MT Coursework A2: Neural Translation Models

A Comparative Evaluation of Encoder–Decoder and Causal LLMs on OPUS-100 (EN–FR)

Student: Giuseppe Di Palma

January 27, 2026

1 Introduction and Dataset

This report presents a systematic evaluation of different neural machine translation (MT) paradigms on the English-to-French subset of the **OPUS-100** dataset. The analysis contrasts *specialized encoder–decoder models* (NLLB and mBART) with *general-purpose causal language models* (Llama-2 and OPT), focusing on three main axes:

- architectural differences,
- the effectiveness of parameter-efficient fine-tuning (PEFT),
- the impact of decoding strategies.

The objective is not only to compare absolute performance, but also to explain *why* certain models and training strategies are better suited for machine translation.

1.1 Dataset and Computational Constraints

Experiments were conducted on the English–French portion of `Helsinki-NLP/opus-100`. Due to limited computational resources, the following constraints were imposed:

- **Maximum sequence length:** 16 tokens for encoder–decoder models and 24 tokens for causal LMs.
- **Quantization:** All models were loaded using 4-bit NormalFloat quantization (NF4), significantly reducing memory usage while preserving most of the representational capacity.

Although these constraints limit absolute performance, they allow for a fair *relative* comparison across models under identical hardware conditions.

1.2 Evaluation Metrics

Four complementary metrics were used:

- **BLEU**, measuring n-gram precision and favoring surface-level overlap;
- **COMET**, a neural metric estimating semantic adequacy and fluency;
- **chrF**, based on character n-grams and robust to morphological variation;
- **ROUGE-L**, measuring longest common subsequence similarity.

The combination of lexical and semantic metrics is particularly important when evaluating large language models, where semantic correctness may not align with exact n-gram matching.

2 Model-Specific Data Processing Pipelines

Although all experiments are conducted on the same OPUS-100 EN–FR data, the **input construction, tokenization, supervision signal, and generation constraints** differ substantially between encoder–decoder models (NLLB, mBART) and causal language models (Llama-2, OPT). These differences are not mere implementation details, but structural consequences of the underlying architectures, and they strongly influence both training stability and evaluation outcomes.

2.1 Encoder–Decoder Models: NLLB and mBART

NLLB and mBART operate in a classical sequence-to-sequence framework, where **source and target sentences are encoded separately** and aligned through cross-attention. This enables a clean separation between conditioning context and prediction targets.

Language-Aware Tokenization Both models rely on explicit language identifiers:

- **NLLB** uses ISO-style language codes (e.g. `eng_Latn`, `fra_Latn`);
- **mBART** uses learned language tokens (e.g. `en_XX`, `fr_XX`).

During inference, the target language is **hard-constrained** via `forced_bos_token_id`, ensuring that decoding always begins in French. This prevents language drift and removes ambiguity, a guarantee unavailable to causal LMs.

Supervision Signal Training samples are constructed as:

$$\texttt{encoder input} = x_{\text{EN}}, \qquad \texttt{decoder target} = y_{\text{FR}}$$

with loss computed exclusively on the decoder side. Padding tokens are ignored automatically, and no masking of the source sequence is required.

Length Control and Filtering Both source and target sequences are truncated to a maximum of 16 tokens. Examples exceeding this limit are removed, ensuring:

- bounded memory usage under 4-bit quantization;
- consistency between training and inference;
- stable beam search behavior.

Overall, the preprocessing pipeline for NLLB and mBART is **translation-native**, directly optimizing the conditional likelihood $p(y \mid x)$ with explicit alignment and language control.

2.2 Causal Language Models: Llama-2 and OPT

Llama-2 and OPT are decoder-only architectures originally trained for next-token prediction. As a result, translation must be reformulated as a **prompted text generation task**, fundamentally altering the data pipeline.

Prompt-Based Input Construction Each example is serialized into a single sequence:

```
Translate from English to French:
English: <source sentence> = French:
```

For few-shot evaluation, multiple demonstration pairs are concatenated in-context, followed by the test source sentence. This leads to:

- variable prompt length depending on the number of shots;
- aggressive truncation when context exceeds the token budget;
- sensitivity to example ordering and formatting.

Training-Time Label Masking When fine-tuning OPT, the input consists of:

$$[\text{prompt} \parallel \text{target}]$$

but the loss is computed **only on target tokens**. Prompt tokens are masked using `-100` labels. Unlike encoder–decoder models, the causal LM must implicitly learn where the translation begins and when to terminate generation. To enforce stopping, a custom `<END>` token is introduced.

Padding and Post-Processing Causal LMs use **left padding** to preserve autoregressive semantics. Generated outputs contain both prompt and prediction, requiring explicit post-processing to remove the prompt and truncate the output before evaluation. This additional step introduces variability absent in encoder–decoder pipelines.

2.3 Implications for Comparative Evaluation

These preprocessing differences explain several empirical observations:

- Encoder–decoder models achieve higher BLEU and chrF due to explicit alignment and bidirectional conditioning.
- Causal LMs exhibit higher variance across prompts and decoding strategies, particularly under tight length constraints.
- Fine-tuning benefits encoder–decoder models more consistently, as their supervision signal is cleaner and task-aligned.

3 Specialized Encoder–Decoder Models

3.1 NLLB-200 Analysis

NLLB is a large-scale multilingual translation model explicitly trained for MT. Table 1 reports results under greedy and beam search decoding.

Status	LR	Decoding	BLEU	COMET	chrF	ROUGE-L	ΔBL	ΔCM
Baseline	-	Greedy	26.77	72.33	50.16	0.490	-	-
Finetuned	5e-5	Greedy	29.72	73.46	51.60	0.513	+2.95	+1.13
Finetuned	1e-4	Greedy	29.81	73.46	51.86	0.514	+3.03	+1.13
Baseline	-	Beam 4	28.54	73.36	51.14	0.503	-	-
Finetuned	5e-5	Beam 4	30.74	74.09	52.28	0.520	+2.20	+0.73
Finetuned	1e-4	Beam 4	**30.78**	**74.12**	**52.34**	**0.520**	+2.24	+0.75

Table 1: NLLB-200 Results

Fine-tuning consistently improves all metrics. Interestingly, the relative BLEU gain is larger under greedy decoding, suggesting that fine-tuning improves the local quality of token predictions, reducing reliance on exhaustive search.

3.2 mBART Analysis

mBART is evaluated using IA3 and LoRA.

Status	PEFT	Decoding	BLEU	COMET	chrF	ROUGE-L	ΔBL	ΔCM
Baseline	-	Greedy	25.79	71.05	48.53	0.478	-	-
Finetuned	IA3	Greedy	27.38	72.04	49.28	0.489	+1.59	+0.99
Finetuned	LoRA	Greedy	28.92	72.75	50.65	0.506	+3.13	+1.69
Baseline	-	Beam 4	26.67	72.16	49.38	0.488	-	-
Finetuned	IA3	Beam 4	28.41	72.71	50.18	0.502	+1.74	+0.55
Finetuned	LoRA	Beam 4	**29.46**	**73.29**	**51.20**	**0.515**	+2.78	+1.13

Table 2: mBART Results

LoRA consistently outperforms IA3, confirming that low-rank updates provide a richer adaptation of the attention mechanism and better semantic preservation.

4 Causal Language Models

4.1 Llama-2 Prompt-Based Translation

Llama-2 relies entirely on in-context learning.

Performance is highly sensitive to prompting. Increasing the number of shots does not monotonically improve results; 5-shot prompting degrades performance due to context dilution, while 10-shot prompting improves semantic adequacy as measured by COMET.

4.2 OPT-1.3B Fine-Tuning

OPT requires fine-tuning to perform translation.

Lower learning rates are crucial for stability. LoRA at 5×10^{-5} achieves the best performance, outperforming Llama-2 prompting despite a much smaller model size.

Configuration	Beams	BLEU	COMET	chrF	ROUGE-L
Llama 1-shot	1	8.44	58.82	33.44	0.264
Llama 1-shot	4	**12.90**	62.16	**39.25**	**0.332**
Llama 5-shot	1	5.80	54.18	26.92	0.198
Llama 5-shot	4	10.97	58.60	34.74	0.285
Llama 10-shot	1	6.63	59.45	28.44	0.225
Llama 10-shot	4	12.88	**63.62**	38.02	0.324

Table 3: Llama-2 Prompting Results

PEFT	LR	Beams	BLEU	COMET	chrF
LoRA	5e-5	1	13.82	63.85	35.42
LoRA	5e-5	4	**17.33**	**66.97**	**38.96**
LoRA	1e-4	1	11.36	61.36	33.06
LoRA	1e-4	4	15.67	65.65	36.80
IA3	5e-5	1	8.48	57.64	28.75
IA3	5e-5	4	12.26	62.42	32.19
IA3	1e-4	1	7.03	55.51	27.03
IA3	1e-4	4	11.04	60.21	30.69

Table 4: OPT-1.3B Results

5 Conclusions and Discussion

This work presented a controlled comparison between encoder–decoder and causal language models for English–French translation on OPUS-100, under identical data and computational constraints. The results show that translation quality is not primarily determined by model scale, but by the alignment between architecture, training objective, and data processing pipeline. Models explicitly designed for sequence-to-sequence learning consistently achieve higher and more stable performance than general-purpose causal LMs.

Encoder–decoder models benefit from structural properties that are inherently well matched to translation: separate source and target representations, cross-attention, explicit language control, and clean supervision on the decoder side. In contrast, causal LMs must approximate translation through prompted next-token prediction, which introduces ambiguity in source–target boundaries, termination behavior, and sensitivity to prompt design. These factors lead to higher variance and weaker surface-level accuracy, even when semantic adequacy is partially captured.

Parameter-efficient fine-tuning substantially improves performance across all models, with LoRA consistently outperforming IA3, especially for attention-based architectures. Decoding strategies also matter, particularly for causal LMs, where beam search is often necessary to obtain coherent outputs under tight length constraints. Overall, this study confirms that, under realistic resource limits, specialized encoder–decoder models remain the most effective and reliable choice for neural machine translation, while causal LMs are better viewed as flexible but suboptimal approximations of the task rather than direct replacements.