Irish-based Large Language Model with Extreme Low-Resource Settings in Machine Translation

Khanh-Tung Tran
University College Cork
Ireland
123128577@umail.ucc.ie

Barry O'Sullivan*
University College Cork
Ireland
b.osullivan@cs.ucc.ie

Hoang D. Nguyen
University College Cork
Ireland
hn@cs.ucc.ie

Abstract

Large Language Models (LLMs) have demonstrated exceptional performances in a wide range of natural language processing tasks. However, their success does not always extend to machine translation, particularly in challenging scenarios such as translating low-resource languages. This study investigates the multilingual capability of LLMs, with a case study on Irish, an extremely low-resource language, focusing on translation tasks between English and Irish. We propose a dynamic, efficient language adaptation framework for English-centric LLMs, which involves layer-specific adjustments and subsequent fine-tuning for machine translation. Our findings highlight several key insights: (1) different layers in the LLM serve distinct functions such as language understanding and task reasoning, (2) effective translation requires extensive pre-training on both source and target languages, and (3) targeted fine-tuning for machine translation leads to significant improvements of 36.7% for English to Irish and 133.4% for Irish to English compared to the previous state-of-the-art.

1 Introduction

Large Language Models (LLMs) have recently revolutionized the field of Natural Language Processing (NLP), demonstrating remarkable performance across a wide range of tasks. These models, built on the transformer architecture, leverage vast amounts of data to achieve exceptional levels of linguistic understanding. However, significant challenges remain, particularly in the domain of machine translation for low-resource languages (Bawden and Yvon, 2023). Traditional approaches for Neural Machine Translation (NMT) are often data-inefficient and rely on large numbers of parallel data pairs to obtain reliable performance, limiting their applicability in low-resource tasks (Ranathunga et al., 2023; Lamar and Kaya, 2023).

This paper seeks to explore the multilingual capabilities of LLMs, specifically focusing on Irish, an extremely low-resource language, and the translation tasks between English and Irish. Irish, classified as an endangered language, poses unique challenges for machine translation. The limited availability of parallel corpora (Lankford et al., 2022; Ojha et al., 2021) and the sparse representation in pre-training datasets (Barry et al., 2022; Tran et al., 2024) make it a vital candidate for investigating the potential of LLMs in low-resource settings. LLMs, such as ChatGPT (OpenAI, 2022, 2024), BLOOM (Workshop et al., 2023), and the Llama series (Touvron et al., 2023a,b), are predominantly English-centric, although pre-trained on multilingual datasets. The extent to which these models can effectively translate between low-resource languages remains an open question.

Our research identifies several key insights to successfully apply LLM to the low-resource scenario, such as the requirement for the LLMs to be bilingual through extensive pre-training on both languages. We propose a novel framework for efficiently adapting English-centric LLMs to a novel unseen language, and further fine-tuning for the task of machine translation. Our approach involves a two-stage training process: dynamic continued pre-training, where we selectively train layers of the LLM based on their language capability, indicated by retrieval scores, and additional fine-tuning on specific machine translation datasets. By focusing on the layers responsible for language understanding and reasoning, we aim to enhance the bilingual capabilities of the LLM while being efficient, requiring only a fraction of the model's total parameters for effective language adaptation. Specifically, we achieve an improvement of up to 46.14 BLEU score for Irish to English translation and 13.22 BLEU score for English to Irish transla-

*Corresponding Author

tion compared to previous state-of-the-art methods on the LoResMT-2021 dataset (Ojha et al., 2021).

Our source code and model weights are made publicly available at https://github.com/ReML-AI/UCCIX for future research and benchmarking purposes.

2 Related Work

2.1 Neural Machine Translation

Neural Machine Translation (NMT) has become the dominant approach in the field of machine translation, largely due to the success of sequence-to-sequence models and the introduction of attention mechanisms. The advent of the Transformer model (Vaswani et al., 2017) offers a more efficient and scalable architecture that relies entirely on attention mechanisms. Transformers have become the backbone for most state-of-the-art (SoTA) NMT systems (Lankford et al., 2021; Team et al., 2022). Despite these advancements, NMT systems still struggle with translating low-resource languages due to the lack of sufficient training data. Various approaches have been proposed to mitigate this issue, such as transfer learning (Zoph et al., 2016; Chen and Abdul-mageed, 2023) and multilingual NMT (Johnson et al., 2017; Dabre et al., 2020). These methods leverage information from high-resource languages or use monolingual data to further improve translation quality for low-resource languages, but significant challenges remain. One of the main challenges is the dependence on parallel data, which is notably deficient for low-resource languages.

In this work, we explore a recent paradigm (Workshop et al., 2023; Bawden and Yvon, 2023), by applying LLMs to the domain of NMT. We leverage the vast amount of pre-training conducted for LLMs and investigate whether their capabilities can be transferred to NMT tasks, particularly for low-resource languages. However, it should be noted that while LLMs can be pre-trained on multiple languages, their pre-training data is mostly monolingual per sample, making it uncertain how well the models can translate across languages.

2.2 Large Language Models

LLMs have garnered attention for their impressive text generation capabilities and versatility across various NLP tasks. However, most of these models, either closed-source ones such as ChatGPT, or open-source models like BLOOM (Workshop et al., 2023), and the Llama series (Touvron et al., 2023a,b) have demonstrated significant proficiency in handling a variety of languages and tasks. However, these models predominantly focus on widely spoken languages like English, leading to a performance disparity when applied to low-resource languages. Recent surveys (Bawden and Yvon, 2023; Hendy et al., 2023) have investigated the capability of LLMs for machine translation tasks, reporting that LLMs can perform well in these scenarios, especially for high-resource languages. However, their effectiveness in low-resource settings, such as in Irish, remains limited due to the lack of adequate training data.

UCCIX (Tran et al., 2024) is a recent LLM developed with a focus on Irish, a Definitely Endangered language as recognized by UNESCO (UNESCO, 2010). Given the limited availability of Irish data, the authors proposed a framework for language adaptation of an English-centric LLM to make it bilingual. Despite this, issues such as catastrophic forgetting of English have been observed, as a consequence of continued pre-training on Irish data.

With the case study on Irish, we investigate the potential usages of such models for the translation tasks between Irish and English, as part of the effort to preserve the Irish language and prevent its loss. We analyze the bilingual capabilities of the LLM and propose an adaptive language adaptation strategy to balance the model's performance between the two languages. This approach aims to enhance the efficiency of adapting LLMs in low-resource settings, ensuring robust performance in both high-resource and low-resource languages.

2.3 Low-Resource Settings

Research on the challenges of addressing low-resource languages in NLP is essential given the diversity of languages and the demand for inclusive technology. Since large annotated datasets are necessary for training strong models, low-resource languages frequently lack them, making it challenging to achieve good performance using traditional techniques. As pointed out in recent survey (Ranathunga et al., 2023), if there are less than 0.5 million parallel sentences in the parallel corpora, a language pair is deemed "low-resource" in an MT scenario, and if there are less than 0.1 million parallel sentences, it is deemed "extremely low-resource".

Irish, an endangered language, fits into the 'extremely low-resource' category. Recent works report a composite dataset from different sources amounting to only 25,000 (Lankford et al., 2022) or 52,000 (Lankford et al., 2021) parallel sentences. Given this limited amount of data, we investigate whether the large amount of available monolingual data can aid in improving performance through the use of LLMs. Our findings highlight the effectiveness of further fine-tuning LLMs for the machine translation task, even with such sparse data.

3 Method

3.1 Preliminary Explorations

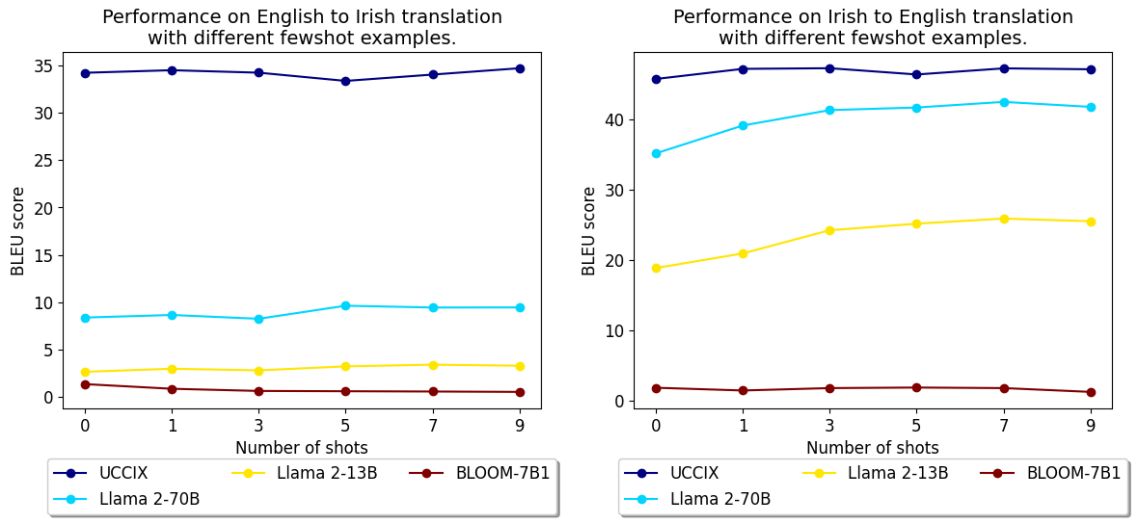
Large Language Models (LLMs) are typically built using the transformer decoder-only architecture, consisting of multiple stacked transformer layers. While LLMs are often trained on large-scale English-dominant text corpora, they often also include a small percentage of texts in multiple languages due to the vast size of the training data. This raises the question of whether LLMs can understand underpresented languages effectively. For instance, in the Llama series of models, the Irish language constitutes less than 0.005% of the training corpus. To explore this, we conduct few-shot prompting experiments with the machine translation task between English (dominant language) and Irish (extremely low-resource language). Few-shot prompting allows LLMs to follow specific input patterns and leverage their pre-trained knowledge for the translation task. We investigate the performance in both directions: Irish to English (assessing LLM's understanding of the low-resource language) and English to Irish (analyzing its capability to generate text in the target language). Table 1 shows examples of the prompts used. The results, presented in Table 2 and Figure 1, highlight the following insights:

- English-centric LLMs may have some understanding of low-resource languages but struggle with text generation in those languages. This is evident by the strong performance of the Irish to English direction, with gpt-3.5-turbo and Llama 2-70B able to outperform the previous task-specific SoTA (Lankford et al., 2021), up to 7.97 BLEU score.
- Efficient translation requires extensive pretraining on both languages, as evidenced by UCCIX outperforming English-centric LLMs, beating the much larger gpt-3.5-turbo model.
- Generally, providing examples (few-shot prompting) helps LLMs follow the task format better (Figure 1), aligning with previous findings (Brown et al., 2020).

To further investigate LLM behavior without relying on few-shot prompting and the variants it created, we analyze the sentence retrieval task. The sentence retrieval task (Artetxe and Schwenk, 2019; Dufter and Schütze, 2020; Yong et al., 2023), aims to identify the closest sentence in English given a representation of a sentence in a new language (Irish). We compute sentence retrieval accuracy at each layer of different pretrained models to understand where and how language understanding capabilities emerge. For this analysis, we focus on the Llama 2 model, a popular and widely-used open-sourced LLM.

Prompt English->Irish	Prompt Irish->English
Aistrigh Béarla go Gaeilge: Béarla: When needed, EMA and the other European regulators take action. Gaeilge: Gníomhaíonn EMA agus rialtóirí Eorpacha eile nuair is gá. Béarla: mumps Gaeilge: na leicní Béarla: woman holding a phone Gaeilge: bean a bhfuil fón aici Béarla: 17 June 2020 – EU COVID-19 vaccines strategy unveiled Gaeilge: 17 Meitheamh 2020 – Straitéis an Aontais um vacsaíní in aghaidh COVID-19 curtha i láthair Béarla: 31 August 2020 - Coronavirus Global Response: The Commission joins the COVID-19 Vaccine Global Access Facility Gaeilge: 31 Lúnasa 2020 – An Fhreagairt Dhomhanda ar an gCoróinvíreas: An Coimisiún páirteach sa tSaoráid Rochtana Domhanda ar Vacsaíní in aghaidh COVID-19 Béarla: {input} Gaeilge:	Aistrigh Gaeilge go Béarla: Gaeilge: Gníomhaíonn EMA agus rialtóirí Eorpacha eile nuair is gá. Béarla: When needed, EMA and the other European regulators take action. Gaeilge: na leicní Béarla: mumps Gaeilge: bean a bhfuil fón aici Béarla: woman holding a phone Gaeilge: 17 Meitheamh 2020 – Straitéis an Aontais um vacsaíní in aghaidh COVID-19 curtha i láthair Béarla: 17 June 2020 – EU COVID-19 vaccines strategy unveiled Gaeilge: 31 Lúnasa 2020 – An Fhreagairt Dhomhanda ar an gCoróinvíreas: An Coimisiún páirteach sa tSaoráid Rochtana Domhanda ar Vacsaíní in aghaidh COVID-19 Béarla: 31 August 2020 - Coronavirus Global Response: The Commission joins the COVID-19 Vaccine Global Access Facility Gaeilge: {input} Béarla:

Table 1: 5-shot prompts used to evaluate pre-trained LLMs on machine translation tasks.



(a) English to Irish translation

(b) Irish to English translation

Figure 1: Effects of a number of fewshot examples during prompting for different models on a) English to Irish translation, and b) Irish to English translation on the LoResMT-2021 dataset.

Model	**BLEU on English->Irish**	**BLEU on Irish->English**
SoTA from LoResMT2021 (Lankford et al., 2021)	**36.0**	34.6
gpt-3.5-turbo	18.64	42.57
Llama 2-70B	9.63	41.66
Llama 2-13B	3.25	25.60
BLOOM-7B1	0.61	1.84
UCCIX	33.34	**46.36**

Table 2: BLEU scores on machine translation tasks for baseline NMT model (Lankford et al., 2021), English-Irish bilingual LLMs (UCCIX (Tran et al., 2024)), and other LLMs.

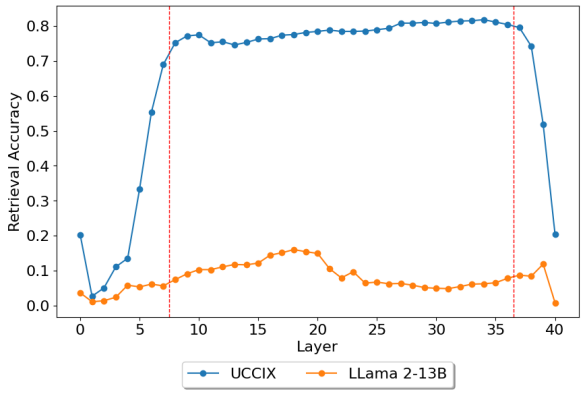


Figure 2: Sentence retrieval accuracy score across layers of UCCIX, English-Irish bilingual LLM and Llama 2-13B, English-centric LLM.

Formally, given $D = \{(s_0^{ga}, s_0^{en}), \ldots, (s_i^{ga}, s_i^{en}), \ldots, (s_{N-1}^{ga}, s_{N-1}^{en}), \}$ a dataset of parallel sentences in Irish (denoted s^{ga}) and English (denoted s^{en}), the sentence retrieval task involves finding the closest English sentence given an Irish sentence representation. A generative (decoder-only) LLM processes the input textual data autoregressively through each transformer layer. The text is first tokenized into subword units and mapped into embeddings through a learned embedding matrix. The embeddings are input to transformer layers, maintaining their dimensions throughout the forward pass. Given the initial input embedding at position j as h_0^j, the corresponding output latent embedding at layer l is computed as:

$$h_l^j = f_l(h_{l-1}^0, \ldots, h_{l-1}^j) \quad (1)$$

with f_l as the transformer block at layer l, $l \in [0, L)$ for an LLM with L layers. For instance, Llama 2-13B and the finetuned version UCCIX have $L = 41$ layers. The sentence representation at each layer is calculated as the average over embeddings at all positions:

$$e_l = \frac{1}{K} \sum_{k=0}^{K-1} h_l^k \quad (2)$$

Thus, the retrieval accuracy for sentence i at

layer l is determined by:

$$\text{accuracy}_{l,i} = \begin{cases} 1 & \text{if } \underset{i \in [0,N)}{\text{argmax}} \cos(e^{ga}_{l,i}, e^{en}_{l,i}) = i \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where cos is the cosine similarity between embeddings.

As observed in Figure 2, intermediate layers of UCCIX (visualized as between the 2 horizontal lines) achieve almost perfect retrieval score, a trend that can also be noticed in the base LLaMA 2 model, although to a lesser extent. This leads us to hypothesize that there are two types of layers in the architecture of LLMs: (1) *interface layers*, which consist of the input layer (the first few layers) that analyze the language of the input text, extracting information such as syntax, lexical structure, and the output layer (the last few layers) that map back to the token space of the target languages, and (2) *reasoning layers*, which are the intermediate layers capable of reasoning and performing the task at hand. As the interface layers contain information about the unique characteristics of each language, they fail in retrieving sentences with the same meaning but written in different languages, hence the low retrieval scores.

3.2 Proposed Framework for Dynamic and Efficient Language Adaptation

Building on our insights, we propose a framework for efficiently adapting LLMs to understand additional languages and to the machine translation task. This framework, as depicted in Figure 3, involves two main stages: dynamic continued pre-training for language adaptation and additional fine-tuning on machine translation data.

Based on our preliminary experiments, we hypothesize that certain layers of the LLM function as interface and reasoning layers for bilingual understanding. Thus, we dynamically identify and train only the relevant layers. Specifically, we use the retrieval score accuracy_l for each layer l of the original English-centric LLM to guide this process.

For the input layers, which are part of the interface layers, we select layers from 0 to the first layer that has a retrieval score larger than α_s:

$$l_{input} = \{l \mid 0 \leq l \leq \underset{l}{\text{argmin}}(\text{accuracy}_l > \alpha_s)\} \tag{4}$$

Similarly, for the output interface layers, we select from the last layer and move backward to the first layer that has a retrieval score smaller than α_e:

$$l_{output} = \{l \mid \underset{l}{\text{argmax}}(\text{accuracy}_l < \alpha_e) \leq l < L\} \tag{5}$$

By focusing training on these identified layers, we aim to enhance the LLM's bilingual capabilities by targeting the layers in charge of language understanding while maintaining the reasoning capabilities of the LLM. Moreover, the proposed training strategy is efficient, as it reduces number of layers that require training.

After the dynamic continued pre-training stage, we further fine-tune the LLM on specific machine translation datasets. This step ensures that the model not only understands both languages but also effectively translates between them. The additional training is performed on both directions: English to Irish and Irish to English with full fine-tuning, as both interface layers and reasoning layers are vital to adapt the LLM to this specific task. During this stage, we compute the training loss solely on the target language sentence, and ignore prediction loss on the task prompt and input sentence.

4 Experiments

4.1 Datasets

For language adaptation with continued pre-training, we utilize the monolingual corpus introduced in UCCIX (Tran et al., 2024). The monolingual dataset includes data from various sources such as CulturaX(Nguyen et al., 2023), Glot500 (ImaniGooghari et al., 2023), Irish Wikipedia, providing valuable content from Irish sites and pages. To ensures a fair comparison, we also choose the base pre-trained LLM to be Llama 2-13B, same as UCCIX. The dataset in total has approximately 500M Irish tokens, significantly smaller than the original 2T tokens used to train Llama 2.

For the fine-tuning phase, we combine the corpus of LoResMT (Ojha et al., 2021) and ga-Health (Lankford et al., 2022), both in-domain datasets for the health domain, resulting in a total of 17k samples for the training set.

For MT evaluation, we report the BLEU score, a common metric in the literature, for both translation directions: English to Irish and Irish to English. The evaluation set from LoResMT comprises 500 samples for English to Irish and 250 samples for Irish to English. For the fine-tuning MT data, we

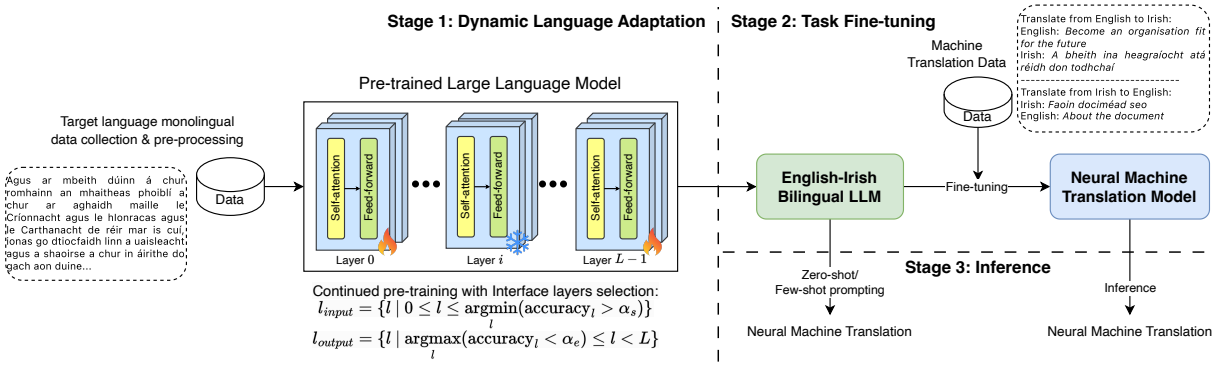


Figure 3: Our main pipeline, including two training stages: 1) dynamic language adaptation of base English-centric to the target language, and 2) further fine-tuning using parallel data for the neural machine translation task.

directly use the corpus from previous works, adhering to the training split of LoResMT to avoid any data contamination issues. Additionally, the pre-training corpus from UCCIX is monolingual, while the evaluation data is parallel between the two languages. Consequently, the problem of data contamination is further mitigated.

4.2 Experimental Setup

For language adaptation with continued pre-training, we start with the base Llama 2-13B model, pre-trained on an English-dominant corpus of 2T tokens. This ensures a fair comparison with the English-Irish bilingual LLM, UCCIX, which also based on Llama 2-13B. In this foundational work, for simplicity, we set both α_s and α_e to 0.075 selecting 11 layers as interface layers for training out of 41 layers in total in Llama 2-13B. This means we train approximately 25% of the total model parameters, ensuring the technique to be efficient, compared to full fine-tuning. Following UCCIX, we also expand the tokenizer to include 10k native Irish tokens before continued pre-training. During this phase, we train the model with the AdamW optimizer for a total of 2 epochs, with a learning rate of $1e-4$, a batch size of 96 samples, each 4096 tokens long. Training is distributed across 6 NVIDIA H100 GPUs with a gradient accumulation step of 8. We leverage DeepSpeed (Rasley et al., 2020) for the training process. The pre-trained LLMs can be used for the machine translation task through few-shot prompting. By default, we design a prompt in Irish, with a description of the task and 5 examples (5-shot prompting), as illustrated in Table 1. The 5 examples are initially randomly chosen from the development subset.

For fine-tuning the machine translation task, we train the model for at most 10 epochs on the training set. We use the AdamW optimizer, setting the learning rate to $1e-4$ for full fine-tuning, and to $1e-3$ for training with parameter efficient methods. These values are chosen through grid search. We also leverage DeepSpeed in this stage, with a batch size of 96 samples, each 4096 token long distributed across 6 H100 GPUs. Each experiment is repeated 3 times across random seeds, and we report the average results for robust evaluation. The model is prompted as illustrated in the right part of Figure 3 for inference.

Model	BLEU on English → Irish	BLEU on Irish → English
Llama 2-13B	3.25	25.60
UCCIX	**33.34**	**46.36**
UCCIX$_{(IA)^3}$	19.53	39.48
UCCIX$_{LoRA}$	26.14	43.65
UCCIX$_{reasoning_layer}$	29.27	42.71
UCCIX$_{interface_layer}$	<u>30.69</u>	<u>46.07</u>

Table 3: Comparison between our framework with other language adaptation techniques: full fine-tuning (UCCIX), parameter-efficient (UCCIX$_{LoRA}$, UCCIX$_{(IA)^3}$) and the ablation study with the training of reasoning layers (UCCIX$_{reasoning_layer}$).

5 Results and Discussion

5.1 Langugage Adaptation Effectiveness

Table 3 demonstrates the efficiency and performance of our proposed dynamic language adaptation approach, UCCIX$_{interface_layer}$. Despite training only 25% of the parameters, our method remains competitive with UCCIX, which employs full fine-tuning. For instance, the performance drop

Model	Acc. on Cloze Test (0-shot)	Acc. on SIB-200 (Irish subset) (10-shot)	Exact-match on IrishQA (ga) (5-shot)	Exact-match on Natural Question (5-shot)	Exact-match on IrishQA (en) (5-shot)	Acc. on Wino-grande (5-shot)	Acc. norm on HellaSwag (10-shot)	Average
gpt-3.5-turbo	N/A	N/A	0.2222	**0.4660**	0.3333	N/A	N/A	N/A
Llama 2-70B	0.63	0.7059	0.2963	0.3806	0.4074	**0.8374**	**0.8701**	0.5897
Llama 2-13B	0.54	0.5343	0.3148	0.3069	0.4444	0.7609	0.8223	0.5319
BLOOM-7B1	0.45	0.1471	0.0000	0.0806	0.1667	0.6519	0.6202	0.3024
UCCIX	**0.75**	0.7794	**0.3889**	0.1668	0.3704	0.7135	0.7758	0.5635
UCCIX$_{(IA)^3}$	0.45	0.7353	0.2222	0.2490	0.4815	0.7435	0.7883	0.5242
UCCIX$_{LoRA}$	0.67	0.7792	0.2963	0.2704	**0.5370**	0.7474	0.7851	0.5836
UCCIX$_{reasoning_layer}$	0.64	0.7304	0.2222	0.1950	0.3889	0.7167	0.7903	0.5262
UCCIX$_{interface_layer}$	0.69	**0.7892**	**0.3889**	0.2404	**0.5370**	0.7451	0.7971	**0.5982**

Table 4: Evaluation results of pre-trained models on curated set of Irish (first 3 columns) and English (the following 4 columns) benchmarking datasets (Tran et al., 2024). We compute *Average* as the mean across all metrics.

is minimal, only 0.29%, for the Irish to English translation task. Compared to other parameter efficient fine-tuning techniques, including LoRA (Hu et al., 2022) and (IA)^3 (Liu et al., 2022), our dynamic interface layers training approach achieves the strongest performance. Additionally, unlike other methods that require injecting additional parameters during training and merging with the original model for efficient inference, our method does not introduce any additional parameters, allowing the model to be ready for use directly after training.

We conduct an ablation study where we train the reasoning layers instead of the interface layers selected in Equation 2 and Equation 3. We denote this as UCCIX$_{reasoning_layer}$. Our approach outperforms this method in both English to Irish and Irish to English translation directions, with a gap of 1.42 and 3.36 BLEU scores, respectively.

To further analyze the bilingual capability of models trained with our proposed approach, both learning the new language and preserving the capability in the original language, we benchmark on the curated set of Irish and English benchmarking datasets introduced in (Tran et al., 2024). This curated set includes diverse tasks such as topic classification and open-ended question answering. The results, as illustrated in Table 4, highlight the balanced performance between the two languages for UCCIX$_{interface_layer}$, where we achieve top-1 average score of 0.5982, surpassing the much larger model Llama 2-70B (0.5897). Our model also achieves SoTA results on 3 out of 7 datasets, namely SIB-200, IrishQA (Irish version), and IrishQA (English version). Furthermore, in benchmarking with Irish tasks, our model performs comparably to UCCIX, which was fully fine-tuned to focus on Irish, while being efficient, as we only trained 25% of the parameters compared to UCCIX. This validates our hypothesis on interface and reasoning layers: fine-tuning interface layers allows the model to understand additional languages without catastrophic forgetting, and freezing the reasoning layers helps maintain the model's usefulness and effectiveness.

Model	BLEU on English → Irish	BLEU on Irish → English
Llama 2-13B-*mt*	31.74	62.52
UCCIX-*mt*	**49.22**	76.44
UCCIX$_{interface_layer}$-*mt*	39.10	**80.74**

Table 5: Fine-tuning results on machine translation tasks for baseline English-centric LLM (Llama 2-13B) and English-Irish bilingual LLMs (UCCIX). Here *-mt* denotes further finetuning for the machine translation task.

5.2 Machine Translation Fine-tuning Result

We carry out further fine-tuning experiments to investigate whether LLMs can be further adapted to this specific task. As shown in the preliminary experiment in Section 3.1, prompting pre-trained LLMs seem to be effective only when they are extensively trained on both source and target languages. In addition to fine-tuning UCCIX and UCCIX$_{interface_layer}$, we also fine-tuned the English-centric LLM Llama 2-13B to investigate whether exposure to a small amount of parallel data can enhance its performance. Results in Table 5 indicate that further fine-tuning helps significantly, with a substantial performance jump for Llama 2-13B, from 3.25 to 34.16 BLEU score for English to Irish, and from 25.60 to 62.52 for Irish to English.

Nevertheless, having a base bilingual pre-trained LLM is important. Fine-tuning results with both UCCIX and UCCIX$_{interface_layer}$ showcase impressive performance gaps. UCCIX-*mt* achieves a new SoTA result for Irish to English translation, with a gap of 13.22 (49.22 compared to 36.0), and on English to Irish task, and UCCIX$_{interface_layer}$-*mt* surpasses the SoTA with a gap of 46.14 BLEU score.

In general, the results convincingly demonstrate the effectiveness of leveraging large-scale pre-trained LLMs for machine translation tasks involving extremely low-resource languages. By further fine-tuning on the limited available parallel data, we can significantly enhance translation performance, even in resource-constrained scenarios.

6 Conclusion

In this work, we investigate the application of LLMs to the domain of neural machine translation, particularly focusing on Irish, an extremely low-resource language. We analyze the bilingual capabilities of LLMs and propose a dynamic language adaptation strategy aimed at balancing the model's performance across multiple languages. We hypothesize that certain layers of the LLM serve as interface layers for language understanding, and reasoning layers, and we develop a novel, efficient training approach that dynamically identifies and trains only the relevant layers. Our experimental results demonstrate the effectiveness of our approach, achieving balanced performance between languages. Moreover, we show that leveraging large-scale pre-trained LLMs and further fine-tuning them on machine translation tasks with limited parallel data can significantly enhance translation performance in resource-constrained scenarios, with performance enhancement up to 46.14 BLEU score. This highlights the potential of our method to improve machine translation tasks involving extremely low-resource languages.

Acknowledgements

We would like to acknowledge CloudCIX Limited for the generous collaborative support of computing resources on their NVIDIA HGX/H100 GPU cluster. This research work has emanated from research conducted with financial support from Science Foundation Ireland under Grant 12/RC/2289-P2 and 18/CRT/6223.

Limitations

In this work, we focus specifically on the Irish language and a bilingual translation scenario. Our experiments are primarily conducted using the LoResMT-2021 dataset for the translation task. While our proposed framework can theoretically be applied to other languages and multilingual scenarios, further experiments are beneficial to verify its generalizability across different languages and diverse datasets.

Ethics Statement

Our work contributes to language technologies that support the digitalization and preservation of low-resource languages, with a particular focus on Irish. We aim to promote linguistic diversity and inclusivity by enhancing the accessibility and usability of endangered languages through advanced machine translation techniques.

References

Mikel Artetxe 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.

James Barry, Joachim Wagner, Lauren Cassidy, Alan Cowap, Teresa Lynn, Abigail Walsh, Mícheál J. Ó Meachair, and Jennifer Foster. 2022. gaBERT — an Irish language model. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4774–4788, Marseille, France. European Language Resources Association.

Rachel Bawden and François Yvon. 2023. Investigating the translation performance of a large multilingual language model: the case of BLOOM. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 157–170, Tampere, Finland. European Association for Machine Translation.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Wei-rui Chen and Muhammad Abdul-mageed. 2023. Improving neural machine translation of indigenous languages with multilingual transfer learning. In *Proceedings of the Sixth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2023)*, pages 73–85, Dubrovnik, Croatia. Association for Computational Linguistics.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A survey of multilingual neural machine translation. *ACM Comput. Surv.*, 53(5).

Philipp Dufter and Hinrich Schütze. 2020. Identifying elements essential for BERT's multilinguality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4423–4437, Online. Association for Computational Linguistics.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation.

Edward J Hu, 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 *International Conference on Learning Representations*.

Ayyoob ImaniGooghari, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Annie Lamar and Zeyneb Kaya. 2023. Measuring the impact of data augmentation methods for extremely low-resource NMT. In *Proceedings of the Sixth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2023)*, pages 101–109, Dubrovnik, Croatia. Association for Computational Linguistics.

Séamus Lankford, Haithem Afli, Órla Ní Loinsigh, and Andy Way. 2022. gaHealth: An English–Irish bilingual corpus of health data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6753–6758, Marseille, France. European Language Resources Association.

Seamus Lankford, Haithem Afli, and Andy Way. 2021. Machine translation in the covid domain: an English-Irish case study for LoResMT 2021. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*, pages 144–150, Virtual. Association for Machine Translation in the Americas.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2023. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages.

Atul Kr. Ojha, Chao-Hong Liu, Katharina Kann, John Ortega, Sheetal Shatam, and Theodorus Fransen. 2021. Findings of the LoResMT 2021 shared task on COVID and sign language for low-resource languages. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*, pages 114–123, Virtual. Association for Machine Translation in the Americas.

OpenAI. 2022. Introducing chatgpt.

OpenAI. 2024. Gpt-4 technical report.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Comput. Surv.*, 55(11).

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 3505–3506, New York, NY, USA. Association for Computing Machinery.

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, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Khanh-Tung Tran, Barry O'Sullivan, and Hoang D. Nguyen. 2024. UCCIX: Irish-eXcellence Large Language Model.

UNESCO, editor. 2010. *Atlas of the world's languages in danger*, 3 edition. Memory of Peoples Series. UNESCO.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

BigScience Workshop et al. 2023. BLOOM: A 176b-parameter open-access multilingual language model.

Zheng Xin Yong, Hailey Schoelkopf, Niklas Muennighoff, Alham Fikri Aji, David Ifeoluwa Adelani, Khalid Almubarak, M Saiful Bari, Lintang Sutawika, Jungo Kasai, Ahmed Baruwa, Genta Winata, Stella Biderman, Edward Raff, Dragomir Radev, and Vassilina Nikoulina. 2023. BLOOM+1: Adding language support to BLOOM for zero-shot prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11682–11703, Toronto, Canada. Association for Computational Linguistics.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.