate better corpus quality.

B.2 Quality Dimension Derivations

The six quality dimensions are rigorously defined as follows:

1. **Semantic Alignment (s_1).** Measured as the mean cosine similarity between source and target LASER embeddings (Artetxe and Schwenk, 2019):

$$s_1 = \frac{1}{N} \sum_{i=1}^{N} \cos\big(E(x_i), E(y_i)\big) \tag{2}$$

2. **Completeness (s_2).** Fraction of sentence pairs with non-empty source and target:

$$s_2 = \frac{1}{N} \sum_{i=1}^{N} I(|x_i| > 0 \text{ and } |y_i| > 0) \tag{3}$$

where $I(\cdot)$ is the indicator function.

3. **Source Uniqueness (s_3), Target Uniqueness (s_4), Pair Uniqueness (s_5).** One minus the duplication rate at source, target, and pair levels:

$$s_3 = \frac{|\text{unique}(\{x_i\})|}{N} \tag{4}$$

$$s_4 = \frac{|\text{unique}(\{y_i\})|}{N} \tag{5}$$

$$s_5 = \frac{|\text{unique}(\{(x_i, y_i)\})|}{N} \tag{6}$$

4. **Domain Relevance (s_6).** Fraction of petroleum-domain terms from a curated lexicon L that appear in the corpus:

$$s_6 = \frac{|\{t \in L \mid \exists (x, y) \in \mathcal{D},\ t \in x \cup y\}|}{|L|} \tag{7}$$

Each term in L counts as matched if it occurs in at least one sentence pair.

All six sub-scores are normalized to $[0, 1]$, with higher values indicating better corpus quality. Equation (1) then aggregates them into a single OQS score, suitable for corpus profiling and comparison across datasets.

B.3 Notes for Reproducibility

- s_1 uses pre-computed LASER embeddings; corpus-level mean is applied.
- s_2 checks non-empty tokens after tokenization consistent with model preprocessing.
- s_3–s_5 count exact string matches after lightweight normalization (lowercasing, whitespace trimming).

- s_6 requires a fixed domain lexicon; in our study, it was curated from NPD resources.

Using this formulation, one can reproduce the OQS 2.0 reported in the main text (**OQS = 0.976** for our raw EN–NO corpus).

C Baseline System Details

C.1 Commercial MT Systems

Microsoft Translator (Azure) Neural MT service with broad language coverage and enterprise features such as terminology glossaries and Custom Translator. We use the default quality-oriented setting without custom glossaries. Strength: stable adequacy and terminology consistency. Limitation: domain adaptation requires additional resources.

DeepL API High-quality MT system optimized for European languages. Default settings without glossaries are used. Strength: fluent and natural outputs. Limitation: domain-specific terminology and units may drift.

Google Cloud Translation (NMT v3) General-purpose neural MT system with wide language coverage. Strength: robustness and scalability. Limitation: weaker terminology consistency in specialized petroleum texts.

OpenAI GPT-4o-mini General-purpose LLM used for zero-shot translation via prompting. Strength: good semantic adequacy. Limitation: non-MT-specific decoding; prone to paraphrasing and inconsistent handling of numbers and units.

C.2 Open-source Model Quick Screening

To obtain a low-cost preliminary ranking, we evaluate five multilingual models on a randomly sampled 100-sentence subset of the test data. This step also serves as a sanity check for tokenization and language tag handling.

Model	BLEU	chrF++	COMET
NLLB-200-1.3B	42.20	66.75	0.8951
M2M100-1.2B	39.18	63.56	0.8816
NLLB-200-distilled-600M	37.46	62.85	0.8768
OPUS-MT	34.40	60.75	0.8576
M2M100-418M	31.67	57.82	0.8178

Table 13: Quick screening results on a 100-sentence subset.

C.3 Model Descriptions

NLLB-200-1.3B (Meta) A 1.3B-parameter multilingual model covering 200 languages with explicit language tags. Strong zero-shot performance but high computational cost.

M2M100-1.2B (Meta) Many-to-many MT model trained on 100 languages, enabling direct non-English translation.

NLLB-200-distilled-600M (Meta) A distilled 600M-parameter variant of NLLB-200, offering a favorable performance–efficiency trade-off.

OPUS-MT (Helsinki-NLP) MarianMT-based model targeting North Germanic languages (`en-gmq`).

M2M100-418M (Meta) Smaller M2M100 variant used as a lower-bound baseline.

D Error Analysis Examples

D.1 Critical Errors

Critical errors altered factual content or technical meaning, with potential operational consequences in petroleum documentation:

Nonsense term (Sample #4): Model produced *sokkelhisten* instead of *sokkelskråningen* (shelf slope), rendering the geological reference incomprehensible.

Terminology confusion (Sample #12): *Nyåpning* (new opening) vs. *Gjenåpning* (reopening)—two operationally distinct concepts with different regulatory implications in petroleum licensing.

Directional error (Sample #16): Model produced *sørvest* (southwest) instead of *sørøst* (southeast), reversing the drilling location relative to the Sleipner Øst field.

Generic temporal reference (Sample #37): *i fjor* (last year) instead of the specific *i 2006*, losing precision required in regulatory production reports.

Year hallucination (Sample #38): Model produced year 2002 instead of the correct 2005, introducing a factually incorrect temporal reference in a production volume statement.

D.2 Major Errors

Major errors preserved core meaning but reduced professional quality:

Variant Mixing + Terminology (Sample #3): Mixed Bokmål (*ble boret*, *havdyp*) and Nynorsk

(*vart bora*, *vanndjubde*) forms, with imprecise geological term *sen paleozo* instead of *sein-paleozoisk*. Comprehensible but stylistically inconsistent.

Grammar + Word Choice (Sample #14): Non-idiomatic phrase (*bak vellykket på*), wrong word choice (*arrangementer* vs. *begivenheter*), and missing punctuation reduced fluency and clarity.

D.3 Minor Errors

Minor errors were primarily lexical or stylistic variations not affecting comprehension:

Word Choice (Sample #1): *Konglomeratzonen* vs. *Den konglomeratiske sonen*—both correct; reference slightly more formal.

Variant Mixing (Sample #9): Mixed Norwegian varieties (*millioner* + *lågare* vs. *millionar* + *mindre*); comprehensible despite inconsistency.

Punctuation (Sample #8): Missing comma after *Barentshavet*; trivial formatting difference.

D.4 Perfect Translations

Nine samples (18%) achieved BLEU=1.0, typically shorter or formulaic sentences with standard terminology:

Sample #5: "Contact in the NDP Øystein Dretvik, tel." → *Kontaktperson i Oljedirektoratet Øystein Dretvik, tlf.*

Sample #7: "The West Alpha is currently drilling well 6406/6-2 for Shell in the Norwegian Sea." → *West Alpha borer for tiden brønn 6406/6-2 for Shell i Norskehavet.*